UNIVERSIDAD DE GRANADA

Departamento de Estadística e Investigación Operativa

REDUCCIÓN DE DIMENSIÓN EN REGRESIÓN LOGÍSTICA FUNCIONAL

TESIS DOCTORAL

Manuel Escabias Machuca

2002

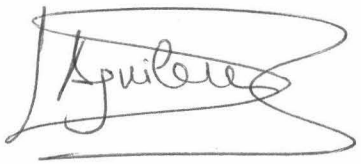
REDUCCIÓN DE DIMENSIÓN EN REGRESIÓN LOGÍSTICA FUNCIONAL

Memoria presentada por:
Manuel Escabias Machuca
para aspirar al Grado
de Doctor.

Doctorando

Fdo: Manuel Escabias Machuca.

V°. B°. de la Directora



Fdo: Dra. Dña. Ana María Aguilera Del Pino.

Departamento de Estadística e
Investigación Operativa

UNIVERSIDAD DE GRANADA

2002

Prólogo

Existen multitud de disciplinas científicas en las que resulta de especial relevancia conocer la probabilidad de ocurrencia de determinados sucesos en función de la ocurrencia de otros. Así por ejemplo en medicina o epidemiología es necesario conocer la probabilidad que tiene uno o un grupo de individuos de padecer determinada enfermedad en función de la incidencia de un conjunto de factores de riesgo. También en meteorología es interesante este hecho pues su objeto de estudio es el de la probabilidad de que se desencadene un determinado fenómeno meteorológico a partir de la información obtenida de la observación de la atmósfera. Las ciencias actuariales también se sirven de la predicción de probabilidades de ocurrencia de sucesos como padecer un accidente de tráfico en términos de los valores de determinadas variables como la edad del conductor, el tipo de vía, etc., para fijar las primas de los seguros. La técnica estadística que analiza situaciones como las descritas es el modelo de regresión logística, cuyo desarrollo sigue proporcionando resultados, sobre todo en medicina, como los obtenidos en los últimos años por Pulkstenis y Robinson (2002) o Paik (2000).

Un problema al que es muy sensible el modelo de regresión logística, y que ya fue puesto de manifiesto en Ryan, (1997), es el de la multicolinealidad o alta dependencia existente entre las covariables del modelo, que hace que no se pueda encontrar solución apropiada a la estimación de los parámetros del mismo. Otro problema que se presenta en regresión logística está relacionado con el principio de parsimonia tan requerido en estadística, y consiste en la necesidad de explicar la variable dependiente del modelo con el menor número de regresores posible. Una propuesta de solución a estos dos problemas será uno de los objetivos de esta tesis.

El análisis en componentes principales (ACP), introducido por Hotelling, es una técnica multivariante que sirve a estos dos problemas de manera admirable. Para ello las componentes principales no son más que un conjunto de variables incorreladas que se obtienen mediante combinaciones lineales de las originales

y que, por el hecho de ser incorreladas, son capaces de evitar problemas como el de la multicolinealidad. Por otro lado el ACP intenta explicar la variabilidad existente en un problema que involucra a un conjunto de variables mediante el menor número de componentes principales posible, con lo que sirve a la reducción de dimensión o criterio de parsimonia indicado anteriormente para los problemas de regresión. Con esta intención introdujo Massy en 1965 los modelos de regresión en componentes principales para el caso lineal, y que nosotros introducimos aquí para el caso de regresión logística mostrando las similitudes y diferencias con aquél.

En otro orden de cosas, en los últimos años se han desarrollado numerosas técnicas referentes a modelizar variables que evolucionan en el tiempo. Históricamente tales situaciones han sido abordadas por modelos de series temporales o modelos de predicción lineal del tipo del filtro de Kalman-Bucy, modelos que, por otro lado, necesitan de restricciones como estacionariedad o la expresión del problema como un modelo en espacio de estados, restricciones muy difíciles de asegurar en determinadas situaciones. Estudios recientes han obtenido metodologías que abordan problemas funcionales dependientes del tiempo con exigencias menos restrictivas que las planteadas anteriormente como es el caso de los trabajos de Deville (1974), Besse (1988) o Saporta (1981) que utilizan el desarrollo de Karhunen-Loève, introducido por Karhunen en 1947, en aplicaciones para las ciencias sociales. Este desarrollo permite expresar las trayectorias de un proceso estocástico como una combinación lineal de funciones determinísticas con coeficientes dados por variables aleatorias incorreladas independientes del tiempo.

El desarrollo de los procesos estocásticos ha permitido la generalización de las técnicas de reducción de dimensión al campo funcional surgiendo el análisis en componentes principales funcional (ACPF) y la aparición de métodos de predicción lineal basados en las mismas como es el caso de los modelos de predicción en componentes principales (PCP) introducidos en Aguilera et al. (1997). Otra generalización al campo funcional de técnicas multivariantes es la del modelo de regresión lineal funcional que permite predecir una variable en términos de la evolución temporal de una magnitud relacionada con ella. Cuando la variable respuesta es dicotómica el modelo lineal no es adecuado para explicar dicha variable a partir de la evolución temporal de otra. Para resolver este problema introduciremos los modelos de regresión logística funcional que es la principal aportación de esta memoria. Además de formular el modelo, proponemos una forma de estimación aproximada del mismo así

como soluciones a los distintos problemas que surgen basadas en análisis de componentes principales funcional y que a continuación se detallan.

Para la presentación de los distintos aspectos de que es objeto la presente memoria se ha dividido la misma en tres capítulos bien diferenciados: en el primero se abordará el problema múltiple, en el segundo el funcional y en el tercero se presentará un estudio de simulación con el que se ilustran los problemas y las soluciones propuestas en los dos primeros.

Como se ha indicado, el primer capítulo aborda el aspecto múltiple de los modelos de regresión logística en componentes principales y se divide en tres partes. La primera trata de la formulación del modelo de regresión logística, la estimación del mismo mediante el método de máxima verosimilitud y el planteamiento de los distintos contrastes tanto sobre los parámetros como de bondad de ajuste. Llegados a este punto se plantean los dos principales problemas del modelo de regresión logística: la separación completa, que para el caso múltiple no presenta dificultades, y la multicolinealidad. En la segunda parte se presenta brevemente el análisis de componentes principales desde el punto de vista muestral, con la definición y obtención de las componentes principales (cc. pp.) de una muestra de observaciones de un conjunto de variables, y las principales propiedades de optimalidad referentes a la reducción de dimensión y de representación de las variables originales. Finalmente se propone el modelo de regresión logística en componentes principales, comenzando con su formulación en términos de todas las componentes principales, el cual proporciona las mismas estimaciones de la variable respuesta que el modelo en términos de las variables originales. Finalizaremos esta última parte y este capítulo con la primera aportación de esta tesis: la formulación del modelo de regresión logística en componentes principales. Una vez formulado estimaremos los parámetros del mismo y propondremos una estimación (reconstrucción) de los parámetros del modelo original a partir éstos, estimación que será más precisa que la obtenida con todas las variables en presencia de multicolinealidad. También se proponen aquí distintos métodos de introducción de las cc. pp. en el modelo a la luz de las distintas investigaciones que indican que no necesariamente las componentes que más variabilidad acumulan son las que mejor explican a la variable respuesta.

El segundo capítulo aborda el aspecto funcional de esta tesis y se divide a su vez en cuatro secciones. En la primera se presentan las principales definiciones relativas a datos funcionales: proceso estocástico, trayectorias muestrales, funciones media y de covarianza, operador de covarianza, etc, así como las princi-

pales propiedades y exigencias: continuidad en media cuadrática y trayectorias de cuadrado integrable. También se introduce la necesidad de utilizar métodos aproximativos para la expresión de las trayectorias del proceso debido a la imposibilidad de observarlas de manera continua y se repasan las dos que se utilizarán en el tercer capítulo: aproximación mínimo-cuadrática e interpolación spline cúbica natural. Seguidamente en la segunda sección proponemos el modelo de regresión logística funcional y una estimación del mismo a partir de la consideración de que tanto las trayectorias como la función parámetro pertenezcan a un espacio de dimesión finita generado por una base. Dicha estimación será poco precisa, como se verá en el tercer capítulo, debido a la multicolinealidad lo que motivaría la utilización de análisis en componentes principales como posible solución. La tercera sección está dedicado al análisis en componentes principales funcional (ACPF). Después de repasar la teoría básica sobre la definición y estimación muestral de las cc. pp., se desarrollan los aspectos de estimación aproximada de las cc. pp. en un espacio de dimensión finita a partir de observaciones discretas de las funciones muestrales. En este caso el ACPF se reduce a un ACP múltiple de cierta transformación de las coordenadas de las funciones muestrales respecto de una base de funciones (Ocaña et al. 2002). Finalmente la cuarta parte muestra la principal aportación de esta tesis: la formulación y propuesta de estimación del modelo de regresión logística funcional. En primer lugar, considerando que las trayectorias muestrales pertenecen a un espacio de dimensión finita generado por una base de funciones, el modelo funcional es reducido a uno múltiple cuya matriz de diseño involucra a la matriz de coeficientes de las trayectorias muestrales con respecto a la base. En segundo lugar, para resolver el problema de multicolinealidad y reducir la dimensión del problema funcional, se formulan los modelos de regresión logística funcional en componentes principales, haciendo hincapié en la utilización de dos tipos de ACPF posibles así como en los dos métodos de introducción de cc. pp. en el modelo.

El tercer y último capítulo está dedicado a desarrollar cuatro estudios de simulación para ilustrar los modelos propuestos, la problemática que presentan y las distintas soluciones que se aportan. Todos estos estudios han sido desarrollados con el paquete S-plus. En primer lugar se desarrolla un primer estudio de simulación correspondiente al caso múltiple. En él se simulan datos de un modelo de regresión logística múltiple, esto es, de las covariables (con alta multicolinealidad) y de la variable respuesta, después de fijar unos parámetros. Posteriormente se ajustan tanto el modelo logístico en términos

de las variables originalmente simuladas como de las cc. pp. con distinto número de componentes y con los distintos métodos de selección de las mismas. Con los distintos modelos ajustados se reconstruyen los parámetros originales y se calculan medidas tanto de bondad de ajuste como de precisión de las estimaciones de los parámetros. Estos pasos se repiten una cantidad grande de veces y se calculan medidas resumen de las mismas que ponen de manifiesto que aquello que se obtuvo en el primer caso se verifica siempre bajo las mismas condiciones. La segunda parte de este último capítulo desarrolla tres ejemplos más de simulación en este caso correspondientes al caso funcional: el primero considerando la simulación directa de las trayectorias en un espacio de splines, el segundo aproximándolas mediante interpolación, a partir de observaciones en instantes discretos, y el tercero utilizando aproximación mínimo cuadrática. En los tres casos funcionales hemos seguido las mismas pautas, hemos simulado las trayectorias explicativas, una función parámetro y a partir de ahí los valores de la respuesta (siempre a través de aproximación en espacios de dimensión finita). Posteriormente se han ajustado los distintos modelos en términos de los valores originales y en términos de las cc. pp., las cuales han sido obtenidas por los dos tipos de ACPF considerados en el segundo capítulo, con la introducción de componentes según los dos métodos propuestos (variabilidad y stepwise), y se han reconstruido los parámetros y calculado las medidas de bondad de ajuste y precisión. Finalmente se han repetido todas estas operaciones una cantidad grande de veces para demostrar que aquello que ocurre en cada caso se mantiene al repetirlo.

Resumiendo, las principales aportaciones de esta Tesis son: formulación de los modelos de regresión logística múltiple en cc. pp. para resolver el problema de multicolinealidad en regresión logística y reducir la dimensión (tercera parte del capítulo 1), formulación, interpretación y estimación de los modelos de regresión logística funcional (segunda parte del capítulo 2), introducción de los modelos de regresión logística funcional en componentes principales (cuarta parte del capítulo 2) y desarrollo de estudios de simulación para ilustrar la precisión de las estimaciones obtenidas con los modelos propuestos y seleccionar criterios de introducción de las cc. pp. en los distintos modelos.

No quisiera concluir este prólogo si mostrar mi más sincero agradecimiento a todos aquellos que han "*sufrido*" la elaboración de este trabajo. En primer lugar a la directora de esta tesis, la profesora Dra. Dña. Ana M. Aguilera Del Pino, sin cuyos consejos tanto profesionales como personales no hubiera salido adelante la misma, en segundo lugar al profesor Dr. D. Mariano J.

Valderrama Bonnet pues él me abrió las puertas al mundo de la investigación e hizo que mi interés por ella creciera, también a mis compañeros y profesores del Departamento de Estadística e I.O. de la Universidad de Granada, ya que de todos ellos hay un poquito en esta tesis. También quisiera agradecer a mi familia la paciencia tenida en estos años y muy especialmente a Helena por su renuncia a tantas y tantas cosas para que este trabajo fuera posible.

Índice General

Capítulo 1

Regresión logística múltiple en componentes principales

1.1 Introducción

Con el modelo de regresión logística se pretende dar solución al problema de predecir una variable respuesta dicotómica en términos de un conjunto de variables explicativas. Como en la mayoría de modelos de regresión, en el logístico suele ocurrir con determinada frecuencia que las variables explicativas presentan una alta dependencia, lo que se conoce como multicolinealidad, que hace que no se obtengan estimaciones precisas de los parámetros del modelo con el consiguiente perjuicio para la interpretación de los mismos. Además también es común encontrarse con un número demasiado alto de variables explicativas que hace que el modelo presente dificultades computacionales en su ajuste.

En este capítulo abordamos la problemática expuesta anteriormente mediante la sustitución de las variables explicativas del modelo de regresión logística por un número reducido de variables incorreladas (sus componentes principales) de manera que podamos dar una estimación de los parámetros de dicho modelo a través de los que se obtienen al estimar el que denominaremos como modelo de regresión logística en componentes principales.

1.2 Regresión logística múltiple

Con el modelo de regresión logística múltiple pretendemos estudiar la relación existente entre una variable aleatoria de respuesta dicotómica (variable binaria

que toma los valores 1 y 0) y un conjunto de variables explicativas no aleatorias que en nuestro caso consideraremos continuas. La regresión logística ha sido objeto de muchas aplicaciones en medicina y epidemiología para explicar generalmente la probabilidad de padecer una determinada enfermedad en función de ciertos factores de riesgo asociados.

1.2.1 Formulación del modelo

Sea Y la variable respuesta que consideraremos dicotómica y $X_1, X_2, ..., X_p$ un conjunto de variables explicativas no aleatorias, todas observables y continuas. Fijados unos valores de estas últimas, tratamos de explicar el valor que tomaría Y. Como se demuestra en Ryan (1997), al ser la variable respuesta dicotómica, el modelo lineal no es adecuado de modo que utilizaremos el de regresión logística.

El modelo de regresión logística múltiple puede escribirse de la forma

$$Y = \pi(X) + \varepsilon \tag{1.1}$$

donde

- $\pi(X) = (\pi_1, \pi_2, \dots, \pi_n)^T$

  con

$$\pi_i = \frac{\exp\left\{\sum_{j=0}^{p} \beta_j x_{ij}\right\}}{1 + \exp\left\{\sum_{j=0}^{p} \beta_j x_{ij}\right\}} \quad i = 1, \dots, n. \tag{1.2}$$

- $Y = (y_1, y_2, \cdots, y_n)^T$ es el vector $n \times 1$ de observaciones de la variable respuesta (vector de unos y ceros).

- $X = (x_{ij})$; $i = 1, \dots, n$; $j = 0, \dots, p$ es una matriz $n \times (p+1)$ que contiene las observaciones de las variables explicativas, esto es, x_{ij} representa la observación i-ésima de la j-ésima variable explicativa, $j \neq 0$ y $x_{i0} = 1 \ \forall i$. Dicha matriz se conoce como matriz de diseño y sus filas se denotarán por x_i.

- $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^T$ vector de parámetros fijo desconocido.

- $\varepsilon = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)^T$ vector de errores que se considerarán variables aleatorias independientes de media cero y varianza $\pi_i(1 - \pi_i)$.

Observemos que las varianzas de los errores son distintas con lo que no podemos utilizar el método de mínimos cuadrados ordinarios para estimar el modelo. Como alternativa podría usarse el método de mínimos cuadrados ponderados en el caso de disponer de observaciones repetidas de la variable respuesta en cada combinación de valores x_i de las variables explicativas. El método de estimación más usualmente utilizado es el de máxima verosimilitud basado en la distribución de Bernouilli de cada una de las observaciones de respuesta (Hosmer y Lemeshow, 1989), $y_i \rightsquigarrow B(\pi_i)$.

De manera análoga, podemos expresar (1.2) de la siguiente forma:

$$l_i = \sum_{j=0}^{p} \beta_j x_{ij} \quad i = 1, \ldots, n$$

donde cada l_i recibe el nombre de transformación logit y corresponde al logaritmo de la ventaja de respuesta $Y = 1$ para el valor observado x_i de las variables explicativas, cuya expresión es la siguiente:

$$l_i = \ln\left[\frac{\pi_i}{1-\pi_i}\right], \; i = 1.\ldots, n.$$

De este modo, el modelo de regresión logística puede verse como un modelo lineal generalizado cuya función ligadura (link) es la transformación logit (ver por ejemplo McCullagh y Nelder, 1983). Así, el modelo matricialmente quedaría

$$L = X\beta$$

con $L = (l_1, \ldots, l_n)^T$, y siendo X la matriz de diseño habitual, y β el vector de parámetros anteriormente definido.

1.2.2 Interpretación de los parámetros

En cualquier investigación en la que se propone un modelo es necesario interpretar los parámetros del mismo de manera que se puedan obtener conclusiones de la relación entre las variables. Los coeficientes de las variables independientes representan la razón de cambio de una función de la variable dependiente, por cada unidad que cambia la independiente. Esto conlleva la necesidad de saber qué función relaciona las variables dependiente e independientes, y definir la unidad de cambio.

En regresión logística la función que relaciona las variables, para el caso univariante, es

$$l(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$$

de manera que

$$\beta_1 = l(x+1) - l(x)$$

por lo que representa el cambio en los logit para un cambio de una unidad en la variable independiente x.

Es decir, cuando x aumenta en una unidad el logaritmo de la ventaja (de $Y = 1$ frente a $Y = 0$) aumenta aditivamente β_1 unidades. La interpretación de este coeficiente depende de las unidades en que se mida la variable.

Como indican Hosmer y Lemeshow (1989) muchas veces un aumento de una unidad en la variable x puede no tener sentido, dependiendo de las unidades de medida, por lo que puede ser más interesante un cambio de c unidades en x, en cuyo caso

$$c\beta_1 = l(x+c) - l(x).$$

En cualquier caso el cambio en el logaritmo de las ventajas no tiene una interpretación intuitiva en la práctica. Por ello se trabaja con el cociente de ventajas que proporciona una interpretación interesante de las exponenciales de los parámetros del modelo de regresión logística.

El cociente de las ventajas (odd ratio) de respuesta uno frente a respuesta cero para dos valores de la variable explicativa que se diferencien en una unidad es de la forma

$$\theta(\Delta x = 1) = \frac{\dfrac{\pi(x+1)}{1-\pi(x+1)}}{\dfrac{\pi(x)}{1-\pi(x)}} = \frac{e^{\beta_0+\beta_1(x+1)}}{e^{\beta_0+\beta_1(x)}} = e^{\beta_1},$$

que quiere decir que al aumentar una unidad la variable x, la ventaja de respuesta uno queda multiplicada por e^{β_1}.

Análogamente, para un cambio de c unidades en x se obtiene que el cociente de ventajas asociado es

$$\theta(\Delta x = c) = \exp(c\beta_1).$$

En el caso del modelo de regresión logística múltiple, la exponencial de cada coeficiente corresponde al cociente de ventajas de respuesta uno cuando incrementamos en una unidad la variable asociada a ese coeficiente y mantenemos fijas las demás. Es decir,

$$\theta(\Delta x_j = 1/x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p) = e^{\beta_j} \quad j = 1, \ldots, p.$$

Esto significa que al aumentar en una unidad una de las variables y controlar las demás, la ventaja de respuesta uno queda multiplicada por la exponencial del coeficiente de la variable incrementada. Esto implica que si la exponencial de un parámetro es mayor que uno (o el parámetro mayor que cero) la probabilidad de respuesta uno aumenta cuando se incrementa la variable asociada y se controlan las demás, mientras que si esta exponencial es menor que uno (parámetro menor que cero) la relación es inversa.

1.2.3 Estimación de parámetros

El primer paso, como en todo modelo de regresión, es estimar los parámetros presentes en el modelo, esto es, encontrar el vector de parámetros β. Este problema es conocido como estimación del modelo y lo vamos a resolver utilizando el método de máxima verosimilitud, método que consiste en encontrar el valor de los parámetros que maximiza la probabilidad de encontrar una muestra como la observada, es decir, maximiza la función de verosimilitud.

Como hemos comentado en la formulación del modelo, las observaciones de la variable independiente se pueden ver como valores de n vv.aa. independientes con distribución de Bernouilli por lo que la probabilidad de la muestra, que vendrá dada por la función de verosimilitud, será

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}.$$

Maximizaremos, como es usual, la log-verosimilitud

$$\mathcal{L}(\beta) = \ln L(\beta) = \sum_{i=1}^{n} \left(y_i \ln \left(\frac{\pi_i}{1 - \pi_i} \right) + \ln (1 - \pi_i) \right).$$

Derivando esta expresión e igualando a cero obtenemos las ecuaciones de verosimilitud que son de la forma

$$\sum_{i=1}^{n} (y_i x_{ij} - \widehat{\pi}_i x_{ij}) = 0, \ \forall j = 0, \ldots, p, \tag{1.3}$$

donde

$$\widehat{\pi}_i = \frac{\exp\left\{\sum_{j=0}^{p} x_{ij}\widehat{\beta}_j\right\}}{1+\exp\left\{\sum_{j=0}^{p} x_{ij}\widehat{\beta}_j\right\}} \tag{1.4}$$

es el estimador de máxima verosimilitud de π_i y $\widehat{\beta}_j$ el estimador de máxima verosimilitud de β_j.

Las ecuaciones de verosimilitud pueden escribirse equivalentemente en forma matricial como

$$X^T\left(Y-\widehat{\pi}\left(X\right)\right)=0 \tag{1.5}$$

donde $\widehat{\pi}(X)=(\widehat{\pi}_1,\ldots,\widehat{\pi}_n)^T$.

En general se ha demostrado que las ecuaciones de verosimilitud tienen solución única salvo cuando existe lo que se conoce como separación completa de los datos que, en el caso univariante, consiste en que todas las observaciones para las que la variable respuesta es cero tienen valores de la variable explicativa menores que aquellos para los que la respuesta es uno. Así, Albert y Anderson (1984) y Santner y Duffy (1986) demuestran que para que existan los estimadores máximo-verosímiles de los parámetros tiene que darse cierto solapamiento en los datos. Afortunadamente, en el caso que estamos tratando, el de variables explicativas continuas, es difícil encontrar separación completa en las observaciones por lo que consideraremos que existen los estimadores máximo-verosímiles. Un estudio más detallado acerca de la separación completa y la existencia de los estimadores de máxima verosimilitud se puede ver en Ryan (1997). Recientemente, Christmann y Rousseeuw (2000) han desarrollado algoritmos para medir la cantidad de solapamiento de un conjunto de observaciones.

Observemos finalmente que las ecuaciones de verosimilitud no son lineales en los parámetros por lo que no se pueden resolver directamente sino que debemos utilizar métodos iterativos para ello. El algoritmo más utilizado para la resolución de estas ecuaciones es el de Newton-Raphson. Veamos brevemente en qué consiste la estimación por máxima verosimilitud mediante este método.

Para obtener una expresión de la solución del sistema (1.5) mediante el método de Newton-Raphson (ver Apéndice) tendremos que calcular la matriz jacobiana de $X^T\left(Y-\pi\left(X\right)\right)$ que estará dada por

$$J\left(\beta\right)=-X^TWX$$

donde

$$W = Diag\left(\pi_1\left(1-\pi_1\right), \ldots, \pi_n\left(1-\pi_n\right)\right).$$

Así que aplicando el método de Newton-Raphson, las estimaciones de los parámetros quedan

$$\widehat{\beta}^{(m+1)} = \widehat{\beta}^{(m)} + \left(X^T W^{(m)} X\right)^{-1} X^T \left(Y - \pi^{(m)}\left(X\right)\right)$$

donde $W^{(m)} = Diag\left(\pi_i^{(m)}\left(1-\pi_i^{(m)}\right)\right)$, con $\pi_i^{(m)}$ la aproximación de $\widehat{\pi}_i$ obtenida en el paso m, y $\pi^{(m)}\left(X\right) = \left(\pi_1^{(m)}, \ldots, \pi_n^{(m)}\right)$.

Como en todo método iterativo, hay que dar un criterio de parada que indique cuándo estamos lo suficientemente cerca de la verdadera solución. En regresión logística este criterio suele tomarse de manera que el valor de la función de verosimilitud en las estimaciones de los parámetros en dos iteraciones consecutivas cambie en menos de una determinada cantidad. También se suele considerar un cambio suficientemente pequeño en los parámetros y en las probabilidades estimadas en dos iteraciones consecutivas. La convergencia del método de Newton-Raphson suele ser rápida, de hecho esta convergencia es de segundo orden. Además del criterio de parada, resulta interesante elegir bien una aproximación inicial de la solución que haga que el método converja de manera rápida. Se ha demostrado que las soluciones que proporciona el análisis discriminante son un buen punto de partida para encontrar la solución rápidamente (Ryan, 1997).

Como se aprecia, no podemos dar una expresión explícita de los estimadores de los parámetros sino una aproximación de la estimación de los mismos que converge al verdadero valor.

Propiedades de los estimadores

Como hemos comentado previamente, no podemos dar una expresión concreta del estimador de los parámetros, con lo que no podemos obtener la distribución muestral del mismo de manera exacta. Lo que sí podemos hacer, puesto que los estimadores son máximo-verosímiles, es dar las propiedades asintóticas de los mismos (ver Rohatgi, 1984). Así obtenemos el siguiente resultado.

Teorema 1 *Sea $\widehat{\beta}$ el estimador máximo-verosímil de β del modelo de regresión logística. Entonces,*

$$\left(\widehat{\beta} - \beta\right) \underset{n\to\infty}{\rightsquigarrow} N_{p+1}\left[0, i\left(\beta\right)^{-1}\right]$$

(convergencia en distribución), siendo $i(\beta)^{-1}$ *la matriz de información de Fisher cuya expresión para el modelo de regresión logística es*

$$i(\beta) = E\left[-\frac{\partial^2}{\partial\beta^2}\mathcal{L}(\beta)\right] = -J(\beta) = X^T W X.$$

La conclusión del teorema es que la sucesión de estimadores de máxima verosimilitud, supuesto que se obtienen como raíz única de las ecuaciones de verosimilitud (y son por tanto una sucesión consistente), tienen asintóticamente distribución $N_{p+1}\left[\beta, (i(\beta))^{-1}\right]$.

Según esto, podemos dar una estimación de la matriz de varianzas-covarianzas del estimador sin más que utilizar las estimaciones de π_i en W. Es decir,

$$\widehat{Cov}\left[\widehat{\beta}\right] = \left(X^T \widehat{W} X\right)^{-1},$$

con $\widehat{W} = Diag(\widehat{\pi}_i(1-\widehat{\pi}_i))$. Además el estimador $\widehat{\beta}$ es asintóticamente insesgado. Observemos que la matriz de covarianzas del estimador aproximado $\widehat{\beta}^{(m)}$ es

$$Cov\left[\widehat{\beta}^{(m)}\right] = \left(X^T \widehat{W}^{(m)} X\right)^{-1}$$

y se obtiene como un subproducto del método de Newton-Raphson.

Una vez obtenidos los estimadores de los parámetros y sus propiedades sería bueno poder evaluar la bondad de dichos estimadores. La primera medida de bondad de un estimador es que sea insesgado, lo cual hemos comprobado asintóticamente en el resultado anterior. Otra buena propiedad de los estimadores máximo-verosímiles es que tienen varianza mínima.

1.2.4 Inferencia

Una vez que se han estimado los parámetros, el siguiente paso es hacer inferencia sobre los mismos y contrastar la bondad del ajuste del modelo. En regresión logística existen diversas formas de hacer contrastes sobre los parámetros, todas ellas basadas en contrastes asintóticos y válidas para tamaños muestrales suficientemente grandes.

Contrastes de bondad de ajuste

El contraste de bondad de ajuste pretende decidir si el modelo de regresión logística propuesto se ajusta bien a los datos. En la literatura podemos encontrar multitud de formas de contrastar la bondad del ajuste en regresión

logística algunas de las cuales se comparan en Hosmer et al. (1997). Nosotros analizaremos aquí tres: test Chi-cuadrado de bondad de ajuste, estadístico de Wilks de razón de verosimilitudes y test de bondad de ajuste de Hosmer y Lemeshow. En todos los casos se pretende contrastar

$$H_0 : \pi_i = \frac{\exp\left\{\sum_{j=0}^{p} x_{ij}\beta_j\right\}}{1+\exp\left\{\sum_{j=0}^{p} x_{ij}\beta_j\right\}},\ i = 1, \ldots, n$$

$$H_1 : \pi_i \neq \frac{\exp\left\{\sum_{j=0}^{p} x_{ij}\beta_j\right\}}{1+\exp\left\{\sum_{j=0}^{p} x_{ij}\beta_j\right\}}$$

esto es, si la esperanza condicional $E\left[Y/X_1 = x_{i1}, \ldots, X_p = x_{ip}\right] = E[y_i] = \pi_i$ se puede expresar según el modelo de regresión logística.

Test chi-cuadrado de Pearson Se trata del test chi-cuadrado de bondad de ajuste habitual. En nuestro modelo cada observación de la variable respuesta ha sido generada por una v.a. de Bernouilli $B(\pi_i)$ o equivalentemente una multinomial $M(1, \pi_i, 1-\pi_i)$, con lo que el estadístico chi-cuadrado de bondad de ajuste al modelo de regresión logística se obtiene como la suma de los estadísticos chi-cuadrado de bondad ajuste a cada una de las n multinomiales bajo la hipótesis nula H_0. El test chi-cuadrado compara los valores observados con los valores esperados bajo la hipótesis nula, esto es, los ajustados según el modelo que estemos utilizando, que en nuestro caso es el logístico. Así, para nuestro caso el estadístico chi-cuadrado de contraste quedaría

$$\mathrm{X}^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\pi}_i)^2}{\widehat{\pi}_i\,(1-\widehat{\pi}_i)}$$

donde $\widehat{\pi}_i$ es la estimación máximo-verosímil de π_i asumiendo el modelo, dada por (1.4).

Bajo la hipótesis nula de que las probabilidades verifiquen el modelo, este estadístico se distribuye como una χ^2_{n-p-1}. Observemos que los grados de libertad se obtienen como la diferencia entre el número de observaciones y el número de parámetros del modelo. Por lo tanto, se rechazará la hipótesis nula al nivel α cuando se verifique

$$\mathrm{X}^2_{Obs} \geq \chi^2_{n-p-1;\alpha},$$

siendo $\chi^2_{n-p-1;\alpha}$ el cuantil de orden $1-\alpha$ de la distribución χ^2_{n-p-1}.

Este estadístico de contraste tiene un problema que consiste en que la distribución de cada estadístico chi-cuadrado es asintótica y nosotros tenemos una única observación en cada uno, por lo que no resultaría acertado. Para evitar este problema, Hosmer & Lemeshow propusieron un test alternativo agrupando las observaciones.

Test de Hosmer y Lemeshow Este test no es más que el test chi-cuadrado que resulta después de agrupar convenientemente las observaciones de las variables explicativas en intervalos.

Para realizar el test se agrupan las observaciones en g grupos o clases, de modo que en cada clase estarán todos los individuos cuyas probabilidades estimadas estén entre dos prefijadas. Se ha comprobado que la mejor opción de agrupamiento es la basada en los deciles de las probabilidades estimadas obteniendo $g = 10$ grupos.

Una vez formados los grupos, se contabilizan en cada uno el número de observaciones que caen en dicho grupo, el número de unos, y la media de las probabilidades estimadas. Entonces el estadístico del contraste se obtiene de nuevo como la suma de los estadísticos chi-cuadrado de bondad de ajuste a las multinomiales asociadas a los distintos grupos

$$\widehat{C} = \frac{\sum_{r=1}^{g} (o_r - n_r \overline{\pi}_r)}{n_r \overline{\pi}_r (1 - \overline{\pi}_r)}$$

donde

- n_r es el número de observaciones en el grupo r,
- $o_r = \sum_{i=1}^{n_r} y_i$,
- $\overline{\pi}_r = \dfrac{1}{n_r} \sum_{i=1}^{n_r} \widehat{\pi}_i$.

Este estadístico se distribuye de nuevo asintóticamente y bajo la hipótesis nula como χ^2_{g-2}.

Para utilizar el estadístico chi-cuadrado hay que exigir un número mínimo de observaciones en cada grupo, y que las frecuencias esperadas superen una determinada cantidad. La regla usual es exigir que el 80% de los mismos tengan frecuencias esperadas mayores que 5 y todas ellas mayores que 1.

Test de razón de verosimilitudes El test de razón de verosimilitudes considera como estadístico el cociente entre el máximo de la verosimilitud bajo la hipótesis nula y el máximo de la verosimilitud en la población. En este sentido, el estadístico sería

$$\Lambda = \frac{\sup\{L_{H_0}\}}{\sup\{L\}}$$

Veamos en primer lugar la estimación de los parámetros sin asumir ningún modelo, esto es, asumiendo que tenemos $y_1, \ldots, y_n$ observaciones independientes de distribuciones $B(\pi_i)$ cada una. La verosimilitud de los datos sería

$$L = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i}$$

de manera que tomando la log-verosimilitud

$$\ln L = \sum_{i=1}^{n} [y_i \ln \pi_i + (1 - y_i) \ln (1 - \pi_i)],$$

derivando e igualando a cero, se tiene

$$y_i - \pi_i = 0, \; i = 1, \ldots, n$$

por lo que la estimación de los parámetros π_i coincidirían con las observaciones. Esto es lo que se conoce como modelo saturado, que no es más que aquel modelo que tiene tantos parámetros como observaciones, con lo que los valores ajustados coinciden con los observados. A partir de ahora denotaremos por L_S al supremo de la verosimilitud bajo el modelo saturado y L_M al supremo de la verosimilitud bajo el modelo logístico de H_0.

Entonces, el cociente de verosimilitudes será

$$\Lambda = \frac{L_M}{L_S} = \prod_{i=1}^{n} \left(\frac{\widehat{\pi}_i}{y_i}\right)^{y_i} \left(\frac{1 - \widehat{\pi}_i}{1 - y_i}\right)^{1-y_i}$$

que tiene distribución desconocida. Esto se soluciona utilizando el estadístico de Wilks que se resume en el siguiente resultado y que se puede ver en Rohatgi, (1984).

Teorema 2 *Dada $X_1, \ldots, X_n$ una muestra aleatoria simple de una población con distribución f_θ con $\theta \in \Theta$, y consideremos el contraste de hipótesis*

$$H_0 \; : \; \theta \in \Theta_0$$
$$H_1 \; : \; \theta \in \Theta.$$

con Θ_0 y Θ espacios paramétricos. Sea el cociente de verosimilitudes

$$\Lambda = \frac{\sup\limits_{\theta\in\Theta_0} f_\theta}{\sup\limits_{\theta\in\Theta} f_\theta}$$

entonces

$$-2\ln\Lambda \underset{n\to\infty}{\overset{H_0}{\rightsquigarrow}} \chi^2_l$$

con l la diferencia entre los tamaños de los espacios paramétricos.

La expresión del estadístico de Wilks de razón de verosimilitudes para el contraste de bondad de ajuste al modelo de regresión logística es de la forma

$$G^2(M) = -2\ln\Lambda = 2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{y_i}{\widehat{\pi}_i}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\widehat{\pi}_i}\right)\right] \underset{n\to\infty}{\overset{H_0}{\rightsquigarrow}} \chi^2_{n-p-1}.$$

Denotando por $\mathcal{L}_M$ y $\mathcal{L}_S$ a los supremos de las log-verosimilitudes del modelo logístico y del saturado respectivamente, este estadístico también se puede expresar de la forma

$$G^2(M) = -2\ln\frac{L_M}{L_S} = 2\left(\mathcal{L}_S - \mathcal{L}_M\right)$$

de manera que se tiene que el estadístico G^2 de Wilks se puede expresar como el doble de la diferencia entre el máximo de la log-verosimilitud bajo el modelo saturado y el máximo de la log-verosimilitud bajo el modelo logístico.

Al estadístico $G^2(M)$ se le suele llamar *deviance* y juega un papel similar a la suma de los cuadrados de los residuos del modelo de regresión lineal.

Medidas globales de bondad de ajuste

Como en todos los modelos de regresión, además de los contrastes de bondad de ajuste resulta interesante analizar medidas que de algún modo informen sobre la bondad del ajuste mediante un dato numérico objetivo. Veamos ahora algunas medidas, análogas a las ampliamente conocidas en el modelo lineal, para nuestro modelo de regresión logística.

Coeficiente R^2 en regresión logística En el caso de regresión lineal es conocido que la bondad del ajuste se mide mediante el coeficiente de determinación que se definine a partir de los valores observados y los predichos por el modelo de la siguiente forma:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}.$$

Para regresión logística no se puede calcular R^2 del mismo modo ya que se ha demostrado que así calculado R^2 puede llegar a tomar valores pequeños cuando el ajuste es casi perfecto (Ryan, 1997). En el caso lineal el coeficiente de determinación también se puede definir en términos de cocientes de verosimilitudes. De manera análoga para el modelo de regresión logística definiremos el coeficiente de determinación a partir del cociente entre el máximo de la verosimilitud bajo el modelo que sólo tiene el término constante (L_0) y el máximo de la verosimilitud bajo el modelo con todos los parámetros (L_M).

Definición 1 *Se define el coeficiente de determinación para regresión logística de la forma*

$$R^2 = 1 - \left(\frac{L_0}{L_M}\right)^{2/n}.$$

El coeficiente R^2 calculado de esta manera será menor que 1 ya que se tiene que el máximo de la verosimilitud del modelo completo es un producto de probabilidades. Por esta razón, el máximo valor que pueda tomar R^2, en lugar de uno, será

$$\max R^2 = 1 - \left[\gamma^{n\gamma} (1-\gamma)^{n(1-\gamma)}\right]^{2/n}$$

siendo γ el porcentaje de unos en el conjunto de datos. Este valor puede ser próximo a cero cuando hay pocos datos con lo que se ha propuesto como medida de la bondad del ajuste lo que se conoce como coeficiente de determinación corregido o ajustado, que se define como

$$\overline{R}^2 = \frac{R^2}{\max R^2}.$$

Tasa de clasificaciones correctas La medida que más se utiliza en regresión logística para evaluar la bondad del ajuste es la tasa de clasificaciones correctas (CCR). Para calcular la medida CCR se elige un punto de corte p_c

para las probabilidades, que usualmente es $p_c = 0.5$ y se considera que un individuo está clasificado correctamente cuando su probabilidad estimada $\widehat{\pi}_i \geq p_c$ e $y_i = 1$ o bien $\widehat{\pi}_i < p_c$ e $y_i = 0$, mientras que en caso contrario se considera clasificado incorrectamente. Así, la tasa de clasificaciones correctas se define como el cociente entre el número de observaciones clasificadas correctamente y el número total de observaciones muestrales.

Aunque usualmente se escoge como punto de corte 0.5, sería más adecuado utilizar aquel punto de corte que maximice la tasa de clasificaciones correctas (Hosmer y Lemeshow, 1989) o incluso tomar como punto de corte la proporción de unos en la muestra.

Contrastes sobre los parámetros

Pretendemos obtener contrastes de significación sobre los parámetros del modelo de regresión logística múltiple considerado, esto es, asumimos que el modelo elegido es el correcto e intentamos identificar las variables que son significativas. Se trata por lo tanto de un contraste de la forma

$$\begin{aligned} &H_0 : \beta_{(1)} = \beta_{(2)} = \cdots = \beta_{(l)} = 0 \\ &H_1 : \text{Algún } \beta_{(i)} \neq 0 \; \forall i = 1, \ldots, l \end{aligned} \tag{1.6}$$

donde $\beta_{(1)}, \beta_{(2)}, \ldots, \beta_{(l)}$ son $l \leq p+1$ parámetros escogidos de entre los $p+1$ de que disponemos.

Dicho de otra manera, pretendemos contestar a la pregunta: ¿las variables $X_{(1)}, X_{(2)} \ldots, X_{(l)}$ ayudan a la explicación de la variable respuesta?. Este problema se puede solucionar de varias formas que se presentan a continuación.

Contrastes de Wald Están basados en la distribución normal asintótica de los estimadores de máxima verosimilitud. Recordemos que la distribución asintótica del estimador máximo-verosímil de los parámetros del modelo logístico era

$$\widehat{\beta} \rightsquigarrow N_{p+1}\left(\beta, (X^T W X)^{-1}\right).$$

Consideremos el vector de parámetros $\beta_{(0)} = \left(\beta_{(1)}, \beta_{(2)}, \ldots, \beta_{(l)}\right)^T$, y su estimador máximo verosímil $\widehat{\beta}_{(0)}$ cuyas componentes son los estimadores máximo-verosímiles $\widehat{\beta}_{(1)}, \widehat{\beta}_{(2)}, \ldots, \widehat{\beta}_{(l)}$ de $\beta_{(1)}, \beta_{(2)}, \ldots, \beta_{(l)}$. Entonces

$$\widehat{\beta}_{(0)} \rightsquigarrow N_l\left(\beta_{(0)}, Cov(\widehat{\beta}_{(0)})\right),$$

siendo $Cov(\widehat{\beta}_{(0)})$ la matriz de covarianzas de $\widehat{\beta}_{(0)}$ obtenida directamente como una submatriz a partir de la del estimador de máxima verosimilitud completo $(X^TWX)^{-1}$.

Entonces

$$\left(\widehat{\beta}_{(0)} - \beta_{(0)}\right)^T \left(Cov(\widehat{\beta}_{(0)})\right)^{-1} \left(\widehat{\beta}_{(0)} - \beta_{(0)}\right) \rightsquigarrow \chi^2_l.$$

Si consideramos esta misma distribución bajo la hipótesis nula impuesta, el estadístico de contraste que se obtiene será

$$\widehat{\beta}^T_{(0)} \left(\widehat{Cov}(\widehat{\beta}_{(0)})\right)^{-1} \widehat{\beta}_{(0)}$$

donde $\widehat{Cov}(\widehat{\beta}_{(0)})$ es el estimador de $Cov(\widehat{\beta}_{(0)})$ obtenido sustituyendo las probabilidades poblacionales π_i por sus estimadores de máxima verosimilitud bajo el modelo, $\widehat{\pi}_i$.

Por lo tanto, se rechazará H_0 con nivel de significación α siempre que

$$\widehat{\beta}^T_{(0)} \left(\widehat{Cov}(\widehat{\beta}_{(0)})\right)^{-1} \widehat{\beta}_{(0)} \geq \chi_{l;\alpha}.$$

Como caso particular podríamos hacer el contraste para todos los parámetros excepto el término independiente. También se pueden plantear contrastes sobre cada uno de los parámetros individuales del tipo

$$\begin{aligned} H_0 &: \beta_j = \beta_j^0, \\ H_1 &: \beta_j \neq \beta_j^0. \end{aligned}$$

Es conocido que el estimador máximo-verosímil de cada parámetro tiene distribución normal con media el parámetro y con varianza el correspondiente elemento diagonal de $\left(X^TWX\right)^{-1}$ que denotaremos por $S\left(\widehat{\beta}_j\right)$. Entonces, bajo la hipótesis nula, se tiene

$$\widehat{\beta}_j \underset{n\to\infty}{\overset{H_0}{\rightsquigarrow}} N\left(\beta_j^0, S\left(\widehat{\beta}_j\right)\right),$$

de ahí que el estadístico del contraste tenga, bajo H_0, distribución

$$Z = \frac{\widehat{\beta}_j - \beta_j^0}{\sqrt{\widehat{S}\left(\widehat{\beta}_j\right)}} \underset{n\to\infty}{\overset{H_0}{\rightsquigarrow}} N[0,1]$$

siendo $\widehat{S}\left(\widehat{\beta}_j\right)$ el correspondiente elemento diagonal de $\left(X^T\widehat{W}X\right)^{-1}$ con $\widehat{W} = diag\left(\widehat{\pi}_i\left(1-\widehat{\pi}_i\right)\right)$.

Por lo tanto se rechazará H_0 al nivel de significación α siempre que

$$|Z| \geq z_{\alpha/2}.$$

El contraste anterior se podría efectuar a través de la distribución chi-cuadrado sin más que considerar el cuadrado del estadístico Z. Tomando $\beta_j^0 = 0$ se obtienen los contrastes de Wald de significación de cada uno de los parámetros individuales del modelo.

Contrastes condicionales de razón de verosimilitudes Constituyen un método alternativo al test de Wald multivariante para el contraste de hipótesis (1.6) que puede expresarse equivalentemente en la forma

$$\begin{aligned} H_0 &: \text{Modelo } M_P \text{ se verifica} \\ H_1 &: \text{Modelo } M_P \text{ no se verifica asumiendo } M, \end{aligned}$$

donde M es el modelo logístico considerado, es decir el que tiene los $p+1$ parámetros y se asume como cierto, y M_P el modelo particular y anidado en M obtenido después de haber hecho cero los l parámetros de la hipótesis nula.

El test de razón de verosimilitudes para este contraste es de la forma

$$\Lambda = \frac{L_{M_P}}{L_M}.$$

siendo L_{M_p} el supremo de la verosimilitud bajo M_p definida por

$$L_{M_P} = \prod_{i=1}^{n} \widehat{\pi}_{i(M_P)}^{y_i} \left(1 - \widehat{\pi}_{i(M_P)}\right)^{1-y_i},$$

con $\widehat{\pi}_{i(M_P)}$ el estimador máximo-verosímil bajo M_p, y L_M el supremo de la verosimilitud bajo M

$$L_M = \prod_{i=1}^{n} \widehat{\pi}_i^{y_i} \left(1 - \widehat{\pi}_i\right)^{1-y_i}.$$

Entonces

$$\Lambda = \frac{L_{M_P}}{L_M} = \prod_{i=1}^{n} \left(\frac{\widehat{\pi}_{i(M_P)}}{\widehat{\pi}_i}\right)^{y_i} \left(\frac{1 - \widehat{\pi}_{i(M_P)}}{1 - \widehat{\pi}_i}\right)^{1-y_i}.$$

La distribución de este estadístico es desconocida. Para poder realizar el contraste acudimos de nuevo a la distribución asintótica del estadístico de

Wilks de razón de verosimilitudes, denotado por $G^2(M_P/M)$, que es de la forma

$$\begin{aligned} G^2(M_P/M) &= -2\ln\Lambda \\ &= -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\widehat{\pi}_{i(M_P)}}{\widehat{\pi}_i}\right) + (1-y_i)\ln\left(\frac{1-\widehat{\pi}_{i(M_P)}}{1-\widehat{\pi}_i}\right)\right] \underset{n\to\infty}{\overset{H_0}{\rightsquigarrow}} \chi^2_l. \end{aligned}$$

Observemos que los grados de libertad de la distribución chi-cuadrado de este estadístico bajo H_0 se obtienen como la diferencia entre la dimensión del espacio paramétrico, que asumiendo el modelo M es $p+1$, y la dimensión del espacio paramétrico bajo la hipótesis nula, que es el número de parámetros del modelo M_P dado por $p+1-l$.

El estadístico anterior tiene la ventaja de poder descomponerse como sigue en términos de los estadísticos de Wilks de razón de verosimilitudes para los contrastes globales de bondad de ajuste de los modelos M_P y M :

$$G^2(M_P/M) = -2\ln\frac{L_{M_P}}{L_M} = 2(\mathcal{L}_M - \mathcal{L}_{M_P}),$$

siendo $\mathcal{L}_M$ y $\mathcal{L}_{M_P}$ los máximos de las log-verosimilitudes de los modelos general y particular respectivamente. Si llamamos $\mathcal{L}_S$ al máximo de la log-verosimilitud del modelo saturado, se tiene que sumando y restando $2\mathcal{L}_S$

$$\begin{aligned} G^2(M_P/M) &= 2(\mathcal{L}_M - \mathcal{L}_{M_P}) - 2\mathcal{L}_S + 2\mathcal{L}_S \\ &= [-2(\mathcal{L}_{M_P} - \mathcal{L}_S)] - [-2(\mathcal{L}_M - \mathcal{L}_S)] \\ &= G^2(M_P) - G^2(M), \end{aligned}$$

lo cual lleva a que el test de razón de verosimilitudes para contrastar dos modelos anidados, uno resultado de hacer cero algunos parámetros en el otro, es la diferencia de los contrastes de razón de verosimilitudes de bondad de ajuste de cada uno de los modelos y su distribución asintótica tiene como grados de libertad la diferencia entre los grados de libertad de ambos estadísticos que corresponde al número de parámetros anulados.

El test de significación de cada uno de los parámetros individuales se puede obtener directamente como un caso particular de contraste condicional de razón de verosimilitudes del modelo que tiene todas las variables menos la asociada al parámetro que se anula, con respecto al modelo con todas las variables de partida.

Intervalos de confianza

Los intervalos de confianza que veremos están basados en la normalidad asintótica de los estimadores de máxima verosimilitud.

Intervalo de confianza para los parámetros Para un parámetro individual, un intervalo de confianza asintótico al nivel $(1-\alpha)$ sería

$$\left(\widehat{\beta}_j \pm z_{\alpha/2}\sqrt{\widehat{S}\left(\widehat{\beta}_j\right)}\right).$$

Intervalo de confianza para las transformaciones logit De manera análoga se puede obtener el intervalo para una combinación lineal de parámetros a partir de la siguiente distribución normal:

$$x_i^T\beta \underset{n\to\infty}{\rightsquigarrow} N\left[x_i^T\widehat{\beta}, x_i^T\left(X^TWX\right)^{-1}x_i\right]$$

con x_i^T la fila i-ésima de X.

Entonces el intervalo de confianza para la transformación logit $l_i = x_i^T\beta$ será

$$\left(x_i^T\widehat{\beta} \pm z_{\alpha/2}\sqrt{x_i^T\left(X^T\widehat{W}X\right)^{-1}x_i}\right).$$

Intervalo de confianza para probabilidades A partir del intervalo anterior para las transformaciones logit, podemos obtener el siguiente intervalo de confianza aproximado al nivel $(1-\alpha)$ para cada una de las probabilidades

$$\pi_i = \frac{\exp\left\{x_i^T\beta\right\}}{1+\exp\left\{x_i^T\beta\right\}} = \frac{e^{l_i}}{1+e^{l_i}},$$

en la forma

$$\left(\frac{\exp A}{1+\exp A}, \frac{\exp B}{1+\exp B}\right),$$

con

$$A = x_i^T\widehat{\beta} - z_{\alpha/2}\sqrt{x_i^T\left(X^TWX\right)^{-1}x_i},$$

y

$$B = x_i^T\widehat{\beta} + z_{\alpha/2}\sqrt{x_i^T\left(X^TWX\right)^{-1}x_i}.$$

1.2.5 Selección de modelos

Hasta ahora hemos ajustado el modelo y hemos realizado los contrastes suponiendo que las variables explicativas consideradas son las mejores a la hora de predecir la respuesta. Pretendemos ahora seleccionar las variables que mejor ayuden a explicar la respuesta. Los métodos más usuales que se utilizan en regresión logística están basados en los contrastes condicionales de razón de verosimilitudes, siguiendo siempre el criterio de parsimonia para seleccionar el modelo con menor número de parámetros que se ajuste bien a los datos.

Método backward o de eliminación progresiva

Es un método que pretende obtener el mejor ajuste de regresión partiendo del modelo que incluye a todas las variables disponibles. En primer lugar, consideramos las p variables disponibles y ajustamos el modelo con todas las variables. A continuación se examinan todas las variables individualmente para ver cuál sería la candidata a abandonar el modelo, mediante los contrastes condicionales de razón de verosimilitudes $H_0 : \beta_j = 0$ $j = 1, \ldots, p$. Elegimos como candidata a ser eliminada, aquella que tenga p-valor máximo entre todos los p-valores mayores que el nivel de significación prefijado α. Si no hay ninguna variable verificando esta condición (todos los p-valores son menores o iguales que α), el modelo con todas las variables es el más acertado y paramos el método. En caso contrario, eliminamos dicha variable y reajustamos el modelo, planteándonos en el siguiente paso la eliminación de cada una de las restantes variables del modelo seleccionado en el paso anterior. Este procedimiento continúa hasta que encontremos que no hay ninguna variable candidata a salir del modelo o bien el modelo resultante de la eliminación no se ajusta bien globalmente. En la Tabla 1.1 se muestra un cuadro resumen del algoritmo de selección del método *backward* con M_j el modelo que resulta de eliminar en cada paso la j-ésima variable del modelo M que será el que tenga las variables que aún no se hayan eliminado en dicho paso.

Éste método tiene el inconveniente del número de cálculos que necesita. Además es irreversible, en el sentido de que una vez eliminada una variable, ésta no vuelve a entrar jamás en el modelo; y podría ocurrir que una variable recupere su poder explicativo en función de otras.

Método forward o de introducción progresiva

Es un método análogo al anterior, pero en este caso pretende obtener el mejor ajuste de regresión partiendo del modelo que incluye sólo al término independiente.

En primer lugar, consideramos las p variables disponibles y ajustamos el modelo con sólo el término independiente. A continuación se examinan todas las variables candidatas a entrar en el modelo individualmente, ajustando todas las regresiones Y/X_j; $(j = 1, \ldots, p)$ para ver cuál sería la candidata a entrar en el modelo, mediante el contraste condicional de razón de verosimilitudes que tiene en la hipótesis nula el modelo con sólo el término independiente y la hipótesis alternativa el modelo que añade la variable correspondiente. Elegimos como candidata a entrar, aquella que tenga el mínimo p-valor de entre todos los p-valores menores o iguales que un nivel de significación elegido α. Si no hay ninguna variable que cumpla esta condición el modelo con sólo el término independiente explica mejor la variable respuesta que en presencia de otras y paramos el método. En caso contrario, la incluímos en el modelo y contrastamos en el siguiente paso la posibilidad de incluir cada una de las restantes variables en el modelo seleccionado en el paso anterior. El procedimiento continúa hasta que encontremos que no hay ninguna variable candidata a entrar en el modelo. La Tabla 1.2 resume el algoritmo de ejecución del método *forward* donde M_j es el modelo resultante de introducir la j-ésima variable en el modelo M, en cada iteración del algoritmo. Este método, al igual que el anterior, necesita muchos cálculos y es irreversible pero en este caso en el sentido de que una vez que una variable entra en el modelo, ésta no vuelve a salir jamás del mismo; y podría ocurrir que una variable pierda su poder explicativo en función de otras.

Método stepwise

Este método aglutina a los dos anteriores, en el sentido de que en cada paso empieza con un método *forward*, y una vez que se ha reajustado el modelo con las variables que hay hasta ese momento, se examina la candidata a salir del modelo antes de volver a probar cuál debería entrar. En este sentido, la elección de los niveles de significación deben ser elegidos cuidadosamente de manera que la variable que entra en un paso no sea eliminada en el mismo paso en la comprobación *backward*. Por eso en general se suele escoger una significación para entrar menor que para salir.

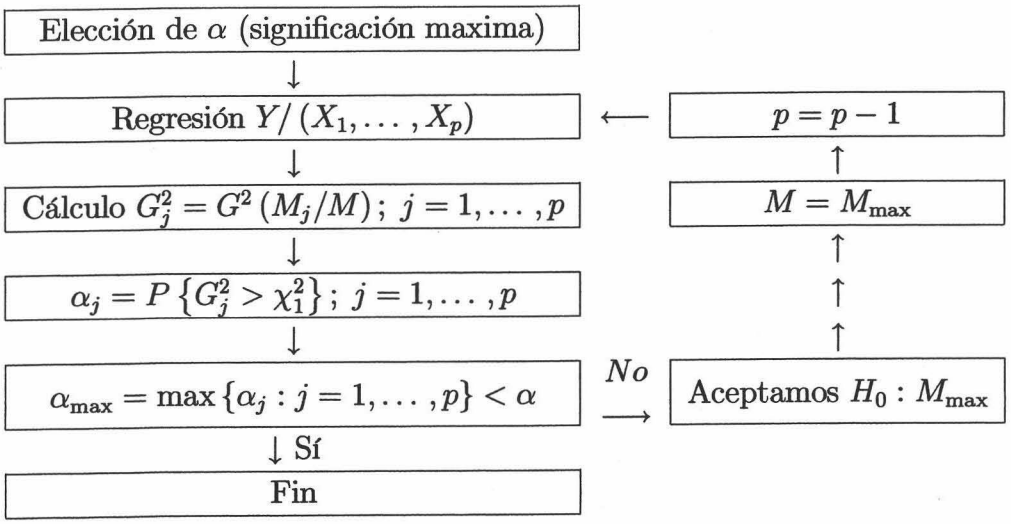


Tabla 1.1: Método Backward o de eliminación progresiva

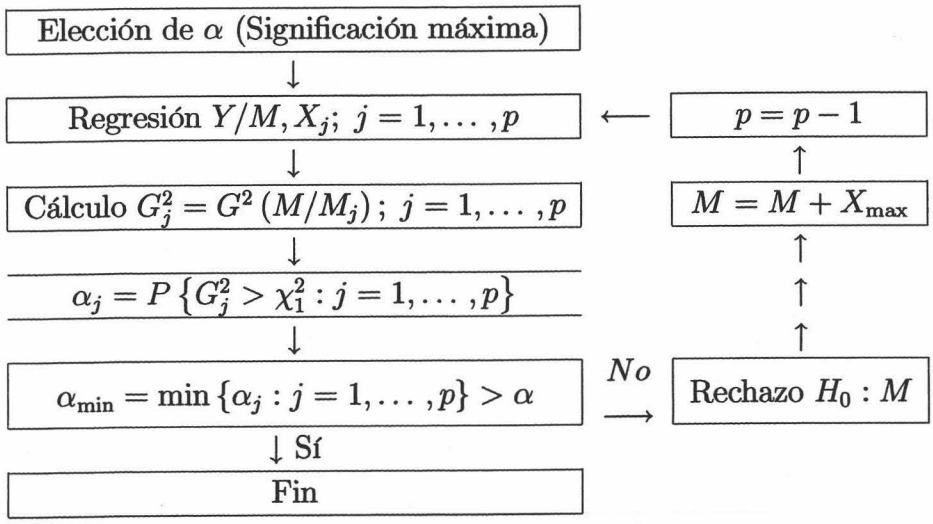


Tabla 1.2: Método Forward o introducción progresiva

1.2.6 Validación del modelo

Una vez estimados los parámetros y contrastada la significación de las covariables, el siguiente paso es la validación del modelo seleccionado, esto es, el análisis de los residuos para estudiar si se verifican las hipótesis del modelo e identificar posibles inconsistencias derivadas de la no verificación de dichas hipótesis. La manera habitual de validar un modelo es mediante el análisis de los residuos, esto es, una medida de la diferencia entre valores observados y ajustados. Estos residuos ponen de manifiesto falta de ajuste, mala elección del modelo, etc. Además los residuos también pueden detectar observaciones extrañas que perjudiquen al modelo. Para el caso de regresión logística se puede elegir entre dos tipos de residuos: residuos de tipo Pearson y los residuos de tipo *deviance*.

Residuos

En el caso lineal los residuos se definen como la diferencia entre el valor observado de la variable dependiente y el valor predicho por el modelo. Para el caso de la regresión logística los valores predichos se definen como

$$\widehat{y}_i = \widehat{\pi}_i,\ i = 1, \ldots, n,$$

de manera que si definimos el vector de valores predichos como

$$\widehat{\pi}(X) = (\widehat{\pi}_1, \widehat{\pi}_2, \ldots, \widehat{\pi}_n)^T,$$

se tiene que los valores predichos verifican

$$X^T (Y - \widehat{\pi}(X)) = 0.$$

Ya estamos en condiciones de definir los residuos siguiendo la misma filosofía que en regresión lineal de modo que vendrán dados por

$$e = Y - \widehat{\pi}(X),$$

y uno de ellos en particular

$$e_i = y_i - \widehat{\pi}_i,\ i = 1, \ldots, n.$$

Sin embargo los residuos que se suelen utilizar en regresión logística no son estos sino aquéllos cuya suma de cuadrados sea el estadístico global de bondad de ajuste del modelo (estadístico chi-cuadrado de Pearson o estadístico de Wilks de razón de verosimilitudes) (ver por ejemplo Ryan, 1997).

Residuos de Pearson

Definición 2 *Se definen los residuos de tipo Pearson como*

$$r_i = \frac{y_i - \widehat{\pi}_i}{\sqrt{\widehat{\pi}_i\left(1-\widehat{\pi}_i\right)}}.$$

Se puede observar que la suma de los residuos al cuadrado coincide con el estadístico chi-cuadrado de Pearson

$$X^2 = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \frac{\left(y_i - \widehat{\pi}_i\right)^2}{\widehat{\pi}_i\left(1-\widehat{\pi}_i\right)}.$$

Residuos de la deviance

Definición 3 *Se definen los residuos de la deviance como*

$$d_i = sign\left(y_i - \widehat{\pi}_i\right)\left\{\left(2\left[y_i \ln\frac{y_i}{\widehat{\pi}_i} + (1-y_i)\ln\frac{1-y_i}{1-\widehat{\pi}_i}\right]\right)^{1/2}\right\}$$

Cada residuo mide la desviación entre la correspondiente componente de la log-verosimilitud del modelo ajustado y la correpondiente componente de la log-verosimilitud que resultaría si cada punto se ajustara perfectamente (modelo saturado).

La suma de los cuadrados de los residuos de la *deviance* es el estadístico G^2 de razón de verosimilitudes

$$G^2 = \sum_{i=1}^{n} d_i^2.$$

Gráficos de los residuos

Además del cálculo de los residuos, para la validación del modelo son útiles algunos gráficos de los mismos que nos informen sobre las hipótesis del modelo. En este sentido se ha propuesto un gráfico de $r_{is}^2 = r_i^2/(1-h_{ii})$ frente a $\widehat{\pi}_i$ (ver Ryan (1997)) donde h_{ii} es el i-ésimo *leverage*, que en regresión logística se define como el i-ésimo elemento diagonal de

$$H = \widehat{W}^{1/2} X \left(X^T \widehat{W} X\right)^{-1} X^T \widehat{W}^{1/2}$$

siendo $\hat{W} = diag\left(\widehat{\pi}_i\left(1-\widehat{\pi}_i\right)\right)$. El motivo de este gráfico es que r_{is}^2 es aproximadamente el cambio que se obtiene en el estadístico chi-cuadrado de Pearson cuando se elimina la i-ésima observación.

Residuos estandarizados

A los residuos

$$r_{is} = \frac{r_i}{\sqrt{1 - h_{ii}}}$$

se les llama residuos de Pearson estandarizados y bajo la hipótesis nula

$$H_0 : r_i = 0$$

tienen distribución $N(0, 1)$. Por ello cada residuo de Pearson estandarizado es utilizado como estadístico del contraste de significación de su residuo asociado, de modo que se rechazará la hipótesis nula de igualdad a cero de dicho residuo cuando $|r_{is}| \geq z_{\alpha/2}$.

De igual manera, las cantidades

$$d_{is} = \frac{d_i}{\sqrt{1 - h_{ii}}}$$

reciben el nombre de residuos de la *deviance* estandarizados y bajo la hipótesis nula

$$H_0 : d_i = 0$$

tienen distribución asintótica $N(0, 1)$, de modo que se utilizan como estadísticos del contraste de significación de los residuos de la *deviance*. Rechazar H_0 implica que hay falta de ajuste en la i-ésima observación, o lo que es igual, que la i-ésima observación es un *outlier*.

Los residuos más utilizados y preferidos son los de la *deviance* ya que la distribución asintótica se aproxima más rápidamente a la normalidad que los Pearson, aunque estos últimos tiene la típica forma de aproximación de una distribución binomial a una normal.

La siguiente es otra modificación que se ha propuesto de los residuos de la *deviance* (ver Ryan, 1997) ya que tienen distribución asintótica normal incluso con tamaños muestrales pequeños: son los residuos de la *deviance* modificados definidos como

$$d_i^* = d_i + \frac{1 - 2\widehat{\pi}_i}{\left[\widehat{\pi}_i \left(1 - \widehat{\pi}_i\right)\right]^{1/2}}.$$

Medidas de influencia

Cuando en los modelos de regresión nos encontramos con puntos que se situan lejos del resto suele ocurrir que este hecho influya de manera desordenada en las estimaciones de los parámetros del modelo. Lo ideal será detectar dichos puntos para eliminarlos y evitar así dicha influencia desordenada sobre las estimaciones de los parámetros. Para detectar dichos puntos existen varias medidas como leverages, definidos previamente, y distancias de Cook. Veamos esta última medida para el modelo de regresión logística.

Distancia de Cook

Podríamos adaptar el estadístico D de Cook del modelo lineal (ver Draper y Smith, 1980) a regresión logística cuya definición sería

$$D_i = \frac{1}{p+1} r_{is}^2 \left(\frac{h_{ii}}{1-h_{ii}} \right).$$

Una modificación de este mismo estadístico es la siguiente (ver Ryan, 1997):

$$D_i^* = \left[\frac{n-p-1}{p+1} \frac{h_{ii}}{1-h_{ii}} \right]^{1/2} |r_{is}^*|$$

donde

$$r_{is}^* = \frac{y_i - \widehat{\pi}_{(i)}}{\sqrt{\widehat{\pi}_{(i)} \left(1 - \widehat{\pi}_{(i)}\right)} \sqrt{1-h_{ii}}}$$

siendo $\widehat{\pi}_{(i)}$ el valor estimado al eliminar la observación i-ésima.

1.3 Análisis en componentes principales

El análisis en componentes principales es una técnica multivariante que tiene como objetivo reducir la dimensión de un problema mediante la construcción de combinaciones lineales de la variables del mismo con varianza máxima e incorreladas, de modo que la variabilidad de las variables es explicada por las componentes principales de mayor varianza.

Los modelos de regresión en componentes principales (Massy, 1965) sustituyen las variables explicativas de un problema de regresión por un subconjunto óptimo de sus componentes principales, persiguiendo un doble objetivo:

- Reducir la dimensión del problema, explicando la respuesta con el menor número posible de variables explicativas. Para ello sustituye el conjunto original de covariables por un conjunto de predictores incorrelados, con el consiguiente ahorro que se produce en la estimación e interpretación de los modelos, así como la eficacia en la elección del mejor modelo.

- Estimar con precisión los parámetros del modelo evitando el problema de la multicolinealidad entre las variables explicativas.

Veamos en primer lugar en qué consiste el ACP como técnica multivariante para posteriormente aplicarlo al modelo de regresión logística. Obviaremos el caso teórico para pasar directamente al aspecto muestral.

Obtención de las componentes principales

Como hemos comentado previamente, pretendemos reducir la dimensión de un problema mediante combinaciones lineales de las variables originales de dicho problema. El ACP va a utilizar determinadas combinaciones lineales que cumplan ciertas condiciones de optimalidad.

Consideremos $X_1, \ldots, X_p$ vectores n-dimensionales que se pueden ver como una muestra de n observaciones de un vector p-dimensional de covariables de un modelo de regresión. Todas estas observaciones se pueden resumir en una matriz de la forma

$$\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

cuyas columnas corresponden a los vectores X_j, $j = 1, \ldots, p$. Las filas de la matriz $\mathcal{X}$ se pueden ver como n vectores p-dimensionales

$$\left\{ x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T ;\ i = 1, \ldots, n \right\}$$

asociados a las observaciones de un conjunto de variables explicativas de un modelo de regresión para cada individuo muestral. De cara a la obtención de las componentes principales (cc.pp.) consideraremos sin pérdida de generalidad que las variables son centradas $\sum_{i=1}^{n} x_{ij} = 0$, $\forall j = 1, \ldots, p$. En otro caso se trabaja con la matriz centrada $\mathcal{X}^*$ cuyas filas son $x_i^* = x_i - \overline{x}$.

Las componentes principales son transformaciones lineales de las variables originales de la forma

$$Z_c = c_1 X_1 + \cdots + c_p X_p = \mathcal{X}c = \left(c^T x_1, c^T x_2, \ldots, c^T x_n\right)^T$$

con $c \in \mathbb{R}^p$, y sus componentes definidas en la forma

$$c^T x_i = c_1 x_{i1} + c_2 x_{i2} + \cdots + c_p x_{ip}.$$

Veamos los momentos muestrales asociados a las combinaciones lineales así construidas

- **Media**

$$\overline{x}_c = \frac{c^T x_1 + c^T x_2 + \cdots + c^T x_n}{n} = c^T \overline{x}$$

con

$$\overline{x} = (\overline{x}_1, \ldots, \overline{x}_p)^T \ y \ \overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \ j = 1, \ldots, p.$$

- **Cuasivarianza.**

$$S_c^2 = \frac{1}{n-1}\left[\left(c^T x_1 - c^T \overline{x}\right)^2 + \cdots + \left(c^T x_n - c^T \overline{x}\right)^2\right] = c^T S c$$

donde S es la matriz de varianzas-covarianzas muestrales.

- **Cuasi-covarianza.** Consideremos otra transformación lineal de la muestra $Z_b = \left(b^T x_1, b^T x_2, \ldots, b^T x_n\right)$ con media y cuasivarianza $\overline{x}_b = b^T \overline{x}$ y $S_b^2 = b^T S b$ respectivamente. La cuasi-covarianza será

$$S_{cb}^2 = \frac{1}{n-1}\left[\left(c^T x_1 - c^T \overline{x}\right)\left(b^T x_1 - b^T \overline{x}\right) + \cdots + \left(c^T x_n - c^T \overline{x}\right)\left(b^T x_n - b^T \overline{x}\right)\right] = c^T S b$$

A continuación se definen las componentes principales como las combinaciones lineales que tienen varianza máxima, están normalizadas a la unidad y son incorreladas.

Definición 4 *Dada una muestra de observaciones de p variables como la definida anteriormente que resumiremos en la matriz $\mathcal{X}$ de dimensión $n \times p$,*

- *se define la primera componente principal como la combinación lineal* $Z_1 = \mathcal{X}\mathcal{V}_1$ *t.q. el vector de coeficientes* $\mathcal{V}_1$ *se obtiene como solución al problema de máximos*

$$\max_{\{l:l^T l=1\}} l^T S l = \mathcal{V}_1^T S \mathcal{V}_1.$$

- *Dadas las* $j-1$ *primeras componentes principales,* $Z_1 = \mathcal{X}\mathcal{V}_1, \ldots, Z_{j-1} = \mathcal{X}\mathcal{V}_{j-1}$, *se define la* j*-ésima componente principal como* $Z_j = \mathcal{X}\mathcal{V}_j$ *tal que*

$$\max_{\left\{l:l^T l=1, \mathcal{V}_k^T S l=0, \forall k=1,\ldots,j-1\right\}} l^T S l = \mathcal{V}_j^T S \mathcal{V}_j \tag{1.7}$$

La condición de normalización se exige para obtener unicidad en los problemas de máximos planteados, ya que si $\mathcal{V}_j$ es tal que $\max\left\{l^T S l\right\} = \mathcal{V}_j^T S \mathcal{V}_j$ entonces $r\mathcal{V}_j$ con $r \in \mathbb{R}$ también hace máxima esa cantidad.

La definición anterior proporciona un método recursivo para la obtención de las componentes principales. El siguiente teorema, cuya demostración puede verse en cualquier texto de análisis de datos multivariantes (por ejemplo en Basilevsky, 1999), proporciona las soluciones de los problemas de máximos planteados.

Teorema 3 *Los vectores* $\mathcal{V}_j$ *soluciones del problema de máximos (1.7) que definen a la componentes principales son las soluciones del problema de valores propios*

$$Sl = \lambda l$$

donde S *es la matriz de varainzas-covarianzas muestral de* $\mathcal{X}$. *Además el valor de dicho máximo es*

$$l^T S l = \lambda.$$

De la obtención de las componentes principales y del hecho de que la matriz de covarianzas muestral sea semidefinida positiva, se tiene que los autovalores (varianzas de las componentes principales) están ordenados aunque no de manera estricta: $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$. Para los autovalores distintos (multiplicidad uno) se tiene que los autovectores son únicos salvo cambio de signo, sin embargo los correspondientes a autovalores con multiplicidad mayor que uno no son únicos (Basilevski, 1999) de modo que las cc.pp. asociadas tampoco lo son.

Por otro lado, y también de la obtención y definición de las cc.pp., se tiene que los autovectores asociados a los autovalores anteriores, son ortonormales (Basilevski, 1999), lo que proporcionará las buenas propiedades que veremos a continuación.

Representación en componentes principales

Una vez obtenidas las componentes principales, éstas se pueden expresar matricialmente en la forma

$$\mathcal{Z} = \mathcal{X}\mathcal{V}$$

con

$$\mathcal{V} = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1p} \\ v_{21} & v_{22} & \cdots & v_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ v_{p1} & v_{p2} & \cdots & v_{pp} \end{pmatrix}$$

la matriz ortogonal que tiene por columnas los p vectores propios de S que reciben el nombre de vectores principales.

El ACP de una muestra de observaciones de un vector aleatorio permite la expresión de las observaciones de cada una de las variables del vector en términos de las observaciones de un conjunto de variables incorreladas, así como la diagonalización de la matriz de varianzas-covarianzas. Ambas propiedades se resumen en el siguiente resultado:

Teorema 4 *Dada una matriz de observaciones de un conjunto de variables* $\mathcal{X} = (X_1, \ldots, X_p)$ *y su matriz de componentes principales* $\mathcal{Z} = (Z_1, \ldots, Z_p)$,

1. *La matriz de covarianzas se factoriza de la forma*

$$S = \mathcal{V}\Delta\mathcal{V}^T.$$

con $\Delta = diag\,(\lambda_1, \ldots, \lambda_p)$ *y* $\mathcal{V}$ *la matriz que tiene por columnas a los vectores principales.*

2. *Podemos reconstruir la matriz de observaciones de la forma*

$$\mathcal{X} = \mathcal{Z}\mathcal{V}^T$$

y cada vector X_j, *mediante la siguiente representación en términos de las cc.pp* Z_k, $k = 1, \ldots, p$

$$X_j = \sum_{k=1}^{p} Z_k v_{jk}$$

siendo v_{jk} *el correspondiente elemento de* $\mathcal{V}$.

Esta representación de las trayectorias es óptima en el sentido de ser la que minimiza el error cuadrático medio

$$\frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \widetilde{x}_{ij})^2 \, , \; j = 1, \ldots, p$$

de entre todas las representaciones lineales mediante variables incorreladas de la forma

$$\widetilde{X}_j = \sum_{k=1}^{p} \alpha_k w_{jk}.$$

Ésta y otras propiedades de optimalidad del ACP se resumen en el siguiente teorema y su demostración puede verse en Basilevski (1999):

Teorema 5 *Sean* $X_1, \ldots, X_p$ *un conjunto de vv.aa. Entonces el ACP tiene las siguientes propiedades de optimalidad:*

1. *Las componentes principales maximizan la traza total (varianza univariante) de las* X_i *dada por* $tr(S) = \sum_{i=1}^{p} \lambda_i$.

2. *Las componentes principales maximizan la varianza generalizada de las* X_i *dada por* $|S| = \prod_{i=1}^{p} \lambda_i$.

3. *Las componentes principales minimizan la entropía total, esto es, maximizan la información contenida en las variables definida por*

$$I = \sum_{i=1}^{p} \lambda_i ln \lambda_i.$$

4. *Las componentes principales maximizan la distancia euclídea.*

5. *Las componentes principales minimizan el error cuadrático medio.*

Al principio de la exposición del tema, hacíamos referencia a dos objetivos del ACP: la reducción de dimensión y la obtención de variables incorreladas para resolver el problema de multicolinealidad en los modelos de regresión. Hasta ahora lo que hemos conseguido son las variables incorreladas. La propiedad que ahora introducimos nos resuelve el segundo objetivo del ACP: la reducción de dimensión.

Consideremos la factorización obtenida anteriormente de la matriz de varianzas-covarianzas mediante ACP, entonces

$$tr\,(S) = tr\,(\Delta)$$

de manera que

$$\sum_{i=1}^{p} Var\,[X_i] = \sum_{i=1}^{p} Var\,[Z_i] = \sum_{i=1}^{p} \lambda_i$$

con lo que la suma de las varianzas de las variables originales coincide con la suma de los autovalores de su matriz de varianzas-covarianzas y con la suma de las varianzas de las componentes principales obtenidas a partir de ellas.

Consideremos la cantidad

$$\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i},\ j = 1, \ldots, p,$$

esta cantidad se puede ver como el tanto por uno del total de varianza de las variables originales que es explicado por cada componente principal.

Ya que las cc.pp. son incorreladas, se tiene que considerando un número reducido s de cc. pp., $(s \leq p)$

$$Var\left[\sum_{i=1}^{s} Z_i\right] = \sum_{i=1}^{s} Var\,[Z_i] = \sum_{i=1}^{s} \lambda_i,$$

de manera que con las s primeras componentes principales se consigue explicar el siguiente porcentaje

$$\left(\frac{\sum_{i=1}^{s} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \times 100\right)$$

del total de variabilidad presente en el experimento que estemos considerando. En la medida en que ese porcentaje sea alto, así de buena será la explicación que se dé al problema que involucra a $\mathcal{X}$, con las s primeras cc.pp.

A partir del teorema de representación visto anteriormente y de esta propiedad se obtiene el siguiente resultado de optimalidad:

Teorema 6 *La siguiente representación aproximada de las observaciones de cada una de las variables de $\mathcal{X}$ mediante las s primeras componentes principales extraidas de $\mathcal{X}$ es óptima en el sentido de minimizar el error cuadrático medio*

$$X_j^s = \sum_{k=1}^{s} Z_k v_{jk} \quad s \leq p,\ j = 1, \ldots, p.$$

Para terminar con las buenas propiedades que proporcionan las componentes principales no podemos olvidar la que permite invertir la matriz de varianzas-covarianzas de una muestra de observaciones $\mathcal{X}$. Así, considerando la matriz $U = \mathcal{V}\Delta^{-1/2}$ donde $\mathcal{V}$ y Δ son las definidas anteriormente, entonces

$$UU^T = S^{-1}.$$

Esta propiedad viene a resolver uno de los problemas planteados al principio de la presente sección: la multicolinealidad, esto es, la alta dependencia existente en las observaciones de una muestra de variables $\mathcal{X}$ que hace que su matriz de varianzas-covarianzas esté mal condicionada y por tanto que su inversa no exista. Inversa que por otro lado será necesaria en la estimación de determinados modelos como es el caso de los de regresión lineal y logística múltiples.

Interpretación de las componentes principales

El problema de las componentes principales es que se trata de combinaciones lineales artificiales de las variables originales, de modo que no son directamente observables. Por ello de cara a las aplicaciones es fundamental su interpretación, para lo que resulta útil conocer la relación lineal entre las componentes principales y las variables originales. Para ello estudiaremos la matriz de correlaciones entre ambas.

Matricialmente, es claro que la matriz de covarianzas entre variables y componentes principales se puede expresar de la forma

$$Cov\,[X, Z] = S\mathcal{V} = \mathcal{V}\Delta\mathcal{V}^T\mathcal{V} = \mathcal{V}\Delta,$$

de modo que la covarianza entre la j-ésima variable y la k-ésima componente principal es

$$Cov[X_j, Z_k] = \lambda_k v_{jk}.$$

Esto implica que la correlación viene dada por

$$Corr[X_j, Z_k] = \frac{\lambda_k v_{jk}}{\sqrt{S_{jj}\lambda_k}} = \left(\frac{\lambda_k}{S_{jj}}\right)^{1/2} v_{jk}.$$

Otro problema relacionado con la puesta en práctica del análisis en componentes principales es la selección del número óptimo de componentes principales necesarias para la reconstrucción de las variables originales. A pesar de que han sido desarrollados diversos criterios teóricos de selección, los más extendidos en la práctica son los criterios empíricos basados en reglas puramente intuitivas. El más usual es elegir un punto de corte (cut-off), usualmente entre el 70 y 95%, y retener las s primeras componentes principales cuyo porcentaje de varianza acumulada sea superior o igual a dicho punto de corte.

1.4 Modelo de regresión logística en términos de las componentes principales

Como han constatado numerosos autores en la literatura, (Hosmer & Lemeshow (1989), Ryan (1997)), el análisis de regresión logística múltiple presenta algunos problemas derivados de la matriz de diseño, que hacen que los resultados obtenidos en dichos modelos sean inconsistentes. El primero y quizás más grave es el problema de la multicolinealidad. Dicho problema consiste en que las variables predictoras estén muy correlacionadas entre sí, lo cual es esperable cuando se trata de investigar la relación entre un conjunto de variables con una variable respuesta, ya que si no todas, muchas de ellas son variables muy cercanas en cuanto a ámbito de estudio. Este problema provoca que la matriz de covarianzas esté mal condicionada y por tanto su inversión presente problemas computacionales. En este sentido sería interesante tener un conjunto de variables "auxiliares" incorreladas que permitieran estimar los parámetros del modelo evitando todos los problemas que provoca la multicolinealidad.

Un segundo problema a abordar en modelos de regresión consiste en el tamaño del experimento. Cuando el número de variables a analizar es muy grande, la selección de las mismas suele ser compleja, por lo que sería interesante reducir en lo posible el número de variables explicativas a utilizar de manera que se consiguiera explicar la práctica totalidad de la variabilidad presente en el modelo, con un conjunto reducido de regresores.

Como solución para los dos problemas comentados anteriormente, proponemos a continuación utilizar como regresores del modelo de regresión logística, las componentes principales de las variables explicativas. De este modo, se evita la multicolinealidad ya que las componentes principales son incorreladas y además se resuelve el problema de la dimensión al poder considerar un número reducido de componentes en lugar de todas las variables. En el capí-

tulo tres se desarrollará un estudio de simulación que evaluará la capacidad de las cc. pp. para resolver de forma eficaz los problemas antes mencionados.

Consideremos que disponemos de un conjunto de posibles regresores o variables explicativas $X_1, X_2, ..., X_p$ y la variable respuesta habitual Y. El modelo de regresión logística se expresaba de la forma

$$Y = \pi(X) + \varepsilon$$

con las matrices, vectores e hipótesis habituales. Supondremos, sin pérdida de generalidad, que la matriz $\mathcal{X}$ de observaciones de las variables explicativas es centrada en media, esto es, que $\sum_{i=1}^{n} x_{ij} = 0,\ \forall j = 1, \dots, p$.

Sea $\mathcal{Z} = \mathcal{X}\mathcal{V}$ la matriz de las componentes principales de $\mathcal{X}$ (*scores*) con $\mathcal{V}$ la matriz que tiene por columnas los autovectores de $\mathcal{X}^T\mathcal{X}$. Como se puede ver en Basilevsky (1983), los autovectores de una matriz no varían si multiplicamos dicha matriz por un escalar, con lo que consideraremos los autovectores de la matriz anterior en lugar de los de la matriz de covarianzas que se obtendría de la anterior dividiendo por $n-1$.

El modelo de regresión logística en términos de las componentes principales consiste en tomar como matriz de diseño del modelo

$$Z = XV$$

donde X es la matriz de diseño del modelo de regresión logística múltiple en términos de las variables originales y V es la siguiente matriz por cajas

$$V = \left(\begin{array}{c|c} 1 & 0_{1\times p} \\ \hline 0_{p\times 1} & \mathcal{V} \end{array}\right).$$

Esta matriz también es ortogonal ya que

$$V^TV = \left(\begin{array}{c|c} 1 & 0_{1\times p} \\ \hline 0_{p\times 1} & \mathcal{V}^T\mathcal{V} \end{array}\right) = I_{p+1}.$$

Además también diagonaliza a X^TX ya que

$$V^TX^TXV = \left(\begin{array}{c|c} n & 0_{1\times p} \\ \hline 0_{p\times 1} & \mathcal{V}^T\mathcal{X}^T\mathcal{X}\mathcal{V} \end{array}\right) = \left(\begin{array}{c|c} n & 0_{1\times p} \\ \hline 0_{p\times 1} & \Delta \end{array}\right) = \Lambda.$$

Entonces el modelo de regresión logística múltiple, considerando como regresores las componentes principales, es de la forma

$$Y = \pi(Z) + \varepsilon \tag{1.8}$$

con

$$\pi(Z) = (\pi_1(Z), \pi_2(Z), \ldots, \pi_n(Z))^T$$

y

$$\pi_i(Z) = \frac{\exp\{z_i^T\gamma\}}{1+\exp\{z_i^T\gamma\}},\ i = 1, \ldots, n$$

siendo z_i^T la i-ésima fila de la matriz Z y $\gamma = (\gamma_0, \ldots, \gamma_p)^T$ el vector de parámetros a estimar.

Como $Z = XV$, la i-ésima fila de esta matriz será

$$z_i^T = x_i^T V$$

con x_i^T la i-ésima fila de la matriz de diseño original. Por lo tanto, se tiene que las probabilidades del modelo en términos de las cc. pp.

$$\pi_i(Z) = \frac{\exp\{z_i^T\gamma\}}{1+\exp\{z_i^T\gamma\}} = \frac{\exp\{x_i^T V\gamma\}}{1+\exp\{x_i^T V\gamma\}} = \frac{\exp\{x_i^T\beta\}}{1+\exp\{x_i^T\beta\}} = \pi_i,\ i = 1, \ldots, n$$

sin más que tomar $\beta = V\gamma$, por lo que las probabilidades del modelo en función de todas las componentes principales y del modelo en función de las variables originales son las mismas.

Equivalentemente, en su forma lineal para las transformaciones logit el modelo de regresión logística en términos de las componentes principales se puede expresar como

$$L = Z\gamma = XV\gamma = X\beta,$$

siendo $L = (l_1, l_2, \ldots, l_n)^T$ el vector cuyas componentes son las transformaciones logit $l_i = \ln[\pi_i/(1-\pi_i)]$, $i = 1, \ldots, n$.

Estimación de parámetros

Consideremos ahora las ecuaciones de verosimilitud del modelo en función de las componentes principales

$$Z^T(Y - \widehat{\pi}(Z)) = 0$$

con

$$\widehat{\pi}_i(Z) = \frac{\exp\{z_i^T\widehat{\gamma}\}}{1+\exp\{z_i^T\widehat{\gamma}\}},\ i = 1, \ldots, n.$$

Entonces esas ecuaciones se pueden expresar de la forma

$$Z^T\left(Y-\widehat{\pi}\left(Z\right)\right)=\left(XV\right)^T\left(Y-\widehat{\pi}\left(Z\right)\right)=V^TX^T\left(Y-\widehat{\pi}\left(Z\right)\right)=0$$

de manera que

$$Z^T\left(Y-\widehat{\pi}\left(Z\right)\right)=0\Leftrightarrow V^TX^T\left(Y-\widehat{\pi}\left(Z\right)\right)=0\Leftrightarrow X^T\left(Y-\widehat{\pi}\left(Z\right)\right)=0$$

que lleva a que en caso de solución única, las probabilidades predichas por el modelo de regresión logística sobre las cc.pp. son las mismas que las predichas por el modelo sobre las variables originales, lo que además es inmediato como se desprende de la propiedad de invarianza de los estimadores de máxima verosimilitud (Rohatgi, 1984).

Por otro lado, si estimamos el modelo de regresión logística múltiple en función de todas las componentes principales, y obtenemos $\widehat{\gamma}$, entonces podemos reconstruir el estimador de máxima verosimilitud del vector de parámetros en términos de las variables originales de la forma

$$\widehat{\beta}=V\widehat{\gamma},$$

de modo que las probabilidades predichas se obtienen en la forma

$$\widehat{\pi}_i\left(Z\right)=\frac{\exp\left\{z_i^T\widehat{\gamma}\right\}}{1+\exp\left\{z_i^T\widehat{\gamma}\right\}}=\frac{\exp\left\{x_i^T\widehat{\beta}\right\}}{1+\exp\left\{x_i^T\widehat{\beta}\right\}}=\widehat{\pi}_i,\ i=1,\ldots,n.$$

El estimador $\widehat{\gamma}$ conserva las propiedades asintóticas de los estimadores de máxima verosimilitud ya estudiados para $\widehat{\beta}$. Es decir,

$$\widehat{\gamma}\underset{n\to\infty}{\rightsquigarrow}N\left(\gamma,Cov\left[\widehat{\gamma}\right]\right)$$

siendo

$$Cov\left[\widehat{\gamma}\right]=\left(Z^TWZ\right)^{-1}=\left(V^TX^TWXV\right)^{-1}.$$

1.5 Modelo de regresión logística en componentes principales

Hasta ahora hemos considerado el modelo de regresión habitual en función de todas las componentes principales, en el que las nuevas observaciones de los

regresores son ortogonales. Veamos a continuación que el modelo de regresión logística en cc. pp. consiste en considerar como regresores sólo algunas de las cc. pp. y no todas.

Consideremos la matriz de diseño del modelo en términos de todas las componentes principales, la de autovectores y el vector de parámetros del modelo dispuestos en cajas de la forma

$$\begin{aligned} Z &= \left(\begin{array}{cccc|ccc} 1 & z_{11} & \cdots & z_{1s} & z_{1s+1} & \cdots & z_{1p} \\ 1 & z_{21} & \cdots & z_{2s} & z_{2s+1} & \cdots & z_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & z_{n1} & \cdots & z_{ns} & z_{ns+1} & \cdots & z_{np} \end{array}\right) \\ &= \left(\begin{array}{c|c} Z_{(s)} & Z_{(r)} \end{array}\right) = \left(XV_{(s)}|XV_{(r)}\right), \\ V &= \left(\begin{array}{cccc|ccc} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & v_{11} & \cdots & v_{1s} & v_{1s+1} & \cdots & v_{1p} \\ 0 & v_{21} & \cdots & v_{2s} & v_{2s+1} & \cdots & v_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & v_{p1} & \cdots & v_{ps} & v_{ps+1} & \cdots & v_{pp} \end{array}\right) = \left(\begin{array}{c|c} V_{(s)} & V_{(r)} \end{array}\right) \\ \gamma &= \left(\gamma_0, \gamma_1, \ldots, \gamma_s | \gamma_{s+1}, \ldots, \gamma_p\right)^T = \left(\frac{\gamma_{(s)}}{\gamma_{(r)}}\right), \; r = p - s. \end{aligned}$$

La expresión del vector de parámetros β (correspondiente al modelo en términos de las variables originales) se puede descomponer entonces en la forma

$$\beta = V\gamma = V_{(s)}\gamma_{(s)} + V_{(r)}\gamma_{(r)}, \tag{1.9}$$

y a su vez se tiene

$$\begin{aligned} \gamma_{(s)} &= (\gamma_0, \gamma_1, \ldots, \gamma_s)^T = V_{(s)}^T\beta \\ \gamma_{(r)} &= (\gamma_{s+1}, \ldots, \gamma_p)^T = V_{(r)}^T\beta. \end{aligned}$$

Observemos que de estas consideraciones se tiene que el modelo de regresión logística en términos de todas las componentes principales se puede expresar a partir de la transformación logit de la forma

$$L = Z\gamma = Z_{(s)}\gamma_{(s)} + Z_{(r)}\gamma_{(r)}$$

Nos proponemos ahora eliminar aquellas componentes principales menos explicativas tomando como regresores las s primeras cc. pp.. Si ajustamos el modelo de regresión logística con unas pocas componentes principales (las que

más variabilidad explican), las s primeras, obtenemos parámetros estimados distintos de las estimaciones de los primeros s parámetros obtenidos con todas las componentes principales. Aun así, se aprecia que se puede obtener una buena aproximación de las probabilidades con una reducción interesante de la dimensión del problema como se podrá apreciar en los ejemploa simuladoa que se incluyen en el Capítulo 3.

Entonces, el modelo de regresión logística en componentes principales será

$$Y = \pi\left(Z_{(s)}\right) + \varepsilon_{(s)}$$

que se ha obtenido eliminando en el modelo (1.8) las componentes principales asociadas a los r menores autovalores, esto es

$$\pi_{i(s)} = \frac{\exp\left\{z_{i(s)}^T \gamma_{(s)}\right\}}{1 + \exp\left\{z_{i(s)}^T \gamma_{(s)}\right\}}, \; i = 1, \ldots, n$$

siendo $z_{i(s)}^T$ la i-ésima fila de la matriz $Z_{(s)}$, y $\gamma_{(s)}$ el vector formado por los primeros s parámetros.

El vector $L_{(s)}$ de transformaciones logit del modelo de regresión logística en términos de las s primeras componentes principales se puede escribir equivalentemente en la forma

$$L_{(s)} = Z_{(s)}\gamma_{(s)},$$

que, igual que en el caso de la regresión lineal en componentes principales, se puede escribir en términos de las variables originales como

$$L_{(s)} = Z_{(s)}\gamma_{(s)} = XV_{(s)}\gamma_{(s)} = X\beta_{(s)},$$

sin más que definir el vector de parámetros

$$\beta_{(s)} = V_{(s)}\gamma_{(s)},$$

que proporciona una reconstrucción aproximada del vector de parámetros poblacionales β en términos del vector de parámetros de las s primeras cc. pp..

Estimación de parámetros

Como hemos comentado previamente, es claro que el estimador de máxima verosimilitud $\widehat{\gamma}_{(s)}$ obtenido resolviendo de forma aproximada mediante Newton-Raphson las ecuaciones de verosimilitud

$$Z_{(s)}^T(Y - \widehat{\pi}(Z_{(s)})) = 0,$$

$$Z_{(s)}^T(Y - \widehat{\pi}(Z_{(s)})) = 0,$$

no tiene como componentes a los estimadores de máxima verosimilitud de los $s+1$ primeros parámetros del modelo con todas las componentes principales. Es decir,

$$\widehat{\gamma}_{(s)} \neq (\widehat{\gamma}_0, \widehat{\gamma}_1, \ldots, \widehat{\gamma}_s)^T$$

con lo que las probabilidades estimadas ajustando el modelo con las s primeras componentes son distintas de las probabilidades estimadas truncando el modelo que contiene todas las componentes principales en la s-ésima componente principal. Esto significa que

$$\widehat{l}_{i(s)} = Z_{i(s)}^T \widehat{\gamma}_{(s)} \neq \sum_{j=0}^{s} z_{ij} \widehat{\gamma}_j,$$

lo que implica reajustar el modelo cada vez que se introduce o elimina una componente principal a diferencia de lo que ocurre en el caso de la regresión lineal en cc. pp..

No obstante, podemos dar una estimación de los parámetros originales del modelo a partir de $\widehat{\gamma}_{(s)}$ que es más precisa que los propios $\widehat{\beta}$ cuando las variables originales están muy correladas, tomando

$$\widehat{\beta}_{(s)} = V_{(s)} \widehat{\gamma}_{(s)}.$$

Finalmente las probabilidades predichas en términos de las s primeras cc.pp. son de la forma

$$\widehat{\pi}_{i(s)} = \frac{\exp\left\{z_{i(s)}^T \widehat{\gamma}_{(s)}\right\}}{1+\exp\left\{z_{i(s)}^T \widehat{\gamma}_{(s)}\right\}} = \frac{\exp\left\{x_i^T \widehat{\beta}_{(s)}\right\}}{1+\exp\left\{x_i^T \widehat{\beta}_{(s)}\right\}}, \; i = 1, \ldots, n.$$

En lo que se refiere a las propiedades de los estimadores, vimos en el caso de regresión lineal que al estimar los parámetros del modelo utilizando los estimadores de las componentes principales se perdían algunas de las buenas propiedades de los mismos.

En el caso de la regresión logística el estimador de los parámetros de las variables originales a través de las componentes principales deja de ser asintóticamente insesgado ya que su esperanza asintótica será $V_{(s)}\gamma_{(s)} = \beta_{(s)} \neq \beta$.

1.6 Selección de componentes principales

Un aspecto esencial en la utilización de las componentes principales como covariables en un modelo de regresión, y tratado ampliamente en el caso lineal,

es la elección del método más idóneo para incluir componentes principales en el modelo. Numerosos autores han tratado este problema en el caso de la regresión lineal. Por ejemplo Aucott y Garthwaite (2000) consideran ideal la introducción de las componentes a partir de los tests condicionales de razón de verosimilitudes en su trabajo de comparación de distintos métodos de regresión en presencia de multicolinealidad, Mansfield et al (1977), Hocking (1976) y Gunst y Mason (1977) también han tratado diferentes métodos en el orden de inclusión de las componentes siendo el anteriormente citado el más popular junto con el método de inclusión de las componentes en orden a su explicación de la variabilidad total de las covariables originales.

En los ejemplos con datos simulados analizados en el Capítulo 3 consideraremos ambos métodos para el caso de la regresión logística lo cual no ha sido analizado antes en la literatura estadística. Llamaremos Método I a la inclusión de las componentes en orden a su porcentaje de varianza explicada, y Método II al de la inclusión en el orden que proporciona el método de selección *stepwise* basado en contrastes condicionales de razón de verosimilitudes poniéndose de manifiesto en todos los casos desarrollados la idoneidad del Método II.

Capítulo 2

Regresión logística funcional en componentes principales

2.1 Introducción

En las últimas décadas y gracias al desarrollo de las nuevas tecnologías ha variado el enfoque que se daba a determinados fenómenos, pasando de un tratamiento que históricamente había sido multivariante a un tratamiento funcional. Así, había fenómenos que a pesar de su naturaleza funcional era necesario darles un tratamiento multivariante por la falta de metodologías eficaces para el análisis a partir de su naturaleza funcional. Tal es el caso de la evolución temporal de magnitudes como determinados índices en un mercado de valores en el ámbito de la economía, o de las temperaturas y precipitaciones en el de la meteorología, o del crecimiento de los jóvenes en medicina, o del agujero de la capa de ozono en física.

A pesar de no ser usual en la práctica tener observaciones de forma continua de magnitudes que tienen esta naturaleza, sino más bien una serie de observaciones en distintos instantes de tiempo, resulta necesario idear técnicas que tengan presente esta naturaleza continua para el análisis de dichos fenómenos. El desarrollo tecnológico también ha posibilitado que podamos disponer de cantidades ingentes de información que hay que sintetizar para poder ser interpretada, por lo que también es necesario, además de examinar la perspectiva funcional de determinados fenómenos, reducir su dimensión sin perder demasiada información. Es por todo esto por lo que se han desarrollado en los últimos años técnicas como el Análisis en Componentes Principales Funcional (ACPF) que sirve a este objetivo de manera adecuada. Desde el

punto de vista metodológico esta teoría tiene su origen en el desarrollo de Karhunen-Loève (K-L) introducido en 1947 por Karhunen y que permite dar una representación ortogonal de las trayectorias de un proceso estocástico en términos de vv. aa. incorreladas y funciones determinísticas.

Además del Análisis en Componentes Principales (ACP) existen otras técnicas generalizables al campo funcional como es el modelo de regresión lineal. Así, cuando se dispone de una variable respuesta Y que queremos explicar a partir de una magnitud que evoluciona en el tiempo y que se modeliza a partir de un proceso estocástico en tiempo continuo, surge el modelo de regresión lineal funcional; tal es el caso de la situación en la que se quieren explicar las precipitaciones totales en un conjunto de estaciones meteorológicas a partir del patrón que presenta la evolución de las temperaturas en dichas estaciones (Ramsay y Silverman, 1997). Así, dado un conjunto de observaciones funcionales $x_1(t), \ldots, x_n(t)$, que constituyen una muestra aleatoria simple de un proceso estocástico $\{X(t) : t \in T\}$, y un conjunto de observaciones de la variable respuesta $y_1, \ldots, y_n$ asociadas a cada una de dichas trayectorias, el modelo de regresión lineal funcional se formula de la forma

$$y_i = \alpha + \int_T x_i(t)\,\beta(t)\,dt + \varepsilon_i,\ i = 1, \ldots, n,$$

siendo el principal problema la estimación de la función parámetro $\beta(t)$.

Para ajustar un modelo como el anterior se presentan varios problemas empezando porque usualmente no se dispone de la expresión explícita de las trayectorias $\{x_i(t),\ i = 1, \ldots, n\}$ sino de un conjunto de observaciones de dichas trayectorias en un conjunto de nodos $t_{i1}, \ldots, t_{im_i},\ i = 1, \ldots, n$; y terminando por el método a utilizar para obtener la estimación de la función parámetro. Para dicha estimación no son adecuados los métodos usuales del caso múltiple: mínimos cuadrados y máxima verosimilitud, como afirman Ramsay y Silverman (1997).

El presente capítulo tiene dos objetivos fundamentales: el primero es generalizar el modelo anteriormente descrito al caso de una variable respuesta dicotómica para la que, al igual que ocurre en el caso multivariante, no es adecuado un modelo lineal, y es por ello por lo que proponemos, como extensión del modelo de regresión logística múltiple, el modelo de regresión logística funcional que tendrá como objetivo estimar una variable binaria a partir de la evolución de una variable continua en el tiempo. El segundo es introducir los modelos de regresión logística funcional en componentes principales para reducir la dimensión del problema y conseguir una estimación precisa de la

función parámetro en presencia de multicolinealidad.

Este capítulo de divide en cinco secciones: la primera la compone esta introducción. La segunda introduce la teoría básica sobre Análisis de Datos Funcionales (ADF) así como los métodos más usuales de aproximación de trayectorias de un proceso en tiempo continuo a partir de la información discreta en un conjunto de nodos de observación. En la sección tres se formula el modelo de regresión logística funcional como extensión del modelo de regresión lineal funcional y el de regresión logística múltiple y se presentan los problemas que se plantean en la estimación de la función parámetro y algunas posibles soluciones. La sección cuarta proporciona un resumen del ACPF para una muestra aleatoria simple de realizaciones de un proceso estocástico, recordando la obtención de las componentes principales, la representación de las funciones muestrales en términos de las compoentes principales y la interpretación de éstas, y la aproximación de las mismas. Por último, en la sección cinco, se propone el modelo de regresión logística funcional en componentes principales como solución de los problemas de estimación de la función parámetro del modelo original a causa de la dependencia existente en los datos funcionales muestrales.

2.2 Análisis de datos funcionales y procesos estocásticos

Como se ha indicado anteriormente, existen muchos campos de aplicación donde los datos observados son funciones en lugar de los vectores del análisis multivariante, funciones que se pueden ver como realizaciones de un proceso estocástico. Igual que cuando se analizan fenómenos desde la perspectiva multivariante la herramienta idónea para su tratamiento son las variables aleatorias, los vectores aleatorios y familias numerables de variables aleatorias; cuando los fenómenos a investigar pertenecen al campo funcional, las herramientas probabilísticas que los representan son los procesos estocásticos que modelizan la evolución de una variable aleatoria en el tiempo. Veamos a continuación, y de forma resumida, algunas definiciones y resultados básicos referentes a procesos estocásticos.

Consideremos un espacio probabilístico $(\Omega, \mathcal{A}, P)$.

Definición 5 *Dado* $(H, \langle , \rangle_H)$ *un espacio de Hilbert separable cualquiera se*

define una variable aleatoria sobre H como una función medible

$$\begin{array}{rccl} X: & \Omega & \longrightarrow & H \\ & \omega & \longrightarrow & X(\omega), \end{array}$$

o lo que es igual, tal que $X^{-1}(B) \in \mathcal{A}$, siendo B un conjunto Borel de H, esto es, de la σ-álgebra de Borel generada por la topología del espacio H.

Como hemos visto en la introducción, el ámbito de esta memoria es el de los procesos estocásticos, y dentro de este campo tienen especial interés aquéllos cuyas funciones muestrales pertenecen al espacio de Hilbert de las funciones de cuadrado integrable sobre un intervalo real T. Por tanto, denotaremos por $L^2(T)$ al conjunto de las funciones reales de cuadrado integrable sobre T:

$$L^2(T) = \left\{ f : T \longrightarrow \mathbb{R} : \int_T |f(t)|^2 \, dt < \infty \right\}.$$

Este espacio, con el producto escalar usual, definido como

$$\langle f, g \rangle_u = \int_T f(t) \, g(t) \, dt, \ \forall f, g \in L^2(T), \tag{2.1}$$

y la norma asociada a dicho producto escalar, tiene estructura de espacio de Hilbert separable.

Una restricción que se suele imponer a las variables aleatorias en el ámbito de los procesos estocásticos, es que tengan al menos momentos de segundo orden finitos. Llamamos $L^2(\Omega)$ al conjunto de las variables aleatorias sobre Ω (en $\mathbb{R}$ o en H) con momentos de segundo orden finitos, esto es, tales que

$$E\left[\|X\|^2\right] = \int_\Omega \|X(\omega)\|^2 \, dP(\omega) < \infty, \ \forall X \in L^2(\Omega).$$

con $\|\cdot\|$ la norma correspondiente al espacio de Hilbert en el que consideremos la v. a. Este conjunto con las operaciones habituales de suma y producto por escalares de variables aleatorias, tiene estructura de espacio vectorial. Considerando sobre $L^2(\Omega)$ el producto escalar

$$\begin{array}{rccl} \langle , \rangle_{L^2(\Omega)}: & L^2(\Omega) \times L^2(\Omega) & \longrightarrow & \mathbb{R} \\ & (X, Y) & \longrightarrow & \langle X, Y \rangle = E[XY] = \int_\Omega X(\omega) \, Y(\omega) \, dP(\omega) \end{array}$$

se le dota a dicho espacio de estructura de espacio de Hilbert separable con la norma engendrada por dicho producto escalar.

Una vez introducidos los elementos probabilísticos necesarios, ya estamos en condiciones de definir lo que es un proceso estocástico.

Definición 6 *Dado $T \subset \mathbb{R}$ un intervalo de la recta real, se define un proceso estocástico continuo como una familia (no numerable) de variables aleatorias $\{X(t) : t \in T\}$ todas definidas sobre el mismo espacio probabilístico $(\Omega, \mathcal{A}, P)$.*

Según sean las variables aleatorias (reales, complejas o hilbertianas), hablaremos de procesos estocásticos reales, complejos o en H, siendo los primeros los que ocuparán nuestra atención en lo que sigue. Por otro lado, si las variables del proceso son de segundo orden, diremos que el proceso es de segundo orden.

Al igual que los principales elementos del análisis multivariante son los dos primeros momentos del vector de variables aleatorias de que es objeto, en el caso funcional también resultan fundamentales dichos momentos debido principalmente al gran número de problemas existentes en física e ingeniería en los que se requiere el conocimiento de tales momentos; y es por todo esto por lo que en general se requiere que los procesos estocásticos sean de segundo orden.

Definición 7 *Sea $\{X(t) : t \in T\}$ un proceso estocástico real de segundo orden. Entonces,*

- *se define su función media como*

$$\begin{array}{rccl} \mu: & T & \longrightarrow & \mathbb{R} \\ & t & \longrightarrow & \mu(t) = E[X(t)] = \int_\Omega X(t,\omega)\, dP(\omega). \end{array}$$

- *Se define su función de covarianza como*

$$\begin{array}{rccll} C: & T \times T & \longrightarrow & \mathbb{R} & \\ & (t,s) & \longrightarrow & C(t,s) & = E[(X(t) - \mu(t))(X(s) - \mu(s))] \\ & & & & = \int_\Omega [(X(t,\omega) - \mu(t))(X(s,\omega) - \mu(s))]\, dP(\omega). \end{array}$$

La mayoría de las técnicas funcionales, y en particular la que es objeto de esta tesis, imponen ciertas restricciones referentes a la continuidad de la función de covarianza. Es por ello por lo que introducimos a continuación el concepto de continuidad en media cuadrática.

Definición 8 *Se dice que un proceso estocástico real $\{X(t) : t \in T\}$ es continuo en media cuadrática si verifica que*

$$\lim_{h \to 0} E\left[(X(t+h) - X(t))^2\right] = 0, \ \forall t \in T.$$

La propiedad de continuidad en media cuadrática de un proceso permite asegurar la continuidad de su función de covarainza como indican los siguientes resultados cuyas demostraciones pueden verse en Todorovic (1992):

Proposición 1 *Un proceso estocástico de segundo orden* $\{X(t) : t \in T\}$ *es continuo en media cuadrática si y sólo si su función de covarianza es continua en todo punto diagonal* (t,t), $\forall t \in T$.

Proposición 2 *Si* $C(t,t)$ *es continua para todo* $t \in T$, *entonces es continua en* $T \times T$.

De la definición de proceso estocástico se tiene que éste se puede ver desde una doble perspectiva: dado un $t_0 \in T$ fijo, $X(t_0)$ es una variable aleatoria con valores en $\mathbb{R}$; y dado $\omega \in \Omega$ fijo, se tiene que $X(\omega)$ es una función de T en $\mathbb{R}$ conocida como trayectoria o función muestral, con lo que un proceso estocástico continuo puede verse como una variable aleatoria sobre un espacio de funciones.

A pesar de las buenas propiedades que proporciona a su función de covarianza la continuidad en media cuadrática de un proceso, no ocurre igual con sus trayectorias muestrales, ya que esta propiedad no es suficiente para establecer la continuidad de las mismas, de hecho el proceso de Poisson homogéneo es continuo en media cuadrática y sus trayectorias son discontinuas; no obstante sí permite asegurar que las trayectorias son al menos de cuadrado integrable como indica el siguiente resultado (ver Todorovic, 1992).

Proposición 3 *Si un proceso es continuo en media cuadrática, existe otro, estocásticamente equivalente a él, cuyas trayectorias son funciones de cuadrado integrable.*

Debido a estas buenas propiedades, a partir de ahora dado un fenómeno de naturaleza continua modelizado por el proceso estocástico real $\{X(t) : t \in T\}$, el ambiente que consideraremos será el dado por:

H_1 $\{X(t) : t \in T\}$ es de segundo orden,

H_2 continuo en media cuadrática,

H_3 y con trayectorias que pertenecen al espacio $L^2(T)$ de las funciones de cuadrado integrable con la métrica usual en $L^2(T)$ definida por (2.1).

Bajo estas hipótesis un proceso estocástico continuo puede verse como una función aleatoria definida sobre $L^2(T)$:

$$\begin{array}{rccl} X: & \Omega & \to & L^2(T) \\ & \omega & \to & X(\omega): \ T \to \mathbb{R} \\ & & & \quad\ t \to X(t,\omega). \end{array}$$

Definición 9 *(Operador de covarianza). Asociado a un proceso estocástico bajo las hipótesis H_1, H_2 y H_3, se define su operador de covarianza en la forma*

$$\begin{array}{rccl} \mathcal{C}: & L^2(T) & \longrightarrow & L^2(T) \\ & f & \longrightarrow & \mathcal{C}(f)(t) = \int_T C(t,s) f(s)\, ds. \end{array}$$

De la perspectiva de proceso estocástico como variable aleatoria en $L^2(T)$ podemos obtener una generalización al caso en que sus trayectorias pertenezcan a un espacio de Hilbert cualquiera H con un producto escalar asociado $\langle , \rangle_H$, siendo entonces un proceso estocástico una variable aleatoria Hilbertiana. En este caso el operador de covarianza se define como sigue:

Definición 10 *Dado $(H, \langle , \rangle_H)$ un espacio de Hilbert, y X una variable aleatoria sobre dicho espacio, se define su operador de covarianza $\mathcal{C}_H : H \to H$ como aquél que verifica que $\forall h_1, h_2 \in H$*

$$\langle \mathcal{C}_H(h_1), h_2 \rangle_H = \int_\Omega \langle X(\omega) - E_H[X], h_1 \rangle_H \langle X(\omega) - E_H[X], h_2 \rangle_H \, dP(\omega) \tag{2.2}$$

siendo $E_H[X]$ el elemento de H que verifica

$$\langle E_H[X], h \rangle_H = \int_\Omega \langle X(\omega), h \rangle_H \, dP(\omega), \ \forall h \in H \tag{2.3}$$

con Ω el espacio muestral sobre el que se define la variable aleatoria.

Igual que en el caso multivariante para el estudio de un fenómeno se toman muestras aleatorias de los vectores que modelizan la situación, dado un proceso estocástico real $\{X(t), t \in T\}$ bajo las condiciones vistas anteriormente en el caso funcional, dispondremos de una muestra aleatoria simple de n realizaciones del proceso (elementos de $L^2(T)$) denotadas por $x_1(t), x_2(t), \ldots, x_n(t)$. A partir de esta muestra podemos definir los dos primeros momentos muestrales.

Definición 11 *Se define la función media muestral como*

$$\overline{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t),$$

y la covarianza muestral como

$$\widehat{C}(s,t) = \frac{1}{n-1}\sum_{i=1}^{n} (x_i(s) - \overline{x}(s))(x_i(t) - \overline{x}(t)).$$

Estas funciones son estimadores insesgados, consistentes y que convergen casi seguramente a sus correspondientes momentos poblacionales (Deville, 1973). La función de covarianza muestral verifica además las siguientes propiedades:

1. Es definida no negativa, esto es, dados $a_1, \ldots, a_m \in \mathbb{R}$ y $t_1, \ldots, t_m \in T$, se tiene que

$$\sum_{k_1=1}^{m}\sum_{k_2=1}^{m} a_{k_1} a_{k_2} \widehat{C}(t_{k_1}, t_{k_2}) \geq 0.$$

2. Tiene simetría hermítica

$$\widehat{C}(s,t) = \widehat{C}(t,s), \ \forall (s,t) \in T \times T.$$

3. Verifica la desigualdad de Schwarz, esto es

$$\left|\widehat{C}(s,t)\right| \leq \left[\widehat{C}(s,s)\,\widehat{C}(t,t)\right]^{1/2}.$$

4. Su norma, definida como

$$\left\|\widehat{C}(s,t)\right\| = \left(\int_T \int_T \left|\widehat{C}(s,t)\right|^2 dsdt\right)^{1/2},$$

es finita por ser el proceso continuo en media cuadrática.

Definición 12 *Un elemento muy importante en el análisis de datos funcionales es el operador de covarianza muestral que se define como sigue:*

$$\widehat{\mathcal{C}}(f)(t) = \int_T \widehat{C}(s,t) f(s)\, ds$$

Este operador verifica las siguientes propiedades:

1. Es un operador acotado sobre $L^2(T)$, esto es,

$$\left\|\widehat{\mathcal{C}}(f)\right\| < k\|f\|,$$

siendo $\|\cdot\|$ la norma engendrada por el producto escalar usual en $L^2(T)$ definido en la ecuación (2.1) y k una constante.

2. Es un operador autoadjunto:

$$\left\langle \widehat{\mathcal{C}}(f), g \right\rangle = \left\langle f, \widehat{\mathcal{C}}(g) \right\rangle, \ \forall f, g \in L^2(T).$$

3. Es un operador positivo:

$$\left\|\widehat{\mathcal{C}}(f)\right\| \geq 0, \ \forall f \in L^2(T).$$

2.2.1 Aproximación de trayectorias en espacios de dimensión finita

Uno de los grandes problemas con que nos encontramos en la práctica, al trabajar con datos funcionales, es el hecho de que usualmente no se dispone de la expresión explícita de las trayectorias muestrales sino de una serie de observaciones de cada una de ellas en un conjunto finito de instantes o nodos. En la literatura podemos encontrar distintas estrategias para reconstruir de forma aproximada las trayectorias muestrales de un proceso estocástico a partir de observaciones discretas. En los libros de Ramsay y Silverman (1997) y Valderrama et. al. (2000) se analizan algunas de estas estrategias que se reducen a dos métodos diferentes dependiendo de que los datos observados se consideren medidos con o sin error, utilizando en el primer caso una aproximación mínimo cuadrática y un método de interpolación en el segundo.

La primera estrategia consiste en considerar que las trayectorias a reconstruir pertenecen a un espacio de dimensión finita generado por una base de funciones y obtener los coeficientes de la expresión de tales trayectorias en términos de los elementos básicos, mediante el método de mínimos cuadrados, a partir de la información conocida en los nodos. De este tipo es el método de proyección ortogonal propuesto en Aguilera et al. (1995).

La segunda consiste en obtener una función que coincida con el valor de cada trayectoria en los nodos de observación, utilizando alguno de los métodos de interpolación conocidos: Lagrange, Newton, Spline, etc. Aguilera et al. (1996) utilizan interpolación spline cúbica natural para la reconstrucción de las

funciones muestrales de un proceso aplicándola para estimar las componentes principales y predecir y modelizar la cotización en bolsa del sector Bancos (Aguilera et al., 1999) o el grado de ocupación hotelera en Granada (Aguilera et al., 1997). Por otro lado Ramsay y Silverman (1997) analizan ambas estrategias en diversos estudios referidos a la evolución de temperaturas en el tiempo o de la estatura de un conjunto de niños, así como la evolución del ángulo de la cadera y de la rodilla al caminar. Todas estas estrategias terminan por discretizar el problema funcional y tratarlo desde la perspectiva multivariante.

Al analizar una variable que evoluciona en el tiempo y se modeliza según un determinado proceso estocástico $\{X(t) : t \in T\}$, es usual tomar una muestra aleatoria simple de trayectorias de dicho proceso, $x_1(t), \ldots, x_n(t)$, y más concretamente un conjunto de observaciones de dichas trayectorias en un conjunto de instantes $t_{i0}, \ldots, t_{im_i}$, $i = 1, \ldots, n$, no necesariamente iguales para cada trayectoria. Denotaremos por x_i al vector que tiene las observaciones de la trayectoria $x_i(t)$ en los nodos $t_{i0}, \ldots, t_{im_i}$,

$$x_i = (x_i(t_{i0}), \ldots, x_i(t_{im_i}))^T,$$

Dada una trayectoria muestral $x(t)$ asociada a un proceso estocástico $\{X(t),\ t \in T\}$, todas las estrategias desarrolladas hasta el momento para reconstruir su forma funcional a partir de datos discretos pasan por considerar que dicha trayectoria pertenece a un espacio de dimensión finita E_p generado por una base de funciones, $\{\phi_1(t), \ldots, \phi_p(t)\}$ de modo que su expresión es de la forma

$$x(t) = \sum_{j=1}^{p} a_j \phi_j(t),$$

siendo diferente en cada caso (aproximación mínimo cuadrática e interpolación) la obtención de los coeficientes a_j.

Un aspecto a tener en cuenta, para la aproximación de las trayectorias, es la elección del espacio más apropiado generado por la base $\{\phi_1(t), \ldots, \phi_p(t)\}$ que dependerá de la naturaleza de las trayectorias. Algunas de las bases más utilizadas son las siguientes:

Funciones indicadoras. Son muy adecuadas para procesos puntuales o de recuento cuyas trayectorias permanecen constantes en intervalos aleatorios. Dada una partición de $T = [T_1, T_2]$, definida por los nodos $T_1 = a_0 < a_1 < \ldots < a_p = T_2$, las funciones básicas serían

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

donde $I_j(t)$ es la función indicadora del intervalo $(a_{j-1}, a_j]$. Observemos que el espacio E_p generado por esta base ortonormal de funciones es el de las funciones constantes en los intervalos de la partición elegida en el intervalo de observación.

Funciones trigonométricas. Son muy utilizadas si se conoce que las trayectorias son regulares (continuas y diferenciables casi seguramente), y sus elementos están definidos de la forma

$$\begin{aligned}
\phi_1(t) &= \frac{1}{(T_2 - T_1)^{1/2}} \\
\phi_j(t) &= \left(\frac{2}{T_2 - T_1}\right)^{1/2} \operatorname{sen}\left(\frac{2\pi j t}{T_2 - T_1}\right), \text{ si } j \text{ es par} \\
\phi_j(t) &= \left(\frac{2}{T_2 - T_1}\right)^{1/2} \cos\left(\frac{2\pi j t}{T_2 - T_1}\right), \text{ si } j \text{ es impar}
\end{aligned}$$

Estas funciones son muy útiles para aproximar trayectorias muy estables en las que no hay un comportamiento local fuerte y cuya curvatura es la misma aproximadamente en todo el intervalo considerado (Ramsay y Silverman, 1997). Por otro lado esta base es inadecuada si se sospecha que existe algún grado de discontinuidad en las trayectorias a aproximar. Observemos que estas funciones constituyen una base ortonormal de $L^2(T)$ de modo que el espacio E_p generado por las p primeras funciones trigonométricas es una aproximación de $L^2(T)$ en el sentido de que la sucesión de proyecciones ortogonales de $L^2(T)$ sobre E_p converge puntualmente a la identidad cuando $p \longrightarrow \infty$ (Riesz y Nagy, 1990).

Funciones polinómicas. Son funciones poco utilizadas porque a pesar de su facilidad de cálculo, presentan grandes oscilaciones y no demasiada suavidad. Los elementos de las bases polinómicas son de la forma

$$\phi_j(t) = (t - \theta)^j, \; j = 0, 1, \ldots$$

con θ un parámetro a elegir. Al igual que las trigonométricas, estas funciones no muestran un comportamiento local adecuado a menos que el grado sea elevado, además los polinomios tienden a ajustarse bien en el centro de los datos y bastante mal en las colas.

B-splines. Son funciones que generan los espacios de las funciones spline, esto es, funciones polinómicas a trozos que se unen de manera suave. Dada una partición del intervalo donde queremos aproximar las trayectorias $\tau_0 < \ldots < \tau_q$, un spline de grado r no es más que una función que en cada intervalo

$[\tau_k, \tau_{k+1}]$ es un polinomio de grado r y que tiene derivada continua en los extremos de dichos intervalos.

Las funciones B-spline de grado r son funciones generadoras del espacio de los splines del mismo grado, y existe un método recursivo que permite obtener las de grado r a partir de las de grado $r-1$, después de ampliar la partición de nodos original en la forma $\tau_{-3} < \tau_{-2} < \tau_{-1} < \tau_0 < \ldots < \tau_q < \tau_{q+1} < \tau_{q+2} < \tau_{q+3}$, dado por

$$\begin{aligned} B_{j,1}(t) &= \begin{cases} 1 & \tau_{j-2} \leq t < \tau_{j-1} \\ 0 & \text{otro caso} \end{cases}, \; j = -1, 0, 1, \ldots, q+4 \\ B_{j,r}(t) &= \frac{t - \tau_{j-2}}{\tau_{j+r-3} - \tau_{j-2}} B_{j,r-1}(t) + \frac{\tau_{j+r-2} - t}{\tau_{j+r-2} - \tau_{j-1}} B_{j+1,r-1}(t) \qquad (2.4) \\ r &= 2, 3, \ldots; \; j = -1, 0, \ldots, q-r+5 \end{aligned}$$

Las funciones spline tienen un mejor comportamiento local que las trigonométricas y polinómicas de ahí su popularidad, siendo las más utilizadas generalmente las de grado 3 que son adecuadas para el caso de trayectorias regulares. Además de la base de B-splines existen otras que generan el espacio de los splines, como por ejemplo la de potencias truncadas (De Boor, 1978) aunque es común utilizar B-splines debido a que tienen soporte compacto, esto es, para los B-splines de grado 3

$$B_{j,4}(t) = 0 \;\; \forall t \notin [\tau_{j-2}, \tau_{j+2}], \; j = -1, 0, \ldots, q+1$$

Una alternativa a los splines que explican muy acertadamente el comportamiento local de las trayectorias son las bases de funciones wavelets utilizadas por Ocaña et al. (1998).

Aproximación mínimo cuadrática en términos de bases de funciones

Para ilustrar el caso en el que los datos se suponen medidos con error, consideremos una trayectoria $x_i(t)$ de un proceso estocástico $\{X(t) : t \in T\}$ y supongamos que conocemos los valores de dicha trayectoria en un conjunto de nodos $t_{i0}, t_{i1}, \ldots, t_{im_i} \in T$. Entonces la información disponible para la aproximación vendrá dada por el vector

$$x_i = (x_i(t_{i0}), x_i(t_{i1}), \ldots, x_i(t_{im_i}))^T$$

Consideremos por otro lado $\phi_1(t), \ldots, \phi_p(t)$ el conjunto de funciones básicas generadoras del espacio donde queremos aproximar dicha trayectoria, entonces

los valores de tales funciones básicas en los nodos de observación se pueden expresar matricialmente en la forma

$$\Phi_i = \begin{pmatrix} \phi_1(t_{i0}) & \phi_2(t_{i0}) & \cdots & \phi_p(t_{i0}) \\ \phi_1(t_{i1}) & \phi_2(t_{i1}) & \cdots & \phi_p(t_{i1}) \\ \cdots & \cdots & \cdots & \cdots \\ \phi_1(t_{im_i}) & \phi_2(t_{im_i}) & \cdots & \phi_p(t_{im_i}) \end{pmatrix}_{(m_i+1)\times p}$$

Al suponer que los datos son observados con error y que la trayectoria está en el espacio generado por la base, podemos suponer el siguiente modelo para las observaciones:

$$x_i(t_{ik}) = \sum_{j=1}^{p} a_{ij}\phi_j(t_{ik}) + \epsilon_{ik},\ k = 0, \ldots, m_i. \tag{2.5}$$

En forma matricial $x_i = \Phi_i a_i + \varepsilon_i$.

Si ajustamos mediante mínimos cuadrados este modelo suponiendo las restricciones habituales sobre los errores ε_{ik} de ser centrados, independientes y de varianza constante, se tiene que el vector de coeficientes estimado $\widehat{a}_i = (\widehat{a}_{i1}, \ldots, \widehat{a}_{ip})^T$ vendrá dado por

$$\widehat{a}_i = \left(\Phi_i^T \Phi_i\right)^{-1} \Phi_i^T x_i$$

y por lo tanto la trayectoria aproximada será

$$\widehat{x}_i(t) = \sum_{j=1}^{p} \widehat{a}_{ij}\phi_j(t).$$

Un caso particular de esta situación ocurre cuando se tienen los mismos nodos de observación $t_0, \ldots, t_m$ para todas las trayectorias observadas, entonces podemos resumir la información disponible en la matriz $\mathcal{X}$ que tiene por filas a los vectores x_i^T,

$$\mathcal{X} = \begin{pmatrix} x_1(t_0) & x_1(t_1) & \cdots & x_1(t_m) \\ x_2(t_0) & x_2(t_1) & \cdots & x_2(t_m) \\ \cdots & \cdots & \cdots & \cdots \\ x_n(t_0) & x_n(t_1) & \cdots & x_n(t_m) \end{pmatrix}_{n\times(m+1)},$$

y la matriz A que tiene por filas los vectores a_i^T de coeficientes de cada trayectoria respecto de las funciones base.

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{np} \end{pmatrix}_{n\times p}$$

de manera que la expresión (2.5) para todas las trayectorias en los mismos nodos de observación quedaría matricialmente

$$\mathcal{X}^T = \Phi A^T + \varepsilon$$

siendo $\Phi_i = \Phi$, $\forall i = 1, \ldots, n$ y ε la matriz de elementos ε_{ij}

$$\varepsilon = \begin{pmatrix} \varepsilon_{01} & \varepsilon_{02} & \cdots & \varepsilon_{0n} \\ \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1n} \\ \cdots & \cdots & \cdots & \cdots \\ \varepsilon_{m1} & \varepsilon_{m2} & \cdots & \varepsilon_{mn} \end{pmatrix}.$$

Ajustando por mínimos cuadrados este modelo, se tiene que los parámetros estimados para todas las trayectorias serían

$$\widehat{A}^T = \left(\Phi^T \Phi\right)^{-1} \Phi^T \mathcal{X}^T. \tag{2.6}$$

En las aplicaciones que se presentarán en el Capítulo 3 de esta tesis se llevará a cabo una aproximación spline cúbica de las trayectorias. Para ello se fijará una partición $\tau_0 < \ldots < \tau_q$ en el intervalo de observación que permita definir una base de B-splines cúbicos de dimensión $(q+3)$ suficientemente menor que el número de valores observados de cada trayectoria.

Aproximación mediante interpolación spline cúbica

Cuando las observaciones que se hacen de las trayectorias se consideran obtenidas sin error, la estrategia para aproximar cada trayectoria consiste en encontrar una función que coincida con ella en los nodos de observación, en otras palabras, interpolar las trayectorias en dichos valores conocidos. Existen muchos métodos de interpolación siendo los más populares, por su simplicidad, los métodos de interpolación polinómica entre los que podemos destacar las fórmulas de Lagrange y Newton o la utilización de polinomios ortogonales.

Sin embargo los polinomios no proporcionan un comportamiento local adecuado y es por esto por lo que surge la interpolación spline. Anteriormente ya hemos definido el concepto de función spline sobre una partición fijada, así como la base de B-splines que las genera. A continuación veamos cómo se obtienen los coeficientes del spline cúbico de interpolación de cada trayectoria sobre los nodos de observación.

Supongamos una función muestral $x_i(t)$ de un proceso estocástico $\{X(t) : t \in T\}$ y, al igual que en el caso anterior, consideremos que conocemos los valores que

toma dicha trayectoria en un conjunto de nodos de observación $t_0 < \ldots < t_m$, de modo que tales valores han sido observados sin error. Entonces la función spline cúbica que interpola a dicha trayectoria en los nodos de observación en términos de la base de B-splines $\{B_{-1,4}(t), B_{0,4}(t), \ldots, B_{m+1,4}(t)\}$ es de la forma

$$\widetilde{x}_i(t) = \sum_{j=-1}^{m+1} \widetilde{a}_{ij} B_{j,4}(t)$$

considerando la correspondiente ampliación de los nodos de observación vista en la sección anterior.

Imponiendo que esta función interpola a las trayectorias en los nodos de observación se obtiene que los coeficientes son las soluciones del siguiente sistema tridiagonal con $m+1$ ecuaciones y $m+3$ incógnitas:

$$x_i(t_k) = \widetilde{x}_i(t_k) = \sum_{j=-1}^{m+1} \widetilde{a}_{ij} B_{j,4}(t_k), \; k = 0, \ldots, m$$

que matricialmente es de la forma

$$\begin{pmatrix} x_i(t_0) \\ x_i(t_1) \\ \cdots \\ x_i(t_m) \end{pmatrix} = \begin{pmatrix} B_{-1}(t_0) & B_0(t_0) & \cdots & B_{m+1}(t_0) \\ B_{-1}(t_1) & B_0(t_1) & \cdots & B_{m+1}(t_1) \\ \cdots & \cdots & \cdots & \cdots \\ B_{-1}(t_m) & B_0(t_m) & \cdots & B_{m+1}(t_m) \end{pmatrix}_{(m+1)\times(m+3)} \begin{pmatrix} \widetilde{a}_{-1} \\ \widetilde{a}_0 \\ \cdots \\ \widetilde{a}_{m+1} \end{pmatrix}$$

Para que este sistema tenga solución única hay que imponer dos condiciones adicionales, resultando distintos tipos de spline cúbicos de interpolación dependiendo de las condiciones fijadas. Una de las más usuales consiste en imponer que la segunda derivada en los extremos del intervalo de interpolación sea nula, considerando así el llamado spline cúbico natural de interpolación y dando lugar al siguiente sistema:

$$x_i = B\widetilde{a}$$

donde $x_i = (0, x_i(t_0), \ldots, x_i(t_m), 0)^T$, $\widetilde{a} = (\widetilde{a}_{-1}, \widetilde{a}_0, \ldots, \widetilde{a}_{m+1})^T$ y

$$B = \begin{pmatrix} B_{-1}^{2)}(t_0) & B_0^{2)}(t_0) & \cdots & B_{m+1}^{2)}(t_0) \\ B_{-1}(t_0) & B_0(t_0) & \cdots & B_{m+1}(t_0) \\ \cdots & \cdots & \cdots & \cdots \\ B_{-1}(t_m) & B_0(t_m) & \cdots & B_{m+1}(t_m) \\ B_{-1}^{2)}(t_m) & B_0^{2)}(t_m) & \cdots & B_{m+1}^{2)}(t_m) \end{pmatrix}_{(m+3)\times(m+3)}$$

donde $B_j^{2)}(t)$ representa la derivada segunda del j-ésimo B-spline. Resolviendo este sistema, que se demuestra que tiene solución única (De Boor, 1978), se tienen los coeficientes de la interpolación en términos de la base de B-splines

$$\widetilde{a} = B^{-1}x_i,\ i = 1, \ldots, n.$$

Considerando ahora una muestra de trayectorias $x_1(t), \ldots, x_n(t)$ y que disponemos de las observaciones de dichas trayectorias en el mismo conjunto de nodos para todas $t_0, \ldots, t_m$, el sistema anterior se convierte en la ecuación matricial siguiente

$$X_I = \widetilde{A}B^T$$

siendo $\widetilde{A}$ a la matriz que tiene por filas los coeficientes de la interpolación de cada trayectoria

$$\widetilde{A} = \begin{pmatrix} a_{-1,1} & \cdots & a_{(m+1),1} \\ \cdots & \cdots & \cdots \\ a_{-1,n} & \cdots & a_{(m+1),n} \end{pmatrix}_{n\times(m+3)},$$

X_I la matriz de interpolación, esto es, la que tiene por filas los valores de cada trayectoria en los nodos y por primera y última columnas los ceros correspondientes a la condición de B-spline cúbico natural

$$X_I = \begin{pmatrix} 0 & x_1(t_0) & \cdots & x_1(t_m) & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & x_n(t_0) & \cdots & x_n(t_m) & 0 \end{pmatrix}_{n\times(m+3)}$$

y B la matriz de los valores de los B-splines en los nodos vista anteriormente. Resolviendo el sistema, tendremos los coeficientes de la interpolación de las trayectorias en términos de la base de B-Splines:

$$\widetilde{A} = X_I\left(B^T\right)^{-1}. \tag{2.7}$$

2.3 Regresión logística funcional

En el primer capítulo del presente trabajo abordamos el modelo de regresión logística múltiple como respuesta al problema de predecir la probabilidad de ocurrencia de un suceso en función de los valores que tomaran un conjunto de

variables relacionadas con dicho suceso. A continuación pretendemos generalizar esta situación al caso en que la información disponible para predecir dicha probabilidad tenga un carácter funcional, esto es, predecir la probabilidad de que ocurra un suceso en función de la evolución en el tiempo de una magnitud relacionada con dicho suceso.

2.3.1 Formulación del modelo

Sea Y una variable aleatoria que representa la variable respuesta binaria objeto de nuestro estudio, y sea $\{X(t) : t \in T\}$ un proceso estocástico que modeliza la magnitud explicativa funcional relacionada con Y. Consideremos $x(t)$ una trayectoria cualquiera de dicho proceso, entonces podemos considerar que

$$Y/[X(t) = x(t)] \rightsquigarrow B(\pi(x(t)))$$

donde $B(\pi(x(t)))$ es una distribución de Bernouilli de parámetro

$$\pi(x(t)) = P\{Y = 1/X(t) = x(t)\} = E[Y/\{X(t) = x(t) : t \in T\}]$$

Consideremos $x_1(t), \ldots, x_n(t)$ una muestra de funciones o realizaciones muestrales del proceso estocástico anterior que supondremos, sin pérdida de generalidad, centradas; y sea $y_1, \ldots, y_n$ una muestra aleatoria simple de observaciones de la variable respuesta dicotómica asociadas a las n funciones muestrales. Tal y como ocurre en el caso múltivariante, el modelo lineal no es adecuado para representar la relación entre las observaciones de la variable dependiente binaria y las trayectorias muestrales explicativas. De forma análoga a la generalización del modelo lineal múltiple al funcional (Ramsay y Silverman, 1997), proponemos la siguiente formulación del modelo logístico funcional:

$$Y = \pi(X(t)) + \varepsilon \tag{2.8}$$

donde

$$\begin{aligned} Y &= (y_1, \ldots, y_n)^T \\ \pi(X(t)) &= (\pi_1, \ldots, \pi_n)^T \\ \varepsilon &= (\varepsilon_1, \ldots, \varepsilon_n)^T \end{aligned}$$

con

$$\pi_i(X(t)) = \frac{\exp\left\{\alpha + \int_T x_i(t)\,\beta(t)\,dt\right\}}{1 + \exp\left\{\alpha + \int_T x_i(t)\,\beta(t)\,dt\right\}},\ i = 1, \ldots, n, \tag{2.9}$$

ε_i, $(i = 1, \ldots, n)$ son los errores que se considerarán centrados e independientes, y con varianza $\pi_i (1 - \pi_i)$, y $\beta(t)$ una función parámetro a determinar. De este modo

$$\pi_i = P\{Y = 1/X(t) = x_i(t)\} = E[Y/X(t) = x_i(t)] = E[y_i], \; i = 1, \ldots, n.$$

En términos de las transformaciones logit, el modelo (2.9) se puede expresar de la forma

$$l_i = \ln\left[\frac{\pi_i}{1 - \pi_i}\right] = \alpha + \int_T x_i(t)\,\beta(t)\,dt, \; i = 1, \ldots, n, \tag{2.10}$$

lo que permite que el modelo logístico se vea como un modelo lineal generalizado funcional con la transformación logit como función link.

Observemos que en el caso de considerar un producto escalar $\langle , \rangle$ distinto del usual en el espacio de las trayectorias, las combinaciones lineales generalizadas de las variables del proceso explicativo serían de la forma $\langle x_i, \beta \rangle$ y la expresión (2.10) para las transformaciones logit se convertirían en

$$l_i = \alpha + \langle x_i, \beta \rangle, \; i = 1, \ldots, n.$$

2.3.2 Estimación aproximada de la función parámetro en un espacio de dimensión finita

Como afirman Ramsay y Silverman (1997) para la estimación de la función parámetro $\beta(t)$ en el caso lineal funcional, la estimación de $\beta(t)$ con los métodos usuales de máxima verosimilitud o mínimos cuadrados ponderados es imposible, ya que $\beta(t)$ contiene un conjunto no numerable de "valores", y disponemos a lo sumo de un número finito de condiciones. Se han apuntado diversas aproximaciones para la solución de este problema en el caso lineal, que pasan en la mayoría de los casos por discretizar el modelo y convertirlo en un modelo de regresión lineal múltiple.

En el caso del modelo logit de respuesta multinomial Cardot et al. (2001) proponen utilizar un método de cuadratura para obtener la integral del modelo (en términos de la transformación logit) en la forma

$$\int_T x_i(t)\,\beta(t)\,dt \approx \sum_{k=1}^{m} \omega_k \beta(t_k)\, x_i(t_k)$$

para un conjunto de nodos $t_1, \ldots, t_m$.

En la sección previa hemos visto cómo podíamos aproximar las trayectorias de un proceso a partir de sus valores observados en un conjunto discreto de nodos de observación. Nos proponemos a continuación dar una respuesta a la estimación de la función parámetro del modelo logístico funcional a partir de aproximaciones como las vistas entonces.

En el epígrafe anterior hemos expresado las trayectorias muestrales en términos de una base $\phi_1, \ldots, \phi_p$ que generaba un espacio de dimensión finita, en la forma

$$x_i(t) = \sum_{j=1}^{p} a_{ij}\phi_j(t), \; i = 1, \ldots, n. \tag{2.11}$$

Supongamos ahora que la función parámetro también se puede expresar de la misma forma en términos de la misma base

$$\beta(t) = \sum_{k=1}^{p} \beta_k \phi_k(t), \tag{2.12}$$

de modo que tendremos determinada la función parámetro cuando conozcamos los coeficientes de esta expresión, $\beta_1, \ldots, \beta_p$.

Consideremos la expresión del modelo logístico funcional en términos de la transformación logit, y sustituyamos tanto las trayectorias como la función parámetro por sus aproximaciones en términos de las funciones básicas

$$\begin{aligned}
\ln\left[\frac{\pi_i}{1-\pi_i}\right] &= \alpha + \int_T x_i(t)\,\beta(t)\,dt \\
&= \alpha + \int_T \left(\sum_{j=1}^{p} a_{ij}\phi_j(t)\right)\left(\sum_{k=1}^{p} \beta_k \phi_k(t)\right) dt \\
&= \alpha + \sum_{j=1}^{p}\sum_{k=1}^{p} a_{ij}\beta_k \int_T \phi_j(t)\,\phi_k(t)\,dt \\
&= \alpha + \sum_{j=1}^{p}\sum_{k=1}^{p} a_{ij}\psi_{jk}\beta_k, \; i = 1, \ldots, n
\end{aligned} \tag{2.13}$$

donde ψ_{jk} son los productos escalares entre las funciones básicas dados por

$$\psi_{jk} = \int_T \phi_j(t)\,\phi_k(t)\,dt, \; j, k = 1, \ldots, p.$$

Entonces el modelo logístico funcional se convierte en un modelo de regresión logística múltiple con matriz de diseño dada por

$$X = (1 \mid A\Psi)$$

y con parámetros $\beta = \left(\alpha, \beta_1, \ldots, \beta_p\right)^T$; siendo Ψ la matriz que contiene los productos escalares de las funciones básicas y A la matriz que tiene por filas los coeficientes de las expresiones de las trayectorias en términos de los elementos básicos. Esta última matriz se aproximará a partir de los datos discretos de las trayectorias por alguno de los métodos vistos anteriormente, siendo su valor aproximado

$$\widehat{A} = \mathcal{X}\Phi\left(\Phi^T\Phi\right)^{-1}$$

en el caso de aproximación mínimo cuadrática, y

$$\widetilde{A} = X_I\left(B^T\right)^{-1}$$

en el de interpolación spline cúbica natural en términos de la base de B-splines.

La matriz de productos escalares Ψ se podrá obtener de manera más o menos fácil dependiendo de cómo sea la base de las expresiones (2.11) y (2.12). Si ésta tiene elementos cuyos productos son integrables de forma sencilla, se obtendrá calculando tales integrales, en caso contrario habría que utilizar algún método de cuadratura apropiado. En concreto, si utilizamos una base de B-splines cúbicos como los vistos anteriormente, podemos emplear la fórmula de cuadratura de Gauss con cuatro nodos propuesta en Ocaña (1996).

En el caso en que la base a utilizar en la aproximación sea ortonormal, como es el caso de la base de funciones trigonométricas o polinomios ortogonales, la matriz Ψ es la identidad y el modelo logístico funcional se convierte en un modelo logístico múltiple con matriz de diseño dada por $\left(1 \mid \widehat{A}\right)$, de modo que la regresión logística funcional es equivalente a la regresión logística múltiple sobre los coeficientes de las trayectorias en términos de las funciones básicas.

Supongamos que hemos aproximado los valores de A y que hemos obtenido la matriz de los productos escalares de los elementos básicos, entonces el modelo logístico funcional aproximado sería, en términos de la transformación logit,

$$L = X\beta$$

donde L es el vector de las transformaciones logit

$$L = \left(\ln\left[\frac{\pi_1}{1-\pi_1}\right], \ldots, \ln\left[\frac{\pi_n}{1-\pi_n}\right]\right)^T.$$

Ajustando por máxima verosimilitud este modelo logístico, como vimos en el primer capítulo, se tendría una estimación de β dada por $\widehat{\beta} = \left(\widehat{\alpha}, \widehat{\beta}_1, \ldots, \widehat{\beta}_p\right)^T$

y de ahí, una estimación aproximada de la función parámetro dada por

$$\widehat{\beta}(t) = \sum_{j=1}^{p} \widehat{\beta}_j \phi_j(t).$$

Recordemos que en el caso multivariante, la alta dependencia en las variables explicativas hacía que los parámetros del modelo logístico no se estimaran de manera precisa. Para resolver esta situación así como para reducir la dimensión del problema proponíamos la utilización de las componentes principales de dichas variables explicativas. Este problema se ve aún más acentuado por la clara dependencia existente en la matriz de diseño X que procede de la observación de una misma trayectoria en distintos nodos. Al igual que en el caso multivariante, proponemos solucionar esta problemática utilizando análisis en componentes principales, en este caso funcional, consiguiendo de este modo, además, reducir la dimensión del problema funcional utilizando un conjunto finito de componentes principales en la estimación de la función parámetro en lugar de los coeficientes de las trayectorias muestrales en términos de las funciones básicas. Observemos que en el caso de la interpolación spline cúbica el número de coeficientes de cada trayectoria supera en dos al número de valores observados, lo que pone de manifiesto la necesidad de reducir la dimensión.

2.3.3 Interpretación de parámetros

Al igual que en el caso múltiple, resultaría interesante dar una interpretación de los parámetros del modelo de regresión logística funcional. Recordemos que en el caso múltiple la exponencial de cada parámetro representaba el cociente de ventajas de respuesta $Y = 1$ frente a respuesta $Y = 0$ cuando aumentábamos en una unidad la correspondiente variable explicativa, permaneciendo constantes el resto.

Consideremos ahora el modelo logístico (2.10) en términos de la transformación logit para una trayectoria muestral $x(t)$

$$\ln\left[\frac{\pi(x(t))}{1-\pi(x(t))}\right] = \alpha + \int_T x(t)\,\beta(t)\,dt = \alpha + \langle x(t), \beta(t)\rangle_u$$

Consideremos que aumentamos una cantidad K de forma constante el valor de la trayectoria en el intervalo $(t_0, t_0 + h)$ de la forma

$$x(t) \to x(t) + I_{\Delta(t)=h}(t) \cdot K$$

siendo

$$I_{\Delta(t)=h}(t) = \begin{cases} 1 & t \in (t_0, t_0+h) \\ 0 & t \notin (t_0, t_0+h) \end{cases},$$

entonces el cociente de ventajas para el incremento descrito será

$$\begin{aligned} \theta^1 &= \theta\left[\Delta X(t) = K/\Delta t = h\right] = \frac{\dfrac{P\left\{Y=1/X(t)=x(t)+I_{\Delta(t)=h}(t)\cdot K\right\}}{1-P\left\{Y=1/X(t)=x(t)+I_{\Delta(t)=h}(t)\cdot K\right\}}}{\dfrac{P\left\{Y=1/X(t)=x(t)\right\}}{1-P\left\{Y=1/X(t)=x(t)\right\}}} \\ &= \frac{\exp\left\{\alpha+\int_T x(t)\beta(t)\,dt+\int_T\left(I_{\Delta(t)=h}(t)\cdot K\right)\beta(t)\,dt\right\}}{\exp\left\{\alpha+\int_T x(t)\beta(t)\,dt\right\}} \qquad (2.14) \\ &= \exp\left\{\int_{t_0}^{t_0+h} K\beta(t)\,dt\right\}, \end{aligned}$$

de manera que la exponencial de la integral en el intervalo (t_0, t_0+h) de la función parámetro multiplicada por K es el cociente de ventajas de respuesta $Y=1$ frente a $Y=0$ cuando la trayectoria aumenta de manera constante una cantidad K en dicho intervalo.

Si en lugar de un incremento constante consideramos un incremento lineal de la forma

$$\Delta X(t) = u + vt \quad \forall t \in (t_0, t_0+h)$$

tendríamos que

$$\theta\left[\Delta X(t) = u+vt/\Delta t = h\right] = \exp\left\{\int_{t_0}^{t_0+h}(u+vt)\beta(t)\,dt\right\}$$

De manera análoga podríamos considerar un incremento más general $\Delta X(t) = g(t)$ pero sería más difícil de precisar en la práctica en qué consiste un incremento de esa forma.

Asumiendo que la función parámetro pertenece al espacio engendrado por las funciones básicas, para un incremento constante en el intervalo (t_0, t_0+h) se tiene

$$\begin{aligned} \theta\left[\Delta X(t) = K/\Delta t = h\right] &= \exp\left\{\int_{t_0}^{t_0+h} K\beta(t)\,dt\right\} \\ &= \exp\left\{\int_{t_0}^{t_0+h} K\sum_{j=1}^{p}\beta_j\phi_j(t)\,dt\right\} \\ &= \exp\left\{K\sum_{j=1}^{p}\beta_j\int_{t_0}^{t_0+h}\phi_j(t)\,dt\right\}, \end{aligned}$$

y para un incremento lineal

$$\begin{aligned}\theta\left[\Delta X\left(t\right)=u+vt/\Delta t=h\right] &= \exp\left\{\int_{t_0}^{t_0+h}\left(u+vt\right)\beta\left(t\right)dt\right\} \\ &= \exp\left\{\int_{t_0}^{t_0+h}\left(u+vt\right)\sum_{j=1}^{p}\beta_j\phi_j\left(t\right)dt\right\}\end{aligned}$$

2.4 Análisis en Componentes Principales Funcional (ACPF)

Como se vio en el primer capítulo el análisis en componentes principales de un vector aleatorio es una técnica multivariante de reducción de dimensión que tiene por objetivo explicar la variabilidad presente en una muestra aleatoria simple de observaciones de dicho vector aleatorio mediante un conjunto reducido de variables incorreladas (componentes principales). Con este fin se definen las componentes principales como combinaciones lineales de las variables del vector con varianza máxima e incorreladas.

El ACP funcional generaliza el caso multivariante a situaciones en las que la información muestral disponible es un conjunto de trayectorias muestrales procedentes de un proceso estocástico en tiempo continuo. Así, las componentes principales funcionales no serán más que combinaciones lineales generalizadas de las variables del proceso con varianza máxima e incorreladas. Dichas componentes principales permitirán una representación ortogonal de las trayectorias del proceso en términos de variables aleatorias incorreladas y funciones determinísticas, conocido en el ámbito probabilísitco como desarrollo de Karhunen-Loève.

Consideraremos sin pérdida de generalidad que el proceso es centrado, ya que de no serlo se trabajaría con el proceso centrado resultante de restar la función media a cada variable de dicho proceso. Como consecuencia dada una muestra aleatoria simple de trayectorias $x_1\left(t\right),\ldots,x_n\left(t\right)$, consideraremos a partir de ahora que $\overline{x}\left(t\right)=0$.

2.4.1 Teoría básica

Como hemos indicado previamente, a continuación se definen las componentes principales mediante combinaciones lineales generalizadas de la variables de un proceso.

Definición 13 *Dada una muestra de trayectorias* $\{x_1(t), \ldots, x_n(t)\}$ *de un proceso estocástico real* $\{X(t), t \in T\}$ *se define una "combinación lineal generalizada" de dichas variables con función peso* $f(t) \in L^2(T)$ *como el vector* ξ_f *de valores*

$$\xi_{if} = \int_T x_i(t) f(t)\, dt,\ i = 1, \ldots, n. \tag{2.15}$$

En el ambiente en que nos moveremos en adelante, las combinaciones lineales generalizadas que vamos a tratar partirán de una muestra aleatoria simple de elementos de un proceso estocástico real (como el definido) verificando las hipótesis vistas $H_1 - H_3$ con lo que las componentes principales muestrales serán vectores de observaciones de las correspondientes variables aleatorias asociadas a sus componentes principales poblacionales.

Para la obtención de las componentes principales, será necesaria la obtención de las distintas funciones peso $f(t)$ que hagan que dichas componentes principales verifiquen determinadas condiciones de optimalidad. Veamos a continuación los momentos más importantes asociados a una combinación lineal así definida.

Dadas dos combinaciones lineales generalizadas como las definidas anteriormente con funciones peso $f(t)$ y $g(t) \in L^2(T)$

$$\begin{aligned} \xi_f &= (\xi_{1f}, \ldots, \xi_{nf})^T;\ \xi_{if} = \int_T x_i(t) f(t)\, dt,\ i = 1, \ldots, n \\ \xi_g &= (\xi_{1g}, \ldots, \xi_{ng})^T;\ \xi_{ig} = \int_T x_i(t) g(t)\, dt,\ i = 1, \ldots, n, \end{aligned}$$

entonces los momentos de primer y segundo orden se obtienen como:

- Media

$$\overline{\xi}_f = \frac{1}{n}\sum_{i=1}^{n} \xi_{if} = \int_T \overline{x}(t) f(t)\, dt = 0. \tag{2.16}$$

- Varianza

$$Var[\xi_f] = \frac{1}{n-1}\sum_{i=1}^{n} \xi_{if}^2 = \int_T f(t) \widehat{C}(f)(t)\, dt = \left\langle f, \widehat{C}(f) \right\rangle_u. \tag{2.17}$$

- Covarianza

$$\begin{aligned} Cov[\xi_f, \xi_g] &= \frac{1}{n-1}\sum_{i=1}^{n} \xi_{if}\xi_{ig} = \int_T f(t) \widehat{C}(g)(t)\, dt = \left\langle f, \widehat{C}(g) \right\rangle_u \\ &= \left\langle \widehat{C}(f), g \right\rangle_u. \end{aligned} \tag{2.18}$$

Como vimos en el primer capítulo y hemos comentado al inicio de esta sección, las componentes principales no son cualesquiera combinaciones lineales generalizadas, sino que deben tener varianza máxima, estar normalizadas y ser incorreladas.

Definición 14 *Dada una muestra aleatoria simple* $x_1(t), \ldots, x_n(t)$ *de elementos de un proceso estocástico real* $\{X(t), t \in T\}$ *verificando las hipótesis* $H_1 - H_3$,

- *se define la primera componente principal* $\xi_1 = (\xi_{11}, \ldots, \xi_{n1})^T$ *como la combinación lineal generalizada*

$$\xi_{i1} = \int_T x_i(t) f_1(t)\, dt,\ i = 1, \ldots, n$$

tal que la función peso $f_1(t)$ *es solución del problema de optimización*

$$\max_{\{\|f\|=1\}} Var\left[\xi_f\right] = \max_{\{\|f\|=1\}} \left\langle f, \widehat{C}(f) \right\rangle_u = \left\langle f_1, \widehat{C}(f_1) \right\rangle_u.$$

- *Suponiendo definidas las componentes* $\xi_1, \ldots, \xi_{j-1}$, *se define la* j*-ésima componente principal* $\xi_j = (\xi_{1j}, \ldots, \xi_{nj})^T$ *como la combinación lineal generalizada de las variables de la muestra, cuyos valores son*

$$\xi_{ij} = \int_T x_i(t) f_j(t)\, dt,\ i = 1, \ldots, n,$$

y tal que la función peso $f_j(t)$ *es solución del siguiente problema de optimización:*

$$\max_{\left\{\|f\|=1; Cov\left[\xi_l, \xi_f\right]=0,\ l=1,\ldots,j-1\right\}} Var\left[\xi_f\right] = \max_{\{\|f\|=1; \langle f_l, C(f)\rangle=0,\ l=1,\ldots,j-1\}} \left\langle f, \widehat{C}(f) \right\rangle_u = \left\langle f_j, \widehat{C}(f_j) \right\rangle_u \quad (2.19)$$

Por la forma de definirse, las componentes principales se obtienen de manera secuencial mediante la solución de los distintos problemas de máximos (2.19) que se resuelven de manera rápida a partir del siguiente teorema cuya demostración puede verse por ejemplo en Aguilera (1993).

Teorema 7 *Las funciones* f_j *que solucionan el problema de máximos (2.19) son las soluciones de la autoecuación*

$$\widehat{C}(f) = \lambda f \quad (2.20)$$

o equivalentemente

$$\int_T \widehat{C}(s,t) f(s)\, ds = \lambda f(t)\,,\ \forall t \in T. \tag{2.21}$$

El valor de dicho máximo se alcanza en

$$\left\langle f, \widehat{C}(f) \right\rangle_u = \lambda \tag{2.22}$$

Una ecuación como la (2.20) se conoce como problema de valores propios funcional con núcleo $\widehat{C}$ (el operador de covarianza muestral en este caso). La resolución de dicha ecuación es equivalente a la resolución de la (2.21) que no es más que una ecuación integral de tipo Fredholm de segunda especie cuyo núcleo es la covarianza muestral. Este tipo de ecuaciones integrales tienen en general como solución la trivial $f(t) = 0$ (Todorovic, 1992), pero bajo determinadas condiciones (núcleo autoadjunto, acotado y positivo) existen soluciones distintas de la trivial. Como consecuencia de las hipótesis de partida de continuidad en media cuadrática del proceso considerado, la función de covarianza muestral verifica tales condiciones por lo que existen soluciones no triviales para este problema. Es más, dado que el rango de dicho operador es $n-1$ podemos asegurar la existencia de $n-1$ soluciones no triviales distintas y ortonormales $f_1(t), \ldots, f_{n-1}(t)$, cada una asociada a valores propios λ reales, positivos y tales que $\lambda_1 > \ldots > \lambda_{n-1} \geq 0$. Los valores de los parámetros para los que existen soluciones de esta ecuación integral reciben el nombre de valores propios o principales, y las soluciones propiamente dichas funciones propias o factores principales.

A pesar de conocer la existencia de solución de la ecuación integral (2.21), la obtención de la misma es complicada, de hecho solución analítica sólo existe para determinados núcleos bien conocidos, convirtiendo la ecuación integral en una ecuación diferencial. En la literatura se han aportado diversas soluciones a este problema, la mayoría de ellas aproximadas, así en Aguilera et al. (1992) se utiliza la fórmula de cuadratura del trapecio para dar una solución, o bien el método de proyección ortogonal aproximando los factores principales en un espacio de dimensión finita generado por una base ortonormal. Otra alternativa consiste en aproximar la solución aprovechando la aproximación de las trayectorias mediante alguno de los métodos vistos en secciones previas: aproximación mínimo cuadrática o interpolación spline cúbica natural, alternativas que han sido desarrolladas por numerosos autores (Ramsay y Silverman, 1997; Aguilera et al., 1996) y que aplicaremos al final del presente capítulo y en los ejemplos que se presentarán en el siguiente.

La principal diferencia entre el ACP multivariante y el funcional es el número de componentes principales muestrales que se pueden extraer. En el caso multivariante dicho número era igual al número de variables originales disponibles mientras que en el caso funcional será, como hemos visto, el rango del operador de covarianza muestral, esto es, $n-1$.

Por otro lado en Deville (1973) se demuestra que las autofunciones y los autovalores del operador de covarianza muestral son estimadores consistentes de las correspondientes autofunciones y autovalores del operador de covarianza del proceso del que se obtuvo la muestra de trayectorias independientes.

Representación en componentes principales.

Al igual que en el caso multivariante, podemos dar una representación de la covarianza en términos de sus autovalores y autofunciones debida a Mercer en 1902, y a partir de ella la correspondiente representación de las trayectorias del proceso en términos de las cc. pp. que puede verse en Todorovic (1992) y que resumimos en el siguiente resultado:

Teorema 8 *Consideremos una muestra de trayectorias* $x_1(t), \ldots, x_n(t)$ *de un proceso estocástico real* $\{X(t) : t \in T\}$ *verificando las condiciones* $H_1 - H_3$. *Consideremos por otro lado las componentes principales* $\{\xi_1, \ldots, \xi_{n-1}\}$ *que se obtienen a partir de las funciones principales* $\{f_1(t), \ldots, f_{n-1}(t)\}$ *cuyos autovalores asociados son* $\{\lambda_1, \ldots, \lambda_{n-1}\}$, *soluciones de la ecuación (2.21). Entonces la función de covarianza muestral se puede expresar de la forma*

$$\widehat{C}(s,t) = \sum_{j=1}^{n-1} \lambda_j f_j(s) f_j(t),$$

y proporciona la siguiente representación en cc. pp. de las trayectorias muestrales:

$$x_i(t) = \sum_{j=1}^{n-1} \xi_{ij} f_j(t); \; i = 1, \ldots, n.$$

Esta representación de las trayectorias del proceso es óptima en el sentido de que es la mejor aproximación en media cuadrática de las trayectorias del proceso por combinaciones lineales de variables escalares (Fukunaga, 1990). Esto es, si consideramos otra representación

$$\widehat{x}_i(t) = \sum_{j=1}^{s} g_j(t) y_{ij}$$

donde

$$y_{ij} = \langle x_i(t), g_j(t) \rangle = \int_T x_i(t)\, g_j(t)\, dt,\ j = 1, \ldots, s$$

y $g_1, \ldots, g_s$ son funciones ortonormales que minimice

$$\sum_{i=1}^{n} \|x_i(t) - \widehat{x}_i(t)\|^2 = \sum_{i=1}^{n} \int_T (x_i(t) - \widehat{x}_i(t))^2\, dt,$$

entonces las variables y_{ij} son las componentes principales y las funciones $g_j(t)$ las autofunciones del operador covarianza muestral. Además el mínimo error cuadrático medio estará dado por

$$\sum_{j=1}^{s} \lambda_j.$$

Uno de los objetivos primordiales del ACPF muestral es explicar la variabilidad presente en una muestra de trayectorias de un proceso estocástico mediante un número reducido de variables. En este sentido, sería interesante disponer de una medida que proporcionara dicha variabilidad. En el caso "poblacional", se define la variabilidad del proceso a través de la traza del operador de covarianza (extensión natural de la traza de la matriz de varianzas-covarianzas en el caso multivariante). Así, en el caso muestral, la variabilidad de la muestra de trayectorias se podrá medir mediante

$$\widehat{V} = \int_T \widehat{C}(t,t)\, dt$$

de manera que aplicando la descomposición del teorema de Mercer vista anteriormente

$$\widehat{V} = \int_T \widehat{C}(t,t)\, dt = \sum_{j=1}^{n-1} \lambda_j.$$

Por lo tanto la variabilidad de los datos funcionales coincide con la suma de los valores propios (varianzas de las componentes principales) del operador de covarianza muestral al igual que en el caso multivariante.

Como hemos dicho anteriormente, el principal objetivo del ACPF muestral es reducir la dimensión del problema, esto es, aproximar las trayectorias muestrales mediante un número finito y reducido de componentes principales que acumulen un alto porcentaje de la variabilidad presente en los datos. En este

sentido, cada componente principal contiene un porcentaje de dicha variabilidad dado por

$$\frac{\lambda_j}{\sum_{j=1}^{n-1}\lambda_j}\times 100$$

de manera que seleccionando las s primeras componentes principales, conseguimos la siguiente representación aproximada de las funciones muestrales

$$x_i^{(s)}(t)=\sum_{j=1}^{s}\xi_{ij}f_j(t)$$

cuya varianza acumulada en % es

$$\left(\frac{\sum_{j=1}^{s}\lambda_j}{\sum_{j=1}^{n-1}\lambda_j}\times 100\right)\%$$

De este modo, cuanto más se aproxime esta cantidad a 100 mejor será la aproximación obtenida con las s primeras componentes principales.

Igual que ocurre en el caso multivariante, uno de los grandes problemas que presenta el ACPF es la interpretación de las componentes principales. Siguiendo la idea del ACP en el caso multivariante, las componentes principales se pueden interpretar a partir de la correlación existente entre ellas y las variables del proceso. En este sentido, y teniendo en cuenta que tanto las componentes principales como las funciones muestrales son centradas, se tiene que la covarianza entre la variable $x(t)=(x_1(t),\ldots,x_n(t))^T$ y la j-ésima componente principal es de la forma

$$\begin{aligned}Cov\left[x(t),\xi_j\right] &= \frac{1}{n-1}\sum_{i=1}^{n}x_i(t)\,\xi_{ij}\\ &= \int_T\left(\frac{1}{n-1}\sum_{i=1}^{n}x_i(t)\,x_i(s)\right)f_j(s)\,ds\\ &= \int_T\widehat{C}(s,t)\,f_j(s)\,ds=\lambda_j f_j(t)\end{aligned}$$

y por lo tanto la correlación será

$$Corr\left[x(t),\xi_j\right]=\frac{Cov\left[x(t),\xi_j\right]}{\sqrt{Var\left[x(t)\right]Var\left[\xi_j\right]}}=\frac{\lambda_j f_j(t)}{\sqrt{\widehat{C}(t,t)\,\lambda_j}}=\sqrt{\frac{\lambda_j}{\widehat{C}(t,t)}}f_j(t).$$

Aunque en la práctica se suele considerar la métrica usual en $L^2(T)$, a veces es necesario realizar transformaciones de los datos que equivalen a cambiar

de geometría en el espacio de las trayectorias. Para abordar este tema se generaliza la definición de componentes principales al caso de un espacio de Hilbert cualquiera H con producto escalar $\langle , \rangle_H$.

Consideremos una muestra aleatoria simple $w_1, \ldots, w_n$ de elementos de un proceso estocástico de segundo orden sobre un espacio de Hilbert $(H, \langle , \rangle_H)$, continuo en media cuadrática y cuyas trayectorias verifican que $\langle w_i, w_i \rangle_H < \infty$, $i = 1, \ldots, n$. Sean dos combinaciones lineales generalizadas con funciones peso $h_1, h_2 \in H$, definidas mediante los vectores de la forma

$$\zeta_{h_j} = \langle w_i, h_j \rangle_H, \; i = 1, \ldots, n, \; j = 1, 2, \tag{2.23}$$

entonces los momentos de primer y segundo orden se obtienen como:

- Media.

$$\overline{\zeta}_{h_1} = 0.$$

- Varianza

$$Var\left[\zeta_{h_1}\right] = \langle h_1, \mathcal{C}_H\left(h_1\right)\rangle_H$$

siendo $\mathcal{C}_H$ el operador de covarianza definido en (2.2).

- Covarianza

$$Cov\left[\zeta_{h_1}, \zeta_{h_2}\right] = \langle h_1, \mathcal{C}_H\left(h_2\right)\rangle_H = \langle \mathcal{C}_H\left(h_1\right), h_2\rangle_H .$$

Una vez definidos los elementos necesarios, veamos la definición de las componentes principales.

Definición 15 *Dada una muestra aleatoria simple $w_1, \ldots, w_n$ de elementos de un proceso estocástico sobre un espacio de Hilbert $(H, \langle , \rangle_H)$ verificando las hipótesis $H_1 - H_3$ y que $\langle w_i, w_i \rangle_H < \infty$, $i = 1, \ldots, n$ sobre dicho espacio, se define la j-ésima componente principal $\zeta_j = \left(\zeta_{1j}, \ldots, \zeta_{nj}\right)^T$ como la combinación lineal generalizada de las variables de la muestra, cuyos valores son $\zeta_{ij} = \langle w_i, h_j \rangle_H$ y tal que la función peso h_j es solución del siguiente problema de optimización:*

$$\max_{\left\{\|h\|_H = 1; \langle h_l, \mathcal{C}_X(h)\rangle = 0, \; l = 1, \ldots, j-1\right\}} \langle h, \mathcal{C}_H\left(h\right)\rangle_H = \langle h_j, \mathcal{C}_H\left(h_j\right)\rangle_H$$

Las funciones peso h_j se obtienen también en este caso diagonalizando el operador de covarianza. Es decir, como solución del siguiente problema de valores propios:

$$\mathcal{C}_H\left(h\right) = \lambda h.$$

2.4.2 Estimación aproximada de las componentes principales en un espacio de dimensión finita

Según se ha visto en las secciones anteriores, la obtención de las componentes principales pasa por la resolución de una ecuación integral de tipo Fredholm. El hecho de conocer la existencia de solución de una ecuación de este tipo no resuelve el problema de la obtención de la misma, lo cual es posible sólo para determinados núcleos bien conocidos. Además de todas estas dificultades, nos encontramos con que usualmente la función covarianza muestral es desconocida, ya que no es usual disponer de la expresión explícita de las trayectorias muestrales de las que proviene, sino de sus valores observados en un conjunto finito de instantes de tiempo no necesariamente iguales para todas las funciones muestrales. Para resolver esta problemática se recurre a métodos de aproximación numérica que reconstruyen la naturaleza funcional de las trayectorias en un espacio de dimensión finita y reducen el problema funcional a uno múltiple. Un estudio pormenorizado de las técnicas desarrolladas por los investigadores en los últimos años puede verse en Valderrama et al. (2000) o Ramsay y Silvermann (1997).

El siguiente resultado muestra cómo el ACPF de un proceso estocástico cuyas trayectorias pertenecen a un espacio de dimensión finita generado por una base de funciones, es equivalente al ACP multivariante de sus coordenadas respecto de una métrica concreta y equivalentemente al ACP de una transformación lineal de dichas coordenadas con respecto a la métrica usual. La demostración del mismo se puede encontrar en Aguilera et al. (2002).

Teorema 9 *Consideremos X una variable aleatoria hilbertiana de segundo orden sobre un espacio de Hilbert E generado por una base $\{\phi_1, \ldots, \phi_p\}$ con un producto escalar definido sobre $(E, \langle , \rangle_E)$. Entonces cualquier elemento de X se expresará como*

$$X(\omega) = \sum_{j=1}^{p} Y_j(\omega)\, \phi_j.$$

Sea el operador

$$\begin{array}{rccc} U: & \mathbb{R}^p & \rightarrow & E \\ & \eta & \rightarrow & U(\eta) = \sum_{j=1}^{p} \eta_j \phi_j \end{array}$$

Consideremos en $\mathbb{R}^p$ el producto escalar $\langle , \rangle_\Psi$ definido

$$\langle \eta_1, \eta_2 \rangle_\Psi = {\eta_1}^T \Psi \eta_2 \quad \forall \eta_1, \eta_2 \in \mathbb{R}^p$$

con Ψ la matriz de productos escalares de los elementos básicos,

$$\Psi = \left(\left\langle \phi_i, \phi_j \right\rangle_E\right)_{p\times p},\ i,j = 1,\ldots,p.$$

Bajo estas condiciones se tiene que los siguientes ACP son equivalentes

1. *ACP de $\Psi^{1/2}Y$ con respecto a $(\mathbb{R}^p, \langle,\rangle)$*

$$\Psi^{1/2}Y\left(\omega\right) = \Psi^{1/2}E\left[Y\right] + \sum_{j=1}^{p}\xi_j\left(\omega\right)\mathcal{V}_j$$

siendo $\langle,\rangle$ el producto escalar usual en $\mathbb{R}^p$ definido por $\langle\eta_1,\eta_2\rangle = \eta_1^T\eta_2,\ \forall\eta_1,\eta_2 \in \mathbb{R}^p$, $\mathcal{V}_j$, $j = 1,\ldots,p$ los vectores propios de la matriz de covarianzas de $\Psi^{1/2}Y$, y el vector aleatorio $Y = (Y_1,\ldots,Y_p)^T$ el de las coordenadas de la variable aleatoria X respecto de la base.

2. *El ACP del vector aleatorio Y con respecto a $(\mathbb{R}^p, \langle,\rangle_\Psi)$*

$$Y\left(\omega\right) = E\left[Y\right] + \sum_{j=1}^{p}\xi_j\left(\omega\right)\Psi^{-1/2}\mathcal{V}_j$$

siendo $E[Y]$ la esperanza de Y en $\mathbb{R}^p$.

3. *ACP de X con respecto a $(E, \langle,\rangle_E)$*

$$X\left(\omega\right) = U\left(E\left[Y\right]\right) + \sum_{j=1}^{p}\xi_j\left(\omega\right)U\left(\Psi^{-1/2}\mathcal{V}_j\right)$$

Observemos que si $\{\phi_1,\ldots,\phi_p\}$ es una base ortonormal, entonces el producto escalar $\langle,\rangle_\Psi = \langle,\rangle$ y $\Psi = I$.

Después del resultado anterior en el que se ve cómo un ACP funcional de un proceso con trayectorias en un espacio de dimensión finita es equivalente a uno múltiple de sus coordenadas, ya estamos en condiciones de resolver la ecuación integral de tipo Fredholm (2.21) del problema de valores propios funcionales que lleva a la obtención de las componentes principales en el ambiente que nos ocupa.

Consideremos un proceso estocástico real $\{X(t) : t \in T\}$ bajo las condiciones consideradas a lo largo del presente capítulo (hipótesis $H_1 - H_3$) y sea $x_1(t),\ldots,x_n(t)$ una muestra aleatoria simple de trayectorias. En lo que sigue

consideraremos que dichas trayectorias muestrales se expresan en términos de una base de funciones $\phi_1(t), \ldots, \phi_p(t)$ en la forma

$$x_i(t) = \sum_{j=1}^{p} a_{ij}\phi_j(t), \; i = 1, \ldots, n$$

con lo que la función de covarianza muestral quedará

$$\widehat{C}(s,t) = \sum_{j=1}^{p}\sum_{k=1}^{p} \frac{1}{n-1} \sum_{i=1}^{n} a_{ij}a_{ik}\phi_j(t)\,\phi_k(s). \tag{2.24}$$

Si aplicamos el teorema anterior para $E = \langle \phi_1(t), \ldots, \phi_p(t) \rangle \subset L^2(T)$ con el producto escalar usual en $L^2(T)$, $\langle , \rangle_E = \langle , \rangle_u$ veremos cómo las autofunciones que definen a las componentes principales se expresan también en términos de la base de funciones que estemos considerando

$$f_j(t) = \sum_{l=1}^{p} f_{lj}\phi_l(t), \; j = 1, \ldots, p, \tag{2.25}$$

con vector de coeficientes f_j dado por una transformación lineal de los autovectores de un determinado ACP múltiple. El siguiente corolario demuestra este hecho.

Corolario 1 *Sea $\{X(t) : t \in T\}$ un proceso estocástico centrado verificando $H_1 - H_3$, y supongamos que las trayectorias de dicho proceso se pueden expresar en términos de los elementos de una base $\{\phi_1(t), \ldots, \phi_p(t)\}$ de la forma (2.11). Consideremos una muestra aleatoria simple $x_1(t), \ldots, x_n(t)$ de dicho proceso tal que $\overline{x}(t) = 0$ siendo A la matriz que tiene por filas los coeficientes de cada trayectoria respecto de la base. Los siguientes ACP muestrales son equivalentes en el sentido de que proporcionan las mismas componentes principales:*

1. *ACP de $A\Psi^{1/2}$ con respecto a $(\mathbb{R}^p, \langle , \rangle)$,*

$$A\Psi^{1/2} = \Gamma G^T$$

siendo G la matriz que tiene por columnas los autovectores $\mathcal{G}_j$ de la matriz de covarianzas de $A\Psi^{1/2}$ y Γ la matriz que tiene por columnas las componentes principales de $A\Psi^{1/2}$.

2. *ACP de A con respecto a $(\mathbb{R}^p, \langle , \rangle_\Psi)$.*

$$A = \Gamma\left(G^T\Psi^{-1/2}\right) = \Gamma F^T$$

con $F = \Psi^{-1/2}G$

3. *ACP de $x_1(t), \ldots, x_n(t)$ con respecto a la métrica usual de $L^2(T)$ sobre el espacio generado por las funciones básicas, $\langle \phi_1(t), \ldots, \phi_p(t) \rangle$.*

$$x_i(t) = \sum_{j=1}^{p} \xi_{ij} f_j(t), \; i = 1, \ldots, n$$

donde las autofunciones $f_j(t)$ del operador de covarianza son de la forma

$$f_j(t) = U(f_j) = \sum_{l=1}^{p} f_{lj}\phi_l(t)$$

con $\mathcal{F}_j = \Psi^{-1/2}\mathcal{G}_j$ *y*

$$\xi_{ij} = \int_T x_i(t) f_j(t)\, dt = \int_T \left(\sum_{k=1}^{p} a_{ik}\phi_k(t)\right)\left(\sum_{l=1}^{p} f_{lj}\phi_l(t)\right) dt$$

$$= \sum_{k=1}^{p}\sum_{l=1}^{p} a_{ik} f_{lj}\left(\int_T \phi_k(t)\,\phi_l(t)\, dt\right) = \sum_{k=1}^{p}\sum_{l=1}^{p} a_{ik} f_{lj}\psi_{kl} = a_i^T\Psi\mathcal{F}_j$$

$$= a_i^T\Psi\Psi^{-1/2}\mathcal{G}_j = a_i^T\Psi^{1/2}\mathcal{G}_j;\; i = 1, \ldots, n$$

donde a_i^T es correspondiente fila de la matriz A y $\mathcal{F}_j$ el vector asociado a la correspondiente componente principal. En forma matricial, la matriz que tiene por columnas las componentes principales será

$$\Gamma = A\Psi F = A\Psi^{1/2}G$$

2.5 Modelo de regresión logística funcional en términos de las componentes principales

En la Sección 3 introdujimos el modelo de regresión logística funcional y pusimos de manifiesto la problemática de la estimación de la función parámetro así como las malas estimaciones que podrían obtenerse con la aproximación de las trayectorias en espacios de dimensión finita debido probablemente al

mal condicionamiento (multicolinealidad) existente en la matriz de diseño del modelo múltiple resultante. Al igual que hicimos en el caso múltiple, a continuación pretendemos disminuir los efectos de dicha multicolinealidad y reducir la dimensión del problema, considerando como variables explicativas un conjunto de componentes principales funcionales asociadas a las trayectorias muestrales en lugar de las propias trayectorias.

Para definir el modelo de regresión logística en términos de todas las componentes principales, consideraremos el conjunto de trayectorias explicativas de dicho modelo $x_1(t), \ldots, x_n(t)$, que, como hemos visto a lo largo del presente trabajo, son una muestra de trayectorias de un proceso estocástico en tiempo continuo verificando las hipótesis $H_1 - H_3$, y asociadas a una muestra aleatoria simple $y_1, \ldots, y_n$ de observaciones de la variable respuesta dicotómica.

En la Sección 2.3 se propuso la estimación del modelo funcional (2.10) en base a un modelo múltiple de matriz de diseño $(\mathbf{1} \mid A\Psi)$ con transformación logit

$$l_i = \ln\left[\frac{\pi_i}{1-\pi_i}\right] = \alpha + \sum_{j=1}^{p}\sum_{k=1}^{p} a_{ij}\psi_{jk}\beta_k,\ i = 1, \ldots, n$$

obteniendo así una estimación de la función parámetro poco precisa (bajo multicolinealidad) en la forma

$$\widehat{\beta}(t) = \sum_{k=1}^{p} \widehat{\beta}_k \phi_k(t)$$

con $\left(\widehat{\alpha}, \widehat{\beta}_1, \ldots, \widehat{\beta}_p\right)^T$ los parámetros estimados de este modelo por máxima verosimilitud.

En base a este modelo múltiple se plantea una alternativa a dicha estimación que consiste en considerar las componentes principales de $A\Psi$ como se hizo en el caso múltiple en la sección 1.4, de modo que si llamamos $\mathcal{Z} = (z_{ij})$ a la matriz que tiene por columnas dichas componentes principales, el modelo anterior quedaría

$$l_i = \ln\left[\frac{\pi_i}{1-\pi_i}\right] = \alpha + \sum_{j=1}^{p} z_{ij}\gamma_j,\ i = 1, \ldots, n \tag{2.26}$$

y matricialmente

$$L = \mathbf{1}\alpha + \mathcal{Z}\gamma$$

con $L = (l_1, \ldots, l_n)^T$; y podremos dar una estimación equivalente del vector de coordenadas de $\beta(t)$ mediante la estimación de γ de la forma

$$\widehat{\beta}_k = \sum_{j=1}^{p} v_{kj}\widehat{\gamma}_j,\ k = 1, \ldots, p \tag{2.27}$$

o en forma matricial

$$\widehat{\beta} = \mathcal{V}\widehat{\gamma}$$

siendo $\mathcal{V}$ la matriz que tiene por columas los autovectores de la matriz de covarianzas de $A\Psi$.

Una vez dada esta solución para la estimación de los parámetros del modelo múltiple considerado, podemos obtener una estimación equivalente de la función parámetro del modelo funcional de la forma

$$\widehat{\beta}(t) = \sum_{k=1}^{p} \widehat{\beta}_k \phi_k(t). \tag{2.28}$$

Consideremos por otro lado la matriz de componentes principales de las trayectorias originales $x_1(t), \ldots, x_n(t)$,

$$\Gamma = \begin{pmatrix} \xi_{11} & \xi_{12} & \cdots & \xi_{1n-1} \\ \xi_{21} & \xi_{22} & \cdots & \xi_{2n-1} \\ \cdots & \cdots & \cdots & \cdots \\ \xi_{n1} & \xi_{n2} & \cdots & \xi_{nn-1} \end{pmatrix}$$

donde

$$\xi_{ij} = \int_T x_i(t) f_j(t)\, dt,\ i = 1, \ldots, n;\ j = 1, \ldots, n-1$$

y $f_j(t)$, $j = 1, \ldots, n-1$ son las funciones propias del operador de covarianza muestral de los datos funcionales.

El modelo logístico funcional (2.10) se puede expresar alternativamente de la forma

$$\begin{aligned} \ln\left[\frac{\pi_i}{1-\pi_i}\right] &= \alpha + \int_T x_i(t)\beta(t)\, dt \\ &= \alpha + \int_T \left(\sum_{j=1}^{n-1} \xi_{ij} f_j(t)\right)\beta(t)\, dt \\ &= \alpha + \sum_{j=1}^{n-1} \xi_{ij} \int_T f_j(t)\beta(t)\, dt \\ &= \alpha + \sum_{j=1}^{n-1} \xi_{ij}\varphi_j,\ i = 1, \ldots, n \end{aligned} \tag{2.29}$$

que es un modelo de regresión logística múltiple en términos de todas las componentes principales con parámetros α y

$$\varphi_j = \int_T f_j(t)\,\beta(t)\,dt,\ j = 1,\ldots,n-1$$

Análogamente, el modelo (2.29) en términos de las componentes principales es equivalente a un modelo funcional con función parámetro

$$\beta(t) = \sum_{j=1}^{n-1} f_j(t)\,\varphi_j,$$

de hecho,

$$\begin{aligned}\ln\left[\frac{\pi_i}{1-\pi_i}\right] &= \alpha + \sum_{j=1}^{n-1}\xi_{ij}\varphi_j = \alpha + \sum_{j=1}^{n-1}\left(\int_T x_i(t)\,f_j(t)\,dt\right)\varphi_j \\ &= \alpha + \int_T x_i(t)\left(\sum_{j=1}^{n-1} f_j(t)\,\varphi_j\right)dt = \alpha + \int_T x_i(t)\,\beta(t)\,dt\end{aligned}$$

Efectivamente, si tomamos dicha función parámetro,

$$\beta(t) = \sum_{j=1}^{n-1} f_j(t)\,\varphi_j \Rightarrow \int_T \beta(t)\,f_i(t)\,dt = \sum_{j=1}^{n-1}\left(\int_T f_i(t)\,f_j(t)\,dt\right)\varphi_j = \delta_{ij}\varphi_j = \varphi_i$$

con δ_{ij} la delta de Kronecker. Así, definiendo esta función parámetro, se obtiene la misma expresión para los parámetros correspondientes al modelo múltiple en términos de las componentes principales que obtuvimos anteriormente.

En sentido inverso si $\varphi_j = \int_T f_j(t)\,\beta(t)\,dt,\ j = 1,\ldots,p$ y asumimos que la función parámetro $\beta(t)$ pertenece al espacio generado por las autofunciones es inmediato que

$$\beta(t) = \sum_{j=1}^{p}\langle f_j, \beta\rangle\, f_j(t) = \sum_{j=1}^{p}\varphi_j f_j(t).$$

En la práctica es usual considerar que las trayectorias $x_1(t),\ldots,x_n(t)$ pertenecen a un espacio de dimensión finita generado por una base de funciones $\{\phi_1(t),\ldots,\phi_n(t)\}$ lo que nos lleva, como vimos anteriormente, a que las autofunciones que definen a las cc. pp. son de la forma

$$f_j(t) = \sum_{k=1}^{p} f_{kj}\phi_k(t),\ j = 1,\ldots,p,$$

con coeficientes f_{kj} dados por los elementos del vector que se obtiene mediante la transformación lineal $\mathcal{F}_j = \Psi^{-1/2}\mathcal{G}_j$ de los autovectores $\mathcal{G}_j$ de la matriz de covarianza de $A\Psi^{1/2}$, y a que las componentes principales sean las asociadas a la matriz $A\Psi^{1/2}$ dadas por

$$\xi_{ij} = a_i^T \Psi \mathcal{F}_j = a_i^T \Psi^{1/2} \mathcal{G}_j;\ i = 1, \ldots, n,\ j = 1, \ldots, p.$$

En este caso la función parámetro del modelo de regresión logística funcional (2.10) se podrá estimar a partir de la estimación de los parámetros del modelo múltiple en términos de las cc. pp. en la forma

$$\widehat{\beta}(t) = \sum_{j=1}^{p} f_j(t)\,\widehat{\varphi}_j = \sum_{j=1}^{p}\sum_{k=1}^{p} f_{kj}\phi_k(t)\,\widehat{\varphi}_j = \sum_{k=1}^{p} \widehat{\beta}_k \phi_k(t)$$

donde

$$\widehat{\beta}_k = \sum_{j=1}^{p} f_{kj}\widehat{\varphi}_j,\ k = 1, \ldots, p,$$

que en forma vectorial queda

$$\widehat{\beta} = F\widehat{\varphi}$$

con $\widehat{\varphi} = \left(\widehat{\varphi}_1, \ldots, \widehat{\varphi}_p\right)^T$, $\widehat{\beta} = \left(\widehat{\beta}_1, \ldots, \widehat{\beta}_p\right)^T$ y F la matriz que tiene los elementos f_{kj}. De este modo se obtiene una estimación de las coordenadas de la función parámetro en el espacio $\langle \phi_1(t), \ldots, \phi_p(t) \rangle$ equivalente a la obtenida en la sección 3.2 donde se proponía la estimación de la función parámetro del modelo logístico funcional mediante la estimación de sus coordenadas en el espacio generado por la base $\{\phi_1(t), \ldots, \phi_p(t)\}$. De este modo el modelo funcional se reducía a un modelo múltiple de matriz de diseño $(\mathbf{1} \mid A\Psi)$.

De todo lo visto podemos concluir que existen dos formas alternativas y equivalentes para la estimación en componentes principales de las coordenadas de la función parámetro de un modelo de regresión logística funcional en el espacio donde se encuentran las trayectorias de dicho modelo y que está generado por una base de funciones:

- La primera consiste en utilizar los autovectores de la matriz de covarianzas de la matriz de diseño del modelo múltiple, $A\Psi$, y los parámetros estimados del modelo logístico múltiple que tiene por covariables a las componentes principales de $A\Psi$. A esta solución la llamaremos ACP1 y como veremos en los resultados que se muestran a continuación, equivale a un ACP funcional de una transformación de los datos funcionales originales con respecto a la métrica usual de $L^2(T)$.

- La segunda consiste en utilizar los vectores $\mathcal{F}_j$ que se obtienen mediante la transformación $\Psi^{-1/2}\mathcal{G}_j$ de los autovectores de la matriz de covarianzas de $A\Psi^{1/2}$ y los parámetros estimados del modelo logístico múltiple que tiene por covariables a las componentes principales de $A\Psi^{1/2}$. A esta solución la llamaremos ACP2 y como se vio en el Corolario 1, equivale al ACP de los datos funcionales originales con respecto a la métrica usual en $L^2(T)$.

En definitiva las dos alternativas consisten en sendos ACPF o bien de los datos originales o bien de una transformación de los mismos que, por otro lado, podría ayudar a la interpretación de las componentes principales así como evitar posibles componentes triviales. El siguiente teorema, debido a Aguilera et al. (2002), nos permitirá encontrar la transformación de los datos originales que lleva al ACP de $A\Psi$.

Teorema 10 *Bajo las condiciones del Teorema 9 y con $D_{p\times p}$ una matriz real no singular, los siguientes ACP son equivalentes:*

1. *ACP de DY con respecto a $(\mathbb{R}^p, \langle,\rangle)$*

$$DY(\omega) = DE[Y] + \sum_{j=1}^{p} \xi_j(\omega)\mathcal{V}_j$$

2. *ACP de Y con respecto a $(\mathbb{R}^p, \langle,\rangle_K)$*

$$Y(\omega) = E[Y] + \sum_{j=1}^{p} \xi_j(\omega) D^{-1}\mathcal{V}_j$$

con $K = D^TD$ y el producto escalar de la forma $\langle\eta_1,\eta_2\rangle_K = \eta_1^TK\eta_2$, $\forall\eta_1,\eta_2 \in \mathbb{R}^p$.

3. *ACP de X con respecto a $(E, \langle,\rangle_L)$*

$$X(\omega) = U(E[Y]) + \sum_{j=1}^{p} \xi_j(\omega) U\left(D^{-1}\mathcal{V}_j\right)$$

donde el producto escalar $\langle,\rangle_L$ está dado por

$$\langle f_1, f_2\rangle_L = \langle L(f_1), L(f_2)\rangle_E = \eta_1^TK\eta_2,\ \forall f_1 = \phi^T\eta_1 \text{ y } f_2 = \phi^T\eta_2$$

siendo $\phi = (\phi_1,\ldots,\phi_p)^T$, $L(f) = \phi^T\Psi^{-1/2}D\eta$, $\forall f = \phi^T\eta$ y $U(\eta) = \phi^T\eta$, $\forall\eta \in \mathbb{R}^p$.

4. *ACP de $L \circ X$ con respecto a $(E, \langle , \rangle_E)$*

$$LX(\omega) = L(U(E[Y])) + \sum_{j=1}^{p} \xi_j(\omega) L\left(U\left(D^{-1}\mathcal{V}_j\right)\right)$$

Observemos que en el caso de una base ortonormal el operador L es de la forma $L(f) = \phi^T D\eta$.

Este teorema establece que llevar a cabo el ACP de una transformación de las coordenadas (DY) con respecto al producto escalar usual en $\mathbb{R}^p$ es equivalente a llevar a cabo el ACPF de una transformación de los datos $(L \circ X)$ con respecto al producto escalar usual en $L^2(T)$.

Para el caso del ACPF muestral de un proceso estocástico real con trayectorias en un subespacio de $L^2(T)$ generado por una base de funciones $E = \langle \phi_1(t), \ldots, \phi_p(t) \rangle$ se tiene el siguiente corolario.

Corolario 2 *Bajo las condiciones del Corolario 1 se tiene que los siguientes ACP son equivalentes:*

1. *ACP de $A\Psi$ respecto a $(\mathbb{R}^p, \langle , \rangle)$*

$$A\Psi = \mathcal{Z}\mathcal{V}^T$$

con $\mathcal{Z} = A\Psi\mathcal{V}$ la matriz que tiene por columnas a las componentes principales de $A\Psi$ con la métrica usual, y $\mathcal{V}$ la matriz que tiene por columnas a los autovectores de la matriz de covarianzas de $A\Psi$.

2. *ACP de A con respecto a $(\mathbb{R}^p, \langle , \rangle_K)$*

$$A = \mathcal{Z}\mathcal{V}^T\Psi^{-1}$$

con $K = \Psi^T\Psi = \Psi^2$ y el producto escalar de la forma $\langle \eta_1, \eta_2 \rangle_K = \eta_1^T\Psi^2\eta_2$, $\forall \eta_1, \eta_2 \in \mathbb{R}^p$.

3. *ACP de $x_1(t), \ldots, x_n(t)$ con respecto a $(E, \langle , \rangle_L)$ definida anteriormente*

$$x_i(t) = \sum_{j=1}^{p} z_{ij} U\left(\Psi^{-1}\mathcal{V}_j\right)(t), \; i = 1, \ldots, n$$

con $\mathcal{V}_j$ la j-ésima columna de $\mathcal{V}$ (j-ésimo vector propio de la matriz de covarianzas de $A\Psi$).

4. ACP de $L(x_1)(t), \ldots, L(x_n)(t)$ *con respecto a* $(E, \langle , \rangle_u)$

$$L(x_i)(t) = \sum_{j=1}^{p} z_{ij} L\left(U\left(\Psi^{-1}\mathcal{V}_j\right)\right)(t), \; i = 1, \ldots, n$$

con $L(x_i) = \phi^T \Psi a_i^T$.

2.6 Modelo de regresión logística funcional en componentes principales

Una vez dada una estimación de la función parámetro del modelo de regresión logística funcional en base a dos tipos distintos de ACPF, vamos a eliminar cc. pp. obteniendo así dos clases de modelos de regresión logística funcional en componentes principales. Al igual que en el caso multivariante, vamos a considerar un número reducido de componentes principales eliminando de la matriz Γ o de la matriz $\mathcal{Z}$ una serie de columnas.

Modelo de regresión logística funcional en componentes principales mediante ACP1

Sea el modelo de regresión logística múltiple (2.13) con matriz de diseño $(\mathbf{1} \mid A\Psi)$ obtenido del funcional (2.8) al considerar que tanto las trayectorias explicativas $x_1(t), \ldots, x_n(t)$ como la función parámetro $\beta(t)$ pertenecen al espacio de dimensión finita generado por la base de funciones $\{\phi_1(t), \ldots, \phi_p(t)\}$. Sea $\mathcal{Z}$ la matriz de componentes principales de $A\Psi$ y $\mathcal{V}$ la de los autovectores correspondientes, entonces, de igual modo que vimos en el caso múltiple (Sección 1.4), podemos considerar la descomposición de la matriz de cc. pp. de la forma $Z = \left(Z_{(s)} \mid Z_{(r)}\right)$, siendo $Z = (\mathbf{1} \mid \mathcal{Z})$ y $Z_{(s)}$ la submatriz formada por las primeras columnas de Z. Bajo estas condiciones se define el modelo de regresión logística funcional en componentes principales mediante ACP1, como

$$Y = \pi\left(Z_{(s)}\right) + \varepsilon_{(s)}$$

donde $Y = (y_1, \ldots, y_n)^T$ es el vector de observaciones de la variable respuesta dicotómica, y $\pi\left(Z_{(s)}\right)$ el vector de componentes

$$\pi_{i(s)}^{1)} = \frac{\exp\left\{\gamma_{0(s)} + \sum_{j=1}^{s} z_{ij}\gamma_{j(s)}\right\}}{1 + \exp\left\{\gamma_{0(s)} + \sum_{j=1}^{s} z_{ij}\gamma_{j(s)}\right\}}, \; i = 1, \ldots, n.$$

Equivalentemente, en términos de la transformación logit se tiene

$$l_{i(s)}^{1)} = \ln\left[\frac{\pi_{i(s)}^{1)}}{1-\pi_{i(s)}^{1)}}\right] = \gamma_{0(s)} + \sum_{j=1}^{s} z_{ij}\gamma_{j(s)},\ i = 1,\ldots,n$$

que no es más que el truncamiento de la expresión (2.26) en los s primeros términos o lo que es igual, el modelo logístico múltiple con matriz de diseño $Z_{(s)}$.

Supongamos que ajustamos este modelo y obtenemos los parámetros estimados $\widehat{\gamma}_{(s)} = \left(\widehat{\gamma}_{1(s)},\ldots,\widehat{\gamma}_{s(s)}\right)^T$, entonces a partir de la expresión (2.27) podemos dar una estimación de las coordenadas de la función parámetro de la forma

$$\widehat{\beta}_{k(s)}^{1)} = \sum_{j=1}^{s} v_{kj}\widehat{\gamma}_{j(s)},\ k = 1,\ldots,p$$

que matricialmente queda

$$\widehat{\beta}_{(s)}^{1)} = \mathcal{V}_{(s)}\widehat{\gamma}_{j(s)}$$

siendo v_{kj} los elementos correspondientes de $\mathcal{V}_{(s)}$ que no es más que la submatriz formada por las primeras s columnas de $\mathcal{V}$. De aquí, la correspondiente estimación de la función parámetro sería

$$\widehat{\beta}_{(s)}^{1)}(t) = \sum_{k=1}^{p} \widehat{\beta}_{k(s)}^{1)}\phi_k(t)$$

Modelo de regresión logística funcional en componentes principales mediante ACP2

De igual modo que en el caso de ACP1, llamemos $\Gamma_{(s)}$ a la submatriz formada por la primera columna de unos y posteriormente las s primeras columnas de Γ con $s < p$. Entonces si llamamos

$$\pi_{i(s)}^{2)} = \frac{\exp\left\{\varphi_{0(s)} + \sum_{j=1}^{s}\xi_{ij}\varphi_{j(s)}\right\}}{1+\exp\left\{\varphi_{0(s)} + \sum_{j=1}^{s}\xi_{ij}\varphi_{j(s)}\right\}},\ i = 1,\ldots,n,$$

el modelo de regresión logística funcional en componentes principales según ACP2 se formula en la forma

$$Y = \pi\left(\Gamma_{(s)}\right) + \varepsilon_{(s)}$$

donde Y es el vector de observaciones dicotómicas de la variable respuesta, $\pi\left(\Gamma_{(s)}\right) = \left(\pi^{2)}_{1(s)}, \ldots, \pi^{2)}_{n(s)}\right)^T$ definido anteriormente y $\varepsilon_{(s)}$ es el vector de errores ya habitual. En términos de la transformación logit el modelo se expresa de la forma

$$l^2_{i(s)} = \ln\left[\frac{\pi^{2)}_{i(s)}}{1-\pi^{2)}_{i(s)}}\right] = \varphi_{0(s)} + \sum_{j=1}^{s} \xi_{ij}\varphi_{j(s)},\ i = 1, \ldots, n \tag{2.30}$$

obteniendo en este caso un modelo de regresión logística más reducido, con menor número de parámetros a estimar. Expresando las cc. pp. en términos de las trayectorias originales se tiene

$$\begin{aligned}
\ln\left[\frac{\pi^{2)}_{i(s)}}{1-\pi^{2)}_{i(s)}}\right] &= \varphi_{0(s)} + \sum_{j=1}^{s} \xi_{ij}\varphi_{j(s)} \\
&= \varphi_{0(s)} + \sum_{j=1}^{s} \left(\int_T x_i(t) f_j(t)\, dt\right) \varphi_{j(s)} \\
&= \varphi_{0(s)} + \int_T x_i(t) \left(\sum_{j=1}^{s} f_j(t)\, \varphi_{j(s)}\right) dt \\
&= \varphi_{0(s)} + \int_T x_i(t)\, \beta^{2)}_{(s)}(t)\, dt;\ i = 1, \ldots, n
\end{aligned}$$

tomando

$$\beta^{2)}_{(s)}(t) = \sum_{j=1}^{s} f_j(t)\, \varphi_{j(s)}$$

que permite dar la siguiente estimación aproximada de la función parámetro en términos de los elementos de la base considerada

$$\widehat{\beta}^{2)}_{(s)}(t) = \sum_{j=1}^{s} f_j(t)\, \widehat{\varphi}_{j(s)} = \sum_{j=1}^{s}\sum_{k=1}^{p} \widehat{\varphi}_{j(s)} f_{jk}\phi_k(t) = \sum_{k=1}^{p} \widehat{\beta}_{k(s)}\phi_k(t)$$

siendo $\widehat{\varphi}_{j(s)}$ el j-ésimo parámetro estimado del modelo (2.30). En forma matricial, la expresión anterior quedará

$$\widehat{\beta}_{(s)} = F_{(s)}\widehat{\varphi}_{(s)}$$

con $F_{(s)}$ la submatriz formada por las primeras s columnas de F, $\widehat{\varphi}_{(s)}$ el vector de elementos $\widehat{\varphi}_{j(s)}$, $j = 1, \ldots, s$ y $\widehat{\beta}_{(s)}$ el vector que tiene las coordenadas de la función parámetro estimadas por el modelo en términos de la base.

Al igual que en el caso de la utilización de todas las componentes principales, los coeficientes de la expansión de la función parámetro, estimados a partir de un número reducido de componentes principales, se obtendrán multiplicando los estimados por estos últimos modelos, por los correspondientes coeficientes de la expresión de las autofunciones en términos de los elementos básicos.

Llegados a este punto debemos notar que las estimaciones que se obtienen de $\widehat{\gamma}_{(s)}$ y $\widehat{\varphi}_{(s)}$mediante estos modelos son completamente diferentes de las que se obtienen truncando en los primeros s términos el vector $\widehat{\gamma} = \left(\widehat{\gamma}_1, \ldots, \widehat{\gamma}_p\right)^T$ (parámetros estimados obtenidas mediante el modelo (2.26)), por lo que la estimación $\widehat{\beta}_{(s)}(t)$ de la función parámetro así obtenida es distinta de la obtenida truncando la expresión (2.28).

Observemos también que si la base del espacio al que consideramos que pertenecen tanto las trayectorias explicativas como la función parámetro es ortonormal, las dos alternativas presentadas aquí para la estimación de dicha función parámetro coinciden.

2.7 Selección de componentes principales

Como ya se explicó en el caso múltiple, la elección del número de componentes principales a utilizar, es un aspecto muy tratado en la literatura. Aucott y Garthwaite (2000), Mansfield et al. (1977), Hocking (1976), Gunst y Mason (1977) son algunos de los autores que han tratado este aspecto en el caso lineal y han dado diversas soluciones, siendo las más populares la elección de las componentes en orden de variabilidad, o bien la elección en base a alguna medida de correlación con la variable a explicar o bien considerar aquéllas que mejor explican la respuesta en el caso de modelos de regresión.

En el caso que nos ocupa, y como ya apuntamos en el modelo múltiple, vamos a considerar el primero y el último de los anteriores métodos. Esto nos permitirá comparar la reducción de dimensión que se consigue con uno y con otro y lo podremos comprobar en la sección siguiente en la que se presentan los ejemplos simulados. También entonces explicamos la importancia de la elección del número óptimo de componentes a elegir, quedando patente que, a efectos de la interpretación de los parámetros, sería conveniente utilizar aquel número de componentes que proporcionara parámetros estimados próximos a los reales. En el caso funcional la interpretación de la función parámetro también es importante de modo que resulta interesante que dicha función se

estime de la manera más exacta posible. Este hecho debería tenerse en cuenta y tomar el número de componentes a utilizar en función del valor de alguna medida de la proximidad entre la función estimada y la real.

Una medida de la distancia entre la función parámetro del modelo y una estimación de la misma (por cualquiera de los métodos presentados en esta memoria) es el error cuadrático medio integrado que se define, para $T = [T_1, T_2]$ y una estimación $\widehat{\beta}(t)$ cualquiera de $\beta(t)$

$$ECMBI\left(\widehat{\beta}(t)\right) = \frac{1}{T_2 - T_1}\int_T \left(\beta(t) - \widehat{\beta}(t)\right)^2 dt$$

que en términos de la base de funciones queda

$$\begin{aligned} ECI\left(\widehat{\beta}(t)\right) &= \frac{1}{T_2 - T_1}\int_T \left(\sum_{j=1}^{p}\beta_j\phi_j(t) - \sum_{j=1}^{p}\widehat{\beta}_j\phi_j(t)\right)^2 dt \\ &= \frac{1}{T_2 - T_1}\int_T \left(\sum_{j=1}^{p}\left(\beta_j - \widehat{\beta}_j\right)\phi_j(t)\right)^2 dt \\ &= \frac{1}{T_2 - T_1}\left(\beta - \widehat{\beta}\right)^T \Psi\left(\beta - \widehat{\beta}\right) \end{aligned}$$

siendo $\beta = (\beta_1, \ldots, \beta_p)^T$ y $\widehat{\beta} = \left(\widehat{\beta}_1, \ldots, \widehat{\beta}_p\right)^T$ los coeficientes de la expansión de las funciones parámetros (real y estimada, independientemente del método y número de cc. pp. utilizado para ello respectivamente) en términos de los elementos de la base de funciones.

Capítulo 3

Ejemplos simulados

En este capítulo se presentan distintos estudios con datos simulados que permitirán estudiar la precisión de los estimadores obtenidos mediante los modelos en componentes principales propuestos en los capítulos 1 (caso múltiple) y 2 (caso funcional). Uno de los objetivos fundamentales de este capítulo es diseñar criterios de selección de las cc. pp. que lleven a una estimación precisa de los parámetros reduciendo al máximo el número de componentes.

3.1 Simulación del caso múltiple

Para ilustrar el modelo de regresión logística múltiple en componentes principales, hemos desarrollado un estudio de simulación en el que pretendemos poner de manifiesto las ventajas que con dicho modelo se obtienen, en presencia de multicolinealidad en las variables explicativas, al estimar los parámetros del modelo y al reducir la dimensión del problema. En dicho estudio hemos considerado una serie de modelos en los que, a partir de una cierta estructura de dependencia conocida de las variables explicativas, se consigue dar una estimación de los parámetros que se aproxima bien a los reales. El método de simulación es similar en todos los casos, si bien se ha variado en cada uno de ellos la estructura de dependencia y los parámetros fijados. Antes de comenzar la explicación de nuestra simulación destacaremos la importancia de la elección adecuada de las variables explicativas y de los parámetros que darán lugar a las probabilidades reales que determinan la distribución binomial de la variable respuesta, para que, computacionalmente, se obtengan estimaciones estables, fruto de la ejecución de los métodos iterativos necesarios para la obtención de tales estimaciones.

Siguiendo los esquemas de simulación de Pulkstenis y Robinson (2002) y de Hosmer et al. (1997) en los que se analizan diversos estadísticos de bondad de ajuste en el modelo de regresión logística, se han seleccionado una serie de modelos de regresión logística y se ha repetido un número considerable de veces la simulación de cada uno de ellos. En cada modelo y en cada repetición del mismo se obtendrán diversas medidas de la bondad de la estimación de los parámetros y las probabilidades. A continuación se exponen detalladamente los pasos seguidos en la repetición de cada uno de los modelos, así como los resultados obtenidos referentes a parámetros estimados, probabilidades, medidas de bondad de ajuste, etc. Posteriormente se analizarán las medidas globales de todas las repeticiones de cada uno de los modelos elegidos.

3.1.1 El proceso de simulación

El primer paso de la simulación consiste en obtener una muestra de las variables explicativas y que presenten una estructura de dependencia conocida. Para ello elegimos inicialmente un número de p variables y simulamos n valores de manera independiente de cada una de las p variables utilizando la distribución Normal estándar en el primer ejemplo y la chi-cuadrado en el segundo. Una vez simuladas estas cantidades, que resumiremos en una matriz N de dimensión $n \times p$, buscamos otra A, de dimensión $p \times p$, que proporcione la estructura de dependencia en las covariables del modelo de regresión logística mediante una transformación lineal de la anterior de la forma NA, y que también se obtiene mediante simulación. Si llamamos $X = (\mathbf{1} \mid NA)$ a la matriz de diseño del modelo, siendo $\mathbf{1}$ un vector $n \times 1$ de unos, tendremos simulados los valores de las variables explicativas. En cada uno de los casos que aquí se muestran se han modificado tanto los tamaños de muestra n como el número de variables p; así como las distribuciones para la obtención de las variables independientes de N y de la matriz de la transformación lineal A.

Siguiendo con la obtención de los datos que nos llevan al ajuste del modelo de regresión logística, fijamos los parámetros del modelo $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ obteniéndolos también a través de la simulación de valores de alguna distribución adecuada. Tales parámetros proporcionan las combinaciones lineales $X\beta$ que nos llevan a calcular las probabilidades reales del modelo de regresión logística

$$\pi_i = \frac{\exp\{x_i^T \beta\}}{1 + \exp\{x_i^T \beta\}}, \; i = 1, \ldots, n,$$

siendo x_i^T la i-ésima fila de la matriz X. En este punto cabe señalar la importancia de la elección adecuada tanto de los valores de N como de A y β, para que las exponenciales que llevan a las π_i sean valores estables, esto es, no demasiado grandes.

Finalmente, y como último paso en la obtención de los datos, simulamos los valores de la respuesta dicotómica a partir de una distribución de Bernouilli con el parámetro dado, en cada uno de los n casos, por la probabilidad correspondiente

$$y_i \sim B(\pi_i), \ i = 1, \ldots, n.$$

Una vez simulados los datos pasamos al ajuste del modelo de regresión logística con estas variables, resultando en general, y como veremos posteriormente, parámetros estimados $\widehat{\beta}$ muy alejados de los reales β, causa que, como ya han apuntado Hosmer y Lemeshow (1989) en regresión logística o Ryan (1997) en regresión lineal y logística, se puede deber a la gran dependencia existente en las variables explicativas de la matriz X (multicolinealidad). Para eludir este problema de estimación, debido fundamentalmente a la necesidad de invertir matrices mal condicionadas para la obtención de los parámetros estimados, decidimos utilizar como regresores, en lugar de las variables originales, las componentes principales de las variables originales, $Z = XV$. De este modo, y debido a que las componentes son ortogonales, se evitan los problemas derivados de la multicolinealidad.

Pero la decisión de utilizar las componentes como regresores va más allá de evitar problemas de multicolinealidad, con ello también es posible reducir la dimensión del problema de regresión logística utilizando, en lugar de todas, un número reducido de componentes que expliquen un determinado porcentaje de variabilidad.

Después de estimar los parámetros del modelo en términos de las s primeras componentes principales $\widehat{\gamma}_{(s)}$, la reconstrucción de los originales será

$$\widehat{\beta}_{(s)} = V_{(s)}\widehat{\gamma}_{(s)},$$

siendo s el número de componentes principales seleccionadas y $V_{(s)}$ la submatriz que tiene las s primeras columnas de V.

Otro de los grandes problemas a tratar en este campo es la elección de un criterio de parada al introducir componentes principales. Así, Aucot et al. (2000) consideran tres criterios: (1) introducir todas las componentes que indique el método Stepwise con los test condicionales de razón de verosimilitudes, (2) elegir aquel número de componentes que minimice el error cuadrático

medio de la predicción, (3) o bien ir introduciendo componentes hasta que el error anteriormente visto aumente sin importar lo que ocurra después.

En nuestro caso el problema más grave que presenta la multicolinealidad de las variables originales es la deficiente estimación de los parámetros del modelo, por lo tanto decidimos elegir una medida de la precisión de los parámetros estimados o reconstruidos con respecto a los reales; y tomamos como tal medida el error cuadrático medio para los parámetros estimados

$$ECMB = \frac{1}{p+1}\sum_{j=0}^{p}\left(\widehat{\beta}_{j(s)} - \beta_j\right)^2 \tag{3.1}$$

siendo $\widehat{\beta}_{(s)}$ el vector de parámetros estimados (reconstruidos) a partir de los obtenidos con el modelo con s componentes principales; con lo que el consiguiente criterio de parada consistirá en elegir el número de componentes (bien en orden de variabilidad o bien en base al procedimiento stepwise) que minimice esta cantidad. Como se pondrá de manifiesto en los ejemplos simulados, generalmente ocurre que el error ECMB va disminuyendo hasta alcanzar un mínimo y después comienza a crecer, produciéndose como consecuencia un empeoramiento en la estimación de los parámetros β.

No obstante, en una aplicación con datos reales, no se conocen los parámetros reales, con lo que se debería elegir algún otro criterio que no tuviese en cuenta dichos parámetros. En los distintos ejemplos de simulación que se presentan, se ha observado que generalmente el mínimo en el ECMB va asociado a un aumento considerable en el valor de la varianza de los parámetros estimados, definida como la suma de los elementos diagonales de la matriz de covarianzas de los parámetros

$$Var\left[\widehat{\beta}_{(s)}\right] = V_{(s)}^T\left(Z_{(s)}^T W_{(s)} Z_{(s)}\right)^{-1} V_{(s)}. \tag{3.2}$$

Además, en los casos en que no se produce dicho salto en la varianza, es usual que el mínimo del error ECMB se alcance en la última componente que entra en el modelo. Es por todo esto por lo que aquel número de componentes anteriores a un crecimiento significativo en la varianza de los parámetros, también es un criterio que se revela como adecuado en las aplicaciones con datos reales.

Finalmente un aspecto importante a tener en cuenta cuando hemos de decidir el número de componentes principales a introducir en un modelo de regresión, es el orden en que éstas se incluyen en el mismo ya que se ha demostrado (ver Jackson, 1991) que las componentes más explicativas no son necesariamente las que mejor predicen la variable respuesta. En este sentido, y

a efectos de comparación, consideraremos para nuestro caso de regresión logística dos formas de introducción de componentes: el que llamaremos Método I que consiste en la introducción de las componentes en orden de variabilidad, y el Método II en el que la inclusión se realiza en el orden que proporciona el método Stepwise basado en los contrastes condicionales de razón de verosimilitudes.

3.1.2 Ejemplo 1

A continuación se presentan los resultados de un ejemplo como el expuesto anteriormente, en el que se considera una muestra de $n = 100$ valores de $p = 10$ variables explicativas. En este caso las variables de la matriz N provienen de una distribución Normal con media 0 y varianza 1, y la matriz de la transformación lineal se ha fijado con valores uniformes en el intervalo $[0, 1]$. Para comprobar que efectivamente existe gran dependencia entre las variables explicativas así simuladas, calculamos la matriz de correlaciones de las 10 variables explicativas contenidas en la matriz NA

$$Corr\,(NA) = \begin{pmatrix} 1 & & & & & & & & & \\ 0.82 & 1 & & & & & & & & \\ 0.89 & 0.60 & 1 & & & & & & & \\ 0.87 & 0.69 & 0.93 & 1 & & & & & & \\ 0.70 & 0.58 & 0.80 & 0.73 & 1 & & & & & \\ 0.77 & 0.56 & 0.90 & 0.79 & 0.84 & 1 & & & & \\ 0.87 & 0.59 & 0.96 & 0.93 & 0.81 & 0.88 & 1 & & & \\ 0.89 & 0.60 & 0.82 & 0.76 & 0.50 & 0.73 & 0.79 & 1 & & \\ 0.92 & 0.63 & 0.89 & 0.83 & 0.65 & 0.76 & 0.85 & 0.83 & 1 & \\ 0.79 & 0.63 & 0.89 & 0.81 & 0.89 & 0.87 & 0.86 & 0.74 & 0.73 & 1 \end{pmatrix}$$

en la que se puede apreciar la gran correlación existente en la mayoría de los casos en los que se supera el 0.8 de correlación siendo en casi todos superior a 0.5. Especial mención requiere la tercera variable que tiene gran dependencia con el resto.

Los parámetros considerados, que llamaremos reales, aparecen en la Tabla 3.1 y proporcionan, a través de las combinaciones lineales con las filas de la matriz de diseño X simulada, las probabilidades reales π_i. A partir de estas probabilidades se obtienen, como explicamos anteriormente, los valores de la variable respuesta binaria proporcionando en este caso valores en los que el porcentaje de unos es del 39%.

Después de obtener los datos ajustamos el modelo de regresión logística múltiple con las variables originales

$$y_i = \pi_i + \varepsilon_i,\ i = 1, \ldots, 100$$

obteniendo como parámetros estimados los de la Tabla 3.1, que difieren mucho de los originales como apuntábamos en la presentación del proceso de simulación. Así, encontramos grandes diferencias en los parámetros 1, 5, 8 y 10 siendo en algunos casos incluso diferentes en el signo como en el caso de los parámetros 2, 3, 4, 6, 7, 8 y 10 con el consiguiente perjuicio para la interpretación de tales parámetros como cocientes de ventajas. A pesar de ello, y como se observa en la Tabla 3.3 (última fila) la tasa de clasificaciones correctas es del 87%. Las otras medidas de bondad de ajuste como son el R^2 y R^2 ajustado son, respectivamente, 0.53 y 0.72, mientras que el estadístico G^2 del contraste de razón de verosimilitudes es 58.44 con un p-valor de 0.99. Todas estas medidas indican que globalmente el modelo se ajusta bien, salvo por el hecho de que los parámetros se estiman bastante deficientemente. Esto significa que el modelo es adecuado para hacer predicciones pero no para interpretar sus parámetros.

Con el objetivo de intentar estimar los parámetros mejor y reducir de algún modo la dimensión del problema, calculamos las componentes principales para posteriormente tomar como variables explicativas tales componentes. Así, la varianza explicada por tales componentes se expresa en la Tabla 3.2 que pone de manifiesto que con las tres primeras componentes se consigue explicar más del 93% del total de variabilidad, y con 6 casi alcanzamos el 99%.

Una vez obtenidas las componentes principales ya estamos en condiciones de utilizarlas como covariables de nuestro modelo de regresión logística. Con el objetivo de reducir la dimensión de nuestro problema y de proporcionar una estimación de los parámetros lo más próxima posible a los reales, hemos ido ajustando los modelos de regresión logística con s componentes principales en cada paso,

$$y_i = \pi_{i(s)} + \varepsilon_{i(s)};\ i = 1, \ldots, 100;\ s = 1, 2, \ldots, 10$$

donde

$$\pi_{i(s)} = \frac{\exp\left\{z_{i(s)}^T \gamma_{(s)}\right\}}{1 + \exp\left\{z_{i(s)}^T \gamma_{(s)}\right\}},$$

$z_{i(s)}^T$ es la i-ésima fila de la matriz de diseño de componentes principales formada por las correspondientes $s+1$ columnas de Z, y $\gamma_{(s)}$ el vector de $s+1$ parámetros; y hemos reconstruido los parámetros originales en la forma

$$\widehat{\beta}_{(s)} = V_{(s)}\widehat{\gamma}_{(s)}$$

lo que proporciona una estimación de los mismos en cada paso ($V_{(s)}$ es la matriz con s columnas de la matriz de vectores propios y la primera columna y fila con los valores $(1, 0, \ldots, 0)$).

En las Tablas 3.3 y 3.4 se muestran distintas medidas de bondad de ajuste para los dos Métodos considerados de introducción de componentes principales en el modelo (I y II). Las medidas presentadas en dichas tablas son las siguientes:

- ECMP hace referencia al error cuadrático medio de las probabilidades estimadas respecto de las reales, esto es,

$$ECMP = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\pi}_{i(s)} - \pi_i\right)^2$$

  siendo $\widehat{\pi}_{i(s)}$ la i-ésima probabilidad estimada por el modelo con s componentes principales. Observando tales cantidades vemos que, a medida que vamos introduciendo componentes principales, este error va disminuyendo quizás hasta que se introduce la última componente en orden de variabilidad en que vuelve a crecer de 0.01 a 0.015. Por otro lado observando la misma cantidad con el Método II se aprecia cómo este error va disminuyendo continuamente. Estos valores indicarían que si tomásemos el criterio propuesto por Aucot et al. (2000) deberíamos seleccionar las nueve primeras componentes principales con el método I y las cuatro que entran con el II, siendo muy relevante en este caso la reducción de dimensión que se consigue con el método II.

- CCR representa la tasa de clasificaciones correctas tomando como punto de corte 0.5. Se aprecia que, salvo levísimas desviaciones, a medida que se introducen componentes las tasas de clasificaciones correctas crecen pasando del 66 al 87%. Comparando las cantidades de esta medida, se puede ver que los dos métodos proporcionan casi iguales tasas de clasificaciones correctas, sólo hay ligeras diferencias y no claramente a favor de uno de los métodos. En el método II se alcanzan tasas de clasificaciones correctas parecidas a las del Método I pero con un número menor de cc. pp.

- ECMB es el error cuadrático medio definido anteriormente en (3.1) y se observa que en el Método I se alcanza el mínimo cuando entra la séptima componente principal con un valor de 0.52. No obstante ese mínimo es muy próximo a los ECMB asociados a $s = 3, 4, 5, 6, 7, 8$ lo que podría indicar que todos ellos podrían ser buenas elecciones como se aprecia en la Tabla 3.5 que presenta los valores de los parámetros reconstruidos. Lo que sí está claro es que al introducir la última c.p. este valor se dispara (40.70) llevando a una estimación muy desviada de los parámetros. Habría que recordar que los parámetros obtenidos con todas las componentes principales son los mismos que los que se obtienen con las variables originales, lo que vuelve a poner de manifiesto la mejora que se consigue al utilizar componentes principales y quedarnos con un número inferior al total. Observando por otro lado los valores de esta medida para el método II se aprecia que el mínimo, 0.487, se obtiene al entrar la componente nueve en el modelo (tercera en entrar según el Método II), lo que vuelve a indicar la reducción de dimensión que se consigue siendo mayor con el método II que con el I. Observemos que al entrar la c.p. 3 el ECMB incrementa empeorando ligeramente la estimación de los parámetros.

- Otra medida que hemos considerado como indicativa de la proximidad entre valores estimados y reales de los parámetros es el máximo de las diferencias en valor absoluto entre tales parámetros

$$Max_j \left| \widehat{\beta}_{j(s)} - \beta_j \right|, \; j = 0, \ldots .p.$$

  En las distintas simulaciones llevadas a cabo se ha observado que generalmente el mínimo valor de estos máximos suele coincidir con el número de componentes en donde se alcanza el valor más pequeño de ECMB tanto en el método I como en el II; si bien cuando existe alguna diferencia, como es este caso, no suele haber gran variación en el número de componentes principales (± 1) ni en el valor de este máximo.

- La siguiente medida que se presenta es la varianza de los parámetros estimados definida en (3.2) y que se utilizaría en los casos reales como medida indicativa del número de componentes a elegir. Podemos ver que tanto con el método I como con el II dicha varianza va aumentando a medida que introducimos componentes en el modelo pasando de 0.06 a 158.95 en el primer caso y de 0.09 a 2.16 en el segundo. Según se ha

indicado en secciones previas, un aumento demasiado grande en dicha varianza, al introducir una variable, suele ser indicativo de que a partir de ahí no se deberían escoger más componentes como ocurre en el método I, en el que pasaríamos de 4.67 al introducir la novena componente a 158.95 cuando introducimos la última. Si observamos los valores del ECMB en la séptima componente, 0.52, y en la novena, 0.60, éstos son muy parecidos con lo que, con el criterio de la varianza, la estimación de los parámetros sería muy parecida a la obtenida con el de mínimo ECMB, al menos como medida global. Cabe decir que con el método II no se observa una diferencia tan grande como en el I con lo que seleccionaríamos todas las componentes que entran según el método Stepwise.

Finalmente notar que en las simulaciones desarrolladas se ha observado que cuando la estimación de los parámetros con las variables originales no es muy exacta, éste hecho lo recoge bien el criterio de la varianza en el método I con una diferencia considerable en el valor de la varianza con respecto al resto, al introducir la última componente, y con el método II o bien se aprecia dicha diferencia y coincide con el punto de mínimo ECMB, o muy próximo a él, o bien no se aprecia y deberíamos quedarnos con todas las componentes que entran en el modelo según el método II coincidiendo también con el criterio de mínimo ECMB.

- Las siguientes columnas indican el valor del estadístico G^2 de razón de verosimilitudes del modelo correspondiente para contrastar lo adecuado del modelo logístico y su correspondiente p-valor que indica en todos los casos que el modelo logístico es acertado con p-valores próximos a uno a partir de la segunda componente principal tanto en el método I como en el II.

- Las últimas columnas calculadas para cada modelo son los coeficientes de determinación común y ajustado que se utilizan también como medida de bondad de ajuste. Como se aprecia ambos coeficientes aumentan al ir introduciendo componentes principales, como no podía ser de otro modo, y en ambos métodos; siendo mayores los valores del ajustado que del común.

Para terminar con la exposición de este ejemplo simulado en las Tablas 3.5 y 3.6 aparecen los valores de los parámetros reconstruidos por cada modelo y con cada método, junto a los valores reales, para hacer comparaciones y

poner de manifiesto que el número de componentes principales elegido por los distintos criterios es adecuado. En estas tablas se aprecia lo próxima que está la elección hecha a través del criterio del mínimo ECMB con respecto a los parámetros reales y la elección hecha con el criterio del salto en la varianza.

Al igual que en el caso anterior se aprecia que de las cuatro componentes que entran en el método II, cuando entra la novena ($s = 3$), que va asociada al mínimo ECMB, es cuando se obtienen mejores estimaciones y que éstas parecen sensiblemente mejores que con el método I.

Repetición de la situación

El segundo paso del proceso de simulación consiste en repetir la simulación del modelo anterior un número considerable de veces, de hecho nosotros hemos considerado 200 repeticiones de este caso. Para ello se fijan los valores que proporcionan la estructura de dependencia de las variables explicativas, contenidos en la matriz A, y de los parámetros que definen el modelo, β, y se simulan en cada repetición unos nuevos valores de N y por tanto de X e Y, así como de las probabilidades reales; obteniendo en todas y cada una de dichas repeticiones los mismos estadísticos y medidas vistas en la sección anterior.

Para analizar lo adecuado del modelo de regresión logística en componentes principales múltiple debemos considerar una serie de medidas resumen de tales estadísticos eligiendo el criterio más adecuado. Así, considerando en cada repetición y con cada método de selección de componentes, el modelo con un número de componentes principales tal que minimiza el ECMB, se calculan la media, varianza, coeficiente de variación, mediana y cuartiles de los estadísticos utilizados, observando valores que se resumen en las Tablas 3.7 y 3.8. Veamos detalladamente algunas de las cantidades que aparecen en dichas tablas y las conclusiones que podemos obtener de ellas.

- En cuanto al número de componentes que entran al considerar como modelo óptimo el de menor ECMB, se obtiene que en el método II tanto la media como la varianza de tal número son mucho menores que en el Método I, lo que demuestra la diferencia en reducción de dimensión que se obtiene con el Método II respecto al I. Además observando los coeficientes de variación se tiene que son prácticamente iguales lo que indica que tales medias son igualmente adecuadas en ambos métodos. Además el resto de medidas que son la mediana y los cuartiles también indican gran reducción de dimensión en el Método II y menor en el I.

- Cabe destacar el alto porcentaje de variabilidad que explican las componentes principales en las distintas repeticiones, de modo que utilizando el punto de corte considerado se tiene un número de componentes en el método I que supera el 98% en media de variabilidad explicada y que en la mayoría de las repeticiones se supera el 99%. Sin embargo con el Método II varía mucho de una repetición a otra dependiendo de las componentes que entran en el modelo. En el ejemplo considerado anteriormente las componentes que entran con el Método II son la 1, 2 y 9 siendo ésta última una de las que menos varaibilidad esplica. Esto pone de manifiesto que no son necesariamente las cc. pp. más explicativas las que mejor explican a la variable respuesta.

- En cuanto a las tasas de clasificaciones correctas se observa que en media ambos métodos proporcionan valores similares en torno al 84% siendo ligeramente superiores en el Método I así como en términos de los cuartiles.

- Observando los valores de la media de los mínimos ECMB se aprecia que tales valores son parecidos (0.40 y 0.46) siendo las varianzas y coeficientes de variación muy similares. La mediana y los cuartiles también indican tales similitudes y de igual magnitud, 0.03 aproximadamente.

- Con los Máximos correspondientes ocurre exactamente igual que con los ECMB, sin embargo la suma de las varianzas de los parámetros estimados de los modelos elegidos son considerablemente más pequeñas con el Método II (0.94) que con el I (4.62), como ilustran la media, mediana y cuartiles. Es más, la varianza de esas cantidades es mucho menor también en el método II que en el I lo que indica que tales medias son más representativas en el método II que en el I.

- Finalmente el resto de medidas de bondad de ajuste como son el estadístico G^2, y los coeficientes de determinación son muy similares en el Método I y en el II con ligeras diferencias a favor del Método I.

Parámetros	Reales	Estimados
β_0	-0.67	-0.81
β_1	-0.95	-14.84
β_2	-0.95	4.96
β_3	-0.97	4.35
β_4	1.40	-0.89
β_5	1.12	8.66
β_6	0.61	-2.70
β_7	-0.24	0.79
β_8	-0.71	7.61
β_9	1.21	4.66
β_{10}	0.93	-5.48

Tabla 3.1: Caso múltiple. Ejemplo 1. Parámetros simulados y estimados con el modelo de regresión logística múltiple en términos de las varaibles originales

Componente	Varianza	%Varianza aumulada
1	29.99	83.03
2	2.23	89.20
3	1.63	**93.70**
4	0.88	96.15
5	0.51	97.57
6	0.45	**98.81**
7	0.25	99.50
8	0.11	99.79
9	0.08	99.99
10	0.001	100.00

Tabla 3.2: Caso múltiple. Ejemplo 1. Variabilidad explicada por las distintas componentes principales y porcentaje acumulado de variabilidad explicada para cada una.

s	ECMP	CCR	ECMB	Max	Var	G^2	p-val	R^2	R^2(Aj)
1	0.076	**66**	0.776	1.34	**0.060**	112.35	0.153	0.19	0.26
2	0.025	78	0.628	1.37	0.149	84.77	0.808	0.39	0.53
3	0.017	79	0.594	1.41	0.214	82.05	0.844	0.40	0.55
4	0.012	80	0.567	1.42	0.339	79.84	0.868	0.42	0.57
5	0.013	78	0.567	1.28	0.490	79.42	0.859	0.42	0.57
6	0.012	78	0.563	**1.26**	0.687	79.40	0.842	0.42	0.57
7	0.010	80	**0.519**	1.47	1.003	78.51	0.841	0.42	0.58
8	0.010	82	0.541	1.32	2.098	74.28	0.899	0.45	0.61
9	**0.010**	86	**0.603**	1.56	**4.672**	61.43	0.991	0.51	0.70
10	0.015	**87**	**40.70**	13.89	**158.949**	**58.44**	**0.995**	**0.53**	**0.72**

Tabla 3.3: Caso múltiple. Ejemplo 1. Medidas de bondad de ajuste para el ejemplo simulado, con la introducción de las cc. pp. en el modelo según el Método I: orden de variabilidad

s	c.p	ECMP	CCR	ECMB	Max	Var	G^2	p-val	R^2	R^2(Aj)
1	2	0.076	72	0.658	1.45	**0.09**	111.45	0.17	0.20	0.27
2	1	0.025	78	0.628	1.37	0.15	84.77	0.81	0.39	0.53
3	9	0.025	85	**0.487**	1.56	1.75	72.74	0.96	0.46	0.62
4	3	0.016	87	0.659	1.48	**2.16**	68.19	0.98	0.48	0.65

Tabla 3.4: Caso múltiple. Ejemplo 1. Medidas de bondad de ajuste para el ejemplo simulado, con la introducción de las cc. pp. en el modelo según el Método II: orden que propociona el método Stepwise

s	1	2	3	4	5	6	**7**	8	9	Real
$\widehat{\beta}_{0(s)}$	-0.43	-0.51	-0.49	-0.43	-0.42	-0.42	**-0.40**	-0.40	-0.66	-0.67
$\widehat{\beta}_{1(s)}$	0.07	-0.32	-0.34	-0.38	-0.31	-0.30	**-0.40**	-1.24	-1.91	-0.95
$\widehat{\beta}_{2(s)}$	0.05	-0.53	-0.32	-0.31	-0.30	-0.32	**-0.31**	0.01	-0.71	-0.51
$\widehat{\beta}_{3(s)}$	0.09	0.26	0.20	0.24	0.25	0.25	**0.49**	0.01	-2.53	-0.97
$\widehat{\beta}_{4(s)}$	0.07	0.03	0.01	0.31	0.21	0.20	**0.33**	0.09	2.32	1.40
$\widehat{\beta}_{5(s)}$	0.05	0.44	0.64	0.65	0.71	0.73	**0.62**	-0.08	1.63	1.12
$\widehat{\beta}_{6(s)}$	0.07	0.43	0.47	0.32	0.41	0.37	**0.41**	0.53	1.21	0.61
$\widehat{\beta}_{7(s)}$	0.09	0.31	0.26	0.42	0.40	0.39	**0.05**	0.78	-0.38	-0.24
$\widehat{\beta}_{8(s)}$	0.06	-0.24	-0.39	-0.64	-0.71	-0.72	**-0.83**	-1.25	-0.07	-0.71
$\widehat{\beta}_{9(s)}$	0.05	-0.11	-0.20	-0.21	-0.08	-0.05	**0.09**	1.04	2.38	1.21
$\widehat{\beta}_{10(s)}$	0.07	0.36	0.48	0.31	0.19	0.21	**0.33**	1.05	0.90	0.93

Tabla 3.5: Caso múltiple. Ejemplo 1. Parámetros reconstruidos por el modelo de regresión logística considerado, con los distintos números de cc. pp. en el modelo. Introducción de las cc. pp. en el modelo según el Método I: orden de variabilidad.

s	1	2	3	4	Real
$\widehat{\beta}_{0(s)}$	–0.53	–0.51	**-0.69**	-0.72	-0.67
$\widehat{\beta}_{1(s)}$	-0.32	-0.32	**-0.96**	-1.09	-0.95
$\widehat{\beta}_{2(s)}$	-0.45	-0.53	**-1.24**	-1.06	-0.51
$\widehat{\beta}_{3(s)}$	0.11	0.26	**-1.96**	-2.37	-0.97
$\widehat{\beta}_{4(s)}$	-0.05	0.03	**1.94**	2.18	1.40
$\widehat{\beta}_{5(s)}$	0.29	0.44	**1.80**	2.29	1.12
$\widehat{\beta}_{6(s)}$	0.26	0.43	**0.98**	1.12	0.61
$\widehat{\beta}_{7(s)}$	0.15	0.31	**-0.60**	-0.80	-0.24
$\widehat{\beta}_{8(s)}$	–0.24	-0.24	**0.85**	0.77	-0.71
$\widehat{\beta}_{9(s)}$	-0.14	-0.11	**1.13**	1.18	1.21
$\widehat{\beta}_{10(s)}$	0.21	0.36	**0.29**	0.46	0.93

Tabla 3.6: Caso múltiple. Ejemplo 1. Parámetros reconstruidos por el modelo de regresión logística considerado, con los distintos números de cc. pp. en el modelo. Introducción de las cc. pp. en el modelo según el Método II: orden proporcionado por el método Stepwise

Estadísticos	Media	Varianza	CV	Me	Q_1	Q_3
N°Comp.	7.51	5.57	0.31	9.00	6.00	9.00
Var. Acum.	**98.39**	10.85	0.03	99.996	98.94	99.996
ECMP	1.69	0.45	0.40	1.59	1.22	2.01
CCR	85.25	10.85	0.04	85.00	83.00	87.00
ECMB	**0.40**	0.03	0.45	0.38	0.24	0.57
Max	1.21	0.12	0.28	1.26	0.92	1.48
Var	**4.62**	160.37	2.74	3.62	0.89	4.69
G^2	65.11	105.36	0.16	63.90	57.96	72.67
R^2	0.50	0.003	0.10	0.51	0.47	0.54
R^2(Aj)	0.68	0.005	0.10	0.69	0.63	0.73

Tabla 3.7: Caso múltiple. Ejemplo 1. Medidas promedio de bondad de ajuste para las 200 repeticiones del ejemplo simulado, con la introducción de las cc. pp. en el modelo según el Método I: orden de variabilidad.

Estadísticos	Media	Varianza	CV	Me	Q_1	Q_3
N°Comp.	3.12	1.00	0.32	3.00	2.00	4.00
ECMP	2.23	2.36	0.69	1.85	1.35	2.48
CCR	83.41	12.10	0.04	84.00	81.00	86.00
ECMB	**0.46**	0.04	0.43	0.56	0.26	0.61
Max	1.26	0.11	0.26	1.39	0.99	1.51
Var	**0.94**	0.73	0.90	0.40	0.23	1.62
G^2	72.33	106.41	0.14	72.12	64.92	78.93
R^2	0.47	0.003	0.12	0.47	0.44	0.50
R^2(Aj)	0.63	0.006	0.12	0.63	0.58	0.68

Tabla 3.8: Caso múltiple Ejemplo 1. Medidas promedio de bondad de ajuste para las 200 repeticiones del ejemplo simulado, con la introducción de las cc. pp. en el modelo según el Método II: orden proporcionado por el método Stepwise.

3.1.3 Ejemplo 2

Con la intención de corrobarar lo visto en el ejemplo anterior, hemos desarrollado otro trabajo de simulación en el que sólo presentaremos las conclusiones obtenidas de las numerosas repeticiones del mismo y no así del ejemplo particular.

La simulación de los valores de las variables explicativas se ha obtenido mediante la ya comentada transformación lineal de valores independientes, en este caso de la distribución chi-cuadrado con un grado de libertad, considerando $n = 200$ valores de $p = 12$ variables con dicha distribución. La matriz A de la transformación es una matriz triangular superior de unos y los parámetros fijados para la simulación de las probabilidades del modelo son

β_0	β_1	β_2	β_3	β_4	β_5	
0.41	-1.23	-0.22	-0.50	-1.83	1.79	
β_6	β_7	β_8	β_9	β_{10}	β_{11}	β_{12}
-0.53	0.85	-0.68	1.46	-1.52	0.10	1.01

Una vez simulados los valores de las variables independientes y los parámetros del modelo, la obtención de las probabilidades y de los valores dicotómicos de la variable respuesta se obtienen como en el ejemplo anterior. Después de la simulación de todos los valores necesarios, ajustamos el modelo, estimamos los parámetros y obtenemos las medidas de bondad de ajuste y de la precisión de los parámetros estimados ya conocidos en términos de las cc. pp.

Repitiendo la simulación y los ajustes antes mencionadas, en este caso 300 veces, obtenemos las medidas resumen de esta repetición ya definidas y que se muestran en las Tablas 3.9 y 3.10. De la observación de las mismas se llega a las siguentes conclusiones.

- En cuanto a la reducción de dimensión que se consigue, vuelve a ser notable si comparamos los Método I y II ya que en el primero el número de componentes promedio de los modelos óptimos es casi 8 y en el segundo sólo 4. Además, estas dos medidas son igualemte representativas a tenor de lo indicado por el coeficiente de variación que presenta en los dos métodos valores cercanos a 0.35.

- Observando los valores promedio de la tasa de clasificaciones correctas se aprecia que los modelos predicen de manera adecuada llegando a ser las tasas del 96.07% en el Método I (con 8 componentes de promedio) y de 95.77 (con 4) en el II. Además si examinamos las medidas de variabilidad

Estadísticos	Media	Varianza	CV	Me	Q_1	Q_3
N. cc. pp.	7.82	7.28	0.35	8.00	5.00	10.00
Var. Acum	**98.22**	2.76	0.02	98.71	97.12	99.49
ECMP	0.01	1.23E-5	0.44	0.01	0.01	0.01
CCR	**96.07**	1.62	0.01	96.00	95.00	97.00
ECMB	**0.85**	0.07	0.32	0.90	0.68	1.02
Max	**1.75**	0.13	0.21	1.76	1.57	1.95
Var	8.43	69.01	0.99	5.65	2.61	11.53
G^2	42.68	122.50	0.26	42.11	34.61	49.57
R^2	0.22	2.94E-3	0.25	0.22	0.18	0.26
R^2(Aj)	0.59	0.01	0.18	0.59	0.51	0.67

Tabla 3.9: Caso múltiple. Ejemplo 2. Medidas de bondad de ajuste para las 300 repeticiones del ejemplo simulado, con la introducción de las cc. pp. en el modelo según el Método I: Orden de variabillidad

calculadas para estos promedios vemos que son bastante bajas con un coeficiente de variación de 0.01 para ambos casos.

- De las medidas de la precisión de las estimaciones podemos concluir que suelen ser ligeramente mejores las obtenidas por el Método I como por ejemplo para el ECMB que obtiene un promedio de 0.85 en el primer caso y 1.02 en el segundo o para el Max con valores de 1.75 y 2.02 respectivamente. Sin embargo esta ganancia en precisión del Método I necesita de demasiadas componentes (el doble). Cabe decir que de las variabilidades de los promedios aquí indicados se concluye que lo aquí observado es tónica general en cada una de las repeticiones ya que dichas variabilidades son pequeñas y similares.

- Finalmente las medidas de bondad de ajuste en este caso vuelven a no indicar una preferencia por uno u otro Método, ya que son muy parecidas en media y variabilidad como se desprende de R^2 con valores promedio en ambos casos de 0.2 aproximadamente.

Estadísticos	Media	Varianza	CV	Me	Q_1	Q_3
N. cc. pp	3.92	2.25	0.38	4.00	3.00	5.00
ECMP	0.01	2.50E-5	0.47	0.00	0.01	0.01
CCR	**95.77**	1.91	0.01	96.00	95.00	96.63
ECMB	**1.02**	0.13	0.36	1.04	0.78	1.24
Max	**2.02**	0.28	0.26	1.98	1.70	2.29
Var	3.43	7.76	0.81	2.79	1.38	4.80
G^2	47.03	167.29	0.28	44.92	37.21	56.11
R^2	0.20	3.76E-3	0.31	0.20	0.16	0.24
R^2(Aj)	0.54	0.02	0.25	0.56	0.47	0.63

Tabla 3.10: Caso múltiple. Ejemplo 2. Medidas de bondad de ajuste para las 300 repeticiones del ejemplo simulado, con la introducción de las cc. pp. en el modelo según el Método II: Orden Stepwise

Conclusiones

De este estudio de simulación podemos concluir que en aquellas medidas de bondad de ajuste en las que hay alguna diferencia a favor del Método I, tales diferencias son muy pequeñas y no significativas con lo que podríamos concluir que la bondad del ajuste del modelo seleccionado con los dos Métodos es similar. Pero si nos centramos en la reducción de dimensión del problema y estimación de los parámetros, el Método II es claramente superior al I como indica el número de cc.pp y la varianza de los parámetros estimados y además a gran distancia con lo que sería éste el método más adecuado de introducción de componentes principales y la utilización de éstas una buena elección para el ajuste de un modelo de regresión logística con variables explicativas continuas y con gran presencia de multicolinealidad.

3.2 Simulación del caso funcional

En el capítulo dos de la presente memoria introdujimos de manera teórica el modelo de regresión logística funcional y los problemas que podrían surgir en la práctica en su estimación, y propusimos una serie de soluciones a los mismos. En el presente pretendemos poner de manifiesto, con datos simulados, toda esta problemática de estimación, así como lo adecuado de las distintas soluciones aportadas.

Como se indicó en el capítulo teórico, al trabajar con datos funcionales el primer problema que se plantea en situaciones reales es la forma en que se presenta la información muestral disponible para la estimación y ajuste de los modelos bajo estudio. Los datos funcionales, desde su perspectiva teórica, son observaciones de trayectorias o funciones muestrales, generalmente dependientes del tiempo, de un proceso estocástico. La imposibilidad práctica de hacer mediciones de manera continua de tales funciones hace que, como mucho, dispongamos de observaciones puntuales en tantos instantes discretos como deseemos, lo que hace necesario el desarrollo de técnicas eficaces que permitan conocer de forma aproximada las funciones de las que se obtienen dichas observaciones. Como afirman Ramsay & Silverman (1997), conoceremos una función cuando seamos capaces de calcular su valor en cualquier punto de su dominio, de ahí que ante la imposibilidad de conocer la expresión analítica de una función, será suficiente con tener la posibilidad de evaluarla en cualquier punto. Analizaremos aquí las dos técnicas indicadas en los capítulos teóricos: interpolación spline cúbica natural y aproximación mínimo-cuadrática.

Según hemos visto en el Capítulo 2, el ajuste de un modelo de regresión logística funcional se reduce a uno múltiple cuya matriz de diseño se obtiene a partir de las observaciones discretas de las trayectorias "explicativas" (así las llamaremos puesto que se utilizan para explicar una variable respuesta binaria); observaciones en las que encontraremos usualmente una gran dependencia por la naturaleza de las mismas y la forma de obtenerlas con el cosiguiente perjuicio para la estimación del modelo. Al igual que en el caso múltiple, hemos propuesto, para resolver estos problemas, la utilización de análisis en componentes principales, en este caso funcional. En los ejemplos que siguen se ilustrará cómo se mejora la estimación del parámetro funcional de nuestro modelo al utilzar ACPF de las trayectorias explicativas.

A continuación vamos a presentar una serie de ejemplos simulados para ilustrar tanto los problemas que surgen en los ajustes de los modelos de regresión logística funcional, como lo adecuado de las distintas soluciones adoptadas.

El primer ejemplo consiste en el ajuste de un modelo de regresión logística funcional en el que las trayectorias explicativas proceden de un proceso estocástico cuyas trayectorias son funciones spline cúbicas. Para ello simularemos una serie de trayectorias y, a partir de ellas, sus correspondientes valores de respuesta, fijando previamente una función parámetro. En los ejemplos segundo y tercero, en lugar de simular las trayectorias propiamente dichas, simulamos los valores de un conjunto de trayectorias de un proceso estocástico conocido, en una partición de nodos de observación; y a partir de los métodos de interpolación y aproximación mínimo cuadrática (ejemplos dos y tres respectivamente) aproximamos las trayectorias explicativas. En cada uno de los casos se obtienen los valores simulados de la respuesta a partir de una función parámetro previamente fijada. En los tres ejemplos repetiremos un número grande de veces la simulación de los valores de la respuesta, los ajustes de los modelos de regresión logística y la estimación de las funciones parámetro; y calcularemos determinadas medidas de bondad del ajuste y de la precisión de las estimaciones.

3.2.1 Simulación de los datos

A continuación expondremos con detalle los pasos seguidos en cada ejemplo de simulación que en todos los casos sigue el mismo esquema, variando en cada uno de ellos determinados aspectos que indicaremos en su momento. Consideraremos en todos los casos que las trayectorias son regulares y pueden aproximarse en el espacio generado por las funciones B-spline definidas a partir de un conjunto de nodos $\tau_0, ..., \tau_q$ mediante la espresión recursiva (2.4), y que por tanto se pueden expresar de la forma

$$x_i(t) = \sum_{j=-1}^{q+1} a_{ij} B_j(t)$$

variando en cada ejemplo los nodos de definición de las funciones básicas (B-splines) y el soporte de definición del proceso.

El primer paso es la obtención de una muestra aleatoria de trayectorias del proceso que se considere en cada ejemplo $x_1(t), ..., x_n(t)$, las cuales quedarán completamente determinadas cuando se conozcan los coeficientes de la expresión anterior para cada trayectoria, coeficientes que se pueden resumir en una matriz de dimensión $n \times (q+3)$

$$A = (a_{ij})_{i=1,...,n;j=-1,...,q+1}.$$

Los distintos ejemplos que se van a presentar muestran tres formas distintas de obtener los coeficientes de la expresión anterior. Así, en el primer ejemplo se consideran los splines directamente mediante la simulación de estos coeficientes de manera aleatoria escogiendo valores de una transformación mediante valores uniformes en $[0,1]$ de la distribución Normal estándar. En el segundo ejemplo se aborda la situación en que la información disponible es un conjunto de observaciones de una muestra de trayectorias del proceso en un conjunto de instantes discretos y que se consideran medidas sin error, en los mismos instantes de definición de los B-spline, esto es, las trayectorias muestrales se aproximan mediante interpolación spline cúbica natural. En el tercer ejemplo se repite la situación anterior, pero con la salvedad de que ahora las observaciones se consideran tomadas con error, y además se dispone de observaciones de las trayectorias en los nodos de definición de los B-splines y en un conjunto de nodos intermedios, por tanto podemos obtener los elementos de la matriz A mediante la aproximación mínimo cuadrática vista en la expresión (2.6).

Una vez simuladas las trayectorias explicativas, el siguiente paso es la simulación del modelo logístico funcional, esto es, de los valores de la variable respuesta a partir de las trayectorias del proceso. Para ello previamente hemos de fijar la función parámetro. En los casos que nos ocuparán, tales funciones serán distintas funciones suaves (sinusoidales generalmente) y que se expresarán también en términos de la base de funciones B-spline definida previamente en la forma

$$\beta(t) = \sum_{j=-1}^{q+1} \beta_j B_j(t).$$

Al igual que en el caso de las trayectorias, la función parámetro simulada se tendrá cuando se conozcan sus coeficientes β_j de manera que, para fijar la función parámetro en los ejemplos que aquí presentamos, hemos obtenido dichos coeficientes a partir de la interpolación de los valores de las funciones en los nodos de definición de los spline.

Después de fijada la función parámetro, ya podemos obtener los valores de la variable respuesta $y_1, ..., y_n$ a partir de la simulación de valores de n distribuciones de Bernouilli cada una con probabilidad dada por

$$\pi_i = \frac{\exp\left\{\alpha + \int_{[\tau_0,\tau_q]} x_i(t)\,\beta(t)\,dt\right\}}{1 + \exp\left\{\alpha + \int_{[\tau_0,\tau_q]} x_i(t)\,\beta(t)\,dt\right\}},\ i = 1, ..., n.$$

Para obtener estas probabilidades se fija en primer lugar un valor de α, y posteriormente, para el cálculo de la integral anterior, se aprovecha la expresión

en términos de la base de funciones de la forma

$$\int_{[\tau_0,\tau_q]} x_i(t)\,\beta(t)\,dt = a_i^T \Psi \beta,\ i = 1, ..., n$$

con

$$\begin{aligned} \Psi_{(q+3)\times(q+3)} &= (\psi_{jk});\ \psi_{jk} = \int_{[\tau_0,\tau_q]} B_j(t)\,B_k(t)\,dt,\ j,k = -1, ..., q+3 \\ \beta &= (\beta_{-1}, \ldots, \beta_{q+3}) \end{aligned}$$

y a_i^T la i-ésima fila de A. Las integrales de esta expresión se han obtenido mediante la fórmula de cuadratura de Gauss con cuatro nodos propuesta por Ocaña (1995).

Cabe aquí destacar, al igual que en el caso múltiple, la importancia que tiene en los ejemplos simulados la elección adecuada de la función parámetro, de las trayectorias explicativas y del parámetro constante α, para que las exponenciales de la expresión anterior no crezcan o dismunuyan excesivamente y nos lleven a valores de la variable respuesta que sean o bien todos cero o bien todos uno, lo que haría al modelo de regresión logística inestable.

Una vez simulados los datos, estamos en condiciones de ajustar el modelo de regresión logística funcional y analizar la precisión de la estimación de la función parámetro así como la ganancia en reducción de dimensión obtenida al utilizar las cc. pp. como variables explicativas. Tanto por la imposibilidad de dar una estimación directa de la función parámetro del modelo como de disponer de las trayectorias explicativas, en su expresión analítica, y poder calcular la integral que involucra el modelo de regresión logística; es necesario dar una aproximación múltiple al problema de regresión logística funcional. Como se vio en el Capítulo 2, esta aproximación pasa por el ajuste de un modelo de regresión logística múltiple cuya matriz de diseño es

$$(\mathbf{1} \mid A\Psi),$$

con $\mathbf{1}$ un vector $n \times 1$ de unos. Si llamamos $\widehat{\beta} = \left(\widehat{\beta}_{-1}, \widehat{\beta}_2, \ldots, \widehat{\beta}_{q+1}\right)^T$ a los parámetros estimados por este modelo múltiple, la estimación de la función parámetro del correspondiente modelo funcional será

$$\widehat{\beta}(t) = \sum_{j=-1}^{q+1} \widehat{\beta}_j B_j(t)$$

Al ajustar este modelo múltiple surgen determinados problemas como los vistos en el primer capítulo debidos a la alta multicolinealidad existente en la

matriz de diseño, y que se ve más acusada en este caso funcional por la forma en que se obtiene ésta a partir de observaciones dependientes del tiempo. Asimismo surgen otros problemas derivados de la aproximación de las trayectorias en el espacio de dimensión finita que genera la base de B-splines, y que comentaremos en cada caso posteriormente. Además de todo esto aparece el problema de la dimensión que consiste en que en muchos casos para que la aproximación de las trayectorias sea adecuada es necesario que el número de funciones básicas que conforman la base sea elevado, con el consiguiente perjuicio en el ajuste del modelo de regresión logística. Para dar respuesta a todos estos probemas hemos propuesto la utilización de análisis en componentes principales funcional, de manera que, en lugar de ajustar el modelo de regresión logística múltiple visto anteriormente, ajustaríamos el modelo que tiene como valores de la variable respuesta binomial los simulados y como valores de las variables explicativas los de las componentes principales.

Existen dos aspectos a tener en cuenta a la hora de utilizar componentes principales como regresores de un modelo de regresión: por un lado el tipo de análisis de componentes principales a utilizar, y por otro el número de componentes más adecuado junto con el orden en que éstas se introducen en el modelo. La respueta a la segunda cuestión ya fue amplimente tratada en el caso múltiple, llegando a lo que convinimos en denomiar Método I (entrada de las componentesen el modelo en orden a su variabilidad) y Método II (entrada mediante selección stepwise). En cuanto a la segunda cuestión debemos notar que existe una doble alternativa: la utilización de las componentes principales de la matriz de diseño del modelo múltiple sin la columna de unos ($A\Psi$), a lo que llamaremos ACP1; o bien considerar el análisis en componentes principales funcional con respecto a la métrica usual en $L^2(T)$, que, como vimos en el Capítulo 2, es equivalente a considerar las componentes principales de la matriz $A\Psi^{1/2}$ y a lo que llamaremos ACP2. En cualquier caso, y como se demuestra en Aguilera et al. (2002), el ACP1 puede verse también como el ACP funcional de una transformación adecuada de los datos respecto de la métrica usual en $L^2(T)$. En los ejemplos que se presentarán posteriormente se han calculado ambos tipos de ACP.

Teniendo todo esto en cuenta y siguiendo las pautas vistas en el caso múltiple, ajustaremos los modelos de regresión logística con distinto número de cc. pp. según los dos tipos de ACP y órdenes de introducción y obtenemos los parámetros estimados $\widehat{\gamma}_{(s)}$ y $\widehat{\varphi}_{(s)}$ en cada caso. A partir de estas estimaciones obtenemos tanto las estimaciones de las coordenadas de la función parámetro

como de la propia función parámetro $\widehat{\beta}^{1)}_{(s)}$, $\widehat{\beta}^{2)}_{(s)}$, $\widehat{\beta}^{1)}_{(s)}(t)$ y $\widehat{\beta}^{2)}_{(s)}(t)$. Para decidir la estimación más adecuada en cada caso obtenemos las medidas propuestas en los capítulos teóricos ECMB y ECMBI eligiendo como adecuada aquellas estimaciones que minimicen estas cantidades.

Otras medidas también muy útiles a la hora de decidir la estimación más precisa posible de la función parámetro y de la bondad del ajuste de los distintos modelos son las vistas en el caso múltiple: error cuadrático medio de las probabilidades (ECMP), tasa de clasificaciones correctas con 0.5 como punto de corte (CCR), máximo de las diferencias en valor absoluto entre los coeficientes de la expresión (3.3) simulados y estimados (Max), varianza de las estimaciones de los parámetros anteriores (Var), estadístico de bondad de ajuste G^2 así como su p-valor y R^2. Todas estas medidas se obtendrán en los distintos ejemplos y nos ayudarán a la elección del mejor ajuste y más preciso.

Finalmente indicar que en problemas con datos reales no se dispone de la verdadera función parámetro con lo que se debe escoger como número de componentes aquel a partir del cual se produzca un incremento considerable en la varianza de las estimaciones de los parámetros desde el punto de vista múltiple como vimos en aquella ocasión.

3.2.2 Ejemplo 1: simulación de un proceso cuyas trayectorias son funciones spline

Con este primer ejemplo pretendemos ilustrar el modelo de regresión logística funcional con trayectorias explicativas pertenecientes a un espacio de dimensión finita en lugar de una aproximación de las mismas. Así, consideraremos un proceso cuyas trayectorias son funciones spline cúbicas expresadas en términos de la base de funciones B-spline cúbicas definidas con los nodos $0, 1, 2, \ldots, 9, 10$; y una muestra aleatoria simple de las mismas de tamaño 100. Para ello simulamos 100 vectores aleatorios de dimensión 13 cada uno y correspondientes a los coeficientes de la expresión de los splines en términos de los B-splines

$$x_i(t) = \sum_{j=-1}^{11} a_{ij} B_j(t), \; i = 1, \ldots, 100$$

En definitiva los valores simulados lo son de una matriz de dimensión 100×13

$$A = (a_{ij})_{i=1,..,100;j=-1,...,11}$$

Para la obtención de esta matriz hemos seguido las mismas pautas del ejemplo múltiple, esto es, hemos simulado 100 valores de una distribución normal de dimensión 13, con vector de medias nulo y la identidad como matriz de varianzas-covarianzas. Posteriormente, y con objeto de que exista multicolinealidad, hemos simulado valores uniformes en el intervalo $[0, 1]$ correspondientes a cada uno de los elementos de una matriz de dimensión 13×13. El producto de estas dos matrices dará lugar a los coeficientes de la matriz A.

Después de disponer de las trayectorias explicativas del proceso, fijamos como función parámetro del modelo logístico la interpolación spline cúbica natural de $sen\,(x + \pi/4)$ sobre los nodos antes mencionados. Su expresión en términos de la base de funciones B-spline es

$$\beta(t) = \sum_{j=-1}^{11} \beta_j B_j(t) \tag{3.3}$$

siendo tales coeficientes

β_1	β_2	β_3	β_4	β_5	β_6	
-1.63	-0.71	0.22	1.22	0.94	-0.09	
β_7	β_8	β_9	β_{10}	β_{11}	β_{12}	β_{13}
-1.04	-1.04	-0.08	0.95	1.12	0.21	-0.70

Una vez fijada la función parámetro y simuladas las trayectorias, calculamos las integrales

$$\int_{[0,10]} x_i(t)\,\beta(t)\,dt = a_i^T \Psi \beta,\ i = 1, ..., 100 \tag{3.4}$$

que nos llevarán a las probabilidades simuladas

$$\pi_i = \frac{\exp\left\{\alpha + a_i^T \Psi \beta\right\}}{1 + \exp\left\{\alpha + a_i^T \Psi \beta\right\}},\ i = 1, \ldots, 100 \tag{3.5}$$

con $\alpha = 0.5$, y finalmente a los valores de la variable respuesta mediante la distribución de Bernouilli de dichas probabilidades

$$y_i = B(\pi_i),\ i = 1, ..., 100.$$

Una vez simulados tanto los valores de la variable respuesta como las trayectorias explicativas, ajustamos el modelo de regresión logística funcional

$$y_i = \pi_i + \varepsilon_i,\ i = 1, ..., 100$$

que, por la forma de considerar tanto las trayectorias como la función parámetro, sabemos que no es más que un modelo de regresión logística múltiple con matriz de diseño

$$(\mathbf{1} \mid A\Psi)$$

siendo A la matriz de dimensión 100×13 simulada y $\mathbf{1}$ el vector 100×1 de unos. El término independiente estimado es $\widehat{\alpha} = 1.54$ frente al real $\alpha = 0.5$ y el resto de parámetros estimados por este modelo aparecen en la penúltima columna de la Tabla 3.11 que, como podemos apreciar, difieren bastante de los reales a pesar de que el modelo se ajusta globalmente bien ($p = 0.9949$). La magnitud de esas diferencias se aprecia mucho más claramente si en lugar de comparar coeficiente a coeficiente las comparamos globalmente mediante las respectivas representaciones gráficas de las funciones parámetro real y estimada que se presentan en la Figura (3.1).

Con la doble intención de mejorar la estimación de la función parámetro y de reducir la dimensión del problema, al igual que hicimos en el caso múltiple, ajustamos los modelos de regresión logística con 1, 2, ... componentes principales mediante ambos análisis (ACP1 y ACP2), introduciendo dichas componentes en el modelo en orden de variabilidad y en el orden dado por el método stepwise (Método I y Método II) respectivamente, y obtenemos tanto las estimaciones que proporciona cada ajuste como las medidas de bondad de ajuste más usuales y las de precisión de las estimaciones de los parámetros propuestas en esta memoria. Con objeto de que la observación de las cantidades sea más cómoda decidimos agrupar todas estas medidas en tablas al final de la sección. Observando estas medidas se aprecia que:

- Variabilidad acumulada: (V. Ac.). En la Tabla 3.12 se ve que al ajustar los modelos con ACP1 se supera el 92% de variabilidad explicada con las cinco primeras componentes principales y que con ocho se alcanza prácticamente el 99%. Estas mismas cantidades son alcanzadas respectivamente con una y cinco componentes en el caso del ACP2 (Tabla 3.14), con lo que parece que para reducir la dimesión resultaría más eficaz el ACP2 que el ACP1 ya que se consiguen mayores porcentajes de variabilidad explicada con menor número de componentes.

- ECMP. Si nuestro objetivo es predecir correctamente las probabilidades de respuesta de la variable dependiente, hemos de observar el error cuadrático medio de las probabilidades estimadas; cantidad que presenta

su menor valor (0.005) cuando introducimos la cuarta componente en el modelo con ACP1 y Método I (Tabla 3.12), lo que indicaría ya una reducción de dimensión en cuanto a este objetivo, sin embargo con el método II el mínimo se alcanza con tres componentes al entrar la cuarta en el método Stepwise (Tabla 3.15) con un valor de 0.014, valor que es superior, lo que indicaría un peor ajuste, pero por contra una mayor reducción de la dimensión. En cuanto al ACP2 (Tablas 3.13 y 3.14), esta medida se comporta de manera muy similar a como lo hacía en el ACP1 tanto individualmente en cada método de entrada de componentes, en los que se alcanzan mínimos para los modelos con cinco y tres componentes y valores de 0.016 y 0.009, como en la comparación de dichos métodos.

- CCR. La tasa de clasificaciones correctas, al igual que el ECMP, es una medida de la capacidad predictiva del modelo, siendo en este ejemplo los valores de dicha tasa generalmente altos rondando en todos los casos el 85% y el 90%, aunque en ningún método de entrada de variables ni tipo de ACP se observa un comportamiente mejor a otro, en el sentido de conseguir un claro incremento en dicha tasa.

- ECMB. Con los valores del error cuadrático medio de los coeficientes estimados, ocurre algo parecido a lo que ocurría con el ECMP en el ACP1, no así en el ACP2. El valor mínimo en el primero (0.263) se alcanza con cuatro componentes en el Método I y con tres (0.288) en el II (valores muy similares aunque ligeramente más pequeño el primero pero también ligera menor reducción de dimensión) según se puede ver en las Tablas 3.12 y 3.15. Sin embargo para el segundo (Tablas 3.13 y 3.14) esos mínimos se alcanzan con una y dos componentes principales y valores mucho más altos de 0.755 y 0.697 para los Métodos I y II respectivamente. Este hecho pone de manifiesto que esta medida es adecuada para decidir lo precisas que son las estimaciones de los parámetros y la reducción de dimensión que se consigue con ACP1, no así con ACP2. De hecho en la Figura **??** se observa más claramente la mala estimación de la función parámetro obtenida con el ACP2 y el criterio del mínimo ECMB mientras que la Figura 3.2 indica una estimación mucho más precisa. Cabe destacar en cuanto a esta medida se refiere, el aumento que experimenta cuando introducimos las últimas componentes en el modelo (Tabla 3.12), llegando a tomar valores del orden de 10^5 y 10^7 lo que pone de manifiesto lo mal que se estiman los parámetros (en este caso los coeficientes de la

expresión (3.3)) cuando se introducen muchas componentes; recordemos que el modelo con todas las componentes principales como regresores proporcionaba iguales parámetros estimados que el modelo con las variables originales en el caso múltiple y por tanto igual ocurrirá ahora.

- Max. El máximo de las diferencias entre parámetros simulados y estimados en valor absoluto era una medida alternativa al ECMB en el caso múltiple y que generalmente se comportaba de manera similar a ésta. El caso funcional también recoge su idoneidad para la decisión en el ACP1 y no así en el ACP2 como se aprecia en las Tablas 3.12, 3.15, 3.14 y 3.13 de tal manera que en el ACP1 tanto con el Método I como con el II, se ve que se alcanzan valores mínimos de dichos máximos para los modelos con 4 y 3 componentes respectivamente y con valores de 1.581 y 1.577; valores que son muy similares como ocurría en el ECMB. Sin embargo para el ACP2 los valores mínimos se alcanzan para los modelos con una sola componente principal y valores 1.756 en ambos casos.

- ECMBI. Recordemos que esta medida era una alernativa al ECMB cuando lo que realmente nos interesa es estimar de manera precisa la función parámetro del modelo de regresión logística funcional y olvidarnos de algún modo del término independiente α. En el caso de ACP1, el error cuadrático integrado se comporta como el ECMB, alcanzando los valores mínimos para los modelos que tienen las 4 primeras cc. pp. con el método I y 3 en el método II con valores respectivamente de 0.272 y 0.564, que se pueden ver en las Tablas 3.12 y 3.15. Sin embargo en el ACP2 (Tablas 3.14 y 3.13) el comportamiento de esta medida no es como en el del ECMB, ya que ahora los valores más pequeños se alcanzan para los modelos con cinco componentes en el Método I (1.313) y con tres en el Método II (1.544), en definitiva un comportamiento análogo al que produce ECMB y ECMBI en el ACP1. Este hecho nos lleva a concluir que ECMB y ECMBI se comportan de igual manera en cuanto a la reducción de dimensión estimando la función parámetro al utilizar ACP1, y siendo esta redución mayor en el Método II que en el I aunque la precisión sea ligeramente menor, lo que además se aprecia en la Figura 3.2. Sin embargo cuando utilizamos ACP2 es el ECMBI la medida más adecuada a pesar de producir peores resultados en cuanto a precisión (1.313 frente a 0.272 en el Método I y 1.544 frente a 0.564 en el II), y como se puede ver en la Figura 3.3, comparada con la 3.2 y en cuanto a reducción de

dimensión (5 frente a 4 componentes en el Método I y 3 frente a 3 en el II).

- Var. Recordemos del caso múltiple que la varianza de las estimaciones de los parámetros (en este caso de los coeficientes de las expresiones en términos de las funciones básicas) era una alternativa al ECMB para decidir el número de componentes más adecuado para una estimación precisa de dichos parámetros y una reducción adecuada de la dimensión del problema, cuando no se conocían los verdaderos parámetros del modelo; de tal manera que habíamos decidido que lo adecuado es elegir aquel número de componentes anterior a un aumento significativo en los valores de estas varianzas. De la observación de los resultados de las distintas alternativas consideradas de ajuste (ACP1 y Método I (Tabla 3.12), ACP1 y Método II (Tabla 3.15), ACP2 y Método I (Tabla 3.14) y ACP2 y Método II (Tabla 3.13)) se tiene que en el Método II debemos quedarnos con todas las componentes que entran en el modelo en ambos tipos de ACP (igual resultado que con ECMB y ECMBI) ya que no se produce un aumento significativo de esta varianza, y que con el Método I, en ambos casos se produce un incremento de valor de la varianza al introducir la séptima componente (de 3.228 a 9.626 en el ACP1 y de 6.216 a 14.45 en el ACP2) por lo que nos quedaríamos con el modelo que tiene 6 componentes que a raiz de los correspondientes valores de ECMB y ECMBI, proporciona estimaciones parecidas a la obtenidas con los modelos que proporcionan dichas medidas como se aprecia además en las Figuras 3.5 y 3.6.

- G^2 y p-val. Para terminar con las medidas, hemos calculado medidas globales de bondad de ajuste para cada modelo en cada situación analizada como son el estadístico deviance y su p-valor asociado que indica lo apropiado del ajuste del modelo logístico para los datos, y en todos los casos se ve que los modelos son convenientes ya que proporcionan p-valores próximos a la unidad incluso en modelos con muy pocas componentes principales.

De todo lo expuesto anteriormente podemos concluir que en este ejemplo cuando ajustamos el modelo con los datos originales o bien con el modelo que tiene todas las componentes principales, sea cual sea el tipo de ACP elegido, se produce una estimación muy mala de la función parámetro. Al utilizar componentes principales en un número menor del total se obtienen muy buenas

mejoras: así al utilizar ACP1 se obtiene una buena estimación de la función parámetro tanto con el Método I con cuatro componentes en el modelo como con el II con tres, de lo cual informan tanto el ECMB como el ECMBI y se puede ver en las Tablas 3.12 y 3.15 y en la Figura 3.2. Al utilizar ACP2 se obtiene una estimación poco acertada con ambos Métodos (I y II) siguiendo el criterio del mínimo ECMB (Tablas 3.14 y 3.13 y la Figura **??**). Finalmente al utilizar ACP2 con cinco componentes en el Método I y tres según el II de acuerdo al mínimo ECMBI (Tablas 3.14 y 3.13 y la Figura 3.3) se obtiene mejor estimación que la anterior pero ligeramente peor que la primera. En definitiva, parece más adecuado utilizar ACP1 que ACP2 a la vista de las respectivas gráficas y en este caso daría igual utilizar ECMB que ECMBI para tomar la decisión del número más adecuado de componentes a utilizar, sin embargo si se decide utilizar ACP2 es conveniente decidir sobre el número de componentes a través del ECMBI. Finalmente hay que llamar la atención sobre las Figuras 3.5 Y 3.6 que representan las funciones parámetro estimadas con 6 y 3 componentes en ACP1 y ACP2 respectivamente, valores que decidiríamos a través de la varianza de los parámetros estimados en un caso real y que no difieren demasiado de los modelos óptimos ya comentados.

Repetición de la situación

Después de ilustrar con un ejemplo la simulación de las trayectorias spline cúbicas vamos a repetir esta situación un número grande de veces con el objeto de mostrar lo adecuado de la utilización de un ACPF a la hora de la estimación precisa de la función parámetro de un modelo de regresión logística funcional. Para ello consideramos las trayectorias simuladas en nuestro ejemplo así como la función parámetro, y a partir de las transformaciones lineales (3.4) y de las probabilidades obtenidas de ellas (3.5) simulamos 350 vectores de dimensión 100 de valores de variables respuesta dicotómica a partir de una distribución de Bernouilli con las probabilidades anteriormente referidas. Con cada uno de estos vectores de respuesta y con la matriz de diseño $(\mathbf{1} \mid A\Psi)$ ajustamos los modelos en terminos de las componentes principles utilizando los dos tipos de análisis en componentes principales: ACP1 y ACP2; así como los dos métodos de introducción de componentes en el modelo: Método I y II. Seguidamente reconstruimos los coeficientes de (3.3) como hemos visto anteriormente y calculamos las medidas de bondad tanto del ajuste de cada modelo como de la precisión de las estimaciones analizadas también en los casos anteriores.

Con objeto de dar medidas que resuman estas 350 repeticiones nos decidi-

mos por considerar el promedio de las medidas que se obtienen del modelo óptimo de cada repetición (ECMB, ECMBI,...), siendo tal modelo óptimo el que minimiza el ECMB. Asimismo repetimos dichos promedios, pero considerando ahora como modelos óptimos los que minimizan en cada repetición el ECMBI, con objeto de analizar si las diferencias detectadas en el caso anterior, en cuanto a la mejor medida para la elección del modelo óptimo con ACP2, se mantienen ahora. Resumiendo, consideramos en cada repetición los posibles modelos óptimos para cada una de las posibilidades: ACP1-Método I, ACP1-Método II, ACP2-Método I y ACP2-Método II; y calculamos en cada una de dichas posibilidades el promedio de: el número de componentes que entran en el modelo (N. cc. pp.), el error cuadrático medio de las probabilidades (ECMP), la tasa de clasificaciones correctas con 0.5 como punto de corte (CCR), el error cuadrático medio integrado (ECMBI), el error cuadrático medio de los coeficientes (ECMB), el máximo de las diferencias de cada coeficiente en valor absoluto (Max), la varianza de las estimaciones de los coeficientes o parámetros (Var) y el estadístico G^2. Las medias de estas cantidades para las ocho situaciones descritas se resumen en las tablas 3.16, 3.17, 3.18 y 3.19. Además del promedio de las medidas citadas anteriormente hemos calculado (y se muestran en dichas tablas) la varianza y el coeficiente de variación como medidas de dispersión asociadas al promedio antes citado con la intención de ver la representatividad de los promedios indicados.

Asimismo hemos calculado en cada una de las ocho situaciones expuestas, y en cada repetición, la reconstrucción de los coeficientes estimados de (3.3) para los modelos óptimos y como estimación final de dichos coeficientes elegimos el promedio de estas 350 estimaciones óptimas lo que nos permitirá analizar el método de inclusión de componentes principales más adecuado y el ACP idóneo para la estimación de la función parámetro del modelo de regresión logística funcional. Los resultados de estos parámetros promedio se muestran en la Tabla 3.20 y en las Figuras 3.7, 3.8 y 3.9.

De las medidas calculadas podemos extraer las siguientes conclusiones:

- En primer lugar hay que observar que con ACP1 es equivalente utilizar ECMB y ECMBI como medida de selección del modelo óptimo lo que se aprecia en las Tablas 3.16 y 3.17 ya que proporcionan ambas prácticamente el mismo número medio de componentes principales óptimo 4.183 y 4.186 respectivamente para el Método I y 2.811 y 2.851 respectivamente para el II; además ésto que ocurre con el promedio del número de componentes ocurre con el resto de medidas: varianza y coeficiente de variación.

Sin embargo con el ACP2 (Tablas 3.18 y 3.19) no ocurre igual ya que por ejemplo el número promedio de componentes óptimo tomando el Método I es 1.146 si consideramos el ECMB y 4.914 si tomamos el ECMBI, valor que se parece bastante a los 4.183 vistos con ACP1. Con el Método II también se aprecian diferencias en el número de componentes óptimo para ACP2 entre ECMB siendo éste 1.191, y ECMBI siendo 2.814 valor parecido también a lo que ocurría para ECMB y ECMBI en el caso de ACP1. Todo ésto viene a corroborar lo visto el en caso anterior; que para ACP1 da igual la medida a utilizar para la selección del mejor modelo en cuanto a estimación de la función parámetro, mientras que cuando se utiliza ACP2 es más conveniente utilizar el ECMBI.

- Otro hecho que ponen de manifiesto las medidas resumen es la reducción de dimensión que se obtiene al utilizar componentes principales en el modelo de regresión logística ya que como se ve en las Tablas 3.16 y 3.17 tanto para ECMB como para ECMBI y en las Tablas 3.18 y 3.19 para ECMBI, el modelo óptimo "medio" será el que tiene unas cuatro componentes en el mismo con el Método I y unas tres con el II, por lo que este último método reduce más la dimensión del problema. Además de tener presente este promedio, sería conveniente examinar alguna medida de dispersión asociada a dicho promedio para asegurarnos, de algún modo, que es representativo de lo que ocurre en la muestra, y estas medidas vienen dadas por la varianza y el coeficiente de variación, en los que podemos ver valores pequeños, especialmente de este último, para aquellos promedios que aparecen en las tablas 3.16 y 3.17.

- En cuanto al resto de medidas cabe destacar las tasas de clasificaciones correctas que se comportan como el ejemplo particular visto anteriormente al tomar valores promedios del 85% de clasificaciones correctas en todos los casos, aunque en los casos en los que no es adecuado el ECMB para la decisión (parte izquierda de las Tablas 3.18 y 3.19) tales clasificaciones bajan a rondar el 80%.

- En cuanto a la variabilidad acumulada por los modelos óptimos con el Método I, se aprecia que también ocurre como en el caso particular, y con ACP1 se obtienen unos promedios de variabilidad explicada de un 88% aproximadamente (Tabla 3.16), mientras que con el ACP2 se llega hasta un promedio del 92%. Si vigilamos los coeficientes de variación de estas cantidades se tiene que son realmente bajos con valores que rondan

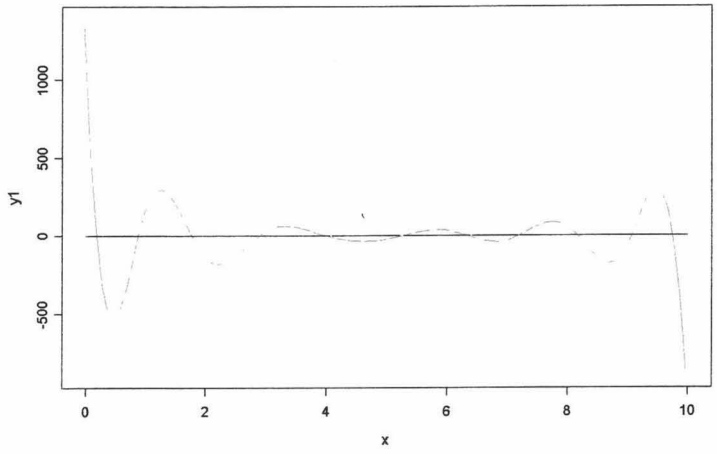


Figura 3.1: Caso funcional. Ejemplo 1. Funciones simulada (Negro) y estimada mediante la aproximación múltiple del modelo de regresión logística funcional (Rojo).

el 0.05, y el 0.01, lo que me indica que en la muestra las variabilidades con estos métodos se comportan más o menos como se ve en el promedio.

- Finalmente quiero destacar la similitud existente en los gráficos de las funciones parámetro estimadas a través de los promedios de los coeficientes que se obtienen mediante las distintas posibilidades de ajuste anteriormente definidas (Figuras 3.7, 3.8 y 3.9) y los presentados anteriormente, lo que vuelve a poner de manifiesto la reducción de dimensión que se consigue al utilizar ACP como regresores, y cómo el ACP1 (Tabla 3.17) proporciona estimaciones ligeramente más precisas que el ACP2, este último tomando como medida para la elección el ECMBI, (Tabla 3.19) y en los dos casos prefiriendo el Método II al I si no por su precisión en las estimaciones, sí por la mayor reducción de dimensión que consigue. Por otro lado el la Figura 3.8 pone de manifiesto lo inadecuado de utilizar ECMB como criterio de selección de modelos óptimos con ACP2.

	ACP1		ACP2					
Met.	I	II	I		II		Variables	
s	4	3	1	5	2	3	Originales	Reales
$\widehat{\alpha}_{(s)}$	0.752	0.762	0.663	0.918	0.827	0.762	1.541	0.500
$\widehat{\beta}_{1(s)}$	-0.050	-0.055	0.125	1.827	0.164	-0.055	1.92E+4	-1.638
$\widehat{\beta}_{2(s)}$	-0.490	-0.542	0.134	-1.675	-0.283	-0.542	-3.01E+3	-0.707
$\widehat{\beta}_{3(s)}$	-0.005	-0.086	0.155	0.604	0.704	-0.086	1.10E+3	0.217
$\widehat{\beta}_{4(s)}$	1.282	1.275	0.118	1.892	0.823	1.275	-5.04E+2	1.116
$\widehat{\beta}_{5(s)}$	1.006	1.127	0.155	-0.151	-0.574	1.127	1.75E+2	0.942
$\widehat{\beta}_{6(s)}$	-0.052	0.144	0.137	0.682	1.062	0.144	-20.315	-0.086
$\widehat{\beta}_{7(s)}$	-0.797	-0.662	0.147	-0.294	0.247	-0.662	-51.957	-1.038
$\widehat{\beta}_{8(s)}$	-0.860	-0.851	0.173	-1.921	-1.405	-0.851	1.07E+2	-1.035
$\widehat{\beta}_{9(s)}$	-0.327	-0.423	0.109	-0.326	-0.386	-0.423	-1.44E+2	-0.080
$\widehat{\beta}_{10(s)}$	0.549	0.410	0.134	0.599	0.323	0.410	2.95E+2	0.945
$\widehat{\beta}_{11(s)}$	1.090	0.977	0.154	1.587	1.064	0.977	-6.39E+2	1.115
$\widehat{\beta}_{12(s)}$	0.569	0.529	0.193	0.228	0.017	0.529	1.80E+3	0.209
$\widehat{\beta}_{13(s)}$	0.043	0.041	0.076	0.209	-0.671	0.041	-1.19E+4	-0.698

Tabla 3.11: Caso funcional. Ejemplo 1. Parámetros reconstruidos de los modelos óptimos con los distintos tipos de ACP, distintos Métodos de selección de cc. pp. y distintos criterios de elección del modelo óptimo.

s	V. Ac.	ECMP	CCR	ECMBI	ECMB	Max	Var	G^2	p-val
1	64.25	0.037	85	4.555	0.710	1.638	0.108	73.14	0.97
2	76.10	0.029	89	4.299	0.689	1.641	0.166	71.69	0.98
3	82.20	0.013	87	1.485	0.393	1.622	0.618	62.96	1.00
4	87.50	**0.005**	86	**0.272**	**0.263**	**1.581**	1.242	58.52	1.00
5	**92.53**	0.010	89	1.297	0.373	1.611	1.986	56.14	1.00
6	95.91	0.010	88	1.263	0.368	1.609	3.228	56.14	1.00
7	97.82	0.012	89	4.737	1.189	1.962	**9.626**	53.98	1.00
8	**98.99**	0.013	90	4.895	1.233	1.950	21.481	53.94	1.00
9	99.43	0.016	87	8.526	5.062	4.972	89.599	53.16	1.00
10	99.82	0.017	90	23.417	14.780	6.548	2.0E+2	51.87	1.00
11	100.0	0.017	88	34.771	35.790	8.850	7.2E+2	51.31	1.00
12	100.0	0.022	89	4.4E+3	4.3E+5	1.87E+3	7.5E+2	48.46	1.00
13	100.0	0.028	90	**3.7E+5**	**3.8E+7**	1.92E+4	8.1E+2	44.80	1.00

Tabla 3.12: Caso funcional. Ejemplo 1. Medidas de bondad de ajuste para los modelos con distinto número de cc.pp. como regresores. Componentes obtenidas con el ACP1. Introducción de componentes según el Método I: orden de variabilidad.

s	cc.pp	ECMP	CCR	ECMBI	ECMB	Max	Var	G^2	p-val
1	1	0.036	87	4.565	0.755	**1.756**	0.110	72.94	0.97
5	5	0.031	87	2.968	**0.697**	1.795	0.980	63.64	1.00
3	3	**0.016**	88	**1.544**	1.236	3.227	2.858	57.06	1.00

Tabla 3.13: Caso funcional. Ejemplo 1. Medidas de bondad de ajuste para los modelos con distinto número de cc.pp. como regresores. Componentes obtenidas con el ACP2. Introducción de componentes según el Método II: orden que proporciona el método Stepwise.

s	V.Ac.	ECMP	CCR	ECMBI	ECMB	Max	Var	G^2	p-val
1	**91.99**	0.036	87	4.565	**0.755**	**1.756**	0.110	72.94	0.97
2	95.56	0.030	89	4.395	0.755	1.779	0.173	71.89	0.97
3	96.90	0.016	85	2.626	0.976	2.741	0.931	67.29	0.99
4	97.90	0.015	85	2.555	1.060	2.908	1.577	67.07	0.99
5	**98.80**	**0.009**	87	**1.313**	1.298	3.459	3.774	55.90	1.00
6	99.47	0.010	88	1.793	1.772	4.280	6.216	54.91	1.00
7	99.75	0.012	89	3.956	1.987	3.760	**14.450**	54.21	1.00
8	99.90	0.012	89	3.908	2.025	3.837	31.450	54.20	1.00
9	99.95	0.013	87	7.964	10.774	9.292	2.4E+2	53.42	1.00
10	99.99	0.017	89	24.357	48.959	19.077	4.6E+2	51.59	1.00
11	100.0	0.017	88	34.628	46.064	9.618	1.4E+3	51.34	1.00
12	100.0	0.022	89	4.4E+3	4.3E+5	1.85E+3	2.4E+6	48.45	1.00
13	100.0	0.028	90	3.7E+5	3.8E+7	1.92E+4	1.7E+8	44.80	1.00

Tabla 3.14: Caso funcional. Ejemplo 1. Medidas de bondad de ajuste para los modelos con distinto número de cc.pp. como regresores. Componentes obtenidas con el ACP2. Introducción de componentes según el Método I: orden de variabilidad.

s	cc.pp	ECMP	CCR	ECMBI	ECMB	Max	Var	G^2	p-val
1	1	0.037	85	4.555	0.710	1.638	0.108	73.14	0.97
2	3	0.023	84	1.848	0.422	1.619	0.521	64.88	1.00
3	4	**0.014**	84	**0.564**	**0.288**	**1.577**	1.137	60.18	1.00

Tabla 3.15: Caso funcional. Ejemplo 1. Medidas de bondad de ajuste para los modelos con distinto número de cc.pp. como regresores. Componentes obtenidas con el ACP1. Introducción de componentes según el Método II: orden que proporciona el método Stepwise.

	ECMB			ECMBI		
Medidas	Media	Varianza	CV	Media	Varianza	CV
N. cc. pp.	**4.183**	**0.918**	**0.229**	**4.186**	**0.834**	**0.218**
V. Ac	**88.092**	19.477	**0.050**	**88.165**	18.395	0.049
ECMP	0.009	3.90E-5	0.661	0.009	3.92E-5	0.661
CCR	**85.931**	8.717	0.034	**85.969**	8.781	0.034
ECMBI	0.964	0.589	0.796	0.950	0.580	0.801
ECMB	0.338	6.68E-3	0.242	0.339	6.76E-3	0.242
Max	1.586	6.06E-4	0.016	1.585	6.25E-4	0.016
Var	0.005	7.65E-3	17.815	0.005	7.65E-3	-17.641
G^2	0.983	1.12E-3	0.034	0.983	1.10E-3	0.034

Tabla 3.16: Caso funcional. Repetición del Ejemplo 1. Promedios y variabilidades de distintas medidas de bondad de ajuste obtenidas de los modelos óptimos. Componentes obtenidas con ACP1 e introducidas según el Método I.

	ECMB			ECMBI		
Medidas	Media	Varianza	CV	Media	Varianza	CV
N. cc.pp.	**2.811**	**0.629**	**0.282**	**2.851**	**0.637**	**0.280**
ECMP	0.019	9.56E-5	0.516	0.019	9.61E-5	0.526
CCR	**84.771**	10.498	0.038	**84.846**	10.526	0.038
ECMBI	1.624	0.995	0.614	1.618	0.992	0.615
ECMB	0.406	0.011	0.258	0.406	0.011	0.258
Max	1.596	8.45E-4	0.018	1.595	8.78E-4	0.019
Var	0.514	0.197	0.863	0.519	0.200	0.861
G^2	0.975	2.622e-3	0.053	0.975	2.615e-3	0.052

Tabla 3.17: Caso funcional. Repetición del Ejemplo 1. Promedios y variabilidades de distintas medidas de bondad de ajuste obtenidas de los modelos óptimos. Componentes obtenidas con ACP1 e introducidas según el Método II.

	ECMB			ECMBI		
Medidas	Media	Varianza	CV	Media	Varianza	CV
N. cc.pp.	**1.146**	0.325	0.498	**4.914**	1.620	0.259
V. Ac.	**92.410**	1.592	**0.014**	98.550	0.900	0.010
ECMP	0.037	1.420e-5	0.101	0.011	0.000	0.605
CCR	**80.820**	11.770	0.042	**86.271**	8.089	0.033
ECMBI	4.552	0.064	0.055	1.531	0.737	0.561
ECMB	0.752	1.45E-4	0.016	1.676	0.370	0.363
Max	1.741	3.77E-3	0.035	3.873	0.524	0.187
Var	0.000	0.000	0.000	-0.141	3.502	-13.272
G^2	0.823	0.026	0.195	0.981	0.001	0.038

Tabla 3.18: Caso funcional. Repetición del Ejemplo 1. Promedios y variabilidades de distintas medidas de bondad de ajuste obtenidas de los modelos óptimos. Componentes obtenidas con ACP2 e introducidas según el Método I.

	ECMB			ECMBI		
Medidas	Media	Varianza	CV	Media	Varianza	CV
N. cc. pp.	**1.191**	0.190	0.365	**2.814**	0.633	0.283
ECMP	0.037	1.249e-5	0.096	0.018	8.502e-5	0.509
CCR	**81.177**	12.822	0.044	**85.140**	10.356	0.038
ECMBI	4.295	0.392	0.146	2.025	0.944	0.480
ECMB	0.738	1.35E-3	0.050	1.450	0.214	0.319
Max	1.745	8.48E-4	0.017	3.435	0.533	0.213
Var	0.037	0.019	3.747	0.475	0.208	0.961
G2	0.847	0.023	0.179	0.978	1.51E-3	0.040

Tabla 3.19: Caso funcional. Repetición del Ejemplo 1. Promedios y variabilidades de distintas medidas de bondad de ajuste obtenidas de los modelos óptimos. Componentes obtenidas con ACP2 e introducidas según el Método II.

	ACP1				ACP2			
	ECMB		ECMBI		ECMB		ECMBI	
Met.	I	II	I	II	I	II	I	II
$\widehat{\alpha}_{(s)}$	0.505	0.495	0.505	0.498	0.412	0.435	0.566	0.559
$\widehat{\beta}_{1(s)}$	-0.045	-0.036	-0.046	-0.036	0.110	0.113	2.241	1.803
$\widehat{\beta}_{2(s)}$	-0.414	-0.328	-0.426	-0.328	0.110	0.050	-1.563	-1.416
$\widehat{\beta}_{3(s)}$	0.172	0.199	0.152	0.204	0.139	0.223	0.229	0.286
$\widehat{\beta}_{4(s)}$	1.379	1.288	1.384	1.291	0.105	0.212	1.589	1.566
$\widehat{\beta}_{5(s)}$	0.890	0.878	0.903	0.873	0.123	0.008	0.429	0.355
$\widehat{\beta}_{6(s)}$	-0.294	-0.165	-0.305	-0.173	0.104	0.257	-0.297	-0.043
$\widehat{\beta}_{7(s)}$	-1.035	-0.833	-1.039	-0.847	0.119	0.137	-0.684	-0.521
$\widehat{\beta}_{8(s)}$	-0.901	-0.727	-0.886	-0.739	0.139	-0.108	-0.866	-0.883
$\widehat{\beta}_{9(s)}$	-0.098	-0.126	-0.100	-0.117	0.099	0.015	0.241	0.341
$\widehat{\beta}_{10(s)}$	0.689	0.462	0.685	0.480	0.122	0.149	0.648	0.462
$\widehat{\beta}_{11(s)}$	1.009	0.751	1.022	0.762	0.139	0.283	1.021	0.686
$\widehat{\beta}_{12(s)}$	0.499	0.383	0.509	0.386	0.163	0.135	0.388	0.282
$\widehat{\beta}_{13(s)}$	0.038	0.029	0.039	0.030	0.068	-0.054	1.101	0.826

Tabla 3.20: Caso funcional. Ejemplo 1. Parámetros promedio de los reconstruidos en cada modelo óptimo con los distintos tipos de ACP, Métodos de selección de cc. pp. y criterios de elección del modelo óptimo.

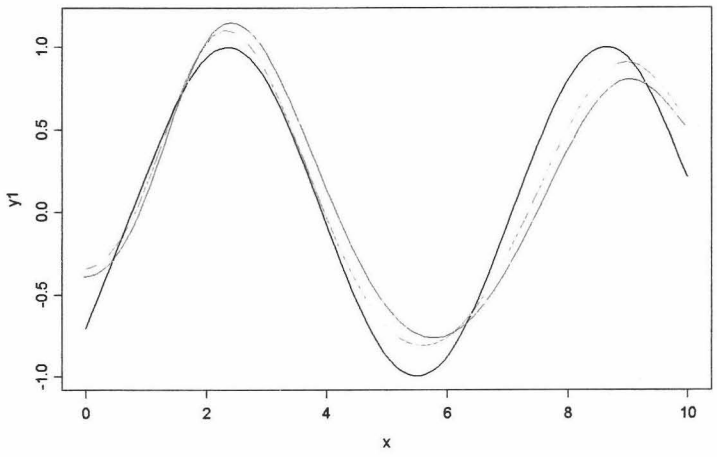


Figura 3.2: Caso funcional. Ejemplo 1. Funciones simulada (Negro) y estimadas utilizando cc. pp: modelo con 4 componentes introducidas según el Método I (Roja), y con 3 introducidas según el II (Azul). Mínimos de ECMB para ACP1.

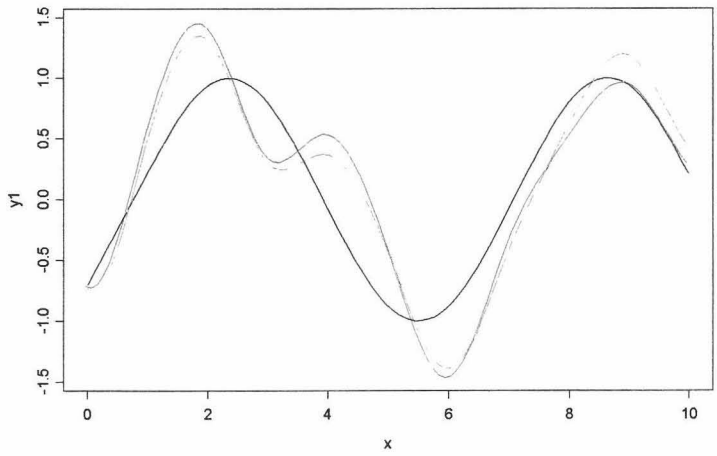


Figura 3.3: Caso funcional. Ejemplo 1. Funciones simulada (Negro) y estimadas utilizando cc. pp: modelo con 5 componentes introducidas según el Método I (Roja), y con 3 introducidas según el II (Azul). Mínimos de ECMBI para ACP2.

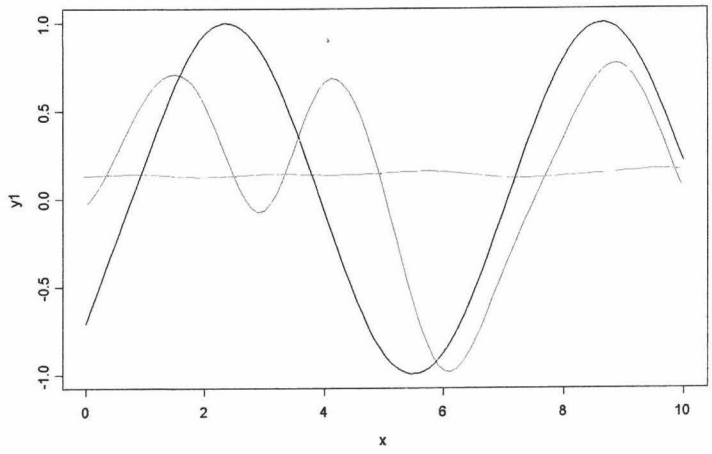


Figura 3.4: Caso funcional. Ejemplo 1. Funciones simulada (Negro) y estimadas utilizando cc. pp: modelo con 1 componente introducida según el Método I (Roja), y con 3 introducidas según el II (Azul). Mínimos de ECMB para ACP2.

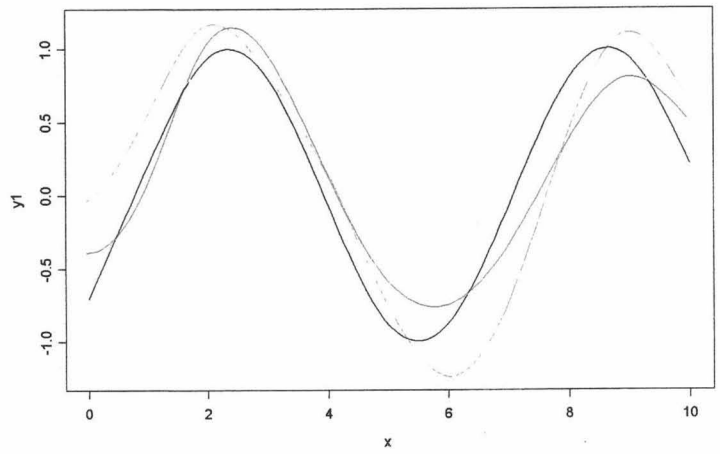


Figura 3.5: Caso funcional. Ejemplo 1. Funciones simulada (Negro) y estimadas utilizando cc. pp: modelo con 6 componentes introducidas según el Método I (Roja), y con 3 introducidas según el II (Azul). Cambio en la Varianza para ACP1.

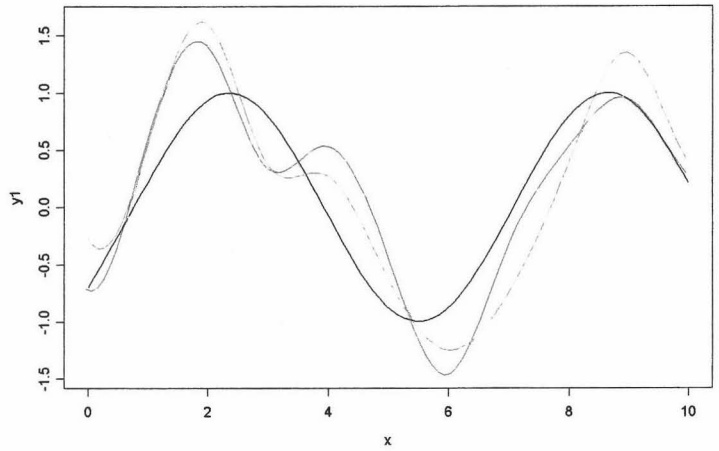


Figura 3.6: Caso funcional. Ejemplo 1. Funciones simulada (Negro) y estimadas utilizando cc. pp: modelo con 6 componentes introducidas según el Método I (Roja), y con 3 introducidas según el II (Azul). Cambio en la Varianza para ACP2.

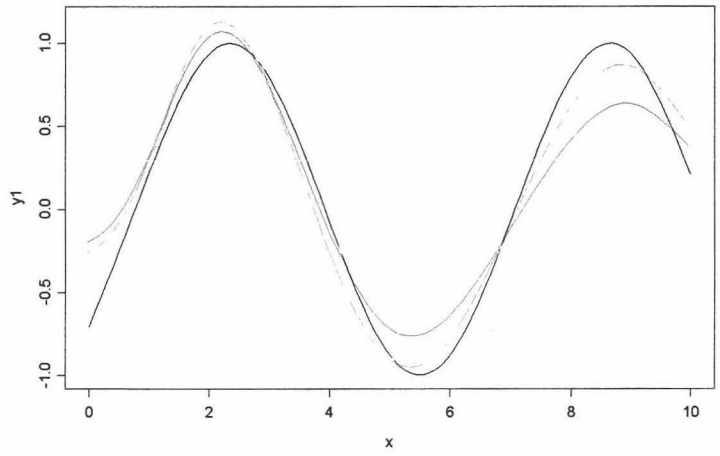


Figura 3.7: Caso funcional. Repetición del Ejemplo 1. Óptimos según ECMB y ECMBI con ACP1. Funciones simulada (Negro) y promedio de las estimadas con los modelos óptimos. Introducción según el Método I (Roja), e introducción según el Método II (Azul).

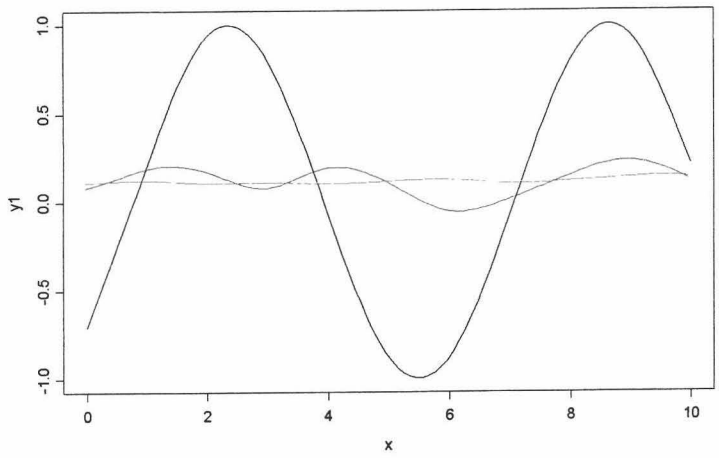


Figura 3.8: Caso funcional. Repetición del Ejemplo 1. Óptimos según ECMB con ACP2. Funciones simulada (Negro) y promedio de las estimadas con los modelos óptimos. Introducción según el Método I (Roja), e introducción según el Método II (Azul).

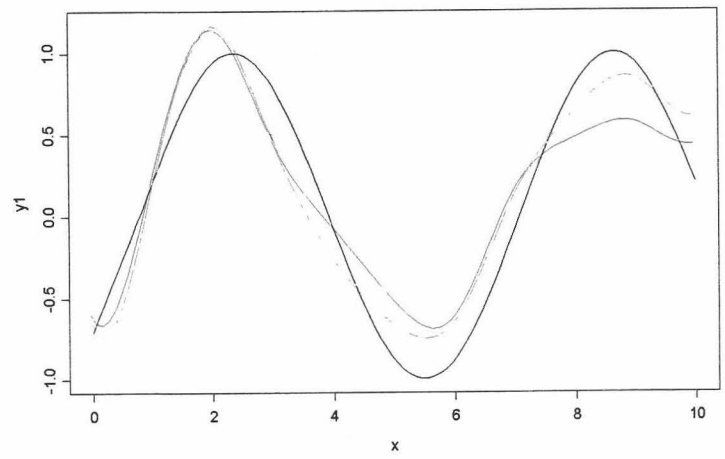


Figura 3.9: Caso funcional. Repetición del Ejemplo 1. Óptimos según ECMBI con ACP2. Funciones simulada (Negro) y promedio de las estimadas con los modelos óptimos. Introducción según los Métodos I (Roja), y II (Azul).

3.2.3 Ejemplo 2. Aproximación de las trayectorias simuladas mediante interpolación spline cúbica

A continuación planteamos una situación muy usual en casos reales en los que generalmente no es posible observar las trayectorias propiamente dichas del proceso bajo estudio y el investigador ha de conformarse con las observación de las mismas en instantes puntuales concretos. Como se indicó al inicio del presente capítulo, en este caso el primer paso es la aproximación de las trayectorias en un espacio de dimensión finita para llevar a cabo el ajuste del modelo de regresión logística funcional, como hemos visto en el ejemplo anterior. Esta aproximación se puede llevar a cabo, como hemos resaltado en repetidas ocasiones, desde una doble perspectiva: a partir de la interpolación de las trayectorias observadas en los nodos discretos si los datos son observados sin error, o bien a partir de una aproximación mínimo cuadrática si los datos han sido observados con error.

En el presente ejemplo vamos a abordar el problema de la interpolación de las trayectorias explicativas del modelo de regresión logística funcional a partir de sus observaciones en nodos discretos, utilizando interpolación spline cúbica, y para ello vamos a considerar la expresión de los spline de interpolación en términos de la base de B-splines que se obtienen a partir de los nodos de observación, de este modo aprovecharemos además dicha expresión para la estimación aproximada del modelo de regresión logística funcional en un espacio de dimesión finita, que en este caso volverá a ser el de los spline cúbicos sobre los nodos de observación.

Como es conocido, cuando se pretende interpolar una función mediante splines cúbicos a partir de un conjunto de nodos discretos, es necesario imponer dos condiciones adicionales para que el problema de interpolación tenga solución única (De Boor, 1978) siendo usual considerar que la segunda derivada de la función que se interpola (que en nuestro caso serán las trayectorias del proceso que explica la variable respuesta) será nula en los extremos de su intervalo de definición de dicha función. Este tipo de interpolación se conoce como interpolación spline cúbica natural. En la práctica, y como veremos en este ejemplo, el hecho de utilizar la solución natural del problema de interpolación spline cúbica hará que no exista solución única para la estimación de los parámetros de la aproximación múltiple del modelo de regresión logística funcional, debido a que dos columnas de la matriz de diseño (las últimas) son combinación lineal de las restantes. Esto hará necesario o bien la elección de algún otro tipo de interpolación spline cúbica, o bien la utilización de análisis

en componentes principales como medio de dar al menos una estimación de dichos parámetros y, por tanto, del modelo de regresión logística funcional.

En primer lugar simularemos una muestra de trayectorias del siguiente proceso considerado en el contexto de la estimación funcional de trimed mean por Fraiman y Muñiz y (2001)

$$X(t) = Z(t) + \frac{t}{4} + 5E \tag{3.6}$$

donde $Z(t)$ es un proceso estocástico Gausiano con función media nula y función de covarainza dada por

$$C(t,s) = \left(\frac{1}{2}\right)^{80|t-s|}$$

y E es una v.a. de Bernouilli de parámetro $p = 0.1$. Para la simulación de las observaciones discretas del anterior proceso tomamos la siguiente partición del intervalo $[0, 12]$ con nodos desigualmente espaciados.

$$\Pi = \{0, 1.1, 2.5, 3.7, 5.1, 7.3, 8.5, 9.6, 12\}$$

y calculamos los valores de la matriz cuadrada de dimensión 9

$$C = (c_{ij});\ c_{ij} = \left(\frac{1}{2}\right)^{80|t_i - t_j|},\ t_i, t_j \in \Pi,\ i, j = 0, \ldots, 8.$$

A partir de esta matriz simulamos una muestra de tamaño 80 de una distribución normal multivariante de dimensión 9 con vector de medias nulo y matriz de varianzas-covarianzas C que denotaremos por (N_{ij}). Seguidamente simulamos una muestra aleatoria de tamaño 80 de valores de la distribución de Bernouilli de probabilidad 0.1: $E_i,\ i = 1, \ldots, 80$. Finalmente, los valores de las observaciones de las 80 trayectorias en los nodos anteriormente definidos se recogen en la matriz $\mathcal{X} = (\mathrm{x}_{ij})$ con valores

$$\mathrm{x}_{ij} = X_i(t_j) = N_{ij} + \frac{t_j}{4} + 5E_i,\ i = 1, \ldots, 80,\ j = 0, \ldots, 8,\ t_j \in \Pi$$

Una vez simulados los valores correspondientes a las observaciones de las trayectorias en los nodos antes definidos, el siguiente paso es la obtención de las trayectorias aproximadas interpolando los valores de cada una de las filas de $\mathcal{X}$. Así, si denotamos por $\widehat{x}_i(t)$ a la interpolación spline cúbica de los valores de la i-ésima fila de $\mathcal{X}$, denotada por x_i^T, dicha función será de la forma

$$\widehat{x}_i(t) = \sum_{j=-1}^{10} a_{ij} B_j(t). \tag{3.7}$$

Para obtener los coeficientes de esta expresión imponemos que $\widehat{x}_i(t_j) = \mathrm{x}_{ij}$, $t_j \in \Pi$, $j = 0, \ldots, 8$ y que la segunda derivada de $\widehat{x}_i(t)$ es nula en los extremos, esto es, que $\widehat{x}_i''(0) = \widehat{x}_i''(12) = 0$, $\forall i$ de manera que dichos coeficientes se obtienen como

$$a_i = B^{-1} x_i$$

donde B es la matriz 11×11 que tiene en cada una de sus primeras 9 filas los valores de las funciones B-spline definidas con la partición Π evaluadas en cada uno de los nodos de observación de Π y en las dos últimas los valores de sus segundas derivadas en el 0 y el 12 respectivamente, x_i es el vector $x_i^T = \left(\mathrm{x}_i^T, 0, 0\right)$ y a_i es el vector de coeficientes de dimensión 11 a determinar.

Repitiendo todo este proceso para cada una de las filas de la matriz $\mathcal{X}$ tendremos la interpolación de cada una de las trayectorias simuladas en los nodos de observación, esto es, los coeficientes de las mismas en términos de los B-splines, que consideraremos por filas en una matriz de dimensión 80×11 que denotaremos por A y que tendrá la forma

$$A = X\left(B^{-1}\right)^T \tag{3.8}$$

donde $X = (\mathcal{X} \mid \mathbf{0}\ \mathbf{0})$ siendo $\mathbf{0}$ un vector de ceros de dimesión 80.

Llegados a este punto, ya estamos en una situación como la descrita en el Ejemplo 1 de esta sección, esto es, con los coeficientes de la expresión de las trayectorias en términos de la base de funciones B-spline obtenidas a partir de Π, que utilizaremos para explicar a la variable respuesta dicotómica en el modelo de regresión logística funcional. En este caso la matriz de productos escalares de los elementos básicos Ψ tiene las integrales en el intervalo $[0, 12]$ en lugar del $[0, 10]$. Como entonces, fijamos una función parámetro $\cos(x - \pi/4)$ considerando su expresión en términos de la base de B-splines como en (3.3), siendo ahora tales coeficientes

α	β_1	β_2	β_3	β_4	β_5
-0.25	-0.994	4.04E–16	0.994	1.074	0.167
β_6	β_7	β_8	β_9	β_{10}	β_{11}
-0.895	-1.129	-0.343	0.824	0.989	1.155

Los coeficientes β han sido obtenidos mediante interpolación spline cúbica natural de la función $\cos(x - \pi/4)$ sobre los nodos de Π. A partir de aquí, simulamos los valores de la variable respuesta igual que detallamos en el Ejemplo 1 y que no repetiremos aquí.

Una vez que disponemos de todos los elementos necesarios, intentamos estimar el modelo de regresión logística funcional mediante el ajuste del modelo múltiple equivalente (2.13), y observamos que las dos últimas columnas de la matriz $A\Psi$ son combinación lineal del resto con lo que no existe solución única al problema de estimación de los parámetros del modelo de regresión logística múltiple indicado. Este hecho está motivado por la elección de la interpolación spline cúbica natural, que hace que las dos últimas columnas de X sean constantes (nulas) y como consecuencia que la matriz de diseño (sin la columna de unos), que es de la forma

$$X\left(B^{-1}\right)^{T}\Psi,$$

no tenga rango completo.

Para evitar esta indeseable situación vamos a cambiar la condición que define a la interpolación spline cúbica natural considerando que las derivadas segundas de las trayectorias en los extremos sean prácticamente cero en lugar de ser exactamente cero y no todas iguales. En definitiva lo que hacemos es obtener una nueva matriz de coeficientes de la expresión de las trayectorias interpoladas en términos de la base de B-splines

$$A^{*} = X^{*}\left(B^{-1}\right)^{T}$$

siendo $X^{*} = (\mathcal{X} \mid u_1\, u_2)$ y u_1, u_2, dos vectores de valores uniformes en el intervalo $[-0.01, 0.01]$, y que hace que la matriz $A^{*}\Psi$ sea de rango completo.

Si a continuación consideramos la funcion parámetro antes definida mediante los coeficientes de su expresión en términos de la base y a partir de ella y la matriz de diseño calculamos las probailidades de (3.5) y seguidamente simulamos valores de una variable con distribución de Bernouilli con dichas probabilidades, tendremos todos los elementos necesarios para el ajuste del modelo logístico múltiple correspondiente y de rango completo. Ajustando dicho modelo logístico obtenemos los siguientes parámetros estimados:

$\widehat{\alpha}$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$
0.229	120517.9	-18087.03	7453.083	-3404.639	1443.072
$\widehat{\beta}_6$	$\widehat{\beta}_7$	$\widehat{\beta}_8$	$\widehat{\beta}_9$	$\widehat{\beta}_{10}$	$\widehat{\beta}_{11}$
-698.5814	-29.557	653.646	-2129.526	6811.884	-34395.72

De este modo obtenemos al menos una estimación de los parámetros de este modelo aunque, como podemos ver, los parámetros estimados difieren mucho

de los reales fijados previamente, hecho que, al igual que hemos visto en situaciones anteriores, se puede deber a la alta multicolinealidad existente en los datos.

Al igual que en el resto de ocasiones en las que la multicolinealidad no permite una estimación precisa de los parámetros de un modelo de regresión logística, vamos a ajustar el modelo en términos de un número adecuado de componentes principales de estas trayectorias aproximadas (las que proporciona la matriz A^*) y a partir de las estimaciones de los parámetros así obtenidos obtendremos una estimación más precisa de los coeficientes y a partir de ellos de la función parámetro.

Tal como se demostró en el Capítulo 2, podemos considerar dos tipos de ACP y dos métodos para la introducción de las componentes principales en el modelo. Las Tablas 3.22, 3.23, 3.24 y 3.25 muestran las medidas resumen que nos ayudan a encontrar el modelo óptimo en el sentido de proporcionar la mejor estimación de la función parámetro. De la observación de dichas tablas se tiene que utilizando ACP1, los valores más pequeños de ECMB se obtienen para el modelo que tiene 5 componentes con el Método I (0.243) y tres con el II (0.253), mientras que los valores más pequeños de ECMBI se tienen para los modelos con cuatro y dos componentes respectivamente. Si por contra nos decidimos por ACP2, los valores mínimos de ECMB son 0.497 para el modelo con 8 componentes y utilizando el Método I y 0.488 con 3 componentes y utilizando el Método II; mientras los de ECMBI son 1.122 del modelo con 5 componentes en el orden dado por el Método I y 1.355 del que tiene 4 en el orden dado por el II.

Según esto, el ACP1 proporciona mayor reducción de dimensión que el 2, y siempre el Método II se comporta mejor en cuanto a reducción de dimensión. Los parámetros estimados de los modelos descritos aquí aparecen en la Tabla 3.21.

Si además de los valores numéricos de las medidas, observamos las gráficas de los modelos que proporcionan las cantidades anteriormente reflejadas, vemos que las funciones parámetro mejor ajustadas son las que proporciona el ACP1 tomando los mínimos de ECMB (Figura 3.11), y que tanto con el Método I como con el II la estimación es la misma prácticamente, sin embargo el Método II reduce más la dimensión. No obstante, si nos decidiésemos por el ACP2, parece claro que las medida más adecuada para la elección del modelo para la estimación lo más precisa posible es el ECMBI, ya que la Figura 3.13 presenta una mejor aproximación a la función simulada que la 3.12 y además se consigue

máyor reducción de dimensión.

En cuanto al resto de medidas de bondad de ajuste: V. Ac., CCR, etc., los resultados que se obtienen son análogos a los vistos en el Ejemplo 1, de aquí que obviemos cualquier comentario acerca de los mismos y nos remitamos a lo visto en aquella ocasión.

Interpolación spline cúbica natural. Solución mediante ACP.

Otra manera de evitar el problema anteriormente citado de la no existencia de solución para los parámetros del modelo logístico aproximado es la utilización de análisis de componentes principales funcional directamente de las trayectorias obtenidas mediante interpolación spline cúbica natural por cualquiera de las dos alternativas citadas en el primer ejemplo. A continuación presentamos otro ejemplo simulado que ilustra el modelo de regresión logística funcional cuando aproximamos las trayectorias del mismo mediante interpolación spline cúbica natural y utilizamos ACP (en las dos alternativas que estamos considerando aquí), para evitar problemas derivados de la multicolinealidad. Mostraremos los resultados que se obtienen al repetir la situación un gran número de veces y obviaremos los resultados para un caso particular.

Consideremos los valores simulados originalmente de las trayectorias en los nodos, del proceso 3.6 y tomemos la aproximación de dichas trayectorias dada por los coeficientes de (3.8) (tomando en X las columnas de ceros). Consideremos la función parámetro fijada al inicio de esta sección y las probabilidades obtenidas a partir de ella y de las trayectorias aproximadas. Simulamos posteriormente 326 vectores de dimesión 80 cada uno con valores de la distribución de Bernouilli de probabilidades las dadas anteriormente, y ajustamos los 326 modelos, obteniendo con cada uno de ellos los correspondientes parámetros estimados y las medidas ya conocidas de bondad de ajuste y de la precisión de las estimaciones. Al igual que en el Ejemplo 1 presentamos como resumen de estos 326 ajustes el promedio de las medidas correspondientes a los modelos considerados como óptimos, siendo tales modelos los que minimizan el ECMB por un lado y los de ECMBI por otro. También presentamos sus varianzas y coeficientes de variación con la intención de evaluar la representatividad de dichos promedios.

De los valores calculados podemos ver en la Tabla 3.26 que las medias del número de componentes que entran en el modelo según el Método I son casi iguales con el ACP1 considerando como modelo óptimo el que proporciona el mínimo del ECMB y el del ECMBI (3.604 y 3.687). Con el resto de medidas

ocurre igual, sin embargo parece que los promedios que proporciona el mínimo del ECMBI son más representativos ya que generalmente las varianzas y coeficientes de variación en este caso son menores (salvo las del ECMB) como por ejemplo para el caso de la variabilidad acumulada que para el ECMB toma una varianza de 150.297 y para el ECMBI 60.564. Este hecho también se observa para el caso de la introducción de componentes por el Método II (Tabla 3.27)

Por otro lado si comparamos las Tablas 3.26 y 3.27 se observa que el número de componentes que entran en el modelo en el segundo caso disminuye ligeramente, pasando de 3.604 a 2.613 con ECMB y de 3.687 a 2.479 con ECMBI, de igual modo que ocurre con el C.V., lo que nos llevaría a concluir que el Método II proporciona mayor reducción de dimensión (en el caso del ACP1) que el Método I. Además, los valores que se obtienen son más homogéneos (tanto para un error como para otro), con lo que claramente las estimaciones son mejores con el método II que con el I.

Consideremos ahora el caso de ACP2 (Tablas 3.28 y 3.29), ahora las medias del número de cc. pp. que entran con el Método I no son iguales como era el caso del ACP1: 3.859 para el mínimo de ECMB y 5.069 para el de ECMBI. Esta tendencia de aumento se repite para el promedio del resto de medidas, excepto para el ECMBI, sin embargo generalmente las variabilidades son inferiores para los promedios de los modelos elegidos con el ECMBI y por tanto las representatividades de estos promedios mejores. Para el Método II no ocurre igual que el I en el ACP2; ahora no se aprecian grandes diferencias ni en promedios ni en variabilidades de modelos seleccionados mediante ECMB y ECMBI. Finalmente si comparamos Método I con II en este caso de ACP2, se aprecia la gran reducción de dimensión que se obtiene de uno a otro, pasando de casi 4 a 2 componentes tomando ECMB y de 5 a casi 3 tomando ECMBI.

Con objeto de decidirnos por el mejor tipo de ACP en cuanto a la estimación de la función parámetro del modelo, y ya que en todos los casos hemos llegado a la conclusión de que la introducción de las componentes en el modelo mediante el Método II es más adecuada que el I, compararemos las Tablas 3.27 y 3.29: haciéndolo con ECMB vemos que si bien el número de componentes que proporcionan los dos tipos son parecidos (2.613 y 2.101), las cantidades de ECMB (0.437 y 0.676) y de ECMBI (1.170 y 3.377) son menores al utilizar ACP1 que al utilizar ACP2. Exactamente igual ocurre con ECMBI.

De la observación de las gráficas podemos ver cómo, generalmente en todos los casos, las funciones que resultan al utilizar los dos métodos de introducción de componentes considerados son parecidas en cuanto a su precisión con

	ACP1				ACP2			
	Método I		Método II		Método I		Método II	
s	4	5	2	3	5	8	4	3
$\widehat{\alpha}_{(s)}$	0.010	-0.094	2.156	0.025	-0.559	-1.382	-0.668	-0.213
$\widehat{\beta}_{1(s)}$	0.052	0.011	0.004	0.010	-1.142	-1.240	-1.126	-0.580
$\widehat{\beta}_{2(s)}$	0.352	0.336	0.160	0.294	-0.044	0.317	-0.018	0.141
$\widehat{\beta}_{3(s)}$	1.097	1.069	0.576	0.956	1.153	2.015	1.191	0.927
$\widehat{\beta}_{4(s)}$	1.217	1.225	0.735	1.204	0.490	-0.483	0.345	0.229
$\widehat{\beta}_{5(s)}$	0.376	0.393	0.2650	0.566	0.209	0.684	0.284	0.506
$\widehat{\beta}_{6(s)}$	-0.980	-1.032	-0.774	-0.937	0.234	-0.611	0.319	0.222
$\widehat{\beta}_{7(s)}$	-1.565	-1.580	-1.240	-1.614	-2.159	-1.796	-2.240	-1.578
$\widehat{\beta}_{8(s)}$	-0.364	-0.305	-0.425	-0.435	0.110	0.137	0.148	0.666
$\widehat{\beta}_{9(s)}$	1.270	1.280	0.811	1.277	1.241	0.940	1.248	0.229
$\widehat{\beta}_{10(s)}$	0.723	0.719	0.474	0.739	0.301	1.076	0.289	0.129
$\widehat{\beta}_{11(s)}$	0.069	0.069	0.044	0.071	-0.846	1.242	-0.881	0.007

Tabla 3.21: Caso funcional. Ejemplo 2. Parámetros reconstruidos de los modelos óptimos con los distintos tipos de ACP (1 y 2), distintos Métodos de selección de cc. pp. (I y II) y distintos criterios de elección del modelo óptimo (ECMB y ECMBI).

respecto a la función simulada, sin embargo al reducirse más la dimesión en el método II, será preferible éste. También se aprecia en las gráficas cómo el ACP1 da mejores aproximaciones que el ACP2 (Figuras 3.14 y 3.17).

Con todo lo visto se tiene que, al igual que en el Ejemplo 1, es clara la mejora que se obtiene al utilizar ACP en la estimación de la función parámetro de un modelo de regresión logística funcional en lugar de las trayectorias propiamente dichas, y más en este ejemplo, en el que pasamos de una situación en la que no se puede dar una estimación de la misma, ni a través de aproximación siquiera, a una en la que no sólo la podemos dar sino que además esta estimación es bastante precisa. Además de poder dar estas estimaciones, vuelve a ponerse de manifiesto que la introducción de las componentes en los distintos modelos en el orden que proporciona el método Stepwise es más adecuada que la introducción por orden de variabilidad, ya sea utilizando ACP1 o ACP2, y que, de preferir alguno de estos, nos quedaríamos con el primero pues las estimaciones que proporciona son más precisas que el segundo.

s	V. Ac.	ECMP	CCR	ECMBI	ECMB	Var	G^2	p-val
1	54.76	0.103	68.75	5.909	0.689	0.100	92.95	0.12
2	66.05	0.060	73.75	3.798	0.753	0.178	76.94	0.48
3	75.58	0.004	87.50	0.467	0.253	1.277	45.67	1.00
4	83.01	0.003	83.75	**0.412**	0.244	1.429	45.23	1.00
5	89.36	0.003	83.75	0.434	**0.243**	1.704	45.20	1.00
6	94.28	0.004	85.00	0.691	0.342	2.529	42.38	1.00
7	97.17	0.004	85.00	0.794	0.397	3.863	42.03	1.00
8	98.84	0.004	88.75	0.768	0.353	7.616	41.61	1.00
9	99.99	0.010	90.00	3.158	2.732	23.187	39.43	1.00
10	100.00	0.015	90.00	1.6E+6	1.4E+8	24.831	36.64	1.00
11	100.00	0.018	91.25	1.5E+7	1.4E+9	24.659	35.29	1.00

Tabla 3.22: Caso funcional. Ejemplo 2: interpolación cuasi-natural. Medidas de bondad de ajuste para los modelos con distinto número de cc.pp. como regresores. Componentes obtenidas con el ACP1. Introducción de componentes según el Método I: orden de variabilidad.

s	cc. pp.	ECMP	CCR	ECMBI	ECMB	Var	G^2	p-val
1	3	0.082	73.75	3.637	1.632	0.162	82.56	0.34
2	2	0.026	83.75	**0.427**	0.720	0.484	59.87	0.93
3	1	0.004	87.50	0.467	**0.253**	1.277	45.67	1.00

Tabla 3.23: Caso funcional. Ejemplo 2: interpolación cuasi-natural. Medidas de bondad de ajuste para los modelos con distinto número de cc.pp. como regresores. Componentes obtenidas con el ACP1. Introducción de componentes según el Método II: orden que proporciona el método Stepwise.

s	V. Ac.	ECMP	CCR	ECMBI	ECMB	Var	G^2	p-val
1	79.28	0.101	70.00	5.823	0.677	0.122	90.99	0.15
2	84.99	0.085	68.75	5.150	0.742	0.206	84.35	0.27
3	89.03	0.057	73.75	3.431	0.708	0.419	74.14	0.54
4	92.71	0.023	85.00	1.355	0.712	1.881	54.08	0.97
5	95.42	0.019	85.00	**1.122**	0.640	2.197	53.93	0.96
6	97.11	0.021	86.25	1.398	0.755	2.636	52.79	0.96
7	98.65	0.023	85.00	1.328	0.873	4.242	50.30	0.98
8	99.44	0.021	87.50	1.342	**0.497**	12.754	49.43	0.98
9	100.00	0.023	87.50	1.890	1.347	32.736	48.65	0.98
10	100.00	0.024	86.25	1.2E+29	1.1E+31	32.067	48.44	0.97
11	100.00	0.398	40.00	3.1E+30	2.7E+32	147.193	41.43	1.00

Tabla 3.24: Caso funcional. Ejemplo 2: interpolación cuasi-natural. Medidas de bondad de ajuste para los modelos con distinto número de cc.pp. como regresores. Componentes obtenidas con el ACP2. Introducción de componentes según el Método I: orden de variabilidad.

s	cc. pp	ECMP	CCR	ECMBI	ECMB	Var	G^2	p-val
1	4	0.102	72.50	4.802	0.992	0.331	89.10	0.18
2	1	0.081	75.00	4.452	0.670	0.843	79.16	0.41
3	3	0.046	78.75	2.517	**0.488**	1.247	67.28	0.75
4	2	0.023	85.00	**1.355**	0.712	1.881	54.08	0.97

Tabla 3.25: Caso funcional. Ejemplo 2: interpolación cuasi-natural. Medidas de bondad de ajuste para los modelos con distinto número de cc.pp. como regresores. Componentes obtenidas con el ACP2. Introducción de componentes según el Método II: orden que proporciona el método Stepwise.

	ECMB			ECMBI		
Medidas	Media	Varianza	C.V	Media	Varianza	C.V
N. cc. pp.	**3.604**	2.548	**0.443**	**3.687**	1.434	**0.325**
V. Ac.	78.308	**150.297**	0.157	79.854	**60.564**	0.097
ECMP	0.026	0.001	1.355	0.015	0.0003	1.144
CCR	82.956	49.959	0.085	85.161	18.801	0.051
ECMBI	1.714	3.874	1.148	1.124	1.205	0.977
ECMB	0.423	0.031	0.415	0.508	0.122	0.688

Tabla 3.26: Caso funcional. Repetición del Ejemplo 2. Promedios y variabilidades de distintas medidas de bondad de ajuste obtenidas de los modelos óptimos, siendo tales aquéllos con el número de componentes correspondiente con los valores mínimos de ECMB y de ECMBI. Componentes obtenidas con el ACP1. Introducción de componentes según el Método I: orden de variabilidad.

	ECMB			ECMBI		
Medidas	Media	Varianza	C.V.	Media	Varianza	C.V.
N. cc. pp.	**2.613**	0.435	**0.252**	**2.479**	0.2688	**0.209**
ECMP	0.020	0.0004	0.994	0.020	0.0002	0.617
CCR	84.551	20.918	0.054	84.651	9.7096	0.037
ECMBI	**1.170**	1.973	1.200	0.539	0.1633	0.749
ECMB	**0.437**	0.064	0.581	0.581	0.2062	0.782

Tabla 3.27: Caso funcional. Repetición del Ejemplo 2. Promedios y variabilidades de distintas medidas de bondad de ajuste obtenidas de los modelos óptimos, siendo tales aquéllos con el número de componentes correspondiente con los valores mínimos de ECMB y de ECMBI. Componentes obtenidas con el ACP1. Introducción de componentes según el Método II: orden proporcionado por el método Stepwise.

	ECMB			ECMBI		
Medidas	Media	Varianza	C.V.	Media	Varianza	C.V.
N. cc. pp.	**3.859**	7.605	0.715	**5.061**	1.725	0.260
V. Ac.	89.130	79.078	0.100	94.935	6.855	0.028
ECMP	0.056	0.001	0.585	0.034	0.0002	0.458
CCR	78.102	89.997	0.121	84.674	14.908	0.046
ECMBI	3.156	5.916	0.771	1.320	0.658	0.614
ECMB	0.665	0.055	0.352	0.902	0.284	0.590

Tabla 3.28: Caso funcional. Repetición del Ejemplo 2. Promedios y variabilidades de distintas medidas de bondad de ajuste obtenidas de los modelos óptimos, siendo tales aquéllos con el número de componentes correspondiente con los valores mínimos de ECMB y de ECMBI. Componentes obtenidas con el ACP2. Introducción de componentes según el Método I: orden de variabilidad.

	ECMB			ECMBI		
Medidas	Media	Varianza	C.V.	Media	Varianza	C.V.
N. cc. pp.	**2.101**	1.870	0.651	**2.954**	2.062	0.486
ECMP	0.069	0.0005	0.315	0.058	0.0005	0.374
CCR	75.598	56.064	0.100	79.804	49.755	0.088
ECMBI	**3.377**	2.229	0.442	2.536	1.749	0.521
ECMB	**0.676**	0.026	0.237	0.870	0.189	0.500

Tabla 3.29: Caso funcional. Repetición del Ejemplo 2. Promedios y variabilidades de distintas medidas de bondad de ajuste obtenidas de los modelos óptimos, siendo tales aquéllos con el número de componentes correspondiente con los valores mínimos de ECMB y de ECMBI. Componentes obtenidas con el ACP2. Introducción de componentes según el Método II: orden proporcionado por el método Stepwise.

	ACP1				ACP2			
	ECMB		ECMBI		ECMB		ECMBI	
Par.	I	II	I	II	I	II	I	II
$\widehat{\alpha}_{(s)}$	-0.572	0.110	-0.760	0.766	-0.335	0.718	-0.870	0.160
$\widehat{\beta}_{1(s)}$	0.007	0.007	0.007	0.005	-0.578	-0.341	-1.043	-0.620
$\widehat{\beta}_{2(s)}$	0.230	0.223	0.256	0.198	-0.044	-0.062	-0.115	-0.077
$\widehat{\beta}_{3(s)}$	0.793	0.757	0.881	0.685	0.540	0.242	0.896	0.514
$\widehat{\beta}_{4(s)}$	0.948	0.952	1.069	0.863	0.832	0.471	1.088	0.754
$\widehat{\beta}_{5(s)}$	0.290	0.404	0.370	0.349	0.072	0.328	0.254	0.317
$\widehat{\beta}_{6(s)}$	-0.717	-0.800	-0.821	-0.786	-0.399	-0.507	-0.493	-0.491
$\widehat{\beta}_{7(s)}$	-1.133	-1.348	-1.336	-1.308	-0.708	-0.572	-1.541	-1.003
$\widehat{\beta}_{8(s)}$	-0.296	-0.381	-0.350	-0.399	-0.134	-0.022	-0.157	0.007
$\widehat{\beta}_{9(s)}$	1.052	1.124	1.231	1.024	0.655	0.424	1.293	0.709
$\widehat{\beta}_{10(s)}$	0.629	0.657	0.737	0.599	0.369	0.154	0.476	0.253
$\widehat{\beta}_{11(s)}$	0.062	0.063	0.073	0.057	0.020	-0.177	-0.522	-0.304

Tabla 3.30: Caso funcional. Ejemplo 2. Parámetros promedio de los reconstruidos en cada modelo óptimo con los distintos tipos de ACP (1 y 2), distintos Métodos de selección de cc. pp. (I y II) y distintos criterios de elección del modelo óptimo (ECMB y ECMBI)

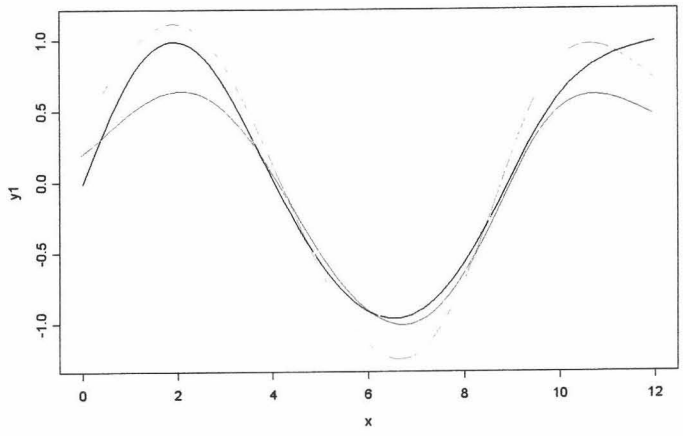


Figura 3.10: Caso funcional. Ejemplo 2. Mínimos de ECMBI y ACP1. Funciones simulada (Negro) y estimadas utilizando cc. pp. en la aproximación del modelo de regresión logística funcional: modelo con 4 componentes e introducción según el Método I (Roja), y modelo con 2 componentes e introducción según el Método II (Azul).

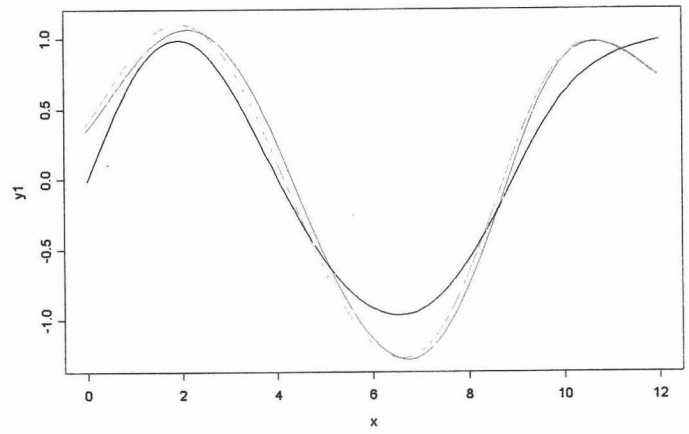


Figura 3.11: Caso funcional. Ejemplo 2. Mínimos de ECMB y ACP1. Funciones simulada (Negro) y estimadas utilizando cc. pp. en la aproximación del modelo de regresión logística funcional: modelo con 5 componentes e introducción según el Método I (Roja), y modelo con 3 componentes e introducción según el Método II (Azul).

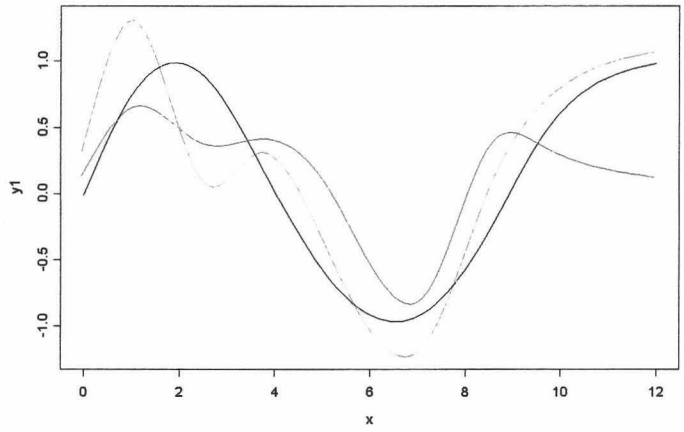


Figura 3.12: Caso funcional. Ejemplo 2. Mínimos de ECMB y ACP2. Funciones simulada (Negro) y estimadas utilizando cc. pp. en la aproximación del modelo de regresión logística funcional: modelo con 8 componentes e introducción según el Método I (Roja), y modelo con 3 componentes e introducción según el Método II (Azul).

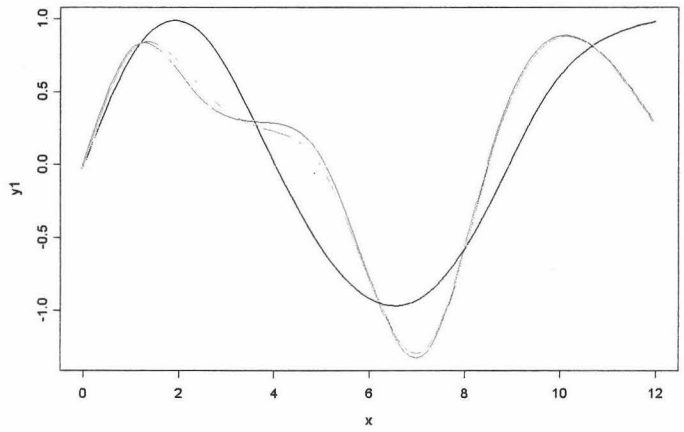


Figura 3.13: Caso funcional. Ejemplo 2. Mínimos de ECMBI y ACP2. Funciones simulada (Negro) y estimadas utilizando cc. pp. en la aproximación del modelo de regresión logística funcional: modelo con 5 componentes e introducción según el Método I (Roja), y modelo con 4 componentes e introducción según el Método II (Azul).

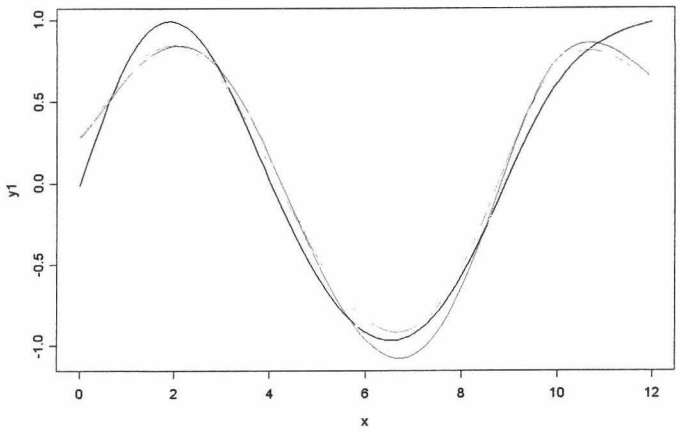


Figura 3.14: Caso funcional. Repetición del Ejemplo 2. Óptimos según ECMB con ACP1. Funciones simulada (Negro) y promedio de las estimadas con los modelos óptimos. Introducción según el Método I (Roja), e introducción según el Método II (Azul).

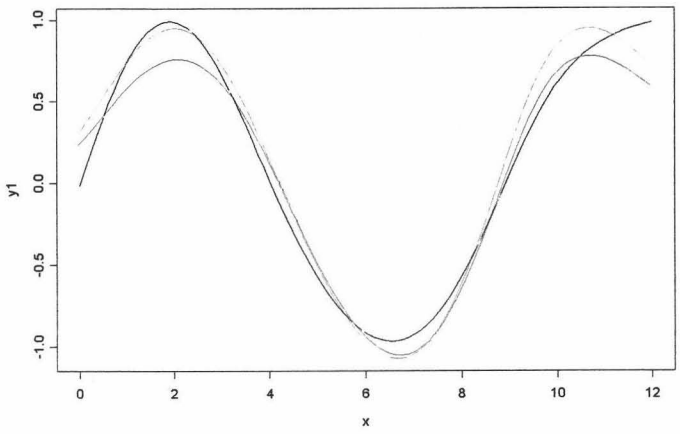


Figura 3.15: Caso funcional. Repetición del Ejemplo 2. Óptimos según ECM-BI con ACP1. Funciones simulada (Negro) y promedio de las estimadas con los modelos óptimos. Introducción según el Método I (Roja), e introducción según el Método II (Azul).

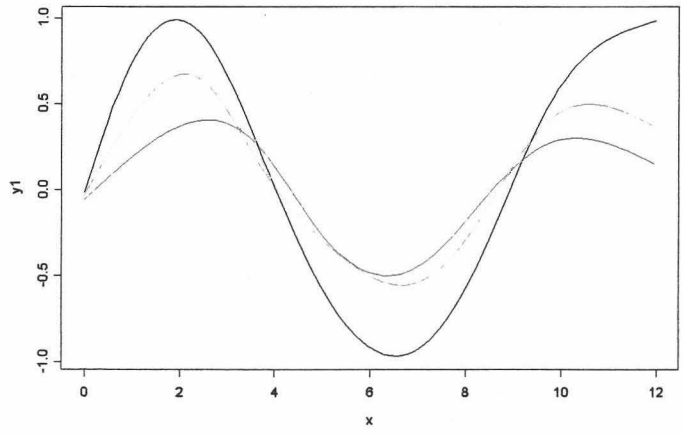


Figura 3.16: Caso funcional. Repetición del Ejemplo 2. Óptimos según ECMB con ACP2. Funciones simulada (Negro) y promedio de las estimadas con los modelos óptimos. Introducción según el Método I (Roja), e introducción según el Método II (Azul).

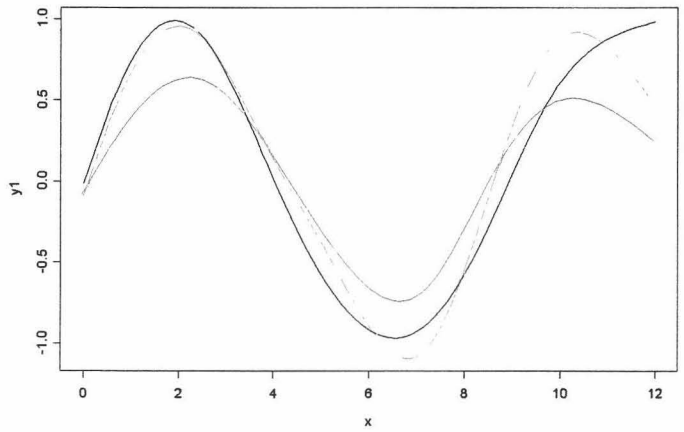


Figura 3.17: Caso funcional. Repetición del Ejemplo 1. Óptimos según ECM-BI con ACP2. Funciones simulada (Negro) y promedio de las estimadas con los modelos óptimos. Introducción según el Método I (Roja), e introducción según el Método II (Azul).

3.2.4 Ejemplo 3: Aproximación mínimo cuadrática de las trayectorias simuladas

Al igual que en el caso anterior, en este ejemplo abordamos la problemática de la aproximación de las trayectorias del modelo de regresión logística funcional, a partir de observaciones de las mismas en instantes discretos. A diferencia de entonces, ahora consideraremos que las observaciones recogidas lo han sido asumiendo un determinado error, con lo que no será apropiado utilizar un método de interpolación para la aproximación anteriormente citada.

En este caso, al igual que en los anteriores, también asumiremos que las trayectorias del modelo se pueden aproximar en un espacio de dimensión finita generado por una base de funciones, siendo diferente el método de aproximación de las mismas por la consideración de la existencia de un error en la recogida de información. También, como en el resto de ejemplos, consideraremos ahora como espacio el generado por una base de funciones B-spline generada a partir de un conjunto de nodos. A diferencia del caso de aproximación mediante interpolación spline cúbica, ahora los nodos que definen a las funciones B-spline no tienen por qué ser los mismos que los nodos de observación, de ahí que hablemos a partir de ahora de lo que llamaremos "nodos de definición", haciendo referencia a los nodos necesarios para definir los B-splines, y "nodos de observación" que serán aquellos en los que tomamos las observaciones en nodos discretos de las trayectorias a aproximar. De hecho, a efectos prácticos, en el ejemplo que nos ocupa, consideraremos como nodos de definición una parte escogida de entre los nodos de observación.

La técnica de aproximación mínimo cuadrática, expuesta en el Capítulo 2 de esta memoria, ha sido tratada por Ramsay & Silvermann (1997) para reconstruir observaciones de tipo meteorológico, en aquella ocasión utilizando una base formada por funciones trigonométricas debido a la estacionalidad de este tipo de datos; nosotros consideraremos la base de funciones B-spline con la intención de aproximar trayectorias regulares.

Para analizar esta situación en el ambiente en que nos movemos del modelo de regresión logística funcional vamos a tomar un conjunto de 21 observaciones discretas en los nodos igualmente espaciados que forman la partición del intervalor $[0, 10]$, de 100 trayectorias simuladas del proceso estocástico que analizamos en el ejemplo anterior y que está dado por la expresión (3.6), observaciones que se han obtenido de igual manera que entonces y que resumiremos en la matriz que denotaremos por X de dimensión 100×21. Seguidamente consideramos la base de funciones B-spline que se obtienen a través de los

nodos que forman la partición del intervalo anterior de tamaño 11 con puntos igualmente espaciados, esto es, hemos considerado una parte de los instantes de observación como instantes de definición, aunque podríamos haber seleccionado cualesquiera otros nodos.

Cabe destacar, a efectos prácticos, la importancia de que la distancia entre los nodos de definición de los B-splines sea mínimamente grande (mayor que 1) cuando se consideran nodos igualmente espaciados, ya que si dichas distancias son menores, la matriz de productos escalares de los elementos básico, Ψ, que en este caso involucran a dichas distancias (Aguilera, 1993), toma valores que en ocasiones son demasiado pequeños y que pueden hacer que al calcular inversas de matrices que involucren a ésta se obtengan resultados computacionalmente inestables.

Tal y como se indicó en el Capítulo 2, la matriz que tiene por filas los coeficientes de la expresión de las trayectorias aproximadas en términos de la base de B-splines, y que seguiremos denotando por A, se obtiene como

$$A = XB\left[\left(B^TB\right)^{-1}\right]^T$$

siendo B la matriz que tiene por columnas los valores de cada B-spline básico en los nodos de observación. Una vez obtenidos los coeficientes de la expresión (3.7), el cálculo de la matriz de diseño del modelo múltiple correspondiente, así como de los valores de la variable respuesta es análoga a la vista en los ejemplos anteriores; y para ello consideramos como función parámetro $sen\,(x+\pi/4)$ o más concretamente la función que se obtiene al interpolar de manera natural los valores obtenidos al evaluar esta función en los nodos de definición. Como en ocasiones anteriores, las estimaciones que se obtienen de la función parámetro a través de la aproximación considerada no es muy precisa por lo que volvemos a utilizar las componentes principales de las trayectorias para mejorar dichas estimaciones.

Con objeto de analizar el modelo que nos ocupa con este tipo de aproximación, hemos simulado 500 vectores de dimensión 100 cada uno de la variable respuesta, y hemos ajustado los correspondientes modelos en términos de las componentes principales y reconstruido los coeficientes de la función parámetro con los distintos métodos que estamos considerando a lo largo del presente capítulo (ACP1 y ACP2, introducción de componentes según los Métodos I y II, y elección del modelo óptimo en cada repetición mediante los valores mínimos de ECMB y ECMBI). Las medidas resumen de esta situación aparecen en las Tablas 3.31, 3.32, 3.33, y 3.34. En ellas lo primero que se observa es

relativo al número medio de componentes que entran en el modelo y es que, análogamente a lo que ocurría en el Ejemplo 1, se aprecia una reducción de dimensión al utilizar las componentes principales, siendo menor al utilizar la introducción de componentes que proporciona el Método II que al usar el I, como se ve al comparar las Tablas 3.31 y 3.32 en las que el número promedio de componentes en la primera es 4.592 y en la segunda 3.298 al utilizar el mínimo de ECMB como criterio de elección de los modelos óptimos. También ocurre igual si comparamos en las mismas tablas las mismas cantidades para el mínimo de ECMBI con los valores 4.496 y 3.550 respectivamente para el número de componentes con los métodos I y II; lo que pone también de manifiesto que al utilizar ACP1, que es el correspondiente a estas dos tablas, la elección de los modelos óptimos es equivalente con las dos medidas que estamos considerando.

Esto no ocurre así con el ACP2 como se ve en las Tablas 3.33, y 3.34, ya que a pesar de que los números promedios de componentes tomando ECMB es de 1.34 y 1.23 para los Métodos I y II respectivamente, la Figura 3.21 y los valores de ECMB (0.854 y 2.035) muestran que las estimaciones en este caso no son muy precisas; y tomando ECMBI ocurre más o menos igual, que las estimaciones son muy malas a pesar de que parezca que existe gran reducción de dimensión.

Todo esto nos lleva a concluir que parece más adecuado, en el caso de aproximación, la utilización de ACP1 que el 2 para la estimación precisa de la función parámetro y, en todo caso, elegir el método II de introducción de componentes porque reduce algo más la dimensión que el I siendo la estimación que se obtiene más o menos igual de precisa. Finalmente decir que en cuanto al resto de medidas, como el ECMP, no se aprecian grandes diferencias. Quizás parezca que los valores para ACP1 son menores que para ACP2, en cada una de las comparaciones posibles: 0.015 y 0.077 para ECMB y Método I o 0.023 y 0.047 para ECMB y Método II, manteniendo las varianzas más o menos equivalentes. Por otro lado las variabilidades explicadas (con el Método I) son muy similares en uno y otro ACP para el mínimo de ECMB, rondando el 80%, no así con el de ECMBI, en el que se tiene un 93% de variabilidad explicada con ACP2 frente al 78% del 1, lo que vendría a indicar que como medida de decisión, el ECMBI parece más adecuada cuando utilizamos ACP2.

	ECMB			ECMBI		
Medidas	Media	Varianza	CV	Media	Varianza	CV
N. cc. pp.	**4.592**	2.622	0.353	**4.496**	1.100	0.233
V. Ac.	78.305	78.563	0.113	**78.550**	27.614	0.067
ECMP	**0.015**	3.172e-4	1.176	0.011	6.467e-5	0.736
CCR	78.828	23.008	0.061	79.562	13.285	0.046
ECMBI	1.140	1.310	1.004	0.907	0.489	0.771
ECMB	0.411	0.058	0.585	0.446	0.113	0.754

Tabla 3.31: Caso funcional. Repetición del Ejemplo 3. Promedios y variabilidades de distintas medidas de bondad de ajuste obtenidas de los modelos óptimos, siendo tales aquéllos con el número de componentes correspondiente con los valores mínimos de ECMB y de ECMBI. Componentes obtenidas con el ACP1. Introducción de componentes según el Método I: orden de variabilidad.

	ECMB			ECMBI		
Medidas	Media	Varianza	CV	Media	Varianza	CV
N. cc. pp.	**3.298**	0.943	0.294	**3.550**	0.412	0.181
ECMP	**0.023**	0.0004	0.894	0.017	0.0002	0.715
CCR	77.774	20.684	0.058	78.824	12.823	0.045
ECMBI	1.270	1.183	0.856	0.940	0.549	0.788
ECMB	0.380	0.029	0.449	0.482	0.127	0.738

Tabla 3.32: Caso funcional. Repetición del Ejemplo 3. Promedios y variabilidades de distintas medidas de bondad de ajuste obtenidas de los modelos óptimos, siendo tales aquéllos con el número de componentes correspondiente con los valores mínimos de ECMB y de ECMBI. Componentes obtenidas con el ACP1. Introducción de componentes según el Método II: orden que proporciona el método Stepwise.

	ECMB			ECMBI		
Medidas	Media	Varianza	CV	Media	Varianza	CV
N. cc. pp.	**1.340**	1.676	0.966	**7.472**	6.903	0.352
V. Ac.	79.443	11.110	0.042	**93.810**	22.026	0.050
ECMP	**0.077**	0.0002	0.191	0.020	6.507e-5	0.395
CCR	64.200	23.772	0.076	79.016	13.254	0.046
ECMBI	4.778	0.533	0.153	1.743	0.329	0.329
ECMB	**0.854**	0.016	0.150	11.658	138.080	1.008

Tabla 3.33: Caso funcional. Repetición del Ejemplo 3. Promedios y variabilidades de distintas medidas de bondad de ajuste obtenidas de los modelos óptimos, siendo tales aquéllos con el número de componentes correspondiente con los valores mínimos de ECMB y de ECMBI. Componentes obtenidas con el ACP2. Introducción de componentes según el Método I: orden de variabilidad.

	ECMB			ECMBI		
Medidas	Media	Varianza	CV	Media	Varianza	CV
N. cc. pp.	**1.230**	0.538	0.596	**3.192**	1.350	0.364
ECMP	**0.047**	6.558e-5	0.171	0.030	1.128e-4	0.359
CCR	71.792	19.684	0.062	77.298	13.805	0.048
ECMBI	2.852	0.175	0.147	2.075	0.245	0.239
ECMB	**2.035**	3.496	0.919	12.273	0.013	0.913

Tabla 3.34: Caso funcional. Repetición del Ejemplo 3. Promedios y variabilidades de distintas medidas de bondad de ajuste obtenidas de los modelos óptimos, siendo tales aquéllos con el número de componentes correspondiente con los valores mínimos de ECMB y de ECMBI. Componentes obtenidas con el ACP2. Introducción de componentes según el Método II: orden que proporciona el método stepwise.

	ACP1				ACP2			
	ECMB		ECMBI		ECMB		ECMBI	
Par.	Met. I	Met. II	Met. I	Met. II	Met. I	Met. II	Met. I	Met. II
$\widehat{\alpha}_{(s)}$	-0.015	0.478	-0.148	0.393	-0.568	0.067	-0.308	-0.253
$\widehat{\beta}_{1(s)}$	0.037	0.030	0.039	0.035	0.191	2.470	-2.873	-1.921
$\widehat{\beta}_{2(s)}$	0.622	0.520	0.658	0.586	0.049	-0.981	0.631	0.071
$\widehat{\beta}_{3(s)}$	1.219	1.009	1.267	1.125	0.173	1.377	1.555	1.628
$\widehat{\beta}_{4(s)}$	0.343	0.212	0.332	0.254	0.040	-0.783	-0.421	-0.697
$\widehat{\beta}_{5(s)}$	-0.465	-0.449	-0.470	-0.474	0.079	0.089	0.050	0.195
$\widehat{\beta}_{6(s)}$	-1.061	-1.005	-1.100	-1.087	0.030	-1.028	-1.269	-1.511
$\widehat{\beta}_{7(s)}$	-0.631	-0.699	-0.709	-0.761	0.058	-0.080	-0.428	-0.175
$\widehat{\beta}_{8(s)}$	0.375	0.342	0.437	0.388	0.064	-0.377	0.059	-0.034
$\widehat{\beta}_{9(s)}$	1.344	1.383	1.532	1.546	0.129	0.435	1.507	1.197
$\widehat{\beta}_{10(s)}$	0.789	0.710	0.820	0.772	0.149	1.075	0.628	0.357
$\widehat{\beta}_{11(s)}$	-0.637	-0.702	-0.712	-0.769	-0.001	-0.548	-0.258	0.533
$\widehat{\beta}_{12(s)}$	-0.447	-0.454	-0.460	-0.516	0.149	0.866	-1.075	-1.590
$\widehat{\beta}_{13(s)}$	-0.022	-0.022	-0.021	-0.027	-0.027	-2.700	5.712	6.237

Tabla 3.35: Caso funcional. Ejemplo 3. Parámetros promedio de los reconstruidos en cada modelo óptimo con los distintos tipos de ACP (1 y 2), distintos Métodos de selección de cc. pp. (I y II) y distintos criterios de elección del modelo óptimo (ECMB y ECMBI)

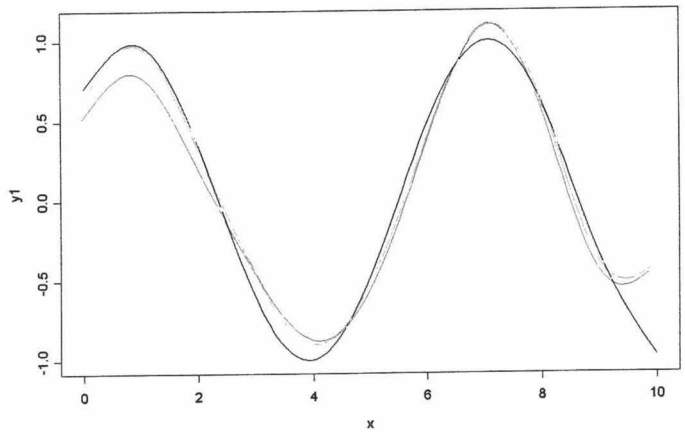


Figura 3.18: Caso funcional. Repetición del Ejemplo 3. Óptimos según ECMB con ACP1. Funciones simulada (Negro) y promedio de las estimadas con los modelos óptimos. Introducción según el Método I (Roja), e introducción según el Método II (Azul).

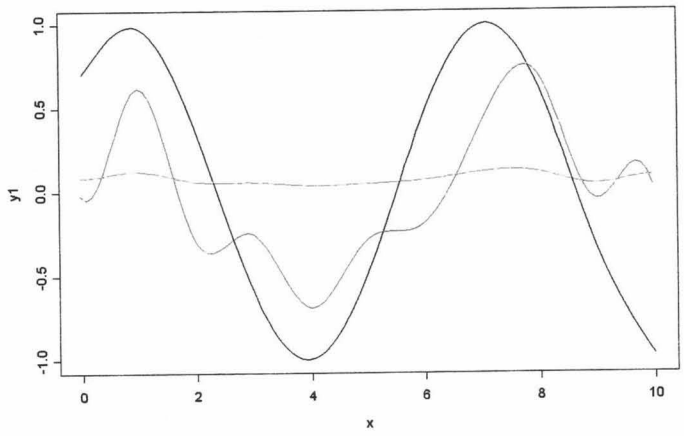


Figura 3.19: Caso funcional. Repetición del Ejemplo 3. Óptimos según ECMB con ACP1. Funciones simulada (Negro) y promedio de las estimadas con los modelos óptimos. Introducción según el Método I (Roja), e introducción según el Método II (Azul).

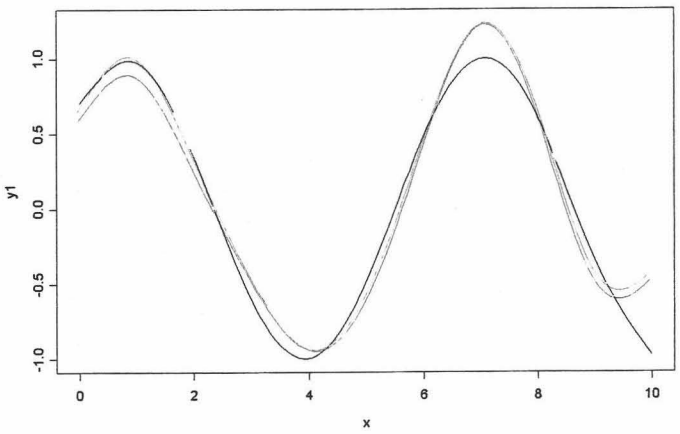


Figura 3.20: Caso funcional. Repetición del Ejemplo 3. Óptimos según ECM-BI con ACP1. Funciones simulada (Negro) y promedio de las estimadas con los modelos óptimos. Introducción según el Método I (Roja), e introducción según el Método II (Azul).

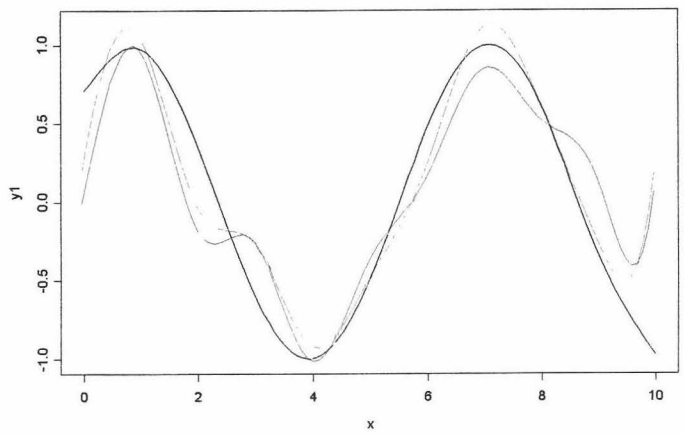


Figura 3.21: Caso funcional. Repetición del Ejemplo 3. Óptimos según ECM-BI con ACP2. Funciones simulada (Negro) y promedio de las estimadas con los modelos óptimos. Introducción según el Método I (Roja), e introducción según el Método II (Azul).

Para finalizar, y como resumen de lo obtenido con los ejemplos simulados, se tiene que en situaciones en las que queremos predecir una variable respuesta dicotómica a partir de información de naturaleza continua, el modelo logístico funcional propuesto es una alternativa acertada. Para su aplicación, es necesaria cierta aproximación múltiple debido inicialmente a la no existencia de solución única (Ramsay & Silverman, 1997) y posteriormente a la disponibilidad, o mejor dicho, no disponibilidad de información muestral de manera continua. Dichas aproximaciones no proporcionan estimaciones muy adecuadas debido, fundamentalmente, a los problemas que provocan la existencia de multicolinealidad en los datos aproximados, hecho que resulta muy común cuando discretizamos situaciones funcionales. El uso de análisis de componentes principales responde bien a las dos cuestiones aquí presentadas: la problemática de la obtención de una estimación de la función parámetro y la de evitar los problemas indeseables de la multicolinealidad. En cuanto a la utilización de ACP, podemos considerar varias alternativas, todas ellas equivalentes, y que proporcionan distintas soluciones; de hecho, considerando las dos tratadas en esta memoria resulta superior utilizar lo que llamamos ACP1 que no es más que realizar un ACP múltiple de los datos aproximados. Finalmente el orden en el que se van introduciendo las componentes en los distintos modelos es un aspecto que también influye, como hemos visto, en la estimación de la función parámetro y en la reducción de dimensión de un problema como éste, siendo más recomendable la introducción de las componentes en el orden que proporciona el método Stepwise que la introducción en orden de variabilidad.

Apéndice

Método de Newton-Raphson

Consideremos el problema de hallar una solución $s = (s_1, s_2, \ldots, s_n)$ de un sistema de n ecuaciones con n incógnitas

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0 \\ f_2(x_1, x_2, \ldots, x_n) &= 0 \\ &\cdots\cdots\cdots\cdots \\ f_n(x_1, x_2, \ldots, x_n) &= 0 \end{aligned}$$

donde $f_1, f_2, \ldots, f_n$ son funciones reales de n variables reales.

El problema se puede plantear en términos vectoriales. Si denotamos por x al vector de $\mathbb{R}^n$ de componentes $x_1, x_2, \ldots, x_n$ y por $f(x)$ al de componentes $f_1(x_1, x_2, \ldots, x_n), f_2(x_1, x_2, \ldots, x_n), \ldots, f_n(x_1, x_2, \ldots, x_n)$, la expresión anterior puede escribirse en forma de ecuación vectorial como

$$f(x) = 0$$

La resolución de sistemas (y de ecuaciones) de manera iterativa se basa en transformar la ecuación vectorial anterior en una de la forma

$$x = g(x)$$

equivalente a ella. Es decir,

$$\begin{aligned} x_1 &= g_1(x_1, x_2, \ldots, x_n) \\ x_2 &= g_2(x_1, x_2, \ldots, x_n) \\ &\cdots\cdots\cdots\cdots \\ x_n &= g_n(x_1, x_2, \ldots, x_n) \end{aligned}$$

Para su resolución aproximada se obtiene una sucesión de vectores $\left\{x^{(m)}\right\}$ en la que se parte de un $x^{(0)}$ (arbitrario en algunos casos o próximo a la solución s en otros) y luego se van formando

$$x^{(m+1)} = g\left(x^{(m)}\right)$$

Muchos métodos iterativos consisten en tomar como función g la siguiente:

$$g(x) = x - A(x) f(x)$$

donde $A(x)$ es una matriz cuadrada de orden n no singular para cualquier valor de x. Esto da lugar a la solución aproximada

$$x^{(m+1)} = x^{(m)} - A\left(x^{(m)}\right) f\left(x^{(m)}\right)$$

La forma particular que tenga $A(x)$ es lo que da lugar a los distintos métodos.

El método de Newton-Raphson consiste en tomar

$$A(x) = J^{-1}(x)$$

donde $J(x)$ es la matriz jacobiana de f.

Para que no surjan dificultades en la aplicación del método habrá que suponer que $J(s)$ es no singular y que las derivadas parciales de f son continuas, para que por esa continuidad exista $J^{-1}(x)$ en un entorno de s. Si además f es de clase C^2 en un intervalo que contenga a s resulta la convergencia del método para todo $x^{(0)}$ de un entorno de s.

Bibliografía

Abellanas, L. y Galindo, A. (1987). *Espacios de Hilbert (geometría, operadores, espectros).* Eudema.

Agresti, A. (1990). *Categorical data analysis.* Wiley.

Agresti, A. (1996). *An introduction to categorical data analysis.* Wiley.

Aguilera, A.M. (1993). *Métodos de Aproximación de Estimadores en el ACP de un Proceso Estocástico.* Tesis Doctoral, Universidad de Granada.

Aguilera, A.M. (2001). *Tablas de contingencia bidimensionales.* Hespérides-La Muralla.

Aguilera, A.M. y Escabias, M. (2000a). Principal component logistic regression. *Proceedings in Computational Statistics 2000, 175-180.* Editado por J.G. Bethlehem y P.G.M. van der Heijden, Physica-Verlag.

Aguilera, A.M. y Escabias, M. (2000b). Reducción de dimensión en regresión logística. *Actas del XXV Congreso Nacional de Estadística e I.O.*, 275-276, Servicio de Publicaciones de la Universidad de Vigo.

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), 57-77.*

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. y Valderrama, M.J. (1996a). Approximation of estimators in the PCA of a stochastic process using B-slines. *Communications in Statistics (Simulation), 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. *Qüestiió, 20 (1), 7-28.*

Aguilera, A.M., Ocaña, F.A. y Valderrama, M.J. (1996c). On a weighted principal component model to forecast a continuous time series. *Proceedings in Computational Statistics 1996, 169-174.* Editado por A. Prat, Physica-Verlag.

Aguilera, A.M., Ocaña, F.A. y Valderrama, M.J. (1997a). An appoximate principal component prediction model for continuous-time stochastic processes. *Applied Stochastic Models and Data Analysis, 13 (1), 61-72.*

Aguilera, A.M., Ocaña, F.A. y Valderrama, M.J. (1997b). Regresión sobre componentes principales de un proceso estocástico con funciones muestrales escalonadas. *Estadísitca Española, 39, No. 142, 5-21.*

Aguilera, A.M., Ocaña, F.A. y Valderrama, M.J. (1999b). Forecasting time series by functional PCA. Discussion of several weighted approaches. *Computational Statistics 14, 443-467.*

Aguilera, A.M., Ocaña, F.A. y Valderrama, M.J. (1999c). Forecasting with unequally spaced data by a functional principal component approach. *Test, 8 (1), 233-253.*

Aguilera, A.M., Ocaña, F.A. y Valderrama, M.J. (1999d). Stochastic modelling for evolution of stock-prizes by means of functional principal component analysis. *Applied Stochastic Models. Business and Industry, 15 (4), 227-234.*

Aguilera, A.M., Ocaña, F.A. y Valderrama, M.J. (2002). Estimating functional principal component analysis on finite-dimensional data spaces. Sometido a *Statistics and computing.*

Albert, A. y Anderson, J.A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika, 71, 1-10.*

Anderson, J.A. (1972). Separate sample logistic discrimination. *Biometrika, 59, 19-35.*

Anderson, T.W. (1963). Assymtotic theory for principal component analysis. *Annals of Mathematical Statistics, 34, 123-148.*

Anderson, T.W. (1984). *An introduction to multivariate statistics analysis.* Wiley.

Anderson, J.A. y Blair, V. (1982). Penalized maximum likelihood estimation in logistic regression and discrimination. *Biometrika, 69, 123-36.*

Aucott, L.S. Garthwaite, P.H. and Currall, J. (2000). Regression methods for high dimensional multicollinear data. *Communications in Statistics, simulation, 29 (4), 1021-1037.*

Baker, C.T.H. (1977). *The numerical treatment of integral equations.* Oxford University Press.

Bartels, R.H., Beaty, J. C. y Barsky, B.A., (1987). *An introduction to Splines for use in computer graphics and geometric modelling.* Morgan Kaufmann Publishers.

Basilevsky, A. (1994). *Statistical factor analysis and related methods. Theory and applications.* Wiley.

Basilevsky, A. (1983). *Applied matrix algebra in the statistical sciences.* North-Holland.

Besse, P. (1980). Deux exemples d'analyses en composantes principales filtrantes. *Statistique et Analyse des Données, 15 (4), 5-15*

Besse, P. (1987). Choix de la métrique pour l'a.c.p. d'événements discrets. *Statistique et Analyse des Données, 12 (3), 1-16.*

Besse, P. (1988). Spline functions and optimal metric in linear principal component analysis. *Component and correspondence analysis. Dimension reduction by functional appoximation,* 81-110. Editado por Van Rijckevorsel, J.L.A. and De Leeuw, Wiley.

Besse, P. (1991). Approximation spline de l'ánalyse in composantes principales d'une variable aléatoire Hilbertienne. *Annales de la Faculté des Sciences de Tolouse, 12, 329-346.*

Besse, P. y Cardot, H. (1994). Approximation spline de la prévision d'un procesus autorégressif Hilbertien d'ordre 1. *Publication du Laboratoire de Statistique et Probabilités de Tolouse, 17.*

Besse, P, Caussinus, H., Ferre, L. y Fine, J. (1986). Some guidelines for principal component analysis. *Proceedings in Computational Statistics 1996, 23-30.* Springer-Verlag.

Besse, P., Caussinus, H., Ferre, L. y Fine, J. (1988). Principal components analysis and optimization of graphical displays. *Statistics, 2, 301-312.*

Besse, P. y Ramsay, J. O. (1986). Principal components analysis of sampled functions. *Psychometrika, 51 (2), 285-311.*

Bojanov, B.D., Hakopian, H.A. y Sahakian, A.A. (1993). *Spline functions and multivariate interpolations.* Kluwer Academic Publishers.

Bonifas, L., Escoufier, Y., Gonzalez, P. L. y Sabatier, R. (1984). Choix de variables en analyse en composantes principales. *Revue Statistique Appliquée, XXXII (2), 5-15.*

Cardot, H., Faivre, R. y Goulard M. (2001). Prédictions fonctionelles de l'occupatios des sols à partir de l'èvolution temporelle d'images satellites. *Actas de la IV Reunión de Trabajo de Predicción Dinámica,* Editado por M. J. Valderrama, Universidad de Granada.

Cardot, H., Ferraty, F. y Sarda, P. (1999). Functional liear model. *Statistics and Probability Letters, 45, 11-22.*

Cardot, H., Ferraty, F., Grimm, Mas, A. y Sarda, P., (2001). Testing hypothesis in the functional linear model. *Scaninavian Journal of Statistics.* En prensa.

Carling, J. B., Wolfe, R., Brown, C.H., y Gelman, A. (2001). A case study on the choice, interpretation and checking of multilevel models for longitudinal binary outcomes. *Biostatistics, 2, 397-416.*

Castro, P.E., Lawton, W.H. y Silvestre, E.A. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics, 28(4), 329-337.*

Chiang, C.T., Rice, J.A. y Wu, C.O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association, 96, 605-619.*

Christensen, R. (1997). *Log-linear models and logistic regression.* Springer-Verlag.

Christmann, A. y Rousseeuw, P. (2000). Measuring overlap in logistic regression. *Proceedings in computational statistics 2002 (Short communications and posters), 19-20.* Editado por Jansen, W. y Bethlehem, J. G., Statistics-Netherlands.

Cox, D. R. y Miller, H. D. (1972). *The theory of stochastic processes.* Chapman and Hall.

Cuadras, C. M. (1981). *Métodos de Análisis Multivatriante.* Eunibar.

Daudin, J. J., Duby, C. y Trécourt, P. (1989). PCA stability studied by the bootstrap and the infinitesimal jacknife method. *Statistics, 20, 2, 255-270.*

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.*

Davis, P.J. (1975). *Interpolation and approximation.* Dover publications.

De Boor, C. (1978). *A practical guide to splines.* Springer-Verlag.

Delves, L.M. y Mohamed, J. L. (1985). *Computational methods for integral equations.* Cambridge University Press.

Deville, J. C. (1973). Estimation of the eigenvalues and of the eigenvectors of a covariance operator. *Note interne de l'INSEE.*

Deville, J. C. (1974). Méthodes estatistiques et numériques de l'analysse harmonique. *Annales de l'INSEE, 15, 3-101.*

Dobson, A.J. (1983). *An introduction to generalized linear models.* Chapman & Hall.

Doob, J. L. (1953). *Stochastic processes.* Wiley.

Draper, N. R. y Smith, H. (1980). *Applied regression analysis.* Wiley.

Firinguetti, L. (1987). Regresión en componentes principales versus regresión "Ridge". *Revista de la Sociedad Chilena de Estadística, 4 (1-2), 14-32.*

Fomby, T.B., Hill, R.C. y Johnson, S.R. (1978). An optimal property of principal components in the context of restricted least squares. *Journal of the American statistical association: theory and methods, 361, 191-193.*

Fraiman, R. y Muñiz, G. (2001). Trimmed means for functional data. *Test, 10 (2), 419-440.*

Frank, I.E. y Friedman, J.H. (1993). A statistical view of some chemometrics regression tools. *Technometrics, 35 (2), 109-148.*

Fukunaga, K. (1990). *Introduction to statistical pattern recognition.* Academic Press.

Golub, G.H. y Van Loan, C. F. (1989). *Matrix computations.* The Johns Hopkins University Press.

Graybill, F.A. e Iyer, H.K. (1994). *Regression analysis: concepts and applications.* Belmont.

Gunst, R.F. y Mason, R.L. (1977). Advantages of examining multicollinearities in regression analysis. *Biometrics, 33, 249-260.*

Gunst, R.F. y Mason, R.L. (1977). Biased estimation in regression: an evaluation using mean squared error. *Journal of the American statistical association: theory and Methods, 359, 616-627.*

Gutiérrez, R. y González, A. (1991). *Introducción al análisis multivariante. Vol. II.* Universidad de Granada.

Gutiérrez, R., Ruiz-Molina, 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.*

Gutiérrez, R. y Valderrama, M.J. (1989). Discussion of two procedures for expanding a vector-valued stochastic process in an orthonormal way. *Linear Algebra and its applications, 120, 617-623.*

Hocking,R.R. (1976). The analysis and selection of variables in linear regression. *Biometrics, 32, 1-49.*

Hoover, D.R., Rice, J.A., Wu, C. O. y Yang, L.P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longotudinal data. *Biometrika, 85 (4), 809-822.*

Hosmer, D.W., Hosmer, T., Le Cessie, S. y Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model, *Statistics in Medicine, 16, 965-980.*

Hosmer, D.W. y Lemeshow, S. (1989). *Applied logistic regression.* Wiley.

Jackson, J.E. (1991). *A user's guide to principal components.* Wiley.

Jolliffe, I.T. (1982). A note on the use of principal components in regression. *Applied statistics, 31, 300-303.*

Kantorovich, L.V. y Akilov, G.P. (1964). *Functional analysis in normed spaces.* Pergamon- McMillan.

Kauermann, G. (2000). Modelling longitudinal data with ordinal response by varying coefficients. *Biometrics, 56, 692-698.*

Kleinbaunm, D.G. (1994). *Logistic regression. A self-learning text.* Springer-Verlag.

Lancaster, P. y Salkauskas, K. (1986). *Curve and surface fitting. An introduction.* Academic Press.

Le-Cessie, S. y Van Houwelingen, J.C. (1991). A goodness-of-fit test for binary regression models bassed on smoothing methods. *Biometrics, 47, 1267-1282.*

Levy, P. (1996). *Processus stochastiques et mouvement Brownien.* Gauthiers Villars.

Little, R.J.A. y Schluchter, M.D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika, 72, 497-512.*

Mansfield, E.R., Webster, J.T. y Gunst, R.F. (1977). An analytic selection technique for principal component regression. *Applied statistics, 26, 34-40.*

Marx, B.D. y Eilers, P.H.C. (1999). Generalized linear regression on sampled signals and curves. A P-spline approach. *Technometrics, 41, 1-13.*

Massy, W.F. (1965). Principal component regression in exploratory statistical research. *Journal of the American statistical association, 60, 234-246.*

McCullagh, P. y Nelder, J.A. (1983). *Generalized Linear Models.* Chapman & Hall.

Montgomery, D.C. y Peck, E.A. (1992). *Introduction to linear regression analysis.* Wiley.

Muirhead, R.J. (1982). *Aspects of multivariate statistical theory.* Wiley.

Obadia, J. (1978). L'analyse en composantes explicatives. *Revue Statistique Appliquée, 26 (4), 5-28.*

Ocaña, F.A. (1995). *Alternativas geométricas en el ACP de una variable aleatoria Hilbertiana.* Tesis Doctoral, Universidad de Granada.

Ocaña, F.A., Aguilera, A.M. y Valenzuela, O. (1998). A wavelet approach to functional principal component analysis. *Proceedings in Computational Statistics 1998, 413-418,* Editado por R. Payne, y P. Green, Physica-Verlag.

Ocaña, F.A., Aguilera, A.M. y Valderrama, M.J. (1999). Functional principal component analysis by choice of norm. *Journal of Multivariate Analysis, 71 (2), 262-276.*

Otter, P.W. y Schuur, J.F. (1982). Principal component analysis in multivariate forecasting of economic time-series. En: *Time Series Analysis: Theory and Practice 1.* Ed. Anderson, O.D., North-Holland.

Pottier, P. (1991). Utilisations de l'analyse en composantes principales pour la prévision statistique en météorologie. *Revue Statistique Appliquée, 39 (1), 37-49.*

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of statistics, 9 (4), 705-724.*

Prenter, P. M. (1975). *Splines and variational methods.* Wiley.

Prentice, R.L. y Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika, 66, 403-11.*

Pulkstenis. E. y Robinson, T.J. (2002). Two goodness-of-fit tests for logistic regression models with continuous covariates. *Statistics, and Medicine, 21, 79-93.*

Qin, J. y Zhang, B. (1997). A goodness-of-fit test for logistic regression models bassed on case control data. *Biometrika, 84 (3), 609-618.*

Ramsay, J.O. (1982). When the data are functions. *Psychometrika, 47 (4), 379-396.*

Ramsay, J. O. y Dalzell, C. J. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society, serie B, 53 (3), 539-572.*

Ramsay, J. O. y Silverman, B.W. (1997). *Functional Data Analysis.* Springer-Verlag.

Ramsay, J. O. y Wang, X. (1995). A functional data analysis of the pinch force of human fingers. *Applied Statistics, 44 (1), 17-30.*

Rao, C.R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya, A, 26, 329-358.*

Rao, C.R. y Toutenburg, H. (1995). *Linear models. Least squares and alternatives.* Springer-Verlag.

Reinhart, H.J. (1985). *Analysis of approxiamation methods for differential and integral equations.* Springer-Verlag.

Rice, J. A. y Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, serie B, 53 (1), 233-243.*

Riesz, F. y Sz-Nagy, D. (1990). *Leçons d'analisé fonctionnelle.* Gauthier-Villars.

Rohatgi, V.K. (1984). *Statistical inference.* Wiley.

Royston, P. (1992). The use of cusums and other techniques in modelling continuous covariates in logistic regression. *Statistics in Medicine,* Vol. 11, 1115-1129.

Ryan, T.P. (1997). *Modern regression methods.* Wiley.

Santner, T.J. y Duffy, D.E. (1986). A note on A. Albert and J. A. Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika, 73, 755-758.*

Saporta, G. (1985). Data analysis for numerical and categorical individual time-series. *Applied Stochastic Models and Data Analysis, 1, 109-119.*

Saporta, G. (1990). *Probabilités analyse des données et statistique.* Editions Technip.

Sen, A. y Srivastava, M. (1990). *Regression analysis. Theory, methods and applications.* Springer-Verlag.

Silverman, B.W. (1996). A smoothed functional principal component analysis by choice of norm. *Annals of Statistics, 24, 1-24.*

Stefanski, L.A., Carroll, R.J. y Ruppert, D. (1986). Optimally bounded functions for generalized linear models with applications to logistic regression. *Biometrika, 73, 413-24.*

Tenenhaus, M. (1977). Analyse en composantes principales d'un ensemble de varaibles nominales ou numériques. *Revue Statistique Appliquée, 25 (2), 39-56.*

Todorovic, P. (1992). *An introduction to stochastic processes and their applications.* Springer-Verlag.

Tsiatis, A.A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika, 67 (1), 250-1.*

Valderrama, M.J. Aguilera, A.M. y Ocaña, F.A. (2000). *Predicción dinámica mediante análisis de datos funcionales.* Hespérides-La Muralla.

Venables, W.N. y Ripley, B.D. (1997). *Modern applied statistics with S-Plus.* Springer-Verlag.

Vera, A. y Alegría, P. (1997). *Un curso de Análisis Funcional.* AVL.

Wang, C.Y. y Huang, Y. (2001). Functional methods for logistic regression on random-effect-coefficients for longitudinal measurements. *Statistics and Probability Letters, 53, 347-356.*

Wedderburn, R.W. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika, 63, 27-32.*

Wegman, J. y Wright, W. (1983). Splines in statistics. *Journal of the American Statistical Associations, 78 (382), 351-365.*

Wilkinson, J.H. (1965). *Algebraic eigenvalues problem.* Oxford University Press.

Wilks, S. (1962). *Mathematical statistics.* Wiley, New York.

Winsberg, S. y Ramsay, J. (1983). Monotone spline transformation for dimension reduction. *Psychometrika, 48, 575-595.*

Wold, S. (1974). Spline functions in data analysis. *Technometrics, 16 (1), 1-11.*

Wong, C.S. y Li, W.K. (2001). On a logistic mixture autoregressive model. *Biometrika, 88 (3), 833-846.*

Wong, E. (1971). *Stochastic processes in information and dynamical systems.* McGraw-Hill.

Wong, E. y Hajek, B.K. (1985). *Stochastic processes in Engineering Systems.* McGraw-Hill.