

memorias

X

V

I

Foro Nacional de Estadística



www.inegi.gob.mx

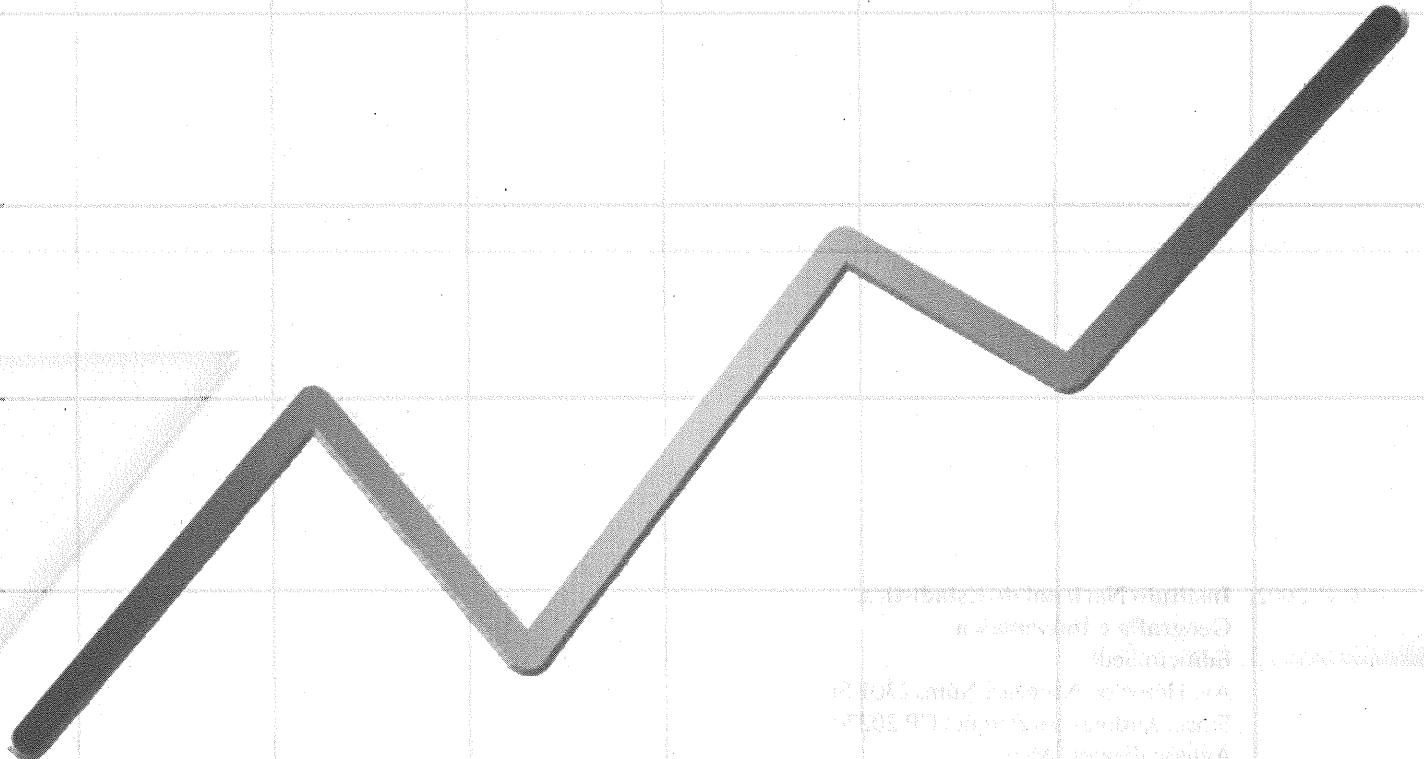
memorias

X

V

I

Foro Nacional de Estadística



FORO NACIONAL
DE ESTADÍSTICA
ESTADÍSTICA
ESTADÍSTICA

DR © 2002, **Instituto Nacional de Estadística,
Geografía e Informática**
Edificio Sede
Av. Héroe de Nacozari Núm. 2301 Sur
Fracc. Jardines del Parque, CP 20270
Aguascalientes, Ags.

www.inegi.gob.mx
atencion.usuarios@inegi.gob.mx

Memorias del XVI Foro Nacional de Estadística

Impreso en México
ISBN 970-13-4076-0

Presentación

El XVI Foro Nacional de Estadística se llevó a cabo de 8 al 13 de octubre de 2001 en el Centro Universitario de Ciencias Exactas e Ingeniería (CUCEI) de la Universidad de Guadalajara, organizado por el departamento de Matemáticas de dicha Universidad.

Se presentaron 66 contribuciones libres, 4 conferencias magistrales y 2 cursos cortos. En estas memorias se presentan resúmenes de dichas contribuciones. Todos los resúmenes recibidos fueron incluidos sin un proceso de arbitraje, aunque con una detallada revisión.

La Asociación Mexicana de Estadística agradece a la Universidad de Guadalajara su apoyo para la realización de este foro y al Instituto Nacional de Estadística, Geografía e Informática el apoyo para la edición de estas memorias.

El Comité Editorial:

Eduardo Castaño Tostado

Rafael Perera Salazar

Karim Anaya Izquierdo

Contenido

Presentación	III
Una Aplicación de Distribuciones Estables	1
<i>Alvarez, E. y Alegria, A.</i>	
Elementos Básicos del Análisis Funcional de Datos:	
Una Aplicación en Tecnología de Alimentos	9
<i>Calva, H. y Castaño, E.</i>	
Estimacion casi no Parametrica en mezclas de modelos	23
<i>Cruz Medina, I.</i>	
A Statistical Method for the Determination of the Appropriate Order in a General Class of Time Series Models	29
<i>Contreras, A. y Gonzalez Barrios, J.</i>	
Construcción de Clusters en Series de Tiempo	37
<i>Domínguez, J. y González Fariás, G.</i>	
Componentes Principales y Medidas de Dependencia	45
<i>González Barrios, J. y Ruiz Velasco, S.</i>	
Modelos Marginales y Condicionales Para Mediciones Repetidas Binarias	57
<i>Gracia Medrano, L. y Ruiz Velasco, S.</i>	

Encuesta Aplicada a Empresas Hoteleras con Categorías de 3, 4 y 5 Estrellas en los Municipios de Bahía de Banderas, Compostela, San Blas y Tepic del Estado de Nayarit.	
(Oct.1998 a Feb.1999)	67
<i>Iñiguez, A. y Gómez, M.</i>	
Análisis de modelos lineales en encuestas complejas	75
<i>Méndez, I., Romero, P. y Eslava, G.</i>	
Introducción al Análisis Bayesiano de Datos Direccionales	85
<i>Nuñez, G. y Gutiérrez Peña, E.</i>	
Hacia una Nueva Pedagogía: El Enfoque Basado en Proyectos para Mejorar el Aprendizaje del Diseño Estadístico	93
<i>Ojeda, M., Morales, E., Caballero, M. y Galeana, N.</i>	
Las Componentes Principales como Análisis de Datos en Poblaciones de Maíz	105
<i>Padrón, E., Méndez, I., Muñoz, A., y Hernández, N.</i>	
Uso de Variable Indicadoras en Predicción Espacial	113
<i>Peraza, F. y González Farias, G.</i>	
Inferencia sobre Valores Récords	121
<i>Perera, R. y Cortina, M.</i>	
Estimación de los Parámetros de Correlación en Modelos para Datos Longitudinales	129
<i>Ruiz Velasco, S.</i>	

Modelos Threshold Autorregresivos y Métodos MCMC	135
<i>Russel, M., González Farias, G. y Huerta Gómez, G.</i>	
Aplicación de la Búsqueda Tabú en Regresión No Lineal	143
<i>Trejos, J. y De los Cobos, S.</i>	

Una Aplicación de Distribuciones Estables

Elia Alvarez
Alejandro Alegría

Instituto Tecnológico Autónomo de México

1. Introducción

Durante mucho tiempo se sostuvo la hipótesis que afirmaba que el comportamiento de los precios en mercados especulativos podía ser descrito por caminatas aleatorias. Originalmente, la distribución de estos cambios era considerada aproximadamente Normal, tiempo después Mandelbrot y Fama formularon la hipótesis que sostenía que la distribución de estos cambios se ajusta mejor mediante las Leyes Estables debido a que en ellas se considera el comportamiento pesado de las colas de la distribución, así como la asimetría. En este trabajo se definen las Distribuciones Estables por medio de su función característica, se presentan sus principales propiedades y finalmente se presenta un modelo regresivo generalizado para la estimación de los parámetros.

2. Distribuciones Estables

Las Distribuciones Estables se definen analíticamente mediante su función característica debido a que sus densidades pueden ser expresadas únicamente por complicadas funciones especiales. Si $F(x)$ es estable, su función característica puede ser escrita de la forma,

$$\log(\varphi_X(t)) = \log E(e^{itX}) = i\gamma t - \delta|t|^\alpha(1 + i\beta \operatorname{sgn}(t) z(t, \alpha)), \quad t \in \mathbb{R}.$$

donde

$$z(t, \alpha) = \begin{cases} \tan\left(\frac{\pi\alpha}{2}\right) & , \text{ si } \alpha \neq 1 \\ (2/\pi)\log|t| & , \text{ si } \alpha = 1 \end{cases}$$

$\gamma \in \Re$ es el parámetro de Localización; $\delta \in \Re$ es el parámetro de Escala; $\alpha \in [0, 2]$ es el Exponente Característico (el parámetro que determina las propiedades básicas de esta clase de distribuciones como son momentos, densidad de las colas, el tipo de distribución estable, etc) y $\beta \in [-1, 1]$ es el parámetro de Simetría. La distribución es asimétrica a la izquierda cuando $\beta > 0$ o a la derecha cuando $\beta < 0$, y es simétrica cuando $\beta = 0$. Si $|\beta| = 1$, entonces la distribución tiene una sola cola.

En algunos casos la función característica corresponde a una función de densidad que puede ser escrita en forma explícita, por ejemplo, con $\alpha = 2$ obtenemos una Distribución Normal($\gamma, 2\delta$) $\alpha = 1$ y $\beta = 0$ corresponde a la Distribución Cauchy, $\alpha = 1/2, \beta = 1$ corresponde a la Distribución Levy.

Una representación alternativa de $\log(\varphi_X(t))$ y que es útil para los propósitos de este trabajo es la siguiente,

$$\log(\varphi_X(t)) = i\gamma t - |t|^\alpha \delta'^\alpha \exp\left\{-i\beta' \frac{\pi}{2} \eta_\alpha sgn(t)\right\}, \text{ con } \eta_\alpha = \min(\alpha, 2 - \alpha)$$

La relación de β' y δ' con los parámetros de la caracterización inicial está dada por

$$\begin{aligned} \beta' &= \frac{2}{\pi\eta_\alpha} \cos^{-1}(\Delta^{-1} \cos(\frac{\pi\alpha}{2})), & \delta' &= (\Delta\delta / \cos(\frac{\pi\alpha}{2}))^{1/\alpha} \\ \Delta^2 &= \cos^2(\frac{\pi\alpha}{2}) + \beta^2 \operatorname{sen}^2(\frac{\pi\alpha}{2}), & \operatorname{sgn}(\Delta) &= \operatorname{sgn}(1 - \alpha), \quad \operatorname{sgn}(\beta') = \operatorname{sgn}(\beta) \end{aligned}$$

El parámetro de localización, así como el exponente característico no cambian en esta representación. La densidad correspondiente a esta función característica puede ser calculada usando la fórmula de inversión de Fourier,

$$f_\alpha(y|\gamma, \delta', \beta') = \frac{1}{\pi} \int_0^\infty \cos\{(\gamma - y)s/\delta' + s^\alpha \operatorname{sen}(\eta'_{\alpha,\beta'})\} \exp\{-s^\alpha \cos(\eta'_{\alpha,\beta'})\} \frac{ds}{\delta'}$$

donde $\eta'_{\alpha,\beta'} = \beta' \eta_\alpha \pi / 2$.

La anterior integral puede ser evaluada numéricamente; así, podemos aproximar la densidad (y por lo tanto la verosimilitud) para la Distribución Estable con parámetros dados. Para lograr una buena aproximación en este trabajo se utilizó la optimización desarrollada por P. Lambert y J.K. Lindsey (1999) implementada en R, que consta de diversos algoritmos basados en un optimizador no lineal. El procedimiento de optimización requiere de valores iniciales que sean razonables para garantizar la convergencia y para permitir una rápida evaluación de la función de verosimilitud.

3. Modelos de Regresión Generalizados

Una vez que hemos calculado la verosimilitud, es posible modelar el parámetro de localización en términos de covariables. Sean $\{X_1, X_2, \dots, X_n\}$ covariables asociadas con las observaciones $\{y_1, y_2, \dots, y_n\}$ y $g(\cdot)$ una función liga que se aplica al parámetro de localización. Se define el siguiente modelo de regresión,

$$g(\gamma) = (g(\gamma_1), \dots, g(\gamma_n))^T = X^T \Psi$$

donde $X = (X_1, X_2, \dots, X_n)^T$ y Ψ denotan la matriz de diseño y el vector de parámetros asociados, respectivamente. Los otros tres parámetros de la Distribución Estable, pueden ser preestablecidos o simultáneamente estimados en caso de ser necesario. Cuando $\alpha = 2$, se recobra el modelo de regresión tradicional.

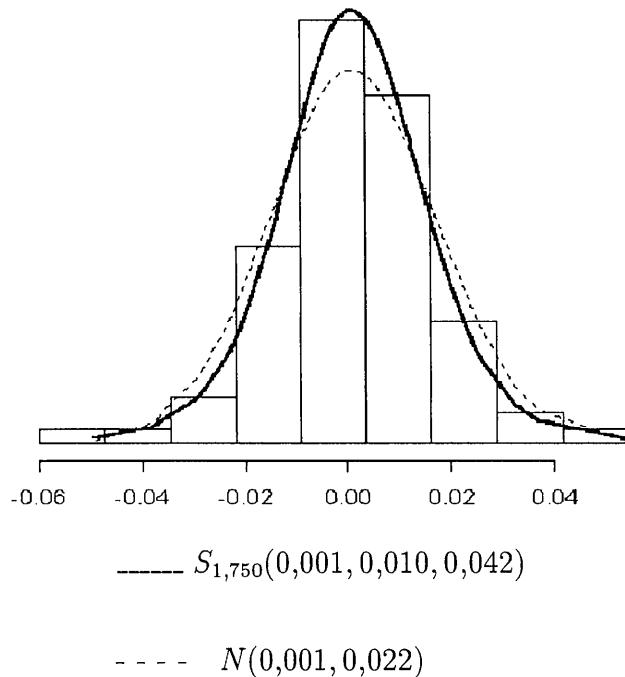
Modelar la moda en vez de el parámetro de localización puede ser en algunos casos más apropiado, ya que la moda es el valor que presenta a su alrededor las regiones de mayor densidad. Desafortunadamente, para muchas distribuciones, no contamos con una expresión cerrada que nos indique cuál es la moda, pero ésto no implica que no la podemos modelar directamente. En la familia de distribuciones estables, para valores fijos de δ' , β' y α , se demuestra fácilmente que $\tilde{y} - \gamma$ no varía con γ , donde \tilde{y} denota la moda de la Distribución Estable $S_\alpha(\gamma, \delta', \beta')$. De esta forma, \tilde{y} es sólo una función, llamémosla $d(\delta', \beta', \alpha)$, de los otros parámetros. Así, el modelo de regresión generalizado para la moda sería $\tilde{y}_i = g^{-1}(X_i^T \Psi) + d(\delta', \beta', \alpha)$, y considerando el caso especial en el que $g(.)$ es la función identidad, tenemos que el modelo se define por $\tilde{y}_i = X_i^T \Psi + d(\delta', \beta', \alpha)$.

Se debe considerar que la moda se encuentra además influenciada por la asimetría y la curtosis de la distribución, por lo que un modelo regresivo para la moda debe de ser interpretado con más cuidado que un modelo para el parámetro de localización . La total independencia de los cuatro parámetros de una Distribución Estable nos da gran flexibilidad para modelar los datos.

4. Análisis de los Rendimientos del índice de Precios y Cotizaciones de la Bolsa Mexicana de Valores

En esta sección se analizarán los rendimientos del IPC usando el modelo de Distribuciones Estables que permite modelar colas más pesadas que las presentes bajo una Distribución Normal. Además, las observaciones extremas son consideradas como una realidad dentro del proceso de cambios en los precios, pero no reciben un tratamiento especial. Se analizaron los rendimientos en el IPC para los 4 primeros meses de 2001, incluyendo en los modelos ajustados la covariable tiempo. En la siguiente gráfica podemos observar que el ajuste proveniente de la Distribución estable, sin dinámica en los parámetros, considera las colas pesadas que

muestra el histograma. En la cima de la distribución la curva que representa la Distribución Estable, con $\alpha \neq 2$, se encuentra muy cercana a los valores más probables en el histograma, lo que no sucede con el ajuste de una distribución normal.



Varios modelos se compararon entre sí de acuerdo al criterio de Akaike. En primer lugar se consideró el modelo estacionario para estos rendimientos. Este modelo asume que la distribución de los rendimientos no evoluciona con el tiempo, es decir, para el periodo considerado los cuatro parámetros, localización, escala, exponente característico y asimetría son constantes. Las estimaciones con sus errores estándar fueron,

$$\begin{aligned}\hat{\gamma} &= 0,00103, & \hat{\delta} &= 0,0099, & \hat{\beta} &= 0,0421, & \hat{\alpha} &= 1,7499 \\ (0,002) & & (0,112) & & (1,094) & & (0,925)\end{aligned}$$

indicando que la distribución es ligeramente asimétrica hacia la izquierda y con colas más pesadas que una Distribución Normal.

Del mismo modo se consideraron modelos no estacionarios, y el mejor de éstos resultó ser el que modela el parámetro de localización por medio de una función cuadrática del tiempo. Los resultados fueron,

$$\begin{aligned}\hat{\gamma} &= 0,0009 - 0,0188t + 0,0346t^2, & \hat{\delta} &= 0,0108, & \hat{\beta} &= 0,9999, & \hat{\alpha} &= 1,9622 \\ &(0,002) \quad (0,015) \quad (0,016) & &(0,087) & &(1,800) & &(2,622)\end{aligned}$$

Aunque el anterior modelo provee de un mejor ajuste, el modelo lineal en el tiempo para el parámetro de localización posee una singular ventaja en cuanto a interpretación, ya que se relaciona directamente con el valor que presenta a su alrededor la región con más densidad, es decir, la moda. Para nuestro ejemplo, tenemos que el modelo para el parámetro de localización, que varía linealmente con el tiempo es $\hat{\gamma}_t = 0,0032 - 0,0054t$. De la ecuación $\tilde{y}_i = X_i^T \Psi + d(\delta', \beta', \alpha)$, tenemos que $d(\delta', \beta', \alpha)$ puede ser determinada restando $\hat{\Psi}_0$ de la moda de la Distribución Estable ajustada, $S_{\hat{\alpha}}(\hat{\gamma}, \hat{\delta}, \hat{\beta})$, cuando $t = 0$. Esto produce el modelo $\tilde{y}_t = 0,0002 - 0,0054t$ para la moda.

5 Conclusiones

Existe evidencia empírica para afirmar que el comportamiento de la distribución de los rendimientos diarios del precio de las acciones es Estable. La familia de Distribuciones Estables tiene elementos cuya densidad se maneja a través de su función característica, y están definidas por cuatro parámetros : localización, escala, asimetría y exponente característico, que pueden variar independientemente uno de otro, lo cual provee de gran libertad a diferencia de otras distribuciones.

En el estudio de las Distribuciones Estables aún queda mucho por investigar, por ejemplo mejorar métodos de estimación, incluir otras covariables, aplicar métodos bayesianos, etc.

El objetivo de este trabajo fue el mostrar una aplicación de estas distribuciones, y se espera que aumente el interés en su estudio.

Referencias

Fama, E.F. & Roll,R. (1968) Some Propieties of Symmetric Stable Distributions. *Journal of the American Statistical Association*, **63**, 817-836.

Fama, E.F. & Roll,R. (1971) Parameter Estimates for Symmetric Stable Distributions. *Journal of the American Statistical Association*, **66**, 331-338.

Fielitz, B.D. & Smith E. W. (1972) Asymmetric Stable Distributions of Stock Price Changes. *Journal of the American Statistical Association*, **67**, 813-814.

Lambert, P. & Lindsey, J.K. (1999) Analysing Financial Returns by Using Regression Models Based on Non-Symmetric Stable Distributions. *Appl. Statist*, **48**, 409-424.

Mandelbrot, B. (1963) The Variation of Certain Speculative Prices, *Journal of Business*, **36**, 394-419.

Fuente de los datos: <http://infoselfinanciero.com.mx/servint>. Infosel Financiero. México. Mayo de 2001.

Elementos Básicos del Análisis Funcional de Datos: Una Aplicación en Tecnología de Alimentos

Hugo Calva Díaz

Eduardo Castaño Tostado

Universidad Autónoma de Querétaro

1. Representando los datos como funciones suaves

En la actualidad hay procesos en los cuales se miden características de unidades experimentales ó de muestreo de manera casi continua, esto motiva a que estas mediciones representen una función de manera discreta. Esto lleva a analizar este tipo de procesos de una manera funcional. La filosofía básica del análisis funcional de datos: un conjunto de datos que representa mediciones repetidas de una unidad bajo estudio, se considera como producto de la manifestación discreta de una función subyacente x . Un registro de una observación funcional x consiste de n pares (t_j, y_j) , donde $y_j = x(t_j)$. Es supuesto siempre que el rango de los valores de interés para el argumento t es un intervalo acotado Υ , además de que x satisface algunas condiciones de continuidad o suavidad sobre Υ . Si tenemos un vector $y = (y_1, \dots, y_n)$ entonces podemos escribir el modelo como sigue: $y_j = x(t_j) + \varepsilon_j$ con $\varepsilon_j \sim (0, \sigma^2)$, donde ε_j es el error, el cual contribuye a la rugosidad del vector de datos. Éste error o ruido se puede dar por aspectos del proceso de medición. Una de las tareas en representar los datos como funciones es intentar filtrar el ruido eficientemente, o suavizar los resultados del análisis. Contando con estos elementos el paso a seguir, es la representación de x como una función suave, para de ahí poder llegar al análisis de datos.

1.1. Representar a x por funciones base

Uno de los métodos o procedimientos más familiares para suavizar es representando la función $\hat{x}(t)$ por una combinación lineal de K funciones base conocidas como ϕ_k , es decir,

$$x(t) = \sum_{k=1}^K c_k \phi_k(t).$$

El grado al cual los datos y_j están suavizados está determinado por K funciones base. El suavizador lineal más simple es obtenido si determinamos los coeficientes de la expansión c_k minimizando el criterio de mínimos cuadrados

$$SMSSE(y | c) = \sum_{j=1}^n \left[y_j - \sum_{k=1}^K c_k \phi_k(t_j) \right]^2,$$

o en términos matriciales,

$$SMSSE(y | c) = (y - \Phi c)' (y - \Phi c) = \|y - \Phi c\|^2,$$

donde el vector c contiene los coeficientes c_k y Φ es una matriz de $n \times K$ donde $\Phi_{j,k} = \{\phi_k(t_j)\}$, es decir son las funciones base evaluadas en los puntos de observación. Este criterio es minimizado por la solución $c = (\Phi' \Phi)^{-1} \Phi' y$. Así tendremos que $\hat{x} = \Phi(\Phi' \Phi)^{-1} \Phi' y$. Haciendo

$S = \Phi(\Phi' \Phi)^{-1} \Phi'$, llegaremos a que $\hat{x} = S y$, finalmente $\hat{x}(t) = \sum_{j=1}^n S_j(t) y_j$. Como hemos

podido observar llegamos a una expresión explícita para representar a la función x por medio de funciones base, a la última expresión del proceso de estimación recibe el nombre de **suavizador lineal**, el cual estima el valor de la función $x(t)$ mediante una combinación

lineal de las observaciones discretas $\hat{x}(t) = \sum_{j=1}^n S_j(t) y_j$. Una característica deseable de las

funciones base, es que concuerden con aquellos aspectos conocidos de las funciones a ser estimadas. Una base debería ser escogida de tal manera que se logre una excelente aproximación

usando valores pequeños de K . La selección de bases es particularmente importante para estimar derivadas. Bases que trabajan bien para estimar la función pueden dar una muy pobre estimación de las derivadas. Esto es debido a que una representación más exacta de las observaciones pueden forzar a \hat{x} , a tener pequeñas oscilaciones, pero con una frecuencia muy alta con terribles consecuencias para sus derivadas.

1.2. Tipos de base

1.2.1. Series de Fourier

Una base muy conocida está dada por series de Fourier: $\hat{x}(t) = c_0 + c_1 \sin \varpi t + c_2 \cos \varpi t + c_3 \sin 2\varpi t + c_4 \cos 2\varpi t + \dots$. Definida por la base $\phi_0(t) = 1$, $\phi_{2r-1}(t) = \sin(r\varpi t)$, $\phi_{2r}(t) = \cos(r\varpi t)$. Esta base es periódica y ϖ determina el período $2\pi/\varpi$, el cual es igual a la longitud del intervalo Υ . La transformada rápida de Fourier (TRF) hace posible encontrar los coeficientes eficientemente; cuando n es una potencia de 2 y los argumentos están equiespaciados, entonces podemos encontrar los coeficientes c_k y los n valores suaves en $x(t_j)$ en el orden de $n \log n$ operaciones [$O(n \log n)$]. Esta es una de las características que ha hecho a la serie de Fourier la base tradicional para series de tiempo muy largas. Una serie de Fourier es especialmente útil para funciones extremadamente estables, sin comportamientos locales fuertes y curvatura uniforme. Esta base produce expansiones las cuales son uniformemente suaves. Una serie de Fourier es como margarina: “es barata y se puede untar donde sea, pero no esperes que el resultado de comerla sea excelente”.

1.2.2. Bases spline

Los splines son funciones construidas juntando polinomios suavemente en los valores τ_k llamados nodos. El número de estos nodos denotado por $K_1 + 1$, donde los nodos de los extremos definen el intervalo Υ sobre el cual la estimación toma lugar y $K_1 - 1$, es el número de nodos interiores. Entre dos nodos adyacentes, un spline polinomial es un polinomio de grado fijo K_2 , pero en un nodo interior, dos polinomios adyacentes son requeridos para coincidir en los valores de un número fijo de derivadas, generalmente $K_2 - 1$. Los splines combinan el cálculo fácil, con la capacidad para cambiar conductas locales y gran flexibilidad, los splines se representan generalmente como sigue: $\phi_k(t) = (t - \varpi)_+^{k_2}$, donde $u_+ = u$ si $u \geq 0$ y 0 en otro caso, y donde sólo los puntos interiores son usados. Esto es llamada la base de potencia truncada. Esta base tiene que ser aumentada por la base monomial t^k , $k = 0, \dots, K_2$, para producir un spline polinomial completo de grado K . El número total de funciones base es $K = K_1 + K_2$.

1.3. Representación de x penalizando la rugosidad

Este método de suavizamiento estima una curva x de las observaciones $y_j = x(t_j) + \varepsilon_j$, haciendo explícito dos posibles metas en la estimación de la curva. Por un lado deseamos asegurarnos que la curva estimada dé un buen ajuste a los datos. Por otro lado no deseamos que el ajuste sea demasiado bueno si esto resulta en una curva x que sea excesivamente localmente variable. De acuerdo al modelo $y_j = x(t_j) + \varepsilon_j$, $E(\varepsilon_j) = 0$, $Var(\varepsilon_j) = \sigma^2$, cada y_j es insesgado de $x(t_j)$, visto como una curva estimada, debe tener mucha variabilidad, si sacrificamos un poco el insesgamiento, un poco más suave será la curva estimada. Requiriendo que el estimado varíe sólo un poco de un valor a otro valor, estaremos tomando información de los valores de datos vecinos, y así expresando nuestra fe en la regularidad de la función subyacente x que estemos tratando de estimar. En suavizamiento por spline, el

error cuadrático medio captura, generalmente lo conocido por pobreza de estimación. Esto puede ser algunas veces dramáticamente reducido, sacrificando algo de sesgo y reducir la varianza muestral, y esta es una razón clave para imponer suavidad sobre la curva estimada. La filosofía de este método es permitir una clase más grandes de funciones x , pero cuantificar la rapidez de variación local de x y hacer un balance explícito entre regularidad y bondad de ajuste de los datos, el cual corresponde a un balance implícito entre varianza y sesgo. Una medida popular de la rugosidad de una función se computa integrando el cuadrado de un operador diferencial lineal, que es de la forma,

$$Lx = w_0x + w_1Dx + w_2D^2x + \dots + w_{m-1}D^{m-1}x + w_mD^mx,$$

y el criterio es:

$$PEN_L(x) = \int \{Lx(s)\}^2 ds = \|Lx\|^2.$$

Este criterio cuantifica la curvatura total en x , o alternativamente, el grado en el cual x se desvía de una línea recta. Entonces podemos definir una suma de cuadrados de residuales penalizada como:

$$PENSSE_\lambda(x | y) = \sum_j \{y_j - x(t_j)\}^2 + \lambda \times PEN_L(x)$$

Nuestro estimado de la función es obtenido al encontrar la función x que minimiza $PENSSE_\lambda(x)$. El parámetro λ es un parámetro de suavidad que mide la tasa de intercambio entre el ajuste de los datos y la variabilidad de la función x . Entre más grande λ , más grande el peso que se le dará a la rugosidad de la estimación. Para λ pequeña la curva tiende a ser más y más variable, ya que hay menos penalidad sobre su rugosidad. Cuando $\lambda \rightarrow 0$ la curva se approxima a una interpolación de los datos satisfaciendo $y_j = x(t_j) \quad \forall j$.

1.4. Representación de x usando bases complementarias

Supongamos que tenemos dos conjuntos de funciones base, ϕ_j , $j = 1, \dots, J$ y ψ_k , $k = 1, \dots, K$ que se complementan una a la otra. Las ϕ_j s son pocas y escogidas para dar una cantidad razonable de características de gran-escala de los datos. Las ψ_k s serán generalmente más en número y están diseñadas para captar lo que no representan las ϕ_j s. Supongamos que cualquier función x de interés puede ser expresada en términos de las dos bases como:

$$x(s) = \sum_{j=1}^J d_j \phi_j(t_j) + \sum_{k=1}^K c_k \psi_k(t_j).$$

1.4.1. Especificando la penalidad de la rugosidad

Enfoque en el que combinaciones lineales de ϕ_j s son completamente suaves, y por lo tanto no contribuyen a la penalidad de la rugosidad, entonces la penalidad de la rugosidad tiene que depender únicamente sobre los coeficientes de las ψ_k s, entonces x es la suma de dos

partes una función “ultrasuave” $x_s = \sum_{j=1}^J d_j \phi_j$ y una función $x_R = \sum_{k=1}^K c_k \psi_k$. Entonces lo que se busca es penalizar x_R . Podemos usar cualquier operador diferencial lineal definiendo.

$$PEN_L(x_R) = \int_{\Upsilon} [Lx_R(s)]^2 ds = \int_{\Upsilon} \left[\sum_{k=1}^K c_k L\phi_k(s) \right]^2 ds.$$

En términos matriciales como $PEN_L(x_R) = c' R c$, donde R es una matriz simétrica que contiene los elementos $R_{kl} = \int_{\Upsilon} L\psi_k L\psi_l ds = \langle L\psi_k, L\psi_l \rangle$. Podemos representar la suma de residuales al cuadrado mediante vectores c y d

$$\sum_i \{y_i - x(t_i)\}^2 = \|y - \Phi d - \Psi c\|^2,$$

donde Ψ es una matriz de $n \times K$ con elementos $\psi_{ik} = \psi_k(t_i)$, y de aquí se sigue el criterio de suavidad:

$$PENSSE_\lambda(x | y) = \|y - \Phi d - \Psi c\|^2 + \lambda c' R c,$$

y minimizando sobre d y c tenemos la curva ajustada x en términos de su expansión como en

$$x(s) = \sum_{j=1}^J d_j \phi_j(t_j) + \sum_{k=1}^K c_k \phi_k(t_j).$$

La solución para d , con c fija esta dada por: $d = (\Phi' \Phi)^{-1} \Phi' (y - \Psi c)$, y consecuentemente $\Phi d = P_\phi(y - \Psi c)$ donde $P_\phi = \Phi(\Phi' \Phi)^{-1} \Phi'$ (matriz de proyección) y sustituyendo d en $PENSSE_\lambda$, definiendo $Q_\phi = I - P_\phi$ y renombrando $Q_\phi \cdot Q_\phi = Q_\phi$ llegamos a la ecuación

$$c = (\Psi' Q_\phi \Psi + \lambda R)^{-1} \Psi' Q_\phi y.$$

Estos cálculos son muy didácticos pero numéricamente son muy inestables, es por eso que hay algunas maneras más estables y económicas de computar las soluciones anteriores.

2. Estadísticas descriptivas de datos funcionales

Las estadísticas descriptivas son una generalización del caso univariado y cumplen con el mismo objetivo de ayudarnos a obtener información acerca de la variabilidad de los datos:

$$\bar{x}(t) = N^{-1} \sum_{i=1}^N x_i(t);$$

$$VAR_X(t) = (N-1)^{-1} \sum_{i=1}^N [x_i(t) - \bar{x}(t)]^2$$

$$COV_X(t_1, t_2) = (N-1)^{-1} \sum_{i=1}^N [x_i(t_1) - \bar{x}(t_1)] [x_i(t_2) - \bar{x}(t_2)]$$

$$CORR_X(t_1, t_2) = \frac{COV_X(t_1, t_2)}{\sqrt{VAR_X(t_1)VAR_X(t_2)}}$$

3. Análisis de componentes principales para datos funcionales

La forma que tendrán los componentes principales funcionales estarán dados de forma análoga al caso multivariado, la ecuación propia resultante en el caso funcional es:

$$\int v(s, t)\xi(t)dt = \langle v(s, \cdot), \xi \rangle = \rho\xi(s),$$

donde $v(s, t)$ es la función covarianza. El lado izquierdo de la ecuación anterior es una transformación V de la función de peso ξ . Esta transformación integral actúa u opera sobre ξ , entonces podemos expresar la ecuación-propia directamente como $V\xi = \rho\xi$ donde ξ es una función-propia. Resumiendo, se encuentra que PCA está definido como la búsqueda de un conjunto de funciones de peso ortogonales y normalizados ξ_m . PCA funcional puede ser expresado como el problema del eigenanálisis del operador covarianza V , definido al usar la función covarianza v como el kernel de una transformación integral.

3.1. Análisis de componentes principales regularizados

Si nosotros queremos incorporar suavidad dentro de nuestro procedimiento anterior, entonces tendremos otra meta, controlar la rugosidad de los componentes principales estimados ξ , para que no sea demasiado grande. La clave de la penalización a la rugosidad es hacer explícito este posible conflicto. En el contexto de que estimando la función componente principal, existe un balance entre maximizar la varianza muestral funcional $\langle \xi, V\xi \rangle$, y cuidar la penalidad de la rugosidad $PEN_L(\xi)$ de que sea demasiado grande. Este balance es controlado, por un

parámetro de suavidad $\lambda \geq 0$, el cual regula la importancia del término de penalidad de la rugosidad. Una manera de penalizar la varianza muestral, es dividirla por $\{1 + \lambda PEN_L(\xi)\}$. Así la varianza muestral penalizada: $PCAPSV = \frac{\langle \xi, V\xi \rangle}{\|\xi\|^2 + \lambda PEN_L(\xi)}$. Supongamos que $\{\phi_i\}$ es una base, y sea c_i el vector de coeficientes de la función de datos $x_i(s)$ en la base $\{\phi_i\}$. Sea V la matriz de varianzas y covarianzas de los vectores c_i y $L = D^2$. Se define la matriz \mathbf{J} como $\int \phi \phi'$, cuyos elementos son $\langle \phi_j, \phi_k \rangle$ y la matriz \mathbf{K} con elementos $\langle D^2 \phi_j, D^2 \phi_k \rangle$. La varianza muestral penalizada puede ser escrita como:

$$PCAPSV = \frac{\langle \xi, V\xi \rangle}{\|\xi\|^2 + \lambda \|D^2\xi\|^2} = \frac{y'Vy}{y'\mathbf{J}y + \lambda y'\mathbf{K}y}.$$

La eigenecuación está dada por: $Vy = \rho(\mathbf{J} + \lambda\mathbf{K})y$. Mediante una factorización $\mathbf{L}\mathbf{L}' = \mathbf{J} + \lambda\mathbf{K}$ y si se define a $S = \mathbf{L}^{-1}$. Podemos encontrar una matriz adecuada \mathbf{L} por una descomposición en valores singulares o por una factorización de Cholesky, luego entonces podemos escribir la ecuación como: $(\mathbf{S}\mathbf{V}\mathbf{S}')(\mathbf{L}'y) = \rho\mathbf{L}'y$. El algoritmo obtenido es como sigue:

1. Obtener los vectores de coeficientes c_i de las funciones x_i expandidas en una base.
2. Resolver el sistema $\mathbf{L}a_i = c_i$ para cada i .
3. Llevar a cabo un ACP estándar sobre los vectores de coeficientes a_i .
4. Con los eigenvectores resultantes u , resolver $\mathbf{L}'y = u$ en cada caso y normalizar los vectores y de tal manera que se tenga que $y'\mathbf{J}y = 1$.
5. Expandir a ξ con los coeficientes estimados.

4. Aplicación

El crecimiento de galletas durante su cocinado está influenciado por el tipo de harina y edulcorante utilizado en la masa de la galleta. La galleta gradualmente crece en diámetro

durante el cocinado hasta alcanzar su diámetro máximo. Los datos obtenidos representan mediciones de tres galletas por tipo de producto realizadas cada minuto, empezando en cero minutos y terminando al minuto 6. Cada galleta proviene de un lote de producción diferente, pero las mediciones en el tiempo son mediciones repetidas sobre la misma galleta. A continuación se pueden observar las representaciones funcionales del crecimiento del diámetro de las galletas usando B-splines de grado cuatro, lo cual me genera polinomios de grado 3. Estas 15 curvas (ver Figura 1) obviamente son crecientes, podemos observar también que del minuto 2 aproximadamente al minuto 3 y medio las curvas crecen en una banda muy pequeña a diferencia de la banda que forman a partir del minuto 3 y medio, es aquí donde se puede suponer que la mayor variabilidad se da en la parte final de la cocción de la galleta. La función correlación (ver Figura 2) nos muestra la fuerte asociación que hay entre los tiempos medios del proceso de cocción. Esto se puede ver, trazando una recta sobre las curvas de nivel de la función, esta recta tiene que ser perpendicular a la recta de 45° , y en una vecindad de un punto en particular fijarse que tan rápido o lento decae la correlación, ayudándose con la recta trazada. Es así que en este experimento se puede observar que los tiempos medios van dependiendo de los segundos anteriores, incluso del minuto anterior, esto es contrastante con los tiempos iniciales, donde la correlación decrece de una manera muy rápida. A partir del minuto 4 la correlación empieza a decrecer igual que en los tiempos iniciales, pero en los tiempos finales la correlación empieza a crecer de manera no tan rápidamente. En el análisis de componentes principales podremos ver donde se encuentra la fuente de mayor variación durante el experimento. Tendremos 7 componentes principales, de los cuales los dos primeros explican el 91.3 % de la variación total. La gráfica del primer componente principal (ver Figura 3) nos dice que la variabilidad en el experimento va aumentando de manera considerable a partir del medio minuto de cocción de las galletas y en la parte final es donde se puede observar la mayor variabilidad.

Figura 1

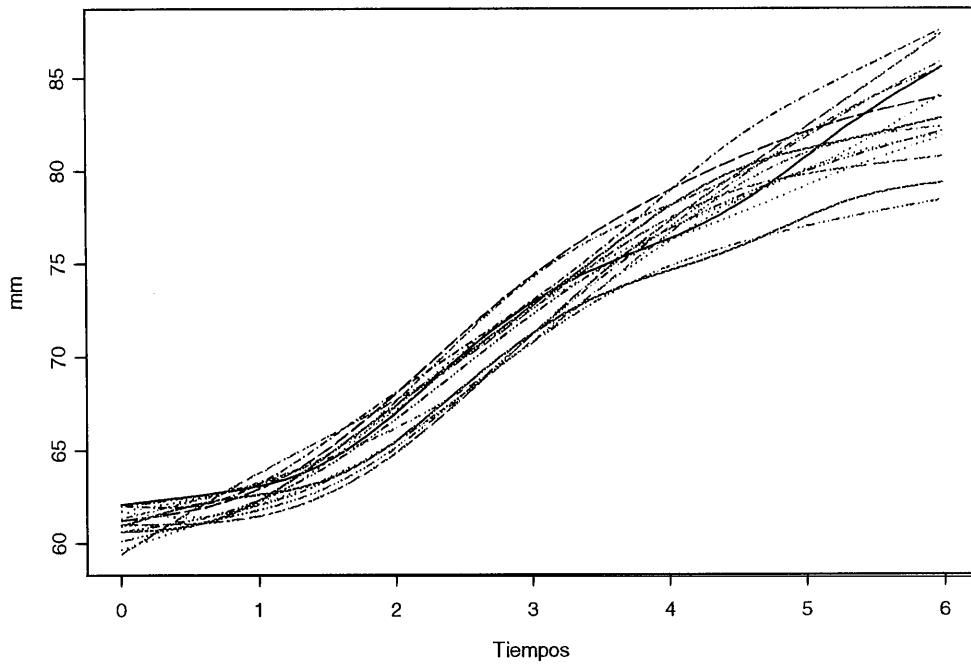


Figura 2

Correlación

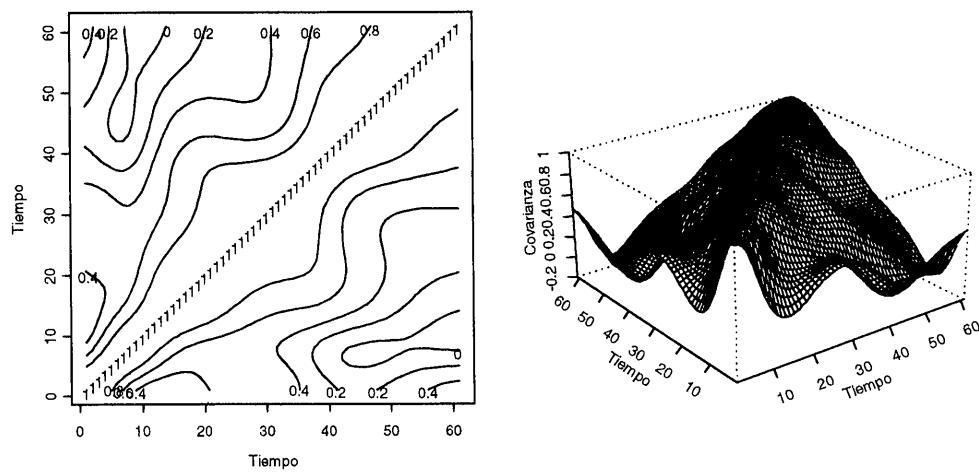
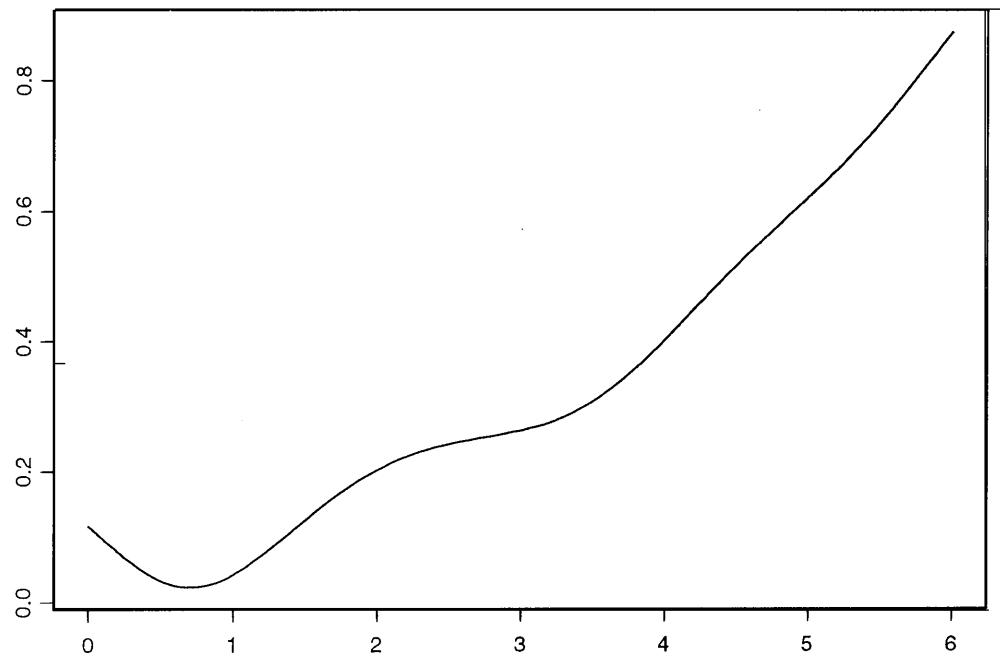


Figura 3

Componente Principal 1



Referencias

Castaño-Tostado, E y Domínguez-Domínguez, J. (2001). *Diseño de Experimentos para el Desarrollo Tecnológico y Mejora Industrial*. Just in Time Press, ISBN 968-7788-16-x.

Ramsay, J.O. y Silverman, B.W. (1997). *Functional Data Analysis*. Springer-Verlag, New York.

Ramsay, J.O. y Dalzell, C.J. (1991). Some tools for Functional Data Analysis. *Journal of the Royal Statistical Society, B*, 3, 539-572.

Ramsay, J.O. (1995). Some tools for the multivariate analysis of functional data. In *Recent Advances in Descriptive Multivariate Analysis*. Edited by Wojtek J. Krajancoswski. Clarendon Press, Oxford.

Ramsay, J.O. y Silverman, B.W. (1999). *S-PLUS Functions for FDA*.

Estimacion casi no Parametrica en mezclas de modelos

Isidro Roberto Cruz Medina

Instituto Tecnológico de Sonora

1. Introducción

Las mezclas de Modelos surgen en innumerables áreas de aplicación en donde se conoce o sospecha que las observaciones provienen de diferentes poblaciones, cada una de las cuales tiene una distribución diferente. Como ejemplo conside que se dispone del registro de estaturas de hombres y mujeres adultos, en este caso la estatura se puede modelar mediante una mezcla de dos distribuciones normales en donde la media de la distribución correspondiente a los hombres es la mayor. En este trabajo se consideran sólo mezclas finitas en donde los componentes provienen de la misma clase paramétrica. Sea Φ la clase de funciones m-dimensionales de donde se formarán las mezclas.

$$\Phi = \{ F(x, \theta), \theta \in \Omega, x \in R^n \}$$

La clase de mezclas finitas se identificará con la clase Ψ de funciones de distribución $G(x)$ definida por:

$$\Psi = \left\{ G(x) : G(x) = \sum_{j=1}^k \pi_j F(x, \theta_j), \pi_j > 0, \sum_{j=1}^k \pi_j = 1, x \in R^n \right\}$$

Se pueden mencionar al menos tres problemas importantes en una mezcla de modelos.

- El primer problema es el de la identificación: ¿Es la mezcla única?
- El segundo problema ocurre cuando el número de componentes es desconocido y se necesita estimarlo.
- El tercer problema consiste en la estimación de parámetros cuando el número de componentes es conocido. Karl Pearson (1894), consideró el problema de estimar a los parámetros de una mezcla de dos distribuciones normales con el método de momentos. Rao (1948) consideró el mismo problema con el método de máxima verosimilitud.

Los problemas computacionales se resolvieron parcialmente con el algoritmo EM desarrollado por Dempster, Laird y Rubin (1977).

2. Estimación con la Distribución Binomial

Hettmansperger y Thomas (2000), proponen un método casi no paramétrico para la estimación del parámetro de la mezcla cuando se dispone de vectores de observaciones. Si Y_j denota el número de veces que las observaciones en el vector j-ésimo es menor o igual que c , y el número de elementos del vector es $m \geq 3$. Entonces para c fija, si $Y_j, j = 1, \dots, n$ es iid con función de probabilidad:

$$h(y) = \lambda b(y; m, F_1(c)) + (1 - \lambda)b(y; m, F_2(c))$$

donde $F_r(c) = \int_{-\infty}^c f_r(x)dx$, $r = 1, 2$ y $b(\cdot; m, p)$ representa la distribución binomial con parámetros m and p . Si $\hat{\lambda}$ denota el EMV (estimador de máxima verosimilitud) de la mezcla de binomiales. Entonces si $n \rightarrow \infty$, por las propiedades asintóticas de los EMV.

$$\sqrt{n}(\hat{\lambda} - \lambda_0) \xrightarrow{D} Z \sim N\left(0, \frac{1}{Eg^2(Y, \lambda_0, c)}\right) \quad (1)$$

donde

$$Eg^2(Y, \lambda_0, c) = \sum_{i=0}^m \frac{[b(i; m, F_1(c)) - b(i; m, F_2(c))]^2}{\lambda_0 b(i; m, F_1(c)) + (1 - \lambda_0) b(i; m, F_2(c))}. \quad (2)$$

Observe que la distribución asintótica del estimador depende sólo débilmente de la mezcla original. La óptima elección de c será el valor que minimice esta varianza asintótica. Si f_1 y f_2 son distribuciones simétricas con la misma forma pero con parámetros de localización diferentes M_1 y M_2 para $\lambda_0 = 0,5$ la esperanza $Eg^2(Y, \lambda_0)$ es simétrica con respecto a $c = \frac{M_1+M_2}{2}$ y tiene un máximo local en este valor.

Gráficas de la varianza asintótica para varias distribuciones muestran que este valor proporciona el mínimo global y que además esta varianza asintótica es poco sensible a la elección del punto de corte c ; esta propiedad es muy importante porque en una situación real el punto de corte c debe ser estimado y estas gráficas muestran que la estimación no necesita ser precisa. El Cuadro 1 presenta la varianza asintótica para algunos valores de m (número de componentes en el vector original de observaciones) y λ para el óptimo valor de c . Note que esta varianza decrece con m para todos los valores de λ_0 y que el punto de corte c es simétrico con respecto a $\frac{M_1+M_2}{2}$ y λ en el sentido de que para $\lambda_0 = ,1$ y $\lambda_0 = ,9$ el punto óptimo es simétrico con respecto a 0.5.

Cuadro 1. Varianzas Asintóticas (*V.A*) de $\sqrt{n}(\hat{\lambda} - \lambda_0)$ para diferentes valores de λ_0 and m .

λ_0/m	3 (<i>c, V.A</i>)	4	10	15
0.1	0.1392, 0.4239	0.1950, 0.3209	0.3409, 0.1528	0.3873, 0.1213
0.3	0.3526, 0.6537	0.3854, 0.5301	0.4406, 0.3069	0.4547, 0.2597
0.5	0.5000, 0.7227	0.5000, 0.5937	0.5000, 0.3560	0.5000, 0.3046
0.7	0.6473, 0.6537	0.6145, 0.5301	0.5593, 0.3069	0.5452, 0.2597
0.9	0.8607, 0.4239	0.8049, 0.3209	0.6590, 0.1528	0.6126, 0.1213

3. Estimación con el Método Multinomial

El método binomial puede generalizarse fácilmente, si incluimos más puntos de corte c tendremos una distribución multinomial. La primera ventaja de este método es que sólo se necesita que el vector de observaciones tenga al menos componentes para la estimación de una mezcla de dos multinomiales (para una mezcla de dos binomiales se requieren al menos tres componentes). La segunda ventaja es que mediante la desigualdad de Cauchy-Schwartz se puede demostrar que para valores pequeños de m :

$$V_{M_{r-1}}\{\sqrt{n}(\hat{\lambda} - \lambda_0)\} \geq V_{M_r}\{\sqrt{n}(\hat{\lambda} - \lambda_0)\}$$

donde V_{M_r} es la varianza asintótica de la distribución multinomial con r clases. Esta desigualdad nos muestra la ventaja de utilizar más clases, esto es, la ventaja de la distribución trinomial sobre la binomial y en general la ventaja de la multinomial con r clases sobre la multinomial con $r - 1$.

Para valores grandes de m se puede demostrar que para la distribución binomial y en general para la multinomial:

$$V\{\sqrt{n}(\hat{\lambda} - \lambda_0)\} = \lambda_0(1 - \lambda_0)$$

Este resultado indica que la máxima información que cualquiera de estos métodos puede proporcionar en una mezcla de modelos es la correcta identificación del número de observaciones de cada componente. Este número tiene una distribución binomial con parámetro λ_0 y la varianza asintótica alcanza la cota inferior de Cramér-Rao para un estimador insesgado dada por Hill (1963).

El Cuadro 2 muestra las varianzas asintóticas para los métodos binomial, trinomial y tetranomial, observe que la máxima disminución de la varianza asintótica se tiene al pasar de la distribución binomial a la trinomial.

Cuadro 2. Varianzas Asintóticas ($A.V$) de $\sqrt{n}(\hat{\lambda} - \lambda_0)$ para los métodos binomial, trinomial, tetranomial y pentanomial

m	Binomial	Trinomial	Tetranomial	Pentanomial
3	0.7227	0.6144	0.5819	0.5675
4	0.5937	0.5106	0.4857	0.4745
6	0.4621	0.4058	0.3889	0.3813
10	0.3560	0.3228	0.3130	0.3087
20	0.2810	0.2677	0.2642	0.2626
30	0.2614	0.2553	0.2539	0.2533
60	0.2508	0.2502	0.2501	0.2501
100	0.2500	0.2500	0.2500	0.2500

4. Conclusiones

La varianza asintótica del parámetro de la mezcla es poco sensible a cambios en el punto de corte c , esto significa que la estimación del punto de corte c no necesita ser muy precisa.

El método trinomial es más eficiente que el binomial y en general el multinomial con r clases es más eficiente que el multinomial con $r - 1$ clases.

Para m (número de componentes en el vector de observaciones) y n (número de vectores) grandes todos los métodos multinomiales son equivalentes.

Referencias

- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, **39**, 1-38.
- Hettmansperger, T. P., and Thomas H. 2000. Almost nonparametric inference for repeated measures in mixture models. *Journal of the Royal Statistical Society B*, **62**, 811-825.
- Hill, B. M. 1963. Information for estimating proportion in the mixtures of exponential and normal distributions. *J. Amer. Statist. Assoc.*, **58**, 918-932.
- Pearson K. 1894. Contributions to the mathematical theory of evolution. *Phil. Trans. A*, **185**, 71-110.
- Rao. C. R. 1948. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society B*, **10**, 159-203.

A Statistical Method for the Determination of the Appropriate Order in a General Class of Time Series Models

Contreras Cristán Alberto

J. M. González Barrios

IIMAS, UNAM

1. Introduction

Let $\{\epsilon_t ; t = 1, 2, \dots\}$ be a sequence of independent and identically distributed random variables and consider the process $X_t = h(\epsilon_t, \epsilon_{t-1}, \dots, \epsilon_{t-q})$, where h is a real valued function and q is a nonnegative integer. When this function is linear, and under the assumption of the existence of second order moments for the noise sequence $\{\epsilon_t\}$, this model is known as linear moving average of order q , (MA(q)). Other choices of the function h exist in the literature (see e.g. Robinson 1977, Robinson and Zaffaroni 1997 and Tong 1990) and the corresponding models are known as nonlinear moving average models NLMA(q). For the MA(q) model, it is well known (see for example Box and Jenkins 1976) that the appropriate order q can be determined by examination of the autocorrelation sequence computed from a sample X_1, X_2, \dots, X_N of size N of such process. The aim of this work is to propose an alternative way to determine the appropriate order q for the time series models described in the last paragraph

1.1. A stochastic dependence statistic

Let (Ω, \mathcal{F}, P) be an arbitrary probability space and let X_1, X_2, \dots, X_n be n random variables defined in this probability space. We define as a multidimensional dependency measure

$$\delta_{X_1, X_2, \dots, X_n} = \sup_{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n} |F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) - \prod_{i=1}^n F_{X_i}(x_i)|, \quad (3)$$

where F_{X_1, \dots, X_n} is the joint distribution function of the X_i 's and F_{X_i} is the distribution function of X_i . Assume we have an n -dimensional sample of size m , $\mathbf{X}_i = (X_{i,1}, X_{i,2}, \dots, X_{i,n})$ for each $i = 1, 2, \dots, m$, coming from a joint distribution $F_{\mathbf{X}}(x_1, x_2, \dots, x_n)$ with margins given by $F_{X_j}(x_j)$, $j = 1, 2, \dots, n$. Denote by $F_m(x_1, \dots, x_n)$ the joint empirical distribution function and by $F_{m,j}(x_j)$ the empirical distributions of X_j for $j = 1, 2, \dots, n$. We will use as a sample multidimensional dependency measure the following,

$$\delta_{X_1, X_2, \dots, X_n}^m := \sup_{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n} |F_m(x_1, x_2, \dots, x_n) - \prod_{j=1}^n F_{m,j}(x_j)|. \quad (4)$$

Then this sample version mimics the population version of the dependency measure defined above. In fact this proposal makes sense since when $m \rightarrow \infty$, F_m approaches the population joint distribution, and $F_{m,j}$ approaches the population margin distribution. For a detailed discussion and study of $\delta_{X_1, X_2, \dots, X_n}^m$ see Fernández-Fernández and González-Barrios (2001).

Given a random sample of size m of n -dimensional random vectors $\mathbf{X}_i = (X_{i,1}, X_{i,2}, \dots, X_{i,n})$ for $i = 1, \dots, m$ where each \mathbf{X}_i comes from a common continuous distribution $F_{\mathbf{X}}(\cdot)$; we want to test whether the corresponding coordinates are independent. The test we propose includes the following steps.

- Evaluate the statistic $\delta_{X_1, X_2, \dots, X_n}^m$ for the given sample.

- Reject independence if the statistic is greater than the $1 - \alpha$ quantile of the exact or approximated distribution. Otherwise, we do not reject the independence hypothesis.

In the next section, we shall use this statistic for $n = 2$ and for m given in terms of the sample size of a given time series data set.

2. Determining the appropriate order of the model

We assume that our time series model is

$$X_t = h(\epsilon_t, \epsilon_{t-1}, \dots, \epsilon_{t-q}), \quad (5)$$

where $\{\epsilon_t\}$ is a sequence of independent and identically distributed continuous random variables, which for the moment will be assumed to have zero mean and unit variance.

Given a sample X_1, X_2, \dots, X_N , the procedure would include the next steps :

- 1 Choose an integer m such that $N > 1 + (2m - 1)(K + 1)$, where K is assumed to be large enough such that $K > q$ and q is the order of the model. The integer K plays the same role as the maximum lag studied for auto-correlations.
- 2 For each $k = 0, 1, 2, \dots, K$ form the bivariate samples of size m in equation (5).

$$\begin{aligned} Y_1^k &= (X_1, X_{1+(k+1)}) \\ Y_2^k &= (X_{1+2(k+1)}, X_{1+3(k+1)}) \\ &\dots \\ Y_m^k &= (X_{1+2(m-1)(k+1)}, X_{1+(2m-1)(k+1)}), \end{aligned} \quad (5)$$

- 3 For each of these bivariate samples test for dependence of their components at significance level α , using the dependence test described in section 2.

- 4 The order of the model q is taken as the maximum integer in the range $1, 2, \dots, K$, where independence of the coordinates of the corresponding sample is rejected at level α .

2.1. Determining the order of MA(q) processes

The MA(q) model is given by

$$X_t = \sum_{j=0}^q \theta_j \epsilon_{t-j}, \quad (6)$$

where $\theta_j \in \mathbb{R}$ for $j = 1, \dots, q$ and $\theta_0 \equiv 1$.

We have simulated these processes for $q = 2$ and for different values of the parameters $\{\theta_j, j = 1, \dots, q\}$. For this simulation we repeated 1000 times the next steps:

- Simulate $N = 400$ observations of the process $\{X_t\}$.
- Compute $K = 4$ terms of the auto-correlation sequence for the sample X_1, \dots, X_{400} .
- For each $k = 0, 1, 2, 3$, build the two dimensional samples in equation (5). From the values of N and K used, we obtain $m = 50$ bivariate vectors per sample. From these compute the value of the $\delta_{Y^k}^m$ statistic for each k .

Thus, for $n = 2$, the distribution of δ^{50} was approximated by means of simulation as described in Fernández-Fernández and González-Barrios (2001).

Using the statistic $\delta_{Y^k}^m$, and for a fixed significance level α , we counted the number of times (out of the 1000 simulations) where independence of the bivariate sample corresponding to

$k = q$ is rejected, denote this number by i_1 . From these i_1 cases, we counted the number of times where independence of the bivariate samples corresponding to $k = q+1$ is not rejected, denote this new counting by i_2 . Then we repeated this last step for $k = q+2, q+3, \dots, K$. The last counting corresponding to $k = K$ gives us the number of times that our test procedure detects the right order of the model. This is the same procedure for detection of the order using auto-correlations (ACF).

For the MA(2) models in Table 1 (corresponding to $\theta_1 = 0,3$ and $\theta_2 = 0,8$) and Table 2 (corresponding to $\theta_1 = 0,5$ and $\theta_2 = 0,3$), we can observe that the empirical power of our method for the detection of the order depends on the parameters of the model. The first of these models follows a more defined structure of stochastic dependence at lag 2, whereas the second presents confusion in between dependence at lag 1 or 2. These features are captured by our method, but since these are linear models, the sample auto-correlation works better in this case.

Cuadro 1: empirical power for MA(2) model with $\theta_1 = 0,3$ and $\theta_2 = 0,8$.

α	ACF	δ_X
0.10	819	653
0.05	901	635

Cuadro 2: empirical power for MA(2) model with $\theta_1 = 0,5$ and $\theta_2 = 0,3$.

α	ACF	δ_X
0.10	815	271
0.05	896	209

2.2. Determining the order of NLMA(q) processes

Using the same sample size N , we repeated the simulation process in the last section for the next nonlinear moving average processes:

- (a) The non-linear moving average process studied by Robinson (1977)

$$X_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-1} \epsilon_t \quad (7)$$

- (b) The non-linear moving average process

$$X_t = \epsilon_t / 10 - \theta_1 \epsilon_{t-1}^3 \quad (8)$$

Table 3 displays results of the simulation and tests for the non-linear process in (7), with parameters $\theta_1 = 0,2$ and $\theta_2 = 1$. In this case our method performs better than the sample auto-correlation. This was expected since for θ_1 close to zero, the X_t in (7) are nearly uncorrelated, but if $\theta_2 \neq 0$ they are not independent (see e.g. Robinson and Zaffaroni 1997).

An important note is that this statistical test requires the estimation of θ_1 and θ_2 . This has been done using the (moment) estimators proposed by Robinson (1977). These estimates may fail to exist according to section 2 of Robinson's paper. From the 1000 simulated series, only in 890 cases these estimates can be computed, so the results in Table 3 are based on these 890 cases.

Table 4 displays results of the simulation for the non-linear process in (8), with parameter $\theta_1 = 1$. As before, for the non-linear model (7) our method seems to perform better. In this case, the estimates for θ_1 are obtained with a similar method to that used for Robinson's estimators for the model (7). These estimates are needed for the statistical tests of significance for the auto-correlation coefficients.

Cuadro 3: empirical power for nonlinear model (6) with $\theta_1 = 0.2$ and $\theta_2 = 1$

α	ACF	δ_X
0.10	12	233
0.05	7	162

Cuadro 4: empirical power for nonlinear model (7) with $\theta_1 = 1$

α	ACF	δ_X
0.10	2	351
0.05	2	340

3. Concluding remarks

We presented a novel method for the statistical selection of the order q in general moving average models. This method does not make any assumptions on the distribution of the noise sequence $\{\epsilon_t\}$ and we certainly think this is a major advantage, since the theory for sample auto-correlations is strongly based on the assumption of the existence of second order moments. For the case of nonlinear models our method provides a new and simple way to determine the order of the model. It follows analogous steps to those followed in the sample-correlation analysis and presents significative improvements against this classical method.

References

- Box, G.E.P. and Jenkins, G.M. (1976). *Time Series Analysis forecasting and control*. Holden-Day, Oakland, CA.

Brockwell, P.J. and Davis, R.A.(1991). *Time Series: Theory and Methods*. Springer-Verlag: New York.

Fernández-Fernández, B. and González-Barrios, J.M. (2001). Multidimensional Dependency Measures. *Preimpreso No. 101, IIMAS*.

Robinson, P.M. (1977). The estimation of a nonlinear moving average model. *Stochastic Processes and their Applications*. **5**, 81-90.

Robinson, P.M. and Zaffaroni, P. (1997). Modelling Nonlinearity and Long Memory in Time Series. In *Nonlinear Dynamics and Time Series*, (Cutler, C.D. and Kaplan, D.T., editors), Fields Institute Communications. Series published by the AMS.

Tong, H. (1990). *Non-linear Time series*, Oxford University Press: Oxford.

Construcción de Clusters en Series de Tiempo

José Ramón Domínguez Molina

Graciela Ma. González Farías

Centro de Investigación en Matemáticas A.C.

1. Introducción

El objetivo de este trabajo es estudiar, diseñar e implementar un entorno adecuado para el estudio de un gran número de series de tiempo por ejemplo, ventas de compañías refresqueras, consumo de bienes energéticos, datos de sismología, información ambiental, etc., de forma tal, que nos permita agrupar las series de tiempo en k grupos lo más homogéneos posible y en cada grupo encontrar una familia de modelos que permita desarrollar una metodología automática de modelización.

En la Sección 2 se presenta una revisión de métodos para formación de conglomerados (clusters) con series de tiempo, en la Sección 3 se propone una medida de disimilitud entre series de tiempo, la cual tomará en cuenta la estructura de dependencias dentro de las series basada en el trabajo de Piccolo (1990). Con esta medida definida, se podrán entonces aplicar los métodos clásicos de agrupación jerárquicos y no jerárquicos. En la Sección 4, se presentan un ejemplo con datos reales de velocidad de viento de 12 estaciones meteorológicas de Irlanda.

Cabe mencionar que se desarrolló el software necesario para su adaptación y las interfaces que lo hacen de fácil acceso a los usuarios incluyendo un módulo de detección de observaciones aberrantes para la modelización automática de pronósticos basado en la propuesta de la Sección 3 y en la extensión de un resultado dado por Díaz-García y González-Farías (2001).

El software fue desarrollado en S-plus para facilitar las interfaces pero es fácilmente transferible a lenguajes de programación básicos como C o Fortran, que agilizan su implementación. El software está disponible bajo requisición en: jrdguez@cimat.mx.

2. Antecedentes

Kakizawa *et al.* (1998) sugieren que las similitudes y diferencias entre series de tiempo pueden ser caracterizados en términos de la estructura de covarianza o equivalentemente por el espectro. La primera es la medida de Kullback-Leiber (KL), la cual está dada por,

$$J(\mathbf{f}_T; \mathbf{g}_T) = \frac{1}{2T} \sum_s (\mathbf{f}_T(\lambda_s) \mathbf{g}_T^{-1}(\lambda_s) + \mathbf{g}_T(\lambda_s) \mathbf{f}_T^{-1}(\lambda_s) - 2)$$

y la segunda es la medida dada por la información simétrica de divergencia de Chernoff

$$JB_\alpha(\mathbf{f}_T; \mathbf{g}_T) = \frac{1}{2T} \sum_s \left(\log \frac{|\alpha \mathbf{f}_T(\lambda_s) + (1-\alpha) \mathbf{g}_T(\lambda_s)|}{|\mathbf{g}_T(\lambda_s)|} + \log \frac{|\alpha \mathbf{g}_T(\lambda_s) + (1-\alpha) \mathbf{f}_T(\lambda_s)|}{|\mathbf{f}_T(\lambda_s)|} \right)$$

donde $\lambda_s = \frac{2\pi s}{T}$, $s = 1, 2, \dots, T$, $\alpha \in (0, 1)$, $\mathbf{f}_T(\lambda_s)$ y $\mathbf{g}_T(\lambda_s)$ son espectros de dos diferentes series de tiempo, calculadas mediante el estimador espectral suavizado.

Piccolo (1990), propone una métrica para series de tiempo donde un modelo *ARIMA* es ajustado a un gran número de series de tiempo con el propósito de hacer predicciones y ajustes estacionales. Utilizando la metodología propuesta por Piccolo es necesario escoger primero un procedimiento estadístico para el ajuste de modelos *ARIMA*,

$$\phi_p(B) \Phi_P(B^s) (1-B)^d (1-B^s)^D Z_t = \theta_q(B) \Theta_Q(B^s) a_t \quad (9)$$

Suponiendo que el modelo (9) es invertible, esto es, que las raíces del polinomio $\theta(B)$ estén fuera del círculo unitario, entonces $Z_t \in \mathcal{L}$, donde \mathcal{L} es de la clase de modelos invertibles

ARIMA. Entonces

$$Z_t = \pi_1 Z_{t-1} + \pi_2 Z_{t-2} + \cdots + a_t \quad (10)$$

donde el modelo $AR(\infty)$ es definido por $\pi(B) = \varphi(B)/\theta(B) = 1 - \pi_1 B - \pi_2 B^2 - \dots$.

La métrica sobre \mathcal{L} la define por la distancia

$$d(X, Y) = \left\{ \sum_{j=1}^{\infty} (\pi_{j,x} - \pi_{j,y})^2 \right\}^{1/2}$$

3. Modificación del procedimiento de Piccolo

La metodología propuesta por Piccolo requiere en primer lugar ajustar un modelo del tipo *ARIMA* para cada una de las series y posteriormente calcular los valores de π_j , $j = 1, 2, \dots$.

Si se requiere hacer esto para un gran número de series, resulta una tarea complicada y pesada computacionalmente. En algunas de las aplicaciones la formación de clusters ayuda a entender también características que no habían sido tomadas en cuenta y permite identificar grupos racionales para la creación de escenarios comunes, que pueden resultar muy útiles en el proceso de predicción.

Por lo anterior, surge la idea de proponer una medida entre series que no requiera modelarlas antes de construir una medida de disimilitud para formar clusters de modelos semejantes. Con esto lo que se pretende es primero formar grupos y después modelar cada grupo con una familia mucho más pequeña de modelos tipo *ARIMA* que se puedan ajustar. Esto incluye el caso de no-estacionariedad en la media, la no inclusión de covariables para limpiar las series antes de formar los grupos, como por ejemplo, variables para modelar valores atípicos, tendencias, etc. Con la representación de Z_t en (10), se tiene que el proceso puede

aproximarse por

$$Z_t = \sum_{l=1}^h \pi_l Z_{t-l} + a_t$$

para algún h , ya que $\pi_j \rightarrow 0$ cuando $j \rightarrow \infty$.

El estimador de máxima verosimilitud condicional de π coincide con el estimador de mínimos cuadrados, y está dado por

$$\hat{\pi} = (X^T X)^{-1} X^T Z^*$$

donde $X = [X_{h+1}, X_{h+2}, \dots, X_n]^T$ y $Z^* = [Z_{h+1}, Z_{h+2}, \dots, Z_n]^T$.

Haciendo esto se obtendrá una aproximación de los coeficientes de π . Por lo que una métrica sobre \mathcal{L} puede definirse ahora por la distancia,

$$d(X, Y) = \left\{ \sum_{j=1}^h (\hat{\pi}_{j,x} - \hat{\pi}_{j,y})^2 \right\}^{1/2}$$

4. Aplicación

Haslett y Rafter (1989) estudian datos de viento de 12 estaciones meteorológicas, tomados cada hora durante los años 1961-1978 en Irlanda. La velocidad del viento fue registrada en knots (1 knot = 0.5148 m/s). El objetivo de su artículo es desarrollar métodos para la evaluación del poder del viento en zonas de Irlanda donde no hay estaciones meteorológicas.

En este trabajo el objetivo sólo será encontrar grupos de ciudades donde las velocidades del viento sean semejantes. Para esto se utilizará un orden de $h = 50$ debido a que las series parecen provenir de un proceso de memoria larga, lo cual ocasiona dependencias en ordenes altos de rezagos. El método de agrupación fue el jerárquico y para el enlace se utilizó la

técnica de la liga completa, el cálculo de la distancia entre los coeficientes π 's fue la Euclídea. En la Figura 1 se tiene el dendograma y la ubicación de los grupos formados de las estaciones meteorológicas de Irlanda respectivamente. Se observa que las ciudades se agrupan por su ubicación geográfica, esto es, Malin Head, Clones, Mullingar y Dublin están al norte; Valentia, Roche's Pt, Shannon, Birr y Kilkenny en el sur; Bellmullet y Claremorris al oeste; y por último Rosslare formo su propio grupo y esta ubicada al este y en la costa. Resulta interesante que Haslett y Raftery (1989) en su artículo eliminan de su análisis a la ciudad de Rosslare debido a que su correlación con las demás estaciones es muy baja y por lo tanto no tiene influencia sobre las demás estaciones, esto se ve reflejado en este análisis al no agruparse con las demás estaciones.

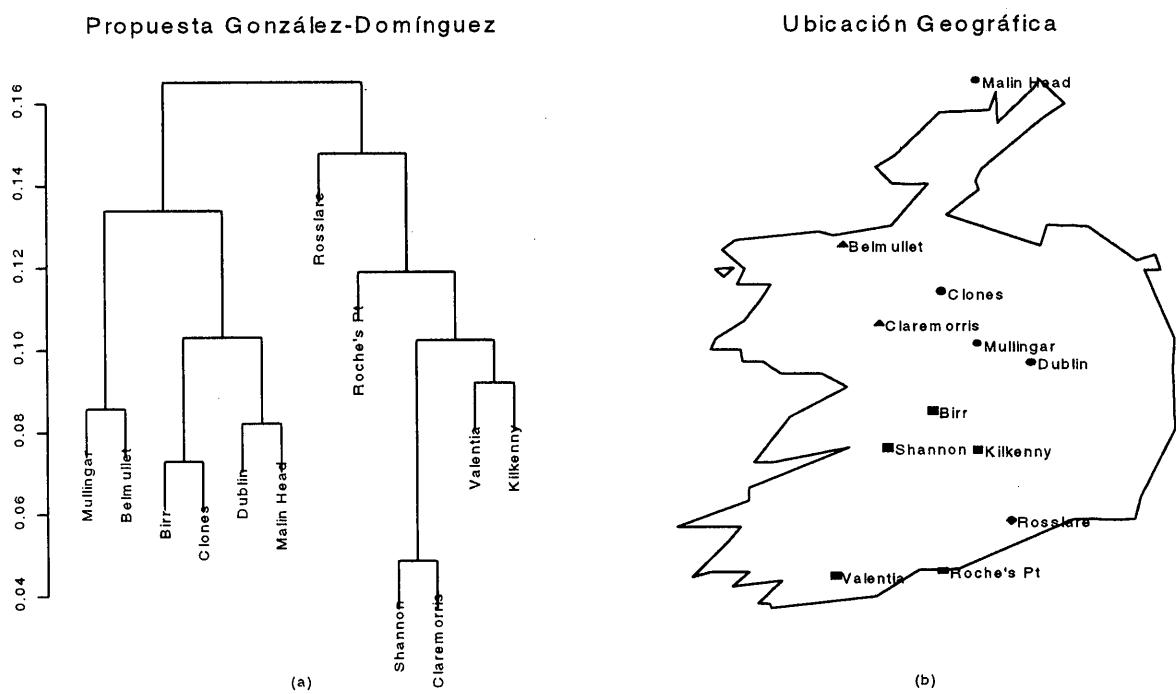


Figura 1. (a) Dendrograma de las estaciones meteorológicas de Irlanda. (b) Estaciones meteorológicas de Irlanda agrupadas.

5. Conclusión

La medida de distancia entre series propuesta en este trabajo toma en cuenta la estructura de correlaciones entre las series, por lo tanto es común que los grupos formados no tengan semejanza a simple vista. Esto se debe a que dos series que provengan del mismo modelo *ARIMA* pueden ser muy distintas cuando se grafican.

Es recomendable pero no necesario estacionarizar las series, ya que con series no estacionarias se requiere un orden de p muy grande para recoger toda la estructura de correlaciones mientras que, con series estacionarias, se pueden obtener modelos que cumplan con el principio de parsimonia. El procedimiento propuesto en este trabajo es muy sencillo de implementar porque solamente es necesario tener una idea general del tipo de modelos *ARIMA* con los cuales las series serían modeladas de una manera adecuada, y con esto proponer un modelo $AR(p)$ que sea lo suficientemente grande para aproximar de una manera adecuada a π . Habiendo decidido el orden p , a través de mínimos cuadrados ordinarios se pueden estimar los coeficientes del $AR(p)$, el cálculo de la matriz de distancias es trivial, entonces cualquier técnica de agrupamiento que utilice una matriz de disimilitudes es aplicable. En este trabajo el objetivo fue definir una medida de disimilitud entre series de tiempo para así poder agruparlas, pero esta medida puede ser utilizada en otros procedimientos que necesiten el concepto de medidas entre series, por ejemplo, escalamiento multidimensional.

Este trabajo se llevó a cabo bajo el proyecto CONACyT No. 32393-E, “Modelos Estadísticos Espacio-Temporales en Medio Ambiente”.

Referencias

Díaz-García J. A. y González-Farías, G. (2001). “A Note on the Cook’s Distance”, *Enviado para su publicación*.

Domínguez-Molina J.R (2001). “Formación de Conglomerados y valores atípicos en Series de Tiempo”. *Tesis no publicada*. Maestría en Estadística. Programa conjunto CIMAT- Universidad de Guanajuato.

Haslett, J. y Raftery, A. (1989). “Space-Time Modelling with Long-Memory Dependence: Assessing Ireland’s Wind Power Resource” *Applied Statistics*, **38**, 1-50.

Kakizawa, Y., Shumway. R. H. y Taniguchi. M. (1988). “Discrimination and Clustering for Multivariate Time Series”, *Journal of the American Statistical Association*, **93**, 328-339.

Piccolo. D. (1990). “A Distance Measure for Classifying ARIMA Models”, *Journal of Time Series Analysis*, **11**, 154-164.

Componentes Principales y Medidas de Dependencia

José M. González Barrios ¹

Silvia Ruiz Velasco ²

IIMAS, UNAM

1. Introducción

Una de las técnicas más utilizadas en la Estadística Multivariada actual es la de componentes principales, cuya finalidad es encontrar combinaciones lineales con mayor variabilidad. Otra de las finalidades de esta técnica es poder reducir el número de variables aleatorias lo que se conoce como reducción de dimensionalidad.

Por otra parte, recientemente se han estudiado un buen número de estadísticas de dependencia multivariada basándose en la Teoría de Cúpulas. En este trabajo analizamos y comparamos una de estas medidas, la estadística δ^m (definida en la siguiente sección), con los resultados que se obtienen al hacer un análisis de componentes principales. Como se puede ver en los resultados observados en este trabajo mediante simulaciones y trabajando con datos reales, los resultados obtenidos mediante componentes principales y los que provienen de la estadística δ^m son muy compatibles en el caso de combinaciones lineales y aproximaciones a éstas. Sin embargo, cuando la relación entre las variables no es lineal (ni siquiera bajo transformaciones) la estadística δ^m detecta esta relación, mientras que la técnica de componentes principales da resultados ambiguos.

¹Investigación parcialmente apoyada por Conacyt Proyecto 32705-E y PAPIIT Proyecto IN-101198

²Investigación parcialmente apoyada por Conacyt Proyecto 32297-E y PAPIIT Proyecto IN-101198

2. Medidas de dependencia multivariadas

Sea (Ω, \mathcal{F}, P) cualquier espacio de probabilidad y sean:

$$X_i : (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R})) \quad \text{para} \quad i \in I_n = \{1, 2, \dots, n\}$$

n variables aleatorias. Definimos:

$$\delta_{X_1, X_2, \dots, X_n} = \sup_{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n} |F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) - \prod_{i=1}^n F_{X_i}(x_i)|, \quad (11)$$

donde F_{X_1, \dots, X_n} es la función de distribución conjunta de las X_i 's y F_{X_i} es la función de distribución de X_i . Es fácil mostrar (Fernández y González (2001)) que $\delta_{X_1, X_2, \dots, X_n}$ satisface

Teorema 2.1 *Sea (Ω, \mathcal{F}, P) un espacio de probabilidad y X_i variables aleatorias reales para $i = 1, 2, \dots, n$, y $\delta_{X_1, X_2, \dots, X_n}$ definida como (1). Entonces $\delta_{X_1, X_2, \dots, X_n}$ satisface:*

a) $\delta_{X_1, X_2, \dots, X_n} = \delta_{X_{\sigma(1)}, X_{\sigma(2)}, \dots, X_{\sigma(n)}}$ para cada σ permutación de I_n .

b) $\delta_{X_1, X_2, \dots, X_n} = 0$ si y sólo si X_1, X_2, \dots, X_n son independientes.

c) Para cada $x_i \in \mathbb{R}$, $i \in I_n$

$$-\left(\frac{n-1}{n}\right)^n \leq F_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) - \prod_{i=1}^n F_{X_i}(x_i) \leq \left(\frac{1}{n}\right)^{\frac{1}{n-1}} \left(1 - \frac{1}{n}\right) < 1.$$

De aquí se sigue que $0 \leq \delta_{X_1, X_2, \dots, X_n} \leq 1$.

d) $0 \leq \delta_{X_1, X_2} \leq \delta_{X_1, X_2, X_3} \leq \dots \leq \delta_{X_1, X_2, \dots, X_{n-1}} \leq \delta_{X_1, X_2, \dots, X_n}$.

Lemma 2.2 Sean $X_1, X_2, \dots, X_n, X_{n+1}$ $n + 1$ variables aleatorias definidas sobre (Ω, \mathcal{F}, P) .

Si X_{n+1} es independiente de (X_1, \dots, X_n) , entonces:

$$\delta_{X_1, X_2, \dots, X_n} = \delta_{X_1, X_2, \dots, X_n, X_{n+1}}.$$

La prueba de éste y los resultados subsecuentes de esta sección pueden encontrarse en (Fernández y González-Barrios (2001)).

Supongamos que se tiene una muestra n -dimensional de tamaño m , $\mathbf{X}_i = (X_{i_1}, X_{i_2}, \dots, X_{i_n})$ para $i = 1, 2, \dots, m$, proveniente de una distribución conjunta $F_{\mathbf{X}}(x_1, x_2, \dots, x_n)$ con marginales dadas por $F_{X_j}(x_j)$, $j = 1, 2, \dots, n$. Denotemos por $F_m(x_1, \dots, x_n)$ la función de distribución conjunta empírica y por $F_{m,j}(x_j)$ la función de distribución empírica de X_j para $j = 1, 2, \dots, n$. Proponemos como una medida muestral de dependencia multivariada la siguiente:

$$\delta_{X_1, X_2, \dots, X_n}^m := \sup_{(x_1, x_2, \dots, x_n) \in \mathbb{R}^n} |F_m(x_1, x_2, \dots, x_n) - \prod_{j=1}^n F_{m,j}(x_j)|. \quad (12)$$

Entonces esta versión muestral imita a la medida poblacional definida arriba. De hecho esta propuesta tiene sentido ya que cuando $m \rightarrow \infty$, F_m se aproxima a la función de distribución conjunta, y $F_{m,j}$ se aproxima a la función de distribución marginal.

Si tenemos m muestras independientes de vectores n -dimensionales cuyas coordenadas son independientes se tiene que:

$$\delta_{X_1, X_2, \dots, X_n}^m \rightarrow 0$$

casi seguramente cuando $m \rightarrow \infty$.

Proposición 2.3 Sea $X_{i_1}, X_{i_2}, \dots, X_{i_n}$ para $i = 1, 2, \dots, m$ y sean $f_j : R \rightarrow R$ para $j = 1, 2, \dots, n$ funciones monótonas estrictamente crecientes. Entonces:

$$i) \delta_{X_1, X_2, \dots, X_n}^m = \delta_{f_1(X_1), f_2(X_2), \dots, f_n(X_n)}^m.$$

ii) Para cada $n \geq 2$ y cualquier $m > n$

$$\delta_{X_{i_1}, X_{i_2}}^m \leq \dots \leq \delta_{X_{i_1}, X_{i_2}, \dots, X_{i_k}}^m \leq \dots \leq \delta_{X_1, X_2, \dots, X_n}^m$$

Para toda $k \leq n$ y cualesquiera subíndices $i_1, i_2, \dots, i_k \in \{1, 2, \dots, n\}$. Además

$$\delta_{X_1, X_2, \dots, X_n}^m \leq \max_{0 \leq k \leq m} \left(\frac{k}{m} - \left(\frac{k}{m} \right)^n \right) =: \mathcal{K}_{m,n}.$$

El resultado principal que se usará en el resto de este trabajo es:

Teorema 2.4 Sea $(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$ una muestra aleatoria de tamaño m proveniente de una distribución conjunta $F_{\underline{X}}(x, y)$, donde $\underline{X} = (X, Y)$. Entonces:

i) Si X y Y son variables aleatorias continuas, entonces $\delta_{X,Y}^m \neq 0$ c.s. para toda $m \geq 2$.

ii) Si $F_m(x_i, y_j) = k/m$, para $0 \leq k \leq m$, $F_{m,1}(x_i) = k_1/m$ y $F_{m,2}(y_j) = k_2/m$. Entonces $\min\{k_1, k_2\} \geq k$ y $m + k - k_1 - k_2 \geq 0$.

iii) Supongamos que $F_{\underline{X}}(x, y) = xy$, es decir, X y Y son variables aleatorias independientes con distribución uniforme $(0, 1)$. Entonces siempre es posible encontrar la distribución de $\delta_{X,Y}^m$ para cada $m \geq 2$.

iv) Supongamos que $F_{\underline{X}}(x, y) = F_X(x)F_Y(y)$, es decir, X y Y son variables aleatorias independientes con funciones de distribución correspondientes $F_X(\cdot)$ y $F_Y(\cdot)$, las cuales supon-

dremos ser continuas. Entonces:

$$\delta_{X,Y}^m \stackrel{\text{dist}}{=} \delta_{F_X(X), F_Y(Y)}^m.$$

Una versión multivariada a diferencia de la versión bivariada del teorema anterior también resulta válida.

Este último Teorema, en el caso de variables aleatorias continuas, prueba que la distribución del estimador $\delta_{X_1, X_2, \dots, X_n}^m$ bajo la hipótesis de independencia, sólo depende del tamaño de muestra y la dimensión, pero no de la distribución subyacente.

Usando la propiedad de la estadística $\delta_{X_1, X_2, \dots, X_n}^m$ se puede implementar una prueba de independencia multivariada utilizando su distribución, o en caso necesario aproximar su distribución mediante simulaciones.

Dada una muestra $\mathbf{X}_i = (X_{i,1}, X_{i,2}, \dots, X_{i,n})$ para $i = 1, \dots, m$ de vectores aleatorios en dimensión n y de tamaño m , donde cada \mathbf{X}_i proviene de una función de distribución continua común $F_{\mathbf{X}}(\cdot)$; queremos probar si las coordenadas correspondientes son independientes. El procedimiento de esta prueba incluirá los siguientes pasos.

- Encontrar la distribución exacta o aproximada por simulaciones de $\delta_{X_1, X_2, \dots, X_n}^m$ en el caso de n variables aleatorias independientes uniformes $(0, 1)$ y de tamaño muestral m .
- Evaluar la estadística $\delta_{X_1, X_2, \dots, X_n}^m$ para la muestra dada.
- Rechazar independencia si la estadística es mayor que el cuantil $1 - \alpha$ de la distribución exacta o aproximada. En caso contrario no se rechaza la hipótesis de independencia.

3. Comparación entre la medida de dependencia y componentes principales

Ejemplos con datos generados.

CASO 1.- Se consideran u_1, u_2 y u_3 variables aleatorias independientes $U(0, 1)$ y $u_4 = 3u_1 - 2u_2 + u_3$ una combinación lineal perfecta de las tres primeras, y 40 observaciones de ellas.

Al ser la cuarta variable una combinación lineal de las tres primeras es claro que la varianza va a ser explicada por los tres primeros componentes. Encontramos el valor de $\delta_{U_1, U_2, U_3, U_4'}^{40} = 0,25051$ que corresponde a un cuantil $> 0,99$, indicando clara dependencia. La tabla 3 nos muestra el análisis de componentes principales con tres variables evaluando la estadística δ y sus respectivos cuantiles. La tabla 4 nos muestra el estudio realizado de dos en dos variables.

Notemos que en este caso la relación entre u_4 y u_1 es la más fuerte, seguida de u_4 y u_2 , y finalmente de u_4 y u_3 , lo cual se desprende de los pesos asignados. Lo que podemos observar del estudio de componentes principales es que por la manera en que fue generada la cuarta variable, está más relacionada con u_1 que con las otras dos. Esto se ve reflejado también en el caso de tres variables. En el caso de la medida de dependencia δ se rechaza la independencia, aún a nivel de $\alpha = 0,01$, al considerar las ternas (u_1, u_2, u_4) y (u_1, u_3, u_4) , al igual que al considerar las parejas (u_1, u_4) y (u_2, u_4) , los resultados son pues similares a los obtenidos por componentes principales, salvo en el caso de u_3 y u_4 , en este caso la medida de dependencia es baja debido a que el coeficiente de u_3 en la combinación lineal es el más bajo.

CASO 2.- Se generaron 40 observaciones de u_1 y u_2 variables aleatorias independientes $U(0, 1)$ y se definieron $u_3 = u_1/u_2$ y $u_4 = \sin(\pi u_2)$.

En este caso la varianza explicada por el último componente es prácticamente de 0 %, lo que sugiere una fuerte dependencia lineal. En este caso la estadística $\delta_{U_1, U_2, U_3, U_4}^{40} = 0,26894$ correspondiente a un cuantil de 1.0.

Al realizar los análisis de las posibles combinaciones de tres variables los resultados se presentan en la Tabla 3. Y al analizar las posibles combinaciones de dos variables se obtuvieron los datos de la Tabla 4.

Lo que es claro es que las dependencias no logran ser detectadas mediante componentes principales. De hecho en el caso de u_2 y u_4 , se podría pensar que hay independencia ver (Figura 1: Valores de u_2 y u_4), lo cual es claramente falso. Sin embargo, en las ternas y parejas en las que aparece la variable u_3 el primer componente explica toda la varianza, esto se debe a una observación aberrante (ver Figura 2: Valores de u_3 y u_4). En cambio al analizar los resultados obtenidos mediante la estadística δ se observa que todas las ternas posibles indican dependencia, lo cual se sigue de la definición de las u_i 's. Pero al analizar las parejas se observa que solamente $u = (u_1, u_2)$ y $u = (u_1, u_4)$ se pueden considerar que tienen coordenadas independientes, mientras que en las otras se rechaza la independencia de coordenadas, esto refleja claramente la forma en que fueron definidas.

La varianza de u_3 es 10506.4, mientras que para las demás variables es menor a uno. Por lo que realizamos el análisis de componentes principales con la matriz de correlación, los resultados con grupos de tres se presentan en la Tabla 5 y los de parejas se presentan en la Tabla 6, es importante notar que la influencia que tenía u_3 en el análisis usando la matriz de covarianzas se anula al considerar la matriz de correlación, pues al estandarizar se cancela el efecto de la varianza. La estadística $\delta_{\underline{u}}^{40}$ en cambio permanece inalterada.

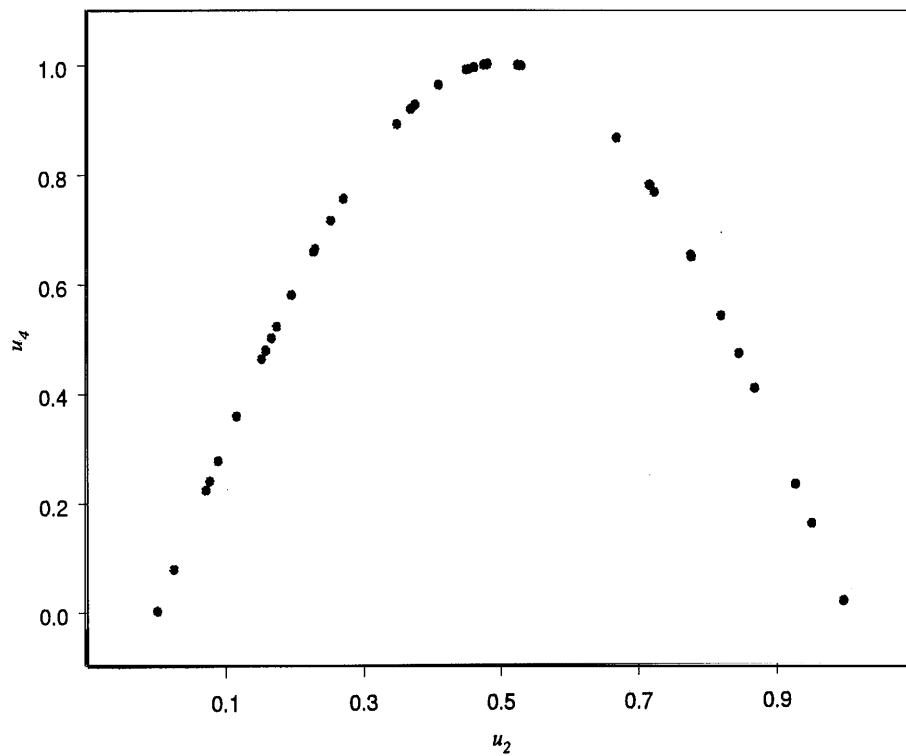


Figura 1: Valores de las Variables u_2 y u_4 en el CASO 2

Tabla 1

Variables	Porcentaje de varianza del primer componente	Valor de la Est. δ_u^{40}	Cuantil de la Distribución
$\underline{u} = (u_1, u_2, u_3)$	46.34 %	0.10494	0.14
$\underline{u} = (u_1, u_2, u_4)$	95.46 %	0.24347	1.0
$\underline{u} = (u_1, u_3, u_4)$	93.85 %	0.24347	1.0
$\underline{u} = (u_2, u_3, u_4)$	91.90 %	0.15125	0.92

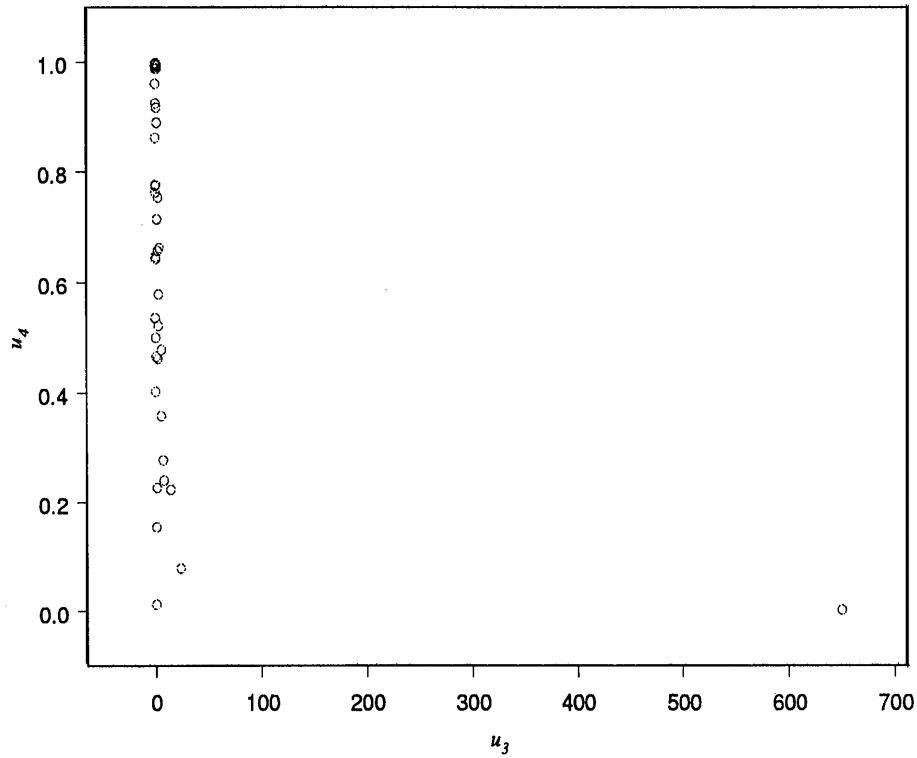


Figura 2: Valores de las variables u_3 y u_4 en el CASO 2

Tabla 2

Variables	Porcentaje de varianza del primer componente	Valor de la Est. δ_u^{40}	Cuantil de la Distribución
$\underline{u} = (u_1, u_2)$	62.60 %	0.0970	0.74
$\underline{u} = (u_1, u_3)$	55.95 %	0.0666	0.11
$\underline{u} = (u_1, u_4)$	98.66 %	0.2365	1.0
$\underline{u} = (u_2, u_3)$	62.33 %	0.1000	0.79
$\underline{u} = (u_2, u_4)$	96.64 %	0.15215	> 0.99
$\underline{u} = (u_3, u_4)$	94.82 %	0.075	0.30

Tabla 3

VARIABLES	Porcentaje de varianza del primer componente	Valor de la Est. δ_u^{40}	Cuantil de la Distribución
$\underline{u} = (u_1, u_2, u_3)$	99.99 %	0.20313	> 0.99
$\underline{u} = (u_1, u_2, u_4)$	44.10 %	0.15938	0.95
$\underline{u} = (u_1, u_3, u_4)$	99.99 %	0.21227	> 0.99
$\underline{u} = (u_2, u_3, u_4)$	99.99 %	0.26894	1.0

Tabla 4

VARIABLES	Porcentaje de varianza del primer componente	Valor de la Est. δ_u^{40}	Cuantil de la Distribución
$\underline{u} = (u_1, u_2)$	65.58 %	0.11313	0.91
$\underline{u} = (u_1, u_3)$	99.99 %	0.20313	1.0
$\underline{u} = (u_1, u_4)$	56.10 %	0.08563	0.52
$\underline{u} = (u_2, u_3)$	99.99 %	0.20313	1.0
$\underline{u} = (u_2, u_4)$	56.87 %	0.15938	> 0.99
$\underline{u} = (u_3, u_4)$	99.99 %	0.13813	0.99

Tabla 5

VARIABLES	Porcentaje de varianza del primer componente	Valor de la Est. δ_u^{40}	Cuantil de la Distribución
$\underline{u} = (u_1, u_2, u_3)$	46.97 %	0.20313	> 0.99
$\underline{u} = (u_1, u_2, u_4)$	45.53 %	0.15938	0.95
$\underline{u} = (u_1, u_3, u_4)$	45.10 %	0.21227	> 0.99
$\underline{u} = (u_2, u_3, u_4)$	49.96 %	0.26894	1.0

Tabla 6

Variables	Porcentaje de varianza del primer componente	Valor de la Est. δ_u^{40}	Cuantil de la Distribución
$\underline{u} = (u_1, u_2)$	65.58 %	0.11313	0.91
$\underline{u} = (u_1, u_3)$	50.63 %	0.20313	1.0
$\underline{u} = (u_1, u_4)$	53.90 %	0.08563	0.52
$\underline{u} = (u_2, u_3)$	63.73 %	0.20313	1.0
$\underline{u} = (u_2, u_4)$	55.41 %	0.15938	> 0.99
$\underline{u} = (u_3, u_4)$	67.34 %	0.13813	0.99

Referencias

- Aitchinson, J. (1986) *The Statistical Analysis of Compositional Data*. Ed. Chapman and Hall, London.
- Fernández-Fernández, B. and González-Barrios, J.M. (2001) Multidimensional Dependency Measures. *Preimpreso No. 101*, IIMAS-UNAM, México D.F.
- Jolliffe, T. (1986) *Principal Components Analysis*. Ed. Springer-Verlag, New York.
- Seber, G.A.F. (1984) *Multivariate Observations*. Ed. John Wiley & Sons, New York.

Modelos Marginales y Condicionales Para Mediciones Repetidas Binarias

Leticia Gracia Medrano Valdelamar

Silvia Ruiz Velasco Acosta

IIMAS, UNAM

1. Introducción

Las mediciones repetidas aparecen muy frecuentemente en estudios longitudinales en medicina, para modelar estas respuestas existen varias formas, dos de ellas son los modelos marginales y los modelos condicionales. En los primeros se especifican las distribuciones marginales y se considera la dependencia que existe entre las respuestas como parámetro de ruido, mientras que en los segundos se especifican las distribuciones condicionales. Una discusión amplia sobre la modelación marginal y condicional de mediciones repetidas aparece en Lindsey y Lambert (1998). Aquí se mencionan además las ventajas en términos de interpretación de los modelos condicionales. Este trabajo se enfocará a comparar dos modelos: los modelos GEE propuestos por Zeger y Liang (1986) y el de Azzalini (1994). Ambos modelos son marginales, según lo define Lindsey (1998), con la diferencia de que en el segundo cuenta con distribuciones condicionales sencillas que forman una cadena de Markov. Para compararlos se utilizan tanto datos simulados como datos reales, en este caso pertenecientes a un registro de ausencias de niños en edad preescolar en una escuela del sur de la ciudad de México.

2. Los Modelos GEE

En el modelo logístico de regresión, se especifican las esperanzas marginales $E(Y_{ij}) = \mu_{ij}$ y se satisface que $\text{logit}(\mu_{ij}) = \mathbf{x}'_{ij}\beta$.

Estos modelos surgen como una generalización de las conocidas “quasi-score function” de Wedderburn (1974):

$$\mathbf{S}_\beta(\beta, \alpha) = \sum_{i=1}^m \left(\frac{\partial \mu_i}{\partial \beta} \right)' Var(\mathbf{Y}_i)^{-1} (\mathbf{Y}_i - \mu_i) = \mathbf{0}$$

Aquí existe una complicación, y es que \mathbf{S}_β depende del parámetro de dependencia α y de β los parámetros del modelo; pues $Var(\mathbf{Y}_i) = Var(\mathbf{Y}_i; \beta, \alpha)$. Donde α es tal que $cor(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}; \alpha)$. Esto puede resolverse si se sustituye por un estimador consistente $\hat{\alpha}(\hat{\beta})$. Se ha mostrado que esta solución es asintóticamente tan eficiente como si α fuese conocida, Liang y Zeger (1986).

Zeger y Liang utilizan una matriz de correlación de trabajo $\mathbf{R}(\alpha)$ que depende del vector de parámetros α , de manera que $Var(Y_i) = \mathbf{A}_i^{1/2} \mathbf{R}_i(\alpha) \mathbf{A}_i^{1/2} / \phi$ Donde i es una matriz diagonal con $g(\mu_{ij})$ como elemento j -ésimo en la diagonal, la inversa de la función liga.

Si $\mathbf{R}(\alpha) = \mathbf{I}_{n_i}$ las observaciones son no correlacionadas. Si

$$\mathbf{R}_i(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^3 & \cdots & \alpha^m \\ \alpha & 1 & \alpha & \alpha^2 & \cdots & \alpha^{m-1} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \alpha^m & \alpha^{m-1} & \cdots & \alpha^2 & \alpha & 1 \end{pmatrix}$$

se le llama estructura autoregresiva. Si

$$\mathbf{R}_i(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha & \alpha & \cdots & \alpha \\ \alpha & 1 & \alpha & \alpha & \cdots & \alpha \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ \alpha & \alpha & \cdots & \alpha & \alpha & 1 \end{pmatrix}$$

se trata de una estructura intercambiable. Si

$$\mathbf{R}_i(\alpha) = \begin{pmatrix} 1 & \alpha_1 & \cdots & \alpha_m & \cdots & 0 \\ \alpha_1 & 1 & \alpha_1 & \cdots & \alpha_m & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & & \cdots & \alpha_1 & 1 & \alpha_1 \\ 0 & \cdots & \alpha_m & \cdots & \alpha_2 & \alpha_1 & 1 \end{pmatrix}$$

es una estructura m-dependiente. Los parámetros de correlación α pueden estimarse de manera simultánea, resolviendo $\mathbf{S}_\beta = \mathbf{0}$ y

$$\mathbf{S}_\alpha(\beta, \alpha) = \sum_{i=1}^m \left(\frac{\partial \eta_i}{\partial \alpha} \right)' H_i^{-1} (\mathbf{W}_i - \eta_i) = \mathbf{0}$$

donde $\mathbf{W}_i = (Y_{i1} Y_{i2}, Y_{i1} Y_{i3}, \dots, Y_{in_i-1} Y_{in_i}, \dots, Y_{i1}^2, \dots, Y_{in_i}^2)'$ y $\eta_i = E(\mathbf{W}_i; \beta, \alpha)$

La elección de la matriz de pesos H_i depende del tipo de respuestas. En el caso de ser la respuesta binaria, se pueden ignorar las últimas n_i entradas del vector \mathbf{W}_i , pues la varianza queda determinada por la media. En este caso

$$H_i(\alpha) = \begin{pmatrix} Var(Y_{i1}Y_{i2}) & 0 & \cdots & 0 \\ \vdots & Var(Y_{i1}Y_{i3}) & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & Var(Y_{in_i-1}Y_{in_i}) \end{pmatrix}$$

Los parámetros de asociación α se pueden formular y estimar de distintas maneras. Zeger y Liang parametrizan la $Var(Y_i)$ en términos de las correlaciones, y usan estimadores de momentos, otros proponen el uso de razones de momios condicionales o marginales.

3. El Modelo de Azzalini

Azzalini presenta un modelo estocástico para estudiar la influencia que tienen ciertas covariables (que dependen del tiempo), sobre la distribución marginal de variables respuestas binarias, que están correlacionadas serialmente. Esta correlación la modela con Cadenas de Markov, a través de las probabilidades de transición.

Considera un proceso estacionario (Y_1, \dots, Y_T) generado por una cadena de Markov que toma valores 0 y 1. La matriz de probabilidades de transición de este proceso está dada por

$$P = \begin{pmatrix} 1 - p_0 & p_0 \\ 1 - p_1 & p_1 \end{pmatrix}$$

donde $p_j = pr(Y_T = 1 | Y_{T-1} = j)$ para $j = 0, 1$.

Busca una parametrización tal que $\theta = E(Y_T)$ no dependa de los parámetros que regulan la dependencia serial. La dependencia entre dos observaciones sucesivas es medida por el

cociente de momios

$$\psi = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \frac{pr(Y_{T-1} = Y_T = 1)pr(Y_{T-1} = Y_T = 0)}{pr(Y_{T-1} = 0, Y_T = 1)pr(Y_{T-1} = 1, Y_T = 0)} \quad (3,1)$$

Con esta forma de medir la asociación los estimadores de los parámetros son insensibles a cambios en el parámetro de asociación.

Dados los valores de (θ, ψ) resuelve la ecuación (2.1) y la ecuación:

$$\theta = \theta p_1 + (1 - \theta)p_0 \quad (3,2)$$

con respecto a p_0 y p_1 . Para el caso en que $\theta_t = E(Y_t)$ varía con t a través de una función como

$$\text{logit}(\theta_t) = x'_t \beta, \quad (3,3)$$

usa la generalización de la ecuación (3.2),

$$\theta_t = \theta_{t-1} p_1 + (1 - \theta_{t-1}) p_0 \quad (t = 2, \dots, T) \quad (3,4)$$

donde p_0 y p_1 varían con t . Resuelve las ecuaciones anteriores y encuentra la expresión para las p_j . Entonces genera un proceso con las propiedades requeridas. Primero toma $pr(Y_1 = 1) = \theta_1$ y después genera Y_2, \dots, Y_T como una cadena de Markov no homogénea con las p_j encontradas, de manera que $E(Y_t) = \theta_t$ para $t = 1, \dots, T$ y la razón de momios para (Y_{t-1}, Y_t) es igual a ψ .

Para una sucesión de valores y_1, \dots, y_T obtiene la función de logverosimilitud para β y $\lambda = \log \psi$ esta es:

$$l(\beta, \lambda) = \sum_{t=1}^T l_t(\beta, \lambda) = \sum_{t=1}^T \{y_t \text{logit}(p_{t-1}) + \log(1 - p_{t-1})\}$$

Para poder usar el modelo con valores faltantes, supone que las observaciones son equiespaciamas y usa la identidad de Chapman-Kolmogorov recursivamente.

También extiende el modelo para mediciones repetidas, cada individuo i es observado en los tiempos $1, \dots, T_i$ y tiene su vector x_{it} de covariables al tiempo t . Lo generaliza usando la suma de las logverosimilitudes de los n individuos. Además da una aproximación para la matriz de varianzas y covarianzas de los estimadores.

4. Simulaciones

Para generar los datos binarios correlacionados con estructura intercambiable y autoregresiva de orden 1 nos basamos en el artículo de Lunn y Davies(1998), mientras que para generar las repeticiones con estructura m-dependiente de orden 1, utilizamos el algoritmo de Oman y Zucker (2001). Se generaron 100 individuos con 5 repeticiones equiespaciadas para cada estructura. Se usó el modelo $\text{logit}(\mu) = 1 + 3 * \mathbf{x}$, donde la variable explicativa \mathbf{x} está formada por uniformes(0,1). Para cada una de las tres estructuras simulamos con tres distintos valores de dependencia, para cada una de las situaciones se hicieron 500 simulaciones y se "corrieron" cuatro modelos: tres GEE cuyas matrices de trabajo de correlación tenían estructura intercambiable, autoregresiva de orden uno y m-dependiente de orden uno y además el modelo de Azzalini. Para el caso de la estructura intercambiable usamos tres valores para la correlación: 0.16,0.36 y 0.64. Encontramos que los parámetros son estimados de manera muy parecida para los cuatro modelos cofirmando la robustez de estos modelos. Las varianzas de los estimadores crecen conforme crece la correlación de los datos. Encontramos que el GEE con estructura 1-dependiente tuvo problemas de convergencia para el caso de correlación de 0.64. Y en todos los casos el parámetro de dependencia fue subestimado por los GEE. En los datos con estructura autoregresiva con valores de 0.4,0.6 y 0.8, encontramos que los parámetros de dependencia fueron estimados bien por el GEE autoregresivo. El modelo 1 dependiente de nuevo presentó problemas de convergencia. También ocurrió que las varianzas de los parámetros estimados crecían con la correlación. Con los datos de estructura 1-dependiente con correlaciones de 0.16, 0.20 y 0.25 (por el algoritmo no pudimos hacerlas mayores a ,5²), el modelo GEE 1-dependiente estimó bien la correlación. Llama la atención

que los cuatro modelos presentan varianzas muy semejantes para todas las estructuras. La estructura autoregresiva, que es la representa una mayor dependencia, es la que presentó las mayores varianzas, y la 1-dependiente la que presentó las menores.

5. Datos de Ausentismo en Preescolares

Se modelaron las ausencias de 55 niños en edad preescolar con 78 repeticiones por individuo. Se ajustó el modelo

$$\text{logit}(E(\text{ausencia})) = \beta_0 + \beta_1 \text{sexo} + \beta_2 \text{edad} + \beta_3 \text{ozeno del día anterior} + \beta_4 \text{temperatura mínima}$$

Los resultados fueron:

Coeficientes	Valor	Desv. Std.	Valor de t
constante	-2.60725	0.944	-2.76
sexo	-0.09720	0.266	-0.365
edad	-0.26927	0.128	-2.090
ozono.1	2.91787	2.159	1.351
temp.min	-0.06988	0.062	-1.117
$\log(\psi)$	1.782	0.4204	4.24

Para los modelo GEE se encontraron prácticamente los mismos parámetros estimados, siendo el estimador de la dependencia intercambiable $\alpha = 0,0073$, en el caso de la estructura autoregresiva de $\alpha = 0,0584$ y en la 1-dependiente $\alpha = 0,0585$

En todos los modelos sólo resultó significativo el parámetro que multiplica a la edad y el término constante. Como se esperaba estos parámetros estimados fueron positivos, es decir que los niños tienden a no faltar y que entre más grandes menos faltan. Lo que llama la atención es que la dependencia estimada con los GEE es muy pequeña, mientras que en el

modelo de Azzalini la λ toma un valor de 1.78, que nos habla de una dependencia positiva alta. Las varianzas estimadas de los estimadores en los cuatro modelos son muy parecidas.

6. Conclusiones

Tanto los GEE como el modelo de Azzalini son robustos en los parámetros del modelo. Cuando los datos presentan estructura intercambiable el parámetro de dependencia no es estimado adecuadamente por los GEE. Las varianzas de los estimadores en general aumentan cuando la correlación crece.

En todos los casos de datos generados el sesgo fue positivo, y este aunque es pequeño también crece cuando la dependencia crece. El valor más alto de λ fue de 4.9 para la estructura autoregresiva con $\rho = 0,8$ y el más bajo de 0.60 para la estructura intercambiable con $\alpha = 0,16$. Los GEE convergen más rápidamente, entonces conviene utilizar los parámetros estimados por los GEE como parámetro inicial en el modelo de Azzalini. El modelo GEE con matriz de correlación de trabajo intercambiable fue el que peor estimó la dependencia. Los GEE en el caso de los datos reales dieron una mala estimación de la dependencia entre las observaciones, tal vez esto debido a la estructura de correlación no se parece a ninguna de las tres con las que se trabajó, mientras que el modelo de Azzalini arrojó una alta dependencia.

Referencias

Azzalini, A. (1994). Logistic Regression for autocorrelated data with application to repeated measures, *Biometrika*, **81**, 767-75.

Lindsey, J.K. y Lambert, P. (1998). On the appropriateness of marginal models for repeated measurements in clinical trials, *Statistics in medicine*, **17**, 447-469.

Lunn, A.D. y Davies, S.J. (1998). A note on generating correlated binary variables, *Biometrika*, **85**, 487-490.

Oman, S.d. y Zucker, D.M. (2001). Modelling and generating correlated binary variables, *Biometrika*, **88**, 287-290.

Wedderburn, R.W.M. (1974) Quasi-likelihoodd functions, generalized linear models and the Gaussian method. *Biometrika*, **61**, 439-447.

Zeger, S.L. y Liang K.Y. (1986). Longitudinal data analysis for discrete and continuos outcomes, *Biometrika*, **42**, 121-130.

**Encuesta Aplicada a Empresas
Hoteleras con Categorías de 3, 4 y 5
Estrellas en los Municipios de Bahía
de Banderas, Compostela, San Blas y
Tepic del Estado de Nayarit.
(Oct.1998 a Feb.1999)**

**Ma. Antonia Iñiguez Valadez
María Margarita Gómez Ramírez**

Universidad Autónoma de Nayarit

1. Introducción

El turismo como fenómeno social, evoluciona paralelamente a la industrialización y en razón directa del concepto empleo del tiempo libre, aunado a las facilidades que el mundo moderno presenta en lo referente al transporte y alojamiento. Los factores técnicos y fenómenos naturales conjugados proporcionan el perfil de la infraestructura turística. El turismo abarca una serie diversa de áreas de operación, cada una de las cuales genera actividades de índole social, cultural, económica y recreativa.

2. Objetivo de la investigación

Debido a la importancia económica que reviste la actividad turística en Nayarit y considerando a la hotelería como parte sustantiva de esta industria, se hace necesario realizar un estudio exploratorio de estas empresas con el objeto de elaborar un diagnóstico.

3. Objetivo de la encuesta

Obtener información precisa que nos permita elaborar un diagnóstico de las empresas hoteleras que operan en el estado de Nayarit. El cuestionario consta de cinco apartados. En el primero se solicitan los datos generales de la empresa, localización, tamaño de empresa, categoría y número de habitaciones. El segundo y tercero tienen por objeto definir el perfil tanto de la empresa como del empresario. El cuarto trata acerca de la operación de la empresa. Y en el quinto trata acerca del uso de la tecnología y la capacitación. Al inicio de esta investigación se exploró la posibilidad de que la selección de la muestra fuera aleatoria y representativa, considerando los siguientes criterios:

- a) Categoría del hotel.- Se descartaron los hoteles con menos de tres estrellas debido a que su promedio de cuartos es muy bajo en comparación al resto de la clasificación;
- b) Localización geográfica.- Dado que los municipios de Bahía de Banderas, Compostela, San Blas y Tepic, concentran la mayor oferta hotelera del estado, estos se seleccionaron para aplicar la encuesta. Sin embargo, para asegurar la representatividad de cada uno de los criterios señalados el tamaño de la muestra requerido era demasiado grande. En consecuencia se descartó la posibilidad de trabajar con una muestra probabilística y se seleccionaron los hoteles para determinar el número de empresas que se entrevistarían respetando en lo posible los criterios anteriores. En principio se planeó aplicar la encuesta a 30 establecimientos de un total de 36 para cubrir el 83 por ciento de la población. El diseñar una muestra y aplicarla para llevar a cabo la investigación son dos cosas distintas, esto es necesario decirlo

porque al levantar los cuestionarios a los hoteles previamente elegidos, se encontró que en la mayoría de los negocios los responsables de dar la información mostraron gran resistencia y en algunos casos ninguna cooperación. Es decir, que en varios de los establecimientos no se permitió el acceso para llevar a cabo la entrevista a los administradores y así informales de la investigación que se estaba llevando a cabo. Debido a estas barreras encontradas se decidió pedir apoyo a la Secretaría de Turismo del Estado de Nayarit para la aplicación de los cuestionarios en los municipios de Bahía de Banderas y Compostela, lográndose levantar 7 cuestionarios con la ayuda de los delegados de esta Secretaría en esos municipios, además de los 10 que se habían levantado antes. Así el número de cuestionarios aplicados sumó 17, por tanto la muestra representa un poco más del 47 por ciento de la población. Para aplicar el cuestionario a estas empresas fue necesario determinar el número total de estos establecimientos. Para ello se consideró el Directorio Hotelero 1998, publicado por la Secretaría de Turismo del Estado de Nayarit. Así, analizando dicho directorio podemos definir que la población en estudio estará compuesta por los hoteles ubicados en los municipios de Bahía de Banderas, Compostela, San Blas y Tepic con categorías de 3, 4 y 5 Estrellas ya que ellos concentran la mayor oferta hotelera como se menciona en el apartado anterior. La encuesta por su distribución geográfica se aplicó en un 11.8 por ciento en Bahía de Banderas, 29.4 por ciento en Compostela, 5.9 por ciento en San Blas y 52.9 por ciento en Tepic. Otro factor que se consideró para aplicar la encuesta en esos cuatro municipios es la limitación de tiempo y recursos. Con esta muestra se obtuvieron algunos indicadores que permiten puntualizar la participación de la hotelería en el desarrollo económico del estado. Pero también, hacer un análisis que permita llegar a construir una caracterización de las empresas hoteleras para la región delimitada anteriormente, a través de medir las siguientes variables: perfil de las empresas encuestadas, perfil de los empresarios y sus expectativas, perfil laboral y operación de la empresa. Dentro de los objetivos de la investigación están los de precisar el perfil de los hoteles que actualmente operan en el estado de Nayarit. Para ello se investigó si dichos hoteles están registrados como: único Propietario, Cooperativa o bien Sociedad Anónima. Asimismo se investigó el papel que tienen en las empresas los socios mayoritarios sobre todo

en la operación de la empresa. Que niveles de dirección ostentan los socios mayoritarios y de dichos socios mayoritarios se precisaron algunos aspectos como su edad para ver la permeabilidad o no de las innovaciones, correlacionándolo esto con el grado de estudios de los empresarios y sobre todo la ocupación anterior de dicho empresario. Además de que la información de si en el momento actual tiene otros ingresos nos permitió detectar hasta donde en forma exclusiva estos empresarios están dedicados a la hotelería. Asimismo, se obtuvo información sobre la visión que ellos tienen con respecto del desempeño de su empresa y de sus motivaciones para dicho negocio y su visión respecto a las perspectivas de la hotelería.

4. Perfil de la empresa hotelera

La muestra se aplicó en 1 hotel de 5 estrellas, 10 de 4 estrellas y 6 de 3 estrellas representando poco más del 22 por ciento del total de 75 establecimientos registrados en el estado. Cabe destacar que dentro de estas 3 categorías el total de hoteles en la región es de 36, por lo que la encuesta representa aproximadamente el 47 por ciento del total de la región. Las empresas encuestadas por su tamaño se clasifican en micro (23.5 %), pequeña (58.8 %) y mediana (17.60 %). En cuanto a la oferta de habitaciones se obtuvo una mediana de 56 cuartos para los establecimientos. Se encontró que del 100 % de nuestras encuestas un 41 por ciento de los hoteles entrevistados tienen un único Propietario y el 59 por ciento restante es Sociedad Anónima.

5. Perfil del empresario y sus expectativas

En cuanto al puesto que ocupa el socio mayoritario o bien el propietario único los resultados indican que más del 90 por ciento es el líder de la empresa, de tal suerte que la conducción de la hotelería de la muestra levantada se lleva a cabo por los propietarios mismos o por

el socio mayoritario de la sociedad. Sin embargo, se puede decir que la participación de estos actuales empresarios hoteleros fue motivada fundamentalmente por consejo de amigos (20 %), y sólo el 33 por ciento tenían antecedentes y conocimientos del negocio, de tal suerte que un poco más del 60 por ciento participan o iniciaron su participación en el negocio sin tener conocimientos del mismo, toda vez que prácticamente ninguno de los actuales empresarios conocía el negocio de la hotelería cuando se inicio empresarialmente en él. En cuanto al origen de los empresarios, el 36 por ciento se dedicaban anteriormente al comercio, más del 28 por ciento eran empresarios en otro giro y el 28.6 por ciento decidieron dejar el trabajo de empleados para dedicarse a las funciones empresariales. Del 100 por ciento de los empresarios solo el 15 por ciento vive exclusivamente de lo que le deja la hotelería. De tal suerte que el 85 por ciento de los empresarios tienen ingresos complementarios a la actividad hotelera en diversas actividades. Asimismo los empresarios consideran que su desempeño es satisfactorio, ya que poco más del 70 por ciento considera buena o excelente la conducción de sus empresas, el 11 por ciento la considera no satisfactoria. En cuanto a las perspectivas de ellos en el negocio hotelero consignan que el 65 por ciento ve buenas perspectivas o excelentes para su negocio. Estos hoteles, independientemente de la opinión en cuanto a las perspectivas de la actividad hotelera en el estado expresadas por los informantes o participantes en la muestra, presentaron un nivel promedio de ocupación anual de 50.75 por ciento para el año de 1997. Cabe mencionar que el promedio de ocupación hotelera a nivel estatal para el mismo año sumó 48.42 por ciento.

6. Perfil laboral

De acuerdo con Kotler et. al. (1997); en la industria de la hospitalidad (está integrada por aquellas empresas que realizan una o más de las siguientes acciones: proporcionan alojamiento, preparan servicio de alimentos y bebidas y ofrecen entretenimiento al viajero.), los empleados forman parte crucial del producto turístico. Es por ello que se hace preciso

analizar la encuesta en lo relativo a los trabajadores, en esto se encontró que la mediana del número de trabajadores de planta es de 30 personas y la de los trabajadores eventuales resultó igual a 7. En cuanto al total de trabajadores (de planta y eventuales) la mediana es 31. De acuerdo a las razones de los empleados para dejar la empresa el 35.3 por ciento reportó que por asuntos personales, el 17.6 por ciento por cambio de empleo y el 29.3 por ciento informó que no hubo bajas. La mediana de los trabajadores retirados resultó igual a 4 personas. El criterio que se utiliza para la contratación es la experiencia, la edad y el sexo. Las razones fundamentales para el reclutamiento de personal se encontró que básicamente es el reemplazo de retirados que representa más del 86 por ciento y sólo un 14 por ciento indica que es por expansión o crecimiento de las actividades hoteleras. La mediana de los trabajadores reclutados en 1997 fue de 6.5 personas.

7. Operación de la empresa hotelera

Las empresas encuestadas indicaron que para 1997, el promedio de ocupación anual fue de 50.75 por ciento. En cuanto al origen de los huéspedes el porcentaje promedio del 81 por ciento correspondió a los nacionales y el resto a los extranjeros. De acuerdo con Hernández Díaz (1990), una de las características del perfil del consumidor es la motivación, y en base a ello puede derivarse una segmentación de mercado la cual se considera como uno de los indicadores importantes para elaborar un producto turístico de acuerdo a los requerimientos de la demanda por captar. Los segmentos motivacionales que maneja este autor son: Negocios o actividades profesionales Vacaciones o uso de tiempo libre Congresos, convenciones y/o reuniones Así, que de acuerdo a lo anterior, la encuesta contempló un apartado con el objeto de conocer de que manera se comporta el mercado para las empresas de la muestra. De los resultados de la encuesta en cuanto la segmentación motivacional de la demanda turística, tenemos entonces, que la visita por vacaciones o uso del tiempo libre tuvo un porcentaje promedio de 64, y el resto corresponde a los segmentos de: negocios y/o actividades

profesionales y congresos, convenciones y reuniones, Con estas cifras se puede decir que la demanda que atienden estas organizaciones son básicamente vacacionistas y que su motivo es la diversión y el descanso. Considerando que la comercialización del producto depende del tipo de mercado donde se compite, la participación de las empresas en el mercado, a juicio de los directivos por factor geográfico puede ser: regional, nacional e internacional, que suma un poco más del 64 por ciento. En otro sentido, el financiamiento o el manejo financiero está dado por aportaciones directas de los socios y se evidencia que hay muy poca vinculación con el sistema financiero. Esto mismo concuerda con lo escrito por Ramírez Blanco (1992), los empresarios de la industria hotelera por lo general, han invertido sus propios recursos y sólo han recurrido a las instituciones financieras para cubrir necesidades en el campo del capital de trabajo.

Referencias

Kotler, P., Boiven,J. y Makens, J.(1997). *Mercadotecnia para la Hotelería y Turismo* 1a. Ed. México: Prentice Hall. p. 319

Ramírez, B.M. (1992). *Teoría General de Turismo*. 2a. Ed. México: Edit. Diana. p. 81.

Análisis de modelos lineales en encuestas complejas

Dr. Ignacio Méndez Ramírez

M. en E. Patricia Romero Mares

Dra. Guillermina Eslava Gómez

IIMAS, UNAM

1. Introducción

En encuestas complejas (diseños polietápicos y estratificados), la muestra no es “representativa”, en el sentido que los momentos de la muestra no son estimadores consistentes de los correspondientes momentos poblacionales. Además, las observaciones no son independientes, es decir, hay correlación entre las observaciones dentro de una misma Unidad Primaria de Muestreo, UPM, entonces los estimadores de varianza que no toman en cuenta el diseño, son sesgados y en ocasiones con subestimaciones muy fuertes. Para valorar esta última característica se usa el “Efecto de diseño”:

$$DEF = \frac{V_{verdadera}(\hat{\Theta}_{diseño})}{V_{iid}(\hat{\Theta}_{iid})}$$

Por lo anterior, aplicar los modelos de regresión por mínimos cuadrados usuales, no es adecuado.

2. Modelos lineales en poblaciones finitas

La regresión entre una variable Y y variables X_1, X_2, \dots, X_p , se puede considerar, a nivel poblacional, como la solución a las ecuaciones normales poblacionales:

$$B = (X'X)^{-1}(X'Y), \quad Y_{N \times 1}, \quad X_{N \times (p+1)}$$

Donde la matriz $(X'X)$ contiene las sumas poblacionales de productos entre las X_j , con $X_{.1} = 1$.

Para estimar B , se deben estimar cada uno de los totales involucrados en las expresiones para B . Así, por Horvitz-Thompson, donde π_i es la probabilidad de inclusión de la unidad U_i :

$$\hat{N} = \sum_{i=1}^n \frac{1}{\pi_i}, \quad \widehat{\sum_{i=1}^N Y_i} = \sum^n \frac{Y_i}{\pi_i} \quad \text{y, en general} \quad \widehat{\sum_{i=1}^N Y_i X_{ji}} = \sum_{i=1}^n \frac{Y_i X_{ji}}{\pi_i}$$

Estas expresiones se sustituyen en las ecuaciones normales y se encuentra la solución para estimar B . Este procedimiento es equivalente a una regresión ponderada, o mínimos cuadrados ponderados, donde los ponderadores son los factores de expansión. Esto corrige por la falta de representatividad y los efectos de conglomeración tienen un impacto muy leve sobre este estimador, por lo que se pueden ignorar en la estimación puntual, pero no es el caso para estimar varianzas. Para estimar varianzas, y corregir las pruebas de hipótesis sobre los coeficientes de regresión, se pueden usar tres procedimientos básicos: 1. Técnicas de remuestreo (no se consideran en este trabajo) 2. Estimadores de varianza vía series de Taylor y 3. Ecuaciones de Estimación Generalizadas (GEE).

3. Linealización

Este procedimiento, descrito entre otros, en Raj (1968), Sarndal et al.(1990), está implementado en los programas PC-CARP y STATA.

4. Ecuaciones de Estimación Generalizadas en Modelos lineales

Método descrito en Zeger y Liang(1986) y es el que usa el programa SUDAAN. Se considera una estructura de covarianzas entre observaciones dentro de los conglomerados, éstas se pueden suponer: a) Independientes, es decir, covarianzas cero, que equivale al método de linealización ó b) Intercambiables, covarianza ρ , igual para cualquier pareja dentro de una UPM.

5. Ajustes a los errores estándar

Skinner et al.(1989) sugieren otro enfoque que consiste en multiplicar los estimadores de los errores estándar de los coeficientes de regresión (al ajustar el modelo con mínimos cuadrados ponderados por los factores de expansión) por la raíz cuadrada del DEFF de la media de la variable dependiente. También se puede usar para cada coeficiente el DEFF de la variable independiente correspondiente.

6. Ajustes por tamaño de muestra efectivo (n_e)

En un diseño con probabilidades de inclusión arbitrarias, se tiene que el tamaño de muestra efectivo es:

$$n_e = \frac{n}{DEFF(\hat{Y})}$$

Existen dos métodos para el ajuste: 1. Ajustar el modelo con mínimos cuadrados ponderados por los factores de expansión, pero modificando los grados de libertad del error en las pruebas de hipótesis sobre cada coeficiente de regresión b_i a $n_{ei} - p$, donde n_{ei} es el tamaño de muestra efectivo para la variable X_i y p el número de parámetros por estimar en el modelo. 2. Ajustar los grados de libertad del error con el tamaño de muestra efectivo que usa el DEFF máximo de las variables independientes.

7. Modelos Anidados Aleatorios

Se considera un modelo de regresión con efectos fijos de las variables independientes agregando dos términos, un efecto fijo de estrato y uno aleatorio de UPM dentro o anidado en el estrato. Esto modela la correlación entre mediciones dentro de las UPM.

8. Encuesta Nacional de Alimentación en el Medio Rural 1996

Se estudiaron todas las localidades del país que tenían entre 500 y 2499 habitantes. En cada estado se formaron estratos homogéneos en aspectos fisiográficos y socioeconómicos

de localidades; en cada estrato se tomaron por m.a.s. entre dos y tres localidades o UPM. Dentro de cada localidad muestreada se tomaron, por m.a.s., 50 familias y en cada familia un máximo de 3 niños menores de 6 años de edad. Se midieron índices de crecimiento de los niños, entre otros aspectos, además de consumo de alimentos e indicadores socioeconómicos. El estado que se estudió en el presente trabajo fue Chiapas, donde se tuvieron 25 estratos, 55 localidades y 1473 niños en muestra con datos completos. Con la variable peso para la edad, PEDZ, como dependiente, se exploraron varios modelos de regresión, sin hacer ajustes por el diseño de muestra y se tomó uno que resultó con 6 variables independientes significativas (con mínimos cuadrados ordinarios), las cuales fueron: **Gasto semanal en alimentos (Gasto)**, **Idioma** (0 si habla lengua indígena, 1 si habla español), **meses de lactancia del niño (Lact)**, **número de miembros en la familia (Miembros)**, **lugar donde calienta alimentos (Calienta)** (0 fogón, 1 estufa), y **lugar donde cocina alimentos (Cocina)** (1 cocina separada, 0 no). Con estas variables más la identificación de estratos, localidades dentro de estrato y los factores de expansión de cada niño-familia, se corrieron regresiones bajo los diferentes métodos.

9. Resultados

Valores de DEFF para cada variable: **PEDZ** 2.93, **Idioma** 25.13, **Gasto** 23.18, **Lact** 4.30, **Miembros** 5.54, **Calienta** 7.55, **Cocina** 3.95.

Las siglas para los diferentes métodos usados son: **MCO**: Mínimos cuadrados ordinarios, i.e. se ignora el diseño. **MCP**: Mínimos cuadrados ponderados por factores de expansión. **DEFF-Y**: Se ajustan los errores estándar multiplicando por la raíz cuadrada del DEFF de PEDZ. **DEFF- X_i** : Se ajusta el error estándar de cada coeficiente dividiendo entre la raíz del DEFF de la variable independiente correspondiente. **S Indep**: Sudaan Independiente. Se corre el SUDAAN con el estimador robusto, suponiendo que las correlaciones son nulas.

S Inter: Sudaan Intercambiable. Se corre el SUDAAN con el estimador robusto, suponiendo que la correlación es la misma y diferente de cero para cualquier pareja de observaciones dentro de la UPM. **PC-CARP:** Se corre el PC-CARP con su método de linealización y ajuste a los grados de libertad. **JMP Anidado:** Se corre el modelo de efectos mixtos, con estratos y UPM anidadas dentro de estratos. **Ajuste n_e min:** Con los resultados de MCP se ajustan los grados de libertad del denominador en las pruebas de t , por el DEFF máximo, que equivale a la n_e mínima. **Ajuste n_e var:** Con los resultados de MCP se ajustan los grados de libertad del denominador en las pruebas de t , por el DEFF, que equivale a la n_{ei} de la variable X_i para su correspondiente coeficiente.

Tabla 1. Valores de los Errores Estándar (EE) de los Coeficientes de Regresión

Término	MCO	MCP	DEFF-Y	DEFF- X_i	S.Independiente	S.Inter	PC-CARP	JMP Anidado
Cte.	0.1410	0.1370	0.2345		0.1391	0.1202	0.1390	0.1610
Idioma	0.0620	0.0620	0.1061	0.3114	0.0903	0.0944	0.0945	0.1160
Gasto	0.0020	0.0020	0.0034	0.0096	0.0020	0.0021	0.0019	0.0021
Lact.	0.0040	0.0040	0.0068	0.0083	0.0049	0.0051	0.0049	0.0039
Miembros	0.0140	0.0140	0.0240	0.0330	0.0167	0.0157	0.0167	0.0140
Calienta	0.0900	0.0860	0.1472	0.2363	0.1518	0.1589	0.1520	0.0930
Cocina	0.0810	0.0750	0.1284	0.1491	0.0732	0.0759	0.0732	0.0810

Tabla 2. Valores de p en las pruebas de nulidad de coeficientes de regresión

Término	MCO	MCP	DEFF-Y	DEFF- X_i	S. Independ.	S. Inter	PC-CARP	JMP Anidado	Ajuste n_e min	Ajuste n_e var
Cte.	0.000	0.000	0.000		0.000	0.000	0.000	0.004		
Idioma	0.000	0.000	0.017	0.418	0.009	0.012	0.009	0.0003	0.006	0.000
Gasto	0.015	0.096	0.365	0.748	0.126	0.376	0.127	0.810	0.127	0.126
Lact.	0.040	0.014	0.180	0.272	0.070	0.076	0.071	0.010	0.027	0.024
Miembros	0.029	0.016	0.167	0.315	0.056	0.036	0.056	0.007	0.021	0.018
Calienta	0.054	0.039	0.226	0.451	0.251	0.236	0.250	0.004	0.044	0.040
Cocina	0.044	0.112	0.350	0.421	0.112	0.079	0.112	0.019	0.116	0.119

Observaciones a las tablas de resultados.

- Los errores estándar (EE) casi no cambian al pasar de MCO a MCP.
- Los ajustes por DEFF inflan demasiado los EE. Salvo para Idioma el Anidado produce EE semejantes a MCP.
- PC-CARP y Sudaan Independiente son prácticamente idénticos. Los valores de los coeficientes de regresión cambian poco de un método a otro, salvo el anidado que aumenta para Idioma y Calienta.
- Lo verdaderamente preocupante es que el MCO reporta todas las variables independientes como significativas, mientras que todos los otros métodos no detectan significativo el Gasto, y tampoco detectan para Calienta y Cocina PC-CARP y SUDAAN.
- Los ajustes por DEFF son demasiado exigentes y no señalan como significativo ningún término, salvo Idioma con DEFF por PEDZ.
- El Idioma es significativo en todos los métodos salvo con DEFF de cada X_i .
- En general el modelo anidado mixto no corrige suficiente los EE y salvo Gasto, reporta menores valores de p .
- Los ajustes por tamaño de muestra efectivo son adecuados para Gasto y Cocina, pero con valores un poco más bajos de p para Lact., Idioma y Miembros, pero para Calienta son mucho más bajos que PC-CARP y SUDAAN.

10. Conclusiones

- Es importante tomar en cuenta el diseño en el análisis de los modelos lineales.
- La falta de ponderación no implica cambios importantes en los estimadores de coeficientes de regresión, pero sí afecta los errores estándar.
- Se requieren los métodos que consideran el impacto del diseño sobre el supuesto de independencia de las observaciones, con estimación adecuada de EE y/o ajuste a los grados de libertad.
- Resultan adecuados los de PC-CARP y SUDAAN, pero se requieren estos paquetes.
- El de tamaño de muestra mínimo resulta fácil de aplicar y es más o menos adecuado para casi todas las variables.
- Los métodos de mínimos cuadrados ponderados por factores de expansión y estimadores de errores estándar por linealización de series de Taylor o las ecuaciones generales de estimación (GEE) son los mejores.
- La corrección de grados de libertad por DEFF fue buena para algunos coeficientes de regresión.

Referencias

Kish, L. (1984). *Survey Sampling*. J. Wiley

Overton, W., Stehman, S. (1995). The Horvitz-Thompson Theorem As a Unifying Perspective for Probability Sampling: with Examples from Natural Resource Sampling, *The American*

Statistician, Vol. 49, No. 3, 261-268.

Raj, Des (1968). *Sampling Theory*. Mc. Graw Hill Co.

Sarndal, C.E., Swensson, B., Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

Skinner,C.J., Holt,D. and Smith, T.M. (1989). *Analysis of Complex Surveys*. J. Wiley.

St-Martin P.(1993). Statistical Analysis of Complex Survey Data. *Curso impartido durante el Foro Nacional de Estadística, Aguascalientes, México*.

Zeger, C.J. and Liang, D.K. (1986). Longitudinal data analysis for discrete and continuous outcomes, *Biometrics*, 42, 121-13.

Introducción al Análisis Bayesiano de Datos Direccionales

Gabriel Nuñez Antonio

Instituto Tecnológico Autónomo de México

Eduardo Gutiérrez-Peña

IIMAS, UNAM

1. Introducción

En muchas aplicaciones se pueden encontrar datos direccionales que no pueden ser tratados adecuadamente con una teoría lineal. Por ejemplo, datos de orientación en biología, datos de pendientes y declinaciones en geología, datos de fluctuaciones en medicina y datos de direcciones de viento en meteorología. La aplicación de técnicas lineales convencionales puede producir paradojas; por ejemplo, la media aritmética de los ángulos 1 y 359 es 180 mientras que por intuición geométrica debería ser 0.

En este trabajo se presenta una introducción al análisis de datos direccionales y se muestran los desarrollos asociados al análisis Bayesiano de un modelo particular para este tipo de datos, basado en la distribución de von Mises.

2. Naturaleza de los datos direccionales

Los datos circulares aparecen en varias formas. Tal vez las dos más relevantes surgen de los más importantes instrumentos de medición circular: *la brújula* y *el reloj*. Observaciones

típicas medidas a través de una brújula incluyen direcciones de viento y direcciones de migración de aves. Como ejemplo de observaciones típicas medidas con el reloj se pueden mencionar datos asociados a tiempos de llegada (sobre un reloj de 24 hrs.). Datos similares aparecen como tiempos a lo largo de un año (o tiempos en meses) de la ocurrencia de algún evento.

El interés en desarrollar técnicas para analizar datos direccionales se remonta a la época en la que Gauss desarrolló la teoría de errores para analizar ciertas medidas direccionales en astronomía. Es un accidente histórico que los errores involucrados fueran suficientemente pequeños para que Gauss usara una aproximación lineal y, como una consecuencia, que desarrollara una teoría lineal en lugar de una teoría direccional de los errores. En muchas aplicaciones, sin embargo, se pueden encontrar datos direccionales que no pueden ser tratados adecuadamente con una teoría lineal.

Como un punto final para entender la naturaleza diferente de los datos circulares con respecto a los datos sobre la línea, se puede ver que el círculo es una curva cerrada pero la línea no, por lo que se pueden anticipar diferencias entre la teoría estadística sobre la línea y sobre el círculo. Por ejemplo, es necesario definir funciones de distribución, funciones características y momentos de tal manera que tomen en cuenta la periodicidad natural del círculo.

3. Medidas Descriptivas

Una forma útil de presentar un conjunto de datos, en particular datos circulares, es a través de medidas descriptivas apropiadas. Una manera adecuada de construir estas medidas para datos circulares es considerar los puntos sobre el círculo como vectores unitarios en el plano y trabajar en términos de las coordenadas polares correspondientes.

Una vez seleccionada una dirección y una orientación inicial, cada punto x sobre el círculo

se puede representar por un ángulo θ o, en forma equivalente, por un número complejo, es decir,

$$\mathbf{x} = (\cos \theta, \sin \theta)^t \quad y \quad z = e^{i\theta} = \cos \theta + i \sin \theta.$$

A continuación se describen brevemente algunas medidas descriptivas para datos circulares. El lector interesado puede consultar Mardia (1972), y Mardia y Jupp (2000).

3.1. Dirección Media

Sean $\mathbf{x}_1, \dots, \mathbf{x}_n$ vectores unitarios correspondientes a los ángulos $\theta_i, i = 1, \dots, n$. La *dirección media*, $\bar{\theta}$, de $\theta_1, \dots, \theta_n$ se define como la dirección de la resultante $\mathbf{x}_1 + \dots + \mathbf{x}_n$ de los vectores $\mathbf{x}_1, \dots, \mathbf{x}_n$. También se define como la dirección del centro de masa, $\bar{\mathbf{x}}$, de $\mathbf{x}_1, \dots, \mathbf{x}_n$,

Como las coordenadas cartesianas de \mathbf{x}_i son $(\cos \theta_i, \sin \theta_i)$ $i = 1, \dots, n$, entonces las coordenadas cartesianas del centro de masa son (\bar{c}, \bar{s}) , donde

$$\bar{c} = \frac{1}{n} \sum_{i=1}^n \cos \theta_i \quad \bar{s} = \frac{1}{n} \sum_{i=1}^n \sin \theta_i.$$

Si se define

$$\bar{R} = (\bar{c}^2 + \bar{s}^2)^{1/2},$$

entonces $\bar{\theta}$ es la solución de las ecuaciones

$$\bar{c} = \bar{R} \cos \bar{\theta} \quad y \quad \bar{s} = \bar{R} \sin \bar{\theta}.$$

Cuando $\bar{R} > 0$, $\bar{\theta}$ se obtiene como

$$\bar{\theta} = \begin{cases} \tan^{-1}(\bar{s} / \bar{c}) & \text{si } \bar{c} \geq 0 \\ \tan^{-1}(\bar{s} / \bar{c}) + \pi & \text{si } \bar{c} < 0, \end{cases}$$

donde $\tan^{-1}(\cdot)$ toma valores en $(-\pi/2, \pi/2)$.

3.2. Medidas de concentración y dispersión

Si $\mathbf{x}_1, \dots, \mathbf{x}_n$ son vectores unitarios, la longitud de la resultante promedio, \bar{R} , se define como la longitud del centro de masa $\bar{\mathbf{x}}$. Como $\mathbf{x}_1, \dots, \mathbf{x}_n$ son vectores unitarios, entonces

$$0 \leq \bar{R} \leq 1.$$

Si las direcciones $\theta_1, \dots, \theta_n$ están muy agrupadas, \bar{R} será cercana a 1 y si están muy dispersas entonces será cercana a 0. Así, \bar{R} es una medida de la *concentración* de un conjunto de ángulos. Hay que notar que cualquier conjunto de datos de la forma $\theta_1, \dots, \theta_n, \theta_1 + \pi, \dots, \theta_n + \pi$ tiene un valor de $\bar{R} = 0$, por lo que si $\bar{R} \approx 0$ no implica que las direcciones estén dispersas alrededor de todo el círculo.

Para propósitos descriptivos e inferenciales, \bar{R} juega un papel más importante que cualquier medida de dispersión. Sin embargo, para tener una comparación con los datos sobre la recta, es útil considerar medidas de dispersión de datos circulares, algunas de las cuales se presentan a continuación.

La más simple de estas medidas es la *varianza circular*, que se define como $V = 1 - \bar{R}$, donde

$$\bar{R} = \frac{1}{n} \sum_{i=1}^n \cos(\theta_i - \bar{\theta}) = (\bar{c} + \bar{s})^{1/2}.$$

En Mardia y Jupp (2000) se muestra que V es invariante bajo rotaciones. La varianza circular V toma valores en $[0, 1]$, a diferencia de la varianza muestral lineal, cuyo rango es $[0, \infty)$. Una transformación apropiada de V al rango $[0, \infty)$ está dada por

$$v = \{-2 \log(1 - V)\}^{1/2}$$

Así definida, v es en cierta manera una medida análoga a la desviación estándar muestral ordinaria. Sin embargo, V y \bar{R} son más útiles que v en investigaciones teóricas.

4. Modelos para datos circulares

Los modelos para datos circulares se pueden clasificar en tres grandes categorías: modelos generados por proyecciones, modelos “envueltos” (wrapped) y modelos tipo von Mises. Aquí sólo se trabaja con el modelo de von Mises, el cuál se define a continuación.

Se dice que un ángulo aleatorio θ tiene una distribución de von Mises, con dirección media μ y parámetro de concentración k , si su función de densidad está dada por

$$M(\theta | \mu, k) = \frac{1}{2\pi I_0(k)} e^{k \cos(\theta - \mu)} I_{[0, 2\pi)}(\theta) I_{[0, 2\pi)}(\mu) I_{[0, \infty)}(k),$$

donde I_0 es la función de Bessel modificada de primer tipo y orden cero.

Esta distribución fue introducida por von Mises en 1918 en el estudio de desviaciones de pesos atómicos medidos a partir de valores enteros.

5. Análisis Bayesiano del modelo de von Mises

El análisis clásico del modelo de von Mises se puede revisar, por ejemplo, en Fisher et al. (1987) y Mardia (1972). La literatura sobre inferencia Bayesiana es menos extensa debido al problema que representa trabajar con este tipo de distribuciones. A continuación se presenta un análisis Bayesiano completo, basado en el trabajo de Damien y Walker (1999).

5.1. Especificación del modelo

Función de densidad

$$M(\theta \mid \mu, k) = \frac{1}{2\pi I_0(k)} e^{k \cos(\theta - \mu)} I_{[0, 2\pi)}(\theta).$$

Distribución Inicial Conjugada

Propuesta por Guttorp y Lockhart (1988).

$$f(\mu, k) = \frac{1}{\{I_0(k)\}^c} \exp \{kR_0 \cos(\mu - \mu_0)\}.$$

Distribución Final

Sea $\theta = (\theta_1, \dots, \theta_n)$ una muestra aleatoria de tamaño n ,

$$f(\mu, k \mid \theta) = \frac{1}{\{I_0(k)\}^{c+n}} \exp \{kR_n \cos(\mu - \mu_n)\},$$

con

$$\begin{aligned} R_n \cos(\mu_n) &= R_0 \cos(\mu_0) + \sum \cos(\theta_i) \\ R_n \sin(\mu_n) &= R_0 \sin(\mu_0) + \sum \sin(\theta_i). \end{aligned}$$

5.2. Inferencias vía *Gibbs sampling*

Damien y Walker (1999) introducen las variables latentes w, v, u y x y, definen

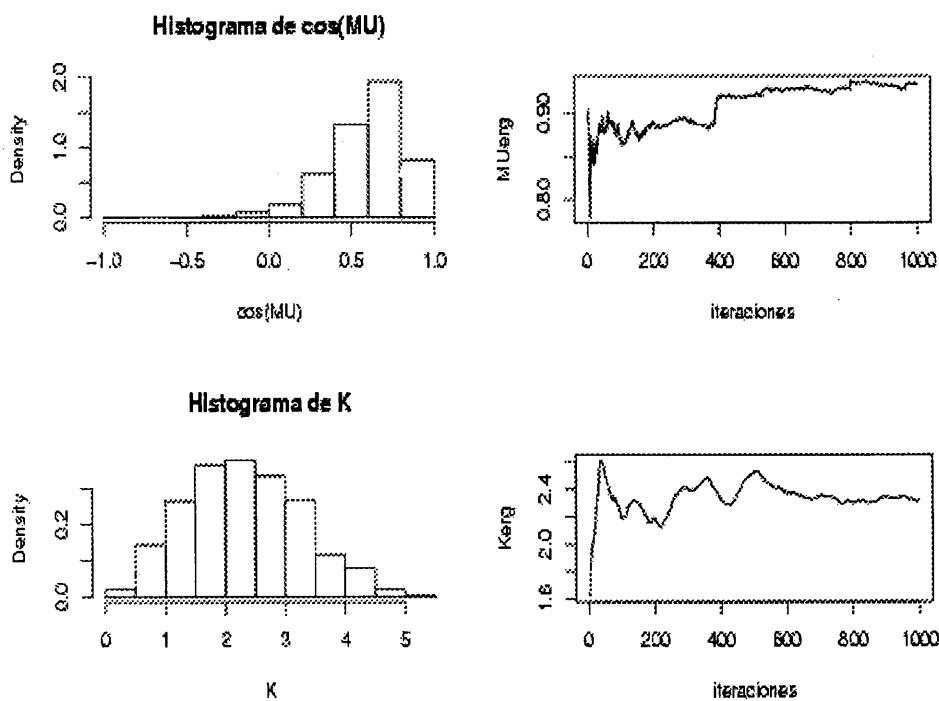
$$\begin{aligned} f(\mu, k, w, v, u, x) &\propto e^{R_n k} I(v < \exp[R_n k \{1 + \cos(\mu - \mu_n)\}], x < w^{m-1}) \\ &\times e^{-w} \prod_{i=0}^{\infty} I(u_i < e^{-w \lambda_i k^{2i}}) \end{aligned}$$

A partir de la definición anterior se pueden obtener todas las densidades condicionales completas y llevar a cabo un *Gibbs sampling* para realizar inferencias sobre los parámetros μ , k .

6. Ejemplo

A continuación se presenta un ejemplo de la metodología presentada en este trabajo. Los datos fueron tomados de Mardia (1972) y consisten de $n = 9$ observaciones asociadas a los resultados de una ruleta. Para estos datos los estimadores correspondientes de máxima verosimilitud son $\hat{\mu} = 0,89$ radianes y $k = 2,08$

Se obtuvieron simulaciones de la distribución final de (μ, k) usando el algoritmo presentado en la sección anterior. La distribución final de $\cos(\mu)$ y de k se muestran a continuación y concuerdan con los obtenidos por Damien y Walker (1999).



Referencias

- Damien, P. y Walker, S. (1999). A full Bayesian analysis of circular data using the von Mises distribution, *The Canadian Journal of Statistics*, **2**, 291-298.
- Fisher, N.I., Lewis, T., y Embleton, B.J.J. (1987). Statistical Analysis of Spherical Data. Cambridge University Press.
- Guttorp, P. y Lockhart, R. A. (1988). Finding the location a signal: A Bayesian analysis. *J. Amer. Statist. Assoc.*, **83**, 322-329.
- Mardia, K. V. (1972). *Statistics of Direccional Data*. Academic Press, London.
- Mardia, K. V. y Jupp, E. P. (2000). *Directional Statistics*. John Wiley and Sons Ltd.

Hacia una Nueva Pedagogía: El Enfoque Basado en Proyectos para Mejorar el Aprendizaje del Diseño Estadístico

Mario Miguel Ojeda

Edgar Morales

Margarita Caballero

Norma V. Galeana

Laboratorio de Investigación y Asesoría Estadística (LINAЕ)

Facultad de Estadística e Informática

Universidad Veracruzana

1. Introducción

Parte de la problemática actual de la educación, es que los esfuerzos están principalmente enfocados a las metodologías para enseñar que al aprendizaje (Behar, 2001), destacándose una urgente necesidad de diseñar y aplicar procesos de innovación educativa en todos los ámbitos (Rivas-Navarro, 2000), que garanticen el aprendizaje significativo, permanente y que produzca habilidades y actitudes para que el estudiante siga aprendiendo (aprender a aprender). Las materias relacionadas con las matemáticas (Resnick y Ford, 1981), y particularmente la estadística (Barahona, 1997; Garfield, 1995), enfrentan serias dificultades por los bajos niveles de aprendizaje de los estudiantes (Rodríguez-Rebustillo y Bermúdez-Sarguera, 2001; Defior-Citola, 1996; Harlen, 1994). Para resolver esta situación los especialistas en educación estadística y los estadísticos (Moore, 1997; Sahai et al., 1997) han recomendado una serie

de medidas: a) garantizar que la estadística se valore por su utilidad para resolver problemas reales; b) garantizar el aprendizaje de conceptos clave y principios, más que enfatizar en fórmulas y procedimientos; c) promover el apropiado uso de los métodos y técnicas en una amplia variedad de actividades científicas y profesionales; y d) incorporar los adelantos tecnológicos e instruccionales para garantizar el desarrollo de competencias. Uno de los enfoques que se ha utilizado para mejorar los niveles de aprendizaje de la estadística es el enfoque basado en proyectos. Diseñar un curso con este enfoque, consiste en que todas las actividades y los contenidos se planean para que los participantes elaboren y desarrollen un proyecto (Gagné y Briggs, 1976). Se utiliza el esquema de conferencias para transmitir la información clave, de manera organizada y en forma suscinta. Estas conferencias se articulan con actividades prácticas que se desarrollan por equipos de hasta tres participantes, para propiciar el aprendizaje cooperativo. Tanto las conferencias como las actividades prácticas se apoyan con un material de estudio, que incluye la información de las conferencias y los lineamientos generales para realizar y desarrollar el proyecto. Tanto en las conferencias como en el material de estudio se incluyen ilustraciones de proyectos, que constituyen los referentes de forma y contenido para que, por una suerte de imitación, los participantes realicen su proyecto. En este sentido se propicia el aprendizaje por transferencia (Postic, 1996; Beltrán, 1996).

La Universidad Veracruzana en sus políticas de desarrollo académico (UV, 1998). Plantea el propósito de promover la investigación entre los docentes, para que ésta sea llevada a los cursos como un instrumento para propiciar el aprendizaje significativo y la formación intelectual (UV, 1999). En este contexto la estadística, como una metodología para el diseño y realización de investigaciones factuales, es fundamental: es parte integral del proceso de investigación, tanto en la realización del protocolo y en la obtención y análisis de datos, como en la elaboración del informe. Se llevó a cabo, durante el mes de julio de 2001, un curso-taller de “Estadística en la Investigación”, que tuvo el propósito de promover el correcto uso del diseño estadístico en la elaboración de protocolos de investigación, utilizando un modelo

simple que pudiese ser transferido a las aulas. Se planteó el reto de cambiar la visión de la estadística como una metodología útil en la fase de diseño de proyectos y no sólo como una herramienta para el análisis de datos.

El propósito de este trabajo es el de evaluar los resultados de esta modalidad de capacitación que se diseñó utilizando el enfoque basado en proyectos. Se analizan los resultados considerando dos encuestas (una al inicio y otra al final) y los proyectos entregados por los participantes como requerimiento para la acreditación. Al final se apuntan algunos comentarios y conclusiones.

2. Estructura del curso-taller

El contenido del curso taller se especifica en la Tabla 1.

3. Implementación del curso-taller

El curso-taller fue promovido ampliamente durante la primera quincena de julio de 2001. La distribución de los académicos de las cinco regiones de la Universidad Veracruzana que se inscribieron (participantes) y de aquellos que realizaron todas las actividades planeadas se presenta en la Tabla 2.

4. Análisis estadísticos

Se realizaron análisis básicos de frecuencias y porcentajes para las variables categóricas con la finalidad de explorar univariadamente las distribuciones en estudio. La variable edad, de naturaleza continua, fue analizada a través de un histograma, de su media y desviación estándar.

Tabla 1. Descripción de actividades y contenidos del programa del curso-taller.

Actividad	Título	Contenido	Tiempo
Videoconferencia 1	La metodología estadística en el proceso de investigación.	Conceptualización de la estadística; Fases de la investigación y metodología estadística en general. Aspectos básicos del diseño estadístico; Aspectos básicos del análisis estadístico; Elaboración del reporte; Conclusiones.	1 hora.
Sesión presencial	Conceptos clave en el diseño estadístico.	Diseño estadístico; Clasificación de estudios estadísticos; Muestreo de poblaciones; Diseños experimentales; Estudios observacionales; Tamaño de muestra.	2.5 horas.
Videoconferencia 2	Lineamientos para elaborar protocolos de investigación.	Partes fundamentales del protocolo; Redacción de objetivos; Traducción a objetivos estadísticos; Elementos para describir adecuadamente el diseño estadístico; Elementos para describir adecuadamente el análisis estadístico; Otros elementos del protocolo; Ilustración.	1 hora.
Sesión presencial	Revisión de ejemplos de protocolos.	Protocolo de un estudio observacional; Protocolo de un estudio experimental; Protocolo de un estudio muestral.	2.5 horas.
Videoconferencia 3	La estadística descriptiva univariada y multivariada.	Un poco de historia; Estadística exploratoria; Las seis reglas en el análisis de datos; Estrategias de análisis de datos en el proceso de investigación; Análisis univariados y bivariados; Análisis comparativos; Análisis multivariados; Conclusiones.	1 hora.
Sesión presencial	Elaboración de protocolos con supervisión.	Delimitación del problema; Antecedentes; Justificación.	2.5 horas.
Videoconferencia 4	La estadística inferencial paramétrica y no paramétrica.	Escalas de medición; Prueba Ji-cuadrada; Análisis de correspondencias.	1 hora.
Sesión presencial	Elaboración de protocolos con supervisión.	Metodología; Cronograma de actividades; Referencias.	2.5 horas.

Tabla 2. Distribución de los participantes en el taller por región.

Región	Total de participantes	Participantes incluidos en la evaluación
Coatza - Mina	86	39
Córdoba - Orizaba	89	33
Poza Rica-Tuxpan	45	36
Veracruz	77	40
Xalapa	70	30
TOTAL	367	178

Se pretende estudiar la asociación entre las variables que caracterizan al académico (sexo, cursos de estadística, años de docencia, máximo grado de estudios, área de profesión, región y experiencias en proyectos de investigación) con el efecto del curso-taller en el concepto de estadística que expresaron (cambio), por un lado, y por el otro con el desempeño logrado en la elaboración del protocolo. Para evaluar la significancia en él, se aplicó la prueba de McNemar. Finalmente, para evaluar de manera conjunta las asociaciones, entre las variables que en los análisis previos resultaron significativas, se realizó un análisis de correspondencia múltiple.

5. Resultados

El curso-taller tuvo una respuesta muy importante en las cinco regiones universitarias, pero hay que reconocer que el porcentaje de aquellos que cumplieron con todas las actividades planeadas (49 %) no resultó satisfactorio. En cuanto a las características de los participantes: estuvo equilibrada por sexo; la edad más frecuente estuvo entre 46 y 52 años, con una distribución aproximadamente normal; los años de antigüedad fueron de más de 20 años para el 45 %, entre 10 y 20 años el 32 %, notándose una baja participación de académicos de poca

experiencia. Por otro lado, en cuanto a su área de trabajo, notamos una mayor participación (40 %) del área técnica y de las ingenierías, segundándola (38 %) el área biológica y de ciencias de la salud; y el resto provinieron de carreras de administración, contaduría, etc. Respecto al nivel de estudios, la mayoría sólo reportó licenciatura (46 %), aunque un 40 % dijo contar con el nivel de maestría, con un 12 % de especialización y una mínima participación (1 %) con el nivel de doctorado. En lo que se refiere a la experiencia previa de los participantes en proyectos de investigación, 7 de cada 10 dijeron que sí han realizado actividades de este tipo. En la Figura 1 se muestra la distribución del número de cursos previos de estadística observándose que 7 de cada 10 han tomado al menos un curso; sin embargo, el concepto de estadística al inicio del curso-taller resultó en un 80 % limitado a caracterizar la estadística relacionándola con el análisis de los datos o bien declarando un concepto difuso. A pesar de esto, 6 de cada 10 declararon necesitar de la metodología estadística frecuentemente o siempre, en el contexto de su trabajo académico. En lo que se refiere al desempeño en la elaboración del protocolo, 3 de cada 10 resultaron deficientes y otros 3 sólo alcanzaron la categoría de regular (ver Figura 2). El resultado más importante del taller nos indica que hubo un cambio significativo en las distribuciones de la forma como se conceptualiza la estadística, evolucionando de manera importante de un concepto difuso a un concepto limitado y, en menor medida, de un concepto limitado a un concepto amplio, aunque esto se dio en un porcentaje bajo de participantes, aproximadamente 2 de cada 10 (ver Tabla 3). La prueba de McNemar detectó significancia estadística ($p=0.025$). Cuando se cruzaron las variables de características del participante con sus resultados en el cambio de concepto no se detectó asociación alguna, pero en el desempeño evaluado por el protocolo entregado se encontraron asociaciones con el máximo grado de estudios ($p=0.014$), región de procedencia ($p\leq0.001$) y experiencia docente ($p=0.025$). Las asociaciones múltiples nos arrojaron una caracterización por regiones, donde Veracruz se destaca por desempeños deficientes en la elaboración del protocolo a pesar de que hay una alta frecuencia de participantes con maestría y doctorado y experiencia docente entre 10 y 20 años. Poza Rica, por otro lado, destaca

por la frecuencia de excelentes protocolos aunque el nivel de estudios predominante es la especialidad (ver Figura 3).

Figura 1: Gráfico de barras representando el número de cursos previos de estadística

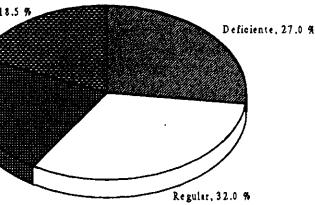
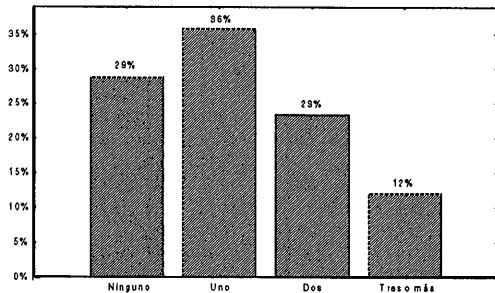


Figura 2: Gráfico circular de la distribución del desempeño en la elaboración del protocolo.

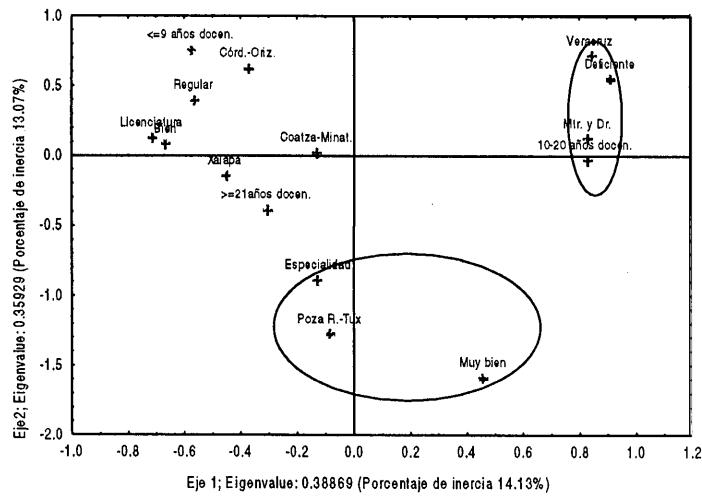


Figura 3: Mapa de correspondencia múltiple.

Tabla 3: Distribución del concepto de estadística (antes y después del curso-taller).

Concepto antes	Concepto después			Totales Renglón
	Difuso	Limitado	Amplio	
Difuso Porcen. renglón	10 27.78 %	25 69.44 %	1 2.78 %	36
Limitado Porcen. renglón	3 2.94 %	76 74.51 %	23 22.55 %	102
Amplio Porcen. renglón	2 5.13 %	8 20.51 %	29 74.36 %	39
Totales	15	109	53	177

6. Conclusiones y discusión

Se obtuvo que el número de cursos previos de estadística y el grado de estudios, no influyeron en el concepto de estadística al inicio del curso-taller, ya que la mayoría de los participantes aportó un concepto difuso o limitado. Esto es coincidente con lo que han señalado diversos autores, como es citado por Behar (2001), en el sentido de que la enseñanza tradicional de la estadística resulta poco efectiva. Considerando que uno de los propósitos del curso-taller fue el de cambiar la visión de la estadística como una metodología útil en la fase de diseño de proyectos, la evaluación permitió conocer la existencia de cambios significativos en los participantes. Esto reafirma lo ya expuesto por Anderson y Sungur (1999), Startking (1997) y Kvam (2000), acerca de que el enfoque basado en proyectos mejora los niveles de aprendizaje de la estadística en general; sin embargo, no se debe sobrevalorar los impactos, ya que desde la perspectiva de los objetivos y metas, los resultados están lejos de ser satisfactorios. En este sentido debemos aceptar que la evidencia sobre el desempeño en la escritura de los protocolos, nos indica que se tiene que reforzar el diseño del taller y garantizar mayor asesoría a los participantes. Posibles medidas a considerar en el futuro serán diseñar una

página WEB (León y Parr, 2000), con diferentes elementos que ayuden al trabajo independiente, como ilustraciones en línea, señalando errores frecuentes, guías didácticas, etc. Algo que se logró indudablemente fue elevar la motivación por el aprendizaje del diseño y análisis estadístico en el contexto del proceso de investigación, ya que casi la totalidad de los participantes declaró desear que se continúe con el taller hasta la fase de escritura del reporte, pasando por la fase de análisis de los datos. En este sentido se logró derribar una importante barrera, como lo menciona Defior-Citoler (1996), la referente a la motivación y a la disposición para abordar las actividades que producen el aprendizaje. Respecto al contexto en el que se desenvuelven los participantes, se logró avanzar en la promoción de una cultura académica que considera al proceso de investigación como un instrumento valioso para la formación de los estudiantes. En este sentido el modelo de protocolo revisado permitirá una gradual transferencia del enfoque basado en proyectos a las aulas universitarias. En lo que se refiere a la organización del taller se requiere en futuras ediciones incrementar el número de monitores y elevar su capacidad de asesoría, ya que la mayoría de las sugerencias de mejora van en ese sentido. Finalmente, un aspecto a analizar con mayor profundidad es lo referente a las diferencias entre regiones, que posiblemente se deba a diferencias entre los auxiliares.

Referencias

- Anderson, J. E. and Sungur E. A. (1999). Community service statistics projects. **The American Statistician**, 53, 132-136.
- Barahona, C. (1997). Biometrical Education: Problems, Experiences and Solutions. In: **Biometric Education: Problems, Experiences and Solutions**, Camacho, J. (Ed). International Biometric Society Network for Central America and the Caribbean, 43-61.
- Behar, R. (2001). **Aportaciones para la Mejora del Proceso de Enseñanza-Aprendizaje**

de la Estadística. Tesis Doctoral, Universidad Politécnica de Cataluña, Barcelona, España.

Beltrán, J. (1996). **Procesos, Estrategias y Técnicas de Aprendizaje**. Editorial Síntesis, Madrid.

Defior-Citoler, S. (1996). **Las Dificultades de Aprendizaje: Un Enfoque Cognitivo**. Ediciones Aljibe, Granada, España.

Gagné, R. M. y Briggs, L. J. (1976). **La Planificación de la Enseñanza: Sus Principios**. Trillas, México D.F.

Garfield, J. (1995). How students learn statistics?. **International Statistical Review**, 63, 25-34.

Harlem, W. (1994). **Enseñanza y Aprendizaje de las Ciencias**. Segunda Edición. Ediciones Morata, Ministerio de Educación y Ciencia, Madrid.

Kvam , P. H. (2000). The effect of active learning methods on student retention in engineering statistics. **The American Statistician**, 54, 136-140.

León, R. V. and Parr, W. C. (2000). Use of course home pages in teaching statistics. **The American Statistician**, 54 (1), 44-48.

Mayor, J., Suengas, A. Y González-Marqués, J. (1996). **Estrategias Metacognitivas: Aprender a Aprender y Aprender a Pensar**. Editorial Síntesis, Madrid.

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. **International Statistical Review**, 65, 123-165.

Ojeda, M. M. (1994). La importancia de una buena cultura estadística en la investigación.

La Ciencia y el Hombre, 17, 143-152.

Ojeda, M. M. and Sahai, H. (1995). **A General proposal for teaching undergraduate statistics service Courses**. Proceeding of the ASA Section on Statistical Education, pp. 311-316. American Statistician Association, Alexandria, Virginia, USA.

Postic, M. (1996). **Observación y Formación de los Profesores**. Segunda Edición. Ediciones Morata, Madrid.

Resnick, L. B. and Ford, W. W. (1981). **The Psychology of Mathematics for Instruction**. Lawrence Erlbaum Associates, Hillsdale, New Jersey, USA.

Rodríguez-Robustillo, M. y Bermúdez-Sarguera, R. (2001). **Psicología del Pensamiento Científico**. Editorial Pueblo y Educación, La Habana, Cuba.

Sahai, H., Behar, R. and Ojeda, M. M. (1997). A reformulation of the problem of statistical education : A learning perspective. In: **Biometric Education: Problems Experiences and Solutions**, Camacho, J. (Ed). International Biometric Society Network for Central America and the Caribbean 75-106.

Startkings, S. (1997). Assessing students projects. In: The Assessmet Challenge in Statistics Education, (Eds. Gal and Garfield). IOS Press Amsterdam.

UV (1998). **Consolidación y Proyección Hacia el Siglo XXI**. Universidad Veracruzana, Xalapa, Ver., México.

UV (1999). **Nuevo Modelo Educativo para la Universidad Veracruzana**. Universidad Veracruzana, Xalapa, Ver., México.

Las Componentes Principales como Análisis de Datos en Poblaciones de Maíz

Emilio Padrón Corral

Universidad Autónoma Agraria Antonio Narro y

Universidad Autónoma de Coahuila

Ignacio Méndez Ramírez

IIMAS, UNAM

Armando Muñoz Urbina

Universidad Autónoma Agraria Antonio Narro

Nidia Hernández Pérez

Universidad Autónoma de Coahuila

1. Introducción

Cuando se desea generar mayor información acerca de los materiales y de las características evaluadas en conjunto, la aplicación de alguna técnica multivariada puede ser, generalmente, una buena opción. Muchas de las técnicas multivariadas permiten valorar correctamente la correlación entre las variables Manly(1990). En particular el Análisis de Componentes Principales(ACP), permite reducir la cantidad de datos por interpretar sin perder mucha de la información que se busca. Por lo tanto, es una herramienta adecuada para analizar

la estructura de observaciones multivariadas cuando se busca evaluar la dependencia entre las distintas variables y se desconocen los patrones de interrelación Mardia et al. (1979). En el programa de mejoramiento genético del Instituto Mexicano del Maíz, “Dr. Mario E. Castro Gil” se aplicó la metodología de selección recíproca recurrente (SRR), Bruce y Lamkey (1993), en cuatro poblaciones sintéticas de maíz, en poblaciones de madurez intermedia Pob.(A) y Pob.(B) en el experimento I, y poblaciones de madurez tardía Pob.(43) y Pob.(23) en el experimento II. Las variables de respuesta utilizadas fueron: rendimiento(REN), por ciento de mazorcas podridas (MZP), por ciento de mazorcas con fusarium (MZF), días a flor masculina (FMA), días a flor femenina (FFE), altura de planta (APL) y altura de mazorca (AMZ) (Muñoz(2000)). Se busca determinar si la variabilidad total puede ser explicada a partir de sólo algunas variables más importantes.

2. Metodología

En el (ACP) se toman p variables X_1, X_2, \dots, X_p con el fin de encontrar combinaciones lineales de ellas para producir índices Z_1, Z_2, \dots, Z_p que no están correlacionados. Los pasos en un (ACP) pueden ser.

- a) Estandarizar las variables X_1, X_2, \dots, X_p para tener media cero y varianza uno.
- b) Calcular la matriz de covarianza, esta es una matriz de correlación si se ha realizado el paso a).
- c) Encontrar los valores propios $\lambda_1, \lambda_2, \dots, \lambda_p$ y los correspondientes vectores a_1, a_2, \dots, a_p de la matriz anterior. Los coeficientes del i -ésimo componente principal son dados por a_i mientras que λ_i es su varianza.

- d) Descartar aquellos componentes principales que sólo representan una pequeña proporción de la variación de los datos.

En el (ACP) los valores propios son estimados de la matriz de covarianza muestral, la cual es simétrica. Las varianzas de los componentes principales están dadas por estos valores propios, hay p valores propios, de los cuales algunos pueden ser iguales a cero. Suponiendo que los valores propios son ordenados como $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$, entonces λ_i es la varianza correspondiente al i -ésimo componente principal, y el i -ésimo componente principal está dado por: $Z_i = a_{i1}X_1 + a_{i2}X_2 + \dots + a_{ip}X_p$ por lo tanto, $Var(Z_i) = \lambda_i$ y las constantes $a_{i1}, a_{i2}, \dots, a_{ip}$ son los elementos correspondientes al vector propio, es decir, las ponderaciones de las variables originales dentro de la i -ésima componente principal.

3. Resultados y Discusión

En la Tabla 1 se presentan los valores propios para los mestizos de la población Pob.(A) Exp.I, en el cual puede observarse que los primeros tres componentes principales son suficientes para explicar el 81.2 por ciento de la variabilidad de los datos. De los elementos del vector propio asociados al primer componente principal (CP1) se observa que las variables FMA (0.477), FFE(0.476) y REN(-0.428), son las que más influyen sobre este componente; se presentó además un comportamiento opuesto entre las variables REN, APL, AMZ y las variables MZP, MZF, FMA y FFE. La variable AMZ(-0.657) proporciona información sobre el segundo componente principal (CP2) y las variables MZF(0.465) y MZP(0.463) son las de mayor aportación al tercer componente principal (CP3). Las poblaciones Pob.(B) Exp.I, y Pob.(43) Exp.II, presentaron un comportamiento similar al de Pob.(A), con respecto a los tres primeros componentes principales (Tablas 2 y 3) en estas poblaciones también se observa que estos tres componentes explicaron más del 80 por ciento de la variabilidad de los datos. Las variables REN, APL y AMZ presentaron un comportamiento opuesto a MZP,

MZF,FMA y FFE. Las variables REN,FMA y FFE son las que más aportan al CP1 mientras que AMZ es la variable que más aporta al CP2. Con respecto al CP3 la variable MZF es la de mayor aportación en la población Pob.(B), y MZP en la población Pob.(43). En el caso de la población Pob.(23) Exp.II, las tres primeras componentes principales explicaron el 78.17 por ciento de la variabilidad de los datos Tabla 4. En esta población, a diferencia de las otras, REN presentó un comportamiento opuesto a APL y AMZ en CP1. La variable APL presentó mayor influencia que AMZ en el CP2. Finalmente se observa que MZP tuvo mayor influencia sobre el CP3. Del análisis de componentes principales de los mestizos del Exp.1 se puede determinar que el CP1 separó a los 154 mestizos de la población Pob.(A) en dos grandes grupos, ubicando a los de mayor precocidad y rendimiento como FMA,FFE y opuestos a REN y MZF, mientras que el CP2 separó a los mestizos de acuerdo a las alturas de mazorca y de planta.

Tabla 1. Elementos del vector propio asociados a los primeros tres componentes principales de los mestizos de la Población Pob.(A) Exp. I

Característica agronómica	Vectores propios		
	CP1	CP2	CP3
REN	-0.428	-0.022	-0.403
MZP	0.247	-0.147	0.463
MZF	0.413	0.010	0.465
FMA	0.477	-0.300	-0.423
FFE	0.476	-0.307	-0.419
APL	-0.302	-0.602	0.141
AMZ	-0.198	-0.657	0.178
Valor Propio	2.567	1.640	1.477
Variación explicada	36.671	23.431	21.099
Variación explicada acumulada	36.671	60.101	81.200

Tabla 2. Elementos del vector propio asociados a los primeros tres componentes principales de los mestizos de la Población Pob.(B) Exp. I

Característica agronómica	Vectores propios		
	CP1	CP2	CP3
REN	0.469	-0.244	0.263
MZP	-0.358	0.029	-0.356
MZF	-0.424	0.210	-0.459
FMA	-0.474	-0.358	0.374
FFE	-0.471	-0.358	0.381
APL	0.134	-0.546	-0.427
AMZ	0.088	-0.584	-0.356
Valor propio	2.544	1.937	1.253
Variación explicada	36.348	27.671	17.897
Variación explicada acumulada	36.348	64.019	81.916

Tabla 3. Elementos del vector propio asociados a los primeros tres componentes principales de los mestizos de la población Pob.(43) Exp.II

Característica agronómica	Vectores propios		
	CP1	CP2	CP3
REN	-0.494	0.186	-0.198
MZP	0.292	-0.131	0.556
MZF	0.421	-0.157	0.419
FMA	0.475	0.322	-0.385
FFE	0.476	0.330	-0.376
APL	-0.197	0.578	0.308
AMZ	-0.047	0.626	0.303
Valores propios	2.711	1.875	1.194
Variación explicada	38.728	26.788	17.051
Variación explicada acumulada	38.728	65.516	82.567

Tabla 4. Elementos del vector propio asociados a los primeros tres componentes principales de los mestizos de la población Pob.(23) Exp.II

Característica agronómica	Vectores propios		
	CP1	CP2	CP3
REN	0.352	0.388	-0.111
MZP	-0.185	-0.100	0.733
MZF	-0.078	-0.432	0.425
FMA	-0.631	0.022	-0.226
FFE	-0.628	0.019	-0.237
APL	-0.118	0.605	0.205
AMZ	-0.171	0.535	0.346
Valor propio	2.299	2.073	1.100
Variación explicada	32.847	29.616	15.709
Variación explicada acumulada	32.847	62.463	78.172

4. Conclusiones

El Análisis de Componentes Principales (ACP) permitió la ordenación de los tratamientos de acuerdo a los tres primeros componentes, los cuales explican más del 80 por ciento de la variación de los datos. Con el ACP se determinó que las variables FMA, FFE y REN tuvieron una mayor influencia sobre el CP1. El CP2 separó a los mestizos con respecto a APL y AMZ. Las variables MZP y MZF presentaron mayor influencia sobre el CP3 y presentaron un comportamiento opuesto a REN, FMA y FFE, es decir, que de acuerdo al porcentaje de variabilidad explicado por este componente existen mestizos con alto rendimiento y tardíos que presentan bajos porcentajes de MZP y MZF.

Referencias

Bruce,J.S., y Lamkey,K.R.(1993). Interpopulation genetic variance after reciprocal recurrent selection in BSSS and BSCB1 maize populations. *Crop Sci*, **33**, 90-95.

Manly,B.F.J.(1986). *Multivariate Statistical Methods*. New York: Chamoman and Hall. p.59-125.

Mardia,K.V., Kent,J.T. and Bibby,J.M.(1979). *Multivariate Analysis*. New York: Academic Press. p.227-243.

Muñoz,U.A. (2000). Selección Reciproca Recurrente en Poblaciones de Maíz para Trópico Seco y Bajío Mexicano. Tesis para obtener el Grado de Doctor en Ciencias en Fitomejoramiento. Universidad Autónoma Agraria Antonio Narro.

Uso de Variable Indicadoras en Predicción Espacial

Felipe Peraza

*Centro de Investigación en Matemáticas y
Universidad Autónoma de Sinaloa*

Graciela González-Farías

Centro de Investigación en Matemáticas

1. Introducción

En el análisis de datos espaciales es usual considerar un modelo que contenga una función de deriva determinística y un error aditivo,

$$y(s) = m(s) + u(s), \quad s \in \mathbb{R}^2.$$

El término residual $u(s)$ puede o no ser estacionario pero tiene esperanza cero. Para cada par de sitios s_i, s_j la covarianza entre $u(s_i)$ y $u(s_j)$ es función de un vector de parámetros θ , se le llamará a esta Ω .

Para Ω conocida, Fuller (1980) muestra el uso de variables indicadoras para predecir simultáneamente con el proceso de estimación de parámetros, en el modelo lineal. Este método puede extenderse a modelos no lineales, sistemas de ecuaciones y series de tiempo autoregresivas. Aquí se muestra que también pueden aplicarse de manera directa a modelos para datos espaciales que contengan una deriva y una función de covarianzas parametrizada en un

campo Gaussiano, como puede ser el caso de modelos con fuentes puntuales Hughes-Oliver y González Farías (1999).

En el caso de Ω desconocida, se requiere explorar métodos para estimar tanto θ , el parámetro de la covarianza, como β el parámetro de la deriva. Cuando β es desconocido, se pueden obtener estimadores menos sesgados para θ usando el método de Máxima Verosimilitud Restringida (REML). Una vez que los parámetros son estimados, el siguiente paso en el análisis es la predicción en varios sitios. El uso de variables indicadoras proporciona un método para estimar β y predecir en cualquier número de sitios simultáneamente. En el presente trabajo se muestra que, como es de esperarse, la inclusión de cualquier número de variables indicadoras no afecta el procedimiento de estimación de los parámetros de la covarianza. El estudio de simulación muestra la ventaja del método de REML sobre Máxima Verosimilitud (MV).

2. Variables Indicadoras

Sea $\{y(s, t); (s, t) \in D\}$ una función aleatoria, donde $D \subset \mathbb{R}^2$, y

$$y(s, t) = \sum_{j=1}^{k+1} f_{j-1}(s, t) \beta_j + u(s, t), \quad (13)$$

donde $\beta = (\beta_0, \dots, \beta_k)'$ es desconocido, las funciones f_0, f_1, \dots, f_k son conocidas y $u(., .)$ es un proceso aleatorio con media cero.

Si se tienen observaciones en n sitios, se puede escribir (13) como

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad (14)$$

donde y es un vector de observaciones de dimensión $n \times 1$ y \mathbf{X} es una matriz de $n \times (k+1)$ con elementos (i, j) dados por $f_{j-1}(s_i, t_i)$, y donde $E(\mathbf{u}) = \mathbf{0}$ y $cov(\mathbf{u}) = \Omega_{11}$.

Para predecir la respuesta en p sitios distintos por ejemplo (s_1^0, t_1^0) , $(s_2^0, t_2^0), \dots, (s_p^0, t_p^0)$ se forma la matriz $\mathbf{X}_0 : p \times (k + 1)$, donde el (i, j) -ésimo elemento está dado por $f_{j-1}(s_i^0, t_i^0)$. Entonces, si \mathbf{y}_0 es el vector no observado de dimensión $(p \times 1)$,

$$\mathbf{y}_0 = \mathbf{X}_0\beta + \mathbf{u}_0. \quad (15)$$

$$cov \begin{pmatrix} \mathbf{u} \\ \mathbf{u}_0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix} = \boldsymbol{\Omega}.$$

Siguiendo Fuller (1980) se combinan (14) y (15) de la siguiente forma

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{0}_{p \times 1} \end{pmatrix} = \begin{pmatrix} \mathbf{X} & \mathbf{0}_{n \times p} \\ \mathbf{X}_0 & -\mathbf{I}_{p \times p} \end{pmatrix} \begin{pmatrix} \beta \\ \mathbf{y}_0 \end{pmatrix} + \begin{pmatrix} \mathbf{u} \\ \mathbf{u}_0 \end{pmatrix} \quad (16)$$

o

$$\mathbf{z} = \mathbf{W}\gamma + \mathbf{e}. \quad (17)$$

Entonces el Estimador de Mínimos Cuadrados Generalizado (EMCG) de γ , (el estimador EMCG para β y el Mejor Predictor Lineal Insesgado para \mathbf{y}_0) está dado por:

$$\hat{\gamma} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\mathbf{y}}_0 \end{pmatrix} = \begin{pmatrix} (\mathbf{X}'\boldsymbol{\Omega}_{11}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}_{11}^{-1}\mathbf{y} \\ \mathbf{X}_0\hat{\beta} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}) \end{pmatrix} \quad (18)$$

Es claro que se necesita conocer la función de covarianzas $\boldsymbol{\Omega}$ para calcular (18) y dado que es desconocida se requiere método para estimarla.

3. Estimación de Parámetros

En esta sección se describe la implementación de los métodos de Máxima Verosimilitud y Máxima Verosimilitud Restringida en términos del modelo (16). Para ello, suponga que el error \mathbf{e} es un campo aleatorio Gaussiano con función de covarianzas parametrizada $\Omega = R(\mathbf{s}_i, \mathbf{s}_j; \theta)$.

En (16), es claro que la dimensión del vector de parámetros $\gamma = (\beta', \mathbf{y}'_0)$ se incrementa como el número de sitios a predecir. Entonces es razonable usar REML para estimar θ y más aún, se puede mostrar que el estimador REML para θ usando γ (modelo: (16)) es igual al estimador REML usando solo β (modelo: (14)). Por lo tanto, la dimensión de \mathbf{y}_0 no tiene efecto en los estimadores REML.

Para encontrar $\hat{\theta}_{REML}$, el estimador REML de θ considerando variables indicadoras (modelo: 17) se maximiza la función:

$$l^*(\theta) = -\frac{1}{2} \log(|\Omega|) - \frac{1}{2} \log(|\mathbf{W}'\Omega^{-1}\mathbf{W}|) - \frac{1}{2} (\mathbf{z} - \mathbf{W}\hat{\gamma})' \Omega^{-1} (\mathbf{z} - \mathbf{W}\hat{\gamma}). \quad (19)$$

Un poco de álgebra prueba que esta función es igual a la verosimilitud restringida sin considerar variables indicadoras (modelo: (14))

$$l_1^*(\theta) = -\frac{1}{2} \log(|\Omega_{11}|) - \frac{1}{2} \log(|\mathbf{X}'\Omega_{11}^{-1}\mathbf{X}|) - \frac{1}{2} (\mathbf{y} - \mathbf{X}\hat{\beta})' \Omega_{11}^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta}) \quad (20)$$

y entonces no hay diferencia entre el REML de θ usando variables indicadoras y la información de los sitios observados, esto es, la información contenida en \mathbf{X} .

El estimador MV con variables indicadoras $\hat{\theta}_{MV_i}$ se obtiene maximizando

$$l(\theta) = -\frac{1}{2} \log(|\Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}|) - \frac{1}{2} \log|\Omega_{11}| - \frac{1}{2} (\mathbf{y} - \mathbf{X}\hat{\beta})' \Omega_{11}^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta}).$$

Si se utiliza sólo la información original, la verosimilitud está dada por:

$$l(\theta) = -\frac{1}{2} \log(|\Omega_{11}|) - \frac{1}{2} (\mathbf{y} - \mathbf{X}\hat{\beta})' \Omega_{11}^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta}).$$

y obtenemos $\hat{\theta}_{MV}$.

Entonces la diferencia entre la log-verosimilitud usando variables indicadoras y la log-verosimilitud sin usarlas, es el término relacionado con el error de estimación. El sesgo se incrementa con el número de puntos a predecir.

Así pues, para n fijo y sin importar en cuantos sitios se quiere predecir, el estimador REML de θ permanece igual.

4. Simulaciones

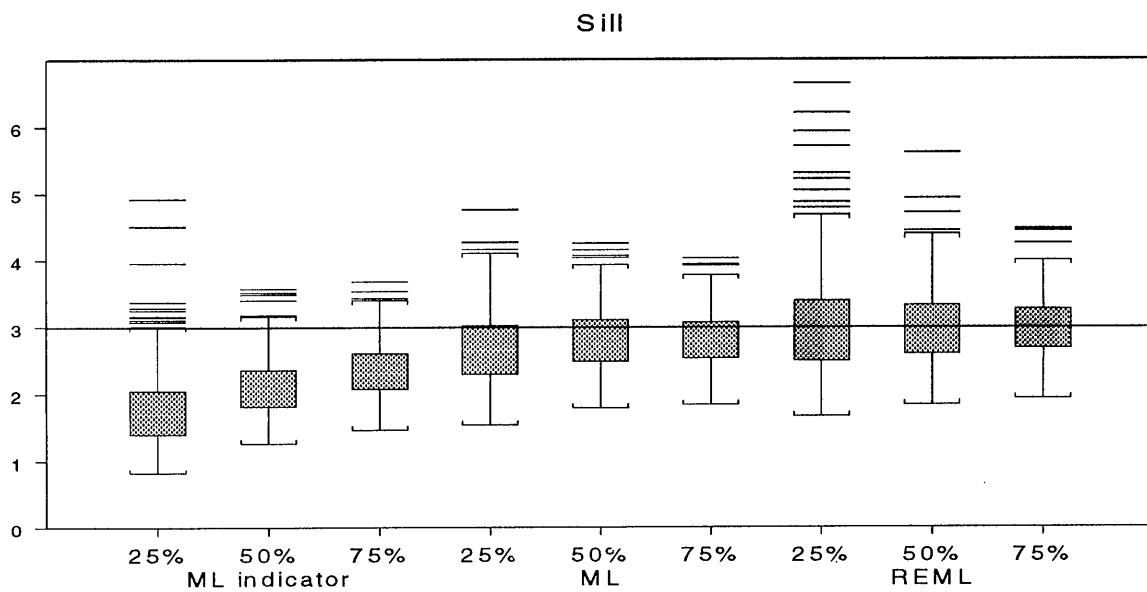
El propósito de este estudio de simulación es confirmar la diferencia entre Máxima Verosimilitud y Máxima Verosimilitud Restringida en la estimación de los parámetros de la covarianza y de regresión usando una estructura de covarianza exponencial y una deriva lineal en un campo aleatorio Gaussiano, usando variables indicadoras. Existen otros estudios en la literatura en los cuales se usa REML, ver por ejemplo Zimmerman (1989), y Zimmerman and Zimmerman (1991). Sin embargo cuando al usarse variables indicadoras se observa que los estimadores REML no son afectados así se usen (14) ó (17) a diferencia de los estimadores MV.

Se generaron 500 muestras en una rejilla regular de 16×16 . Los valores de los parámetros fueron $\theta_1 = 3$ para la meseta y $\theta_2 = 1$ para el rango en la función de covarianzas exponencial, $C(h) = \theta_1 e^{-\theta_2 h}$. Se utilizó la deriva $m(\mathbf{x}) = 1 + 2x_1 + 3x_2$.

El 25 % de los datos ($k = 64$) se reservaron para hacer las predicciones y se usaron tres

diferentes tamaños de muestra $n = 192, 128$ y 64 . Se estimaron los parámetros usando REML y MV y tanto la matrix \mathbf{W} como \mathbf{X} . Los resultado por ambos métodos son idénticos para todos los tamaños muestrales, como se esperaba de (19).

Los resultados interesantes se obtienen al obtener estimadores para $Var(\hat{\beta})$ y $Var(\hat{y}_0 - y_0)$. Por ejemplo, al considerar los elementos a_i de la diagonal de la matriz $Var(\hat{y}_0 - y_0)$. Si se calculan las verdaderas varianzas de los errores de predicción $a_i = \sigma_{p,i}^2, i = 1, 2, \dots, 64$ y se toman las diferencias con respecto al promedio de la varianza del error de predicción ($\bar{\sigma}_{\hat{p},i}^2 = \frac{1}{500} \sum_{j=1}^{500} \hat{a}_{ij}, i = 1, 2, \dots, 64$) sobre las 500 réplicas para los 64 sitios y para cada uno de los estimadores $\hat{\theta}_{MV}$, $\hat{\theta}_{MVi}$ y $\hat{\theta}_{REML}$, se observa que los estimadores usando REML son mejores en el sentido que son cercanos a cero y tienen menor varianza, seguido de MV. El sesgo en MV usando variables indicadoras, hace que las estimaciones para los errores de predicción, sean muy malas.



5. Conclusiones

Como se ha mostrado, el uso de variables indicadoras para datos espaciales con matriz de covarianza parametrizada y desconocida puede implementarse fácilmente usando cualquier software estándar y permite obtener estimadores tanto de la deriva como de las predicciones en cualquier número de sitios. Es importante comentar, que estos resultados no se restringen a procesos con error estacionario o solo a datos espaciales.

Referencias

Cressie, N. **1991**. *Statistics for Spatial Data*, Wiley, New York.

Fuller, W. A.**1980**. The use of indicator variables in computing predictions. *J. Econometrics*, *2*, 231-243.

Hughes Oliver, J. and González- Fariás G.**1999**. Parametric Covariance models for shock-induced stochastic processes. *Journal of Statistical Planning and Inference*, 51-72.

Journel, A.G. and Huijbregts, C.J. **1978**.*Mining Geostatistics*, Academic Press, London .

Zimmerman, D.L. **1989**. Computationally efficient restricted maximum likelihood estimation of generalized covariance functions. *Mathematical Geology*, *21*, 655-672.

Zimmerman, D.L. and Zimmerman, M.D.**1991**. A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. *Technometrics*, *33*, 77-81.

Inferencia sobre Valores Récords

Rafael Perera Salazar

Instituto Tecnológico Autónomo de México

Mario Cortina Borja

University College London

1. Introducción

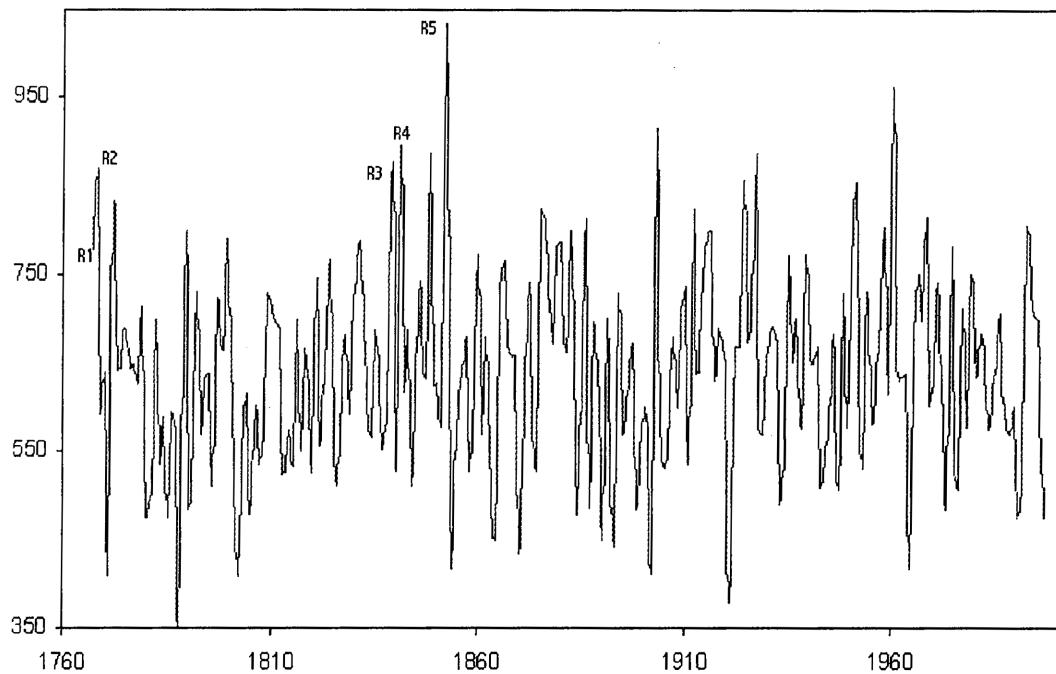
Hace 50 años K.N. Chandler estableció las bases para el estudio de los valores record y sus propiedades. Contrario a lo que podría esperarse, hay pocas referencias en la teoría de records y mucho menos aplicaciones de la misma. Para un documento actual ver Arnold, Balakrishnan y Nagaraja (1998). Una excelente introducción al tema se puede encontrar en Glick (1978).

La teoría estándar empieza asumiendo una sucesión infinita de observaciones independientes e identicamente distribuidas X_1, X_2, \dots de la distribución continua F_X . Una observación X_j es un *record superior* si su valor excede al de todas las observaciones previas, por ejemplo, si $X_j > X_i$ para toda $i < j$; los *records inferiores* se definen de forma análoga. Asumiendo que X_j es observado al tiempo j , entonces la *sucesión de records en el tiempo* $T_n|n \geq 0$ se define como $T_0 = 1$ y para $n \geq 1$, $T_n = \min\{j|X_j > X_{T_{n-1}}\}$. La sucesión de valor de records R_n , se define como $R_n = X_{T_n}$ para $n = 0, 1, 2, \dots$. Nótese que n es el número observado de records, no el número de observaciones de F_X disponibles; a éste último lo denotaremos como n_0 .

La hipótesis de que las observaciones son i.i.d. no es cierta en muchas situaciones prácticas; aún así, tan pronto como se trata de imprimir realismo al modelo, la teoría se complica

rápidamente. Otro problema es que los records son intrínsecamente raros. Por ejemplo en el caso i.i.d., es fácil mostrar que para $n_0 = 1,000,000$ el valor esperado y la varianza del número de records son 14,39 y 12,75, respectivamente.

La figura 1 muestra la lluvia anual en Oxford entre 1767 y 1996. Hay cinco records en 230 observaciones la presente siendo 1034mm observados en 1852.



Figural: Total de lluvia anual en Oxford: 1767–1996

Una pregunta concierne a cuándo esperaríamos que ocurriera el siguiente record. El valor esperado de la distribución inter-records diverge cuando $n \rightarrow \infty$, aunque es posible mostrar que la mediana de la distribución de tiempos de espera para el próximo record en este ejemplo es aproximadamente 183 años.

2. Estimación y Predicción

Sea R_0, R_1, \dots, R_n los records observados para cualquier distribución con parámetros de localización μ y de escala σ con distribución acumulada $F(\frac{x-\mu}{\sigma})$ y función de densidad $f(x; \mu, \sigma) = \frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$. La función de log-verosimilitud puede ser escrita como sigue:

$$l(\mu, \sigma) = -(n+1)\log(\sigma) - \sum_{i=1}^{n-1} \log\{1 - F(r_i)\} + \sum_{i=0}^n \log f(r_i)$$

Hay fórmulas disponibles para los mejores estimadores lineales insesgados basados en los records observados para esta familia de distribuciones. En algunos casos, notablemente para las distribuciones Uniforme Continua y de Valor Extremo, hay cantidades pivotales que nos dan resultados exactos para intervalos de confianza condicional para cada parámetro. Los detalles aparecen en Arnold, Balakrishnan y Nagaraja (1998).

Estimaciones basadas en valores records no son muy eficientes como resultado de lo pequeño que es n . Aún así éstos son útiles para realizar predicciones puntuales y obtener intervalos de records futuros. Como ejemplo, considere los records de la información de lluvia en Oxford. Los valores record son 777,1, 861,3, 877,7, 895,9, 1034,7 El mejor predictor lineal insesgado (BLUP) para el siguiente record es 1071,04, y la predicción del intervalo de confianza condicional de los records pasados es [1035,5, 1199,9].

3. Records con tendencia lineal

Si se considera el siguiente modelo

$$X_{n_0} = cn_0 + Y_{n_0} \tag{21}$$

donde $n \in \mathbb{Z}, c > 0$ y Y_n es una sucesión de variables independientes e idénticamente distribuidas. El porcentaje de records en una muestra se define como $P(n_0) = \frac{1}{n_0} \sum_{j=1}^{n_0} I_{A_j}$ donde A_j es el evento de que el record a ocurra en el tiempo j e I es una función indicadora. La variable P es un indicador de la tendencia c ya que

$$\frac{\sum_{j=1}^{n_0} I_{A_j} - \log n_0}{\sqrt{\log n_0}}$$

converge en distribución a una $N(0, 1)$, entonces si $c = 0$ tenemos que

$$\lim_{n_0 \rightarrow \infty} P(n_0) = 0.$$

Asumiendo el modelo (1) tenemos que $\lim_{n_0 \rightarrow \infty} P(n_0) = p \in (0, 1)$ y por lo tanto $\sqrt{n_0}(P(n_0) - p)$ converge en distribución a $N(0, \sigma^2)$, donde $\sigma^2 = p(1 - p)$. Este resultado implica que si el intervalo de confianza para p contiene al cero, entonces no se puede rechazar que no hay tendencia. La cantidad p es la *tasa asintótica del record* y es estimada por el número observado de records divididos entre n_0 .

Diversos estimadores de σ^2 han sido propuestos (ver Ballerini y Resnick (1987)). Todos éstos están basados en un estimador consistente $\bar{\gamma}_{n_0}$ de la función de autocorrelación para I_{A_j} . Para el presente trabajo se utilizó:

$$\bar{\sigma}^2 = \bar{\gamma}_{n_0}(0) + 2 \sum_{k=1}^{m_1} \left(1 - \frac{k}{n_0}\right) \bar{\gamma}_{n_0}(k)$$

donde $m_1 = \lfloor n_0^{0.499} \rfloor$. Si el intervalo de confianza para p contiene a cero, entonces la tendencia será estadísticamente significativa.

Como un ejemplo analizamos el máximo y mínimo de temperaturas diarias en Oxford entre 1961 y 1994. Consideramos 12 series de medias mensuales para estos 34 años. La tabla 1 muestra los valores de \hat{p} ; el asterisco indica que es significativa al 5 %.

	Jan	Feb	Mar	Apr
MaxTemp	0,15*	0,09	0,03	0,12*
MinTemp	0,15*	0,03	0,09*	0,03
	May	Jun	Jul	Aug
MaxTemp	0,15*	0,12*	0,21*	0,12*
MinTemp	0,12	0,15	0,15*	0,12*
	Sep	Oct	Nov	Dec
MaxTemp	0,06	0,12	0,12*	0,18*
MinTemp	0,06	0,12	0,09	0,24*

Tabla 1

Encontramos que ocho y seis meses tenían una tendencia significativa para el máximo y mínimo respectivamente. Esto podría ser tomado como una leve indicación de un calentamiento secular en Oxford, considerando un número pequeño de observaciones disponibles.

4. Pronosticando Records

Ahora buscamos un modelo de predicción para records utilizando Mínimos Cuadrados Generalizados. Considerando las más bajas (los más rápidos) temperaturas (tiempos) en un mes (año de competencia). Definiendo esta medida como X_t con $t = 1, 2, \dots, T$. Como se mencionó anteriormente ésta es una sucesión de variables aleatorias i.i.d. con distribución común F_X y densidad f_X . Se está interesado sólo en los records $R : n$ definidos como $R_n = \min\{X_1, X_2, \dots, X_n\}$.

Sea $X_t = \mu + \sigma_X Z_t^*$, donde Z_t^* es una variable aleatoria estandarizada con propiedades conocidas, entonces las R_t s pueden ser vistas como funciones lineales de la primera estadística de orden de la Z_t^* s, con Z_t conocido, de tal manera que el valor esperado del mínimo de Z en

una muestra aleatoria de tamaño t puede ser utilizado como una variable independiente en un modelo de regresión con parámetros (μ, σ_X) . Los estimados de estos parámetros se obtienen utilizando mínimos cuadrados generalizados con pesos dados por la matriz de varianzas y covarianzas V de los valores esperados de las estadísticas de orden Z_t ; ver detalles en Tryfos y Blackmore (1985). Utilizando estas estimaciones podemos obtener mejores predictores linealmente insesgados (BLUPs) al igual que intervalos de predicción para records futuros.

Año	400 mt	800 mt
2000	43,176	101,09
2001	43,172	101,07
2002	43,168	101,051
2003	43,164	101,032
2004	43,16	101,013
2005	43,156	100,995
2006	43,152	100,978
2007	43,148	100,961
2008	43,145	100,944
2009	43,141	100,927
2010	43,137	100,911
2011	43,133	100,896
2012	43,129	100,88
2013	43,125	100,865
2014	43,122	100,85

Tabla 2

Ahora presentamos dos ejemplos de esta aproximación. Buscamos los tiempos más rápidos para las competencias de 400 metros y 800 metros. Estos valores pueden ser considerados como las realizaciones de una distribución de valores extremos. En el caso de los 400 metros se tenía la información de 1900 a 1999. Había en total 14 records más bajos siendo éstos

43,18 segundos. Para los 800 metros se tenían datos de 1966 a 1999 y el record más bajo era 101,11 segundos. En la Tabla 2 se muestra la predicción de tiempos para los siguientes 15 años, del 2000 al 2014.

Referencias

Arnold, B. C., N. Balakrishnan, y H. N. Nagaraja, *Records*. Ed. Wiley, New York, 1998.

Ballerini, R., y S. L. Resnick, Records in the presence of a linear trend, *Advances in Applied Probability*, 19: 801–828, 1987.

Chandler, K. N., The distribution and frequency of record values, *Journal of the Royal Statistical Society, series B*, 14: 220–228, 1952.

Glick, N., Breaking records and breaking boards. *American Mathematical Monthly*, 85: 2–26, 1978.

Tryfos, P. y R. Blackmore, Forecasting Records, *Journal of the American Statistical Association*, 80: 46 – 50, 1985.

Estimación de los Parámetros de Correlación en Modelos para Datos Longitudinales

Silvia Ruiz-Velasco Acosta

IIMAS, UNAM

1. Introducción

Los modelos lineales generalizados han sido ampliamente tratados en la literatura (Nelder & Wedderburn (1972)), constan de tres componentes: un componente aleatorio miembro de la familia exponencial:

$$f(y; \theta, \phi) = \exp \{ [y\theta - b(\theta)] / a(\phi) + c(y, \theta) \},$$

un componente sistemático o predictor lineal

$$\eta = \sum_{j=1}^p \beta_j x_j$$

y una función liga que relaciona los componentes sistemático y aleatorio

$$\eta = g(\mu),$$

donde $\mu = E(y)$ y g es una función monótona y diferenciable.

En particular este modelo cumple que $E(y) = \mu = b'(\theta)$ y $var(y) = b''(\theta) a(\phi)$. Este tipo de modelos incluye entre otros al de regresión, modelos loglineales, modelos logístico.

Una forma de ajustar los modelos lineales generalizados es utilizando la función de cuasiverosimilitud. Si y es un vector n dimensional, tal que $E(y) = \mu$ y $cov(y) = \phi V(\mu)$. La función de cuasiverosimilitud esta definida como

$$D^T V^{-1} (y - \mu) = 0$$

donde D es una matriz de $n \times p$ cuyos elementos son

$$\partial \mu_i / \partial \beta_{if}(y_{ij}; \theta, \phi) = \exp\{[y_{ij}\theta_{ij} - b(\theta_{ij})]/a(\phi) + c(y_{ij}, \theta_{ij})\}$$

2. Datos longitudinales

En el contexto de datos longitudinales, una generalización de los modelos lineales generalizados, esta dada por el siguiente planteamiento. Si

$$Y_i = (y_{i1}, \dots, y_{in})$$

es el vector respuesta para el individuo i , con

$$f(y_{ij}; \theta, \phi) = \exp\{[y_{ij}\theta_{ij} - b(\theta_{ij})]/a(\phi) + c(y_{ij}, \theta_{ij})\}$$

Donde $E(Y_{ij}) = \mu_{ij}$ y $cov(Y_i) = \sum_i = \phi V(\mu)$, asociado a cada Y_{ij} existe x_{ij} vector de covariables, además, suponemos que $\mu_{ij} = h(x'_{ij}\beta)$. McCullagh (1983) propuso estimar β utilizando la función de cuasiverosimilitud, definida como:

$$\sum_{i=1}^K D_i^T V_i^{-1} (y_i - \mu_i) = 0$$

En 1986, Liang & Zeger propusieron la estimación de β por medio de las llamadas Generalized Estimating Equations, el método consiste en expresar la matriz de varianza como

$$\sum = A_i^{-1/2} R(\alpha) A_i^{-1/2},$$

donde A_i es una matriz diagonal con elementos m'_{ij} , $R(\alpha)$ es una matriz de correlación parametrizada por el vector α .

Entre las estructuras mas comunes para la matriz de correlación están:

$R(\alpha) = I$, independencia,

$corr(y_{it}, y_{il}) = \alpha$, intercambiable,

$corr(y_{it}, y_{il}) = \alpha^{|t-l|}$, autorregresiva,

$corr(y_{it}, y_{it+k}) = \alpha_k \quad k = 1, \dots, \mu$, m-dependiente.

Liang y Zeger proponen una estimación iterativa en dos pasos, es decir estimar β dado el valor de α , actualizar el valor de α dado β , hasta convergencia. En el artículo original probaron que si el estimador de α es consistente, el estimador de β es insesgado y eficiente y, además, asintóticamente normal.

En 1996 Crowder probó que los estimadores de momentos para α , propuestos por Liang & Zeger no son siempre consistentes, por lo tanto las propiedades del estimador de β no son necesariamente ciertas.

El requisito fundamental es que el estimador de α exista y converga a algún valor, cuando el tamaño de muestra tiende a infinito. Si además la estructura de correlación no es muy diferente de la estructura real, el estimador de β será razonablemente eficiente.

En una serie de artículos Chaganty & Shults propusieron estimar por medio de lo que llamaron “Cuasi mínimos cuadrados”

$$Q(\alpha, \beta) = \sum_{i=1}^K (y_i - \mu_i)^T A_i^{-1/2} R_i^{-1}(\alpha) A_i^{-1/2} (y_i - \mu_i)$$

Dando como resultado la ecuación de cuasiverosimilitud para estimar β y

$$\sum_{i=1}^K (y_i - \mu_i)^T A_i^{-1/2} \frac{dR_i^{-1}(\alpha)}{d\alpha} A_i^{-1/2} (y_i - \mu_i) = 0$$

para estimar α . El estimador de α es consistente aunque sesgado. El estimador de β es insesgado y asintóticamente normal.

Nuestra propuesta es encontrar el estimador de α que minimiza la varianza de β . La varianza de β puede ser estimada por:

$$\text{var}(\hat{\beta}) = K \left(\sum D_i^T V_i^{-1} D_i \right)^{-1} \left\{ \sum D_i^T V_i^{-1} \text{cov}(Y_i) V_i^{-1} D_i \right\} \left(\sum D_i^T V_i^{-1} D_i \right)^{-1}$$

En particular se propone minimizar la traza de $\text{var}(\beta)$ o minimizar $|\text{var}(\beta)|$.

Basado en la teoría de consistencia de Ecuaciones de Estimación de Crowder (1996), es posible demostrar que el estimador de α es consistente y en particular, en el caso de que la estructura de correlación es la correcta, insesgado.

Para comparar nuestros resultados con los de Chaganty & Shults realizamos una serie de simulaciones. El modelo simulado es

$$y_i = 120 - 12,88x_i$$

Las estructuras de correlación simuladas fueron intercambiable, uno dependiente y autorregresivo de orden uno. Simulamos tres cinco y ocho repeticiones por individuo, y para el parámetro de correlación utilizamos los valores de 0.1 y 0.5. Además simulamos para dos tamaños de muestra 15 y 100.

3. Conclusiones

Cuando el número de repeticiones es 3 y el valor de $\alpha = 0,1$, no existe una diferencia entre los métodos de ajuste propuestos y los resultados obtenidos por GEE, además la eficiencia de β ante una estructura errónea es alta.

Cuando el número de repeticiones crece, y el número de individuos es alto, esto sigue siendo válido, sin embargo, si el número de individuos no es alto, se empieza a ver una subestimación de la varianza de β con GEE, además de una pérdida de eficiencia con respecto a la estructura correcta

Con $\alpha = 0,5$, cuando el tamaño de muestra es 15 y el número de repeticiones cinco, existen diferencias en los ajustes con los métodos propuestos y GEE. En particular, con los métodos propuesto existe mayor eficiencia en la estimación de β con la estructura correcta. Asimismo la varianza del estimador de β es menor con el método propuesto que con GEE o los métodos de Chaganty.

Cuando el número de repeticiones crece a ocho, este comportamiento continúa, teniendo incluso problemas de convergencia en el ajuste de modelo dependiente cuando este no es el correcto.

Para tamaños de muestra 100, con cinco y ocho repeticiones los comportamiento son similares.

Referencias

Chaganty N.R. (1997) An alternative approach to the analysis of longitudinal data via generalized estimating equations. *Journal of Statistical Planning and Inference* 63.

Chaganty N.R., Shults J. (1999) On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter. *Journal of Statistical Planning and Inference* 65.

Crowder M. (1995) On the use of a working correlation matrix in using generalised linear models for repeated measure. *Biometrika* 82.

Crowder M. (2001) On repeated measure analysis with misspecified covariance structure. To appear *JRSSB*.

Liang K.Y., Zeger S.L. (1986) Longitudinal Data Analysis using generalized linear models. *Biometrika* 73.

Shults J. Chaganty N.R. (1998) Analysis of Serially Correlated data using Quasi-Least Squares. *Biometrics* 54.

Modelos Threshold Autorregresivos y Métodos MCMC

María Gpe. Russell Noriega

Graciela González Farías

Gabriel Huerta Gómez

Centro de Investigación en Matemáticas, A.C.

1. Introducción

Durante las últimas tres décadas se han propuesto un gran número de modelos no lineales para el análisis de las series de tiempo, esencialmente para series económicas. Los modelos más populares en la literatura, son los llamados modelos threshold autorregresivos (TAR) y modelos bilineales. Los modelos TAR fueron introducidos por Tong (1978) y Tong y Lim (1980) como modelos alternativos para describir series de tiempos periódicas. La principal característica de estos modelos es que estos incluyen ciclos límites, frecuencias dependientes de amplitudes y fenómenos de saltos, los cuales en general no pueden ser capturados por modelos lineales. Algunas dificultades que presentan estos modelos es que no son fácilmente desarrollados en la práctica, ya que la variable threshold es difícil de identificar, así como la estimación del valor de threshold y su distribución de muestreo asociada. La gran mayoría de los procedimientos propuestos para dicha estimación resultan muy complicados y muy costosos computacionalmente hablando.

Una serie de tiempo Y_t se dice sigue un modelo TAR de primer orden con r regímenes y variable threshold Z_{t-d} si satisface la siguiente ecuación:

$$Y_t = \alpha_k Y_{t-1} + e_t^{(k)} \text{ para } \gamma_{k-1} \leq Z_{t-d} < \gamma_k \quad (1)$$

donde $\{e_t^{(k)}\}$, $k = 1, 2, \dots, r$ es una secuencia de variables aleatorias independientes e idénticamente distribuidas, normal con media cero y varianza σ_k^2 , con $\{e_t^{(i)}\}$ y $\{e_t^{(j)}\}$ independientes si $i \neq j$. Los valores $\gamma_1 < \gamma_2 < \dots < \gamma_r$ son los valores de los threshold, y α_i son los coeficientes autorregresivos en el régimen i . Z_{t-d} es la variable threshold y d es un entero positivo fijo, usualmente identificado como el parámetro de “delay” de la variable Z_t y en nuestro caso, conocido. Cuando la variable threshold Z_t es una función de los rezagos de la variable Y_t el modelo se dice TAR en si mismo y se denota por SETAR.

En el ámbito económico se sabe que muchas de las principales series de tiempo macroeconómicas están mejor representadas por modelos con raíces unitarias. Sin embargo, en teoría, algunas de las series de tiempo económicas como por ejemplo, las variables medidas en tasas no pueden tener todas las características de un proceso de raíz unitaria, aún cuando las pruebas estándar de raíz unitaria no rechacen la hipótesis nula. González y Gonzalo (1997) introducen una clase de modelos threshold capaces de replicar el comportamiento de variables económicas tales como desempleo, inflación y tasas de interés. Dependiendo de los valores de la variable threshold estos modelos pueden tener o bien una raíz unitaria o una raíz estable y a pesar de la raíz unitaria prueban que dichos modelos son estacionarios y geométricamente ergódicos, y se denotan como modelos TUR. González y Gonzalo (1997) presentan una aplicación a datos reales modelando la relación existente entre tasas de interés e inflación a través de un modelo TUR.

En este trabajo estudiamos modelos threshold autorregresivo de dos regímenes desde una perspectiva Bayesiana, por medio del uso de procedimientos MCMC para estimar el parámetro threshold, así como los coeficientes autorregresivos y varianzas asociadas al modelo. Consideramos distintos comportamientos para los coeficientes autorregresivos por medio de una distribución mezcla la cual engloba los casos de Modelos AR estacionarios, UR, Modelos TAR y Modelos threshold autorregresivos de raíz unitaria (TUR). Utilizamos el conjunto de datos analizados por González y Gonzalo (1997) para establecer una comparación entre los

resultados obtenidos a partir del modelo TUR propuesto por dichos autores y los resultados obtenidos del modelo mezcla.

2. Modelo y Estimación

Consideramos un caso particular del modelo (1), donde el número de regímenes es dos, i.e. un TAR(2;1,1) dado por:

$$Y_t = \begin{cases} \alpha_1 Y_{t-1} + e_t^{(1)}, & Z_{t-d} \leq \gamma \\ \alpha_2 Y_{t-1} + e_t^{(2)}, & Z_{t-d} > \gamma \end{cases} \quad (2)$$

Sea $\mathbf{Y} = \{y_1, y_2, \dots, y_n\}$ el conjunto de las observaciones de interés. Asuma la primer observación como fija y sea τ_i el i -ésimo índice en el tiempo más pequeño del conjunto de observaciones $\{Z_{p+1-d}, \dots, Z_{n-d}\}$. Condicionando en la primer observación, podemos escribir la función de verosimilitud como:

$$L(\alpha_1, \alpha_2, \sigma_1^2, \sigma_2^2, \gamma | \mathbf{Y}_{-1}) \propto \sigma_1^{-s} \sigma_2^{-(n-p-s)} \exp \left\{ -\frac{1}{2\sigma_1^2} \sum_{i=1}^s (Y_{\tau_i+d} - \alpha_1 Y_{\tau_i+d-1})^2 - \frac{1}{2\sigma_2^2} \sum_{t=s+1}^{n-p} (Y_{\tau_i+d} - \alpha_2 Y_{\tau_i+d-1})^2 \right\}, \quad (3)$$

donde s satisface $Z_{\tau_s} \leq \gamma < Z_{\tau_{s+1}}$. Los parámetros del modelo (2) a estimar son α_1 , α_2 , σ_1^2 , σ_2^2 y γ .

Especificaciones Iniciales

Para implementar el análisis propuesto, es necesario derivar las distribuciones condicionales finales de cada uno de los parámetros dados los otros. Para este fin, elegimos las siguientes distribuciones a priori.

Sean α_1 y α_2 , con distribución inicial uniforme en $(-1, 1)$. Las especificaciones para los parámetros α_1 y α_2 , son:

- **Caso 1:** Modelo AR(1) $\equiv \pi_1, \alpha_1 = \alpha_2, |\alpha_1| < 1, |\alpha_2| < 1$

- **Caso 2:** Modelo UR $\equiv \pi_2, \alpha_1 = \alpha_2, \alpha_2 = 1$

- **Caso 3** Modelo TAR(1) $\equiv \pi_3, \alpha_1 \neq \alpha_2, |\alpha_1| < 1, |\alpha_2| < 1$

- **Caso 4** Modelo TUR(1) $\equiv \pi_4, \alpha_1 \neq \alpha_2, |\alpha_1| < 1, \alpha_2 = 1,$

La distribución inicial para (α_1, α_2) es la mezcla: $p(\alpha_1, \alpha_2) = \pi_1 I_{\{\alpha_2\}}(\alpha_1) U_{(-1,1)}(\alpha_2) + \pi_2 I_{\{1\}}(\alpha_2) I_{\{1\}}(\alpha_1) + \pi_3 U_{(-1,1)}(\alpha_1) U_{(-1,1)}(\alpha_2) + \pi_4 I_{\{1\}}(\alpha_2) U_{(-1,1)}(\alpha_1).$

La distribución inicial para σ_1^2 , y σ_2^2 independientes es $IG(v_i/2, v_i \lambda_i/2)$, $i = 1, 2$, con v_i, λ_i conocidos. Para γ se supone una distribución $U(q_1, q_2)$, con q_1 , y q_2 fijos.

Distribuciones Posterioras

1) La distribución condicional final para α_1 y α_2 esta formada de cuatro componentes dependiendo de cada uno de los elementos de la distribución inicial $p(\alpha_1, \alpha_2)$, y dada por:

$$\begin{aligned} f(\alpha_1, \alpha_2 | \mathbf{Y}, \sigma_1^2, \sigma_2^2, \gamma) = & K^{-1}((\pi_1/2)I_{\{\alpha_2\}}(\alpha_1) N_t(m, M) + (\pi_3/4) \times N_t(a/c, \sigma_1^2/c) \times N_t(e/d, \sigma_2^2/d) \\ & + \pi_2 \times I_{\{1\}}(\alpha_2) \exp\{-e/2\sigma_2^2(1-b/e)^2\} \times I_{\{1\}}(\alpha_1) \exp\{-c/2\sigma_1^2(1-a/c)^2\}) \\ & + (\pi_4/2) \times I_{\{1\}}(\alpha_2) \exp\{-d/2\sigma_2^2(1-b/e)^2\} \times N_t(a/c, \sigma_1^2/c)), \end{aligned}$$

donde $m = (\sigma_2^2 a + \sigma_1^2 b) / (\sigma_2^2 c + \sigma_1^2 e)$ y $M = \sigma_1^2 \sigma_2^2 / (\sigma_2^2 c + \sigma_1^2 e)$ y $N_t(m, M)$ denota la distribución normal truncada y K la constante de marginalización. Las cantidades a, b, c y e están dadas por: $a = \sum_{i=1}^s y_{\tau_i+d} y_{\tau_i+d-1}$, $b = \sum_{i=s+1}^{n-p} y_{\tau_i+d} y_{\tau_i+d-1}$, $c = \sum_{i=1}^s y_{\tau_i+d-1}^2$, y $e = \sum_{i=s+1}^{n-p} y_{\tau_i+d-1}^2$.

2) La distribución condicional final de σ_i^2 independiente de σ_j^2 , para $i \neq j$, es $IG(v_i + n_i/2, (n_i s_i^2 + v_i \lambda_i)/2)$ donde $n_1 = \sum_{i=1}^{n-p} I_{\{Z_{\tau_i} \leq \gamma\}}$, $n_2 = \sum_{i=1}^{n-p} I_{\{Z_{\tau_i} > \gamma\}}$, $s_1^2 = n_1^{-1} \sum_{i=1}^s (y_{\tau_i+d} - \alpha_1 y_{\tau_i+d-1})^2$ y $s_2^2 = n_2^{-1} \sum_{i=s+1}^{n-p} (y_{\tau_i+d} - \alpha_2 y_{\tau_i+d-1})^2$.

3) La función de probabilidad final de γ es proporcional a la verosimilitud, ec. (3).

Todas las densidades condicionales tienen forma cerrada y son fácilmente identificables a excepción de la densidad condicional para γ . Para considerar γ , desarrollaremos un algoritmo de Metrópolis cuya idea es similar a Chen (1998).

3. Una aplicación a Tasas de Interés

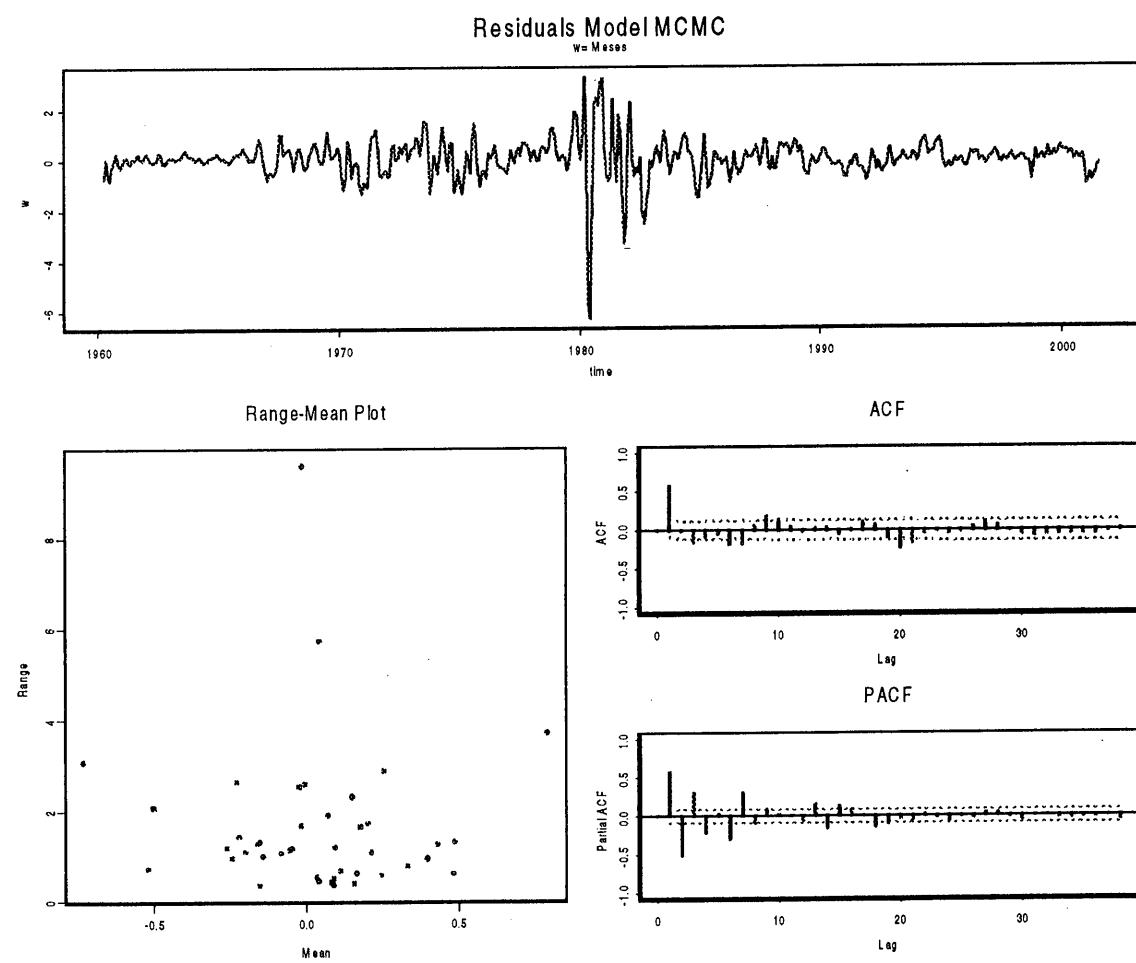
Se realizó un amplio ejercicio de simulación con el interés de estudiar el poder del método para identificar los verdaderos valores de los parámetros, así como la correcta identificación del modelo generador de los datos analizados. En general, independientemente del modelo generador de los datos, se obtuvo que el modelo mezcla es capaz de identificar adecuadamente el modelo que originó los datos, además de estimaciones muy cercanas de los verdaderos valores de los parámetros en el modelo. Dicho análisis se realizó para diferentes tamaños de muestra y los resultados fueron similares.

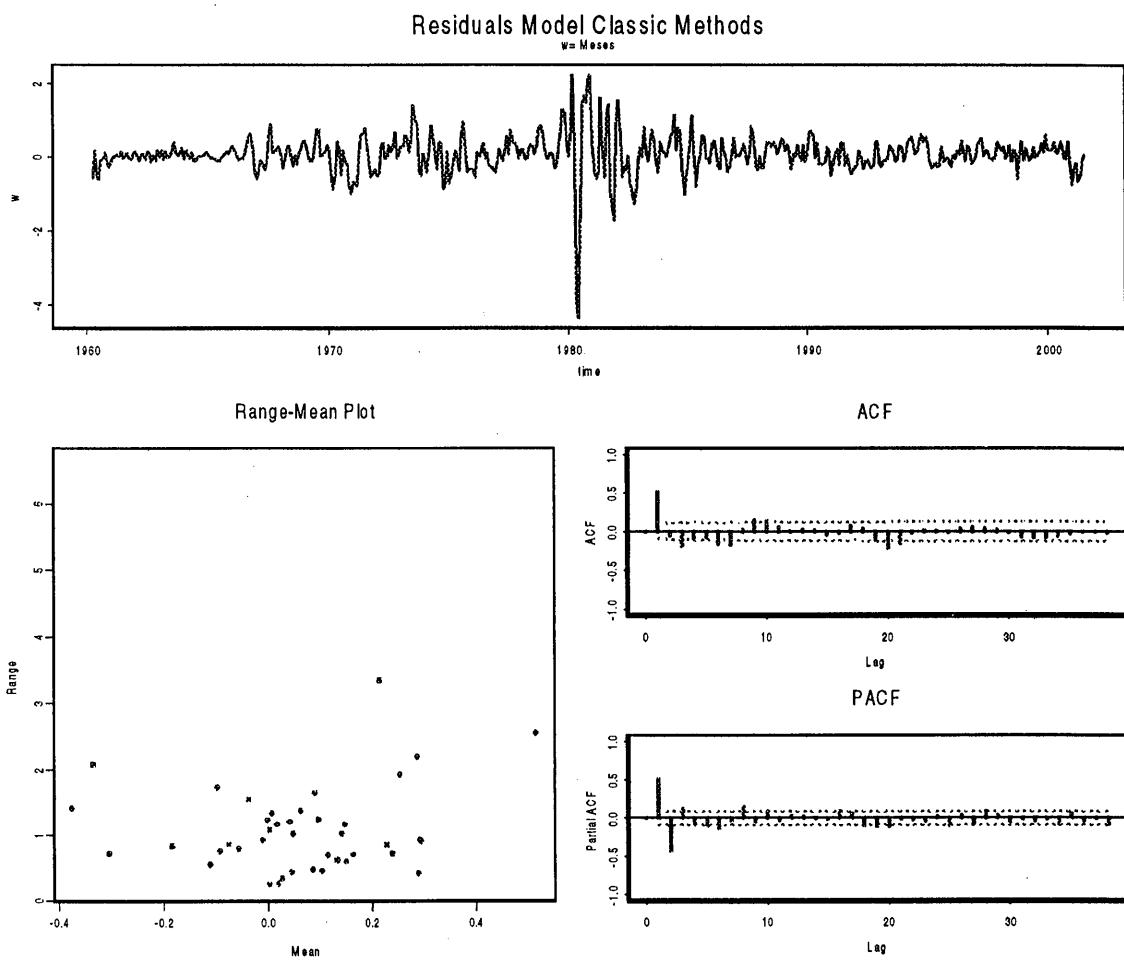
La aplicación a datos reales consiste en analizar la relación entre tasas de interés como la variable a modelar y rezagos de los cambios de la variable inflación como la variable threshold (datos mensuales de la Federal Reserve Board and BLS:PPI, enero 1960-julio 2000). Presentamos tanto los resultados obtenidos por González y Gonzalo (1997), así como los resultados del modelo mezcla.

El modelo mezcla estimado resultante del MCMC consta esencialmente de la componente AR con coeficiente autorregresivo cercano a la unidad y la componente TUR y dado por:
$$Y_t = 0,986 \times I(Z_{t-1} \leq 0,0095) + 0,996 \times I(Z_{t-1} > 0,0095).$$

El modelo TUR estimado propuesto por González y Gonzalo (1997): $Y_t = 0,081 + 0,926 \times I(Z_{t-1} \leq -0,003) + 0,992 \times I(Z_{t-1} > -0,003) + 0,37\Delta X_{t-1} - 0,191\Delta X_{t-2}$.

Es importante mencionar que las desviaciones de los residuales en el régimen uno es 0.6 y en el régimen dos 0.82. El modelo mezcla considera varianzas distintas en cada régimen y se obtienen resultados similares con sólo usar un AR(1) para cada régimen sin otros términos dinámicos, ver gráficas de análisis de residuales para ambos modelos al final del documento. Pensamos además que agregar una constante en el modelo ayudará a tener una mejor aproximación entre los modelos. En trabajo futuro se presentará este ajuste así como una comparación de las capacidad predictivas de los modelos de mezclas.





Referencias

Chen, Cathy W.S. (1998). A Bayesian analysis of generalized threshold autoregressive models. *Statistics & Probability Letters*. **40**, 15-22.

González, Martín and Gonzalo, Jesús (1997). Threshold Unit Root Models. *Working Paper, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, Spain*.

Huerta, Gabriel and West Mike (1997). Priors and component structures in autoregressive time series models. *J.R. Statist. Soc. Ser. B.* **61**, parte 4 881-889.

Tong, H. (1978) On a Threshold Model in Pattern Recognition and Signal Processing (ed. C.H.Chen). Amsterdam: Sijhoff & Noordhoff.

Tong, H., Lim, K.S., (1980). Threshold autoregression, limit cycles and cyclical data (with discussion). *J. Roy. Statist. Soc. Ser. B.* **42**, 245-292.

Aplicación de la Búsqueda Tabú en Regresión No Lineal

Javier Trejos

Universidad Autónoma Metropolitana, Iztapalapa y

Universidad de Costa Rica

Sergio de los Cobos

Universidad Autónoma Metropolitana, Iztapalapa

1. Introducción

En regresión no lineal se encuentra el problema, frecuente en distintos métodos estadísticos, de la obtención de mínimos locales (Draper y Smith, 1968), Tomassone *et al.*, 1992). En efecto, los métodos clásicos de regresión no lineal están basados en la búsqueda de un óptimo local, por lo que es razonable pensar en la implementación de técnicas de optimización global. Entre estas técnicas, podemos citar el recocido simulado, la búsqueda tabú y los algoritmos genéticos (Reeves, 1993), las cuales han sido ampliamente usadas en distintos problemas estadísticos, de investigación operacional o de ingeniería. En particular, en Análisis Estadístico de Datos han sido usados en particionamiento para el análisis de conglomerados con resultados muy superiores a los de métodos clásicos, como el de las k -medias y el de Ward, en clasificación bimodal, en escalamiento multidimensional, en rotaciones varimax oblicuas y en la obtención de conjuntos aproximados (conocidos como *rough sets* en inglés). Todas estas implementaciones de las técnicas de optimización mencionadas han mejorado o al menos igualado los resultados conocidos. En regresión no lineal ya se ha aplicado la técnica de recocido simulado (ver Trejos y Villalobos, 1999 y 2000) con resultados satisfactorios.

En el presente trabajo abordamos el uso de la técnica de búsqueda tabú en regresión no lineal. Conviene señalar que nos restringimos al caso de la regresión con una variable explicativa, pudiendo generalizarse fácilmente el trabajo al caso de varias variables explicativas; además, no abordamos los problemas relativos a la estimación estadística, como la obtención de intervalos de confianza para los parámetros.

2. Regresión no lineal

Dadas dos variables cuantitativas \mathbf{x} y \mathbf{y} observadas sobre n objetos, donde \mathbf{x} es una variable explicativa y \mathbf{y} es una variable a explicar que depende de \mathbf{x} , se quiere describir la relación de dependencia de \mathbf{y} respecto a \mathbf{x} mediante una función f ; es decir, se quiere establecer la relación funcional $\mathbf{y} = f(\mathbf{x}) + \epsilon$, donde ϵ es un término de error³. La función f depende generalmente de ciertos parámetros, cuyo vector denotaremos $\vec{\theta}$, por lo que escribiremos a la función de regresión $f_{\vec{\theta}}$. Se quiere utilizará el criterio de mínimos cuadrados, el cual mide la calidad de la aproximación funcional propuesta:

$$S(\vec{\theta}) = \|\mathbf{y} - f(\mathbf{x})\|^2 = \sum_{i=1}^n [y_i - f_{\vec{\theta}}(x_i)]^2 \quad (22)$$

donde $\mathbf{x} = (x_1, \dots, x_n)^t$ y $\mathbf{y} = (y_1, \dots, y_n)^t$ son los vectores de las observaciones de las variables, y $\|\cdot\|$ es la norma Euclídea usual. Hacemos notar que para el trabajo con las heurísticas de optimización combinatoria (recocido simulado, búsqueda tabú) que usaremos, podemos usar otros criterios, como el de la norma L_1 : $S_1(\vec{\theta}) = \|\mathbf{y} - f(\mathbf{x})\|_1 = \sum_i |y_i - f_{\vec{\theta}}(x_i)|$. Esta extensión no presenta ninguna dificultad ya que nuestros métodos no necesitan ningún tipo de diferenciabilidad del criterio. Notemos también que en el criterio (22) se pueden hacer ponderaciones tanto de los objetos como de las variables. Sin embargo, para no hacer pesada

³En este trabajo no supondremos que ϵ sigue alguna distribución en particular.

la presentación, nos restringiremos aquí al caso en que los objetos tienen el mismo peso y las variables también.

Salvo cuando $f_{\vec{\theta}}$ es una función lineal, no se conoce una solución general a este problema. Debe notarse que en algunos casos la función $f_{\vec{\theta}}$ no es en sí lineal pero el problema de regresión se puede *linealizar*. En este trabajo no nos ocuparemos de los modelos que son linealizables, ya que éstos se resuelven fácilmente mediante la regresión lineal clásica.

El método más ampliamente usado para abordar la regresión no lineal es el de Gauss-Newton, que consiste en usar una aproximación lineal de la función $f_{\vec{\theta}}$ mediante un polinomio de Taylor de primer orden alrededor del punto $\vec{\theta}^0 \in \mathbb{R}$. Luego se realiza una regresión lineal múltiple, y se itera el método hasta que converja (aunque no se puede garantizar la convergencia, ver Draper y Smith, pp.464-465.). Véase que esto requiere que la función en cuestión sea diferenciable. Claramente, la convergencia puede ser hacia un óptimo local de $S(\vec{\theta})$. Por ello se ha pensado en el uso de técnicas de optimización combinatoria que traten de evitar esos mínimos locales del criterio, entre las que está el recocido simulado y la búsqueda tabú.

En Trejos y Villalobos (1999) y (2000) se hace una amplia descripción de la implementación de la técnica de recocido simulado en regresión no lineal; allí se reportan resultados similares a los obtenidos con el método de Gauss-Newton sobre datos reales de crecimiento de cultivos.

Otra técnica ampliamente usada en optimización es la llamada de descenso del gradiente. Se trata de un método iterativo que busca la dirección de máximo descenso en cada punto de la iteración. Debe notarse que si bien teóricamente el método converge, esta convergencia puede ser muy lenta. Marquardt (1963) propuso un método que lleva su nombre y que trata de mejorar los defectos de los métodos de Gauss-Newton y de descenso de gradiente. El método está basado en una interpolación entre las direcciones que escogen esos dos métodos en cada iteración.

3. Regresión No Lineal usando Búsqueda Tabú

La Búsqueda Tabú (BT) es una técnica iterativa de optimización combinatoria basada, al igual que el recocido simulado, en el examen de los vecinos de un estado actual (Glover *et al.*, 1993). Tiene la particularidad de que hace uso extensivo de la memoria, lo cual no tiene el recocido simulado. De esta forma, trata de evitar vecindades donde pueden encontrarse los mínimos locales y también trata de evitar los ciclos.

En la BT, el estado actual del problema de optimización posee una vecindad \mathcal{V}_i el cual puede ser generado en su totalidad en cada iteración, o bien se puede generar únicamente una muestra de vecinos. Se maneja una lista tabú T que contiene un número $|T|$ de vecinos que no pueden ser accesados durante un cierto número de iteraciones. Debe notarse que la longitud de esta lista puede ser dinámica. Se procede a escoger el vecino $j^* \in \mathcal{V}_i$ tal que su valor de la función objetivo es el mejor entre todos los elementos de \mathcal{V}_i (o aquéllos generados, en caso de usar solamente una muestra), excepto si j^* pertenece a T . En la BT, se aplica un criterio llamado de *aspiración* para acceder a un estado j^* si es un estado en la lista tabú pero con el mejor valor de la función objetivo que se haya encontrado durante las iteraciones. Este criterio de aspiración tiene aplicación cuando los nuevos estados son generados mediante un tipo de movimiento y la codificación de los estados en T se hace mediante el movimiento inverso, y no usando explícitamente cada estado. Existen otros tipos de criterios de aspiración, como puede consultarse en Reeves (1993). Se puede codificar de alguna forma la región que se ha explorado, con el fin decidir posteriormente explorar regiones del espacio de estados no exploradas aún. Se itera durante un número m de veces, dado por el usuario. Así, la búsqueda tabú usa principalmente dos parámetros: la longitud $|T|$ de la lista tabú y el número máximo de iteraciones m .

La BT tiene múltiples aplicaciones en investigación operacional. En Análisis Multivariado de Datos, también ha sido usada en el problema de clasificación por particiones para el análisis

de conglomerados Trejos *et al.* (1998), así como en el escalamiento multidimensional Groenen *et al.* (2000), con excelentes resultados.

Al igual que para el uso del recocido simulado, discretizaremos el espacio de parámetros mediante una malla de ancho h_k , siendo h_k más fina conforme se avance en las iteraciones. Así, un estado es un vector $\vec{\theta}$ del espacio de parámetros. Un vecindario se genera por el desplazamiento en la malla hacia alguno de los estados contiguos, mediante la modificación por h_k en alguno de los parámetros del vector $\vec{\theta}$. Así, si el estado $\vec{\theta}$ es un vector p dimensional, entonces tiene $2p$ vecinos. El movimiento consiste entonces en la selección de uno de los parámetros y en la selección de una dirección. De esta forma, si a la l -ésima coordenada de $\vec{\theta}$ se le añade h_k , lo cual podemos denotar como el movimiento $(l, +1)$, entonces codificaremos en la lista tabú el movimiento inverso $(l, -1)$ con el fin de impedir el regreso a la posición anterior de $\vec{\theta}$. Usaremos una lista tabú de tamaño p y emplearemos el criterio de aspiración antes descrito.

4. Resultados Comparativos

Considérese el siguiente ejemplo didáctico (Antoniadis *et al.*, 1992), en el que se quiere ajustar el modelo $\mathbf{y} = \theta_1 e^{-\theta_2 \mathbf{x}}$ para el conjunto de datos siguiente: $(-2,5,1)$, $(-1,1,1)$, $(1,-1,1)$, $(2,0,2)$. La función presenta dos mínimos locales de $S(\vec{\theta})$: $\vec{\theta}^* = (0,669, 0,214)$ con $S(\vec{\theta}^*) = 1,968$, y $\vec{\theta}^1 = (-0,764, -0,0298)$ con $S(\vec{\theta}^1) = 3,436$. El método de búsqueda tabú encontró el óptimo $\vec{\theta}^*$ el 92 % de los casos (en 100 corridas), el de recocido simulado el 98 % de los casos y el de Gauss-Newton el 32 % de los casos. Por falta de espacio no podemos presentar resultados comparativos exhaustivos, pero mencionaremos que en muchos otros conjuntos de datos hemos obtenido resultados similares a los del recocido simulado y del Gauss-Newton. Tal es el caso, por ejemplo, en datos sobre reacciones químicas con

el modelo de Michaelis-Menten (Tomassone *et al.*, 1992) los datos de Puromycin (Bates & Watts, 1988) con modelo de Michaelis-Menten, con los datos (Antoniadis *et al.*, 1992) para el modelo de Micherlich, y para varios de los ejemplos presentados en Draper y Smith (1968), incluyendo modelos con varias variables explicativas.

A manera de conclusión diremos que obtenemos resultados similares a los que da el Método de Gauss-Newton y el recocido simulado, que nuestro método no necesita que la función del modelo sea derivable, y si lo es no se necesita conocer su derivada. Aún deben hacerse comparaciones más amplias, estudiando la sensibilidad a los parámetros de la búsqueda tabú (longitud de la lista tabú, número máximo de iteraciones), trabajo que será emprendido en el futuro cercano.

Referencias

- Antoniadis, A., Berruyer, J. y Carmona, R. (1992). *Régression Non Linéaire et Applications*. Paris: Economica.
- Bates, D.M. y Watts, D.G. (1988). *Nonlinear Regression Analysis and its Applications*. New York: Wiley.
- Draper, N.R. y Smith, H. (1968). *Applied Regression Analysis*. New York: Wiley.
- Glover, F., Taillard, E. y de Werra, D. (1993). A User's Guide to Tabu Search. *Annals of Ops.Res.*, 41.
- Groenen, P.J.F., Mathar, R. y Trejos, J. (2000). Global Optimization Methods for Multidimensional Scaling Applied to Mobile Communications. In *Data Analysis. Scientific Modeling and Practical Application* (eds. W. Gaul *et al.*). pp. 459-469. Berlin: Springer-Verlag.

Marquardt, D.W. (1963). An Algorithm for Least Squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, **2**, 431-441.

Reeves, C. (Ed.) (1995). *Modern Heuristic Techniques for Combinatorial Problems*. McGraw-Hill, London.

Tomassone, R., Audrain, S., Lesquoy, E. y Millier, C. (1992). *La Régression. Des Nouveaux Regards sur une Ancienne Méthode Statistique*. Masson, Paris.

Trejos, J., Murillo, A. y Piza, E. (1998). Global Stochastic Optimization for Partitioning. In *Advances in Data Science and Classification* (eds. A. Rizzi *et al.*), pp. 185-190. Berlin: Springer-Verlag.

Trejos, J. y Villalobos, M. (1999). Optimización Mediante Recocido Simulado en Regresión No Lineal. In *Memorias del XII Foro Nacional de Estadística* (eds. J. Sierra *et al.*), pp. 183-190. Monterrey.

Trejos, J. y Villalobos, M. (2000). Optimización con Sobrecaleamiento Simulado en Regresión No Lineal: algoritmo y software. *Investigación Operacional*, **21**(3), 236-246.

Esta publicación consta de 410 ejemplares y se terminó de imprimir en el mes de septiembre de 2002 en los talleres gráficos del **Instituto Nacional de Estadística, Geografía e Informática**
Av. Héroe de Nacozari Núm. 2301 Sur, Acceso 11, P.B.
Fracc. Jardines del Parque, CP 20270
Aguascalientes, Ags.
México