

# MEMORIA DEL



septiembre 22-26, 1997

iimas, unam  
méxico, d.f.



# MEMORIA DEL



septiembre 22-26, 1997

iimas, unam  
méxico, d.f.



DR © 1998, **Instituto Nacional de Estadística,  
Geografía e Informática**  
Edificio Sede  
Av. Héroe de Nacozari Núm. 2301 Sur  
Fracc. Jardines del Parque, CP 20270  
Aguascalientes, Ags.

<http://www.inegi.gob.mx>  
usuario@cis.inegi.gob.mx

**Memoria del XII Foro Nacional de Estadística  
Septiembre 22-26 de 1997  
IIMAS, UNAM México, D.F.**

Impreso en México

## Presentación

La Asociación Mexicana de Estadística organizó el *12 Foro Nacional de Estadística* en la semana del 22 al 26 de septiembre de 1997. La sede de este evento fue el Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS) de la Universidad Nacional Autónoma de México (UNAM).

Entre otras actividades, se presentaron más de 60 contribuciones libres. Los 31 resúmenes que se recogen en esta Memoria muestran una buena parte de las áreas de interés de nuestra comunidad. Todos los resúmenes que fueron recibidos se incluyeron sin un proceso de arbitraje, pero después de una detallada revisión.

El Comité Editorial agradece a Élida Estrada todo el apoyo que brindó para la recopilación y la transcripción de los trabajos. Asimismo, deseamos reconocer la eficiente labor de formación y edición que realizó Alberto Molina. Sin su valiosa ayuda, esta Memoria no hubiera sido posible. También queremos agradecer al personal de la Subdirección de edición y Diseño del Instituto Nacional de Geografía e Informática la detallada revisión de una primera versión de este manuscrito, que permitió mejorar la calidad de su presentación.

Finalmente, la Asociación Mexicana de Estadística agradece al Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas de la UNAM, a la Dirección General de Estudios de Posgrado de la UNAM, al Instituto Nacional de Estadística, Geografía e Informática y al Colegio de Posgraduados su apoyo en la realización del *12 Foro Nacional de Estadística* que dio origen a esta Memoria.

Raúl Rueda  
Silvia Ruiz-Velasco  
José Villaseñor

Julio 1998



# Contenido

Una Aplicación de Modelado Gráfico en el Colegio de Ciencias y Humanidades de la UNAM <i>Miguel Ángel Abréu Hernández</i> .....	1
Modelos Lineales Generalizados en Actuaría <i>Alejandro Alegria y Evangelina Martínez</i> .....	7
Construcción de un Modelo para un Estudio Multicéntrico en Hospitales de Tercer Nivel <i>Lilia Benavides, Alejandro Aldama y Héctor-Javier Vázquez</i> .....	13
Métodos de Detección de Observaciones Influyentes Multivariadas ante Varias Observaciones Aberrantes <i>Eduardo Castaño Tostado</i> .....	19
Estimación de la Probabilidad de No-Pago de la Cartera Hipotecaria del Sistema Bancario Mexicano <i>Ma. de Lourdes de la Fuente D. y José Manuel Pelayo C.</i> .....	24
Redes Neuronales Probabilísticas: Perspectivas en Clasificación y Reconocimiento de Patrones <i>Sergio de los Cobos, John Goddard, Miguel A. Gutiérrez y Blanca R. Pérez</i> .....	30
Comparación de Métodos para Modelar la Varianza en Procesos Industriales <i>Jorge Domínguez Domínguez y Hortensia Moreno Macías</i> .....	35
Classification Using Graphical Models <i>Guillermina Eslava y Leticia Cañedo</i> .....	40
Clustering based on rules <i>versus</i> Knowledge Discovery of Data: An Application to Astronomy <i>Karina Gibert y Ulises Cortés</i> .....	45
Modelación Gráfica en el Control de la Validez de Construcción de Test Psicológicos <i>Adalberto González Debén, Jesús E. Sánchez García y Ma. Odette Lobato Calleros</i> ....	52
Distribuciones de Referencia para Ciertas Familias Exponenciales <i>E. Gutiérrez-Peña y R. Rueda</i> .....	60

Simulación de Procesos de Riesgo en Ambiente Markoviano <i>Luis Fernando Hoyos Reyes</i> .....	67
Análisis del Comportamiento de la Salinidad en una Laguna Costera por Medio de Métodos de Suavización no Paramétrica <i>Jorge M. López Reynoso, Isaías H. Salgado Ugarte y Ma. José Marques Dos Santos</i> ..	71
Analizando la Distribución de Índices de Ozono de la Ciudad de México por Medio de Estimadores de Densidad por Kernel <i>María José Marques Dos Santos e Isaías H. Salgado Ugarte</i> .....	78
Estimación de la Eficiencia de un Método no Paramétrico para Probar la Multimodalidad de Datos Univariados <i>Juana Martínez R., Erika Mayorga S. e Isaías H. Salgado-Ugarte</i> .....	84
Estadística Ji-Cuadrada para Observaciones Emparejadas <i>Andrzej Matuszewski</i> .....	93
Modelos de Rasch y Log-lineales para Minería de Datos dentro del Formalismo de Dempster-Shafer <i>Andrzej Matuszewski u Guillermo Morales</i> .....	97
Definition of some Integral U-statistics for Tests of Independence <i>David Mayer-Foulkes</i> .....	102
Inferencia Bayesiana para el Cociente de las Medias de Dos Poblaciones Normales con Varianzas Distintas <i>M. Mendoza y E. Gutiérrez-Peña</i> .....	108
Inferencia Bayesiana a Partir de Distribuciones Finales Multimodales <i>L. E. Nieto-Barajas y E. Gutiérrez-Peña</i> .....	114
Regresión Bayesiana: Análisis y Comparación de Modelos Lineales Generalizados <i>Gabriel Nuñez Antonio</i> .....	119
Errores de Redondeo Vistos como Recursión <i>Federico O'Reilly y Raúl Rueda</i> .....	125

Análisis y Corrección de Valores Críticos de un Método para Evaluar Contrastes en Factoriales no Replicados <i>Jorge Olgún Uribe y Patricia Romero Mares</i> .....	130
Estimación de Componentes de Varianza en un Modelo con Partición de Efectos Apli- cado a Ensayos de Híbridos de Maíz <i>Emilio Padrón Corral, Angel Martinez Garza y Ma. Cristina Vega S.</i> .....	136
Sobre la Convergencia del Método de Ascenso por Pendiente Máxima <i>Blanca R. Pérez S., Sergio de los Cobos S. y Miguel A. Gutiérrez A.</i> .....	145
Modelo Lineal Difuso (una aplicación) <i>José C. Romero Cortés y Arturo Aguilar Vázquez.</i> .....	149
Uso de Histogramas Desplazados Promedio y Estimadores de Densidad por Kernel para el Análisis de la Frecuencia de Tallas de Datos Biológico-Pesquero <i>Isaiás H. Salgado Ugarte y Ma. José Marques dos Santos.</i> .....	155
Sobre la Identificación de Observaciones Influyentes en el Análisis de Supervivencia <i>Belem Trejo Valdivia</i> .....	166
Co-Integración en Series de Tiempo <i>Alfredo Troncoso V. y Alejandro Alegría H.</i> .....	172
Estudio de Encuesta Sobre La Producción Industrial de Cereales <i>H. J. Vázquez</i> .....	178
On Information Functionals and Priors <i>Francisco Venegas-Martínez</i> .....	183



# Una Aplicación de Modelado Gráfico en el Colegio de Ciencias y Humanidades de la UNAM

Miguel Ángel Abreu Hernández

*UAM-Azcapotzalco*

## 1 Introducción

El objetivo del estudio consiste en encontrar variables y modelos que expliquen la alta deserción y bajos promedios en los estudiantes del CCH; para tratar de detectar, desde su ingreso a estudiantes con esta problemática. Esta situación no es privativa de esta institución, por lo que el estudio puede resultar de aplicación más general.

El estudio presenta problemas prácticos siendo uno de los más importantes la naturaleza de los datos. Una parte importante de los mismos proviene del cuestionario de ingreso a la UNAM (64 variables de antecedentes socioeconómicos, familiares y escolares en 4,793 alumnos). Se da un índice alto de no respuesta. La mayoría de las variables continuas han sido “discretizadas” y se manejan en modelos “categóricos” que requieren muestras muy grandes lo que se contrapone con el problema de no respuesta que reduce substancialmente el número de casos. Además, se pierde normalidad. Sin embargo, condicionando sobre no respuesta y sobre algunas variables, y al construir promedios se suele recuperar normalidad y obtener resultados para subpoblaciones. Algo similar sucede con la otra fuente de datos constituida por las calificaciones de 18,561 alumnos del CCH en las asignaturas cursadas de las 32 posibles. Que a su vez, están categorizadas como NP, NA, S, B, MB. También se cuenta con la clasificación de las asignaturas por semestre y tipo. Los datos son de la generación 1992 a 1995.

Es necesario hacer notar que no se invertirán recursos en el proyecto por lo que, en esta fase, es retrospectivo, transversal, descriptivo y observacional. En esta situación los modelos estadísticos gráficos lineales pueden ser una herramienta muy útil, dado que requieren “pocos” supuestos. En este trabajo se estudia el comportamiento de los promedios semestrales en grupos de alumnos determinados por su grado de avance. El estudio general incluye varias etapas.

## 2 Modelos Gráficos

Un modelo gráfico es un conjunto de vértices, uno por cada variable en el modelo y un conjunto de aristas, una para cada dos variables interrelacionadas. El modelo representa una estructura de independencia condicional entre las variables bajo estudio. De todos los modelos posibles se busca el modelo más simple consistente con los datos, obteniéndose así las relaciones funcionales entre las variables. Los modelos fueron seleccionados condicionando sobre algunas variables y utilizando el procedimiento “stepwise” del programa MIM (Mixed Interaction Modelling) de Edwards (1990), que ajusta modelos estadísticos de interacción mixtos jerárquicos y “de descomposición” a los datos.

Los gráficos se interpretan de acuerdo al conocimiento de la situación del CCH (enfoque de sistemas); a la regla fundamental para interpretar gráficos de independencia llamada “la propiedad de Markov global”: Si dos conjuntos de variables  $U$  y  $V$  están separados por un tercer conjunto de variables  $W$  en el sentido de que todos los senderos que conectan  $U$  y  $V$  interceptan  $W$ , entonces  $U$  y  $V$  son condicionalmente independientes dado  $W$ . En la figura 2,  $U = A$ ,  $V = B$  y  $W = C$ ; y, en el caso continuo, (caso normal multivariado) observando la inversa de la matriz de covarianza: dos variables son independientes dado el resto de las variables, sí y sólo sí el correspondiente elemento de la covarianza inversa es cero. (Whittaker, 1990; Edwards, 1990,1995).

Todo modelo gráfico tiene una fórmula de modelo y viceversa como se observa en las figuras 1 y 2. En casos con más variables estas fórmulas se basan en las subgráficas triangulares del modelo gráfico llamadas camarillas o clanes. Este enfoque, muy resumido aquí, se aplica al caso CCH de la UNAM.

La figura 1 muestra la dependencia entre las variables  $A$  y  $B$ :  $AB$ . En la figura 2 se observa que dicha relación es espuria pues desaparece al incluir la variable clínica,  $C$ , que era una variable oculta o de confusión en la primera figura. Es decir,  $A$  es condicionalmente independiente de  $B$  dado  $C$ . Lo que se denota como Este fenómeno se conoce como la paradoja de Simpson (Bishop, Fienberg, y Holland, 1975). En consecuencia es necesario asumir un punto de vista multivariado y analizar conjuntamente todas las variables involucradas. Este es el enfoque del modelado gráfico.

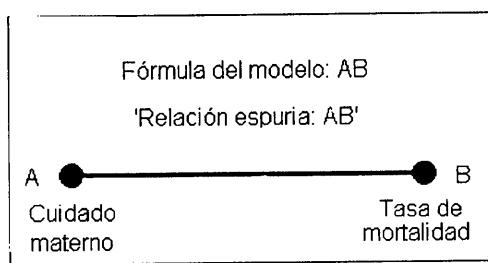


Figura 1

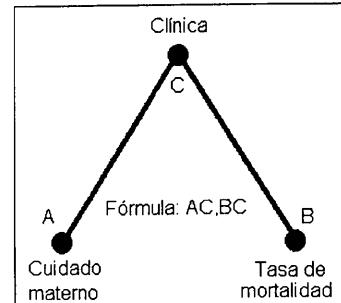


Figura 2

### **3 El Enfoque o Pensamiento de Sistemas**

La búsqueda de variables ocultas que pueden constituirse en variables de confusión no es un proceso directamente estadístico. La disciplina que está realizando una investigación, que a su vez utiliza estadística multivariada, es la que puede sugerir otras variables que se requieran utilizar y el modelado gráfico puede ayudar a decidir que variables deben incluirse en el estudio. Es un proceso complementario. En este proceso de búsqueda el enfoque de sistemas puede ser fundamental.

Esquemáticamente, el enfoque de sistemas trata un problema, dentro de un sistema complejo y dinámico, de forma interdisciplinaria y considerando la manera en que sus componentes contribuyen al propósito general del mismo; a su vez, lo estudia como parte de sistemas mayores. Por ejemplo, puede haber estudiantes que mejoren su promedio semestral consistentemente pero por razones no académicas sino por ciertos comportamientos aprendidos, esto no puede ser detectado dentro del sistema CCH, se requieren medidas externas como exámenes diagnóstico al ingreso y egreso al CCH. Es decir, puede existir un factor de confusión. En el contexto sistémico se pueden localizar puntos críticos para resolver problemas nodales y no secundarios, “principio de la palanca”, (Senge, 1990). Por caso, el detectar estudiantes con deficiencias en su preparación previa, para subsanarla y mejorar su autoestima, puede influir de manera decisiva y permanente en contra de la deserción y a favor del aprovechamiento escolar.

Por otra parte, no es posible manipular 96 variables como un solo modelo gráfico, hay que considerar la naturaleza de las variables y si representan antecedentes socioeconómicos, calificaciones, etcétera, para tener hipótesis y modelos de fácil manejo estadístico e interpretación adecuada. La búsqueda de relaciones no espurias entre características de un fenómeno o problema, dentro de un sistema, requiere el uso de la estadística dentro de un enfoque de sistemas según la materia de que se trate.

### **4 El Caso CCH**

$PROMCCHT = (10MB + 8B + 6S + 3NA + 0NP)/32$ . MB, número de asignaturas con calificación “Muy bien ó 10”; B, número de asignaturas con calificación “Bien u 8”; S, “Suficiente ó 6”; NA, “No acreditada ó 3, como promedio entre 0 y 5.99” y NP, “No presentada ó 0”. 32 asignaturas que se deberían haber cursados en tres años.

Al dividir entre 32 el grado de avance queda implícito en el promedio final “Promccht” que toma un valor entre cero y diez. Por caso, un alumno sin materias aprobadas en los tres años tiene un “promccht” entre 0 y 3; un alumno que aprobó las 32 materias tiene un “promccht” entre 6 y 10. Lo que permite suponer que la normalidad se presenta dentro de subpoblaciones al condicionar sobre el grado de avance, determinado como el número de asignaturas aprobadas en el transcurso de tres años. Lo cual se probó gráficamente y me-

diante pruebas de bondad de ajuste Kolmogorov-Smirnov, pero no se observa normalidad en los promedios semestrales.

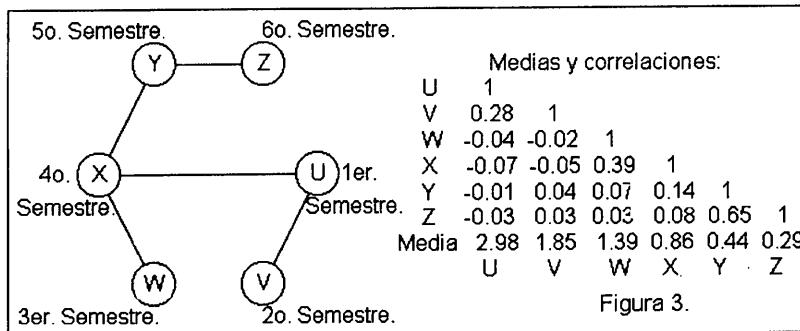


Figura 3: Modelo gráfico con fórmula //UV,UX,WX,XY,YZ y matriz de correlaciones de los promedios semestrales de los alumnos con grado de avance: 0 a 5 asignaturas aprobadas. 914 casos (5%).

Se seleccionó, por “stepwise backward” al nivel 5, un modelo gráfico para cada una de tres subpoblaciones, según grado de avance.

Por restricciones de espacio sólo se presentan algunos resultados. Se incluye el modelo gráfico o cualitativo, respaldado por la matriz de correlaciones y los promedios semestrales, modelo cuantitativo; aunque también es importante la matriz de covarianza inversa, etcétera.

Figura 3: Prácticamente sólo existe relación entre un promedio y el siguiente, lo cual también se observa en las correlaciones. Poca consistencia en el comportamiento global de los promedios semestrales a lo largo de los tres años.

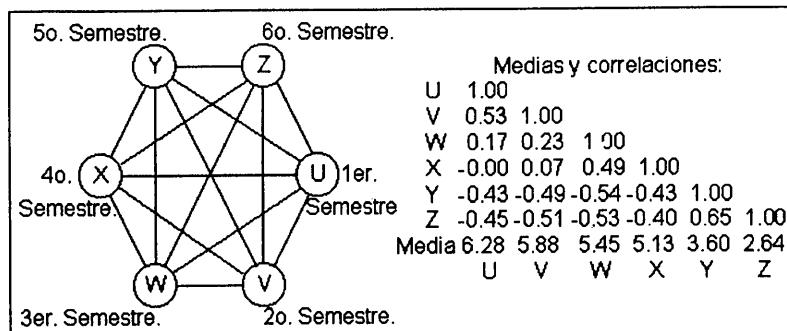


Figura 4: Modelo gráfico con fórmula //UVWXYZ y matriz de correlaciones de los promedios semestrales de los alumnos con grado de avance: 16 a 20 asignaturas aprobadas. 1946 casos (10.5%).

Figura 4: Gran consistencia en comportamiento global de los promedios pero las correlaciones entre los primeros semestres y los últimos es negativa, esto no es observable

en el modelo gráfico. Figura 5: La falta de la arista entre el primero y último semestres podría deberse a que, la mayoría, mejoran su promedio de manera uniforme y consistente. Se observa una “leve pero consistente mejoría” en los sucesivos promedios semestrales que provoca que la relación entre U (primer semestre) y Z (último semestre) sea la “más pequeña”: 0.32 y por ello, esa arista el procedimiento de selección la elimina. Lo cual corrobora la hipótesis de que la mayoría, de los estudiantes que terminan el CCH en tres años mejoran su promedio de manera uniforme y consistente. Aunque, esta prueba aún no es definitiva. Las tres subpoblaciones muestran comportamientos muy bien definidos por el modelo gráfico, la matriz de correlaciones y los promedios.

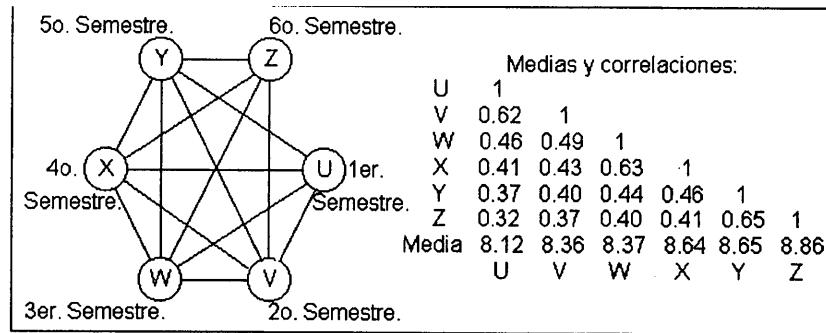


Figura 5: Modelo gráfico con fórmula //UVWXY, UWXYZ y matriz de correlaciones de los promedios semestrales de los alumnos con grado de avance: 32 de 32 asignaturas aprobadas. 4996 casos de 5013 (27% de 18,561 alumnos inscritos).

## 5 Algunas Conclusiones

De los resultados y perspectivas: Se llegan a observar subpoblaciones bien definidas, incluso el “promccht” del 1er. semestre puede ayudar a predecir el comportamiento futuro de un alumno. Sin embargo, se pretende predecirlo desde su ingreso. Por lo cual, se retrocederá para buscar relaciones de las subpoblaciones con los antecedentes socioeconómicos, familiares y escolares, lo que implica pasar a modelos categóricos y mixtos. La violación de algunos supuestos de los modelos gráficos parece influir poco en la pérdida de confiabilidad de los resultados, aún no se pueden sacar conclusiones pues esta parte del trabajo aún está en proceso de elaboración por medio de simulaciones. Se aplicarán otras técnicas estadísticas para cumplir con otros objetivos de este estudio en el CCH.

De las ventajas y limitaciones de los modelos gráficos: No indican directamente la fuerza ni la dirección de la relación. Al considerar un gran número de variables pueden encontrarse relaciones funcionales que no tengan ningún significado para la disciplina de que se trata y hay que buscar nuevas variables que permitan confirmar o rechazar un significado “real”. El modelado gráfico es una excelente herramienta como análisis exploratorio en la

búsqueda de relaciones entre variables o características de entes de un sistema toda vez que éstos modelos requieren, en general, pocos supuestos. Lo anterior siempre y cuando se respalde con matrices de correlación, modelos paramétricos, etcétera. La investigación por medio de modelos gráficos, sin el apoyo del pensamiento sistémico, que permite tener objetivos claros, puede volverse una búsqueda muy desordenada, lenta e incluso infructuosa.

## Referencias

- Bishop, Y.M., Fienberg, S. y Holland, P. (1975). *Discrete Multivariate Analysis*. Cambridge: MIT Press.
- Edwards, D.E. (1990). Hierarchical Interaction Models, (with discussion). *J. R. Statist. Soc. B*, 52: 1, 3-20.
- Edwards, D.E. (1995). *Introduction to Graphical Modelling*. New York: Springer-Verlag.
- Senge, P. M. (1990). *La quinta disciplina*. Buenos Aires: Ediciones Juan Granica.
- Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. New York: Wiley.

# Modelos Lineales Generalizados en Actuaría

Alejandro Alegría y Evangelina Martínez

*Departamento de Estadística, ITAM*

## 1 Modelos Lineales Generalizados

Los modelos lineales generalizados, introducidos en 1972 por Nelder y Wedderburn, involucran el análisis de las relaciones entre dos tipos de variables. Por un lado, las variables respuesta o variables dependientes del modelo son consideradas como variables aleatorias  $Y_1, Y_2, \dots, Y_n$ , cuyos valores cambian en función a las llamadas variables explicativas o independientes del modelo,  $X_1, X_2, \dots, X_n$ . Estas últimas son consideradas como dadas, y pueden representar medidas en un continuo, en cuyo caso reciben el nombre de covariables, o bien pueden dar lugar a una clasificación de las observaciones en  $k$  categorías o niveles, recibiendo entonces el nombre de factores. Los modelos lineales generalizados son una extensión de los modelos lineales clásicos y se definen de acuerdo a tres componentes.

### Componente Aleatorio.

La distribución de cualquier  $Y_i$  ( $i = 1, 2, \dots, n$ ), pertenece a la familia exponencial, es decir,  $f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}$ , donde  $\phi$  es el parámetro de dispersión y  $\theta_i$  es el parámetro canónico, si  $\phi$  es conocido. Se puede demostrar que

$$\mu_i = E(Y_i) = \frac{\partial b(\theta_i)}{\partial \theta_i} = b'(\theta_i) \quad \text{y} \quad V(Y_i) = a(\phi) \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2} - a(\phi) b''(\theta_i),$$

donde  $b'(\theta_i)$  recibe el nombre de *función varianza* y se denota por  $V(\theta_i)$ .

### Componente Sistemático.

Las variables explicativas del modelo,  $X_1, X_2, \dots, X_n$ , ya sean covariables o factores, producen un *predictor lineal*  $\eta^T = (\eta_1, \eta_2, \dots, \eta_n)$  dado por  $\eta_i = \sum_{j=1}^p \beta_j X_{ij}$ , o bien  $\eta = \mathbf{X}\beta$ , donde  $\beta^T = (\beta_1, \beta_2, \dots, \beta_p)$  es el vector de parámetros y  $\mathbf{X}$  es la *matriz de diseño* formada por las variables explicativas. En la matriz de diseño se incluyen los valores de las observaciones correspondientes a cada covariable y se introducen tantas variables *indicadoras* como sean necesarias para representar los niveles asociados a cada factor.

### Función Liga.

El predictor lineal puede expresarse como una función conocida del valor esperado de  $Y_i$ , es decir,  $g(\mu_i) = g[E(Y_i)] = \eta_i$ , donde  $g(\cdot)$  es una función monótona y diferenciable llamada *función liga*. Cuando  $g(\mu_i) = \theta_i = \eta_i$ , la función recibe el nombre de *liga canónica*.

## 2 Estimación de $\beta$

Una vez especificado el modelo de acuerdo a los tres componentes anteriores, se procede a estimar los parámetros  $\beta_1, \beta_2, \dots, \beta_p$  a partir de los datos haciendo uso del método de *máxima verosimilitud*. El *estimador máximo verosímil*  $\hat{\beta}$  será la solución al siguiente sistema de ecuaciones:

$$U_j = \frac{\partial l(\theta, \phi; \mathbf{y})}{\partial \theta_j} = \sum_{i=1}^n \frac{\partial l(\theta, \phi; y_i)}{\partial \theta_j} = \sum_{i=1}^n \frac{(y_i - \mu_i)}{Var(Y_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right) = 0, j = 1, 2, \dots, p$$

donde  $l(\theta, \phi; \mathbf{y})$  es el logaritmo de la función de verosimilitud.

Debido a que las ecuaciones anteriores son no lineales, métodos como el de *Newton-Raphson* y el de *scoring* son de los más usuales para obtener el estimador máximo verosímil de  $\beta$ .

## 3 Medidas de Bondad de Ajuste

**Devianza.** Mide la discrepancia entre el modelo ajustado y el *modelo saturado* con  $n$  parámetros en el cual las estimaciones obtenidas corresponden exactamente a los datos. De acuerdo a lo anterior, la *devianza*  $D$ , cuya distribución muestral es una  $\chi^2_{(n-p)}$ , está dada por  $D = 2[l(\hat{\beta}_{max}; \mathbf{y}) - l(\hat{\beta}; \mathbf{y})]$ , donde  $\hat{\beta}_{max}$  es el estimador máximo verosímil de  $\beta$  bajo el modelo saturado y  $\hat{\beta}$  es el estimador máximo verosímil de  $\beta$  bajo el modelo ajustado.

La devianza como medida de bondad de ajuste resulta particularmente útil al momento de elegir la estructura del predictor lineal. Por otro lado, para el caso en el que se han ajustado modelos de acuerdo a diferentes funciones liga, un criterio para elegir el más adecuado está dado por la magnitud relativa de las devianzas para cada modelo.

**Residuales.** Los residuales son otras medidas de bondad de ajuste y para los modelos lineales generalizados existen varias definiciones, dos de las más usuales se presentan a continuación.

El *residual de Pearson*, que corresponde a las raíces de las contribuciones de cada observación a la *estadística de Pearson* de bondad de ajuste, se define como  $r_p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\mu_i)}}$ , donde  $V(\mu_i)$  es la función varianza y  $\hat{\mu}_i$  es el estimador máximo verosímil de  $\mu_i$ .

Otra medida es el *residual de devianzas*,  $r_D$ , el cual se define en términos de la contribución individual  $d_i$  de cada observación a la devianza del modelo, es decir,  $r_D = sgn(y_i - \mu_i) \sqrt{d_i} = sgn(y_i - \mu_i) \sqrt{2[l(\hat{\beta}_{max}; \mathbf{y}) - l(\hat{\beta}; \mathbf{y})]}$ .

Gráficas de residuales contra valores ajustados y covariables resultan de gran utilidad para detectar observaciones atípicas, para determinar si el modelo describe adecuadamente los efectos de las covariables en él consideradas e incluso para determinar si es necesaria la introducción de términos adicionales al predictor, o más covariables al modelo.

## 4 Graduación con Respecto a la Edad

Se entiende por graduación al conjunto de métodos por medio de los cuales es ajustado un grupo de probabilidades observadas con el fin de proporcionar una base adecuada para hacer inferencias y cálculos. La justificación de la graduación de la experiencia de un grupo está en el supuesto de que las tasas reales de mortalidad de cada edad pueden ser representadas por una función matemática razonablemente simple y suave. De esta forma, se definen las siguientes tasas de mortalidad.

La *probabilidad de muerte* entre las edades  $x$  y  $x + 1$ , denotada por  $q_x$ , se calcula como  $q_x = \frac{A_x}{R^i}$ , donde  $A_x$  es el número de muertes registradas a edad  $x$  y  $R^i$  son los *expuestos iniciales* al riesgo de muerte, es decir el número de personas que entran en observación a edad exacta  $x$  y continúan hasta sobrevivir a edad exacta  $x+1$ .

Por otro lado, la *fuerza de mortalidad* o *tasa instantánea de mortalidad* a edad  $x$ , denotada por  $\mu_x$ , está dada por  $\mu_x = \frac{A_x}{T}$ , donde  $T$  corresponde a los *expuestos centrales*, es decir al período total de exposición al riesgo de muerte, incluyendo el tiempo al que estuvo expuesto cada individuo.

### 4.1 Graduación de $q_x$

Para la graduación de probabilidades de muerte se considera a  $A_x$  como la variable aleatoria respuesta del modelo y a la edad  $x$  como la única covariante. Si suponemos que la muerte o sobrevivencia de cada individuo bajo observación es independiente de la de los demás, se tiene que  $A_x \sim Bin(R^i, q_x)$ , distribución que pertenece a la familia exponencial.

La estructura del componente sistemático o predictor lineal del modelo puede tomar diferentes formas, a continuación se presentan algunas de las más usuales.

Los predictores *polinomiales* son de la forma  $\eta_x = \sum_{j=0}^r \beta_j h_j(\frac{x-a}{b})$ , donde  $\{h_j\}$  es una base ortogonal de polinomios introducida para evitar problemas de colinealidad, y las constantes  $a$  y  $b$  se introducen para cambiar la escala de la edad y por facilidad de cómputo.

Predictores Gompertz-Makeham y Logit Gompertz-Makeham. Sean  $r$  y  $s$  enteros no negativos y  $\alpha$  y  $\beta$  vectores de coeficientes. Se define el predictor *Gompertz-Makeham* como  $\eta_x = GM_x(r, s) = \sum_{i=0}^{r-1} \alpha_i h_j(\frac{x-a}{b}) + \exp\{\sum_{j=0}^{s-1} \beta_j h_j(\frac{x-a}{b})\}$ , mientras que el predictor *Logit Gompertz-Makeham*, denotado por  $LGM(r, s)$ , está dado por  $\eta_x = GM_x(r, s) = \frac{GM_x(r, s)}{1+GM_x(r, s)}$ .

Una elección natural para la especificación de la función liga es la logit, ya que resulta ser la liga canónica para la distribución binomial; es decir,  $\eta_x = \ln(q_x/(1 - q_x))$ . El uso de esta función, en combinación con predictores polinomiales, ha servido como base para las graduaciones elaboradas por el *Continuous Mortality Investigation (CMI) Bureau*.

Por otro lado, Renshaw (1991) propuso el uso de la función log-log complementaria como liga en combinación con un predictor polinomial, es decir  $\eta_x = \ln(-\ln(1 - q_x))$ , obteniendo como fórmula de graduación:  $q_x = 1 - \exp\{-\exp[\sum_{j=0}^r \beta_j h_j(\frac{x-a}{b})]\}$ .

## 4.2 Graduación de $\mu_x$

Supongamos que un grupo de personas se observa entre las edades  $x$  y  $x+1$ , para varios períodos comprendidos en ese año; que la fuerza de mortalidad es constante para todas las personas durante el año e igual a  $\mu_{x+1/2}$ , y que la muerte o sobrevivencia de cada individuo es independiente de la de los demás. Sea  $T$  el tiempo total en el que son observadas las personas, entonces  $A_x \sim Po(T\mu_{x+1/2})$  (Sverdrup, 1965).

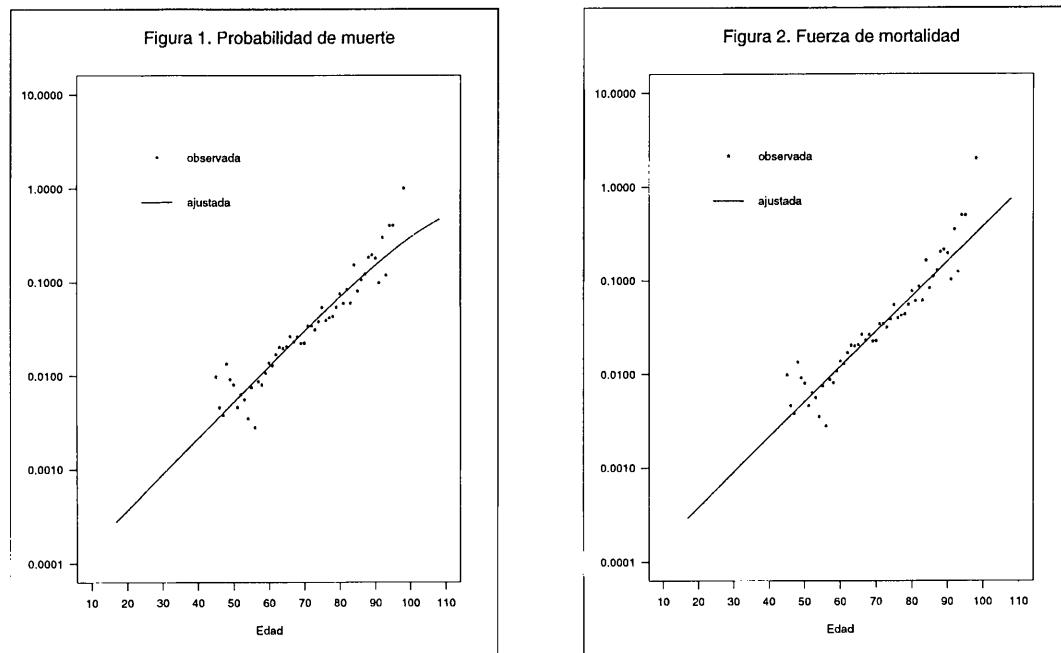
Para la graduación de la fuerza de mortalidad, se utilizan predictores Gompertz-Makeham en combinación con la función logaritmo, que es la liga canónica para la distribución Poisson; es decir,  $\eta_x = \ln(T\mu_{x+1/2}) = \ln(T) + \ln(\mu_{x+1/2})$ . En este caso, el sumando  $\ln(T)$  es un término denominado *offset*, cuyo coeficiente estimado es 1.

Una alternativa para la elección de la función liga es la identidad, utilizada por Forfar, *et al.* (1988) en combinación con predictores Gompertz-Makeham, es decir,  $\eta_x = T\mu_{x+1/2} = GM_x(r, s)$ .

## 4.3 Ejemplos (Forfar, *et al.*, 1988 y Renshaw, 1991).

Los datos corresponden a las muertes registradas (692 en total) y expuestos al riesgo de muerte (28,386.5 años, en total) de viudas de pensionados de 1979 a 1982 en Gran Bretaña. Las edades oscilan entre los 17 y 108 años y no hay muertes registradas a edades menores de 45 y mayores a 98 años.

Se especificaron predictores polinomiales de grado 1 con polinomios ortogonales y edad escalada para las graduaciones tanto de  $q_x$  como de  $\mu_x$  y como ligas las funciones logit y logaritmo, respectivamente. En las figuras 1 y 2 se presentan los valores observados y las graduaciones resultantes.



## 5 Comparación Entre la Mortalidad de Fumadores y No Fumadores

En este caso, el objetivo es la graduación de la fuerza de mortalidad  $\mu$  para fumadores y no fumadores, en base a la clasificación de los individuos de acuerdo a la covariable edad,  $x$  (11-15,16-20,...,96-100) y a los factores sexo (i), hábito (j) con tres niveles (no fumador, fumador y no diferenciado) y estado (k) con dos niveles (médico y no médico). De esta forma, las unidades o individuos quedan especificados como  $u = \{i,j,k,x\}$ . La variable respuesta del modelo  $A_u$ , número de muertes, se modela como una variable aleatoria Poisson sobredispersa (Forfar *et al.*, 1988, Renshaw, 1991, 1992), ya que los datos se basan en pólizas y no en asegurados, es decir,  $E(A_u) = T\mu_u$  y  $\text{Var}(A_u) = \phi T\mu_u$ ,  $\phi > 1$ .

Renshaw (1994) usó la función logaritmo como liga junto con predictores polinomiales, quedando especificado el modelo como  $\eta_u = \ln(T_u) + \sum_{v=0}^p z_{uv}\beta_v$ , donde  $p$  es el número de factores y  $z_{uv}$  es un polinomio en  $x$ . Este modelo lo ajustó a datos correspondientes a muertes registradas entre 1988 y 1989, obteniendo como fórmula de graduación:  $\mu_{xijk} = GM_x(0,4) = \exp\{\alpha + (\tau + \theta_i + \psi_j)x + \beta x^2 + \gamma x^3\}$ . Como puede observarse en la expresión anterior, solamente se tomaron en cuenta para la graduación los factores correspondientes al sexo (i) y al hábito (j) de la persona, sin importar su estado médico.

Como resultado de la graduación realizada, se observó que la tasa de mortalidad para los fumadores es consistentemente mayor que la correspondiente a los no fumadores en todas las edades. Asimismo, se registraron tasas mayores en mujeres fumadoras que en hombres no fumadores.

## 6 Conclusiones

La graduación de la mortalidad de la experiencia de un grupo de personas por medio de los modelos lineales generalizados aquí presentados, representa una alternativa interesante frente a otros métodos de graduación. Esto se debe a que es posible aplicar los resultados inferenciales que se tienen para dichos modelos a las estimaciones obtenidas. Asimismo, la alternativa aquí presentada admite la inclusión de diferentes variables explicativas, ya sea factores o covariables, en la graduación.

El siguiente paso en esta investigación es aplicar las técnicas de modelos lineales generalizados a la experiencia mexicana, así como explorar las aplicaciones de estos modelos en otros campos de la ciencia actuarial, como lo es la determinación de primas para seguro de automóviles.

Cabe destacar que existen otros temas actuariales en los que se han aplicado satisfactoriamente los modelos lineales generalizados, entre los que se encuentran las distribuciones de pérdida y la clasificación de riesgo para asegurados con respecto a la mortalidad y a la terminación prematura de una póliza (Haberman y Renshaw, 1996).

## Referencias

- Bowers, N.L., Gerber, H.U., Hickman, J.C., Jones, D.A. y Nesbit, C.J. (1986). *Actuarial Mathematics*. Society of Actuaries.
- Forfar, D.O., McCutcheon, J.J. y Wilkie, A.D. (1988). On Graduation by Mathematical Formula. *Journal of the Institute of Actuaries*, 115, 1-135.
- Haberman, S. y Renshaw, A.E. (1996). Generalized Linear Models and Actuarial Science. *The Statistician*, 45, 407-436.
- Mc.Cullagh, P. y Nelder, J.A. (1983). *Generalized Linear Models*. London: Chapman and Hall.
- Nelder, J.A. y Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, A*, 135, 370-384.
- Renshaw, A.E. (1991). Actuarial Graduation Practice and Generalized Linear and Non-linear Models, *Journal of the Institute of Actuaries*, 118, 295-312.
- Renshaw, A.E. (1994). A Comparison Between the Mortality of Smoking and Non-smoking Assured Lives in the U.K. *Journal of the Institute of Actuaries*, 121, 561-571.
- Sverdrup, E. (1965). Estimated and Test Procedure in Connection with Stochastic Models for Deaths, Recoveries and Transfers between different States of Health. *Skandinavisk Aktuarietidskrift*, 48, 184.

# Construcción de un Modelo para un Estudio Multicéntrico en Hospitales de Tercer Nivel

Lilia Benavides

Alejandro Aldama

*Dept. de Sistemas Biológicos*

*UAM-Xochimilco*

*Dept. de Sistemas*

*UAM-Azcapotzalco*

Héctor-Javier Vázquez

*Dept. de Sistemas*

*UAM-Azcapotzalco*

## 1 Introducción

La rápida emergencia y diseminación de los gérmenes resistentes a los antibióticos se ve favorecida por la presión selectiva que ejerce el uso indiscriminado y generalizado de estos agentes. A través de mecanismos Darwinianos de selección de células resistentes que involucran una gran diversidad de mecanismos genéticos, los agentes patógenos bacterianos están en franca reaparición ; tal es el caso del *S. aureus* meticilina resistente, del enterorococo resistente a la vancomicina, del *M. bacterium* multiresistente y de los neumococos que alguna vez fueron una de las bacterias más sensibles a la penicilina. La diseminación tan veloz de las clonas resistentes y la emergencia de nuevas variantes de mecanismos de resistencia demandan un cambio de enfoque : del reconocimiento del problema a la intervención para modularlo. Las consecuencias mas importantes de la resistencia bacteriana a los antibióticos es que puede conducir a un aumento en la incidencia de enfermedades.

Para comprender y controlar este problema, es necesario desarrollar programas de Vigilancia Epidemiológica, realizar estudios epidemiológicos y modelos que permitan describir los diferentes mecanismos y proponer métodos de intervención. Este trabajo presenta resultados (estadísticos) obtenidos durante el proceso de elaboración de un modelo descriptivo de los perfiles de resistencia bacteriana a los antibióticos en los hospitales del D.F. e identificar los factores que la modulan.

## 2 Descripción del Sistema y su Modelación

La información epidemiológica acerca de la resistencia a los antibióticos proviene de tres fuentes principales: de la vigilancia, de la investigación de brotes y de estudios prospectivos. En muy pocas ocasiones se colectan datos sobre la resistencia a los antibióticos como parte del sistema de vigilancia; cuando se llegan a reunir, son útiles para identificar las tendencias en la frecuencia de la resistencia a los antimicrobianos. Sin embargo dada la falta de una metodología integral, estos estudios no pueden utilizarse para elucidar factores asociados con la emergencia, persistencia y transmisibilidad de las bacterias resistentes.

La resistencia bacteriana se estudia con frecuencia en las infecciones asociadas al hospital. Se ha observado que en las instituciones hospitalarias se presentan episodios periódicos de resistencia a ciertos antibióticos que involucran a una variedad de microorganismos. Durante estos procesos, se han identificado algunos factores que afectan la resistencia como son: las características del microorganismo, los reservarios naturales (ambientales o humanos) en los cuales los genes de la resistencia o los microorganismos resistentes persisten, los distintos tipos de cambios sociales y tecnológicos que influyen en transmisibilidad de los organismos resistentes así como las modalidades del uso de los antimicrobianos. Por otra parte, dado que las formas de utilización de los antibióticos varían de hospital a hospital, y dentro de un mismo hospital, de un tiempo a otro, se ha sugerido que cada hospital debe identificar sus problemas con la prescripción de antibióticos y su modelo de prescripción. La obtención de estos modelos será el primer paso a seguir en la elaboración y aplicación de cualquier programa dirigido a hacer más racional la terapia antimicrobiana y disminuir con esto la resistencia bacteriana a los antibióticos.

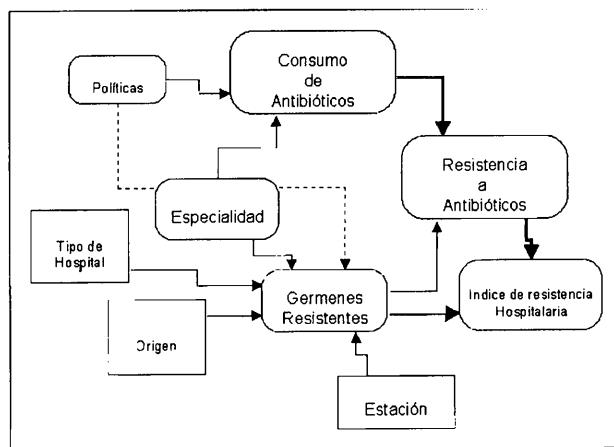


Figura 1

Con base a diversos estudios y a los conocimientos obtenidos en el presente proyecto se propone un primer modelo (figura 1) con el fin de explicar el proceso de evolución de la resistencia en las instituciones hospitalarias. Este modelo no es definitivo y no

pretende cuantificar la relaciones de causa-efecto. Antes de terminar la construcción de un modelo es necesario aplicar procesos iterativos que permitan validar la existencia de las relaciones entre variables mediante estudios estadísticos uni y multivariados. Un objetivo posterior será obtener un modelo que describa de una manera adecuada las diferentes variables de acuerdo a sus distintas características y funciones : exógenas y/o endógenas, relaciones entre variables independientes y dependientes, observadas o no observadas así como los diferentes efectos: directos, indirectos, covariación e interacciones, Susser (1973), Hanneman (1997).

Los resultados que fueron obtenidos provienen de un estudio farmacoepidemiológico con las siguientes características:

Tipo de estudio: observacional, descriptivo, transversal y multicéntrico. Tamaño de la muestra: seis hospitales de los Institutos Nacionales de Salud, III nivel, ubicados en el D.F., (Benavides y Aldama, 1996). Especialidades: Cardiología, Oncología, Medicina Interna, Neumología, Neurología y Pediatría. 100 casos de infección nosocomial con germen aislado correspondientes a 1994- 1995. 859 aislamientos 68 cepas distintas. 3 instancias del hospital encuestadas con cuestionario y entrevista: Comité de Control de Infecciones. Laboratorio de microbiología diagnóstica y farmacia. Las formas de reportar la información sobre niveles de sensibilidad de los gérmenes fueron cualitativamente como: R = resistente N = no evaluado S = si evaluado El consumo de antibióticos: en DDDp o dosis diarias dispensadas en farmacias de cada antibiótico, por paciente egresado.

### 3 Análisis y Resultados

Una primera etapa fue la de realizar un estudio de cada variable sobre las 859 observaciones. Algunos resultados muestran lo siguiente:

Distribución de observaciones por especialidad: Cardiología (122), Oncología (61), Medicina Interna (227), Neumología (172) Neurología (194), Pediatría (83). Distribución de observaciones por año: año 1994 (383), año 1995 (475) Distribución de cepas : Se detectaron 68 cepas bacterianas distintas, correspondientes a 18 géneros. 15 son Cepas responsables del 80% de resistencia: Escherichia coli (135), Pseudomonas aeruginosa (110), Staphylococcus aureus (72), Staphylococcus epidermidis (66), Klebsiella pneumoniae (57). Pseudomonas sp(43) , Enterobacter sp (40), Staphylococcus coagulasa negativa (36), Enterococcus faecium (29), Enterobacter cloacae (26), Acinetobacter calcoaceticus (23), Staphylococcus coagulasa positiva (17), Serratia marcescens (15), Citrobacter freundii (14), Enterococcus faecalis (14) Distribución de antibióticos : Se obtuvo información sobre sensibilidad a 50 diferentes antibióticos (Abs), la información fue predominantemente cualitativa (modalidades : Resistente, Sensibilidad media , Sensibilidad). Los antibióticos contra los que se detectó mayor frecuencia de Rs fueron : Cf ,Ak,Gm, Am, T/S, Pip, Caz, Cip, Ti, Cft/ctx, Crm/Cxm, To, Cax, Cfz, Tim, Imp, Aug, Nx, Cfx/Fox, P, E, Of. El consumo

intrahospitalario se valoró en Dosis Diarias Dispensadas por paciente o DDDp. Del 66% (33) de los antibióticos se dispuso de información en la farmacia y en el laboratorio de microbiología y del 34% (17) sólo se obtuvo sobre dispensación (14) ó sobre resistencia (3). 38% de los Abs se dispensaron en dosis igual ó > a 1 DDDp, de éstos, el 18% presentó valores de R de 3 dígitos, 14% de 2 dígitos, 6% de un dígito y 0% sin R. En los Abs reportados como dispensados en niveles < a 1 DDDp (28%), se observó una R elevada de 3 dígitos en el 42% y de 2 dígitos en el 28%. Antibióticos con mayor consumo (DDDp de 1.33 al 11.50) :Gm, Cfz,Cd,Dx, Ak, T/S, P, Cf, Tubs, Cxl, Cax, Dox, Rif/Ra, Am, Caz, Cip, Va, Cfx/Fox.

De este estudio uni y bivariado se obtiene una visión de la distribución de frecuencias de las modalidades de cada variable, así como los valores dominantes (mediante el uso de curvas de Pareto). Estos resultados indican niveles de R importantes a la mayoría de los antibióticos consumidos. El hecho de observar resistencia aun en los antibióticos dispensados en bajas dosis podría explicarse por un subregistro del consumo real de estos fármacos. En lo que respecta a las posibles interrelaciones entre variables, un estudio bivariado frecuencial permitió distinguir las distribuciones y detectar relaciones interesantes: (Hospital vs. Origen), (Cepa vs. Especialidad), (Especialidad vs. Origen), (Cepa vs. Origen), (Consumo Antibióticos vs. Perfil de Resistencia). Un análisis multivariado (Análisis de Correspondencias Múltiples) sobre las 859 observaciones no permite reducir el número de variables (se obtienen 141 valores propios, 30% de la varianza se concentra en los cinco primeros). Aunque una presentación visual de los dos primeros planos, responsables del 16% de varianza, permite a primera vista clasificar las observaciones en 5 grupos. Dada la dificultad de encontrar una reducción del número de variables, se orientó el estudio hacia el análisis de las unidades estadísticas: Hospital, Cepas, Antibióticos. Únicamente se presentan los resultados del análisis multivariado del perfil de antibióticos:

Del análisis de conglomerados con los métodos de Ward, Método Jerárquico Completo, Método Jerárquico por Centroides, Método Jerárquico por Promedios, resultan asociados los siguientes antibióticos: (Ak-Gm), (Caz-Cip), (Crm,Cip), (Aug, Tim, Imp, Nxn), (Azt, Ofi, Fna), (Dx, Cd), (Amo, Va), (Cdz, F/M), (Ra, Cxl, Dox).

En la Figura 2a se presentan los resultados del estudio de conglomerados con el método de promedios, donde se observan 5 grupos. El método de promedios pareció el mas adecuado por el compromiso ofrecido respecto a las otras opciones y por su mejor calidad para visualizar los resultados gráficos (programa JMP, SAS, 1997).

En la figura 2b se presentan los resultados (95% de varianza en dos ejes) del Análisis de Componentes Principales (Lebart *et al.*, 1984).

Como se puede observar visualmente es posible identificar grupos de antibióticos. Los grupos asociados parecen resultar del efecto de dos variables principales: Número de cepas resistentes o % de resistencia con respecto a la muestra total y las DDDp. Resultando 4 grupos identificados por los dos métodos multivariados. Grupo 1: Ak, Gm, Caz-Cip con un % elevado de r's y de consumo de antibióticos. Grupo 2: Crm-Cft, Aug-Tim-Imp-Nxn

con un % bajo de r's y de consumo de antibióticos. Amo-Va-Cdz-FM-Ra-Cxl-Dox con un porcionto bajo de consumo de antibióticos

Grupo 3: Dx-Cd con un consumo elevado de antibióticos y un % baio de r's.

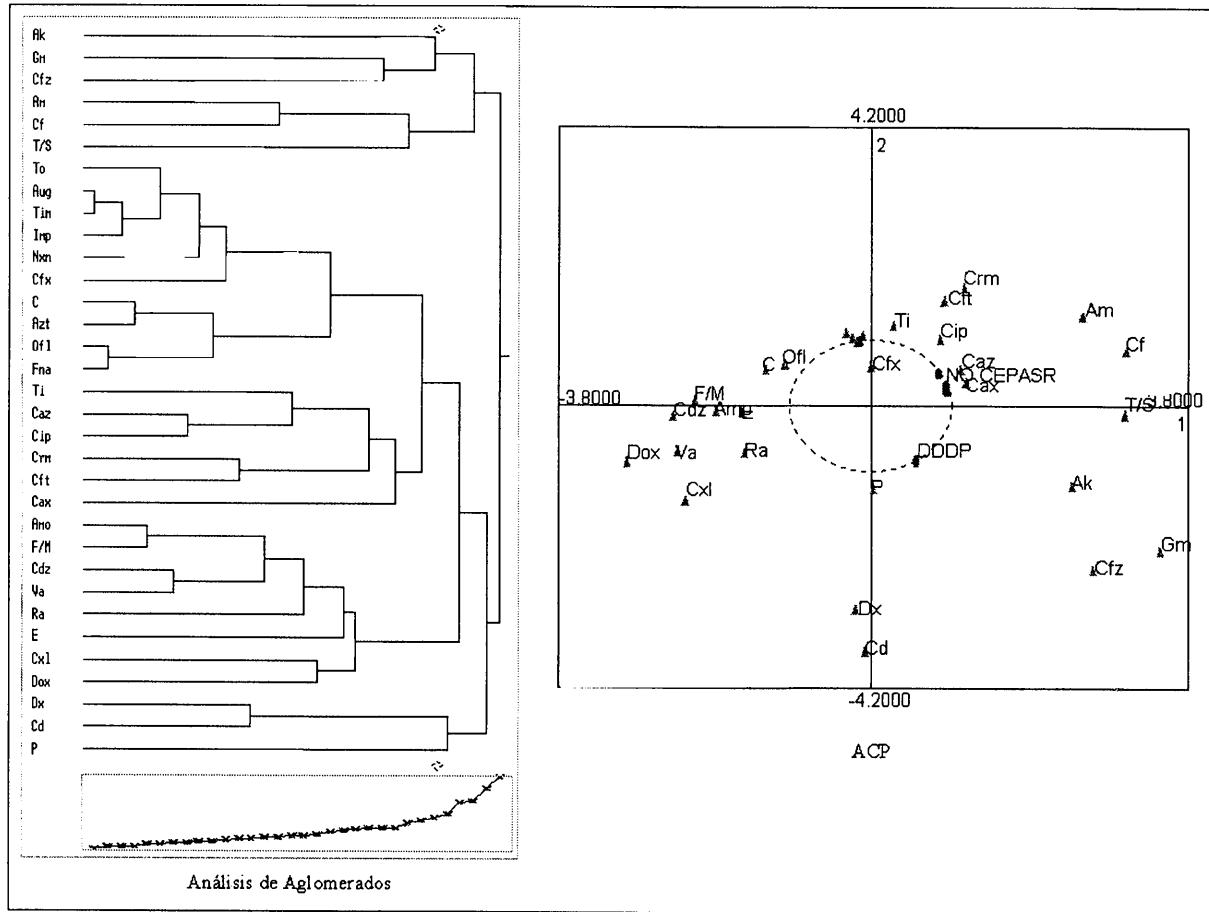


Figura 2a y 2b

## 4 Conclusiones

Los resultados han permitido comprender la jerarquía entre diferentes modalidades de las variables estudiadas e identificar grupos con características bien definidas. Con esto es posible visualizar, en el caso de la variable Antibiótico, relaciones cepas resistentes y consumo de antibiótico por paciente. Por ejemplo de la figura 2 ( ACP), se puede observar un grupo a la derecha de antibióticos con altos índices de CepasR y altos consumos, mientras que del lado izquierdo se encuentra un grupo con bajos índices de CepasR y bajos consumos.

Una vez identificados los grupos se pretende estudiar las relaciones inter e intra grupos respecto a un número reducido de características, con el fin de identificar mejor los efectos delineados en el modelo.

El proceso de modulación es largo y ha requerido un estudio detallado de cada variable antes de poder postular y validar posibles interrelaciones entre ellas. Sin embargo este estudio a sido muy benéfico en resultados. Estos ya permiten comprender el efecto sobre la resistencia del consumo de ciertos antibióticos. Es importante resaltar que aun cuando se conoce la presión selectiva y algunas veces inductiva del uso de antimicrobianos sobre las cepas bacterianas, es necesario identificar qué antibióticos en particular (no grupos de antibióticos) están ejerciendo mayor presión, sobre cuáles géneros bacterianos, de manera tal que se disponga de elementos que fundamenten las políticas de uso de antimicrobianos en cada institución hospitalaria. Por otro lado, una vez asignados pesos a las distintas variables, se podrán incorporar las mas importantes en el diseño de un programa que permita hacer el seguimiento periódico de la resistencia a los antibióticos no solo en cada hospital sino a nivel de cada localidad y nacionalmente. Esto redundara en un uso mas racional de los antibióticos y una disminución del fenómeno de la resistencia a estos fármacos.

## Referencias

- Benavides P. L. *et al.* (1996). Estudio Multicéntrico sobre Resistencia a los Antibióticos en Hospitales de Tercer Nivel en el Distrito Federal. XI Foro Nacional de Estadística, México.
- Hanneman, R. A. (1997) *Multivariate Analysis Course*. Riverside: University of California.
- (<http://wizard.ucr.edu/rhannema/soc203a/diagram.html>)
- Lebart L. *et al.* (1984) *Multivariate Descriptive Analysis*. New York: Wiley.
- Suser, M. (1973). Causal Thinking in Health Sciences: Concepts and Strategies in Epidemiology. New York: Oxford University Press.

# Métodos de Detección de Observaciones Influyentes Multivariadas ante Varias Observaciones Aberrantes

Eduardo Castaño Tostado

*Universidad Autónoma de Querétaro*

## 1 Introducción

Uno de los problemas básicos de la Estadística es la obtención óptima de valores estimados  $\hat{\theta}$  de parámetros desconocidos  $\theta$  de una distribución  $F$  asociada a una variable aleatoria  $X$ . En teoría, se parte en general de una muestra aleatoria  $X_1, \dots, X_n$  de la que al obtenerse su distribución empírica  $F_n$  se tiene que  $\hat{\theta} = \hat{\theta}(F_n)$ , atendiendo a criterios de optimalidad estadística. Este procedimiento de estimación no debe ser usado en la práctica de una manera dogmática, por lo que es necesario ejercer un diagnóstico sobre el grado en que las suposiciones son válidas en el conjunto de datos bajo análisis. En este contexto una pregunta relevante es si las  $n$  piezas de información muestral ejercen la misma influencia en el procedimiento de estimación de  $\theta$ . Si  $\hat{\theta}$  es altamente dependiente de un pequeño subconjunto de la muestra, se deberán de tomar precauciones en la inferencia estadística derivada. En general las respuestas a este tipo de preguntas se agrupan en lo que se conoce como detección de observaciones influyentes. Por otra parte, se ha reconocido en varios contextos los peligros de los métodos clásicos de análisis estadístico ante la presencia de múltiples "outliers" u observaciones aberrantes, por la posibilidad del efecto de enmascaramiento (masking effect).

En esta nota comparamos dos enfoques de detección de observaciones multivariadas influyentes en la estimación de una matriz de covarianzas, ante la presencia de varias observaciones aberrantes.

## 2 La Función de Influencia

Un enfoque para la detección de observaciones influyentes utiliza la función de influencia (Hampel, 1974). Desde el punto de vista teórico, considérese la perturbación de  $F$  en la dirección de una  $z$  en el dominio de  $F$ ,

$$F \rightarrow F(\epsilon) = (1 - \epsilon)F + \epsilon\delta_z, \epsilon > 0, \delta_z = \begin{cases} 1 & \text{en } z \\ 0 & \text{o.c.} \end{cases}$$

De esta manera la función de influencia teórica de  $\theta$  se define por el conjunto

$$\theta^{(1)} : \{\theta_z^{(1)} = \lim_{\epsilon \downarrow 0} \frac{\theta(F(\epsilon)) - \theta(F)}{\epsilon}\}_z.$$

Para análisis de datos se tienen dos versiones muestrales; la primera denominada la función de influencia empírica

$$\hat{\theta}^{(1)} : \{\hat{\theta}_{x_i}^{(1)} = \lim_{\epsilon \downarrow 0} \frac{\hat{\theta}(F_i(\epsilon)) - \hat{\theta}(F_n)}{\epsilon}\}_{x_i},$$

donde  $F_i(\epsilon)$  es  $(1 - \epsilon)F_n + \epsilon\delta_{x_i}$ . La segunda versión muestral se denomina la función muestral de influencia, que es el conjunto

$$\{-(n - 1)(\hat{\theta}(F_{(i)}) - \hat{\theta}(F_n))\}_{x_i}$$

donde  $\hat{\theta}(F_{(i)})$  representa el valor estimado de  $\theta$  previamente, omitiendo la  $i$ -ésima observación del proceso de estimación,  $i = 1, \dots, n$ . El cómputo de esta función muestral de influencia es relativamente ineficiente comparando con el cómputo de la función empírica; esta última es la que utilizaremos en la comparación de enfoques. Influencia empírica ha sido utilizada en la detección de observaciones influyentes en métodos de estimación del análisis de regresión, por ejemplo vea Cook y Weisberg(1982), y en varios métodos de estimación de métodos multivariados de análisis, por ejemplo, Critcley(1985) y Tanaka (1988) y sus colaboradores. En un caso básico, cuando  $X$  representa un vector de dimensión  $p$  con vector de medias  $\mu$  y matriz de varianzas y covarianzas  $\Sigma$ , la función de influencia teórica de  $\Sigma$  es (Critchley,1985),

$$\Sigma^{(1)} = (X - \mu)(X - \mu)^T - \Sigma,$$

con sus consecuentes símiles muestrales. En el caso muestral, dado que la función empírica de influencia es una matriz, para propósitos prácticos de detección de observaciones influyentes se computan normas como la de Frobenius

$$\|\hat{\Sigma}^{(1)}\|$$

o la llamada distancia generalizada de Cook (GCD)

$$GCD(\hat{\Sigma}^{(1)}) = (vech(\hat{\Sigma}^{(1)})^T acov(\hat{\Sigma}))^{-1} vech(\hat{\Sigma}^{(1)})$$

donde  $acov(\hat{\Sigma})$  es la matriz de varianzas asintóticas de los elementos de  $\Sigma$  con elementos  $\frac{1}{n-1}s_{ik}s_{jl} + s_{il}s_{jk}$ ,  $s_{ik}$  denotando la covarianza muestral entre la variable  $i$  y variable  $j$  (Anderson, 1984).

### 3 Influencia Local

Cook (1986) introdujo el concepto de influencia local en el contexto de detección de observaciones influyentes. Denotemos por  $L(\theta)$  al logaritmo de la verosimilitud con base de modelo de probabilidad subyacente y a  $L(\theta | \omega)$  al logaritmo de la verosimilitud perturbada a través de  $\omega$ . Con esto se puede construir lo que llamó Cook la gráfica de influencia ( $\omega, L(\theta) - L(\theta | \omega)$ ). Entonces se propuso encontrar direcciones de curvatura máxima en la gráfica de influencia, siendo equivalente a encontrar el eigenvector dominante  $l_{max}$  de

$$F = \left( \frac{\delta^2 L(\theta | \omega)}{\delta \theta \delta \omega^T} \right)^T \frac{\delta^2 L(\theta)}{\delta \theta \delta \theta^T} \frac{\delta^2 L(\theta | \omega)}{\delta \theta \delta \omega^T}.$$

Así, una observación  $x_i$  si es influyente tendrá un valor grande en el elemento correspondiente de  $l_{max}$ . Kim (1996) aplicó lo anterior para detectar observaciones influyentes en  $\hat{\Sigma}$ ; para ello utilizó como esquema de perturbación

$$X_j \sim N(\mu, \Sigma/\omega_j), j = 1, \dots, n.$$

A partir de este esquema de perturbación mostró que el elemento  $(r, s)$  de la matriz  $\tilde{F}$  es

$$\frac{-1}{n} (MD_{rs} + \frac{1}{2} MD_{rs}^2), i, j = 1, \dots, n,$$

donde  $MD_{rs} = (x_r - \bar{x})^T \dot{\Sigma}^{-1} (x_r - \bar{x})$ .

### 4 Influencia y Enmascaramiento

La presencia de observaciones aberrantes puede distorsionar el proceso de estimación, por lo que la detección de observaciones influyentes puede también verse afectada; esto se puede apreciar ya que la detección de observaciones influyentes depende de los valores estimados de manera clásica (no robusta) de los parámetros de interés. Entonces es recomendable estimar de manera robusta y proceder posteriormente a la detección de observaciones influyentes.

En el contexto de estimación de vectores de medias y de la matriz de varianzas y covarianzas de un vector de dimensión  $p$ , Rousseeuw y van Zomeren (1990) presentan el llamado estimador de elipsoide de volumen mínimo (MVE) como una alternativa robusta cuando los datos provienen de distribuciones elípticas. El MVE es un estimador que tiene ciertas propiedades de equivarianza y cuenta con un punto de quiebre grande que le da la característica de robusticidad ante una gran contaminación por aberrantes. Sus propiedades teóricas y de implementación computacional han recibido ya tratamiento, vea por ejemplo los trabajos de Lopuhaä y Rousseeuw (1991) y de Hawkins (1994). En este

trabajo utilizamos el algoritmo de remuestreo propuesto por Rousseeuw y Van Zomeren (1990).

En la siguiente sección comparamos los dos enfoques para detección de observaciones influyentes ante la presencia de varias observaciones aberrantes.

## 5 Comparación de Enfoques de Influencia

El conjunto de datos base para la comparación es denominado Hawkins- Bradu-Kass con  $n = 75$  observaciones en  $p = 3$  variables, conjunto que contiene 14 observaciones claramente aberrantes. El objetivo es entonces evaluar la influencia de cada observación en el procedimiento de estimación clásica utilizando los dos enfoques de influencia ya presentados, obsevar efectos de enmascaramiento y utilizar estimación robusta MVE de la matriz de covarianzas para enfrentar efectivamente el enmascaramiento. Los resultados de la cuantificación usando estimación clásica se muestran en la Tabla 1 mientras que en la Tabla 2 se muestran los resultados usando estimación MVE. Debe notarse que sólo presentamos los resultados para las primeras 14 observaciones aberrantes que en este contexto de estimación de  $\Sigma$ , son influyentes. Las influencias de las demás observaciones son mucho menores.

A partir de estas dos Tablas observamos que  $\|\Sigma^{(1)}\|$  no sufre de enmascaramiento pero que  $GCD(\hat{\Sigma}^{(1)})$  y  $l_{max}$  se ven afectados por la presencia de múltiples aberrantes, ya que sólo detectan como influyente a la observación 14. De la Tabla 2 se observa que tal situación de enmascaramiento se ve completamente corregida. Entonces la selección de una norma que involucre a  $\hat{\Sigma}^{-1}$  puede provocar problemas de enmascaramiento al utilizar la función empírica de influencia. Por su parte el enfoque de influencia local sufre el efecto enmascaramiento.

Tabla 1. Influyentes y estimación clásica

obs.	$\ \hat{\Sigma}^{(1)}\ $	$GCD(\hat{\Sigma}^{(1)})$	$l_{max}$
1	471	17438	-.007
2	515	95367	-.018
3	619	3247	-.003
4	680	103761	-.007
5	644	54745	-.006
6	543	19272	-.010
7	549	63508	-.021
8	490	18585	-.005
9	619	58982	-.004
10	553	25781	-.002
11	954	243310	-.014
12	1051	8132	-.013
13	994	623019	-.095
14	1378	8256873	-.989

Tabla 2. Influyentes y estimación robusta MVE

obs.	$\ \hat{\Sigma}^{(1)}\ $	$GCD(\hat{\Sigma}^{(1)})$	$l_{max}$
1	1097	28836271	-.206
2	1153	32494379	-.217
3	1281	39096036	-.241
4	1358	45029730	-.256
5	1313	41698305	-.247
6	1188	33575132	-.222
7	1196	34427050	-.224
8	1121	30201193	-.211
9	1282	40039091	-.242
10	1198	34975406	-.227
11	1691	69836336	-.319
12	1805	77999260	-.340
13	1738	74374547	-.324
14	2170	135293257	-.387

## Referencias

- Cook, R.D. y Weisberg S. (1982). *Residuals and influence in regression*. Chapman Hall.
- Critchley, F. (1985). Influence in principal component analysis. *Biometrika*, 72, 627-636.
- Hawkins, D. M. (1993). A feasible solution algorithm for the minimum volume ellipsoid estimator in multivariate data. *Computational Statistics*, Vol. 8, 95-107.
- Kim, M.G. (1996). Local influence in multivariate normal data. *Journal of Applied Statistics*, Vol. 23, No. 5, 535-541.
- Lopuhaä, H.P. and Rousseeuw, P.J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, Vol. 19, No. 1, 229-248.
- Rousseeuw, P.J. y van Zomeren, B.C. (1990). Unsmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, Vol. 85, 633-651.
- Tanaka, Y. (1988). Sensitivity analysis in principal component analysis: influence on the subspace spanned by principal components. *Communications in Statistics, Series A*, 17, 3157-3175.

# Estimación de la Probabilidad de No-Pago de la Cartera Hipotecaria del Sistema Bancario Mexicano

Ma. de Lourdes de la Fuente D. y José Manuel Pelayo C.

*Comisión Nacional Bancaria y de Valores*

## 1 Introducción

A raíz de la crisis económica de 1994 el sistema financiero mexicano ha enfrentado un problema de cartera vencida. El aumento de la cartera vencida ha provocado un incremento en el riesgo tanto de otorgamiento como de seguimiento de los créditos.

Debido a lo anterior es necesario obtener mediciones adecuadas del riesgo de las carteras de crédito con objeto de evitar la descapitalización de la banca. Los modelos de medición de riesgo utilizados hoy en día presentan algunos problemas, como por ejemplo el hecho de que la información que utilizan es en general subjetiva, los modelos de valuación no tienen los fundamentos estadísticos y económéticos necesarios y muchos de ellos, como los modelos de *Credit Scoring* y *Behavioral Scoring* tienen altos costos de implementación.

El propósito de este trabajo es proponer un modelo que permita cuantificar el nivel de riesgo de la cartera hipotecaria en base a la experiencia de pago y a las características del acreditado.

El procedimiento que se siguió para el análisis es el siguiente: en primer lugar se estableció un esquema de muestreo para obtener la información referente a las características del crédito y a la experiencia de pago. Posteriormente se realizó un análisis de la información a partir de la cual se estimó un índice de experiencia de pago y finalmente se estimó la probabilidad de no-pago de la cartera utilizando la metodología de modelos de elección cualitativa y de matrices de transición.

Los resultados de la estimación permitieron estimar la probabilidad de no-pago para cada acreditado, así como identificar aquellas características de los créditos que determinan el que un crédito pueda caer en una situación de morosidad.

## 2 Determinación de la Muestra y Variables Analizadas

El universo total de créditos hipotecarios del sistema bancario es aproximadamente de 900,000 créditos. A partir de una muestra aleatoria de 7500 créditos proporcionada por las

tres instituciones bancarias más importantes en México, se obtuvo una muestra aleatoria de 2500 créditos en base a la cual se realizó el análisis.

A partir de la información disponible, se analizaron las siguientes variables:

**Variables Cuantitativas:** Deuda Total, Saldo Vencido, Saldo Insoluto, Valor de la Garantía, Monto Concedido, Plazo y Experiencia de Pago.

**Variables Cualitativas:** Tipo de Crédito, Región y Destino del Crédito.

donde,

*Saldo Insoluto:* Incluye el monto del capital vigente más el margen diferido o refinanciamiento, además de los intereses de la mensualidad vigente aún no pagados por el acreditado.

*Saldo Vencido :* Principal no pagado en la amortización más los intereses de la misma.

*Deuda Total:* Suma de los dos conceptos anteriores.

*Valor de la Garantía :* Importe actualizado por INPC (sector vivienda) del valor de la garantía.

*Monto Concedido:* Monto del crédito original.

*Plazo:* Número de meses de vigencia contratados.

*Experiencia de Pago:* Perfil de pagos del acreditado. A partir de esta información se construyó una variable dicotómica definida como 1 si en el mes correspondiente se registró pago y 0 en otro caso.

Posteriormente, se construyeron variables dicotómicas como indicadores de algunas características relevantes en el análisis, las cuales se presentan a continuación:

Para el caso del destino del crédito.

*Dadqui:* Definida como 1 cuando el crédito fue otorgado para la adquisición de inmuebles y 0 en otro caso.

*Dcons:* Definida como 1 cuando el crédito fue otorgado para la construcción de inmuebles y 0 en otro caso.

*Dliqui:* Definida como 1 cuando el crédito fue otorgado para liquidez y 0 en otro caso.

Para el caso de tipo de crédito:

*Dmedres:* Definida como 1 cuando el tipo de crédito fue para media residencial y 0 en otro caso.

*Dintso:* Definida como 1 cuando el tipo de crédito fue para interés social y 0 en otro caso.

Para el caso de las Garantías:

En este caso se definieron variables dicotómicas que indican el porcentaje de la deuda total cubierto por la garantía, este porcentaje puede ser desde el 10% hasta el 200%.

*Dgadek*: Definida como 1 cuando el valor de la garantía cubre al menos el  $k\%$  de la deuda total y 0 en otro caso donde  $k = 10, 20, \dots, 200$ .

Antes de proceder a la estimación de los modelos, se hizo un análisis de la información muestral con objeto de determinar si existían diferencias significativas en la cobertura de la muestra con respecto al sistema bancario, en términos de su distribución por actividad económica, entidad federativa y tipo de crédito. En todo los casos hay evidencia de que las proporciones que representa la muestra no son significativamente diferentes de las del sistema.

### 3 Estimación de un Índice de Experiencias de Pago

A partir de la información correspondiente al perfil de pagos de cada acreditado, para un periodo de 12 meses (julio de 1995 a junio de 1996), se construyó una variable dicotómica definida como 1 en el caso en que en el mes correspondiente, el acreditado haya pagado cualquier cantidad positiva de su deuda y 0 en otro caso. Para construir el Índice de Experiencia de Pago, se utilizó la metodología de Matrices de Transición (Cadenas de Markov), ver Narayan (1972).

Sea  $X_m$  la ocurrencia del estado  $X$  en el periodo  $m$ , la probabilidad de pasar del estado  $i$  al estado  $j$  en un periodo se define como:

$$P_{ij}^{(m,m+1)} = P(X_{m+1} = j | X_m = i)$$

Se supondrá que  $P_{ij}^{(m,m+1)}$  es independiente de  $m$ , es decir, es homogénea en el tiempo.

La matriz de transición está dada por:

$$P = \begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix}$$

Donde,  $P_{00}$  se refiere a la probabilidad de que un acreditado que se encuentre al corriente en un periodo, siga al corriente en el siguiente periodo.  $P_{01}$  es la probabilidad de que un acreditado que se encuentre al corriente en un periodo, presente un vencimiento en el siguiente periodo, etc.

Si se eleva la matriz de transición a la potencia  $M$ , la matriz resultante contiene las probabilidades de pasar del estado  $i$  al estado  $j$  en  $M$  periodos de tiempo.

A partir de la información muestral se obtuvieron los estimadores de Máxima Verosimilitud para las probabilidades de transición y a partir de la matriz de transición  $P$ , se estimaron las matrices de transición para los demás periodos. El índice de experiencia de

pago se obtuvo como sigue:

$$P_{cvnj12} = \frac{\sum_{n=1}^{10} P_{i^n5}^n}{10}$$

donde,  $n$  se refiere a los periodos que conforman el historial de pago de cada acreditado.

$P_{i^n5}^n$  es la probabilidad de que un acreditado que se encuentra en el estado  $i$ , pase al estado 1 (moroso) en  $n$  periodos.

## 4 Estimación de la Probabilidad de No-Pago

A partir del índice de experiencia de pago y de las características de cada acreditado, se estimó un modelo de elección cualitativa tipo Probit (Greene 1993, Gujarati 1992).

$$I_i = F^{-1}(P_i) = \beta X + \epsilon$$

donde:  $X$  representa el vector de características de cada crédito,  $\beta$  es el vector de parámetros que mide el impacto de dichas variables sobre la probabilidad de no pago  $\epsilon$ , es el error aleatorio del modelo e  $I_i$  es un índice no observable asociado a la variable dicotómica dependiente definida como:

$$Y = \begin{cases} 1 & \text{cuando el acreditado presenta dos vencimientos en el último bimestre,} \\ 0 & \text{en otro caso.} \end{cases}$$

Los resultados de la estimación del modelo fueron los siguientes:

PROBIT	//	Dependent	Variable	is	Y
Sample:	1 2500	Included observations:	2500		
Convergence	achieved	after	5	iterations	
Variable	Coefficient	Std. Error	T-Statist	Prob.	
C	-7.6536	0.2725	-28.0811	0.0000	
IEP	16.9699	0.6083	27.8935	0.0000	
DINTSO	-0.7442	0.0971	-7.6575	0.0000	
DCONS	0.6222	0.3872	1.6069	0.1082	
Log likelihood	469.2218				
Obs with Dep=1	1435				
Obs with Dep=0	1065				

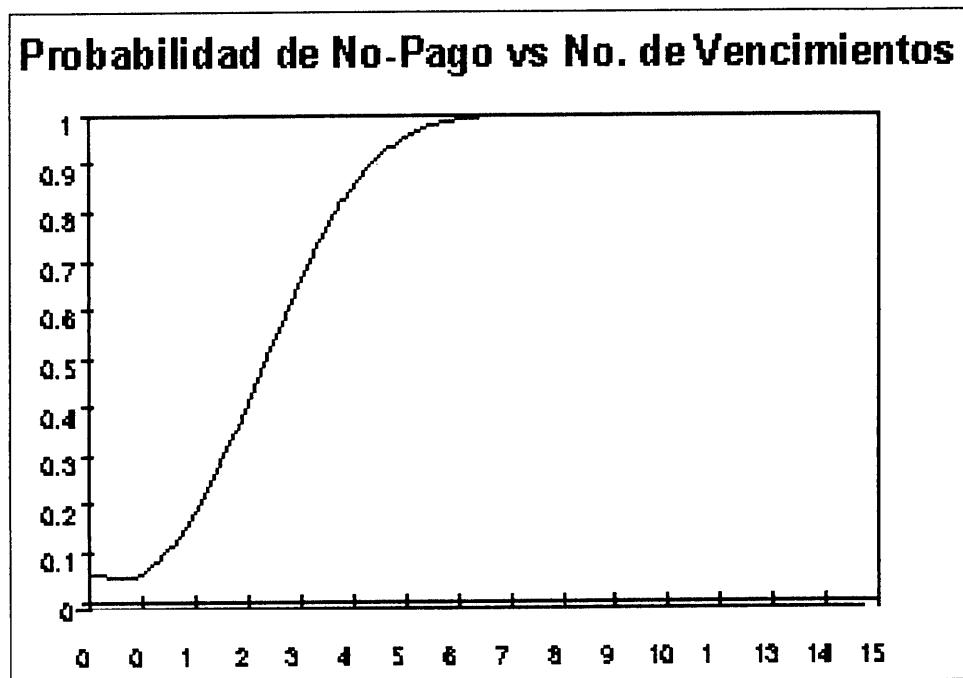
En los resultados de la estimación se observa que las únicas variables que resultaron significativas fueron el índice de experiencia de pago, y las variables Dintso y Dcons, las cuales indican que los créditos cuyo destino es para construcción y del tipo de interés social presentan un efecto diferencial sobre la probabilidad de no-pago, con respecto al resto de los créditos. El coeficiente positivo de la variable Dcons, indica que este tipo de créditos

tienen una probabilidad de no-pago mayor que aquellos que se destinan a la adquisición o liquidez.

La experiencia de pago tiene un impacto positivo, lo cual indica que un crédito con mala experiencia de pago tiene mayor probabilidad de presentar una situación de morosidad y por último la variable Dintso cuyo signo es negativo, indica que los créditos de interés social tienen menos riesgo que los de media residencial.

Otro resultado importante de la estimación es que la variabilidad explicada por el modelo es aproximadamente de 72.48%, sin embargo la experiencia de pago contribuye a esta explicación en un 70% lo cual indica que la contribución marginal de las variables restantes es de solo el 5%.

A partir de los resultados del modelo, se estimaron las probabilidades de no pago para cada crédito. La siguiente gráfica muestra los resultados donde puede observarse que en general un crédito con más de 7 vencimientos tiene una probabilidad de no-pago mayor al 90%.



## 5 Conclusiones

Los resultados del análisis reflejan la importancia de la experiencia de pago en la estimación de la probabilidad de no-pago para la cartera de créditos hipotecarios. La metodología

utilizada es de fácil implementación además de que permite una continua actualización de los resultados. Por último, el modelo otorga una percepción real del riesgo del acreditado<sup>1</sup>

## Referencias

- Greene, W. (1993). *Econometric Analysis*. 2a. Ed. New York: McMillan.
- Gujarati, D. (1992). *Econometría*. 2a Ed. New York: McGraw-Hill
- Narayan, U. (1972). *Elements of Applied Stochastic Processes*. New York: Wiley.

---

<sup>1</sup>La información contenida en este trabajo es responsabilidad de los autores y no refleja la posición de la Comisión Nacional Bancaria y de Valores.

# Redes Neuronales Probabilísticas: Perspectivas en Clasificación y Reconocimiento de Patrones

Sergio de los Cobos S.

*UAM-Iztapalapa, México*

Miguel A. Gutiérrez A.

*UAM-Azcapotzalco, México*

John Goddard C.

*UAM-Iztapalapa, México*

Blanca R. Pérez S.

*UAM-Iztapalapa, México*

## 1 Introducción

El paradigma de las redes neuronales trata en general, de imitar el proceso de solución de problemas que realiza el cerebro humano. Al igual que el ser humano aplica el conocimiento obtenido de experiencias pasadas para resolver problemas nuevos, las redes neuronales a través de ejemplos previamente resueltos construyen un sistema de “neuronas” (entrenadas en esos ejemplos) para realizar nuevas decisiones, clasificaciones y pronósticos.

## 2 Clasificaciones de Patrones Mediante la Estrategia de Bayes

Una de las estrategias utilizadas para clasificar patrones es aquella que minimiza el riesgo esperado. A esta estrategia se le conoce como la estrategia de Bayes, la cual puede aplicarse a problemas que contengan cualquier número de categorías. Considere por ejemplo, una situación de dos categorías en la que el estado de la naturaleza  $E$  sólo puede tomar cualesquiera de dos valores  $E_1$  o  $E_2$ . Se desea decidir cuándo  $E = E_1$  o  $E = E_2$  basados en un conjunto de medidas  $X = [x_1, \dots, x_j, \dots, x_p]$ , la regla de decisión de Bayes será:

$$d(X) = \begin{cases} E_1 & \text{si } h_1 l_1 f_1(X) > h_2 l_2 f_2(X) \\ E_2 & \text{si } h_1 l_1 f_1(X) < h_2 l_2 f_2(X), \end{cases}$$

donde  $f_i(X)$  son las funciones de densidad de la categoría  $i=1,2$ ;  $l_i$  es la pérdida asociada con la decisión  $d(X) = E_i$  cuando  $E = E_j$  para  $j=1,2$ ,  $i \neq j$ ;  $h_1$  es la probabilidad a priori de ocurrencia del estado  $E_1$ ,  $h_2 = 1 - h_1$ , y las pérdidas asociadas con la decisión correcta se toma igual a cero. Lo anterior puede extenderse de manera análoga al caso multivariado.

Cacoullos (1966) propone para el caso particular del kernel Gaussiano, los estimadores multivariados:

$$f_k(X) = \frac{1}{(2\pi)^{p/2} h^p} \frac{1}{m} \sum_{i=1}^m \exp -\frac{(X - X_{ki})^t (X - X_{ki})}{2h^2}$$

donde:  $k$  = categoría,  $i$  = número de patrón (muestra),  $m$  = número total de patrones de entrenamiento,  $X_{ki}$  =  $i$ -ésimo patrón de entrenamiento de la categoría  $k$ ,  $h$  = parámetro de suavizamiento y  $p$  = dimensión del espacio de medida.

### 3 Redes Neuronales Probabilísticas

En términos generales, existen dos tipos de redes neuronales: redes neuronales supervisadas y no supervisadas. Las redes neuronales supervisadas son aquellas en las que se presentan patrones de entrada a la red y ésta compara las salidas resultantes con respecto de las deseadas y entonces ajusta los pesos de la red de manera tal que se reduzca su diferencia, en cambio, en las redes neuronales no supervisadas, los patrones de entrada se aplican y la red se organiza ajustando sus pesos mediante un algoritmo bien definido.

Specht (1996) ha introducido las *redes neuronales probabilísticas* (RNP) como alternativa al paradigma de las redes neuronales supervisadas con la ventaja de que no requieren entrenamiento sino la asignación de pesos en una sola pasada. Como indica Specht “Las RNP son la versión clasificatoria que se obtiene cuando la estrategia de Bayes para la realización de la decisión se combina con un estimador no paramétrico para las funciones de densidad de probabilidad.

La topología de la RNP consta de una capa de entrada, dos capas intermedias, la de unidades patrón y la de unidades suma, y una capa de salida formada por una sola unidad. La capa de entrada contiene unidades de distribución, las cuales proporcionan los datos de la variable  $X$  a todas las neuronas de la segunda capa, unidades patrón. Esta capa contiene tantas unidades como la dimensión del espacio de medida. En las unidades patrón se realiza el producto y antes de salir su nivel de activación hacia las unidades suma, se realiza una operación no lineal, en particular puede ser:  $\exp\{-\frac{(X-X_{ki})^t(X-X_{ki})}{2h^2}\}$ . En esta capa se tienen tantas unidades como patrones de entrenamiento.

Las unidades suma, las cuales son tantas como categorías, realizan la suma de las salidas de las unidades patrón que corresponden a la categoría para la cual los patrones de entrenamiento fueron seleccionados. Finalmente, en la unidad de salida se realiza la decisión aplicando la regla de Bayes.

## 4 El Problema de Clasificación

Básicamente el objetivo del problema de clasificación (llamado también análisis de conglomerados (AC)), es el de obtener una partición de un conjunto de objetos basada ésta en las similitudes, o en las “distancias” entre los objetos de forma que, los objetos agrupados en la misma clase (grupo o conglomerado) sean similares o cercanos entre sí, y que los conglomerados estén bien diferenciados entre ellos.

Existen dos grandes familias de métodos: jerárquicos y clasificatorios ver por ejemplo de los Cobos *et al.* (1997b).

Han surgido diferentes heurísticas para atacar el problema de clasificación. Sin embargo, los métodos usuales de particionamiento se topán con el problema de la explosión combinatoria, además de que con frecuencia estos métodos caen en entrampamientos sub-optimales o valles profundos, por lo que ha inducido a investigadores a plantearse la necesidad de mejorar los algoritmos existentes mediante los llamados métodos de programación estocástica como son entre otros: algoritmos genéticos, búsqueda tabú y recocido simulado, ver por ejemplo de los Cobos *et al.* (1997a, 1997b).

## 5 Resultados Numéricos

En Murillo (1996) se considera el problema de clasificación y se ataca mediante búsqueda tabú. Piza y Trejos (1996), utilizan la técnica de recocido simulado y obtienen la misma clasificación para los casos de 3 y 4 clases para la sociomatriz de Thomas (matriz de 24x24, ver Murillo (1996)). Sin embargo, existe diferencia respecto de la clasificación realizada por Trejos *et. al.* (1996) utilizando algoritmos genéticos. Cabe mencionar que para el caso de 5 clases los agrupamientos son por completo diferentes en los tres trabajos, pero un resultado interesante es que tanto en Murillo (1996) como en Trejos *et al.* (1996), para el caso de 5 clases, aunque las particiones son diferentes, el valor de la función objetivo es el mismo utilizando el criterio de minimizar la inercia intra-clase.

Las particiones encontradas se proporcionan en la tabla 1. La sociomatriz de Thomas junto con las particiones dadas en la tabla 1, se corrieron en NEUROSHELL © bajo un modelo RNP, con número de neuronas: en la capa de entrada y en la segunda capa igual a 24, y de  $k$  en la capa de salida ( $k=4,5$ ), obteniéndose los resultados de la tabla 2.

Clases	Murillo	Piza y Trejos	Trejos et. al.
4	$\{1, 11, 12, 14, 21\}$ $\{15, 16\}$ $\{3, 5, 6, 7, 9, 17, 19, 23\}$ $\{2, 4, 8, 10, 13, 18, 20, 22, 24\}$	$\{1, 11, 12, 14, 21\}$ $\{15, 16\}$ $\{3, 5, 6, 7, 9, 17, 19, 23\}$ $\{2, 4, 8, 10, 13, 18, 20, 22, 24\}$	$\{3, 6, 7, 19\}$ $\{2, 4, 8, 10, 12, 13, 20, 22, 24\}$ $\{5, 21, 23\}$ $\{1, 9, 11, 14, 15, 16, 17, 18\}$
5	$\{1, 11, 12, 14, 21\}$ $\{15, 16\}$ $\{13, 18\}$ $\{3, 5, 6, 7, 9, 17, 19, 23\}$ $\{2, 4, 8, 10, 20, 22, 24\}$	$\{1, 11, 12, 14, 21\}$ $\{15, 16\}$ $\{7, 9, 17, 22, 23\}$ $\{3, 5, 6, 19\}$ $\{2, 4, 8, 10, 13, 18, 20, 24\}$	$\{1, 9, 14, 16, 21\}$ $\{7, 11, 15, 17\}$ $\{2, 8, 12, 13, 18, 22\}$ $\{3, 5, 23\}$ $\{4, 6, 10, 19, 20, 24\}$

tabla 1. Particiones de la sociomatriz de Thomas.

Clases	Murillo			Piza y Trejos			Trejos et. al.		
	1	2	3	1	2	3	1	2	3
4	24		.542	24		.542	23	9	.876
5	24		.542	24		.748	21	6, 7, 11	.977

tabla 2. Subcolumnas 1: número de elementos bien clasificados,  
2: elementos mal clasificados y 3: calibración.

## 6 Conclusiones

De los resultados de la tabla 2, se puede observar que una medida de la “bondad o estabilidad” de la partición es el factor de calibración, el cual es obtenido por la RNP de forma tal que se minimiza el número de elementos mal clasificados mediante un mapeo del error cuadrado medio en el intervalo [0,1].

Un aspecto interesante es el caso de la partición en 5 clases, en donde el valor de la función objetivo es el mismo (valor de inercia intra-clase = 202.58) tanto en Murillo (1996) como en Pizza y Trejos (1996). Parece que utilizando búsquedas tabú se obtiene una solución más “estable” respecto al factor de calibración de RNP utilizado por Neuroshell ©. Es notable observar la partición nueva después de utilizar RNP en la partición obtenida por algoritmos genéticos, en donde los elementos 6 y 7 pasan de los conglomerados 5 y 2 respectivamente al conglomerado 4 y el elemento 11 pasa del conglomerado 2 al conglomerado 1. Esta última partición concuerda con la partición original obtenida por búsquedas tabú.

Lo anterior no significa que siempre búsquedas tabú proporcione mejores resultados, lo que nos interesa en este trabajo es presentar por una parte una forma alternativa de evaluar la “bondad o estabilidad de una partición”, y por otra parte, el de poder identificar “elementos característicos” que proporcionen máxima información sobre el comportamiento de la partición, lo cual representa nuevas líneas de investigación.

## Referencias

- Cacoullos T. (1966). Estimation of a Multivariate Density. *Annals of Institute of Statistical Math. (Tokyo)*, **18-2**, pp. 179-189.
- De los Cobos S.S., Pérez S. B. y Gutiérrez A. M. (1997a). Programación Estocástica en Optimización. *Memorias del X Simposio Internacional de Métodos Matemáticos Aplicados a las Ciencias*, Liberia, Costa Rica, 3-7 febrero, (Castillo W. y Trejos J., Eds.), Universidad de Costa Rica - Instituto Tecnológico de Costa Rica, pp. 31-45.
- De los Cobos S.S., Pérez S. B. y Gutiérrez A. M. (1997b). Programación Estocástica: una Alternativa al Estudio de Conglomerados. *Memorias del XI Foro Nacional de Estadística*, Universidad Autónoma de Sinaloa, INEGI-AME, pp. 45-49.

Neuroshell 2 ©, release 3.0, Ward System Group, Inc.

Murillo A. (1996). Particionamiento Usando Búsqueda Tabú. *IV Encuentro Centroamericano de Investigadores en Matemática*, enero 17-19, Antigua Guatemala.

Piza V. E. y Trejos Z. J. (1996). Clasificación Automática Particionamiento mediante Sobrecalentamiento Simulado. *IV Encuentro Centroamericano de Investigadores en Matemática*, enero 17-19, Antigua Guatemala.

Specht D.F. (1996). Probabilistic Neural Networks and General Regression Neural Networks. En *Fuzzy Logic and Neural Network Handbook*, (Chen C.H., Ed.) New York: McGraw-Hill. pp. 3.1-3.44.

Trejos Z. J., Piza V. E. y Figueroa M. G. (1996). Clasificación Automática mediante un Algoritmo Genético: Resultados Numéricos. *IV Encuentro Centroamericano de Investigadores en Matemática*, enero 17-19, Antigua Guatemala.

# Comparación de Métodos para Modelar la Varianza en Procesos Industriales

Jorge Domínguez Domínguez y Hortensia Moreno Macías

*CIMAT*

*IIMAS, UNAM*

## 1 Introducción

Frecuentemente, se han usado las técnicas de regresión para identificar que variables tienen un efecto en el desempeño de un proceso. Generalmente, se trata de indagar cuál de las variables tiene influencia sobre la media de una característica del proceso, tal como se plantea en el siguiente modelo:

$$E(Y_i) = \mu(\beta) = g(X_i, \beta). \quad (1)$$

Sin embargo, es importante determinar si alguna variable tiene efecto sobre la variabilidad en la respuesta del proceso. En particular, cuando  $\sigma^2$  es no constante para cada valor de  $X$  se dice que existe heteroscedasticidad. Con la finalidad de reducir la variabilidad de la característica de un producto derivado de un proceso industrial, desde mediados de los 80 a la fecha se ha mostrado considerable interés en modelar la variabilidad, lo que es equivalente a estimar los efectos de dispersión.

Los factores de control  $X$  pueden tener efecto tanto en la media como en la variabilidad, los factores no controlados  $Z$ , denominados de ruido, se considera que afectan la variabilidad, así, la variable de respuesta  $Y$  es afectada por los factores  $X$  y  $Z$ . Completando el modelo 1, la varianza se modela mediante la siguiente función  $Var(Y) = \sigma^2(\theta) = \sigma^2(Z, \theta, \mu(\beta))$ , entonces, la respuesta se expresa por:  $Y = \mu(\beta) + \sigma(\theta)\varepsilon$ .

Se plantean varias situaciones de interés estadístico para estimar eficientemente los parámetros de los modelos, por ejemplo, cuando se conocen o se desconocen la función  $g$  y la distribución de probabilidad  $\varepsilon$ , o si existe alguna relación entre los factores de control y ruido. El objetivo en este trabajo es hacer una breve descripción de los procedimientos desarrollados en la estimación de los parámetros cuando la función  $g$  y la distribución de probabilidad  $\varepsilon$  son conocidas y existe efecto de dispersión, éstos se ilustran con un ejemplo de la literatura. La justificación de este trabajo es evaluar mediante simulación los métodos aquí descritos ante otras situaciones no estudiadas en estadística. En la siguiente sección se formaliza el planteamiento del efecto de dispersión y se presentan los métodos de estimación, a continuación se realiza la comparación de éstos mediante un ejemplo.

## 2 Efectos de Dispersion

Considere que la variable  $Y$  describe la característica de un producto, ésta se ve afectada en su valor promedio y su variabilidad debido la influencia de covariables. La revisión que se hará en este apartado, se establece mediante el planteamiento propuesto por Davidian y Carroll (1987), su proposición formaliza el modelo de regresión expresado por:

$$Y^\lambda = f^\lambda(X, \beta) + h(u_i)\varepsilon, \quad (2)$$

tal que para  $\lambda = 1$  (este es el caso de interés en el presente trabajo), la esperanza de las respuestas se plantean en el modelo 1, y la varianza por

$$\text{Var}(Y_i) = h^2(u_i) = \sigma^2 g^2(Z_i, f(X_i, \beta), \theta), \quad (3)$$

donde  $f$  y  $g$  son funciones conocidas. Conviene puntualizar que la  $\text{Var}(Y_i)$  depende de dos tipos de covariables, estas son  $Z_i$  y  $X_i$ . Diferentes expresiones o connotaciones sobre la  $\text{Var}(Y_i)$  han dado lugar a diferentes procedimientos estadísticos. Davidian y Carroll (1987) discuten varios procedimientos para la estimación de la función varianza. Una descripción con más detalle de estos procedimientos la presentan Carroll y Ruppert (1988).

Logothetis (1990), escribe la  $\text{Var}(Y_i) = f(X_i, \beta)^\theta$ , tomando el logaritmo de esta expresión, realiza la siguiente regresión :  $\log(\text{Var}(Y_i)) = \log \phi + \theta \log f(X, \beta) + \omega$ , donde  $\omega$  sigue una distribución de probabilidad, el valor estimado de  $\theta$  permite usarlo como  $\lambda$ , en la expresión (2).  $\beta$  se estima por mínimos cuadrados (MC), con los datos iniciales. Con el valor de  $\lambda$  se transforman las observaciones originales, la finalidad de este planteamiento es eliminar la variabilidad. Cabe observar que la varianza es proporcional a la media de la respuesta,  $\log \phi$  es una constante, y considerarla como tal, provoca inconsistencia en los estimadores de  $\theta$ , (Engel, 1992).

Se toma el logaritmo en cada observación  $(Y_i - \hat{Y})^2/(n - 1)$  como nueva respuesta, la estimación del parámetro que indica el efecto de dispersión viene dado por el modelo de regresión,

$$\log \left( \frac{(Y_i - \hat{Y})^2}{n - 1} \right) = \log \sigma^2 + \log g^2(Z_i, f(X_i, \beta), \theta) + \omega, \quad (4)$$

la función  $g^2$ , se expresa por  $g^2(Z_i, f(X_i, \beta), \theta) = \exp(\theta X)$ , entonces  $\theta$  se estima por MC, en este caso la función no depende de  $Z$ . Donde  $\omega$  es una variable aleatoria que sigue una distribución de probabilidad, si la distribución de  $Y_i$  es una normal,  $\omega$  es el log de una  $\chi^2$ . En este sentido se sabe que  $\beta$  es el parámetro de localización y  $\theta$  el parámetro de dispersión. La estimación de éstos sigue así, primero se estima el parámetro  $\beta$  en (1) por MC, y luego se estima  $\theta$  en (4) análogamente por MC. A este procedimiento desarrollado por Harvey (1976), se le conoce por la estimación en dos etapas.

Una propuesta distinta para analizar efectos de localización y dispersión es mediante modelos lineales generalizados, las ideas centrales sobre estos modelos aparecen en

McCullagh y Nelder (1989), considerando los modelos siguientes:  $E(Y_i) = f(X_i, \beta) = \mu_i$ , y  $Var(Y_i) = \sigma^2 V(\mu_i)$ , ellos proponen la estimación mediante la función de Cuasi-Verosimilitud (CV), se puede observar que la varianza depende de la media, este planteamiento es un caso particular de los modelos (1) y (3). La generalización de este modelo la estudian Nelder y Pregibon (1987), escriben la varianza de la respuesta como:

$$Var(Y_i) = \phi_i V(\mu_i) \quad y \quad g(\phi_i) = Z_i \theta, \quad (5)$$

en este planteamiento  $\phi_i$  dependen de otras covariables  $Z_i$ . Ellos proponen el procedimiento de Cuasi-Verosimilitud Extendida (CVE) para estimar los parámetros, éste permite proponer diferentes comparaciones de las componentes en un modelo lineal generalizado.

Considerando estas ideas, Engel (1992) desarrolla un algoritmo que mejora al procedimiento de Logothetis, donde las diferencias centrales son considerar  $\log \phi_i$  como un intercepto no constante cuando se estima  $\theta$  usando la expresión (4) y  $\gamma$  por  $\log \phi_i = Z_i \gamma$ . La segunda es estimar por mínimos cuadrados ponderados (MCP)  $\beta$ , donde los pesos son  $w_i^{-1} = \hat{\phi}_i V(\mu_i, \hat{\beta})$ .

## 2.1 Métodos propuestos

A continuación se hace una breve descripción de algunos de los métodos propuestos para estimar los parámetros de los modelos 1 y 3. En este caso los métodos están basados en ponderaciones. El algoritmo a seguir es el de mínimos cuadrados generalizados, el procedimiento se describe en Carroll y Ruppert (1988), la parte fundamental del algoritmo, es calcular los pesos  $\hat{w}_i = 1/g^2(\mu_i(\hat{\beta}_*), z_i, \hat{\theta})$  los cuales permiten reestimar iterativamente hasta la convergencia los parámetros del modelo, donde  $\hat{\beta}_*$  es una estimación preliminar por MC y  $\theta$  es estimado por medio de la solución de la ecuación de la forma :  $0 = \sum H(y_i, z_i, \mu_i(\hat{\beta}_*), \theta)$ .

Las diferencias entre los métodos se encuentran básicamente en la forma de las ecuaciones a resolver planteadas por la función  $H$  y el método de estimación de  $\hat{\beta}_*$ , bajo el supuesto de que el modelo para la varianza de los datos es de la forma:  $Var(y_i) = \exp(\theta_0 + \theta_1 x_i)$ .

### Estimación por mínimos cuadrados ponderados MCP(sin repeticiones)

Este proporciona los estimadores de los parámetros  $\beta$  y  $\theta$  aplicando el algoritmo anteriormente descrito.

### Estimación por Máxima Verosimilitud MV

Se plantea la función de verosimilitud:

$$\log L = l(\beta, \theta, \sigma) = -\frac{1}{2} \sum (\log\{2\pi V_\theta(Y)\}) - \frac{1}{2} \sum \left( \frac{Y - \mu(\beta)}{V_\theta(Y)} \right)^2 \quad (6)$$

Se obtienen los parámetros  $\beta$  y  $\theta$ , optimizando la expresión anterior donde la varianza y la media están dadas por las expresiones (1) y (3).

### Estimación por Pseudoverosimilitud PV

Se obtiene un estimación previa del parámetro  $\beta$ , por ejemplo por MCP, se sustituye en la ecuación (6) y  $\theta$  se estima por MV.

### Estimación por Cuasi-Verosimilitud Extendida CVE

Los estimadores para  $\beta$  y  $\theta$  se obtienen por MV apartir de la expresión (7), la varianza de  $Y$  se representa por 5.

$$Q^+(y, \mu) = -\frac{1}{2} \sum (\log\{2\pi\phi V_\theta(Y)\} - \frac{1}{2}\phi^{-1}D_\theta(y, \mu)) \quad (7)$$

donde  $D_\theta(y, \mu)$ , es la función de desviancia definida según la distribución de interés, Nelder-Pregibon (1987).

## 3 Comparación

Con la finalidad de mostrar la eficiencia de los métodos en la estimación de los parámetros cuando existe efectos de dispersión, se empleará un ejemplo reportado en Neter, Wasserman y Kutner (1989), ellos ilustran el problema de heteroscedasticidad a través de un conjunto de datos correspondientes a la presión sanguínea diastólica observada en 54 personas con edades entre entre 20 y 60 años. En un sencillo diagrama de dispersión se observa de manera inmediata que conforme aumenta la edad, aumenta también la varianza de los datos.

Los autores hacen la estimación de los parámetros de un modelo de regresión únicamente para la media por el método de MCP. El peso que asignan a cada observación es el recíproco de la varianza estimada de la población a la que ésta pertenece. Para llevar a cabo tal estimación, hacen una partición de los datos en 4 grupos de edad para considerar igual número de puntos diseño con observaciones repetidas. Esta partición puede resultar un tanto subjetiva, aquí se consideran otras particiones y el conjunto de datos completo.

Los procedimientos se programaron en Gauss v3.14, los resultados usando el mismo conjunto de datos se muestran en la Tabla 1.

TABLA 1  
Estimación de los parámetros por diferentes procedimientos

Estim	MCO	MCP <sub>4</sub> *	MCP <sub>8</sub>	MCP <sub>2</sub>	MCPs/r	MV	PV	QVE
$\hat{\beta}_0$	56.157	56.13	56.285	56.099	56.113	55.846	55.847	55.958
$\hat{\beta}_1$	.58	.587	.568	.585	.5812	.589	.589	.5857
$\hat{\theta}_0$					.8989	1.053	1.059	1.055
$\hat{\theta}_1$					.0289	.0715	.0714	.0715
SCE	3450.4	42.48	48.386	40.076	1517.06	53.999	53.986	53.998

\* Se armaron 4 grupos para estimar la varianza de los datos. En las dos columnas siguientes, el número de particiones fueron 8 y 2 respectivamente.

Cabe observar que también, se hizo una transformación Box-Cox para homogeneizar las varianzas.  $\lambda = -2$  es el valor que minimiza la  $SCE=2843.95$ , sin embargo, el estimador de  $\beta_0$  dio un valor demasiado alto.

El análisis de los residuos muestra que para los procedimientos MCO, MCP<sub>8</sub> y MCP<sub>2</sub> se conserva la forma del “embudo” haciendo evidente la falta de homoscedasticidad. En el resto de los casos, este problema se corrigió.

En una breve conclusión, se puede observar una notable mejoría en la precisión de la estimación, esta situación ha permitido plantear con mayor detalle la evaluación de los métodos aquí discutidos (ver Moreno, 1998).

## Referencias

- Davidan M. and Carroll R.J. (1987). Variance Function Estimation *JASA*, 82, 1079-1091.
- Carroll y Ruppert (1988). *Transformation and Weighting in Regression*. London: Chapman and Hall.
- Engel J. (1992). Modelling Variation in Industrial Experiments. *Appl. Statist.* 41, 579-593.
- Gauss System Version 3.11 (1990-1995). Washington: Aptech Systems.
- Logothetis N. (1990). Box-Cox Transformations and the Taguchi Method. *Appl. Statist.* 39, 31-48.
- Harvey A.C.(1976). Estimating Regression Models with Multiplicative Heteroscedasticity. *Econometrica*, 44, 461-465.
- McCullagh P. and Nelder J.S. (1988) *Generalized Linear Models*. (2nd. ed.) London: Chapman and Hall.
- Moreno, M.H. (1998). Modelación de la Varianza en Procesos Industriales. *Reporte de Tesis*.
- Nelder J.A. and Pregibon D. (1987). An Extended Quasi-likelihood Function. *Biometrika*, 74, 221-232.
- Neter J., Wasserman W., Kutner M.H. (1995). *Applied Linear Statistical Models*. Illinois: Irwin

# Classification Using Graphical Models

Guillermina Eslava y Leticia Cañedo  
*IIMAS, UNAM*      *UACPyP, UNAM*

## 1 Introduction

It is known that in many cases Linear discriminant analysis is better than quadratic discriminant analysis particularly with small data sets. This mainly because the number of parameters to be estimated when dealing with  $p$  variables is  $p$  times the number of groups for the group means and only one common covariance matrix in the linear case, whereas in the quadratic function apart from the group means, it is necessary to estimate one covariance matrix for each group. The question is whether the parameter parsimony by using Graphical gaussian models improves the precision and has lower rates of misclassification than the quadratic function; partly because the number of parameters to be estimated in using a Graphical model is lower than in the case of the quadratic discriminant function. We do a simulation study of a particular case to show that it is the case.

## 2 The model

Consider the multinormal density function

$$f_i(\mathbf{x}) = (2\pi)^{-p/2} |\Sigma_i|^{-1/2} \exp[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)' \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)] \quad (1)$$

$i = 1, \dots, g$ , as the density function modelling each of the  $g$  populations. Assuming equal prior probabilities for each group, and equal cost of misallocation, the discriminant score for the  $i$ th population is computed as  $S_i = \log f_i(\mathbf{x})$ , an observation  $\mathbf{x}$  is assigned to the population for which the discriminant score is highest. With only two populations or groups, we assign  $\mathbf{x}$  to group 1 if  $S_1 - S_2 = Q > 0$ .

$Q$  is called the discriminant function,

$$\begin{aligned} Q(\mathbf{x}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2) &= \frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|} + \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_2)' \Sigma_2^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) \\ &\quad - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_1)' \Sigma_1^{-1} (\mathbf{x} - \boldsymbol{\mu}_1). \end{aligned}$$

Rearranging terms,  $Q$  can be written as the quadratic form

$$Q(\mathbf{x}, \mu_1, \mu_2, \Sigma_1, \Sigma_2) = \mathbf{x}' \mathbf{A} \mathbf{x} + \mathbf{B} \mathbf{x} + \mathbf{C},$$

where  $\mathbf{A} = \frac{1}{2}(\Sigma_2^{-1} - \Sigma_1^{-1})$ ,  $\mathbf{B} = \mu_1' \Sigma_1^{-1} - \mu_2' \Sigma_2^{-1}$ , and  $\mathbf{C} = \frac{1}{2}(\mu_2' \Sigma_2^{-1} \mu_2 - \mu_1' \Sigma_1^{-1} \mu_1) + \frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|}$ .

Under the assumption of equal covariance matrices for group 1 and 2, the quadratic term vanishes and the discriminant function becomes linear,

$$L(\mathbf{x}, \mu_1, \mu_2, \Sigma) = (\mu_1 - \mu_2)' \Sigma^{-1} \mathbf{x} + \frac{1}{2}(\mu_2' \Sigma^{-1} \mu_2 - \mu_1' \Sigma^{-1} \mu_1). \quad (2)$$

Given a sample  $\mathbf{x}_1, \dots, \mathbf{x}_{n_i}$  from each group  $i$ ,  $i = 1, 2$ , the maximum likelihood estimators are calculated as follows

$$\hat{\mu}_i = \bar{\mathbf{x}}_i, \quad \hat{\Sigma}_i = \frac{(\mathbf{x} - \bar{\mathbf{x}}_i)(\mathbf{x} - \bar{\mathbf{x}}_i)'}{n_i}, \quad \hat{\Sigma} = \frac{n_1 \hat{\Sigma}_1 + n_2 \hat{\Sigma}_2}{n_1 + n_2}$$

(Anderson, 1984, p63).  $\hat{\mu}_i$  is an unbiased estimator, but  $\hat{\Sigma}_i$  is biased; unbiased estimators for the covariance matrices are the following

$$S_i = \frac{(\mathbf{x} - \bar{\mathbf{x}}_i)(\mathbf{x} - \bar{\mathbf{x}}_i)'}{n_i - 1}, \quad S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

In the presence of conditional independencies within groups, we are interested in investigating how the quadratic discriminant function  $Q$  compares in terms of misclassification rates when using a) the unbiased estimator  $S_i$  and when using, b) the gaussian graphical maximum likelihood estimator  $\tilde{\Sigma}_i$ ,  $i = 1, 2$ .

The gaussian graphical maximum likelihood estimators are in general obtained by solving numerically a system of simultaneous equations (Lauritzen, 1996, p 133). However, when the structure of  $\Sigma^{-1}$  in (1) induces a decomposable graphical model, the estimators can be calculated analytically in a closed expression (see Lauritzen, 1996 p 145). In the non-decomposable case there are two algorithms already implemented to compute maximum likelihood estimators : i) the iterative proportional scaling IPS, described in Frydenberg and Edwards (1989), and in a modified version implemented in the program MIM (Edwards, 1993); and ii) the algorithm presented by Wermuth and Scheidt (1977). Although a version of the IPS algorithm is available in MIM, it is not feasible to combine it with a simulation program. The implementation of one of the two algorithms could be done to estimate the concentration matrix in the nondecomposable case. For the present study we consider a decomposable model in seven dimensions. Take the following concentration matrix

$$\Sigma^{-1}(\rho) = \begin{pmatrix} 1 & -\rho & 0 & 0 \\ -\rho & 1 + \rho^2 & -\rho & 0 \\ 0 & \cdots & -\rho & 1 + \rho^2 & -\rho \\ 0 & \cdots & 0 & -\rho & 1 \end{pmatrix}$$

Choose concentration matrices  $\Sigma_1^{-1}$  and  $\Sigma_2^{-1}$  as:  $\Sigma_1^{-1} = \Sigma^{-1}(\rho_1)$ ,  $\Sigma_2^{-1} = \alpha \Sigma^{-1}(\rho_2)$ . As a measure of the difference between the two matrices we take

$$\text{trace}((\Sigma_1^{-1} - \Sigma_2^{-1})'(\Sigma_1^{-1} - \Sigma_2^{-1})) = \sum (\sigma_1^{ij} - \sigma_2^{ij})^2,$$

where  $\Sigma_k^{-1} = \{\sigma_k^{ij}\}$ ,  $k = 1, 2$ ;  $i, j = 1, \dots, p$ .

And as difference of the two populations, when  $\Sigma_1 = \Sigma_2 = \Sigma$  take the generalized distance  $D$ , with  $D^2 = (\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2)$ .

In the case where the group means are  $\mu_1 = (0, \dots, 0)$ , and  $\mu_2 = (d, 0, \dots, 0)$ ,  $D^2 = d^2$ .

## Simulations

There are various situations to consider in the simulation, we consider the model described by the graph 1 presented in the previous section and restrict to the following cases. Dimensionality of  $p = 7$ ; sample size of  $n_1 = 40$  for group one and  $n_2 = 80$  for group two; means  $\mu_1 = (0, \dots, 0)$  and  $\mu_2 = (d, 0, \dots, 0)$  with  $d = 0.75, 1$ , and  $1.5$ ;  $\alpha = 1$ .

For each comparison we perform 1000 times the following steps: i) Generate  $n_1$  random points from a multinormal distribution  $N(0, \Sigma_1)$  and  $n_2$  from  $N(\mu, \Sigma_2)$ . ii) Compute two sets of coefficients for the quadratic function (2), the first by using  $S_i$ , and the second by using  $\tilde{\Sigma}_i$ ,  $i = 1, 2$ . iii) Generate another sample as in i). iv) Classify the sample with the quadratic function and calculate the error rate for each set of coefficients.

The error rate of misclassification is calculated simply by the proportion of observations misclassified.

The simulation study was done with a program in fortran 77, in a Sun computer. The subroutines to generate the random points, *Boxnrm* and *Unif*, were taken from Bratley *et al.* (1987, p. 311 and 332); and the one to perform a Cholesky decomposition and inverse of a symmetric matrix, *choldc*, was taken from Press *et al.* (1992, p. 90). Numerical results are presented in Table 1.

*Table 1. Dimension 7. Mean values of the error rates based on 1000 simulations. Computations of the discriminant functions were based on samples of sizes  $n_1 = 40$ ,  $n_2 = 80$ ; and misallocation rates were computed from a different samples of equal sizes.  $\Sigma_1^{-1}$  equal*

to  $\Sigma_2^{-1}$  with zero values except in the three central diagonals. Population means are at a distance  $d$  on the first axis.

Function	Percentage of misclassification			
$\alpha$	1			
$\rho_1, \rho_2$	.3, -.3	.5, -.5	.8, -.8	.8, -.3
$\sum(\sigma_1^{ij} - \sigma_2^{ij})^2$	4.3	12.0	30.7	16.0
$d = 1.5$				
$Q(S_1, S_2)$	0.196	0.119	0.041	0.089
$Q(\tilde{\Sigma}_1, \tilde{\Sigma}_2)$	0.182	0.110	0.038	0.082
$L(S_1, S_2)$	0.250	0.262	0.287	0.268
$d = 1.0$				
$Q(S_1, S_2)$	0.242	0.144	0.045	0.104
$Q(\tilde{\Sigma}_1, \tilde{\Sigma}_2)$	0.227	0.133	0.042	0.096
$L(S_1, S_2)$	0.332	0.341	0.356	0.346
$d = .75$				
$Q(S_1, S_2)$	0.267	0.151	0.047	0.109
$Q(\tilde{\Sigma}_1, \tilde{\Sigma}_2)$	0.251	0.141	0.044	0.100
$L(S_1, S_2)$	0.381	0.386	0.395	0.386

### 3 Comments and Conclusions

From the numerical results reported in table 1, we observe the following. The percentage of misclassification using  $Q(\tilde{\Sigma}_1, \tilde{\Sigma}_2)$  is systematically lower than the one using  $Q(S_1, S_2)$ , though this difference is rather small.

The percentage of misclassification using  $Q(S_1, S_2)$  is always smaller than the one using  $L(S_1, S_2)$ .

These results are in theory expected, what is interesting to observe is the following.

a) Using the crossproduct estimators,  $S_1, S_2$ , and  $\Sigma_1 \neq \Sigma_2$  it is clear that a quadratic function in theory will better discriminate the two populations. However in practice, in spite of the fact that we are working in dimension seven with sample sizes relatively small,  $n_1 = 40, n_2 = 80$ , the quadratic function is still better than the linear function. Notice in table 2, the number of parameters estimated for  $Q(S_1, S_2)$  is 28 more than for  $L(S_1, S_2)$ .

b) In the quadratic function, one should expect that the parsimony of parameters by using the graphical model estimators  $\tilde{\Sigma}_1, \tilde{\Sigma}_2$  instead of  $S_1, S_2$ , that means 40 parameters instead of 70, is reflected in a lower misallocation rate, and it is in fact the observed case but the difference is very small.

*Table 2. Number of estimated parameters in the discriminant functions for two groups in dimension seven.*

$Q(S_1, S_2)$	$2p$	$p(p+1)$	$p^2 + 3p = 70$
$Q(\tilde{\Sigma}_1, \tilde{\Sigma}_2)$	$2p$	$4p - 2$	$6p - 2 = 40$
$L(S_1, S_2)$	$2p$	$\frac{p(p+1)}{2}$	$\frac{p^2+5p}{2} = 42$

The simulation study illustrate the use of graphical models in discriminant analysis and gives estimators for the gain obtained in terms of percentage of misclassification for a particular case. For the linear case  $L(S_1, S_2)$ , we have used  $S = \frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2}$  as estimator of the common covariance dispersion matrix, it can also be used a graphical model estimator as follows  $\tilde{S} = \frac{(n_1-1)\tilde{\Sigma}_1+(n_2-1)\tilde{\Sigma}_2}{n_1+n_2-2}$ . This means that the same conditional independencies are present in each group. Finally, it should be illustrative to present the performance of the use of  $Q(S_1, S_2)$ ,  $L(S_1, S_2)$ ,  $Q(\tilde{\Sigma}_1, \tilde{\Sigma}_2)$  with a genuine data set. Some work is being done by the author in this line.

## References

- Anderson T.W. (1984). *An introduction to multivariate statistical analysis*. Second edition. New York: Wiley.
- Bratley P., Bennett L. F., and Schrage L. E. (1987). A guide to simulation. Second edition. New York: Springer-verlag.
- Edwards D. (1993). Graphical modelling with MIM 2.1, Hypergraph Software, Bymarken 38, DK-4000 Roskilde, Denmark.
- Frydenberg, M. and Edwards, D. (1988). A modified iterative proportional scaling algorithm for estimation in regular exponential families, *Computational Statistics and Data Analysis*, **8**, 143-153.
- Lauritzen, L. S. (1996). *Graphical Models*. Oxford: University Press.
- Press W.H., Teukolsky S.A., Vetterling W.T., and Flannery B.P. (1992). Numerical recipes. Cambridge: University Press. Second esdition.
- Wermuth, N. and Scheidt, E. (1977). Fitting a covariance selection model to a matrix. Algorithm AS105, *Applied Statistics*, **26**, 88-92.

# Clustering based on rules *versus* Knowledge Discovery of Data: An Application to Astronomy<sup>1</sup>

Karina Gibert & Ulises Cortés

*Department of Statistics*      *Departament of Software.*  
*and Operation Research.*

*Universitat Politècnica de Catalunya. Barcelona, SPAIN.*

## 1 Introduction

It is clear that nowadays analysis of complex systems is an important handicap either for Statistics, Artificial Intelligence, Information Systems... In real applications, it is usual to work with *ill-structured domains (ISD)* (see [6] for a characterization) as sea sponges[8], galaxies... Actually, *ISD* refers to complex systems where the consensus among experts is weak — or even non-existent; where experts use to have some *prior knowledge* on their structure — which should be taken into account; where *non-homogeneous* data bases, with both qualitative and quantitative variables become very common.

Describing the structure or obtaining knowledge of complex systems is known as a difficult task (clustering give bad performances, knowledge-based systems (*KBS*) give low predictive capacity...). Combination of Data Analysis techniques (like clustering), inductive learning (Knowledge-based systems), management of data bases and multidimensional graphical representation must produce benefits on this line.

*Clustering based on rules (CBR)* is a methodology developed with the aim finding the structure on *ISDs*, giving better performance than traditional clustering algorithms or *KBS* approach. A combination of clustering and inductive learning is focussed to the problem of finding and interpreting special patterns (or concepts) from large data bases, in order to extract useful knowledge to represent real-world domains. Actually, *CBR* can be seen as a process of building a knowledge model for a given domain. That is why it is also connected with Knowledge Discovery of Data (*KDD*) and Data Mining (*DM*) [3].

The scope of this paper is to present the methodology as well as to show how *CBR* fits in the context of *KDD*. First of all, the methodology is presented in section §2; the connection points with *KDD* are emphasized. Although *CBR* has been also used in other applications, with real data and great amounts of objects, a simulated study is presented

---

<sup>1</sup>This research has been partially supported by the project TIC'96-0878

in §3 for clarity. The last section presents some conclusions and future work.

## 2 The Methodology: Clustering Based on Rules

In this section, *clustering based on rules* is described. As most *KDD* systems, it combines prior knowledge from the expert with a data mining method (automatic clustering). It is an iterative and interactive process, structured in two major phases which finally organize the objects into a set of classes that are presumed to be *interpretable*: initially, there is a process of acquisition of the available background knowledge **even if it is not a complete definition of the domain**, followed by the clustering process *strictu sensu*. On the other hand, this methodology helps the user to explicit his prior knowledge relevant to the problem.

The main idea is to allow the user to introduce logic *constraints* — which may be based on *semantic* arguments... — on the formation of classes. Therefore, the conditions imposed by the expert induce a sort of *super-structure* on the domain. Clustering will be performed *within* this structure. Finally, all the elements are integrated altogether in a global structure. Hierarchical clustering is especially suited for our purposes, mainly considering that the expert can provide heterogeneous knowledge, *i.e.* very specific knowledge of small parts of the domain, together with more general knowledge about other parts.

At the end of this process, the system has acquired the knowledge needed to organize the domain, and the expert has succeeded in making explicit his knowledge in a relatively friendly way. In [6] there is a detailed description of the methodology. Here we provide only a general description:

- Built a Knowledge Base (*KB*) with logic rules provided by the expert (or experts). Only available knowledge is collected, even if it is a partial description of the domain.
- Calculate the *partition induced by the rules* on the domain (put in a *residual class* those objects which either don't satisfy any rule or satisfy contradictory rules).
- For each element of that partition perform the hierarchical clustering and build a prototype. To deal with heterogeneous data matrices significant work has been required on two specific points of the clustering: class representation and distance between individuals. Details on that are introduced in [7] and, including a definition of a new family of metrics that can measure distances with messy data.
- Cluster the prototypes together with the elements of the residual class to obtain a unique hierarchy.
- Use additional interpretation-oriented tools to study the *meaning* of the results. Reformulate *KB* and iterate if needed. Among these tools, a method to identify relevant

variables for a certain class was proposed. Also, an index  $\delta(\mathcal{P}_1, \mathcal{P}_2) \in [0, 1]$  to evaluate the differences between two classifications was defined. Significance test on that index is actually in progress.

Especially good results are obtained when analyzing data from *ISD*. Taking into account that Fayyad defines a *KDD* process as the “*overall process of finding and interpreting patterns from data, typically interactive and iterative, involving repeated application of specific data mining methods or algorithms and the interpretation of the patterns generated by these algorithms*”[3], it is clear that *clustering based on rules* fits closely this definition, and that the data mining technique is, in our case, the clustering method. Following Fayyad, two important key points of *KDD* are: *i*) using domain knowledge and *ii*) domain characterization. It can be seen that those elements take important part in the methodology presented here. This methodology has been successfully implemented in a system called **Klass** [8], [7], and applied to very different domains (sea sponges [8], stars of Milky Way, thyroid tests...). In the following sections, details on one of those applications are presented.

### 3 An Application in Astronomy

In collaboration with a team of experts, we are studying star populations. First, tests with simulated data were done. This will enable a better evaluation of the system performance, since the classes to be distinguished were previously known.

The study of synthetic stars gives to **Klass** the role of a tool oriented towards the acquisition of background knowledge (from an AI point of view, this is equivalent to a supervised learning process). This knowledge is the input to the second part of this research (currently in progress): a study on a sample of stars taken from the **Hipparcos**<sup>2</sup> input catalogue, where the real class of the stars is unknown. New tests with real data are being performed at the moment using as background knowledge the set of rules obtained here.

Both simulated and real stars are described by eight variables, all of them directly measurable: *parallax, radial velocity, galactic longitude, galactic latitude, proper motions in galactic longitude and latitude, apparent magnitude and spectral type*.

The goal of the work presented now is to separate two populations — *halo* and *disk* of the galaxy — using the original data matrix. To do this, the *CBR* methodology was applied.

In this paper, we emphasize the cognitive aspects, including the real experts’ opinions. The main idea is to show how *CBR* can help the experts to make their knowledge about the domain structure explicit and real classes can be finally recognized.

---

<sup>2</sup>**Hipparcos** is a satellite. Its input catalogue is the most reliable at present.

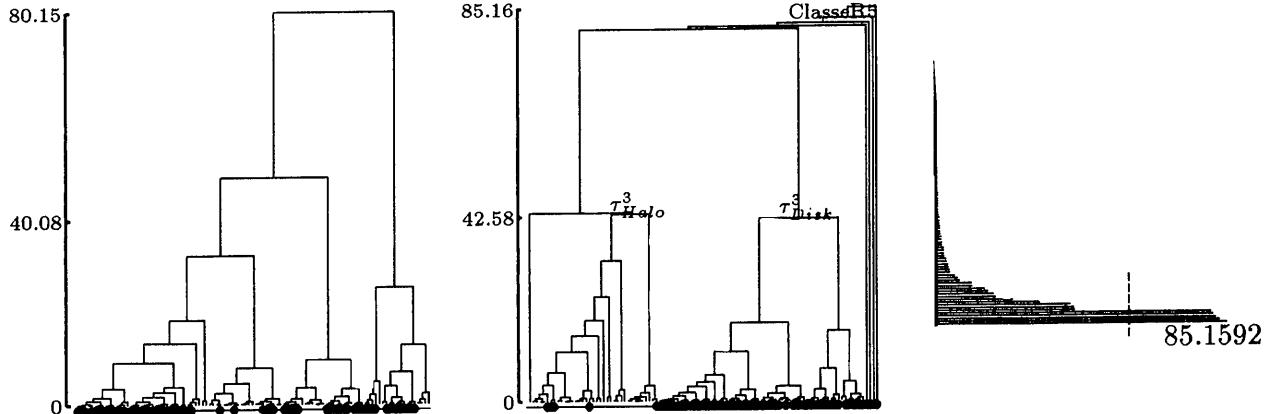


Figure 1: (left) without rules; (center) with rules (final step);(right) aggregation level diagram.

### 3.1 Study

In the classical clustering (without rules, see figure 3.1 (left), where *disk* stars are indicated with  $\bullet$ , whereas *halo* stars are indicated with  $-$ ), a 0.58% of stars were well-classified. As usual when dealing with *ISD*, this seems a random clustering.

The experts were asked to study the results and to provide some rules. They provided rules referring to transformations on the observed variables: the rule  $r_1 : (|W| \geq 40) \rightarrow Halo$  ( $W$  being the vertical component of the velocity, which is not directly measurable) express that *halo* stars have a movement *across* the galactic plane (while *disk* stars describe rotations *inside* it). The partition induced by the *KB*  $\{r_1\}$  on the stars has only one class, namely *halo*, with 15 objects. For this class, a description is produced  $(\bar{\tau}^1_{Halo})$ . The rest of objects formed the residual class  $C_0^1$ . Then, a new clustering was performed on the set  $C_0^1 \cup \{\bar{\tau}^1_{Halo}\}$ , giving mass 15 to  $\bar{\tau}^1_{Halo}$ . In the final hierarchy, well-classified stars increased to the 0.57%. Although the rule is correctly selecting the *halo* stars (there are no *disk* stars satisfying  $r_1$ ), it is not powerful enough to distinguish the two classes.

From the experts opinion, Figure 3.1 (left) suggests that information about the *disk* stars needs to be included in the *KB*. Inspecting the objects that satisfied contradictory rules in iteration 2, the experts remembered<sup>3</sup> that the quickest stars belong to *halo*, while the slowest stars belong to the *disk*. In terms of velocities: **If the velocity module is high then the star belongs to halo; If the velocity module is low then the star belongs to disk.** The rules were:  $r_{18} : (\sqrt{\bar{U}^2 + \bar{V}^2 + \bar{W}^2} > 75) \rightarrow halo$ ;  $r_{19} : (\sqrt{\bar{U}^2 + \bar{V}^2 + \bar{W}^2} < 60) \rightarrow disk$ , where  $\bar{U}, \bar{V}, \bar{W}$  are the velocity coordinates with respect to the Sun velocity.

---

<sup>3</sup>This illustrates that *CBR* contributes to *make explicit* the expert's *implicit* knowledge.

Class	Halo		Disk		E61-1	E58-1	E13-1	E7-1
Var.	$\bar{x}$	s	$\bar{x}$	s				
PAR	5.5139	5.0984	14.5932	14.1032	6.0	6.0	2.0	34.0
VR	8.1486	92.5322	-3.6559	18.0381	62.5	-27.9	4.3	-0.8
GALON	207.6789	117.7034	209.8683	88.0957	163.64328.07	72.65	27.14	
GALAT	-12.5908	44.8745	-1.6064	33.5009	-7.97	9.26	-66.51	-63.9
MLCB	0.0191	0.1369	0.0022	0.06532	0.004	-0.05	0.013	-0.073
MB	-0.024375	0.15212	-0.0129	0.0846	0.023	-0.025	0.017	-0.33
MV	7.2176	0.9898	7.2237	1.3343	7.63	6.935	6.987	7.928
SP	A0 = 1/72 A4-A9 = 1/36 F0-F4 = 1/36 G0-G7 = 1/36 G8-G9 = 1/72	K0-K4 = 23/72 K5-K6 = 41/72	B0-B4 = 11/59 B5-B9 = 2/59 A0 = 7/59 A2-A3 = 8/59 A4-A9 = 3/59	F0-F4 = 5/59 F5-F9 = 9/59 G0-G7 = 8/59 G8-G9 = 1/59 K0-K4 = 5/59	A0	K0-K4	B0-B4	G0-G7
	O7B0B5 A2A4F0F5 G0G8K0K5K7M0M5 O9B4B9A0A3A9F4F9 G7G9K4K6K9M4 +		O7B0B5 A2A4F0F5 G0G8K0K5K7M0M5 O9B4B9A0A3A9F4F9 G7G9K4K6K9M4 +					
n	37		59		1	1	1	1

Table 1. Class description (6-class cut).

The set  $\{r_{18}, r_{19}\}$  was used as  $KB$  in the third iteration. The resulting dendrogramme is in figure 3.1 (*center*). Figure 3.1 (*right*) clearly suggests a 6-classes cut. A significant improvement of well-classified stars (0.88) was found.

Representative of the classes are shown in table 3.1. From the experts point of view, characterization of the partition is clear: Stars classified as *halo* have short parallaxes, high radial velocities and old spectral types (class of old stars with large distances to Sun and with movements not contained in the galactic plane); the class of *disk* stars is composed of stars with long parallaxes (nearer stars), lower radial velocities and major concentration in shortest spectral types (younger stars that have not left the galactic plane yet). Finally, it is proved that enlarging the rules set with the union of all the rules used in previous steps doesn't change the results.

## 4 Conclusions and Future Work

In this paper, the methodology of *clustering based on rules (CBR)* is presented. It successfully combines Artificial Intelligence techniques with Statistical methods, for finding the structure of *ill-structured* domains (see §2).

Clear connections between *clustering based on rules* and *KDD* are shown along the paper: taking into account prior knowledge, applying a repeated Data Mining technique (clustering), including tools interpretation-oriented to help the user to find the classes *meaning*... are some of the features that remain common between *KDD* and *CBR*.

*CBR* uses the background expert's knowledge — which classifiers themselves are unable to capture — to guide the clustering process. The use of rules in the clustering process con-

tributes (acting as a *semantic* bias) to increase the classification quality (and to decrease the computational cost [5]). In fact, the rules act as selectors that cluster objects which could be considered similar by the experts experience. The resulting classes (and their prototypes) tend to be more meaningful to the expert's eye. In most of the cases, this process helps the expert to make his knowledge about some parts of the domain explicit. When synthetic data is used, the system behaves somewhat like a supervised machine learning system.

Clustering of heterogeneous data (in metric spaces) matrices can be done, without transforming the original matrix, using the mixed metrics defined in [7]. In several applications this metrics has shown a good behaviour when dealing with *ISD*.

In the presented application, interaction with the experts was especially emphasized. In particular the cognitive aspects preceding the construction of the final set of rules (making the relationship among variables explicit). In this case it can be said that the rules with better performance are those expressing qualitative aspects of the studied phenomenon rather than quantitative ones. Also, relationships between variables are more powerful than descriptions of the behaviour of isolated variables. It is also shown that the quality of the results does not depend on the size of the Knowledge Base, but on the expressive power of its components.

**Acknowledgement:** To Dr. Manuel Hernández Pajares, from the *Applied Mathematics and Telematics* Department at **UPC** his collaboration concerned with the specific application presented.

## References

- [1] Gowda, K. C., Diday, E. (1992). Symbolic clustering using a new similarity measure. *IEEE Trans. on systems, man, and cib.*, **22**(2), 368–378.
- [2] Everitt, B. (1974) *Cluster analysis*. London: Heinemann Educational Books Ltd.
- [3] Fayyad, U., Piatetsky-Shapiro, G. Smyth, P. (1996) From Data Mining to Knowledge Discovery: An overview *Advances in KDD& DM* (eds. Fayyad, U. *et al.*) R. AAAI/MIT, 1996.
- [4] Frawley, W. *et al.* (1992). KDD: An overview. *AI Magazine* **14**(3), 57–70.
- [5] Gibert, K. (1996) On the Uses and Costs of CBR. In *COMPSTAT'96*, 265–270. Bna: Springer.
- [6] Gibert, K., Cortés, U. (1998). Clustering based on rules and Knowledge Discovery in ill-structured domains. *Computación y Sistemas, revista americana de computación* **2**(2).

- [7] — (1997). Weighing quant. and qual. variables in clustering methods. *MATHWARE* **4**(3), 251–266.
- [8] — (1994). Combining a KBS with a clustering method for an inductive construction of models. In *LN on Statistics 89*. (eds P. Cheeseman *et al.*), 351 – 360. NY: Springer-Verlag.
- [9] Robin, A., Crézé, M. (1986) Stellar popul. in the Milky.... *Astronomy & Astrophysics*, **157**, 71–90.

# Modelación Gráfica en el Control de la Validez de Construcción de Test Psicológicos

Adalberto González Debén      Jesús E. Sánchez García

*ICIMAF, Cuba*

Ma. Odette Lobato Calleros

*UIA, México*

## 1 Introducción

“Un test es un procedimiento científico rutinario para la investigación de una o más características de personalidad, delimitadas empíricamente, con el objetivo de hacer una afirmación, lo más cuantitativa posible, acerca del grado relativo de expresión de la característica en el individuo”, (Lienert, 1969).

Se espera de un buen test que sea objetivo y que tenga precisión experimental, entendiéndose por esto último que sea confiable y válido.

En este trabajo se propone la utilización de los modelos mixtos de interacción, Edwards (1990), como una vía para el análisis de uno de los tres tipos de validez, esto es: la del constructo.

## 2 El Control de la Validez de un Test

Un test es válido cuando mide efectivamente la característica de personalidad que debe medir. Se han diferenciado tres tipos de validez:

1. De contenido (lógica).
2. Con respecto al criterio (empírica).
3. De construcción (del constructo).

La validez de contenido está dada por el criterio de los expertos que, mediante un examen sistemático de los ítems, aseguran o no que estos comprenden una muestra representativa de la característica que se desea medir. Su uso es común en los tests de conocimientos escolares.

La validez relativa al criterio es la más relevante desde los puntos de vista histórico y práctico ; de hecho, en algunos textos se refieren únicamente a ella como validez. Se trata de la relación entre los puntajes del test con los del criterio y por eso también, a veces, se conoce como empírica. Según el tipo de criterio se puede clasificar en externa o interna (si se trata de un criterio objetivo o de la misma característica medida por otro test) y en concurrente o predictiva (si el criterio se mide simultáneamente o después de un intervalo de tiempo).

En muchas ocasiones la característica que se desea medir es muy compleja y no se puede registrar operacionalmente de manera inmediata. La validez de construcción consiste en la aclaración teórica de qué es lo que mide el test y por lo tanto está más relacionada con la investigación fundamental acerca del verdadero significado psicológico de los resultados.

### **3 El Control de la Validez de Construcción de un Test**

Para la validación del constructo puede ser necesario tanto el análisis lógico como la aplicación de principios experimentales y empírico-estadísticos. En dependencia del grado de elaboración de sus fundamentos teóricos, se pueden utilizar diversas herramientas estadísticas para analizar relaciones entre variables (observables o latentes) en un nivel exploratorio o confirmatorio, a saber: análisis de correlaciones, análisis de regresión, análisis de senderos, análisis de ecuaciones estructurales, análisis factorial, y análisis de estructura de covarianzas.

Todas ellas son útiles para el estudio de modelos que se basan en hipótesis de independencia o en el conocimiento cualitativo acerca de la dirección y magnitud de relaciones de dependencia ; que son los más utilizados en ciencias sociales (Cox y Wermuth ,1995, Jöreskog ,1989), pero no contemplan las hipótesis de independencia condicional (Wermuth y Lauritzen, 1989, Cox y Wermuth, 1995).

### **4 Modelación Gráfica**

La modelación gráfica no es más que una forma de análisis multivariado, que usa grafos de independencia condicional para representar los modelos (Edwards, 1995).

Los modelos gráficos (Lauritzen y Wermuth, 1989, Lauritzen, 1989) son modelos probabilísticos para observaciones aleatorias multivariadas cuya estructura de independencia se caracteriza por un grafo. Sus antecedentes son:

1. El análisis de senderos (Wright, 1921).
2. Los modelos de selección de covarianzas (Dempster, 1972).
3. Los modelos loglineales (Bishop, Fienberg y Holland, 1975).

#### 4. Los conceptos de independencia e independencia condicional (Dawid, 1979).

Entre sus principales atractivos está la posibilidad de tratar de manera conjunta variables continuas y discretas, así como que, en los casos de un solo tipo de variables, coincide con modelos ya conocidos aunque no relacionados anteriormente (Wermuth, 1976). Cuando todas las variables son discretas se reduce a la clase de modelos loglineales gráficos, ampliamente utilizados en las ciencias sociales (Wickens, 1989); en el caso continuo se reduce a los modelos de selección de covarianzas, que han tenido menos aceptación y popularidad (Whittaker, 1990).

Asimismo, y precisamente por la característica de donde proviene su nombre, resaltan sus bondades en cuanto a simplificación e interpretabilidad, dos de las propiedades más apreciadas por los que se enfrentan a la difícil tarea de estudiar relaciones complejas entre variables.

Los modelos jerárquicos de interacción (Edwards, 1990), son una generalización de los modelos gráficos, pues son una combinación de los modelos jerárquicos loglineales y los de selección de covarianzas, también llamados modelos gráficos gaussianos. Pese a que es relativamente reciente, ya existen tres libros de texto que presentan el estado del arte de este tema (Whittaker, 1990, Edwards, 1995 y Lauritzen, 1996).

## 5 Ejemplo

### 5.1 Modelo de las Características del Puesto

El modelo de motivación de las características del puesto (Hackman y Oldham, 1975), establece que la motivación interna del trabajador, depende de tres estados psicológicos críticos: que perciba su puesto como significativo, que se sienta responsable por los resultados de su trabajo y que conozca los resultados de su trabajo.

A su vez, estos estados psicológicos críticos dependen de las características objetivas del puesto de trabajo, también llamadas dimensiones centrales del trabajo: variedad de habilidades, identidad de la tarea, importancia de la tarea, autonomía, y retroalimentación del puesto.

Además de la motivación interna, se incluyen otros dos resultados : satisfacción de la necesidad de autorealización y satisfacción general del trabajo. Por último se considera un grupo de variables moderadoras de estas relaciones: destrezas y habilidades, intensidad de la necesidad de autorealización y satisfacción con los factores del contexto.

### 5.2 Estudios Sobre la Validez del Modelo

La validez de este modelo ha sido revisada por muchos autores en diferentes contextos : Roberts y Glick (1981), Loher *et al.* (1985), Fried y Ferris (1987), Hogan y Martell (1987),

Fortea *et al.* (1994).

No obstante todos estos estudios, queda por dilucidar entre otras interrogantes :

1. La dimensionalidad de las características centrales del trabajo.
2. El carácter interventor de los estados psicológicos críticos.
3. Si realmente hace falta la presencia de los tres estados psicológicos críticos
4. El carácter moderador de las necesidades de crecimiento y la satisfacción con el contexto

En este trabajo, se abordan los dos últimos aspectos, pues los datos provienen de una muestra muy homogénea y no parece muy útil el empleo de la información correspondiente a las características objetivas del trabajo.

### 5.3 Aplicación a maestros de la UIA

Se tomó una muestra de 81 profesores de tiempo completo de la Universidad Iberoamericana, Plantel Santa Fe, México, D.F; 36 que se dedican sólo a la docencia y 45 que, además, realizan investigación. Para este análisis exploratorio se tuvieron en cuenta las variables correspondientes a los tres estados psicológicos críticos y los tres resultados, así como una variable dicotómica para señalar los dos grupos de maestros. En la Tabla 1 se muestran los estadísticos descriptivos de cada una de las variables en ambos grupos.

**TABLA 1**  
Medias y desviaciones (entre paréntesis) de las  
variables en ambos grupos

Variable	Etiqueta	No investigan	Investigan
Significado	A	5.71 (0.88)	5.68 (0.78)
Responsabilidad	B	5.70 (0.64)	5.65 (0.81)
Conocimiento	C	4.98 (1.09)	4.85 (0.97)
Satisf. general	D	5.72 (0.78)	5.69 (0.66)
Motiv. interna	E	5.49 (0.78)	5.23 (0.93)
Satisf. crecimiento	F	5.78 (0.79)	5.69 (0.93)

En las Tablas 2 y 3 se muestran las correlaciones marginales (triángulo inferior) y las correlaciones parciales (triángulo superior).

**TABLA 2**

Correlaciones marginales y parciales (grupo de maestros que no investigan)

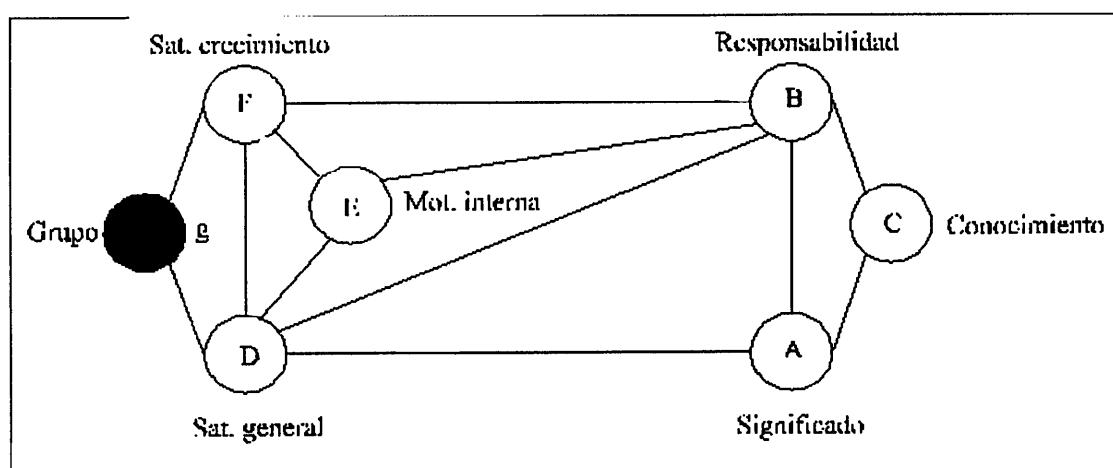
	A	B	C	D	E	F
A	-	0.26	0.35	0.60	0.04	-0.24
B	0.67	-	0.15	0.07	0.48	0.33
C	0.6	0.49	-	-0.09	-0.22	0.18
D	0.79	0.67	0.51	-	-0.08	0.53
E	0.28	0.53	0.07	0.26	-	0.01
F	0.56	0.67	0.48	0.74	0.29	-

**TABLA 3**

Correlaciones marginales y parciales (grupo de maestros que investigan)

	A	B	C	D	E	F
A	-	0.55	0.13	0.09	0.03	0.12
B	0.77	-	0.33	-0.09	0.31	0.08
C	0.63	0.69	-	0.22	0.07	0.14
D	0.46	0.47	0.53	-	0.44	0.25
E	0.5	0.59	0.51	0.58	-	-0.31
F	0.32	0.29	0.33	0.31	0.03	-

Como se rechazó la prueba de homogeneidad de varianzas ( $p=0.0012$ ), se procedió a seleccionar un modelo mixto adecuado. (Ver Gráfica 1).



Gráfica 1

Interpretación:

1. La variable "c" es independiente de las variables "d", "e", "f" y "g" dadas "a" y "b".
2. Las variables "a", "b", "c" y "e" son independientes de "g" dadas "d" y "f".

En la Tabla 4 se muestran los resultados del modelo para toda la muestra y para cada grupo por separado.

**TABLA 4**

Prueba de bondad de ajuste del modelo

	Ambos grupos	No investigan	Investigan
L.R.	4.030	4.480	6.130
p	0.545	0.483	0.294

A manera de resumen se puede afirmar que el estado psicológico "Conocimiento de los Resultados" es el menos importante para explicar los resultados. Asimismo, el hecho de investigar o no está relacionado con la satisfacción; tentativamente, en este contexto particular, pudiera interpretarse como una manifestación de la variable moderadora Necesidades de Crecimiento.

## 6 Conclusiones

La modelación gráfica resulta muy útil en la fase exploratoria de análisis de tests, por cuanto saca a la luz la estructura de independencias condicionales. En este sentido es la única técnica que pone de manifiesto este aspecto importante para el establecimiento de relaciones entre variables. El número de variables no debe ser muy grande, ya que de ser así la representación gráfica puede resultar engorrosa y por tanto dificultarse la interpretación.

## Referencias

- Bishop, Y.M., Fienberg, S. y Holland, P. (1975). *Discrete Multivariate Analysis*. Cambridge: MIT Press.
- Cox, D.R. y Wermuth, N. (1995). *Multivariate dependencies, models, analysis and interpretation*. London: Chapman and Hall.
- Dawid, A.P. (1979). Conditional independence in statistical theory (with discussion). *J. R. Stat. Soc. B*. 41, 1-31
- Dempster, A.P. (1972). Covariance selection. *Biometrics* 28, 157-75

- Edwards,D (1990). Hierarchical interaction models (with discussion) *J.R. Stat. Soc. B* 52, 3-20.
- Edwards, D. (1995). *Introduction to graphical modelling*. New York: Springer-Verlag.
- Fried, Y. y Ferris, G.R. (1987). The validity of the job characteristics model: a review and meta-analysis. *Personnel psychology*, 40, 287-322
- Fortea A., Fuertes, F. y Agost, M.R. (1994). Evaluación del modelo motivacional de las características del puesto a partir de una muestra variada. *Psicología del trabajo y organizaciones*, 10 35-52
- Hackman J.R. y Oldham, G.R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology*, 55, 259-286
- Hogan E.A. y Martell, D.A. (1987). A confirmatory structural equations analysis of the job characteristics model. *Organizational behavior and human decision processes*, 39, 242-263
- Jöreskog, K.G. (1989). Discussion of S.L. Lauritzen's paper Mixed graphical association models. *Scand. J. Statist.* 16, 301-304
- Lauritzen, S.L. (1989). Mixed graphical association models (with discussion). *Scand. J. Statist.* 16, 273-306
- Lauritzen, S.L.(1996). *Graphical models*. Oxford: University Press.
- Lauritzen, S.L. y Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann.Stat.* 17, 31-57
- Lienert, G.A. (1961). *Testaufbau und Testanalyse*. 3. Aufl. Weinheim- Basel: Beltz.
- Loher, B.T., Noe, R.A., Moeller, N.L. y Fitzgerald, M.P. (1985). A meta-analysis of the relation of job characteristics to job satisfaction. *Journal of Applied Psychology*, 70, 280-89
- Roberts, K.H. y Glick, W. (1981). The job characteristics approach to task design: A critical review. *Journal of Applied Psychology*, 66, 193- 217
- Wermuth, N. (1976). Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics* 32, 95-108
- Wermuth, N. y Lauritzen, S.L. (1990). On substantive research hypotheses, conditional independence graphs and graphical chain models (with discussion). *J. R. Stat. Soc. B* 52, 21- 72.

- Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics* New York: Wiley.
- Wickens, T.D. (1989). Multiway contingency tables analysis for the social sciences. Hillsdale: Lawrence Erlbaum Associates.
- Wright, S. (1921). Correlation and causation. *J. Agric. Res.* 20, 557-585.

# Distribuciones de Referencia para Ciertas Familias Exponenciales

E. Gutiérrez-Peña y R. Rueda

*IIMAS, UNAM*

## 1 Introducción

El enfoque Bayesiano requiere de la especificación de una distribución de probabilidad sobre los parámetros desconocidos del modelo a analizar. En ocasiones es conveniente o deseable expresar información inicial vaga sobre los valores de los parámetros a través de distribuciones iniciales no informativas. Entre los métodos propuestos en la literatura para encontrar dichas distribuciones, el de Berger y Bernardo (1992a) ha demostrado ser el más satisfactorio pues distingue entre parámetros de interés y de ruido, además de dar soluciones razonables a problemas controversiales que aparecen por el uso de distribuciones impropias. Sin embargo, la implementación de dicho método dista de ser trivial en el caso multiparamétrico. Bernardo y Ramón (1996) dan una solución particular que, bajo ciertas condiciones, permite encontrar la distribución de referencia de manera sencilla. En este trabajo, se extiende este resultado y se muestra que una clase amplia de familias exponenciales lo satisfacen.

## 2 Distribuciones de Referencia para Parámetros Agrupados

Sea  $X$  una cantidad aleatoria con función de densidad  $p(x|\theta)$ , donde  $\theta \in \Theta \subseteq \mathbb{R}^k$  denota a un parámetro desconocido. Berger y Bernardo (1992a) motivan y describen un algoritmo general para encontrar distribuciones de referencia para  $\theta$ , el cual es usualmente muy difícil de instrumentar en la práctica. Sin embargo, bajo ciertas condiciones, tales como la normalidad asintótica de la distribución final de  $\theta$  (el llamado *caso regular*), dicho algoritmo se simplifica.

Por fortuna, muchos de los modelos utilizados en las aplicaciones satisfacen las condiciones mencionadas anteriormente. Este es el caso, en particular, de la mayoría de los problemas que involucran a las llamadas familias exponenciales.

Sea

$$F = F(\theta) = -E_{x|\theta} \left[ \frac{\partial^2 \log p(x|\theta)}{\partial \theta' \partial \theta} \right]$$

la matriz de información de Fisher para el modelo  $p(x|\theta)$ . Supongamos que  $F$  es invertible y definamos  $S = S(\theta) = F(\theta)^{-1}$ . Supongamos también que  $\theta = (\theta_1, \dots, \theta_k)$  se partitiona en  $m$  grupos de tamaños  $n_1, \dots, n_m$ , respectivamente, i.e.  $\theta = (\theta_{(1)}, \dots, \theta_{(m)})$ , donde  $\theta_{(j)} = (\theta_{N_{j-1}+1}, \dots, \theta_{N_j})$  con  $N_j = n_1 + \dots + n_m$  ( $j = 1, \dots, m$ ).

Esta partición debe hacerse en términos de la importancia inferencial de cada uno de los grupos  $\theta_{(j)}$ . En particular,  $\theta_{(1)}$  debe corresponder al parámetro de interés. Resulta conveniente definir, para  $j = 0, 1, \dots, m$ , a  $\theta_{[j]} = (\theta_{(1)}, \dots, \theta_{(j)})$  y  $\theta_{[\sim j]} = (\theta_{(j+1)}, \dots, \theta_{(m)})$ , con la convención de que  $\theta_{[0]}$  es vacuo y  $\theta_{[\sim 0]} = \theta$ . Denotemos por  $S_j$  a la matriz superior izquierda de dimensión  $N_j \times N_j$  de  $S$  (de manera que  $S_m = S$ ), y definamos a  $h_j = h_j(\theta)$  como la matriz inferior derecha de dimensión  $n_j \times n_j$  de  $H_j = S_j^{-1}$ .

Finalmente, supongamos que  $\Theta = \Theta_{(1)} \times \dots \times \Theta_{(m)}$  y que la distribución final de  $\theta$  es asintóticamente normal con matriz de varianzas-covarianzas  $S(\hat{\theta})$ , donde  $\hat{\theta}$  denota al estimador de máxima verosimilitud para  $\theta$ . A continuación se presenta una versión simplificada del algoritmo de Berger y Bernardo (1992b) para obtener distribuciones de referencia.

1. Defínase

$$\begin{aligned} \pi_m(\theta_{[\sim(m-1)]} | \theta_{[m-1]}) &= \pi_m(\theta_{(m)} | \theta_{[m-1]}) \\ &= \frac{|h_m(\theta)|^{1/2} \mathbf{1}_{\Theta_{(m)}}(\theta_{(m)})}{\int_{\Theta_{(m)}} |h_m(\theta)|^{1/2} d\theta_{(m)}}. \end{aligned}$$

donde  $|h_m(\theta)|$  denota el determinante de la matriz  $h_m(\theta)$ .

2. Para  $j = m-1, m-2, \dots, 1$ , defínase

$$\begin{aligned} \pi_j(\theta_{[\sim(j-1)]} | \theta_{[j-1]}) &= \\ \frac{\pi_{j+1}(\theta_{[\sim j]} | \theta_{[j]}) \exp \left\{ \frac{1}{2} E_j [\log |h_m(\theta)|] \right\} \mathbf{1}_{\Theta_{(m)}}(\theta_{(m)})}{\int_{\Theta_{(j)}} \exp \left\{ \frac{1}{2} E_j [\log |h_m(\theta)|] \right\} d\theta_{(j)}} &, \end{aligned}$$

donde

$$E_j[f(\theta)] = \int_{\Theta_{[\sim j]}} f(\theta) \pi_{j+1}(\theta_{[\sim j]} | \theta_{[j]}) d\theta_{[\sim j]},$$

con  $\Theta_{[\sim j]} = \Theta_{(j+1)} \times \dots \times \Theta_{(m)}$ . En particular, para  $j = 1$ ,  $\pi(\theta) = \pi_1(\theta_{[\sim 0]} | \theta_{[0]})$ .

Si la distribución inicial  $\pi(\theta)$  da lugar a una distribución final propia entonces  $\pi(\theta)$  es la distribución inicial de referencia para  $\theta$ .

*Comentario.* Si alguna de las distribuciones  $\pi_j(\theta_{[\sim(j-1)]}|\theta_{[j-1]})$  ( $j = m, m-1, \dots, 2$ ) es impropia, se requiere una sucesión  $\Theta^1 \subset \Theta^2 \subset \dots$  de subconjuntos compactos de  $\Theta$  tal que

$$\bigcup_{l=1}^{\infty} \Theta^l = \Theta,$$

con  $\Theta^l = \Theta_{(1)}^l \times \dots \times \Theta_{(m)}^l$  para todo  $l = 1, 2, \dots$

Utilizando a  $\Theta^l$  en vez de  $\Theta$  en el algoritmo descrito anteriormente se obtiene una sucesión de distribuciones,  $\pi^l(\theta)$ . La distribución inicial de referencia se define entonces como

$$\pi(\theta) = \lim_{l \rightarrow \infty} \frac{\pi^l(\theta)}{\pi^l(\theta^*)},$$

donde  $\theta^*$  es algún punto en  $\Theta^1$ .

Desafortunadamente, en general  $\pi(\theta)$  depende de la elección particular de la sucesión de compactos. En particular, algunas de estas sucesiones podrían dar lugar a distribuciones de referencia impropias.

El siguiente resultado proporciona una forma de evitar este problema en algunos casos.

*Proposición 1.* En el caso regular, si  $\Theta = \Theta_{(1)} \times \dots \times \Theta_{(m)}$  y los determinantes de las funciones  $h_j(\theta)$  ( $j = 1, \dots, m$ ) se pueden factorizar como

$$|h_i(\theta)| = a_i(\theta_{(i)}) b_i(\theta_{[i-1]}, \theta_{[\sim i]})$$

para algunas funciones positivas  $\{a_i : i = 1, \dots, m\}$  y  $\{b_i : i = 1, \dots, m\}$ , entonces

$$\pi(\theta) \propto \prod_{j=1}^m a_j(\theta_{(j)})^{1/2}$$

independientemente de la sucesión de compactos.

*Demostración.* Notemos primero que

$$\pi_m(\theta_{[\sim(m-1)]}|\theta_{[m-1]}) = \frac{a_m(\theta_{(m)})^{1/2}}{\int_{\Theta_{(m)}} a_m(\theta_{(m)})^{1/2} d\theta_{(m)}}.$$

Ahora, para  $j = m-1, m-2, \dots, 1$ ,

$$\begin{aligned} E_j[\log |h_j(\theta)|] &= \log a_j(\theta_{(j)}) + \\ &\quad \int_{\Theta_{(m)}} \log b_j(\theta_{[j-1]}, \theta_{[\sim j]}) \pi_{j+1}(\theta_{[\sim j]}|\theta_{[j]}) d\theta_{[\sim j]}, \end{aligned}$$

donde el segundo término no dependen de  $\theta_{(j)}$ . Por lo tanto,

$$\pi_j(\theta_{[\sim(j-1)]} | \theta_{[j-1]}) = \frac{\pi_{j+1}(\theta_{[\sim j]} | \theta_{[j]}) a_j(\theta_{(j)})^{1/2}}{\int_{\Theta_{(j)}} a_j(\theta_{(j)})^{1/2} d\theta_{(j)}}.$$

Procediendo inductivamente, se obtiene

$$\pi(\theta) = \pi_1(\theta_{[\sim 0]} | \theta_{[0]}) = \prod_{j=1}^m a_j(\theta_{(j)})^{1/2}.$$

Dada la simplicidad de la Proposición 1, resulta de interés buscar condiciones bajo las cuales dicho resultado puede aplicarse. El resto de este trabajo aborda este problema en el contexto de las familias exponenciales.

### 3 Familias Exponenciales con Cortes

Barndorff-Nielsen (1978, p.50) define el concepto de *corte* en el contexto de la inferencia marginal para un parámetro de interés en presencia de un parámetro de ruido. Sea  $\mathcal{F} = \{p(x|\theta) : \theta \in \Theta\}$  una familia paramétrica de modelos de probabilidad y sea  $s = s(x)$  una estadística. Se dice que  $s$  es un corte si y sólo si existe una parametrización  $\phi = (\phi_1, \phi_2)$  de  $\mathcal{F}$  tal que: (i)  $\Phi = \Phi_1 \times \Phi_2$  y (ii)  $p(x|\theta) = p(s|\phi_1)p(x|\phi_2, s)$ .

Consideremos a una familia exponencial en  $\mathbb{R}^k$ , con función de densidad de la forma

$$p(x|\theta) = B(x) \exp\{\theta' t(x) - M(\theta)\}, \quad \theta \in \Theta, \quad (1)$$

con  $\Theta = \{\theta \in \mathbb{R}^k : \int B(x) \exp\{\theta' t(x)\} dx < \infty\}$ .

Denotemos por

$$\mu = E[t|\theta] = \frac{\partial M(\theta)}{\partial \theta}$$

al parámetro medio y por

$$V(\mu) = \left. \frac{\partial^2 M(\theta)}{\partial \theta' \partial \theta} \right|_{\theta=\theta(\mu)} \quad \mu \in \Omega,$$

a la función de varianza del modelo (1), donde  $\theta(\cdot)$  denota al mapeo inverso de  $\mu(\theta) = \frac{\partial M(\theta)}{\partial \theta}$  y  $\Omega = \mu(\Theta)$ . La pareja  $(V(\cdot), \Omega)$  caracteriza a la familia exponencial (1); ver, por ejemplo, Morris (1982).

Sea  $t = (t_{(1)}, t_{(2)})$  una partición de  $t$ , con  $\dim(t_{(1)}) = n_1$  y  $\dim(t_{(2)}) = n_2 = k - n_1$ . Sean  $\theta = (\theta_{(1)}, \theta_{(2)})$  y  $\mu = (\mu_{(1)}, \mu_{(2)})$  las particiones correspondientes de  $\theta$  y  $\mu$ . Finalmente, sea

$$V(\mu) = \begin{bmatrix} V_{11}(\mu) & V_{12}(\mu) \\ V_{21}(\mu) & V_{22}(\mu) \end{bmatrix}$$

donde  $V_{11}(\mu)$  es la submatriz de dimensión  $n_1 \times n_1$  que corresponde a  $t_{(1)}$ .

Supongamos ahora que  $t_{(1)}$  es un corte de la familia exponencial (1). Entonces el Teorema 3.1 de Barndorff-Nielsen y Koudou (1995) implica que

$$p(t_{(1)}|\mu_{(1)}) = B_1(t_{(1)}) \exp\{t'_{(1)}\theta_{(1)}^*(\mu_{(1)}) - M_1(\theta_{(1)}^*(\mu_{(1)}))\}$$

y

$$p(t_{(2)}|\theta_{(2)}, t_{(1)}) = B_2(t_{(2)}|t_{(1)}) \exp\{t'_{(2)}\theta_{(2)} - M_2(\theta_{(2)}|t_{(1)})\},$$

con  $M_2(\theta_{(2)}|t_{(1)}) = K(\theta_{(2)}) - t(1)'G(\theta_{(2)})$ , para algunas funciones  $\theta_{(1)}^*(\cdot)$ ,  $M_1(\cdot)$ ,  $K(\cdot)$  y  $G(\cdot)$ . Esto a su vez implica que  $V_{11}(\mu) = V_1^*(\mu_{(1)})$  para todo  $\mu \in \Omega$ , donde

$$V_1^*(\mu_{(1)}) = \left. \frac{\partial^2 M_1(\theta_{(1)}^*)}{\partial \theta_{(1)}^{*\prime} \partial \theta_{(1)}^*} \right|_{\theta_{(1)}^* = \theta_{(1)}^*(\mu_{(1)})}$$

Es fácil ver que la matriz de información de Fisher para la *parametrización mixta*  $(\mu_{(1)}, \theta_{(2)})$  está dada por

$$H(\mu_{(1)}, \theta_{(2)}) = \begin{bmatrix} V_1^*(\mu_{(1)})^{-1} & O \\ O & \frac{\partial k(\theta_{(2)})}{\partial \theta_{(2)}'} - \frac{\partial \{g(\theta_{(2)})'\mu_{(1)}\}}{\partial \theta_{(2)}'} \end{bmatrix}$$

donde

$$k(\theta_{(2)}) = \frac{\partial K(\theta_{(2)})}{\partial \theta_{(2)}} \quad \text{y} \quad g(\theta_{(2)}) = \frac{\partial G(\theta_{(2)})'}{\partial \theta_{(2)}}.$$

Bernardo y Smith (1994, Secc. 5.3) han establecido la normalidad asintótica de la distribución final del parámetro canónico  $\theta$  de una familia exponencial. Por su parte, Mendoza (1994) discute la preservación de la normalidad asintótica bajo transformaciones suaves de los parámetros. Estos dos resultados permiten demostrar que la distribución final del parámetro mixto  $(\mu_{(1)}, \theta_{(2)})$  es también asintóticamente normal.

## 4 Distribuciones de Referencia para Familia Exponentiales con Cortes

Dado que  $h_1(\mu_{(1)}, \theta_{(2)}) = V_1^*(\mu_{(1)})^{-1}$ , entonces el determinante se factoriza trivialmente como

$$|h_1(\mu_{(1)}, \theta_{(2)})| = a_1(\mu_{(1)}) b_1(\theta_{(2)}),$$

con  $a_1(\mu_{(1)}) = |V_1^*(\mu_{(1)})|^{-1}$  y  $b_1(\theta_{(2)}) \equiv 1$ . Por otra parte,

$$h_2(\mu_{(1)}, \theta_{(2)}) = \frac{\partial k(\theta_{(2)})}{\partial \theta'_{(2)}} - \frac{\partial \{g(\theta_{(2)})'\mu_{(1)}\}}{\partial \theta'_{(2)}}.$$

En general no es fácil determinar si  $|h_2(\mu_{(1)}, \theta_{(2)})|$  puede factorizarse de manera similar. Sin embargo, si  $k(\theta_{(2)})$  es una función constante entonces podemos distinguir los siguientes casos:

(a)  $n_1 = 1, n_2 = 1$ . En este caso  $\mu_{(1)}, \theta_{(2)}, g(\cdot)$  y  $h(\cdot, \cdot)$  son todos escalares y

$$|h_2(\mu_{(1)}, \theta_{(2)})| = a_2(\theta_{(2)}) b_2(\mu_{(1)}), \quad (2)$$

con  $a_2(\theta_{(2)}) = \left\| \frac{\partial g(\theta_{(2)})}{\partial \theta_{(2)}} \right\|$  y  $b_2(\mu_{(1)}) = \|\mu_{(1)}\|$ , donde  $\|\cdot\|$  denota a la función valor absoluto.

(b)  $n_1 = 1, n_2 > 1$ . En este caso  $\mu_{(1)}$  es un escalar y por lo tanto se da la factorización (2) con  $a_2(\theta_{(2)}) = \left\| \left| \frac{\partial g(\theta_{(2)})}{\partial \theta_{(2)}} \right| \right\|$  y  $b_2(\mu_{(1)}) = \|\mu_{(1)}\|^{n_2}$ .

Por lo tanto, si  $n_1 = 1$  entonces se satisfacen las hipótesis de la Proposición 1 y

$$\pi(\mu_{(1)}, \theta_{(2)}) \propto |V_1^*(\mu_{(1)})|^{-1} \left\| \left| \frac{\partial g(\theta_{(2)})}{\partial \theta_{(2)}} \right| \right\|.$$

No hemos podido demostrar un resultado similar para el caso  $n_1 > 1$ , aunque hasta el momento tampoco hemos podido encontrar un contraejemplo en el que la factorización no se produzca en este caso. De hecho, es posible demostrar que el resultado es válido incluso en el caso  $n_1 > 1$  si nos restringimos a la clase de las familias exponenciales con función de varianza cuadrática simple (Casalis, 1996), la cual contiene a varias de las familias más utilizadas en la práctica, incluyendo a la normal y a la multinomial.

## Referencias

- Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. Chichester: Wiley.
- Barndorff-Nielsen, O.E. and Koudou, A.E. (1995). Cuts in Natural Exponential Families. *Theory of Probability and its Applications*, **40**, 361-372.
- Berger, J.O. y Bernardo, J.M. (1992a). On the Development of Reference Priors (with discussion). En *Bayesian Statistics 4*. (J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds.) Oxford: Clarendon Press, pp. 35-60.
- Berger, J.O. y Bernardo, J.M. (1992b). Ordered Group Reference Priors with Applications to the Multinomial Problem. *Biometrika*, **79**, 25-37.

- Bernardo, J.M. y Smith, A.F.M. (1994). *Bayesian Theory*. Chichester: Wiley.
- Bernardo, J.M. y Ramón, J.M. (1996). An Introduction to Bayesian Reference Analysis: Inference on the Ratio of Multinomial Parameters. Paper presented at the *Workshop on Default Bayesian Methodology*, Purdue University, November 1996.
- Casalis, M. (1996). The  $2d + 4$  Simple Quadratic Natural Exponential Families on  $R^d$ . *Annals of Statistics*, **24**, 1828-1854.
- Mendoza, M. (1994). Asymptotic Normality under Transformations: a Result with Bayesian Applications. *Test*, **3**, 173-180.
- Morris, C.N. (1982). Natural Exponential Families with Quadratic Variance Functions. *The Annals of Statistics*, **10**, 65-80.

# Simulación de Procesos de Riesgo en Ambiente Markoviano

Luis Fernando Hoyos Reyes  
*UAM-Azcapotzalco*

## 1 Introducción

Cuando una compañía de seguros expide una póliza, el total de ingresos más el capital inicial debe exceder la suma de los montos de las reclamaciones en cualquier momento de la duración de la cobertura, a la probabilidad de que esto no ocurra se denomina probabilidad de ruina.

Resulta evidente la necesidad de modelar estocásticamente el comportamiento de una póliza: el monto de cada reclamación es una variable aleatoria, el número de reclamaciones en un intervalo determinado de tiempo también es variable aleatoria, con estos elementos se construye un proceso de riesgo.

Por otra parte, existen factores que inciden en el comportamiento cualitativo de un fenómeno aleatorio por ejemplo, el número de accidentes viales aumenta en temporada de lluvias, el número de incendios forestales en temporada de sequía,etc.

Al proceso de riesgo se le incorpora un proceso de salto markoviano que describe cambios en el comportamiento cualitativo en función del tiempo.

El objetivo de este trabajo consiste en estimar mediante simulación la probabilidad de ruina de un proceso de riesgo en ambiente markoviano con dos estados y distribuciones exponenciales para los montos de las reclamaciones.

## 2 Definiciones y Notación

Sean  $u$  el capital inicial y  $\{X_n; n = 0, 1, \dots\}$  el proceso de los montos de las reclamaciones; el proceso de llegada de las reclamaciones es un proceso de Poisson  $\{N_t; t \geq 0\}$  con tasa  $\lambda$  y tiempos entre reclamaciones  $\{T_i; i = 1, 2, \dots\}$ .

El proceso acumulado de reclamaciones  $Z_t = \sum_{n=0}^{N_t} X_n$  es la suma de los montos de todas las reclamaciones ocurridas desde el inicio hasta el momento  $t$ .

Si consideramos a  $c$  la constante de ingresos por unidad de tiempo, definimos un proceso de riesgo como:  $U_t = u + ct - Z_t$  (Grandell ,1991), lo que modela la utilidad de la compañía en cualquier momento  $t$ .

Típicamente  $c = (1 + \rho) E(Z_t)$ , es decir la constante de ingresos es mayor que la pérdida promedio, ya que el factor de recargo  $\rho$  es mayor estrictamente que cero.

Este factor puede interpretarse como el porcentaje de utilidad promedio de la aseguradora.

La ruina ocurre cuando en algún momento  $\tau$  el proceso de riesgo  $U_\tau$  es negativo y denotamos a la probabilidad de ruina con horizonte finito  $\Psi(u, T) = P\{\exists \tau \in [0, T] : U_\tau < 0\}$ , si definimos  $Y_t = Z_t - ct$  podemos escribir  $\Psi(u, T) = P\{\exists \tau \in [0, T] : Y_\tau > u\}$ .

### 3 Construcción del Proceso de Saltos Markoviano

Una forma de generalizar un proceso de riesgo consiste en considerar  $\{Y_t : t \geq 0\}$  con la propiedad de que la tasa de llegada de  $\{N_t : t \geq 0\}$  y la distribución de los montos de las reclamaciones no estén fijas en el tiempo, sino que dependan del estado de un proceso de salto markoviano  $\{M_t : t \geq 0\}$  en el que no aceptamos transiciones instantáneas de estado, es decir, el tiempo de permanencia en cada estado es estrictamente positivo.

Denotaremos los tiempos de transición por  $0 = H_0 < H_1 < H_2 < \dots$ , los tiempos de permanencia por  $B_n = H_{n+1} - H_n$  y la secuencia de estados visitados por  $S_0, S_1, \dots$

Sea  $E$  el espacio de estados discreto del proceso de salto markoviano.

Existen 2 fenómenos interesantes: i) el proceso puede absorberse en un estado  $i$

ii) el tiempo de explosión  $\omega(\Delta) = \sup_n H_n$  es finito.

En el primer caso existe un último  $H_n$  finito y simplemente definimos

$$B_n = B_{n+1} = \dots = \infty, \quad S_n = S_{n+1} = \dots = i.$$

El segundo caso sugiere la necesidad de una construcción mínima en términos de las intensidades  $\lambda(i)$ , que nos permita caracterizar el fenómeno de explosión.

Construimos  $\{M_t : t \geq 0\}$  hasta el tiempo de explosión simplemente revirtiendo la construcción de  $S_k, B_k$ , es decir: Sean  $H_0 = 0$ ,  $H_n = \sum_{i=0}^{n-1} B_i$ ,  $M_t = \begin{cases} S_k & H_k \leq t < H_{k+1} \\ \Delta & t \geq \omega(\Delta) \end{cases}$

Es un resultado conocido que  $\{M_t : t \geq 0\}$  es un proceso de salto markoviano con espacio de estados  $E_\Delta = E \cup \{\Delta\}$ .

Definimos la matriz de intensidad del proceso  $\Lambda = (\lambda(i, j))_{i,j \in E}$ , donde  $\lambda(i, j) = \lambda(i) q_{ij}$  cuando  $j \neq i$  y  $\lambda(i, i) = -\lambda(i)$ .

Sabemos por la condición de explosión de Reuter que un proceso de salto markoviano es no explosivo si y sólo si la única solución acotada no negativa del sistema de ecuaciones  $\Lambda k = k$  es  $k = 0$ .

## 4 El Proceso de Riego en Ambiente Markoviano

La naturaleza de los fenómenos a modelar sugieren que el proceso de salto markoviano sea irreducible: el mal clima no dura para siempre, al periodo de lluvias siempre sucede la temporada de sequía, etc.

Esto resulta relevante ya que un proceso irreducible no-explosivo es ergódico si y sólo si existe una solución a  $\pi\Lambda = 0$ , con  $\pi = (\pi_1, \pi_2, \dots, \pi_n)$  vector de probabilidad.

El proceso de salto markoviano se refleja en nuevas expresiones para  $c$  y para el margen de seguridad  $\rho$ : puede verse fácilmente que  $E_\pi \frac{Y_t}{t} = E_\pi Z_t - c = \sum_{i \in E} \pi_i \lambda(i) E_i X - c$ , por lo tanto  $c = (1 + \rho) \sum_{i \in E} \pi_i \lambda(i) E_i X$  y  $\rho = \frac{c}{\sum_{i \in E} \pi_i \lambda(i) E_i X} - 1$

La distribución de los tiempos de permanencia  $B$  es exponencial con tasa de transición  $\alpha_i$ .

En este trabajo consideramos un espacio de dos estados  $E = \{1, 2\}$ , donde cada estado tiene un comportamiento cualitativo diferente: en el estado 1 la utilidad crece en promedio, mientras que en el estado 2 decrece.

En nuestro problema cuando  $M_t = 1$  los montos de las reclamaciones se distribuyen exponencialmente con parámetro  $\beta_1$ , cuando  $M_t = 2$  los montos tienen la misma distribución, pero con diferente parámetro  $\beta_2$ .

Sea  $\Lambda = \begin{pmatrix} -\alpha & \alpha \\ 2\alpha & -2\alpha \end{pmatrix}$  la matriz de intensidad de  $\{M_t : t \geq 0\}$ , luego las transiciones del estado 1 al estado 2 ocurren con tasa  $\alpha$ , las transiciones del estado 2 al estado 1 ocurren con tasa  $2\alpha$ .

Por supuesto que la solución a  $\Lambda k = k$  es  $k = 0$ , por lo que el proceso de salto es no-explosivo.

Ahora calculamos la distribución estacionaria:  $\pi\Lambda = 0$  entonces  $\pi = (\frac{2}{3}, \frac{1}{3})$ .

Sabemos que  $E_\pi \frac{Z_t}{t} = \frac{\pi_1 \lambda(1)}{\beta_1} + \frac{\pi_2 \lambda(2)}{\beta_2}$  por lo que basta que  $\lambda(1) < \beta_1 \cdot c$  y  $\lambda(2) > \beta_2 \cdot c$  para garantizar el diferente comportamiento cualitativo.

## 5 Estimación de la Probabilidad de Ruina con Horizonte Finito

Para las simulaciones se consideró  $\lambda(1) = 0.45$ ,  $\lambda(2) = 1.8$  que cumplen con las condiciones mencionadas tomando en cuenta el resto de los parámetros de la tabla de resultados.

	$\alpha$	$T$	$u$	$\beta_1$	$\beta_2$	$c$	$\widehat{\Psi}_1(u, T)$	$\widehat{\Psi}_2(u, T)$	$N$
1	10	30	30	1	1	1	0.033	0.033	1000
2	1	30	30	1	1	1	0.010	0.010	5000
3	1/10	30	30	1	1	1	0.031	0.068	5000
4	1/64	250	50	1	1	1	0.208	0.394	1000
5	1/25	100	50	1	1/10	7	0.529	0.864	5000
6	1/25	100	100	1	1/10	7	0.434	0.745	5000

$\widehat{\Psi}_1(u, T)$  y  $\widehat{\Psi}_2(u, T)$  son los estimadores de la probabilidad del proceso de riesgo en ambiente markoviano dado que  $S_0 = 1$  y  $S_0 = 2$  respectivamente. Se seleccionó  $c = 1$  en los ejemplos del 1 al 4 y  $c = 7$  en los ejemplos 5 y 6 porque implican un factor de recargo típico según Embrechts y Wouters (1990) del 11%. ( $\rho = 0.11$ ) El algoritmo de simulación está dominado por la tasa de transición  $\alpha$ : cuando ésta crece, el tiempo de ejecución aumenta, por lo que para el ejemplo 1 sólo se efectuaron 1000 simulaciones. El ejemplo 4 fue planteado por Asmussen (1989) que obtuvo  $\widehat{\Psi}_1(50, 250) = 0.193$  y  $\widehat{\Psi}_2(50, 150) = 0.399$ , usando una técnica basada en procesos conjugados.

## 6 Conclusiones

En general  $\frac{1}{\alpha}$  puede considerarse como una medida del grado de modulación markoviana: entre más grande  $\alpha$  las medias de los tiempos de permanencia serán menores, provocando que las estimaciones de probabilidad de ruina sean casi independientes del estado inicial. Si disminuimos la tasa de transición, las medias de los tiempos de permanencia crecen, haciéndose significativa la diferencia cualitativa entre ambos estados. La complejidad de esta técnica aumenta considerablemente con el número de estados del proceso de salto markoviano subyacente, Asmussen y Rolski (1991) propusieron una generalización, pero asumiendo que la distribución de los montos de las reclamaciones es tipo-fase en el sentido de Neuts.

## Referencias

- Asmussen S. (1989). Risk Theory in a Markovian Environment. *Scandinavian Actuarial Journal*, 1989, 69-100
- Asmussen S. y Rolski, G. (1991). Computational methods in risk theory: A matrix-algorithmic approach. *Insurance : Mathematics and Economics* 10, 259-274
- Embrechts, P. y Wouters, L. (1990). Simulating risk solvency. *Insurance : Mathematics and Economics* 9, 141-148
- Grandell J. (1991). *Aspects of Risk Theory*. New York: Springer-Verlag.

# Análisis del Comportamiento de la Salinidad en una Laguna Costera, por Medio de Métodos de Suavización no Paramétrica

Jorge Manuel López Reynoso      Isaías H. Salgado Ugarte

María José Marques Dos Santos

*FES - Zaragoza, UNAM*

## 1 Introducción

México se encuentra localizado en una región geográfica donde las lagunas costeras son características de las fronteras continentales con el mar. Una laguna costera se distingue de otros ambientes acuáticos por ser una zona de transición entre las aguas dulces de los ríos y las aguas marinas. Por esta razón, uno de los parámetros hidrológicos más importantes en los estudios de estos cuerpos lagunares es la salinidad, cuyos cambios determinan su productividad y diversidad biológicas. Tales cambios pueden ocurrir como pequeñas variaciones aleatorias o se presentan como consecuencia de fuertes alteraciones ambientales. Sin embargo, la mayoría de las veces los datos de una investigación hidrobiológica no se encuentran claramente en alguno de estos extremos, por lo que deben analizarse en forma exhaustiva y preferentemente con herramientas analíticas cuantitativas.

Actualmente se dispone de un número grande de pruebas estadísticas, las que se pueden aplicar en una gran variedad de investigaciones. No obstante, se presenta con frecuencia el caso que la estructura de los datos no concuerda con los supuestos que requiere alguna prueba en particular. Lo más grave es que el resultado del análisis estadístico en ocasiones contradice el comportamiento real de la variable estudiada y se acepta sin discusión crítica.

En particular en este trabajo se analiza la salinidad de las lagunas de Chacahua, Oax., para mostrar la forma de la distribución de esta variable con técnicas no paramétricas de suavización, en específico los estimadores de densidad por kernel. En el caso de que tal distribución fuera multimodal, se descartaría la posibilidad de un análisis estadístico confirmatorio de esta información bajo el supuesto de normalidad.

## 2 Métodos

Los datos de salinidad analizados provienen de un grupo de trabajos realizados acerca del plancton de las lagunas de Chacahua (ver Zárate-Vidal, 1985; Ortiz-Ortiz y Teodoro-Salvador, 1990).

Para hacer un análisis más detallado, el conjunto de datos se dividió en dos lotes, uno con 96 observaciones correspondientes a la laguna Chacahua y el otro con 144 observaciones obtenidas en la laguna Pastoría. Con cada lote se determinaron diversos valores de amplitud de banda para la construcción de las estimaciones de densidad.

En algunos casos la amplitud de banda calculada proporcionaría una estimación sobreavuizada y en otros se tendría una estimación óptima de la densidad de la variable. También se determinaron amplitudes de banda a través del procedimiento de validación cruzada sesgada (*BCV*) y el de la validación cruzada por mínimos cuadrados (*L2CV*). Esta parte del análisis se hizo siguiendo la metodología propuesta por Salgado-Ugarte *et al.* (1995b).

A continuación se hicieron las estimaciones de densidad propiamente dichas, entre las que destacan los histogramas desplazados promedios (*ASH*) y las estimaciones de densidad por kernel (*EDK*). Lo anterior se logró usando los programas adaptados al paquete estadístico *Stata* que realizan el procedimiento *WARP* (*weighted average of rounded points*) elaborados por Salgado-Ugarte *et al.* (1995a), de acuerdo con las ideas presentadas por Härdle y Scott (1988), Härdle (1991) y Scott (1992), en conjunción con los resultados previos.

La parte final del análisis consistió en efectuar la prueba de multimodalidad propuesta por Silverman (1981), basada en el procedimiento *bootstrap* suavizado, para determinar el número significativo de modas en cada lote, utilizando los programas implementados en *Stata* por Salgado-Ugarte *et al.* (1997), junto con el comando para muestreo repetitivo incluido en el mismo paquete.

## 3 Resultados

Se presentan sólo los resultados obtenidos con el lote de datos de la laguna Chacahua, que mostró una estructura más compleja que la laguna Pastoría.

### CUADRO 1

Algunas reglas prácticas para elección de número y amplitud de intervalos  
y amplitud de banda para histogramas y EDK's

Regla de Sturges	7.5850
Amplitud de intervalo óptima Gaussiana de Scott	6.4848
Amplitud de banda sobreavuizada para kernel Gaussiano de Scott	3.8954

## CUADRO 2

Prueba de Silverman: amplitudes críticas de banda  
y niveles estimados de significancia

No. de modas	Amplitud de banda crítica	Valor de $P$
1	5.41	0.01
2	3.91	0.00
3	2.56	0.03
4	1.56	0.19
5	0.93	0.55

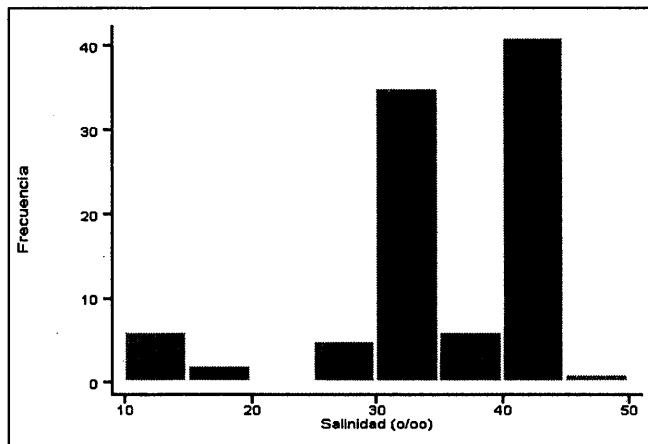


Figura 1: Histograma con la regla de Sturges

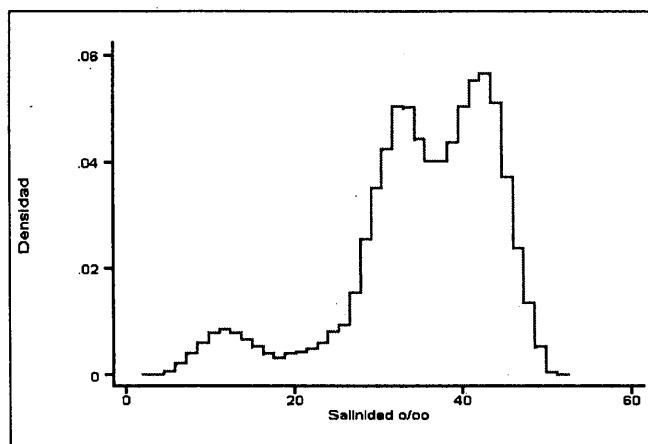


Figura 2: Histograma desplazado promedio con banda óptimo de Scott

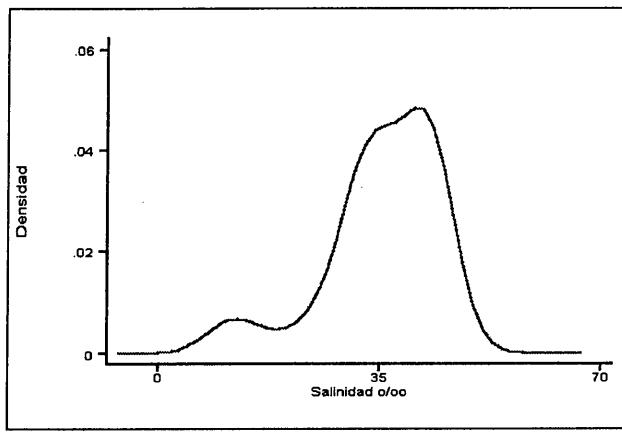


Figura 3: EDK Gaussiano con banda sobresuavizada de Scott ( $h=3.895$ )

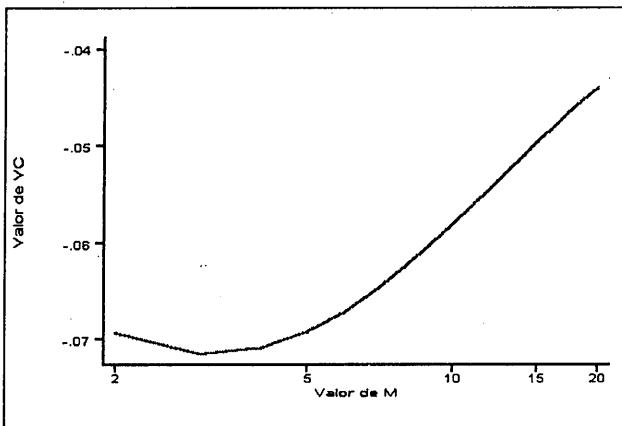


Figura 4: Valor de validación cruzada por mínimos cuadrados. El valor mínimo corresponde a una banda de  $h=0.519$

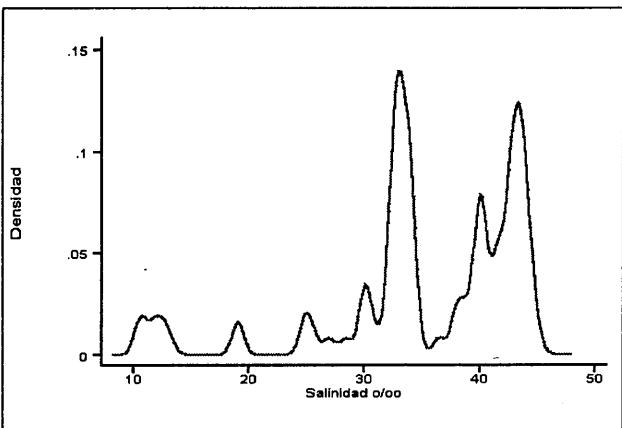


Figura 5: EDK Gaussiano con la banda sugerida por VC por mínimos cuadrados  $h=0.519$

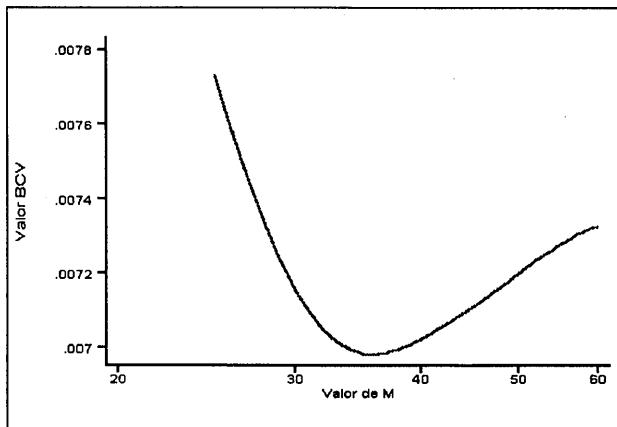


Figura 6: Valor de validación cruzada sesgada. El valor mínimo corresponde a una banda gaussiana  $h=4.173$

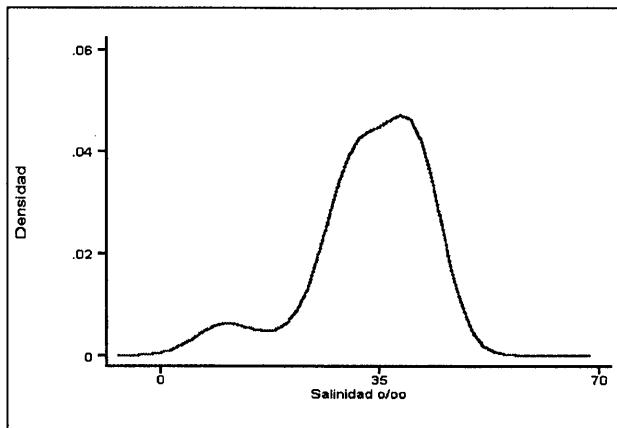


Figura 7: EDK Gaussiano con la banda sugerida por VC sesgada ( $h=4.173$ )

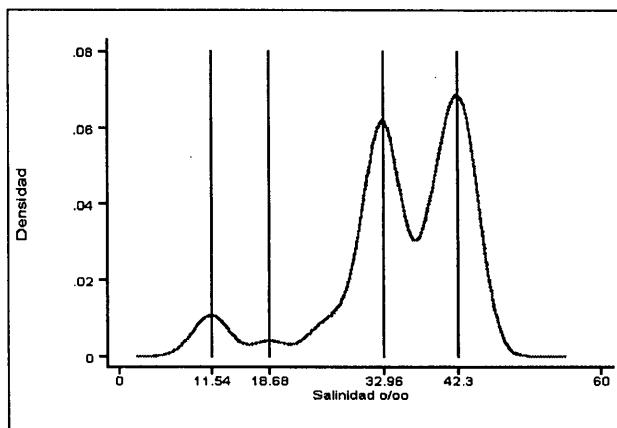


Figura 8: EDK Gaussiano con la amplitud de banda sugerida por la prueba de Silverman ( $h=2.06$ )

## 4 Discusión y Conclusiones

Anteriormente se ha discutido, en forma breve, cómo se distribuye la salinidad en las Lagunas de Chacahua y la posibilidad de aplicar un análisis de varianza a esta información (López-Reynoso, 1995), pero ahora el análisis se basa en los estimadores de densidad por kernel.

Una estimación poco detallada de la densidad de la salinidad en Chacahua es el histograma, construido de acuerdo con la regla de Sturges (figura 1), donde puede observarse una distribución trimodal. El aspecto de la gráfica podría ser accidental debido a una selección arbitraria del origen de la gráfica, pero todas las otras estimaciones muestran una estructura similar. Por ejemplo, se puede ver el *ASH* construido con la amplitud de banda gaussiana recomendada por Scott (figura 2), el cual muestra también tres modas, o la gráfica de *EDK* gaussiano con amplitud de banda sobresuavizada dada por Scott (figura 3) en la que destacan dos modas, aunque la mayor tiene un “hombro”, el cual es un probable indicio de una moda adicional. Buscando una óptima estimación de la densidad, se hizo la determinación de la amplitud de banda con el procedimiento de validación cruzada por mínimos cuadrados (figura 4), pero la gráfica *EDK* resultó ser demasiado “ruidosa” (figura 5). Por otra parte, se obtuvo la amplitud de banda a través de la validación cruzada sesgada (figura 6), con la que resultó una gráfica sobresuavizada (figura 7), similar a la sobresuavizada de Scott. La última estimación de densidad se hizo conforme a los resultados obtenidos en la prueba de Silverman, con los que se puede concluir que existen al menos cuatro modas en la distribución de salinidad en la laguna Chacahua (figura 8).

Al considerar en conjunto los resultados anteriores, se tiene una fuerte evidencia contraria a la hipótesis de que la salinidad en Chacahua tiene una distribución unimodal. Hay tres modas conspicuas que están relacionadas con contingencias ambientales que se pudieron observar durante las campañas de muestreo. Un suceso importante fue el cierre gradual de la boca que comunica a la laguna con el mar. Cuando la laguna tuvo amplia comunicación con el mar, su salinidad fue similar a la del océano adyacente (de 30 a 35 partes por mil), pero cuando cesó dicha comunicación a causa de procesos litorales de transporte de material, la salinidad aumentó notablemente (más de 40 partes por mil) por evaporación y falta de aportes dulceacuícolas. Otro suceso importante fue la temporada de lluvias. Estas introdujeron una considerable cantidad de agua dulce a la laguna, lo que provocó una disminución considerable en la salinidad (menos de 15 partes por mil).

En síntesis, la salinidad de la laguna costera Chacahua no se ajusta a un modelo normal, lo que haría inapropiado analizarla con algún procedimiento estadístico paramétrico (basado en el supuesto de normalidad) en busca de diferencias significativas.

El análisis de los resultados obtenidos con el lote de datos correspondiente a la laguna Pastoría, permite llegar a las mismas conclusiones generales, aunque en particular sólo se detectaron dos modas, ya que no hubo interrupción de la comunicación de la laguna con el mar.

## Referencias

- Härdle, W. (1991). *Smoothing Techniques with implementation in S*. New York: Springer-Verlag.
- Härdle, W. y D.W. Scott. (1988). Smoothing in low and high dimensions by weighted averaging using rounded points. *Technical report 88-16*, Rice University.
- López-Reynoso, J.M. (1995). *Estudio de algunas relaciones del fitoplancton con su ambiente en dos lagunas costeras de Oaxaca, México*. Tesis profesional, FES Zaragoza, U.N.A.M.
- Ortiz-Ortiz, J.O. y Teodoro-Salvador, M.E. (1990). *Algunos aspectos ecológicos del zoopláncton en las lagunas de Chacahua y Pastoría, Oax*. Tesis profesional, ENEP Zaragoza, U.N.A.M.
- Salgado-Ugarte, I.H., M. Shimizu y T. Taniuchi. (1995a). spn6.1: ASH, WARPing and kernel density estimation for univariate data. *Stata Technical Bulletin* 26, 23-31.
- (1995b). spn6.2: Practical rules for bandwidth selection in univariate density estimation. *Stata Technical Bulletin* 27, 5-19.
- (1997). spn13: Nonparametric assessment of multimodality for univariate data. *Stata Technical Bulletin* 38, 27-35.
- Scott, D.W. (1992). *Multivariate density estimation: theory, practice, and visualization*. New York: Wiley.
- Silverman, B.W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society, Series B* 43, 97-99.
- Zárate-Vidal, S.E. (1985). *Algunos aspectos ecológicos del ictioplancton en las lagunas de Chacahua y Pastoría, Oax*. Tesis profesional, ENEP Zaragoza, U.N.A.M.

# Analizando la Distribución de Índices de Ozono de la Ciudad de México por Medio de Estimadores de Densidad por Kernel

María José Marques Dos Santos e Isaías H. Salgado Ugarte  
*FES-Zaragoza, UNAM*

## 1 Introducción y Objetivos

La densidad en un punto  $x$  puede definirse como el límite de la altura de una barra de histograma centrada en  $x$  cuando la amplitud del intervalo tiende a cero. De manera formal, el intervalo es una función de peso que asigna un valor positivo a cada dato que abarca e iguala a cero a todos aquellos fuera de él. El histograma utiliza una función uniforme centrada en las marcas de clase de los intervalos definidos a partir de un origen único. En la práctica, histogramas con diferente origen pueden mostrar distribuciones con diferentes características. Los estimadores de densidad por kernel (EDK) centran la estimación de densidad en cada observación, por lo que no tienen un origen fijo y pueden hacer uso de funciones de ponderación que varían gradualmente, disminuyendo del centro hacia sus extremos (como una función gaussiana). Por estas características, los EDK's son particularmente útiles para estudiar la distribución de variables aleatorias continuas.

A pesar de su simplicidad, los estimadores de densidad por kernel no se propusieron sino hasta 1956 por Rosenblatt debido principalmente a la abrumadora cantidad de operaciones requeridas para su cálculo y por la misma razón su uso ha sido limitado.

Silverman (1986) presenta la monografía más completa que ha servido como base a los trabajos más recientes sobre el tema, entre los que destaca el libro de Scott (1992) quien extiende la aplicación de los EDK's a más de dos variables. También recientemente se ha incluido a estos estimadores en libros generales de estadística aplicada como por ejemplo Chambers, *et al.* (1983). Es hasta que se cuenta con computadoras veloces de bajo costo que su empleo se está generalizando para el análisis de datos en diferentes campos.

En el presente trabajo se pretende mostrar una aplicación de las técnicas de suavización no paramétrica utilizando los índices metropolitanos de calidad del aire de ozono. Se seleccionó este conjunto de datos por ser uno de los problemas que afecta a los habitantes de la ciudad de México. Los índices metropolitanos de la calidad del aire (IMECA) para el ozono se consideran favorables de 0 a 100, aceptables de 101 a 200, peligrosos de 200 a 300

y muy peligrosos de 300 en adelante. Los datos ( $n = 184$ ) consisten en los valores diarios del índice IMECA de ozono correspondientes a los meses de julio, agosto y septiembre de 1994 y 1995.

Para el presente trabajo se plantearon los siguientes objetivos:

1. Aplicar los estimadores de densidad por kernel para analizar la distribución de los índices IMECA de la Cd. de México.
2. Utilizar las reglas prácticas para elección de la amplitud de banda para estimadores de densidad por kernel gaussiano.
3. Aplicar los métodos estadísticos de cálculo intensivo para determinar el número significativo de modas de la distribución de los índices.

## 2 Metodología

Sorprendentemente, los EDK's no están incluidos en varios de los paquetes estadísticos comunes. Para el cálculo de los estimadores de densidad por kernel se utilizaron los programas para el paquete estadístico *Stata* presentados en Salgado-Ugarte, *et al.* (1995a). Estos programas hacen las estimaciones por medio del algoritmo basado en el promedio ponderado de puntos redondeados (WARP, por sus siglas en inglés) y una función ponderal gaussiana. Las reglas para la elección de banda fueron calculadas por los programas para *Stata* de Salgado-Ugarte *et al.* (1995b) que utilizan las expresiones de Silverman (1986), Härdle (1990) y Scott (1992). Estas reglas producen estimaciones de valores óptimos y sobresuavizados para la amplitud de banda. Las bandas óptimas son las adecuadas en el caso de distribuciones cercanas a la gaussiana pero muy amplias en caso de sesgo o multimodalidad. Las reglas de valores sobresuavizados proporcionan un límite superior útil para la amplitud de la ventana utilizada en la estimación de densidad. Estas reglas resultan en estimaciones muy suaves y por tanto son menos propensas a mostrar estructuras falsas. Si en estas estimaciones aparece sesgo o multimodalidad, el analista puede tener un alto grado de confianza en su autenticidad (Terrell, 1990). Asimismo, la validación cruzada por mínimos cuadrados y la sesgada se calculó con las rutinas que utilizan adaptaciones de los programas en el lenguaje C de Härdle (1990) integradas como ejecutables en Turbo Pascal al paquete *Stata* (Salgado-Ugarte *et al.*, 1995b). La prueba de multimodalidad de Silverman (1981) que hace uso del *bootstrap* suavizado (Efron, 1982), se realizó con la versión automatizada presentada recientemente por Salgado-Ugarte *et al.* (1997).

### 3 Resultados

Cuadro 1. Reglas prácticas de elección de amplitud de banda para estimadores de densidad por kernel

---

Amplitud de banda óptima para kernel Gaussiano de Silverman	15.3287
Amplitud de banda óptima “mejorada” para kernel Gaussiano de Härdle	18.0538
Amplitud de banda sobresuavizada para Kernel Gaussiano de Scott	19.4845

---

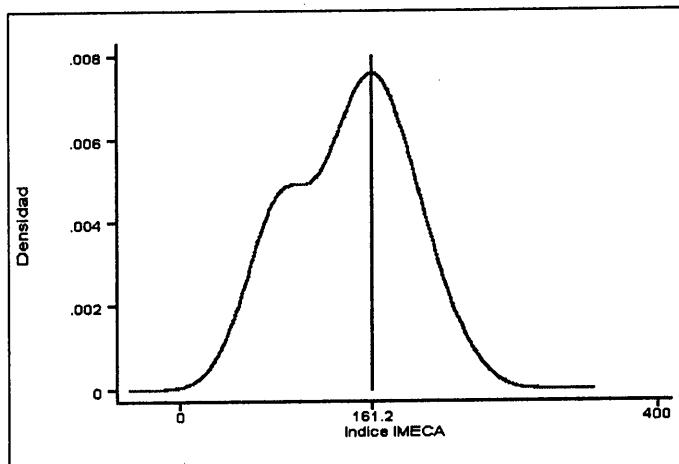


Figura 1: Estimación de densidad por kernel gaussiano con amplitud de banda sobresuavizada de Scott ( $h=19.5$ )

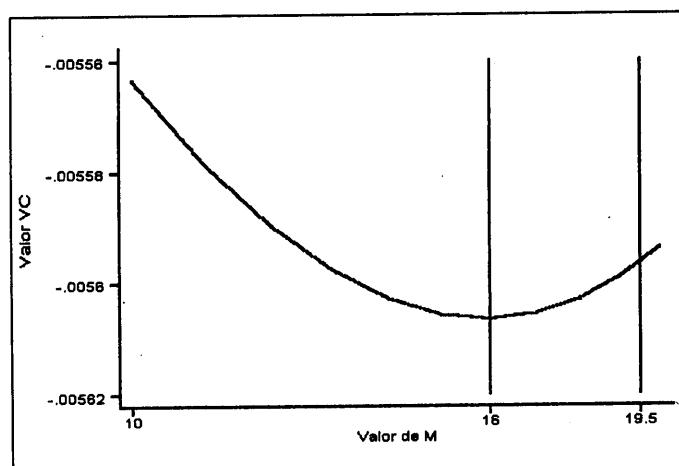


Figura 2: Valor de validación cruzada por mínimos cuadrados para los datos de índices IMECA. La amplitud de banda sobresuavizada se indica por la línea en 19.5. El valor de M corresponde directamente a h

Cuadro 2. Prueba de Silverman: amplitudes críticas de banda y niveles estimados de significancia

Número de modas	Banda crítica	Valor de $P$
1	19.29	0.08
2	7.10	0.95
3	6.48	0.84

Nota: los valores de  $P$  fueron obtenidos de  $B = 100$  repeticiones bootstrap de tamaño 184.

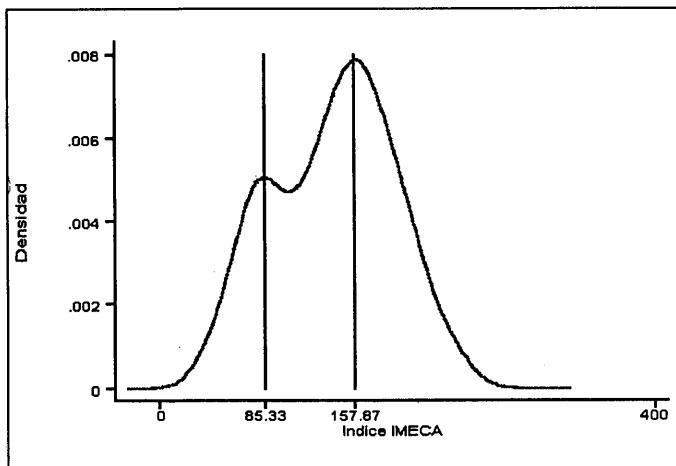


Figura 3: Estimación de densidad por kernel gaussiano con amplitud de banda recomendada por validación cruzada de mínimos cuadrados ( $h = 16$ )

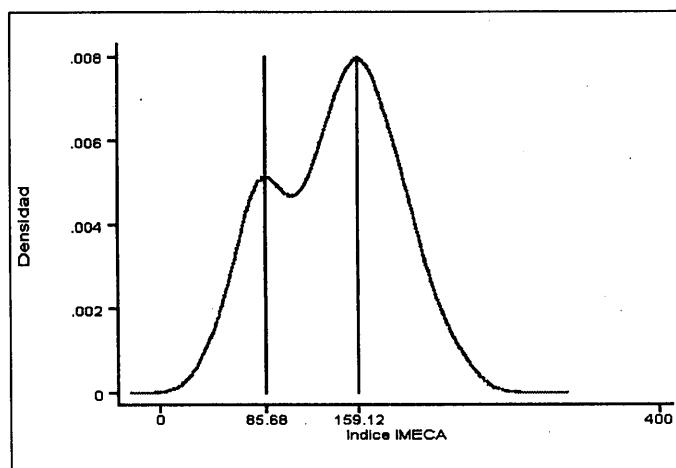


Figura 4: Estimación de densidad por kernel gaussiano con amplitud de banda recomendada por la prueba de multimodalidad de Silverman ( $h = 13.2$ )

## 4 Discusión y Conclusiones

El cuadro 1 presenta los valores de amplitud de banda para tres reglas de elección de banda. La estimación sobresuavizada (Fig. 1) muestra la existencia de una moda (161.2) y un hombro, por lo que se sugiere fuertemente la existencia de una estructura más complicada de los datos. La validación cruzada por mínimos cuadrados produjo un error mínimo (Fig. 2) utilizando una amplitud un poco menor ( $h = 16$ ); la correspondiente estimación de densidad (Fig. 3) es claramente bimodal con modas de 85.33 y 157.87. La aplicación de la prueba de Silverman con 100 muestras repetidas para cada banda crítica condujo a los resultados del Cuadro 2 que indica un número significativo de dos modas. Como dos modas se obtienen con bandas menores de 19.29 y hasta 7.10, se utilizó el valor intermedio de 13.2 para estimar la densidad (Fig. 4), la cual también es bimodal con modas en 85.68 y 159.12. Todo lo anterior sugiere que los valores de los índices IMECA de ozono en los meses de verano de 1994 y 1995 pertenecieron a dos clases: valores predominantes alrededor de 160 y aquellos cercanos a 85. Esto indica que aunque no se registraron valores indicadores de alarma para la Fase 1 de contingencia (mayores de 200), si manifiesta la predominancia de valores altos de ozono.

Un estudio adicional sería la búsqueda de variables correlacionadas con estos dos grupos de valores altos y bajos encontrados para detectar que condiciones o medidas se corresponden con valores bajos y propiciarlas.

El uso de los estimadores de densidad por kernel y la aplicación de las diversas reglas y pruebas basadas en procedimientos de cómputo intensivo (validación cruzada y bootstrap) representan un conjunto de herramientas muy poderoso para la extracción de información de distribuciones univariadas.

## Referencias

- Corona, R. y G. Calva. (1989). Contaminación atmosférica en la ciudad de México: causas, concentraciones y efectos. *Tópicos de Investigación y Posgrado*. 1, 10-21.
- Chambers, J.M., W.S. Cleveland, B. Kleiner y P.A. Tukey, (1983). Graphical Methods for Data Analysis. Belmont: Wadsworth.
- Efron, B. (1982). *The jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: SIAM.
- Härdle, W. (1991). Smoothing Techniques with Implementation in S. New York: Springer-Verlag.
- Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27, 832- 837.

- Salgado-Ugarte, I.H., M. Shimizu y T. Taniuchi. (1995a). ASH, WARPing, and kernel density estimation for univariate data. *Stata Technical Bulletin*, 26, 23-31.
- Salgado-Ugarte, I.H., M. Shimizu y T. Taniuchi. (1995b). Practical rules for bandwidth selection in univariate density estimation. *Stata Technical Bulletin*, 27, 5-19.
- Salgado-Ugarte, I.H., M. Shimizu y T. Taniuchi. (1997). Nonparametric assessment of multimodality for univariate data. *Stata Technical Bulletin*, 38, 27-35.
- Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. New York: Wiley.
- Silverman, B.W. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society, Series B*, 43, 97-99.
- Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.
- Terrell, G.R. (1990). The maximal smoothing principle in density estimation. *Journal of the American Statistical Association*, 85, 470-477.

# Estimación de la Eficiencia de un Método no Paramétrico para Probar la Multimodalidad de Datos Univariados

Martínez Ramírez Juana Mayorga Saucedo Erika

Isaiás Hazarmabeth Salgado-Ugarte

*FES-Zaragoza, UNAM*

## 1 Introducción

Los investigadores que trabajan con formas complejas de distribución de datos univariados han vuelto su interés en los años recientes hacia técnicas no paramétricas tales como la estimación de densidad por kernel (Silverman, 1986). Las distribuciones multimodales pueden interpretarse como mezclas de componentes individuales los cuales pueden detectarse por medio de la identificación de las modas (máximos locales) en la distribución (Izenman y Sommer, 1988). El número y localización de las modas puede no corresponder a cada uno de los componentes. Existe una dependencia entre el espaciamiento de las modas y las formas relativas de la distribución de los componentes. Sin embargo, en un considerable número de casos, la existencia de más de una moda es evidencia de una distribución mezclada. En la literatura estadística existen varias pruebas para detectar la multimodalidad en una distribución. Por ejemplo, la prueba DIP propuesta por Hartigan y Hartigan (1985) para aceptar o rechazar la hipótesis de unimodalidad; Good y Gaskins (1980) usaron el método de verosimilitud penalizada para la estimación de densidad en conjunto con otras técnicas estadísticas; Silverman (1981a) combinó la estimación de densidad por kernel con un procedimiento jerárquico de prueba basado en un muestreo repetido (bootstrap). Estos dos últimos métodos son no paramétricos, basados en los datos y de cómputo intensivo. El contexto específico de los datos desempeña en ocasiones un papel prominente en relacionar las modas empíricas con componentes mezclados plausibles. La frecuencia multimodal de tallas de los peces puede resultar de la mezcla de peces con edad semejante (grupos de edad o cortes) y por tanto puede contener información importante acerca de su crecimiento.

En cuanto a los procedimientos no paramétricos, la elección de la amplitud de banda es uno de los problemas centrales en la estimación de la densidad. Existen varias formas de seleccionar una amplitud de banda apropiada para histogramas, polígonos de frecuencia,

histogramas desplazados promediados y estimadores de densidad por kernel (KDE por sus siglas del inglés *Kernel Density Estimator*). Un breve resumen de algunos procedimientos para la selección de la amplitud de banda y algunos programas para su cálculo se encuentra en Salgado-Ugarte *et al.* (1995b). En el presente trabajo se presentan la amplitud de banda óptima para distribución gaussiana de histogramas y polígonos de frecuencia (Scott, 1979, 1985, 1992) y la amplitud de banda óptima para KDE gaussiano (Silverman, 1986). Además, por medio del uso del procedimiento ASH-WARP es posible calcular validación cruzada por mínimos cuadrados y sesgada (L2CV y BCV respectivamente por sus siglas del inglés) para la selección de banda en estimadores de densidad por kernel (Härdle, 1991).

Estas reglas en conjunto con las bandas sobresuavizadas (Terrel, 1990) representan una poderosa herramienta para elegir la amplitud de intervalos en histogramas y polígonos de frecuencia y la amplitud de banda en los estimadores de densidad por kernel (Scott, 1992).

La prueba de Silverman utiliza un EDK gaussiano de acuerdo a los siguientes pasos: identificación de las amplitudes de banda críticas compatibles con la hipótesis de un número dado de modas; obtención de muestras repetidas suavizadas (*smoothed bootstrap*) para cada banda crítica; estimación de las correspondientes densidades; cálculo de la significancia para el número de modas como el cociente (valores de  $p$ ) del número de estimaciones mostrando más modas que el número indicado por la banda crítica utilizada entre el número total de repeticiones. Para una descripción detallada ver Silverman (1981b, 1986) e Izenman y Sommer (1988). Una implementación computarizada y ejemplos de la aplicación de esta prueba al análisis de frecuencia de tallas en peces se presenta en Salgado-Ugarte (1995) y Salgado- Ugarte *et al.* (1997). Un procedimiento relacionado es el de Wong (1985).

## 2 Material y Métodos

Para evaluar la eficacia del método de Silverman, se simularon 25 distribuciones de tamaño  $n = 100$  de cada uno de los siguientes tipos:

1. Unimodal estándar  $N(0,1)$  .
2. Bimodal  $[N(0,1)+N(4,1)]/2$ .
3. Trimodal  $[N(0,1)+2N(4,1)+N(8,1)]/4$ .

Cada muestra simulada fue sometida al procedimiento de Silverman. En primer lugar se determinaron las bandas críticas: para las distribuciones unimodales se determinaron bandas críticas para una, dos y tres modas; para las bimodales de una hasta cuatro modas y para las trimodales se determinaron bandas críticas de una hasta cinco modas. Posteriormente, y considerando la recomendación de Wong (1985) se utilizó un total de  $B = 120$  repeticiones (*bootstrap* suavizado), con las cuales se estimó la significancia (valor de  $P$ ) del número de modas probado. En total, se examinaron 36,000 estimaciones de densidad.

Como guía inicial se consideraron dos niveles de significancia 0.10 propuesto por Wong (1985) y 0.40 propuesto por Izenman y Sommer (1988). Finalmente se contó el número de veces que se llegó a la distribución original simulada con base en los criterios citados. Para todos los cálculos se utilizaron los programas para el paquete estadístico *Stata* escritos por Salgado-Ugarte *et al.* (1993, 1995a, 1995b y 1997)

### 3 Resultados

Los resultados obtenidos se presentan de manera resumida en los siguientes cuadros

**CUADRO 1**  
Frecuencia del número significativo de modas sugerido por  
la prueba de Silverman. Nivel de significancia = 10%

<i>Modas</i>	1	2	3	4	5	<i>Ninguna</i>
1	<b>22</b>	3	0	0	0	0
2	0	<b>25</b>	0	0	0	0
3	10	5	<b>10</b>	0	0	0

En el cuadro 1 se observa que, con 10% de nivel de significancia, para distribuciones unimodales y bimodales se llega a la conclusión correcta en la mayoría de las veces (88% y 100% respectivamente). No obstante, para las distribuciones trimodales la proporción de aciertos es de sólo un 40%.

**CUADRO 2**  
Frecuencia del número significativo de modas sugerido por  
la prueba de Silverman. Nivel de significancia = 40%

<i>Modas</i>	1	2	3	4	5	<i>Ninguna</i>
1	<b>13</b>	5	3	0	0	4
2	0	<b>19</b>	3	1	0	0
3	1	1	<b>19</b>	3	0	1

En el cuadro 2, que muestra los resultados obtenidos con un nivel de 40% de significancia, se observa que el 52% de veces se acertó en la unimodal y para bimodal y trimodal la conclusión correcta se alcanzó en un 76% de las veces.

Las Figuras 1 a 6 muestran las estimaciones de densidad utilizando las bandas sugeridas por la prueba en dos grados decrecientes de ajuste con la original: para el caso unimodal Figuras 1 y 2, bimodal Figuras 3 y 4 y trimodal 5 y 6.

## 4 Discusión y Conclusiones

Estos resultados muestran que la prueba se comporta de manera diferente para cada valor de significancia. Con el primero, la prueba es más eficiente para distribuciones unimodales y bimodales; con el segundo, la eficiencia se manifiesta en mayor grado para distribuciones trimodales. Para buscar un valor que represente el mejor compromiso para la identificación correcta de distribuciones unimodales y multimodales se probaron varios valores diferentes.

El cuadro 3 contiene aquél sugerido por nuestros resultados:

**CUADRO 3**  
Frecuencia del número significativo de modas sugerido por  
la prueba de Silverman. Nivel de significancia = 20%

<i>Modas</i>	1	2	3	4	5	<i>Ninguna</i>
1	<b>18</b>	7	0	0	0	0
2	0	<b>24</b>	0	1	0	0
3	3	2	<b>19</b>	1	0	0

El valor de 20% para la significancia de referencia identifica con certeza en la mayor parte de las veces el número de modas, es decir se observa que el 72% de veces se acertó en la unimodal, el 96% en la bimodal y el 76% en la trimodal. Por lo anterior parece conveniente sugerir el valor de significancia del 20% para su empleo general en la determinación de la modalidad de distribuciones univariadas.

Aunque la prueba ha mostrado su utilidad con datos reales, se requiere de investigación adicional para tener una idea completa de su eficacia.

Como conclusión general puede afirmarse que si la muestra contiene información multimodal, el uso de los EDK's en combinación con reglas para determinación de amplitud de banda óptima, sobresuavizada y validación cruzada, en adición al uso de procedimientos no-paramétricos para la evaluación de multimodalidad tales como la prueba de Silverman parecen ser una valiosa herramienta para su extracción.

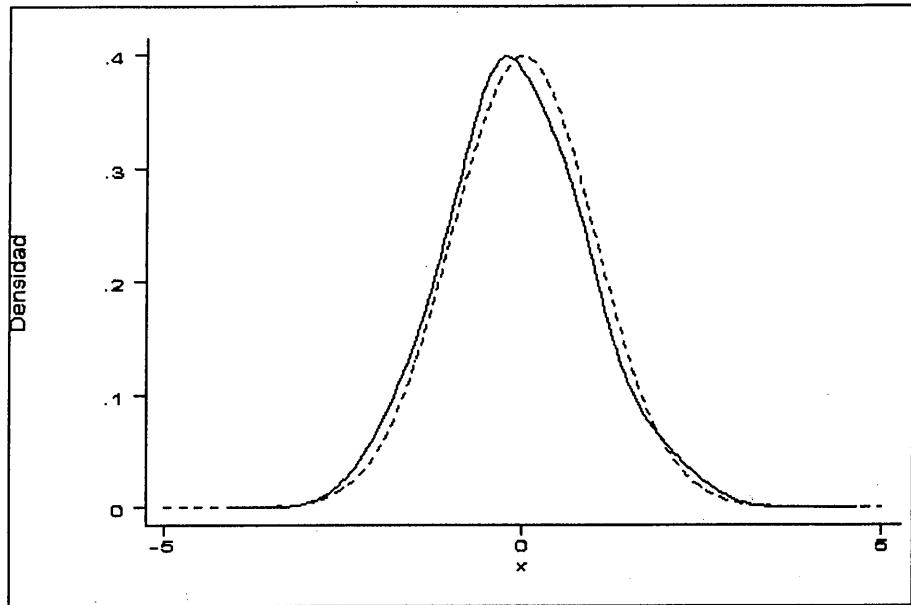


Figura 1: Estimación de densidad por kernel con la amplitud de banda sugerida por la prueba de Silverman comparada con la original (punteada)

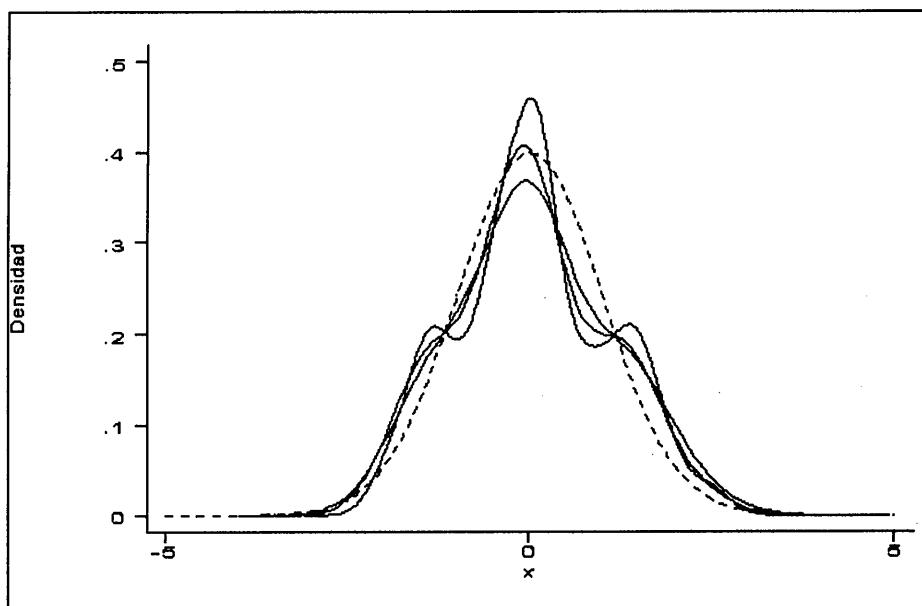


Figura 2: Estimaciones de densidad por kernel con la amplitud de banda sugerida por la prueba de Silverman comparada con la original (punteada)

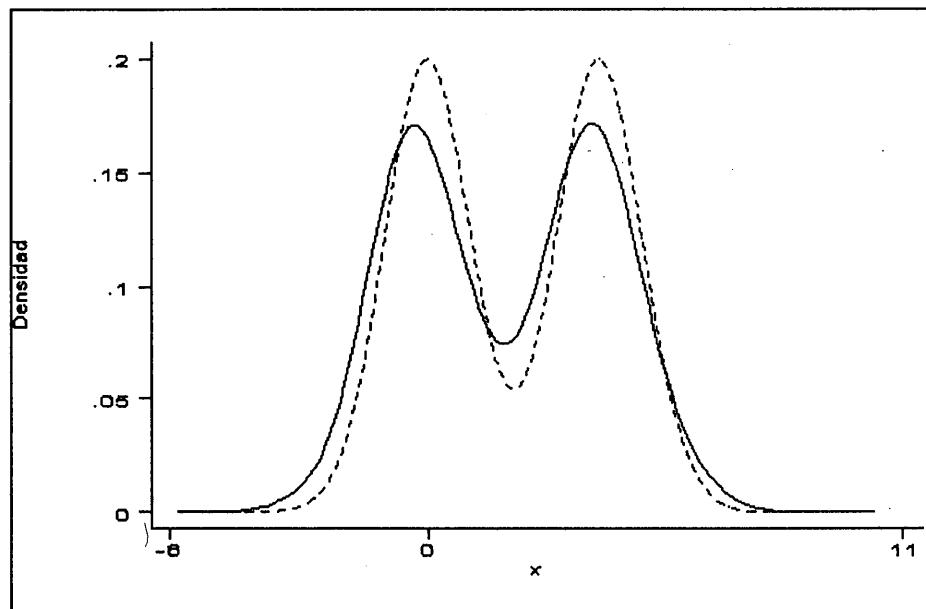


Figura 3: Estimación de densidad por kernel con la amplitud de banda sugerida por la prueba de Silverman comparada con la original (punteada)

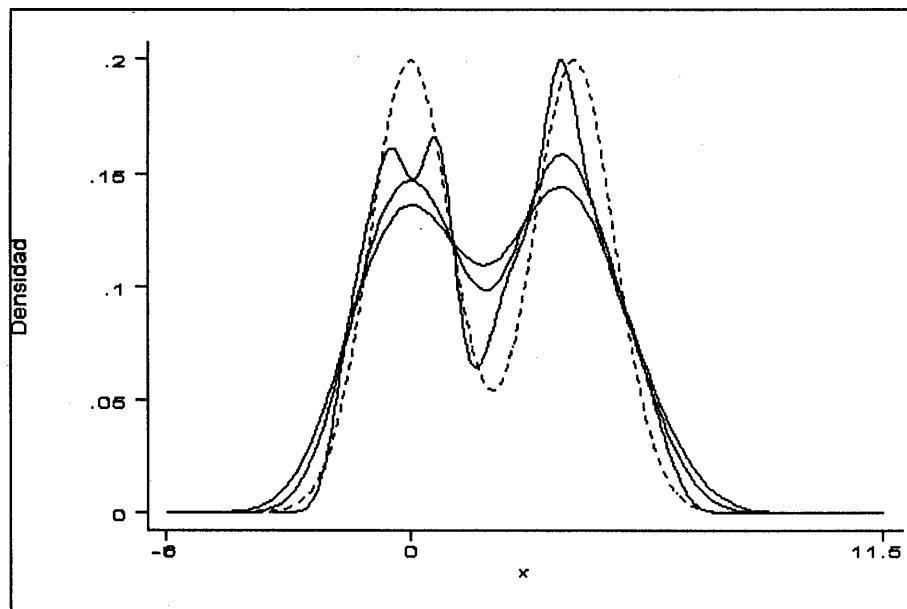


Figura 4: Estimaciones de densidad por kernel con la amplitud de banda sugerida por la prueba de Silverman comparada con la original (punteada)

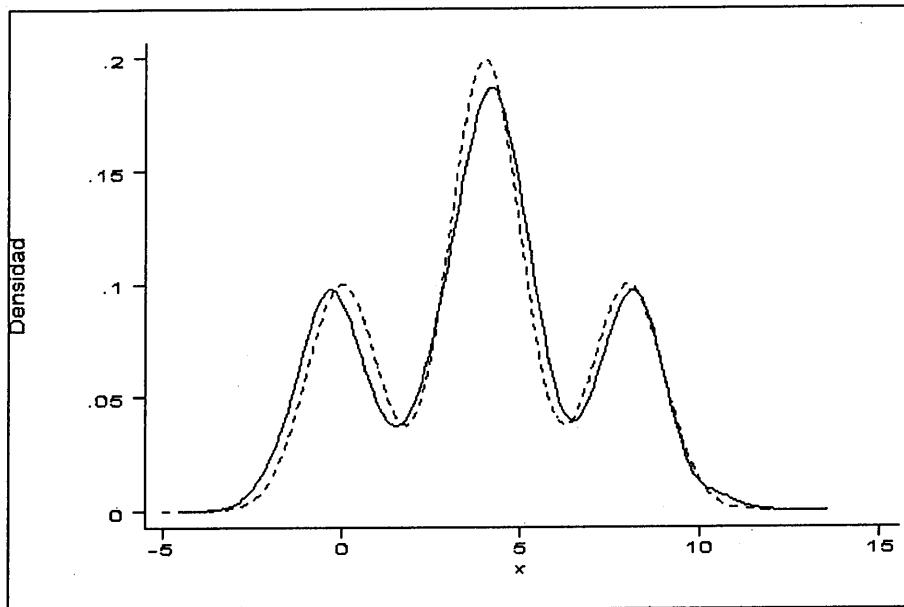


Figura 5: Estimación de densidad por kernel con la amplitud de banda sugerida por la prueba de Silverman comparada con la original (punteada)

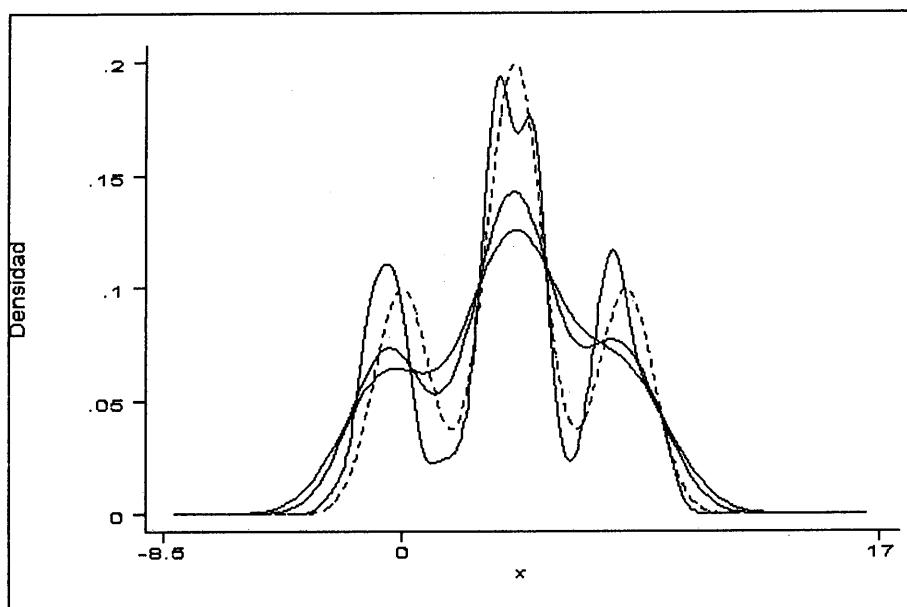


Figura 6. Estimación de densidad por kernel con la amplitud de banda sugerida por la prueba de Silverman comparada con la original (punteada)

## Referencias

- Good, I.J. y R.A. Gaskins (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association*, 75, 42-73.
- Härdle, W. (1991). *Smoothing Techniques. With Implementations in S.* New York: Springer-Verlag.
- Hartigan, J.A. y P.M. Hartigan, (1985). The Dip test of unimodality. *The annals of Statistics*, 13, 70-84.
- Izenman, A.J. y C. Sommer, (1988). Philatelic mixtures and multimodal densities. *Journal of the American Statistical Association*, 83, 941-953.
- Salgado-Ugarte, I.H., (1995). Nonparametric methods for fisheries data analysis and their application in conjunction with other statistical techniques to study biological data of the Japanese sea bass *Lateolabrax japonicus* in Tokyo Bay. Unpublished Ph.D. Dissertation. University of Tokyo, Faculty of Agriculture, Dept. of Fisheries.
- Salgado-Ugarte, I.H., M. Shimizu y T. Taniuchi. (1993). Exploring the shape of univariate data using kernel density estimators. *Stata Technical Bulletin* 16, 8-19.
- Salgado-Ugarte, I.H., M. Shimizu y T. Taniuchi, (1995a). ASH, WARPing, and kernel density estimation for univariate data. *Stata Technical Bulletin* 26, 2-10.
- Salgado-Ugarte, I.H., M. Shimizu y T. Taniuchi, (1995b). Practical rules for bandwidth selection in univariate density estimation. *Stata Technical Bulletin* 27, 5-19.
- Salgado-Ugarte, I.H., M. Shimizu y T. Taniuchi, (1997). Nonparametric assessment of multimodality for univariate data. *Stata Technical Bulletin* 38, 27-35.
- Scott, D.W. (1979). On optimal and data-based histograms. *Biometrika*, 66, 605-610.
- Scott, D.W. (1985). Frequency polygons: Theory and application. *Journal of the American Statistical Association*, 80, 348-354.
- Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* New York: Wiley.
- Silverman, B.W. (1978). Choosing the window width when estimating a density. *Biometrika*, 65, 1-11.
- Silverman, B.W. (1981a). Density estimation for univariate and bivariate data. In *Interpreting Multivariate Data*, (V. Barnett, ed.). Chichester: Wiley, 37-53.

- Silverman, B.W. (1981b). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society, B*, 43, 97-99.
- Silverman, B.W. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.
- Terrell, G.R., (1990). The maximal smoothing principle in density estimation. *Journal of the American Statistical Association*, 85, 470-477.
- Wong, M.A., (1985). A bootstrap testing procedure for investigating the number of subpopulations. *Journal of Statistical Computation and Simulation*. 22, 99-112.

# Estadística Ji-Cuadrada para Observaciones Emparejadas

Andrzej Matuszewski

*Institute of Computer Science, Polish Academy of Sciences*

## 1 Introducción

Se tienen dos instrumentos de medición que conllevan un error normal. El mejor (en calidad) es el instrumento con la menor varianza. Emparejamiento en este contexto se puede entender así. Cada objeto -supongamos  $n$  de ellos- se mide por ambos instrumentos dos veces. Las mediciones del mismo objeto hechas por el mismo instrumento tiene una alta correlación positiva. La calidad de cada instrumento está calculada usando la suma de los  $2n$  cuadrados de los errores. Se han desarrollado pruebas de hipótesis sobre la igualdad de varianzas. Se presentan propuestas sobre la generalización del problema y sus posibles soluciones.

Para facilitar nuestra presentación consideraremos el caso más clásico: los errores satisfacen el supuesto de normalidad y tienen esperanza igual a cero. Sin embargo, tomando en cuenta que la principal herramienta que empleamos en este trabajo es la estadística clásica de Ji-cuadrada, es sencillo aplicar nuestros resultados también para los casos de errores no normales y con una esperanza desconocida.

Para no trivializar el problema es conveniente introducir algún tipo de dependencia entre los errores. Vamos a suponer que el número de errores observados es par y que existe una alta correlación en un conjunto dado de parejas.

Existe una diferencia importante entre errores emparejados y medias emparejadas a las cuales se aplica la conocida estadística del tipo  $t$ -Student para muestras dependientes. Dicha diferencia tiene un aspecto práctico importante que a continuación se describe.

## 2 Varianza de la Ji-Cuadrada

Consideremos el caso estándar de estadística ji- cuadrada de valores de la muestra completa (i.e. de  $2n$  valores). El único parámetro desconocido de tal estadística ( $Y$ ) será la correlación ( $\rho$ ) entre las observaciones emparejadas.

$$Y = \sum_{i=1}^n X_i^2 + \sum_{i=1}^n (\rho X_i + \sqrt{1-\rho^2} X_{n+1})^2, \quad (1)$$

donde

$$\begin{aligned} Dist[X_1, X_2, \dots, X_{2n}]^1 &= N(O, I_{2n}) \\ I_k - \text{la matriz unitaria del orden k.} \end{aligned}$$

Obviamente

$$E[Y] = 2n. \quad (2)$$

Para calcular la varianza de  $Y$  hay que desarrollar los cuadrados en la parte derecha de la ecuación

$$Var[Y] = E(Y - 2n)^2 = E\left(\sum_{i=1}^n (X_i^2 - 1) + \sum_{i=1}^n ((\rho X_i + \sqrt{1-\rho^2} X_{i+n})^2 - 1)\right)^2.$$

El resultado final es que

$$Var[Y] = 4n(1 + \rho^2), \quad (3)$$

el cual se obtiene empleando relaciones conocidas entre variables normales.

Estos resultados relativamente simples contradicen nuestra intuición que proviene del típico análisis estadístico: comparación de esperanzas para muestras emparejadas a través de la estadística del tipo  $t$ -Student. A continuación consideramos los dos análisis subrayando ciertos detalles.

Por un lado la propiedad (2) parece contradecir nuestra intuición de que el emparejamiento debe incrementar la esperanza de la suma Ji-cuadrada. Esto tiene la siguiente consecuencia. En la derivación de Bose (1935) se concluye que una estadística muy similar a  $Y$  tiene una distribución muy conectada con una mezcla infinita de distribuciones. Johnson y Kotz (1972) sugieren entonces que la estadística  $Y$  tiene también una distribución que es una mezcla de distribuciones en donde todas ellas tienen esperanza mayor que  $2n$ . Esto es claramente imposible. Este razonamiento nos lleva a la conclusión de que vale la pena obtener la distribución exacta de  $Y$ .

Por otro lado, la varianza de  $Y$  (por (3)) es mayor o igual a la varianza de la suma de cuadrados de  $2n$  componentes independientes ya que la varianza de esta suma es igual a  $4n$ . Es decir, esto es lo contrario de lo que se observa en la práctica de la comparación de medias emparejadas. Allí la varianza de la diferencia entre dos medias es usualmente menor que para muestras independientes.

### 3 Problema de Dos Muestras

El problema más importante que nos interesa desarrollar en este trabajo es el siguiente: Se tienen dos muestras que representan dos instrumentos de medición:  $U, W$ . Para poder comparar la calidad de estos instrumentos hay que obtener algún tipo de información sobre las varianzas del mecanismo de los instrumentos.

El procedimiento estadístico típico que maneja parámetros del mecanismo (es decir de la población) es el procedimiento de pruebas de hipótesis. En este trabajo vamos a restringirnos a la solución del siguiente problema:

Empezamos definiendo en forma clásica una estadística que tiene como propósito comparar las varianzas de dos poblaciones. Luego se deriva la distribución de tal estadística bajo la hipótesis nula de igualdad de calidad de  $U$  y  $W$ .

A continuación se deriva la distribución de la estadística  $G$  bajo dicha hipótesis nula, donde  $G$  está definida como sigue

$$G = \frac{n_W \sum_{i=1}^{2n_U} U_i^2}{n_U \sum_{i=1}^{2n_W} W_i^2},$$

donde  $\text{Distr}[\{U_i\}, \{W_i\}] = \text{Normal}$  y las muestras representan los instrumentos mencionados.

Suponemos que todos los componentes de ambas muestras tienen esperanza igual a cero. Obviamente las componentes no están estandarizadas y debemos suponer que

$$\forall i = 1, 2, \dots, 2n_U, \text{Var}[U_i] = \sigma_U^2,$$

$$\forall i = 1, 2, \dots, 2n_W, \text{Var}[W_i] = \sigma_W^2.$$

El emparejamiento de los datos dentro de cada muestra está formalizado por

$$\forall i = 1, 2, \dots, 2_U \text{Cor}[U_i, U_{nu+1}] = \rho_U,$$

$$\forall i = 1, 2, \dots, 2_W \text{Cor}[W_i, W_{nw+1}] = \rho_W.$$

Cualquier otra pareja de componentes de ambas muestras tiene correlación igual a cero.

### **PROPIEDAD 1.**

$[H_0 \Leftrightarrow \sigma_U = \sigma_W = \sigma] \Rightarrow Distr[G] = F_{2n_U, 2n_W}$  no depende de  $\sigma$

**PROPIEDAD 2.** Bajo la misma hipótesis nula

a)  $\rho_U = \rho_W = 0 \Rightarrow Distr[G] = F_{2n_U, 2n_W}$

b)  $\rho_U = 1 \wedge \rho_W = 0 \Rightarrow Distr[G] = F_{n_U, n_W}$

Claro que se cumple también la propiedad simétrica

c)  $|\rho_U| = |\rho_W| = 1 \Rightarrow Distr[G] = F_{2n_U, 2n_W}$

d) Asintóticamente ( $n_u \rightarrow +\infty, n_W \rightarrow +\infty$ )

El último punto es una consecuencia de los puntos anteriores y de (3).

La **PROPIEDAD 2** nos dice que no tiene importancia el signo de la correlación. Esto no es así para el problema de comparación de esperanzas emparejadas (*t*-Student).

En base a este desarrollo se nos presentan dos alternativas:

1. El punto d) de la PROPIEDAD 2 nos indica que el procedimiento tiene un tamaño de prueba aproximado. Llamaremos a esta prueba: "realista".

2. Si uno quiere garantizar el nivel de significancia -es decir- se desea tener una prueba conservadora, debe aplicarse el punto c) de la misma propiedad. Esto es equivalente a tomar las medias de las dos componentes emparejadas de  $G$  y obtener así tamaños de muestras dos veces más cortas.

Parece razonable buscar que la solución sea un cierto compromiso entre los enfoques realista y conservador. Los trabajos de Mathai, Moschopoulos (1992) y Ong (1995) serán escenciales para realizar los cálculos analíticos y computacionales. Es muy probable que se quiera emplear simulación Monte-Carlo.

## **Referencias**

Bose, S.S. (1935). On the distribution of the ratio of variances of two samples drawn from a given normal bivariate correlated population. *Sankya*, 2, 65-72.

Johnson, N.L., Kotz, S. (1972). *Multivariate distributions*. New York: Wiley.

Mathai, A.M., Moschopoulos, P.G. (1992). A form of multivariate gamma distribution. *Ann. Inst. Statist. Math.*, 44, 97-106.

Ong, S.H. (1995). Computation of bivariate gamma and inverted beta distribution function. *JSCS*, 51, 153- 163.

# Modelos de Rasch y Log-lineales para Minería de Datos dentro del Formalismo de Dempster-Shafer<sup>1</sup>

Andrzej Matuszewski

y

Guillermo Morales

*Inst. de Ciencias Computacionales  
Academia Polaca de Ciencias  
Varsovia, Polonia*

*Centro de Investigación y  
Estudios Avanzados del IPN  
México, D.F., México*

## 1 Introducción

En trabajos anteriores - (Kłopotek, 1995), (Matuszewski y Kłopotek, 1995, Kłopotek, Matuszewski, Wierzchon, 1996 y Matuszewski, 1997) - hemos definido la noción de *independencia condicional* para variables D-S. Esta noción tiene un contenido no-trivial solamente para tres o más variables, en tanto que la noción de independencia para dos variables D-S tiene el mismo significado que el de independencia para dos variables clásicas.

En este trabajo hay una propuesta de formalizar el tipo de dependencia entre dos variables D-S. La formalización no es completamente conmutativa. Habrá entonces dos nombres para las dos variables.

La dependencia entre dos variables D-S es importante no sólo para modelar un esquema que explique el comportamiento de estas variables, sino también para plantear la hipótesis de que tal dependencia probada es una consecuencia de una tercera variable. Esto promueve trabajos constructivos sobre tres variables incluyendo las pruebas de independencia condicional.

## 2 Motivación

Vamos a considerar una típica configuración de datos que abre la posibilidad de utilizar el formalismo de Dempster-Shafer (e.g. Dempster, 1967). Acaso nuestra experiencia ordinaria adquirida como resultado de los contactos con la medicina será suficiente para presentar los detalles más significativos.

Consideraremos dos géneros de variables del tipo D-S, lo cual significa que ambos son multifunciones. A la población a la cual corresponden ambas variables la llamaremos

<sup>1</sup>Trabajo realizado con el patrocinio conjunto del Komitet Badań Naukowy (KBN) de Polonia y el Consejo Nacional de Ciencia y Tecnología (CONACyT) de México.

*pacientes*. Para tal población se puede considerar por un lado las *diagnosis* (de cierto tipo) y los *síntomas* por otro.

Supondremos que el conjunto de variables  $\{s(i)\}_i$  representa un grupo de síntomas indicados por números consecutivos. En el caso muy particular de este grupo, pueden pertenecer a él todos los síntomas que se consideran para la población específica de pacientes.

Los síntomas tienen dos características importantes:

1. Son de la mayor simplicidad posible. Cada síntoma existe o no. Asumen pues solamente dos valores (SI, NO).
2. Un valor SI es más informativo que un NO desde el punto de vista de poder establecer la diagnosis de un paciente.

El grupo de síntomas y las diagnosis - a las que denotaremos  $d(i)$  - pueden tomar más de un valor, vistas como variables del tipo D-S. En otras palabras, un paciente (digamos, el  $i$ -ésimo) puede tener dos o más valores de ambas diagnosis y síntomas. Sin embargo hay ciertas diferencias entre la variable  $s(i)$  - síntoma generalizado - y la variable  $d(i)$  - diagnosis.

Es teóricamente posible que unos pacientes no tengan ningún síntoma pues  $s(i)$  puede representar un tipo de síntomas y no todos. Es más, esto pasa en raras ocasiones, por lo que agrupamos en una sola variable D-S  $s(i)$  a todos los posibles síntomas que se considera para las enfermedades en cuestión. Por otro lado, no tiene sentido considerar pacientes que no tengan una o más diagnosis ya que el único objetivo de correlación es buscar qué tipo de influencia sobre los síntomas tienen las diferentes diagnosis. Claramente, el grupo de control que exista en varios experimentos se puede tratar como una diagnosis específica.

### 3 Modelos

El modelo que representa todas las posibles correlaciones para el conjunto de pacientes se puede definir en esta forma:

$$\log p_\beta(i, j, k) = \theta_{i,\beta} + c_\beta + \epsilon_{\beta,jk} \quad (1)$$

donde los índices tienen el significado siguiente:

- $i$  - paciente,  $i \leq N$ ,
- $j$  - diagnosis,  $j \leq D$ ,
- $k$  - síntoma,  $k \leq S$ ,
- $\beta$  - posibles eventos de  $p$  como una función de probabilidad, i.e.  

$$\sum_\beta p_\beta(i, j, k) = 1, \quad \forall i, j, k$$

Para los eventos que pueden ocurrir con probabilidad  $p$ , el índice  $\beta$  puede tomar los siguientes valores para  $i, j, k$  dados:

- 0 - el paciente  $i$  no tiene la diagnosis  $j$  ni tampoco el síntoma  $k$ ,
- 1 - el paciente  $i$  posee la diagnosis  $j$  mas no tiene el síntoma  $k$ ,
- 2 - al revés,
- 3 - el paciente  $i$  sí tiene los dos: la diagnosis  $j$  y el síntoma  $k$ , pero también algún otro síntoma (es el evento de la coincidencia general),
- 4 - el paciente  $i$  tiene la diagnosis  $j$  y el único síntoma  $k$  (es una coincidencia específica).

A tal índice lo vamos a llamar el *tipo de coincidencia*.

En otras palabras, la probabilidad básica que modelamos corresponde a tipos de coincidencia. La suma de probabilidades para estos tipos es pues igual a uno para cualquier paciente y cualquier combinación de diagnosis - síntoma.

Hay una cierta simetría entre las componentes en la ecuación (1), *i.e.*  $\forall i : \theta_{i0} = 0$ , y también  $\forall j, k : \epsilon_{0jk} = 0$ . Incluso sería posible imponer restricciones sobre las sumas de estas componentes.

La siguiente restricción tiene una gran importancia para las conclusiones prácticas del procedimiento. Las que discutiremos enseguida tienen una importancia mayor en cuanto al uso del modelo que en cuanto a las propiedades teóricas.

Las restricciones a los parámetros que más nos interesan son:  $\forall \beta, \sum_{j,k} \epsilon_{\beta jk} = 0$ .

De especial importancia es también

$$\epsilon_{\beta jk} > 0 \text{ para } \beta \in \{3, 4\}, \quad (2)$$

pues documentan coincidencias entre respectivas diagnosis y síntomas. En casos prácticos se puede decir que la diagnosis tiene influencia sobre el síntoma, aunque tal interpretación no está impuesta dentro del modelo. Como es práctica común en el análisis de correlaciones clásicas, la interpretación que tienen una y otra restricciones son: magnitud y nivel de significancia ( $p$ -valor) del parámetro. No es necesario imponer restricciones adicionales sobre las componentes del modelo (1) que simbolicen la magnitud de influencia sobre probabilidades que tengan los pacientes individuales. Esto es propiamente una consecuencia del siguiente segmento del razonamiento y del algoritmo computacional mismo.

El modelo (1) es un caso particular del modelo del Rasch. Históricamente (Rasch, 1961) la idea del Rasch (un sicólogo-estadístico danés) apareció como una herramienta de la llamada teoría de “tests” (con aplicaciones principalmente en la sicometría), pero pronto se encontró como una teoría dentro del centro mismo de estadística (Fischer y Molenaar, 1995). Para la estadística, lo importante es poder generalizar. En otras palabras, es más importante describir el “test” que a una persona en particular.

## 4 Algoritmo

Lo más sofisticado y fructífero es la eliminación de los parámetros de los objetos de la población. En nuestro lenguaje, esto significa que hemos de eliminar los parámetros

$$\theta_{i\beta}, \forall i, \beta \quad (3)$$

que simbolizan la influencia de los pacientes. El primer, y más importante, paso del procedimiento conectado con el modelo del Rasch se basa en el siguiente razonamiento consistente de dos puntos.

1. En la función de verosimilitud reemplazamos los parámetros (3) por sus respectivas estadísticas suficientes. Obtendremos así una forma bien conocida (Lehmann, 1959), ya que la distribución que hemos definido en (1) pertenece a la familia exponencial.
2. La estimación de los parámetros (2) y de los restantes va a tener un carácter condicional (estimadores de máxima verosimilitud condicional). Las propiedades óptimas de tales estimadores no son peores, en lo general, que las que corresponden a estimadores incondicionales (*e.g.* Andersen, 1970).

Para realizar cálculos que nos conduzcan a estimadores de los parámetros de interés y a las pruebas de hipótesis correspondientes (Andersen, 1972) es necesario preparar a los datos como sigue:

La primera matriz de frecuencias empíricas es un arreglo en cinco dimensiones dado por  $[n_{r_0r_1r_2r_3r_4}]$ , donde los índices han de cumplir con la restricción  $\sum_\alpha r_\alpha = D \cdot S$ . Cada entrada de la matriz se define como una frecuencia,

$$n_{r_0r_1r_2r_3r_4} = \text{card}\{i | \forall \alpha = 0, \dots, 4 : \text{card}(C_\alpha(i)) = r_\alpha\},$$

donde

- $i$  - índice del paciente,
- $\text{card}$  - cardinalidad (número de elementos) del conjunto,
- $C_\beta(i) = \{(j, k) \text{ de tipo } \beta | j \in d(i) \wedge k \in s(i)\},$
- $d(i)$  - multifunción de diagnosis del paciente  $i$ ,
- $s(i)$  - multifunción de síntomas del paciente  $i$ .

El otro conjunto de datos es más intuitivo. Consideremos el arreglo tridimensional

$$[m_{jk\beta}]_{jk\beta} \quad \text{con} \quad j \leq D, k \leq S, \beta = 0, \dots, 4, \quad (4)$$

donde,

$$m_{jk\beta} = \text{card}\{i | d(i) = j, s(i) = k, \beta = \text{tipo de incidencia } (i, j, k)\}$$

Tal matriz puede ser analizada mediante el modelo log-lineal. Sin embargo hay algunas dificultades en la construcción de algoritmos para este caso. Hay dependencias en la matriz (4). Una de las posibilidades para resolver estos problemas es introduciendo modelos *longitudinales y datos omisos* (Little, 1985, Conaway *et al.*, 1992).

## Referencias

- Andersen, E. B. (1972). Asymptotic properties of conditional maximum likelihood estimates. *J. R. Statist. Soc.*, B, 32, 283-301.
- Andersen, E. B. (1972). The numerical solution of a set of conditional estimation equations. *J. R. Statist. Soc.* B, 34, 42-54.
- Conaway, M. R., Waternaux, C., Allred, E., Bellinger D. y Leviton. (1992). Pre-natal blood lead levels and learning difficulties in children: an analysis of nonrandomly missing categorical data. *Statistics in Medicine*, 11, 799-811.
- Dempster, A. P. (1967). Upper and lower probabilities induced by a multi-valued mapping. *Ann. Math. Stat.*, 38, 325-339.
- Fischer, G. H. y Molenaar, I.W. (1995). *Rasch models: foundations, recent developments and applications*, New York: Springer-Verlang.
- Kłopotek, M. A. (1995). *Interpretation of belief function in Dempster Shafer theory*, Found. Comp. & Dec. Sc., 20, 289-306.
- Kłopotek, M. A., Matuszewski, A. y Wierzchon, S. T. (1996). *Overcoming negative-valued conditional belief functions when adapting traditional knowledge acquisition tools to Dempster-Shafer theory*, Proc. ESA96 IMACS Multiconference, 2, 948-953
- Lehmann, E. (1959). *Testing statistical hypotheses*. New York: Wiley.
- Little R. (1985). Nonresponse adjustments in longitudinal surveys: models for categorical data. *Bull. Int. Stat. Inst.*, 15, 1-15.
- Matuszewski, A., (1997). Consecuencias estadísticas de análisis de tres variables dentro del formalismo de Dempster-Shafer. *Memorias del XI Foro Nacional de Estadística*, Sinaloa, México. Aguascalientes:INEGI, 99-104.
- Matuszewski, A. y Kłopotek, M., (1995). Factorization of Dempster-Shafer belief functions based on data. *ICS PAS Reports No 798*, Warsaw.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proc. Fourth Berkeley Symp.*, 4, 321-333.

# Definition of some Integral U-statistics for Tests of Independence

David Mayer-Foulkes

*Centro de Investigación y Docencia Económicas*

## 1 Introduction

The interest in statistics capable of detecting non-linear dynamics is now well-established. Developing Grassberger and Procaccia's (G&P) (1983) Correlation Dimension (CD), Brock, Dechert and Sheinkman (BDS) defined a statistic testing the IID null whose applications include testing for non-linearity in stochastic processes. Combining these two approaches Mayer (1995) defined the Correlation Dimension Ratio (CDR) (or Statistical Correlation Dimension), a statistic which tests the IID null, calculates dimensions greater than 1, and eliminates a downward bias present in the G&P and BDS statistics. In a parallel development, B. Mizrach defined the Simple Non-parametric Test (SNT), a simpler version of these U-statistics which can be applied for the same purposes and involves less calculation. The numerical methods introduced by Mayer (1995) to calculate the distance histogram  $C(m, \varepsilon)$  used in these statistics obtains it for many distance values  $\varepsilon$  simultaneously, and recursively in the dimension  $m$  (see the definitions in the next section), leading to the question whether the information thus obtained can be used more effectively.

In this report, we define some *integral U-statistics*, which take averages along the  $\varepsilon$  variable. In addition, a homogenization process is introduced previous to the application of these statistics, rendering their distributions independent of the stationary process being tested. *The objective is to define statistics capable of testing for non-linearity for which confidence intervals can be obtained universally, either by theoretical means or by Monte-Carlo experiments.* Here we only define the integrals and prove some general properties. A Windows computer program to calculate these statistics is available from the author<sup>1</sup>, who has conducted Monte-Carlo tests on these statistics.

---

<sup>1</sup>Program development supported by CONACyT research project 3416P-S9607.

## 2 Definitions

Let  $\mathbf{Z}^p = (Z_1^p, \dots, Z_m^p)$ ,  $p = 1, \dots, N$  be  $N$  copies of an  $m$ -dimensional multivariate random variable  $\mathbf{Z}$ . Let  $I$  be the indicator function,  $I(x, y) = 1$  if  $x \leq y$ ,  $I(x, y) = 0$  if  $x > y$ . Define the sets  $\mathfrak{B}^j = \{h_i^j, c_i^j, H^j, C^j\}$  of *order j* “building block” random variables as follows. Each  $g^j \in \mathfrak{B}^j$  is given by

$$g^1(\mathbf{Z}, \mathbf{z}_0, \varepsilon, N) = \frac{1}{N} \sum_{p=1}^N I(E_p, \varepsilon), \quad g^2(\mathbf{Z}, \varepsilon, N) = \frac{2}{N(N-1)} \sum_{p < q} I(E_{pq}, \varepsilon),$$

where the expressions  $E_p$ ,  $E_{pq}$  are given by cases in the following table:

$g^1 \in \mathfrak{B}^1$	$E_p^1$	$g^2 \in \mathfrak{B}^2$	$E_{pq}$
$h_i^1$	$Z_i^p - z_{0i}$	$h_i^2$	$Z_i^p - Z_i^q$
$c_i^1$	$ Z_i^p - z_{0i} $	$c_i^2$	$ Z_i^p - Z_i^q $
$H^1$	$\max_{1 \leq i \leq m} (Z_i^p - z_{0i})$	$H^2$	$\max_{1 \leq i \leq m} (Z_i^p - Z_i^q)$
$C^1$	$\max_{1 \leq i \leq m}  Z_i^p - z_{0i} $	$C^2$	$\max_{1 \leq i \leq m}  Z_i^p - Z_i^q $

$i = 1, \dots, m$ . These random variables can be used to define other statistics, such as the SNT, BDS, CD and CDR statistics. In the case of the order 1 random variables we shall omit the variable  $\mathbf{z}_0$  unless explicitly needed.

## 3 Homogenization of Multivariate Random Variables

We shall write  $\mathfrak{U} : [0, 1] \rightarrow [0, 1]$  and  $\mathfrak{N} : \mathbb{R}_E \rightarrow [0, 1]$  for the accumulated density functions of the standard uniform and normal distributions respectively,  $\mathfrak{U}(z) = z$ ,  $\mathfrak{N}(z) = \frac{2}{\sqrt{\pi}} \int_{-\infty}^z \exp(-t^2) dt$ , where  $\mathbb{R}_E = \mathbb{R} \cup \{-\infty, \infty\}$  is the extended real line with the one-point compactification topology on each extreme (we can thus write  $\mathfrak{N}(\pm\infty)$ ,  $\mathfrak{N}^{-1}(0)$ ,  $\mathfrak{N}^{-1}(1)$ , and write about the uniform and normal distributions in the same terms).

**Definition 1**  $\mathbf{Z}$  is ammenable to homogenizations if each function  $p_i(z) = P(Z_i \leq z)$ ,  $i = 1, \dots, m$  is continuous and surjective from its domain (possibly  $\mathbb{R}_E$ ) onto  $[0, 1]$ . ■

For any continuous, increasing surjective functions  $G_i : I_i \rightarrow [0, 1]$ , where  $I_i \subseteq \mathbb{R}_E$  are closed intervals,  $i = 1, \dots, m$  define the random variables

$$X_i = (G_i^{-1} \circ p_i)(Z_i), \quad i = 1, \dots, m,$$

with  $G_i^{-1}$  an increasing, semi-continuous function defined by  $G_i^{-1}(y) = \inf\{x \mid G(x) = y\}$  satisfying  $G \circ G_i^{-1} = id$ . Observe that the random variables  $X_i$  have accumulated density function  $G_i$ , since

$$\begin{aligned} P(X_i \leq x) &= P(G_i^{-1}(p_i(Z_i)) \leq x) = P(Z_i \leq p_i^{-1}(G_i(x))) \\ &= p_i(p_i^{-1}(G_i(x))) = G_i(x) \end{aligned}$$

(where  $p_i^{-1}$  are defined like  $G_i^{-1}$ ). We thus have the following definition.

**Definition 2** Write  $\mathbf{G} = (G_1, \dots, G_m)$ . For any  $\mathbf{Z}$  ammenable to homogenizations, let its  $\mathbf{G}$ -homogenization be the multivariate random variable  $\mathbf{X} = (X_1, \dots, X_m)$ . We write  $\mathbf{X} = \mathfrak{H}_G(\mathbf{Z})$ . ■

In particular, if  $G_i$  is  $\mathfrak{U}$  or  $\mathfrak{N}$ ,  $X_i$  is uniformly distributed on  $[0, 1]$  or follows the standard normal distribution.

**Theorem 1** If  $Z_1, \dots, Z_m$  are identical random variables,  $G_1 = \dots = G_m = G$ , and  $\mathbf{z}_0 = 0$  or  $\mathbf{z}_0 = (1, \dots, 1)$ , then the order 1 building block random variables are invariant under homogenization:

$$g^1(\mathbf{Z}, \mathbf{z}_0, \varepsilon, N) = g^1(\mathbf{X}, \mathbf{x}_0, G^{-1}(p(\varepsilon)), N), \text{ for any } g^1 \in \mathfrak{B}^1$$

where  $\mathbf{x}_0 = \varphi(\mathbf{z}_0)$  and  $p$  is any of the functions  $p_i$  defined above.

*Proof:* Apply the equalities

$$I(Z_i, \varepsilon) = I(X_i, G^{-1}(p(\varepsilon))), \quad I(\max_{WD} Z_{1 \leq i \leq m} Z_i, \varepsilon) = I(\max_{WD} X_{1 \leq i \leq m} X_i, G^{-1}(p(\varepsilon))). ■$$

For order two statistics, if we suppose that the functions  $p_i$  defined above and  $\mathbf{G}$  are diffeomorphisms the homogenizing transformation  $\mathfrak{H}_G$  is equivalent to applying the diffeomorphism  $\varphi = (G_1^{-1} \circ p_1, \dots, G_m^{-1} \circ p_m)$ . Applying Brock's results (1986a, b), dimension measures such as the CD and CDR are invariant under homogenization.

## 4 Homogenized Integral U-Statistics

We explore variants of the SNT and BDS statistics when they are applied *after* one of the homogenization processes  $\mathfrak{H}_{\mathfrak{U}}$  or  $\mathfrak{H}_{\mathfrak{N}}$ . The idea is that, after the random variables  $\mathbf{Z}$  are transformed to  $\mathbf{X}$ , each  $X_i$  becomes a uniform or normal random variable. Thus the properties of these homogenized statistics will no longer depend on the particular distribution of  $\mathbf{Z}$ . Under some conditions the distributions of  $C^j(\mathbf{Z}, \varepsilon, N)$  and  $H^j(\mathbf{Z}, \varepsilon, N)$  will be known in the IID case and thus the confidence intervals of the homogenized statistics will be simpler to obtain. Even if the distributions are unknown, results derived by Monte-Carlo methods will be simultaneously applicable to all ammenable distributions  $\mathbf{Z}$ .

The statistics we define are *integral* because we shall consider different kinds of averages along  $\varepsilon$ . Suppose the components  $Z_1, \dots, Z_m$  of  $\mathbf{Z}$  are identical. Let  $\mathfrak{B}^j = \{c^j, h^j, C^j, H^j\}$  be the set of building block random variables of order  $j$  (we omit the index  $i$  of  $c_i^j, h_i^j$ , since now these random variables are identical). Given a partition  $\varepsilon_0 < \dots < \varepsilon_I$  of the interval  $[\varepsilon_0, \varepsilon_I]$ , where  $I \in \mathbb{N}$ , let

$$\Delta g^j(\mathbf{Z}, \varepsilon_k, N) = g^j(\mathbf{Z}, \varepsilon_k, N) - g^j(\mathbf{Z}, \varepsilon_{k-1}, N) \text{ for } g^j \in \mathfrak{B}^j,$$

for  $k = 1, \dots, I$ . Write  $\mathfrak{b}^j$  for  $(c^j, h^j, C^j, H^j)$  (the vector of random variables) and let  $f_1 : \mathbb{R} \times R^4 \rightarrow \mathbb{R}$  be any function. Suppose that  $f_1(\mathbf{Z}, \varepsilon, N) = f_1(\varepsilon, \mathfrak{b}^j(\mathbf{Z}, \varepsilon, N))$  is increasing in  $\varepsilon$ , and satisfies  $[f_1(\mathbf{Z}, \varepsilon_0, N), f_1(\mathbf{Z}, \varepsilon_I, N)] = [a, b]$ . Let  $f_2 : \mathbb{R} \times R^8 \rightarrow \mathbb{R}$  be any function. Now define

$$\mathfrak{S}_{f_1, f_2}^1(\mathbf{Z}, a, b, N) = \sum_{k=1}^M f_2(\varepsilon_k, (\mathfrak{b}^1, \Delta \mathfrak{b}^1)(\mathbf{Z}, \varepsilon_k, N)) \Delta f_1(\varepsilon_k, \mathfrak{b}^1(\mathbf{Z}, \varepsilon_k, N)), \quad (1)$$

where  $\Delta \mathfrak{b}^1 = (\Delta c^1, \Delta h^1, C \Delta^1, \Delta H^1)$ . It is clear that

$$\mathfrak{S}_{f_1, f_2}^1(\mathbf{Z}, \mathbf{z}_0, a, b, N) = \mathfrak{S}_{f_1, f_2}^1(\mathbf{X}, \mathbf{x}_0, a, b, N)$$

and that

$$\lim_{I \rightarrow \infty} \lim_{N \rightarrow \infty} \mathfrak{S}_{f_1, f_2}^1(\mathbf{Z}, a, b, N) = \lim_{M \rightarrow \infty} E(\mathfrak{S}_{f_1, f_2}^1(\mathbf{Z}, a, b, N)) = \mathcal{I}_{f_1, f_2}^1(\mathbf{Z}, a, b),$$

where

$$\mathcal{I}_{f_1, f_2}^1(\mathbf{Z}, a, b) = \int_{f_1(\varepsilon, \mathfrak{b}^1(\mathbf{Z}, \varepsilon)) \in [a, b]} f_2(\varepsilon, (\mathfrak{b}^1, \Delta \mathfrak{b}^1)(\mathbf{Z}, \varepsilon)) df_1(\varepsilon, \mathfrak{b}^1(\mathbf{Z}, \varepsilon))$$

and

$$g^1(\mathbf{Z}, \varepsilon) = \lim_{N \rightarrow \infty} g^1(\mathbf{Z}, \varepsilon, N) = E(g^1(\mathbf{Z}, \varepsilon, N))$$

for  $g^1 \in \mathfrak{b}^1$ . Thus the first-order  $U$ -statistics given by these sums are invariant under  $\mathfrak{H}_G$  and converge to the corresponding integrals.

In the case of order 2 statistics we shall use

$$\mathfrak{S}_{f_1, f_2}^2(\mathbf{Z}, a, b, N) = \mathfrak{S}_{f_1, f_2}^2(\mathfrak{H}_G(\mathbf{Z}), a, b, N)$$

as a definition instead of as a result, with the right hand side defined by (1) (with 1's corresponding to the order of the statistic replaced by 2's).

In some cases, nonlinear functions of several of these sums will be used. We shall refer to any of these functions as integral  $U$ -statistics. Amongst the examples we consider in our Monte-Carlo studies are approximations to the following integrals, where the limit  $N \rightarrow \infty$  has been taken on the right hand side:

$$\begin{aligned}\mathfrak{I}_1^j(\mathbf{Z}) &= \|C^j(\mathbf{Z})/c^j(\mathbf{Z})^m - 1\|_{L^p} & \mathfrak{I}_4^j(\mathbf{Z}) &= \|Ln(C^j(\mathbf{Z})) - mLn(c^j(\mathbf{Z}))\|_{L^p} \\ \mathfrak{I}_2^j(\mathbf{Z}) &= \|C^j(\mathbf{Z}) - c^j(\mathbf{Z})\|_{L^p}^m & \mathfrak{I}_5^j(\mathbf{Z}) &= \|C^j(\mathbf{Z})\|_{L^p} / \|c^j(\mathbf{Z})^m\|_{L^p} \\ \mathfrak{I}_3^j(\mathbf{Z}) &= \left\| \frac{Ln(C^j(\mathbf{Z}))}{mLn(c^j(\mathbf{Z}))} - 1 \right\|_{L^p} & \mathfrak{I}_6^j(\mathbf{Z}) &= \|Ln(C^j(\mathbf{Z}))\|_{L^p} / \|mLn(c^j(\mathbf{Z}))\|_{L^p}.\end{aligned}$$

The  $L^p$  measures are integrals along the variable  $\varepsilon$ , which has been omitted. These correspond to choosing in our definition  $f_1(\varepsilon, c, h, C, H) = H$ , which means we use the measure  $dH^j(\mathbf{Z}, \varepsilon, N)$ , and functions  $f_2$  given by  $|C/c^m - 1|^p$ ,  $|C - c^m|^p$ , etc. In practice we calculate approximations of the integrals, obtained using the Riemann sums. Let us refer to some instance of these approximations by  $\mathfrak{S}_i^j(\mathbf{Z}, \mathbf{N})$ ,  $i = 1, \dots, 6$ .

**Theorem 2** *Let  $Z_i$  be a strictly stationary process which is absolutely regular. (A definition is omitted for brevity. There are alternative conditions on the rate of decay of dependence over time yielding the same result, See Denker and Keller, 1983, p. 507). Generically, on intervals  $[a, b]$  on which  $f_1$  and  $f_2$  are non-singular (and therefore smooth) the statistics  $\mathfrak{S}_i^j(\mathbf{Z}, \mathbf{N})$ ,  $i = 1, \dots, 6$ , are asymptotically normal as  $N \rightarrow \infty$ . If  $\mathbf{Z}$  are IID then  $E(\mathfrak{S}_i^j(\mathbf{Z}, \mathbf{N}))$  is 0 for  $i = 1, \dots, 4$  and 1 for  $i = 5, 6$ .*

*Proof:*  $\mathfrak{S}_i^j(\mathbf{Z}, \mathbf{N})$  are asymptotically normal because they are smooth functions of the building block statistics, which are themselves asymptotically normal (Brock, 1986a, b). The means are obtained by replacing the component statistics with their means, and using

$$E(H^j(\mathbf{Z}, \varepsilon, N)) = E(h^j(\mathbf{Z}, \varepsilon, N))^m, E(C^j(\mathbf{Z}, \varepsilon, N)) = E(c^j(\mathbf{Z}, \varepsilon, N))^m. \blacksquare$$

## References

- Brock, W. A. (1986a). Distinguishing Random and Deterministic Systems: Abridged Version. *Journal of Economic Theory* 40 168-195.
- Brock, W. A. (1986b). Theorems on Distinguishing Deterministic from Random Systems. *Dynamic Econometric Modelling, Proceedings of the third International Symposium in Economic Theory and Econometrics* (Edited by W. A. Barnett, E. R. Berndt and H. White) Cambridge: University Press. 247-265.
- Denker, M. and Keller, G. (1983). On U-statistics and V. Misses Statistics for Weakly Dependent Processes. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 64, 505-522.
- Grassberger, P. and Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica 9D*, 189-208.

Mayer-Foulkes, D. (1995). A Statistical Correlation Dimension. *Journal of Empirical Finance* 2, 277-293.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

# Inferencia Bayesiana para el Cociente de las Medias de Dos Poblaciones Normales con Varianzas Distintas

M. Mendoza y E. Gutiérrez-Peña

*ITAM*              *IIMAS, UNAM*

## 1 Introducción

Sean  $\mathbf{X} = (X_1, \dots, X_n)$  y  $\mathbf{Y} = (Y_1, \dots, Y_m)$  dos muestras aleatorias independientes tales que

$$\begin{aligned} X_i &\sim N(x|\mu, \sigma^2); & i = 1, \dots, n \\ Y_j &\sim N(y|\eta, \sigma^2); & j = 1, \dots, m \end{aligned} \tag{1}$$

donde  $\mu$ ,  $\eta$  y  $\sigma^2$  son desconocidos. Bajo el supuesto de que  $\eta \neq 0$ , el propósito es producir inferencias sobre el parámetro  $\phi = \mu/\eta$  que describe la magnitud relativa de las medias. Usualmente  $\mu$  y  $\eta$  tienen el mismo signo y puede suponerse, sin pérdida de generalidad, que son positivas. Desde un punto de vista frequentista, la estimación puntual de  $\phi$  es simple. En la estimación por intervalos, sin embargo, existen algunos problemas metodológicos (Fieller, 1954). Resulta sorprendente que, a pesar de estas dificultades, la fórmula de Fieller y algunas de sus generalizaciones siguen siendo utilizadas (véase Raftery y Schweder 1993, como un ejemplo reciente).

Este problema ha sido abordado, desde una perspectiva Bayesiana, por Kappenman, Geisser y Antle (1970) y Bernardo (1977), entre otros. Naturalmente, estas contribuciones no tienen los problemas del enfoque frequentista. Un supuesto común en estas contribuciones es el de homoscedasticidad.

## 2 El Modelo Básico

### 2.1 El método de Cox

Como una generalización del modelo (1), Cox (1985) supone que  $X_1, \dots, X_n$  y  $Y_1, \dots, Y_m$  son dos muestras aleatorias independientes tales que

$$\begin{aligned} X_i &\sim N(x|\mu, \sigma_1^2); & i = 1, \dots, n \\ Y_j &\sim N(y|\eta, \sigma_2^2); & j = 1, \dots, m \end{aligned} \tag{2}$$

con  $\sigma_1^2 = \rho^2(c + \mu)^k$  y  $\sigma_2^2 = \rho^2(c + \eta)^k$ , donde  $c$  y  $k$  son constantes conocidas. Cox afirma que el valor de  $c$  debe garantizar que las dos varianzas son positivas y sugiere utilizar  $c = 0$  o  $c = 1$ . Sin embargo, en la mayoría de los casos tanto  $\mu$  como  $\eta$  pueden suponerse positivas y, en consecuencia, la constante  $c$  no es necesaria. De cualquier manera, (1) es un caso particular de este modelo con  $c = k = 0$ .

Cox encuentra intervalos de confianza para  $(c + \mu)/(c + \eta)$ , que coincide con el cociente de medias sólo si  $c = 0$ , y sólo discute en detalle el caso  $k = 2$ . Es interesante notar que en esta situación las dos poblaciones tienen el mismo coeficiente de variación,  $\rho$ . La contribución de Cox se basa por completo en la fórmula de Fieller.

Aquí la atención se centra en el cociente de medias,  $\phi = \mu/\eta$ , y sólo se considera el caso  $k = 2$  ya que el caso  $k > 2$  típicamente no es de interés práctico y presenta más dificultades que el modelo general.

## 2.2 Un análisis Bayesiano del modelo de Cox

En este problema el parámetro de interés es  $\phi = \mu/\eta$ , aún cuando el modelo está parametrizado en términos de  $\mu, \eta$  and  $\rho$ . Así, la función de verosimilitud para  $(\phi, \eta, \rho)$  está dada por

$$L(\phi, \eta, \rho) \propto \rho^{-(n+m)} \phi^{-n} \eta^{-(n+m)} \\ \times \exp \left\{ -\frac{1}{2\rho^2\phi^2\eta^2} \left( \sum_{i=1}^n (x_i - \phi\eta)^2 + \phi^2 \sum_{j=1}^m (y_j - \eta)^2 \right) \right\}. \quad (3)$$

Cualquier inferencia Bayesiana sobre  $\phi$  debe partir de la correspondiente distribución marginal final. Para obtener la final conjunta para el vector completo de parámetros, se debe asignar una conjunta inicial sobre  $(\phi, \eta, \rho)$ . Para facilitar la comparación con los resultados frecuentistas, es conveniente utilizar una inicial no informativa. Existen distintos procedimientos para producir iniciales de este tipo. De entre ellos, el método de iniciales de referencia propuesto por Berger y Bernardo (1992) reconoce que los parámetros de interés tienen un papel distinto del que corresponde a los parámetros de ruido, además de ofrecer otras ventajas.

La aplicación del método de referencia requiere un ordenamiento de los parámetros de acuerdo con su importancia inferencial. Aquí se considera la parametrización ordenada  $(\phi, \eta, \rho)$ , tomando en cuenta de esta manera que  $\phi$  es el parámetro de interés. Se puede comprobar que la inicial de referencia para este problema está dada por

$$\pi(\phi, \eta, \rho) \propto \phi^{-1} \eta^{-1} \rho^{-1}; \quad \phi > 0, \eta > 0, \rho > 0, \quad (4)$$

que conduce a la final conjunta

$$\pi(\phi, \eta, \rho | \mathbf{x}, \mathbf{y}) \propto \rho^{-(n+m+1)} \phi^{-(n+1)} \eta^{-(n+m+1)} \\ \times \exp \left\{ -\frac{1}{2\rho^2\phi^2\eta^2} \left( \sum_{i=1}^n (x_i - \phi\eta)^2 + \phi^2 \sum_{j=1}^m (y_j - \eta)^2 \right) \right\}. \quad (5)$$

Desafortunadamente, esta final es impropia para *todo* valor de  $n$  y  $m$  y carece, por tanto, de todo valor inferencial. A pesar de este hecho, el resultado sugiere una inicial de la forma

$$\pi(\phi, \eta, \rho) \propto \phi^{-a} \eta^{-b} \rho^{-c}; \quad \phi > 0, \eta > 0, \rho > 0,$$

donde  $a$ ,  $b$ , y  $c$  son constantes apropiadas. Existen distintos valores que conducen a una distribución final propia. En particular, basta con tomar  $a = 1$ ,  $b < 1$  y  $c = 1$ . En adelante se considerará el caso  $b = 1/2$ , aunque se puede utilizar cualquier otro valor de  $b$  cercano a uno. La final que corresponde a esta elección está dada por

$$\begin{aligned} \pi(\phi, \eta, \rho | \mathbf{x}, \mathbf{y}) &\propto \rho^{-(n+m+1)} \phi^{-(n+1)} \eta^{-(n+m+1/2)} \\ &\times \exp \left\{ -\frac{1}{2\rho^2} \left[ \frac{n}{\phi^2 \eta^2} \{s_x^2 + (\phi\eta - \bar{x})^2\} + \frac{m}{\eta^2} \{s_y^2 + (\eta - \bar{y})^2\} \right] \right\}. \end{aligned} \quad (6)$$

con  $ns_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2$  y  $ms_y^2 = \sum_{j=1}^m (y_j - \bar{y})^2$ , donde  $\bar{x}$  y  $\bar{y}$  son las medias muestrales. En esta expresión  $\rho$  puede ser integrada analíticamente, de manera que se obtiene la final marginal

$$\begin{aligned} \pi(\phi, \eta | \mathbf{x}, \mathbf{y}) &\propto \phi^{-(n+1)} \eta^{-(m+n+1/2)} \\ &\times \left[ \frac{n}{\phi^2 \eta^2} \{s_x^2 + (\phi\eta - \bar{x})^2\} + \frac{m}{\eta^2} \{s_y^2 + (\eta - \bar{y})^2\} \right]^{-(n+m)/2} \end{aligned} \quad (7)$$

para  $\phi$  y  $\eta$ . En esta expresión  $\eta$  no puede eliminarse de la misma manera. En consecuencia, la distribución marginal  $\pi(\phi | \mathbf{x}, \mathbf{y})$  debe aproximarse numéricamente. Con la finalidad de exhibir el tipo de resultados que se pueden obtener con este enfoque, se analiza ahora un conjunto de datos previamente considerado por Cox (1985).

*Ejemplo 1.* En la Tabla 1 se presenta una muestra de datos que se refieren al porcentaje de grava fina en dos tipos de suelo. Para estos datos,  $\bar{x} = 10.91$ ,  $\bar{y} = 3.94$ ,  $s_x^2 = 34.39$  y  $s_y^2 = 5.96$ .

Tabla 1: Datos de grava

Bueno	5.9	3.8	6.5	18.3	18.2	16.1	7.6
Malo	7.6	0.4	1.1	3.2	6.5	4.1	4.7

Se generó una muestra de tamaño 2000 de la distribución final (7) utilizando el algoritmo de Metropolis-Hastings (Gilks, Richardson y Spiegelhalter, 1996). Con la finalidad de obtener una distribución más parecida a una normal el modelo fue reparametrizado en términos de  $\psi = \log \phi$  y  $\lambda = \log \eta$ . Se utilizó el muestreador de independencia y una aproximación normal bivariada para  $\pi(\psi, \lambda | \mathbf{x}, \mathbf{y})$  con media igual a la verdadera moda final y matriz de covarianzas igual a menos 1.5 veces la inversa del Hessiano evaluado en la moda. El correspondiente intervalo de máxima densidad al 95% (Bayesiano I) aparece en la Tabla 2.

### 3 El Modelo General

Cox supone que el valor de  $k$  es arbitrario pero conocido y sólo considera el caso  $k = 2$ . Cox no presenta una solución explícita para  $k > 2$  y en cualquier caso, no considera el problema de asignar el valor de  $k$ .

En general,  $k$  es desconocido. Sin embargo, el problema inferencial asociado resulta intratable. Además, si  $k$  es desconocido, la relación entre la media y la varianza es cuestionable. Como alternativa, el modelo de Cox puede considerarse un caso particular del modelo en que las varianzas son distintas. De hecho, el análisis de este modelo general da lugar a una solución más robusta del problema original que no involucra ninguna hipótesis estructural.

En este trabajo se analiza el modelo general (2) donde  $\sigma_1 > 0$  y  $\sigma_2 > 0$  son desconocidas y no se supone ninguna relación funcional con las medias. Para este modelo se tiene la función de verosimilitud

$$L(\phi, \eta, \sigma_1, \sigma_2) \propto \sigma_1^{-n} \sigma_2^{-m} \exp \left\{ -\frac{1}{2} \left( \frac{1}{\sigma_1^2} \sum_{i=1}^n (x_i - \phi\eta)^2 + \frac{1}{\sigma_2^2} \sum_{j=1}^m (y_j - \eta)^2 \right) \right\}. \quad (8)$$

Si se considera la parametrización ordenada  $(\phi, \eta, \sigma_1, \sigma_2)$ , la inicial de referencia está dada por

$$\pi(\phi, \eta, \sigma_1, \sigma_2) \propto \phi^{-1/2} \sigma_1^{-1} \sigma_2^{-1}; \quad \phi > 0, \eta > 0, \sigma_1 > 0, \sigma_2 > 0, \quad (9)$$

que produce la final

$$\begin{aligned} \pi(\phi, \eta, \sigma_1, \sigma_2 | \mathbf{x}, \mathbf{y}) &\propto \phi^{-1/2} \sigma_1^{-(n+1)} \exp \left\{ -\frac{1}{2\sigma_1^2} \sum_{i=1}^n (x_i - \phi\eta)^2 \right\} \\ &\times \sigma_2^{-(m+1)} \exp \left\{ -\frac{1}{2\sigma_2^2} \sum_{j=1}^m (y_j - \eta)^2 \right\}. \end{aligned} \quad (10)$$

Esta final es propia, en contraste con la correspondiente para el modelo de Cox. Las desviaciones estándar  $\sigma_1$  y  $\sigma_2$  pueden integrarse analíticamente para obtener la marginal final  $\pi(\phi, \eta | \mathbf{x}, \mathbf{y})$  a partir de la cual un procedimiento de integración numérica conduce a la final del parámetro de interés,  $\pi(\phi | \mathbf{x}, \mathbf{y})$ . La final para  $(\mu, \eta, \sigma_1, \sigma_2)$  inducida por (10) está dada por

$$\begin{aligned} \pi(\mu, \eta, \sigma_1, \sigma_2 | \mathbf{x}, \mathbf{y}) &\propto \mu^{-1/2} \sigma_1^{-(n+1)} \exp \left\{ -\frac{1}{2\sigma_1^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \\ &\times \eta^{-1/2} \sigma_2^{-(m+1)} \exp \left\{ -\frac{1}{2\sigma_2^2} \sum_{j=1}^m (y_j - \eta)^2 \right\}, \end{aligned} \quad (11)$$

donde  $\mu > 0$ ,  $\eta > 0$ ,  $\sigma_1 > 0$  y  $\sigma_2 > 0$ . Es claro que  $\mu$  y  $\sigma_1$  son independientes de  $\eta$  y  $\sigma_2$  a posteriori. De hecho, (11) sugiere un procedimiento alternativo para calcular la distribución final del parámetro de interés. Esta idea se explora a continuación.

*Ejemplo 2.* Para ilustrar el análisis del modelo general, se considera el mismo conjunto de datos del Ejemplo 1. Se emplea un enfoque basado en muestreo propuesto por Smith

and Gelfand (1992) y la idea es generar una muestra aleatoria de la distribución marginal final de  $\phi$ . Es posible generar muestras de las distribuciones finales de  $(\mu, \sigma_1)$  y  $(\eta, \sigma_2)$  por separado. A partir de estas muestras se puede obtener una muestra de  $\phi$ .

Puesto que  $\tau_1 = 1/\sigma_1^2$ , la final para  $(\mu, \tau_1)$  puede escribirse como

$$\pi(\mu, \tau_1 | \mathbf{x}) \propto \mu^{-1/2} N_T(\mu | \bar{x}, (n\tau_1)^{-1}) \text{Ga}(\tau_1 | (n-1)/2, ns_x^2/2) \quad (12)$$

donde  $N_T$  denota una densidad normal truncada en cero. Es conveniente considerar la parametrización alternativa  $\theta = \log \mu$  y  $\varphi = \log \tau_1$ . A diferencia de  $\pi(\mu, \tau_1 | \mathbf{x})$ , la densidad final para  $(\theta, \varphi)$  es unimodal. Sea  $\hat{\pi}(\theta, \varphi | \mathbf{x})$  una aproximación normal bivariada para  $\pi(\theta, \varphi | \mathbf{x})$  con media igual a la verdadera moda final y matriz de covarianzas igual a menos 1.5 veces la inversa del Hessiano evaluado en la moda. De esta forma se puede generar una muestra  $\{(\tilde{\theta}^i, \tilde{\varphi}^i)\}_{i=1}^N$  de  $\hat{\pi}(\theta, \varphi | \mathbf{x})$  para entonces remuestrear con pesos proporcionales a

$$w_i = \frac{\pi(\tilde{\theta}^i, \tilde{\varphi}^i | \mathbf{x})}{\hat{\pi}(\tilde{\theta}^i, \tilde{\varphi}^i | \mathbf{x})}.$$

Para simular muestras de  $\pi(\eta, \tau_2 | \mathbf{y})$  se puede utilizar un procedimiento completamente análogo. En este ejemplo se generó una muestra de tamaño 2000 de la distribución final (11) utilizando el método descrito.

Tabla 2: Intervalos del 95% para  $\phi$

Procedimiento	Límites	Longitud
Cox	1.27 6.68	5.41
Elston	1.13 7.85	6.72
Bayesiano I	1.26 4.16	2.90
Bayesiano II	0.63 5.29	4.66

En la Tabla 2 se presenta el intervalo de máxima densidad al 95% (Bayesiano II) junto con dos intervalos de confianza reportados por Cox. El primero se basa en el modelo discutido en la Sección 2.1, mientras que el segundo se deriva de un análisis frecuentista del modelo general debido a Elston (1969). Como es de esperarse, tanto los intervalos frecuentistas como los Bayesianos tienen una longitud mayor si se considera el modelo general. Por otra parte, los intervalos Bayesianos tienen menor longitud que sus contrapartes frecuentistas tanto para el modelo de Cox como para el modelo general. Finalmente, el intervalo Bayesiano para el modelo general tiene una longitud menor que el intervalo frecuentista para el modelo de Cox.

**Agradecimientos.** Este trabajo fue realizado con el apoyo del Sistema Nacional de Investigadores, México.

## Referencias

- Bernardo, J.M. (1977). Inferences about the Ratio of Normal Means: A Bayesian Approach to the Fieller-Creasy Problem. En *Recent Developments in Statistics*. (J.R. Barra et al. eds.) Amsterdam: North-Holland, pp. 345-350
- Berger, J.O., y Bernardo, J.M. (1992), On the Development of Reference Priors (with discussion). En *Bayesian Statistics 4*. (J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds.) Oxford: Clarendon Press, pp. 35-60.
- Cox, C.P. (1985). Interval Estimates for the Ratio of the Means of two Normal Populations with Variances Related to the Means. *Biometrics*, **41**, 261-265.
- Elston, R.C. (1969). An Analogue to Fieller's Theorem Using Scheffe's Solution to the Fisher-Behrens Problem. *The American Statistician*, **23**, 26-28.
- Gilks, W.R., Richardson, S., y Spiegelhalter, D.J. (1996). Introducing Markov Chain Monte Carlo. En *Markov Chain Monte Carlo in Practice*. (W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, eds.) London: Chapman and Hall, pp. 1-19.
- Kappenman, R.F., Geisser, S., y Antle, C.E. (1970). Bayesian and Fiducial Solutions for the Fieller-Creasy Problem. *Sankhya B*, **32**, 331-340.
- Smith, A.F.M. y Gelfand, A.E. (1992). Bayesian Statistics without Tears: a Sampling-resampling Perspective. *American Statistician*, **46**, 84-88.

# Inferencia Bayesiana a Partir de Distribuciones Finales Multimodales

L. E. Nieto-Barajas y E. Gutiérrez-Peña  
*ITAM*                    *IIMAS, UNAM*

## 1 Aproximación Normal Asintótica

Sea  $p(\theta | \mathbf{x})$  la distribución final de  $\theta \in \Re^k$ , donde  $p(\theta | \mathbf{x}) \propto p_x(\theta) = p(\theta)p(\mathbf{x} | \theta)$ ,  $p(\theta)$  es la distribución inicial y  $p(\mathbf{x} | \theta)$  es la función de verosimilitud.

Para poder obtener de una forma rápida características de la distribución final de  $\theta$ , tales como momentos o densidades marginales, es común utilizar la aproximación normal asintótica, la cual se obtiene a partir de una expansión en series de Taylor alrededor de la moda e ignora los términos de orden mayor a dos, obteniéndose el siguiente resultado:

$$p_x(\theta) \approx p_x(\hat{\theta}) \exp \left\{ \frac{- (\theta - \hat{\theta})' \hat{\mathbf{V}}^{-1} (\theta - \hat{\theta})}{2} \right\},$$

donde  $\hat{\theta}$  es la moda de la distribución final, y  $\hat{\mathbf{V}} = - \left\{ \frac{\partial^2 \log p_x(\hat{\theta})}{\partial \theta' \partial \theta} \right\}^{-1}$ .

Por lo tanto,

$$p(\theta | \mathbf{x}) \approx N(\theta | \hat{\theta}, \hat{\mathbf{V}}).$$

La precisión de esta aproximación normal depende de varios aspectos importantes: 1) tamaño de muestra, 2) la parametrización que se tenga y 3) que la distribución final tenga una sola moda dominante. Una demostración formal de esta aproximación se puede encontrar en Bernardo y Smith (1994).

## 2 Caso Multimodal

En algunas aplicaciones Bayesianas, surgen como resultado del análisis distribuciones finales multimodales. Para poder obtener valores esperados o distribuciones marginales, se tiene que recurrir a aproximaciones, ya que en algunos casos ni siquiera la constante de marginalización es fácil de obtener.

## 2.1 Aproximación básica

Una generalización de la aproximación normal se encuentra en O'Hagan (1994), basada en una mezcla de densidades normales. La aproximación consiste en lo siguiente:

Sea  $p(\theta | \mathbf{x}) \propto p_x(\theta)$  una densidad multimodal, y sean  $\hat{\theta}_1, \dots, \hat{\theta}_d$  las  $d$  modas, con  $\mathbf{V}_1(\hat{\theta}_1), \dots, \mathbf{V}_d(\hat{\theta}_d)$  las correspondientes matrices de dispersión, dadas por  $\mathbf{V}_i(\hat{\theta}_i) = \hat{\mathbf{V}}_i = -\left\{ \frac{\partial^2 \log p_x(\hat{\theta}_i)}{\partial \theta' \partial \theta} \right\}^{-1}$ . Entonces, generalizando la aproximación normal unimodal se tiene que

$$p_x(\theta) \approx \sum_{i=1}^d p_x(\hat{\theta}_i) \exp \left\{ \frac{- (\theta - \hat{\theta}_i)' \hat{\mathbf{V}}_i^{-1} (\theta - \hat{\theta}_i)}{2} \right\}.$$

Por lo tanto,

$$p(\theta | \mathbf{x}) \approx \sum_{i=1}^d w_i N(\theta | \hat{\theta}_i, \hat{\mathbf{V}}_i), \quad (1)$$

$$\text{donde } w_i = \frac{p_x(\hat{\theta}_i) |\hat{\mathbf{V}}_i|^{\frac{1}{2}}}{\sum_{j=1}^d p_x(\hat{\theta}_j) |\hat{\mathbf{V}}_j|^{\frac{1}{2}}}.$$

Esta aproximación será mejor si las modas están suficientemente separadas, en cuyo caso las características de  $p(\theta | \mathbf{x})$  se aproximan por las correspondientes características de la aproximación básica.

**Ejemplo 1.** Sean  $X_1, \dots, X_n$  v.a.i.i.d. de la densidad Cauchy( $\theta, 1$ ), es decir,  $p(x | \theta, 1) = \frac{1}{\pi[1+(x-\theta)^2]}$ . Si la distribución inicial para  $\theta$  es  $p(\theta) \propto 1$ , entonces la distribución final de  $\theta$  es  $p(\theta | \mathbf{x}) \propto \prod_{i=1}^n \left\{ \frac{1}{\pi[1+(x_i-\theta)^2]} \right\}$ . Se sabe que cuando  $n = 2$  y se cumple que  $|x_1 - x_2| > 2$  entonces, la función de verosimilitud para  $\theta$  es bimodal. En este caso como la densidad final es proporcional a la verosimilitud se tiene que la densidad final de  $\theta$  es también bimodal. Gráficas de la aproximación básica, para distintos valores de  $x_1$  y  $x_2$  se presentan en la Figura 1.

Si la función de densidad que se quiere aproximar es una mezcla de densidades normales, la aproximación básica no reproduce a la verdadera función de densidad (a diferencia de la aproximación normal unimodal). Una característica importante de esta aproximación es el hecho de que en el caso de que la función de densidad que se quiera aproximar sea unimodal, la aproximación básica se reduce a la aproximación normal asintótica.

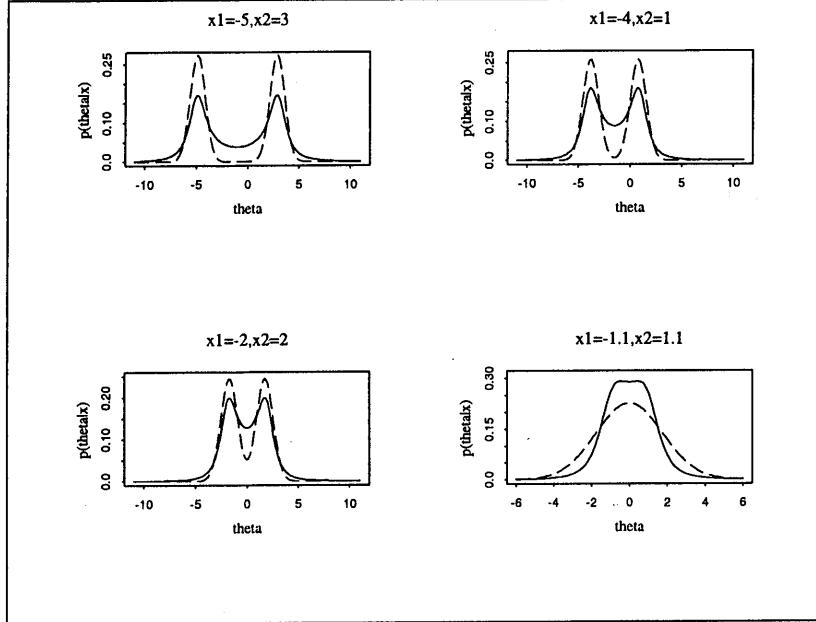


Figura 1: Densidad final de  $\theta$  y Aproximación básica

## 2.2 Nueva aproximació

Una nueva aproximación basada en mezcla de densidades normales es la siguiente. Sea  $p(\theta | \mathbf{x}) = cp_x(\theta)$ , donde  $c$  es la constante de marginalización. Sean  $\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_m$  un conjunto de puntos cualesquiera en la región de mayor densidad de  $p_x(\theta)$ .

La densidad final  $p(\theta | \mathbf{x})$  se puede aproximar mediante

$$\hat{p}(\theta | \mathbf{x}) = \sum_{i=1}^m w_i N_d \left( \theta \mid \tilde{\theta}_i, h \mathbf{I}_d \right) \quad (2)$$

donde,  $h$  es un parámetro de suavizamiento y  $w_i, i = 1, \dots, m$  son la solución al sistema de ecuaciones lineales

$$\begin{pmatrix} p_x(\theta_1) & N_{1,n} - N_{1,1} & \cdots & N_{1,n} - N_{1,n-1} \\ p_x(\theta_2) & N_{2,n} - N_{2,1} & \cdots & N_{2,n} - N_{2,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ p_x(\theta_n) & N_{n,n} - N_{n,1} & \cdots & N_{n,n} - N_{n,n-1} \end{pmatrix} \begin{pmatrix} c \\ w_1 \\ \vdots \\ w_{n-1} \end{pmatrix} = \begin{pmatrix} N_{1,n} \\ N_{2,n} \\ \vdots \\ N_{n,n} \end{pmatrix} \quad (3)$$

con  $N_{i,j} = N \left( \tilde{\theta}_i \mid \tilde{\theta}_j, h \mathbf{I}_d \right)$ .

**Ejemplo 2.** Con referencia al Ejemplo 2.1.1, para un tamaño de muestra  $n = 2$ , y con  $x_1 = -5$  y  $x_2 = 3$ , se cumple que la densidad final para  $\theta$  es bimodal.

Gráficas de la nueva aproximación a esta densidad, para distintos valores de  $n$ , se presentan en la Figura 2.

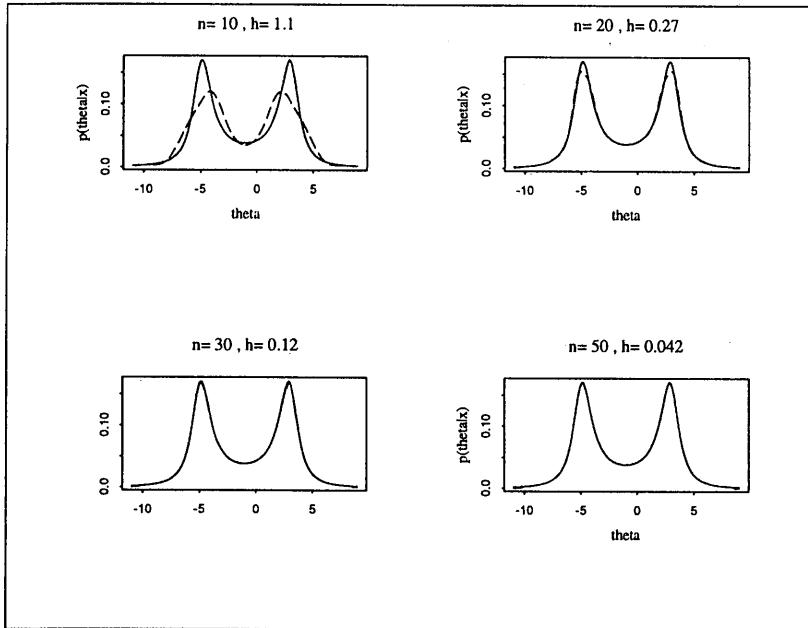


Figura 2: Densidad final de  $\theta$  y Nueva Aproximación

La precisión de la aproximación (2) depende de varios aspectos: 1) la parametrización que se tenga, 2) el número de puntos  $m$ , que a su vez depende del comportamiento de la función que se quiere aproximar, y 3) el parámetro de suavizamiento  $h$ .

Los pesos obtenidos mediante el sistema de ecuaciones (3) siempre suman uno, pero en algunos casos, para ciertos valores de  $h$ , los pesos pueden tomar valores negativos e incluso mayores a uno en valor absoluto. En el caso de que los puntos sean equidistantes, una posible selección inicial de  $h$  para el caso univariado sería  $h \geq (\frac{l}{2})^2$ , donde  $l$  es la distancia entre cada punto.

La ventaja de esta aproximación con respecto a la aproximación básica es que no es necesario encontrar las modas ni evaluar en segundas derivadas.

### 3 Conclusiones

La aproximación propuesta es fácil de implementar porque no se tiene que utilizar un algoritmo para encontrar las modas de la función y, además de ser una aproximación a densidades, se puede ver también como una regla de integración. Por otra parte, la nueva aproximación no distingue si la densidad es unimodal o multimodal. Sin embargo, el número de puntos necesario para tener una buena aproximación aumenta exponencialmente con la dimensión del parámetro.

## Referencias

- Bernardo, J.M. y Smith A.F.M. (1994). *Bayesian Theory*. New York: Wiley.
- O'Hagan A. (1994), *Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference*, Cambridge: University Press.

# Regresión Bayesiana: Análisis y Comparación de Modelos Lineales Generalizados

Gabriel Nuñez Antonio

*UACPyP, UNAM*

## 1 Introducción.

Las técnicas de regresión usuales proponen alguna forma paramétrica para el predictor lineal y proceden a analizar el modelo como si éste fuera el verdadero, sin considerar la discrepancia entre la forma real y el predictor paramétrico asumido. En el presente trabajo se analiza un modelo de regresión generalizado semiparamétrico que toma en cuenta la discrepancia mencionada anteriormente y se propone un enfoque predictivo para la selección de modelos lineales generalizados desde un punto de vista Bayesiano

## 2 Descripción General del Problema.

Sea  $\mathcal{M} = \{M_1, \dots, M_k\}$  un conjunto de modelos paramétricos, donde

$$M_i = \{p_i(\mathbf{y}|\boldsymbol{\theta}_i), p_i(\boldsymbol{\theta}_i)\}; \quad \mathbf{y} \in \mathbf{Y}, \boldsymbol{\theta}_i \in \Theta.$$

Así, el modelo  $M_i$  se define por la verosimilitud  $p_i(\mathbf{y}|\boldsymbol{\theta}_i)$  y la correspondiente distribución inicial  $p_i(\boldsymbol{\theta}_i)$ .

*Problema:* Seleccionar uno de los modelos en  $\mathcal{M}$ , dada la información inicial y una muestra de observaciones de  $\mathbf{Y}$ , con fines predictivos.

En la literatura existen varias formas de plantear el problema de selección y comparación de modelos. En Bernardo y Smith (1994) se discuten los llamados enfoques  $\mathcal{M}$ -cerrado y  $\mathcal{M}$ -abierto.

## 3 Soluciones Tradicionales: Enfoque $\mathcal{M}$ -cerrado.

### 3.1 Factores de Bayes.

En la literatura existen (ver por ejemplo, Bernardo y Smith, 1994) técnicas basadas en los llamados factores de Bayes para tratar los problemas de selección de modelos y pruebas de

hipótesis. Sin embargo, el problema de los factores de Bayes es que éstos dependen fuertemente de la especificación de las distribuciones iniciales de los parámetros. Específicamente, los factores de Bayes no están bien definidos si se utilizan distribuciones iniciales impropias. Berger y Pericchi (1996) proponen un criterio llamado *factor de Bayes intrínseco*; por su parte O'Hagan (1995) utiliza los llamados *factores de Bayes fraccionales*. En ambos casos se pretende resolver algunas de las desventajas de los factores de Bayes.

### 3.2 Criterios Predictivos.

Otras propuestas que existen en la literatura consideran criterios predictivos para solucionar el problema de Comparación de Modelos. San Martini y Spezzaferri (1984) atacan el problema asumiendo el llamado enfoque  $\mathcal{M}$ -cerrado y consideran una función de utilidad score-logarítmica (la cual resulta adecuada en estos casos).

En la práctica asumir el enfoque  $\mathcal{M}$ -cerrado es poco realista. En Bernardo y Smith (1994) se propone una solución aproximada, que involucra lo que se conoce como *validación cruzada*, para aproximar la utilidad esperada considerando el enfoque  $\mathcal{M}$ -abierto.

## 4 Propuesta Semi-paramétrica.

### 4.1 El modelo lirreal rormal.

Supóngase que el conjunto de datos  $\{(y_i, x_i), i = 1, \dots, n\}$  sigue el modelo

$$y_i = \eta(x_i) + \varepsilon_i, \quad (1)$$

donde  $\eta(x)$  es una función desconocida o su forma funcional es extremadamente complicada. En la práctica es frecuente aproximar  $\eta(x)$  a través de una función simple  $p_k(x) = \beta_0 + \beta_1 h_1(x) + \dots + \beta_k h_k(x)$ . Las técnicas de regresión usuales suponen que los  $\{\varepsilon_i\}$  son variables aleatorias independientes, con  $\varepsilon_i \sim N(0, \sigma^2)$  ( $i = 1, \dots, n$ ), y analizan el modelo (1) considerando que  $\eta(x)$  es igual a  $p_k(x)$ . Así, el modelo usual clásico ignora la discrepancia entre  $\eta(\cdot)$  y  $p_k(\cdot)$ . En su lugar, Blight & Ott (1975) proponen un modelo alternativo de la forma

$$y_i = p_k(x_i) + \delta_i + \varepsilon_i, \quad (2)$$

donde  $\delta_i = \delta(x_i) = \eta(x_i) - p_k(x_i)$  es llamado el *error determinístico* del modelo. Los errores  $\{\delta_i\}$  son determinísticos y difieren del componente de error aleatorio  $\varepsilon_i$  en el sentido de que  $\delta_i = \delta_j$  si  $x_i = x_j$ , mientras que dos errores aleatorios para observaciones en el mismo punto de diseño son independientes.

El modelo (2) se puede reescribir como:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \quad (3)$$

con  $\mathbf{y} = (y_1, \dots, y_n)'$ ,  $\beta = (\beta_0, \dots, \beta_k)'$ ,  $\delta = (\delta_1, \dots, \delta_n)'$ ,  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ ,  $\mathbf{X}$  la matriz con renglones  $x_i' = (1, x_{i1}, \dots, x_{ik})$  y  $\Sigma = \sigma^2 W$ . Aquí  $W$  es una matriz diagonal de pesos con elemento  $j$ -ésimo  $m_j^{-1}$ .

## 4.2 Modelos Lineales Generalizados.

Una generalización de los modelos lineales clásicos son los llamados *modelos lineales generalizados*, los cuales tienen como casos especiales a los modelos de regresión lineal, de análisis de varianza, los modelos logit, probit, loglineales, modelos de respuesta multinomial y algunos modelos para datos de supervivencia.

Los modelos lineales generalizados se pueden especificar a través de los siguientes componentes:

**Componente Aleatoria.**  $Y_1, \dots, Y_n$  son variables aleatorias independientes cuya distribución es un miembro de la familia exponencial, con función de densidad de probabilidad dada por

$$f(y_i | \theta_i, \sigma^2) = b(y_i, \sigma^2/m_i) \exp(m_i[y_i\theta_i - a(\theta_i)]/\sigma^2) \quad (4)$$

con  $a(\cdot)$  y  $b(\cdot, \cdot)$  ciertas funciones específicas, donde los  $\{m_i\}$  son pesos conocidos, asociados a cada observación. Si  $\sigma^2$  se conoce, entonces (4) es un modelo que pertenece a la *familia exponencial natural* (ver, por ejemplo, Morris 1988) con *parámetro canónico*  $\theta_i$ . El parámetro  $\sigma^2$  es conocido como el *parámetro de dispersión*.

**Componente sistemática.** Para cada respuesta  $Y_i$ , se tiene asociado un vector de covariables  $x_i = (x_{i1}, \dots, x_{ir})'$ , con el cual se obtiene el predictor lineal  $\eta_i = \eta(x_i) = p_k(x_i) + \delta$ .

**Liga.** Las componentes aleatoria y sistemática se relacionan vía la función liga, de tal manera que  $\eta_i = g(\mu_i)$ . Una liga particularmente importante se obtiene cuando  $g^{-1}(\cdot) = a'(\cdot)$ . En este caso  $\theta_i = \eta_i$  y  $g(\cdot)$  se denomina la *liga canónica*.

### Especificación del Modelo Semiparamétrico

*Nivel 1.* Condicionalmente sobre  $\beta$  y  $\delta$ ,  $Y_1, \dots, Y_n$  son variables aleatorias independientes con  $y_i \sim f(y_i | \theta_i, \sigma^2)$  y  $\theta_i = t(x_i'\beta + \delta_i)$  ( $i = 1, \dots, n$ ) . Lo cual produce la siguiente verosimilitud aproximada

$$l(y | \beta, \delta) \propto \exp \left\{ \frac{1}{\sigma^2} \sum_{i=1}^n m_i [y_i t(x_i'\beta + \delta_i) - a(t(x_i'\beta + \delta_i))] \right\}$$

*Nivel 2.* Condicionalmente sobre  $\rho^2$  y  $\lambda$ , los parámetros  $\beta$  y  $\delta$  son independientes y

$$\begin{aligned} \beta &\sim N(b_0, B_0^{-1}) \\ \delta &\sim N(\mathbf{0}, \rho^2 \Lambda_\lambda) \end{aligned}$$

La densidad final correspondiente a esta especificación inicial es de la forma

$$\begin{aligned} p(\beta, \delta | \mathbf{y}) &\propto \exp \left\{ \frac{-1}{2} (\gamma - t_0)' T_0 (\gamma - t_0) \right\} \\ &\times \exp \left\{ \frac{1}{\sigma^2} \sum_{i=1}^n m_i [y_i t(x_i' \beta + \delta_i) - a(t(x_i' \beta + \delta_i))] \right\} \\ &= \exp \left\{ \frac{-1}{2} (\gamma - t_0)' T_0 (\gamma - t_0) + \frac{1}{\sigma^2} \sum_{i=1}^n m_i [y_i t(D_i \gamma) - a(t(D_i \gamma))] \right\} \end{aligned} \quad (5)$$

donde  $\gamma = (\beta', \delta')'$  es un vector de  $(p+n) \times 1$ , con  $\delta' = (\delta_1, \dots, \delta_n)$  un vector de  $n \times 1$ .  $t_0 = (b_0, \mathbf{0})'$  un vector de  $(p+n) \times 1$ ,  $T_0$  una matriz de  $(p+n) \times (p+n)$  con  $[B_0, (\rho^2 \Lambda_\lambda)]$  como elementos de su diagonal. Aquí,  $D_i = (x_i', 1_i')'$  es un vector de  $(p+n) \times 1$ , con  $1_i$  un vector de  $n$  componentes todos iguales a cero excepto en el lugar  $i$ -ésimo donde tiene un valor de uno.

No en todos los casos se pueden hacer inferencias analíticas exactas basadas en (5). Las inferencias clásicas para modelos lineales generalizados se basan en la estimación de parámetros y las propiedades distribucionales asintóticas de los estimadores. Sin embargo, en general la maximización de la verosimilitud requiere de métodos numéricos. Por otro lado, los métodos de Monte Carlo vía Cadenas de Markov producen una forma relativamente directa para hacer inferencias Bayesianas para una clase amplia de modelos lineales generalizados.

Dellaportas y Smith (1993) utilizan el muestreo de Gibbs en este contexto, para hacer inferencias sobre la densidad final (5); también pueden emplearse distintas versiones del algoritmo de Metropolis-Hastings (ver, por ejemplo, Smith y Roberts, 1993), tales como los algoritmos de *Caminata Aleatoria e Independencia*.

### 4.3 Comparación de Modelos Lineales Generalizados.

Sea  $\mathcal{M} = \{M_1, \dots, M_k\}$  un conjunto de modelos de regresión paramétricos, donde

$$M_i = \{p_i(y|\beta_i), p_i(\beta_i)\} \quad y \in \mathbf{R}, \quad \beta_i \in \mathbf{R}^{q_i}.$$

En este caso

$$p_i(y|\beta_i) = f(y | \theta(\beta_i), \sigma^2)$$

y

$$p_i(\beta_i) = N_{q_i}(\beta_i | b_{0i}, \mathbf{B}_{0i}^{-1}),$$

donde  $f(y | \theta(\beta_i), \sigma^2) = b(y, \sigma^2/m) \exp(m[y\theta - a(\theta)]/\sigma^2)$  denota una familia exponencial y  $\theta(\beta_i) = t(\mathbf{h}_i(x)'\beta_i)$  define tanto al predictor lineal como a la función liga.

*Problema:* Seleccionar uno de los modelos en  $\mathcal{M}$  con fines predictivos.

#### 4.3.1 Solución al problema de Selección de Modelos Lineales Generalizados.

El enfoque semiparamétrico para el problema de selección de modelos, considera éste como un problema de decisión con los siguientes elementos.

**Espacio de decisiones:**

$$\mathbf{D} = \mathcal{M}$$

**Espacio de 'sucisos inciertos':**

$$\mathbf{E} = \{\eta : \eta \text{ es una función suave sobre } \mathbf{R}^r\}$$

**Función de utilidad sobre  $\mathbf{D} \times \mathbf{E}$ :**

$$u(M_i, \eta) = \int \log p_i(y_*|\mathbf{y}) f(y_*|\eta(x_*)) dy_*,$$

donde

$$p_i(y_*|\mathbf{y}) = \int p_i(y_* | \beta_i) p_i(\beta_i | \mathbf{y}) d\beta,$$

es la distribución predictiva final de una observación futura  $y_*$ , considerando el  $i$ -ésimo modelo. Aquí  $f(y_*|\eta(x_*))$  representa un modelo de la familia exponencial.

**Distribución inicial sobre  $\mathbf{E}$ :**

*Nivel I.* Condicional sobre  $\beta$

$$\eta(x) \sim \mathcal{N}(\mu_\beta^*(x), \Sigma^*(x, x)) \quad (\text{Proceso Gaussiano})$$

con

$$\begin{aligned} \mu_\beta^*(x) &= \mathbf{h}(x)' \beta \\ \Sigma^*(x, x) &= \rho^2 \boldsymbol{\Lambda}_\lambda \end{aligned}$$

es decir, dado  $\beta$

$$\eta(x) = \mathbf{h}(x)' \beta + \delta(x), \quad \text{donde } \delta(x) \sim \mathcal{N}(0, \rho^2 \boldsymbol{\Lambda}_\lambda).$$

*Nivel II.*

$$\beta \sim N_q(b_0, ^2 \mathbf{B}_0^{-1}).$$

*Solución:* Seleccione aquel modelo que maximice la utilidad esperada final,

$$\begin{aligned} \bar{u}(M_i) &= E_{\eta|\mathbf{y}} [u(M_i, \eta)] \\ &= E_{\eta|\mathbf{y}} [\int \log p_i(y_*|\mathbf{y}) f(y_* | \eta(x_*)) dy_*] \end{aligned}$$

Así, el modelo  $M_i$  será preferido al modelo  $M_j$  si y sólo si

$$\bar{u}(M_i) > \bar{u}(M_j).$$

**Agradecimientos.** Deseo agradecer profundamente al Dr. Eduardo Gutiérrez-Peña por su directa participación en la dirección y elaboración de este trabajo.

## Referencias

- Berger, J. O. y Pericchi, L. R. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *J. Am. Statist. Assoc.* 91, 109-122.
- Bernardo, J. M. y Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.
- Blight, B. J. N. y Ott, L. (1975). A Bayesian Approach to Model Inadequacy for Polynomial Regression. *Biometrika*, 62, 79-88.
- Dellaportas, P. y Smith, A. F. M. (1993). Bayesian Inference for Generalised Linear and Proportional Hazard Model via Gibbs Sampling. *Appl. Statist.* 42, 443-459.
- Morris, C. N. (1988). Aproximating Posterior Distributions and Posterior Moments (with discussion). *Bayesian Statistics 3* (eds. J. M. Bernardo *et al.*), 327-344. Oxford: University Press.
- O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison (with discussion). *J. of the Roy. Statist. Soc. B*, 57, 99-138.
- San Martini, A. y Spezzaferri, F. (1984). A Predictive Model Selection Criterion. *J. of the Roy. Statist. Soc. B*, 46, 296-303.
- Smith, A. F. M. y Roberts G. O. (1993). Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods. *J. of the Roy. Statist. Soc. B*, 55, 3-23.

# Errores de Redondeo Vistos como Recursión

Federico O'Reilly y Raúl Rueda  
*IIMAS, UNAM*

## 1 Introducción

Sea  $F$  la función de distribución de la variable aleatoria  $X$  y sea  $F_{D_n}$  la función de distribución correspondiente a la variable aleatoria discretizada en la “etapa  $n$ ”, que resulta de redondear los valores de  $X$  en los intervalos de discretización de longitud común  $a_n$ . Esto es, dada  $y$  la marca de clase ( centro ) del intervalo  $(y - a_n/2, y + a_n/2)$ ,  $F_{D_n}$  se define como

$$F_{D_n}(x) = \begin{cases} F(y - a_n/2) & \text{si } x < y \\ F(y + a_n/2) & \text{si } x \geq y \end{cases}$$

Considere que se refina el redondeo, manteniendo las marcas de clase de los intervalos previos (y agregando las marcas adicionales necesarias), pasando de la etapa  $n_1 = n$  a  $n_2 = 4n_1$  y haciendo que este refinamiento cambie las longitudes de los intervalos de redondeo de  $a_{n_1} = a_n$  a  $a_{n_2} = (a_{n_1})/2$ , de modo que  $\sqrt{n_2} \cdot a_{n_2} = \sqrt{n_1} \cdot a_{n_1}$  y llame  $\Delta$  a ese valor común.

La discrepancia  $\delta_n(x) = \sqrt{n}\{F_{D_n}(x) - F(x)\}$  mide el error en el uso de la función de distribución equivocada ( $F$ ) en lugar de la correcta ( $F_{D_n}$ ) habiendo redondeado. Dicha diferencia está multiplicada por  $\sqrt{n}$  pues así aparece en una aplicación de bondad de ajuste en O'Reilly y Rueda (1996) ya que bajo refinamientos en los redondeos del tipo mencionado,  $F_{D_n}$  converge puntualmente a  $F$  y esa diferencia, si no estuviera escalada convergería a cero. El comportamiento límite de  $\delta_n(x)$  es el problema de interés.

La idea básica es el poder relacionar los valores de las discrepancias  $\delta_{n_2}(x)$  con  $\delta_{n_1}(x)$ , para así definir un  $n_3 = 4 \cdot n_2$ , un  $a_{n_3} = (a_{n_2})/2$  etc., y ver si existe un patrón en la sucesión de discrepancias  $\delta_{n_j}$ , para  $j = 1, 2, \dots$ , que permitan afirmar algo sobre su comportamiento asintótico.

## 2 La Recursión

Sea  $y$  un centro (o marca de clase) de un intervalo de redondeo de longitud  $a_n$  y sea  $x$  un punto dentro de ese intervalo. La función  $\delta_n(x)$  toma el valor  $(\sqrt{n})\{F(y + a_n/2) - F(x)\}$

si  $x > y$  o bien,  $(\sqrt{n})\{F(y - a_n/2) - F(x)\}$  si  $x < y$ . En el primer caso la expresión resulta positiva y en el segundo, negativa.

Para  $n_2$ , el refinamiento descrito lleva a una de las siguientes tres posibilidades: que ese mismo punto  $x$  pertenezca al nuevo intervalo centrado en  $y$ , o bien a la parte inferior del nuevo intervalo centrado en  $y + a_n/2$  o a la parte superior del nuevo intervalo centrado en  $y - a_n/2$ .

Al evaluar  $\delta_{n_2}(x)$  resulta

$$\delta_{n_2}(x) = \begin{cases} \sqrt{n_2}\{F(y + a_n/4) - F(x)\} & \text{si } x \text{ está en } (y, y + a_n/2) \\ \sqrt{n_2}\{F(y - a_n/4) - F(x)\} & \text{si } x \text{ está en } (y - a_n/2, y) \end{cases}$$

Se hace la observación de que la cantidad  $a_n$  será en general chica (al aplicar la recursión, la sucesión  $a_{n_j}$  tiende a cero) y la expresión para  $\delta_n(x)$  puede aproximarse por el teorema del valor medio, con  $\Delta \cdot f(x)\{(y - x)/a_n + 1/2\}$  si  $x > y$  o bien con,  $\Delta \cdot f(x)\{(y - x)/a_n - 1/2\}$  si  $x < y$ , habiendo usado en la aproximación, el que la densidad  $f$  evaluada entre los puntos  $x$  y  $y$ , se mantiene esencialmente constante. El error cometido en esta aproximación es  $O(a_n)$ .

Obsérvese que en la anterior expresión para  $\delta_n$ , las cantidades que pueden ser considerables (de orden  $O(1)$ ) son  $(y - x)/a_n \pm 1/2$ . Denótese a estas cantidades por  $u$ .

Haciendo un análisis parecido para expresar aproximadamente a  $\delta_{n_2}$ , se puede observar con facilidad que  $\delta_{n_2}(x) \cong (\Delta) \cdot f(x) \cdot G(u)$ , con  $G$  un mapeo del intervalo  $[-1/2, 1/2]$  en si mismo, dado por

$$G(u) = \begin{cases} 2u + 1/2 & \text{si } u \in [-\frac{1}{2}, 0] \\ 2u - 1/2 & \text{si } u \in [0, 1/2] \end{cases}$$

en otras palabras, se tiene  $\delta_{n_1} \approx \Delta \cdot f(x) \cdot u$ ,  $\delta_{n_2} \approx \Delta \cdot f(x) \cdot G(u)$  y así sucesivamente.

Un mapeo como  $G$ , cuando se aplica sucesivamente produce iterandos que se comportan de manera caótica. Esto es, dados dos valores  $u$  y  $u'$  distintos, al observar las dos sucesiones formadas al aplicar repetidamente  $G$  tanto a  $u$  como a  $u'$ , se nota que después de un número inicial de aplicaciones, los iterandos se comportan de modo muy distinto (sin importar qué tan parecidos hubieren sido  $u$  y  $u'$ ). Un mapeo como  $G$  se encuentra muy relacionado con el “tent map” discutido en Isham (1991).

El siguiente resultado muestra que bajo casi cualquier suposición distribucional que se tuviera sobre los valores de  $\delta_n(x)$ , al ir refinando los redondeos, los subsiguientes valores de  $\delta_{n_j}(x)$  tenderán a tener una distribución uniforme en  $[-1/2, 1/2]$ .

### Proposición

Sea  $U$  variable aleatoria con distribución continua en  $[-1/2, 1/2]$  con densidad  $h$  y sea  $W_j$  la variable aleatoria resultante de aplicar  $G$ ,  $j$  veces a  $U$ , o sea  $W_j = G^j(U)$ , entonces la densidad de  $W_j$  tiende a una uniforme al tender  $j$  a  $\infty$ .

### Demostración.

Dada  $v \in [-\frac{1}{2}, \frac{1}{2}]$ , la imagen inversa  $G^{-1}(v)$  corresponde a dos valores: el primero ( $< 0$ ) dado por  $u = \frac{v}{2} - \frac{1}{4}$  y el segundo por  $u' = u + \frac{1}{2}$ .

Aplicando recursivamente  $G^{-1}$ ,  $G^{-j}(v)$  está formado por los puntos  $\frac{v}{2^j} + \frac{l}{2^{j+1}}$  para  $l = \pm 1, 3, 5, 7, \dots, 2^j - 1$

Estos  $2^j$  puntos (que equidistan  $\frac{1}{2^j}$  entre sí) están dentro de los intervalos generados por los puntos

$$-\frac{2^{j-1}}{2^j}, -\frac{2^{j-1} + 1}{2^j}, \dots, -\frac{2}{2^j}, -\frac{1}{2^j}, 0, \frac{1}{2^j}, \frac{2}{2^j}, \dots, \frac{2^{j-1}}{2^j},$$

y dichos intervalos forman una partición de  $[-\frac{1}{2}, \frac{1}{2}]$ .

Si  $h_j$  denota la densidad de  $W_j$ , evaluada en  $\omega \in [-\frac{1}{2}, \frac{1}{2}]$   $h_j(\omega) = \frac{1}{2^j} \sum_l h(\frac{\omega}{2^j} + \frac{l}{2^{j+1}})$  en donde la suma es sobre las  $2^j$  valores de  $l$  descritos anteriormente. Entonces la densidad  $h_j(\omega)$  es simplemente la aproximación de Riemann a  $\int_{-1/2}^{1/2} f(u) du = 1$ .

## 3 Curiosidad Numérica

En esta sección se quiere ilustrar el comportamiento numérico de la recursión  $G$ , haciendo algunas observaciones.

Para empezar, los valores  $0, \pm 1/2, 1/4, 3/4, 1/8, 3/8, 5/8, 7/8, \dots$  y en general fracciones con denominador que sea potencia de dos, producen iterandos que eventualmente se vuelven idénticos a  $1/2$ .

Por otro lado, para elaborar un programa que produzca muchos iterandos dado un valor inicial  $u$ , se debe construir un algoritmo con una aritmética que no produzca errores acumulables.

Lo más deseable es construir un algoritmo para evaluar  $G$  que utilice precisión exacta. Para ello se debe limitar el dominio de  $G$  a valores  $u$  descritos en aritmética entera como  $u = m/(2 \cdot N)$  ya que  $G(u) = m'/(2 \cdot N)$ , con el nuevo numerador dado por la sencilla recursión  $m' = 2 \cdot m + N$  si  $m$  es negativo y  $m' = 2 \cdot m - N$  si  $m$  es positivo. En principio el valor de  $m$ , entero, debe pertenecer al conjunto

$$J = \{-N + 1, -N + 2, \dots, -1, 1, 2, \dots, N - 1\}$$

que no contiene al entero  $N$  ya que daría  $u = 0$  y es atraído a  $G(u) = 1/2$ .

¿Para qué valor de  $N$  puede uno iterar con  $G$  de modo que la recursión tenga un ciclo máximo?

Para empezar, si  $N$  se elige par,  $N/2$  es un posible valor del recorrido que daría  $u = 1/2$  y los iterandos se vuelven constantes, por lo que  $N$  debe ser impar. Por otro lado, con  $N$  impar y dada la recursión,  $m'$  es impar siempre por lo que no se puede esperar hacer un

ciclo recorriendo a todos los elementos del conjunto  $J$ ; a lo más que se puede esperar es a tener un  $N$  con el que se recorra a todos los impares de  $J$ . Esta sucesión de impares entre  $-N + 1$  y  $N - 1$ , puede utilizarse por su naturaleza caótica (sumándoles  $N$  y dividiendo por  $2N$  cada uno de los términos de la sucesión), para producir números seudo-uniformes en el intervalo unitario. La dependencia (funcional) de un número y su consecutivo puede eliminarse casi completamente si del ciclo de números generados se van tomando éstos a cierta distancia (digamos saltando nueve), y de todos modos el algoritmo recorrería el ciclo.

A continuación una serie de intentos para  $N$  en los que para algunos valores se logra un ciclo máximo. El valor al final es el más grande que se consiguió para correr la recursión en una PC con Fortran.

Valores de $N$ (primo)	Ciclo Máximo?	Valores de $N$ (primo)	Ciclo Máximo?
3	sí	99907	sí
5	sí	99923	sí
7	no	99929	no
11	sí	99961	no
13	sí	99971	sí
17	no	99989	sí
19	sí	99991	no
.	.	1,073,741,789	sí

## 4 Comentario

En Widrow y Kollár (1996) se discuten aspectos estadísticos de los errores debidos a redondeo, en un contexto de cuantización de señales en el que se analizan similitudes entre sus momentos con las de los que se suponen distribuidos uniformemente. En ese trabajo y otros previos ahí citados, los errores se modelan como uniformes y esto queda justificado si la densidad satisface ciertas condiciones. En contraste, el resultado en este trabajo sobre la uniformidad asintótica de los errores de redondeo es independientemente de la distribución que estos pudieran haber tenido para el valor inicial de  $n$ , y proporciona una justificación teórica para utilizar como modelo para su comportamiento, la distribución uniforme, siempre que pueda suponerse un refinamiento como el descrito.

En la aplicación hecha por O'Reilly y Rueda (1996), en el contexto de bondad de ajuste bajo redondeos, se verificó que la teoría asintótica de la estadística propuesta reconstruye con asombrosa fidelidad los resultados de las simulaciones hechas. En ese trabajo, la uniformidad de  $\delta_n(x)$  y la independencia entre  $\delta_n(x)$  y  $\delta_n(x')$ , proporcionó la herramienta básica para encontrar dicha teoría.

## Referencias

- Isham, V. (1991) Statistical aspects of chaos: a review. *Proc. Séminaire Européen de Statistique*. London: Chapman and Hall
- O'Reilly, F. y Rueda, R. (1996) Goodness of fit under rounding of data. *Preimpreso 47*. México: IIMAS, UNAM.
- Widrow, B. y Kollár I. (1996) Statistical theory of quantization. *IEEE Transactions on instrumentation and measurement*, 45, 353-361.

# Análisis y Corrección de Valores Críticos de un Método para Evaluar Contrastes en Factoriales no Replicados

Jorge Olguín Uribe y Patricia Romero Mares  
*IIMAS, UNAM*

## 1 Introducción

En investigaciones industriales es común listar un número grande de variables o factores que se piensa que podrían tener algún efecto sobre determinada característica de calidad de un producto o proceso. La experiencia ha mostrado que con mucha frecuencia se cumple el *Principio de Pareto*, que, en este contexto, establece que la mayor parte de la variabilidad de la respuesta se debe a un número reducido de factores. Cuando el principio de Pareto se cumple, se dice que hay *esparcidad de efectos*. En estos casos, los experimentos factoriales no replicados resultan de gran utilidad, pues permiten estudiar simultáneamente un número grande de factores y, en caso de que los haya, identificar aquéllos cuyos efectos son relevantes. Este procedimiento se denomina *criba de efectos*.

Estos experimentos comúnmente se basan en arreglos ortogonales con  $n$  condiciones experimentales de donde, una vez efectuado el experimento, se obtienen  $m=n-1$  contrastes estimados y se quiere decidir cuáles de ellos tienen un tamaño estadísticamente significativo. Los contrastes que resultan significativos son llamados *contrastes activos*, de otro modo se consideran como *contrastes inertes o nulos*.

Uno de los problemas para evaluar los contrastes estimados es la ausencia de un estimador del error experimental, que en otros experimentos se obtiene con base en réplicas. Lenth (1989) propuso un estimador de la desviación estándar del error y a partir de ahí desarrolló el método que se describe a continuación.

## 2 El Método de Lenth

Sean  $k_j, j = 1, \dots, m$  los contrastes de interés y  $c_j, j = 1, \dots, m$  sus estimaciones. Bajo las suposiciones básicas del modelo de análisis de varianza, las  $c_j, j = 1, \dots, m$  son observaciones independientes de distribuciones  $N(k_j, \sigma^2), j = 1, \dots, m$ . Es decir, se supone que las distribuciones muestrales de las  $c_j$  son normales con medias posiblemente diferentes pero

todas con la misma varianza. En algunos casos será necesario estandarizar los contrastes para que se cumpla esta última condición.

Sea

$$S_0 = 1.5 \times \text{mediana}_j |c_j|.$$

Se define el pseudo error estándar de los contrastes como

$$PSE = 1.5 \times \text{mediana}_{\substack{|c_j| \leq 2.5S_0|c_j|}}.$$

Bajo *esparcidad de efectos*  $PSE$  es un estimador razonablemente bueno de la desviación estándar de los contrastes (ver Lenth, 1989).

Este resultado se utiliza para obtener un *margen de error* ( $ME$ )

$$ME = PSE \times t_{.975,d},$$

donde  $t_{.975,d}$  es el percentil 0.975 de la distribución  $t$  de Student con  $d = m/3$  grados de libertad. Los grados de libertad se aproximan ajustando las distribuciones empíricas de  $PSE^2$  con distribuciones  $\chi^2$ . Tomando en cuenta que se realizan  $m$  inferencias simultáneamente, se define también un *margen de error simultáneo* ( $SME$ )

$$SME = PSE \times t_{\gamma,d} \text{ donde } \gamma = (1 + 0.95^{1/m})/2.$$

Para el cálculo de  $ME$  y  $SME$ , Lenth presentó una tabla con los percentiles correspondientes de  $t$  (valores críticos) para experimentos con  $m = 7, 15, 31, 63, 127$  y  $255$  contrastes. Para el cálculo utilizó un algoritmo que permite obtener percentiles de la distribución  $t$  con grados de libertad fraccionados.

Se sugiere entonces construir una gráfica de barras mostrando los contrastes y añadir líneas de significancia en  $\pm ME$  y  $\pm SME$ . La primera debería permitir probar la hipótesis individual  $H_0 : k_i = 0$ , para el  $i$ -ésimo contraste aisladamente, es decir, sin realizar pruebas sobre los  $m-1$  restantes, utilizando un nivel de significancia de 0.05 y la segunda permitiría la prueba simultánea de las hipótesis  $H_j : k_j = 0$   $j = 1, \dots, m$ . Sin embargo, por las razones que se dan en la siguiente sección, la tabla proporcionada por Lenth está lejos de cumplir con su propósito.

### 3 Revisión del Método

Las bases sobre las que Lenth sustenta la obtención de sus valores críticos tienen deficiencias. La más aparente es la expresión para  $\gamma$  utilizada en los percentiles con  $t_{\gamma,d}$ : con  $\gamma = (1 + 0.95)^{1/m}/2$ , se pretende utilizar un nivel de significancia de 0.05 para la prueba simultánea de las  $m$  hipótesis nulas  $H_j : k_j = 0$   $j = 1, \dots, m$ . Es decir, 0.05 debe ser la probabilidad de rechazar al menos una de las  $m$  hipótesis nulas, bajo la suposición de que todas ellas son ciertas. A esta probabilidad le llamamos PER.

Es fácil ver que la expresión para  $\gamma$  usada por Lenth se obtiene de

$$\gamma = (1 - \theta/2)$$

donde  $\theta$  es la solución de la ecuación  $0.05 = 1 - (1 - \theta)^m$ .

Esto implica que fue obtenido considerando las  $m$  estadísticas de prueba como si fueran independientes. Sin embargo, la suposición de independencia es insostenible por la sencilla razón de que las  $m$  estadísticas de prueba comparten el mismo valor de PSE utilizado como estimador de la desviación estándar de los contrastes.

Otra deficiencia está en el uso de la distribución t para aproximar los valores críticos, pues éstos resultan muy conservadores sobre todo para experimentos con menor número de contrastes. En un estudio comparativo de varios métodos para el análisis de factoriales no replicados, Olguín (1994) mostró que el nivel de significancia real que producen los valores críticos proporcionados por Lenth para las pruebas de hipótesis sobre contrastes individuales para  $m=7$  y  $m=15$  son de 0.020 y 0.029 respectivamente; esto en lugar del pretendido 0.05. A este nivel de significancia (probabilidad de error tipo I) para hipótesis individuales le llamamos EPE.

## 4 Valores Corregidos del Método de Lenth

Olguín y Fearn (1997) mostraron que los valores críticos presentados por Lenth son imprecisos y obtuvieron correcciones para los casos de  $m=7, 15$  y  $31$  contrastes. Para esta presentación hemos completado el cálculo de valores críticos para los tamaños de factoriales no replicados que con mayor frecuencia se utilizan en la práctica. Estos se presentan en la Tabla 1. Por completez, la Tabla incluye los valores obtenidos previamente y con fines comparativos se presentan también los de Lenth. Nuestro propósito principal es hacer públicos estos valores para que esta útil herramienta propuesta por Lenth pueda ser utilizada por estadísticos y experimentadores en general. Los valores críticos fueron obtenidos por simulación de Monte Carlo de la manera siguiente: a partir de un percentil inicial de la distribución empírica de las estadísticas de prueba, se calculó el valor de PER o EPE, según el caso, y entonces, se procedió de manera iterativa hasta obtener el nivel de significancia de 0.05. Los valores que se presentan en la Tabla 1 se obtuvieron con simulaciones de 100,000 muestras cada uno.

## 5 Ejemplo

Para ilustrar el uso del procedimiento de Lenth, tomemos un experimento reportado por Box *et al.* (1978) y por Box y Meyer (1993). Se utilizó una fracción factorial  $2^{8-4}_{IV}$  para investigar los posibles efectos de ocho factores (A,...,H) en el encogimiento de piezas manufacturadas en un proceso de moldeo por inyección. En la Tabla 2 se presentan los

15 contrastes estimados así como su patrón de confusión considerando solamente hasta interacciones de dos factores.

**TABLA 1.**

Percentiles de la distribución  $t$  con  $d = m/3$  g.l. (dados por Lenth) y cuantiles de la distribución empírica obtenidos de 100,000 muestras simuladas.

(m) No. de contrastes	LENTH $t_{.975,d}$ $\gamma = (1 + 0.95)^{1/m}/2$	$t_{\gamma,d}$	VALOR	EMPIRICO
			.05	EPE .05
7	3.76	9.01	2.30	4.86
8	-	-	2.20	4.84
11	-	-	2.21	4.44
15	2.57	5.22	2.15	4.23
17	-	-	2.13	4.19
19	-	-	2.12	4.10
23	-	-	2.09	4.01
26	-	-	2.08	3.97
27	-	-	2.07	3.96
31	2.22	4.22	2.06	3.94

**TABLA 2.**

Contrastes estimados y patrón de confusión (hasta interacciones de dos factores) del ejemplo de moldeo por inyección.

Contrastes estimados	Nombres de los contrastes
-0.7	A
-0.1	B
5.5	C
-0.3	D
-3.8	E
-0.1	F
0.6	G
1.2	H
-0.6	AB+CG+DH+EF
0.9	AC+BG+DF+EH
-0.4	AD+BH+CF+EG
4.6	AE+BF+CH+DG
-0.3	AF+BE+CD+GH
-0.2	AG+BC+FH+DE
-0.6	AH+BD+CE+FG

Para usar el procedimiento de Lenth hemos elaborado un programa en lenguaje de

programación GAUSS que, a partir de los contrastes estimados y los valores críticos correspondientes al número  $m$  de contrastes, obtiene una gráfica del tipo propuesto por Lenth.

Con fines comparativos, para el ejemplo del moldeo por inyección obtuvimos una gráfica utilizando los valores críticos sin corregir (Figura 1) y otra utilizando los valores críticos corregidos (Figura 2).

La interpretación de estas gráficas es como sigue: aquellos contrastes cuya barra sobrepasa las líneas externas ( $\pm SME$ ) son claramente activos; aquéllos cuya magnitud no sobrepasa las líneas interiores ( $\pm ME$ ) son considerados como inertes; por último, los contrastes cuya magnitud está entre las líneas internas y externas, se encuentran en una situación de ambigüedad y requerirán de información adicional para ser considerados ya sea como activos o inertes.

Cuando se utilizaron los valores críticos sin corregir (Figura 1) se observa que hay dos contrastes cuyas barras sobrepasan las líneas exteriores, uno más está en la región de ambigüedad, mientras que los doce restantes son considerados como inertes. Cuando se utilizan los valores críticos corregidos (Figura 2) las barras correspondientes a los tres contrastes de mayor magnitud rebasan claramente las líneas externas y por lo tanto son considerados como contrastes activos. Esta última interpretación está más de acuerdo con el análisis original de los autores y coincide con la aplicación de otros tres métodos empleados.

Fig. 1 Procedimiento de Lenth (sin corregir)

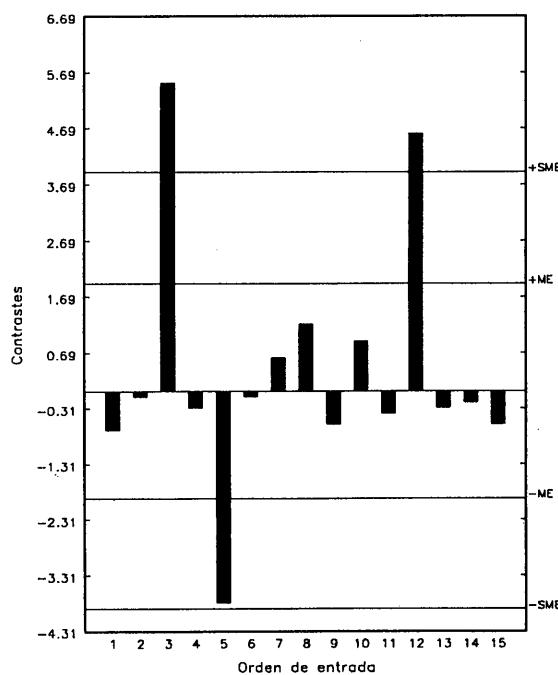
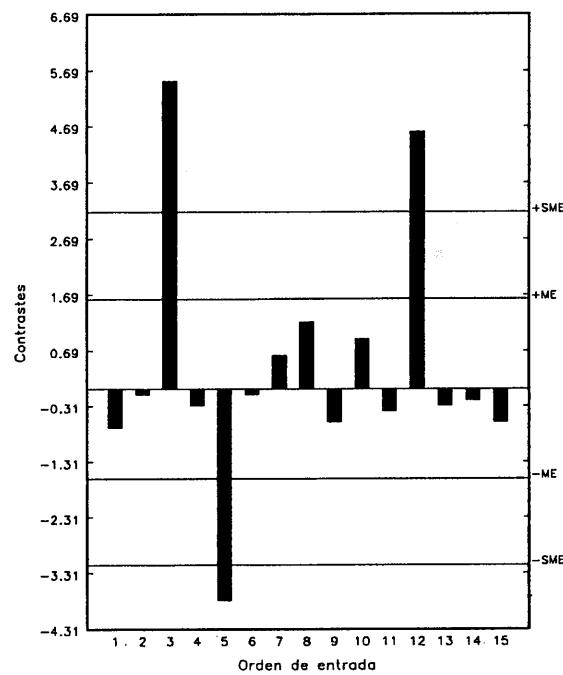


Fig. 2 Procedimiento de Lenth corregido



## 6 Comentario Final

El procedimiento presentado por Lenth (1989) es muy atractivo pues permite analizar gráficamente la magnitud de los contrastes y decidir con facilidad acerca de su significancia. Desafortunadamente las aproximaciones usadas por Lenth son imprecisas, sobre todo en experimentos chicos, haciendo que el método resulte conservador. Esperamos que con la tabla de valores críticos corregidos, presentada en este trabajo, aumente la aceptación del método entre los experimentadores. Consideramos importante señalar, sin embargo, que el análisis de experimentos factoriales no replicados, requiere de un conocimiento del fenómeno bajo investigación, así como el análisis cuidadoso de los datos, contrastes, patrones de confusión y residuos. La utilización de métodos como el analizado aquí, solamente debe de ser una ayuda al momento de decidir qué contrastes deberían ser considerados como activos.

## Referencias

- Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978). *Statistics for Experimenters*. New York: Wiley.
- Box, G.E.P. and Meyer, R.D. (1993). Finding the Active Factors in Fractionated Screening Experiments. *Journal of Quality Technology*, 25, 94-105.
- Lenth, R.P. (1989). Quick and Easy Analysis of Unreplicated Factorials. *Technometrics*, 31, 469-473.
- Olgún, J. (1994). The Analysis of Unreplicated Factorial Experiments. Ph. D. Thesis, University of London, University College London.
- Olgún, J. and Fearn, T. (1997). A New Look at Half-normal Plots for Assessing the Significance of Contrasts for Unreplicated Factorials. *Applied Statistics*, 46, 449-462.

# Estimación de Componentes de Varianza en un Modelo con Partición de Efectos Aplicado a Ensayos de Híbridos de Maíz

Emilio Padrón Corral

*Dept. de Estadística y  
Cálculo UAAAAN Saltillo, Coah.*

Angel Martinez Garza

*ISEI Programa de Estadística,  
Colegio de Postgraduados,  
Montecillo, Edo. de México*

Ma. Cristina Vega S.

*Inst. Mexicano del Maíz  
UAAAAN Saltillo, Coah.*

## 1 Introducción

Cuando se analizan varios tipos de modelos estadísticos, existen diferentes efectos los cuales se descomponen en fijos, aleatorios y mixtos; aquí se hará referencia a los aleatorios y de ellos nos interesa lo concerniente a sus varianzas. Además un programa de mejoramiento genético debe estar apoyado en los métodos estadísticos para avanzar con mayor firmeza en la selección de plantas (genotipos) que coadyuven a incrementar la producción por unidad de superficie, al seleccionar la técnica adecuada que conlleva a una mejora en la toma de decisiones. Por lo tanto en este trabajo se estudia el comportamiento de genotipos de maíz en diferentes ambientes, y se logra el verdadero efecto de sus varianzas contando para ello con la herramienta fundamental de las componentes de varianza estimadas, en base a la técnica del análisis de varianza (ANOVA) como se aprecia en Searle (1987), también autores como Brownlee (1984), nos comenta cómo Bennet y Franklin elaboraron un procedimiento para obtener los valores de esperanzas de cuadrados medios en situaciones parcialmente jerárquicas. Searle *et al.* (1992), comentan que las sumas de cuadrados del análisis de varianza para datos desbalanceados siguen siendo los mismos que para datos balanceados excepto que en lugar de tener  $n$  se debe tener  $n_i$ , y en lugar de  $N$  se debe tener  $\sum n_i$ , y el que los datos sean desbalanceados no elimina la posibilidad de obtención de estimadas negativas de  $\sigma^2_t$  (componentes de varianza para tratamientos) en el análisis de varianza. La teoría desarrollada en este trabajo se aplicó a un experimento de campo

titulado *Estimación de la Habilidad Combinatoria General y Específica de Líneas S<sub>1</sub> de Maíz (Zea may L.) para Condiciones de Humedad Restrictiva*. Esta investigación forma parte del programa de mejoramiento genético del Instituto Mexicano del Maíz (Mario E. Castro Gil) de la U.A.A.A.N. y consta de dos localidades, Narigua Municipio de General Cepeda y Parras de la Fuente, Coahuila, respectivamente, además de 145 híbridos tanto experimentales como comerciales, utilizados estos últimos como testigos.

## 2 Descripción del problema

Dado un modelo estadístico con varios factores agronómicos e interacción se descompondrá como sigue: Genotipos se partirá en cruzas y éstas en líneas dentro de probador uno y líneas dentro de probador dos; cada línea se parte en grupos uno, dos, tres y cuatro dentro de probador uno y lo mismo dentro de probador dos; se hará el contraste de líneas dentro del probador uno contra líneas dentro del probador dos; también se logrará obtener la información de testigos, el contraste de cruzas contra testigo y todos estos efectos se interactuarán con localidades; de cada uno de ellos se desarrollarán las sumas de cuadrados y se obtendrán las esperanzas de cuadrados medios para conocer sus correspondientes componentes de varianza estimadas y por lo tanto, saber en cuanto están contribuyendo de acuerdo a los resultados del experimento de campo.

## 3 Metodología

Se aplicará la técnica de la esperanza de cuadrados medios en el modelo que a continuación se presenta:

$$Y_{ijk} = \mu + L_k + R_{j/k} + G_i + (LG)_{ki} + E_{ijk}$$

donde

$$\begin{aligned} i &= 1, 2, 3, \dots, t && \text{genotipos} \\ j &= 1, 2, 3, \dots, r && \text{repeticiones} \\ k &= 1, 2, 3, \dots, l && \text{localidades.} \end{aligned}$$

- $Y_{ijk}$  : Variable aleatoria observable de la  $k$ -ésima localidad en la  $j$ -ésima repetición del  $i$ -ésimo genotipo
- $\mu$  : Media general
- $L_k$  : Efecto de la  $k$ -ésima localidad
- $R_{j/k}$  : Efecto de la  $j$ -ésima repetición dentro de la  $k$ -ésima localidad

- $G_i$  : Efecto del  $i$ -ésimo genotipo  
 $(LG)_{ki}$  : Efecto conjunto de la  $k$ -ésima localidad y del  $i$ -ésimo genotipo  
 $E_{ijk}$  : Componente aleatoria asociada con la  $k$ -ésima localidad en la  $j$ -ésima repetición del  $i$ -ésimo genotipo

De cada uno de los términos del modelo se obtendrán las esperanzas de cuadrados medios, así como de cada uno de los efectos de la partición de los tratamientos, con el objeto de obtener las estimaciones de componentes de varianza de los datos del trabajo de campo ya mencionado.

Por lo tanto de acuerdo a dicho modelo, el cuadrado medio de la interacción Localidad–Genotipo es el cuadrado medio del error apropiado para probar genotipos. Además se asume que las esperanzas de efectos son cero, es decir,

$$E[L_k] = E[R_{j/k}] = E[G_i] = E[(LG)_{ki}] = E[E_{ijk}] = 0$$

También se supone que las esperanzas de productos cruzados de los diferentes efectos son cero, se tiene además que

$$E[L_k^2] = \sigma_L^2 \quad E[R/L]^2 = \sigma_{r/L}^2 \quad E[G_i^2] = \sigma_G^2 \quad E[(LG)_{ki}^2] = \sigma_{LG}^2 \quad E[E_{ijk}^2] = \sigma_e^2$$

En seguida se listan una serie de términos, semejantes a los que presenta Rodríguez (1992) los cuales serán de gran ayuda para los desarrollos del tema.

$$\begin{aligned} E[\sum_{ijk} Y_{ijk}^2] &= rtl\mu^2 + rtl\sigma_L^2 + rtl\sigma_{r/L}^2 + rtl\sigma_G^2 + rtl\sigma_{LG}^2 + rtl\sigma_e^2 \\ \frac{E[Y^2]}{rtl} &= rtl\mu^2 + rt\sigma_L^2 + t\sigma_{r/L}^2 + rl\sigma_G^2 + r\sigma_{LG}^2 + \sigma_e^2 \\ \frac{E[\sum_k Y_{jk}^2]}{rt} &= rtl\mu^2 + rtl\sigma_L^2 + tl\sigma_{r/L}^2 + rl\sigma_G^2 + rl\sigma_{LG}^2 + l\sigma_e^2 \end{aligned}$$

$$\frac{E[\sum_{jk} Y_{jk}^2]}{t} = rtl\mu^2 + rtl\sigma_L^2 + rtl\sigma_{r/L}^2 + rl\sigma_G^2 + rl\sigma_{LG}^2 + rl\sigma_e^2$$

$$\frac{E[\sum_i Y_{i..}^2]}{rl} = rtl\mu^2 + rt\sigma_L^2 + t\sigma_{r/L}^2 + rtl\sigma_G^2 + r\sigma_{LG}^2 + t\sigma_e^2$$

$$\frac{E[\sum_{ik} Y_{i..k}^2]}{r} = rtl\mu^2 + rtl\sigma_L^2 + lt\sigma_{r/L}^2 + rtl\sigma_G^2 + rl\sigma_{GL}^2 + lt\sigma_e^2$$

Antes de continuar se definirán algunos conceptos agronómicos referentes a la partición de los tratamientos.

- Genotipos = individuos, plantas, animales etc.(En este caso se usaron plantas).
- Cruzas = Serie de nuevas plantas sometidas a ensayo con una característica deseable conocida.
- Probadores = Plantas con características deseables ya conocidas por pruebas estadísticas y agronómicas preliminares.
- $Lin/p_1$  = Número de plantas que se están probando en cruce con la característica germoplásmica uno.
- $(Lin/p_1 \text{ vs } Lin/p_2)$  = Contraste o grado de potencialidad entre las plantas de línea dentro de probador uno con las plantas de línea dentro de probador dos.
- $g_1/p_1$  = Número de plantas de la población uno que provienen de la característica germoplásmica uno.
- Testigos = Plantas comerciales (que ya están en el mercado).
- Cruza vs Testigo = Este es un contraste que mide el grado de potencialidad entre cruzas nuevas y las cruzas comerciales.

Todos estos efectos ya mencionados interactúan con localidad con el objeto de analizar su contribución correspondiente.

## 4 Resultados

A continuación se obtienen las esperanzas de cuadrados medios de cada una de las fuentes de variación correspondientes al modelo, así como de sus particiones, en este caso sólo se presenta el resultado final, pero en el apéndice se describe el desarrollo de algunos efectos para un mejor entendimiento del tema.

$$\begin{aligned}
E[CM(Loc)] &= rt\sigma_L^2 + r\sigma_{LG}^2 + t\sigma_{r/L}^2 + \sigma_e^2 \\
E[CM(Rep/Loc)] &= t\sigma_{r/L}^2 + \sigma_e^2 \\
E[CM(Gen)] &= rl\sigma_G^2 + r\sigma_{LG}^2 + \sigma_e^2 \\
E[CM(Cruza)] &= rl\sigma_C^2 + r\sigma_{LC}^2 + \sigma_e^2 \\
E[CM(Lin/p_1)] &= rl\sigma_{l/p_1}^2 + r\sigma_{L(l/p_1)}^2 + \sigma_e^2 \\
E[CM(g_1/p_1)] &= rl\sigma_{g_1/p_1}^2 + r\sigma_{L(g_1/p_1)}^2 + \sigma_e^2 \\
E[CM(g_2/p_1)] &= rl\sigma_{g_2/p_1}^2 + r\sigma_{L(g_2/p_1)}^2 + \sigma_e^2 \\
E[CM(g_3/p_1)] &= rl\sigma_{g_3/p_1}^2 + r\sigma_{L(g_3/p_1)}^2 + \sigma_e^2 \\
E[CM(g_4/p_1)] &= rl\sigma_{g_4/p_1}^2 + r\sigma_{L(g_4/p_1)}^2 + \sigma_e^2 \\
E[CM(g/p_1)] &= rl\sigma_{g/p_1}^2 + r\sigma_{L(g/p_1)}^2 + \sigma_e^2
\end{aligned}$$

$$\begin{aligned}
E[CM(Lin/p_2)] &= rl\sigma_{l/p_2}^2 + r\sigma_{L(l/p_2)}^2 + \sigma_e^2 \\
E[CM(g_1/p_2)] &= rl\sigma_{g_1/p_2}^2 + r\sigma_{L(g_1/p_2)}^2 + \sigma_e^2 \\
E[CM(g_2/p_2)] &= rl\sigma_{g_2/p_2}^2 + r\sigma_{L(g_2/p_2)}^2 + \sigma_e^2 \\
E[CM(g_3/p_2)] &= rl\sigma_{g_3/p_2}^2 + r\sigma_{L(g_3/p_2)}^2 + \sigma_e^2 \\
E[CM(g_4/p_2)] &= rl\sigma_{g_4/p_2}^2 + r\sigma_{L(g_4/p_2)}^2 + \sigma_e^2 \\
E[CM(g/p_2)] &= rl\sigma_{g/p_2}^2 + r\sigma_{L(g/p_2)}^2 + \sigma_e^2 \\
E[CM(l/p_1 \text{ vs } l/p_2)] &= rl(\sigma_{l/p_1}^2 + \sigma_{l/p_2}^2 - \sigma_c^2) + \\
&\quad r(\sigma_{L(l/p_1)}^2 + \sigma_{L(l/p_2)}^2 - \sigma_{Lc}^2) + \sigma_e^2 \\
E[CM(Tes)] &= rl\sigma_t^2 + r\sigma_{Lt}^2 + \sigma_e^2 \\
E[CM(Cruza \text{ vs } test)] &= rl(\sigma_c^2 + \sigma_t^2 - \sigma_G^2) + \\
&\quad r(\sigma_{Lc}^2 + \sigma_{Lt}^2 - \sigma_{LG}^2) + \sigma_e^2 \\
E[CM(Gen \times Loc)] &= r\sigma_{LG}^2 + \sigma_e^2 \\
E[CM(Cruza \times Loc)] &= r\sigma_{Lc}^2 + \sigma_e^2 \\
E[CM((l/p_1) \times Loc)] &= r\sigma_{L(l/p_1)}^2 + \sigma_e^2 \\
E[CM((g_1/p_1) \times Loc)] &= r\sigma_{L(g_1/p_1)}^2 + \sigma_e^2 \\
E[CM((g_2/p_1) \times Loc)] &= r\sigma_{L(g_2/p_1)}^2 + \sigma_e^2 \\
E[CM((g_3/p_1) \times Loc)] &= r\sigma_{L(g_3/p_1)}^2 + \sigma_e^2 \\
E[CM((g_4/p_1) \times Loc)] &= r\sigma_{L(g_4/p_1)}^2 + \sigma_e^2 \\
E[CM((g/p_1) \times Loc)] &= r\sigma_{L(g/p_1)}^2 + \sigma_e^2 \\
E[CM((l/p_2) \times Loc)] &= r\sigma_{L(l/p_2)}^2 + \sigma_e^2 \\
E[CM((g_1/p_2) \times Loc)] &= r\sigma_{L(g_1/p_2)}^2 + \sigma_e^2 \\
E[CM((g_2/p_2) \times Loc)] &= r\sigma_{L(g_2/p_2)}^2 + \sigma_e^2 \\
E[CM((g_3/p_2) \times Loc)] &= r\sigma_{L(g_3/p_2)}^2 + \sigma_e^2 \\
E[CM((g_4/p_2) \times Loc)] &= r\sigma_{L(g_4/p_2)}^2 + \sigma_e^2 \\
E[CM((g/p_2) \times Loc)] &= r\sigma_{L(g/p_2)}^2 + \sigma_e^2 \\
E[CM((l/p_1 \text{ vs } l/p_2) \times Loc)] &= r(\sigma_{L(l/p_1)}^2 + \sigma_{L(l/p_2)}^2 - \sigma_{Lc}^2) + \sigma_e^2 \\
E[CM(Tes \times Loc)] &= r\sigma_{Lt}^2 + \sigma_e^2 \\
E[CM((Cruza \text{ vs } Tes) \times Loc)] &= r(\sigma_{Lc}^2 + \sigma_{Lt}^2 - \sigma_{LG}^2) + \sigma_e^2 \\
E[CM(Error)] &= \sigma_e^2
\end{aligned}$$

En seguida se presentan cada uno de los valores de las componentes de varianza estimadas(C.V.E) de efectos principales e interacciones, así como de sus particiones para

la variable rendimiento de mazorca en  $ton/ha^{-1}$  del experimento de Ramos Maceda (1994) como se aprecia en el Cuadro 1.1.

Cuadro 1.1: Estimados de Componentes de Varianza por Fuente de Variación.

F.V.	C.V.E.
localidad	0.62237
rep/Loc	0.23390
Genotipos	11.96200
cruzas	0.18670
lin/p1	0.17400
g1/p1	0.07250
g2/p1	0.07570
g3/p1	0.03450
g4/p1	0.67150
g/p1	1.45770
lin/p2	0.25220
g1/p2	0.16650
g2/p2	0.16170
g3/p2	0.04570
g4/p2	0.61720
g/p2	2.00670
lin/p1 vs lin/p2	0.47870
testigos	3.33300
cruzas vs testigos	2.22770
Gen * Loc	3.35850
cruzas * Loc	0.13100
(lin/p1) * Loc	0.13850
(g1/p1) * Loc	0.16050
(g2/p1) * Loc	0.14650
(g3/p1) * Loc	0.14250
(g4/p1) * Loc	0.30400
(g/p1) * Loc	0.14500
(lin/p2) * Loc	0.12250
(g1/p2) * Loc	0.14150
(g2/p2) * Loc	0.14200
(g3/p2) * Loc	0.15600
(g4/p2) * Loc	0.06650
(g/p2) * Loc	0.61750
(lin/p1 vs lin/p2) * Loc	0.73700
testigos * Loc	0.14300
(cruza vs test) * Loc	0.06450
error	1.26800
total	17.44477

En este trabajo se obtuvo no sólo la magnitud relativa de las varianzas, sino también los porcentajes de la suma de las varianzas de cada uno de ellos, como se observa en el cuadro 1.2.

Cuadro 1.2 : Porcentajes de Componentes de Varianza para los efectos considerados en el modelo.

$\sigma_L^2$	representa	$0.62237 \times 100$ 17.44477	3.56765%
$\sigma_{r/L}^2$	representa	$0.23390 \times 100$ 17.44477	1.34080%
$\sigma_G^2$	representa	$11.962 \times 100$ 17.44477	68.57069%
$\sigma_c^2$	representa	$0.18670 \times 100$ 17.44477	1.07023%
$\sigma_{l/p_1}^2$	representa	$0.17400 \times 100$ 17.44477	0.99743%
$\sigma_{q_1/p_1}^2$	representa	$0.07250 \times 100$ 17.44477	0.41559%
$\sigma_{q_2/p_1}^2$	representa	$0.07570 \times 100$ 17.44477	0.43394%
$\sigma_{q_3/p_1}^2$	representa	$0.03450 \times 100$ 17.44477	0.19776%
$\sigma_{q_4/p_1}^2$	representa	$0.67150 \times 100$ 17.44477	3.84929%
$\sigma_{q_1/p_2}^2$	representa	$1.45770 \times 100$ 17.44477	8.35608%
$\sigma_{q_2/p_2}^2$	representa	$0.25220 \times 100$ 17.44477	1.44571%
$\sigma_{q_3/p_2}^2$	representa	$0.16650 \times 100$ 17.44477	0.95444%
$\sigma_{q_4/p_2}^2$	representa	$0.16170 \times 100$ 17.44477	0.92692%
$\sigma_{q_1/p_1}^2$ vs $l/p_2$	representa	$0.04570 \times 100$ 17.44477	0.26196%
$\sigma_{q_4/p_2}^2$	representa	$0.61720 \times 100$ 17.44477	3.53802%
$\sigma_{q_1/p_2}^2$	representa	$2.00670 \times 100$ 17.44477	11.50316%
$\sigma_{l/p_1}^2$ vs $l/p_2$	representa	$0.47870 \times 100$ 17.44477	2.74408%
$\sigma_t^2$	representa	$3.33300 \times 100$ 17.44477	19.10601%
$\sigma_c^2$ vs $t$	representa	$2.12770 \times 100$ 17.44477	12.77002%
$\sigma_{G*L}^2$	representa	$3.3585 \times 100$ 17.44477	19.25218%
$\sigma_{c*L}^2$	representa	$0.13100 \times 100$ 17.44477	0.75094%
$\sigma_{(l/p_1)*L}^2$	representa	$0.13850 \times 100$ 17.44477	0.79393%
$\sigma_{(g_1/p_1)*L}^2$	representa	$0.16050 \times 100$ 17.44477	0.92005%
$\sigma_{(g_2/p_1)*L}^2$	representa	$0.14650 \times 100$ 17.44477	0.83979%
$\sigma_{(g_3/p_1)*L}^2$	representa	$0.14250 \times 100$ 17.44477	0.81686%
$\sigma_{(g_4/p_1)*L}^2$	representa	$0.30400 \times 100$ 17.44477	1.74264%
$\sigma_{(g_1/p_2)*L}^2$	representa	$0.14500 \times 100$ 17.44477	0.83119%
$\sigma_{(l/p_2)*L}^2$	representa	$0.12250 \times 100$ 17.44477	0.70222%
$\sigma_{(g_1/p_2)*L}^2$	representa	$0.14150 \times 100$ 17.44477	0.81113%
$\sigma_{(g_2/p_2)*L}^2$	representa	$0.14200 \times 100$ 17.44477	0.81399%
$\sigma_{(g_3/p_2)*L}^2$	representa	$0.15600 \times 100$ 17.44477	0.89425%
$\sigma_{(g_4/p_2)*L}^2$	representa	$0.06650 \times 100$ 17.44477	0.38120%
$\sigma_{(g_1/p_2)*L}^2$	representa	$0.61750 \times 100$ 17.44477	3.53974%
$\sigma_{(l/p_1) \text{ vs } l/p_2)*L}^2$	representa	$0.73700 \times 100$ 17.44477	4.22476%
$\sigma_{t*L}^2$	representa	$0.14300 \times 100$ 17.44477	0.81973%
$\sigma_{(c \text{ vs } t)*L}^2$	representa	$0.06450 \times 100$ 17.44477	0.36974%
$\sigma_e^2$	representa	$1.26800 \times 100$ 17.44477	7.26865%

## 5 Conclusiones

En lo que respecta a este trabajo, se obtuvieron los grados de libertad y el estadístico apropiado para probar las hipótesis nulas correspondientes, además se observa que fueron los efectos del ambiente los que más contribuyeron en la respuesta . Y esto lo que significa es que la localidad de General Cepeda, fue más rendidora que la de Parras de la Fuente, Coahuila., respectivamente. En base a las estimaciones de componentes de varianza y a la futura explotación comercial de los híbridos en diferentes localidades, se espera encontrar la mejor combinación que facilite al investigador genetista seleccionar nuevo germoplasma con mayor grado de confiabilidad.

## Apéndice

Con el objeto de dar una idea más enfocada de lo que se está haciendo, aquí se presentan desarrollos de las esperanzas de cuadrados medios de algunos de los efectos estimados, tales como los de localidad y repetición dentro de localidad, el desarrollo de las restantes estimaciones que se presentan en este trabajo son semejantes, y en todos ellos se involucran las ecuaciones dadas en la Sección 3.

Como ya se mencionó, para encontrar la esperanza de los cuadrados medios, se inicia obteniendo la esperanza de la suma de cuadrados del efecto correspondiente y después se divide dicha esperanza entre sus respectivos grados de libertad.

$$\begin{aligned} E[SC(LOC)] &= \frac{E[\sum_k Y_{..k}^2]}{rt} - \frac{E[Y_{...}^2]}{rlt} \\ &= rtl\mu^2 + rtl\sigma_L^2 + tl\sigma_{r/L}^2 + rl\sigma_G^2 + rl\sigma_{LG}^2 + l\sigma_e^2 \\ &\quad - (rtl\mu^2 + rt\sigma_L^2 + t\sigma_{r/L}^2 + rl\sigma_G^2 + r\sigma_{LG}^2 + \sigma_e^2) \\ &= rt(l-1)\sigma_L^2 + r(l-1)\sigma_{LG}^2 + t(l-1)\sigma_{r/L}^2 + (l-1)\sigma_e^2 \end{aligned}$$

$$\begin{aligned} E\left[\frac{SC(Loc)}{l-1}\right] &= E[CM(Loc)] \\ &= rt\sigma_L^2 + r\sigma_{LG}^2 + t\sigma_{r/L}^2 + \sigma_e^2 \end{aligned}$$

$$\begin{aligned} E[SC(Rep/Loc)] &= \frac{E[\sum_{jk} Y_{.jk}^2]}{t} - \frac{E[\sum_k Y_{..k}^2]}{tr} \\ &= rtl\mu^2 + rtl\sigma_L^2 + rtl\sigma_{r/L}^2 + rl\sigma_G^2 + rl\sigma_{LG}^2 + rl\sigma_e^2 \\ &\quad - (rtl\mu^2 + rtl\sigma_L^2 + tl\sigma_{r/L}^2 + rl\sigma_G^2 + rl\sigma_{LG}^2 + l\sigma_e^2) \\ &= tl(r-1)\sigma_{r/L}^2 + l(r-1)\sigma_e^2 \end{aligned}$$

$$E\left[\frac{SC(Rep/Loc)}{(r-1)l}\right] = E[CM(Rep/Loc)] = t\sigma_{r/L}^2 + \sigma_e^2$$

## Referencias

- Brownlee, K.A. (1984). *Statistical Theory and Methodology, In Science and Engineering.* Second edition. Malabar: Krieger.
- Ramos, M.L. (1994). Estimación de la Habilidad Combinatoria General y Específica de Líneas de Maíz (*Zea mays L.*) para Condiciones de Humedad Restrictiva. Tesis licenciatura UAAAN.
- Rodríguez, B.L. (1992). Esperanzas de Cuadrados Medios de un Diseño de Bloques al Azar con arreglo Factorial Combinatorio y Partición de Efectos. Tesis postgrado UAAAN.
- Searle, S.R. (1987). *Linear Models for Unbalanced Data.* New York: Wiley.
- Searle, S.R., Casella, G. y McCulloch, C.H.E. (1992) *Variance Components.* New York: Wiley.

# Sobre la Convergencia del Método de Ascenso por Pendiente Máxima

Blanca R. Pérez S.

Sergio de los Cobos S.

*UAM-Iztapalapa*

*UAM-Iztapalapa*

*Depto. de Matemáticas*

*Depto. de Ingeniería*

Miguel A. Gutiérrez A.

*UAM-Azcapotzalco*

*Depto. de Sistemas*

## 1 Introducción

La metodología de superficies de respuestas tiene como finalidad estimar la máxima producción de un proceso y el punto donde ésta se alcanza. Si la función de respuesta está dada por la expresión

$$E(Y|\mathbf{x}) = \eta(\mathbf{x}) = \beta_0 + \mathbf{b}^T \mathbf{x} + \mathbf{x}^T B \mathbf{x}$$

con  $B$ , una matriz negativa definida, entonces la respuesta máxima es igual a

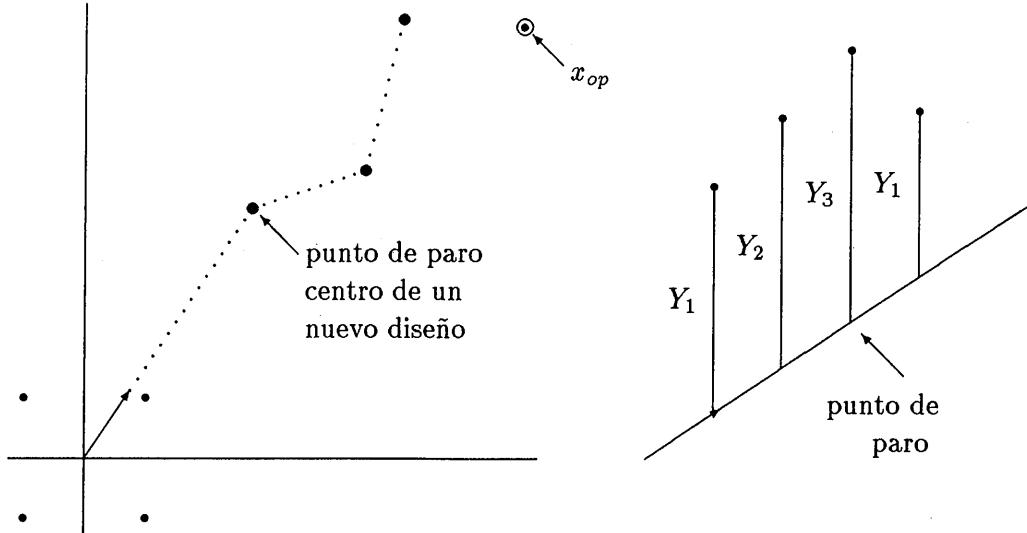
$$\eta_{op} = \beta_0 - \frac{1}{4} \mathbf{b}^T B^{-1} \mathbf{b},$$

y se alcanza en

$$\mathbf{x}_{op} = -\frac{1}{2} B^{-1} \mathbf{b}.$$

El método de ascenso por pendiente máxima permite acercarse a  $\mathbf{x}_{op}$  por medio de movimientos sucesivos que se hacen a lo largo de una ruta de ascenso.

La ruta de ascenso se estima con la aproximación lineal de la función de respuesta, circunvecina al diseño. Se tiene interés en acercarse al óptimo para después estimarlo con un diseño cercano usando una función cuadrática. Una buena ruta de ascenso es aquélla que nos acerca más rápidamente al óptimo. El método de ascenso por pendiente máxima es el más conocido para estimar la ruta de ascenso, y esta ruta depende de los parámetros de la función de respuesta y de la estandarización de los datos, como se verá en las siguientes secciones.



Acercamiento al óptimo en tres etapas, con cada nuevo diseño se estima nuevamente la ruta de ascenso.

El punto de paro se encuentra cuando las sucesivas observaciones sobre la ruta de ascenso presentan una caída.

## 2 La Estandarización de los Datos

Para tener un diseño experimental conocido se deben estandarizar los datos, esto es: si las observaciones originales son:

$$\xi^T = (\xi_1, \xi_2, \dots, \xi_m)$$

y la función de respuesta con estas variables es

$$\eta(\xi) = \alpha_0 + \mathbf{a}^T \xi + \xi^T A \xi,$$

el vector  $\xi$  se estandariza, centrándolo en el origen y reescalándolo mediante la transformación

$$\mathbf{x} = \Lambda(\xi - \mathbf{c}).$$

con  $\Lambda$  una matriz diagonal y  $\mathbf{c}$  un vector. En estas condiciones, la ruta de ascenso con las variables estandarizadas (el vector  $\mathbf{x}$ ) está dado por

$$\{\mathbf{x} \mid \mathbf{x} = t\mathbf{b}; \quad t \in \mathbb{R}\}$$

mientras que la misma ruta de ascenso con las variables originales (el vector  $\xi$ ) es igual a

$$\{\xi \mid \xi = \mathbf{c} + t\Lambda^{-2}(\mathbf{a} + A\mathbf{c}); \quad t \in \mathbb{R}\}.$$

Como se puede ver, esta ruta de ascenso depende de la estandarización de los datos, y nos puede acercar más o menos rápido al vector  $\xi_{op}$ , de acuerdo a que tan bien se haya elegido la posición del diseño inicial y el escalamiento. Obsérvese además que la ruta se calculó sin considerar los errores aleatorios de las observaciones.

### 3 La Convergencia del Método de Ascenso por Pendiente Máxima

Se debe recordar que la finalidad del método es acercarse al vector:

$$\mathbf{x}_{op} = -\frac{1}{2}B^{-1}\mathbf{b}$$

o a su equivalente, el vector:

$$\xi_{op} = -\frac{1}{2}A^{-1}\mathbf{a}.$$

Usando un diseño factorial  $3^n$ , se calcularon las rutas de ascenso en las subsecuentes iteraciones. La ruta de ascenso en la primer etapa es igual a  $\mathbf{v} = \mathbf{b}$ , y el punto donde se alcanza el óptimo en esa dirección es igual a

$$\mathbf{c}_1 = -\frac{1}{2} \frac{\mathbf{b}^T \mathbf{b}}{\mathbf{b}^T B \mathbf{b}} \mathbf{b}.$$

En general, la ruta de ascenso en la  $(i+1)$ -ésima iteración está dada por:

$$\mathbf{v}_{i+1} = \mathbf{c}_i + t\mathbf{b}_{i+1}, \quad t \in \mathbb{R}.$$

y el punto donde se alcanza la máxima respuesta en esa dirección.

$$\mathbf{c}_{i+1} = \mathbf{c}_i - \frac{1}{2} \frac{\mathbf{b}_{i+1}^T \mathbf{b}_{i+1}}{\mathbf{b}_{i+1}^T B \mathbf{b}_{i+1}} \mathbf{b}_{i+1}$$

que están definidos por la relación de recurrencia  $\mathbf{b}_1 = \mathbf{b}$ ;  $\mathbf{c}_0 = \mathbf{0}$  y  $\mathbf{b}_{i+1} = \mathbf{b}_i + \frac{\mathbf{b}_i^T \mathbf{b}_i}{\mathbf{b}_i^T B \mathbf{b}_i} B \mathbf{b}_i$ .

Entonces en cada iteración, la diferencia entre  $\mathbf{c}_i$  y  $\mathbf{x}_{op}$  se puede escribir como

$$\mathbf{c}_i - \mathbf{x}_{op} = \frac{1}{2} \left( B^{-1} - \frac{\mathbf{b}_i^T \mathbf{b}_i}{\mathbf{b}_i^T B \mathbf{b}_i} \right) \mathbf{b}_i.$$

Es fácil ver que si  $\mathbf{b}$  es vector propio de la matriz  $B$ , o si  $B = I$ , se llega al óptimo en el primer paso.

En otros casos, se puede probar que

$$\|\mathbf{c}_{i+1} - \mathbf{x}_{op}\| < \|\mathbf{c}_i - \mathbf{x}_{op}\|,$$

pero si los valores propios de la matriz  $B$  son muy diferentes entre sí, la convergencia puede ser lenta.

## 4 Conclusiones

Los resultados se obtuvieron sin considerar el factor aleatorio intrínseco en las observaciones, considerando esto, se encontró que la convergencia del método de ascenso por pendiente máxima hacia la respuesta óptima puede ser muy rápida, inclusive en el primer paso.

Pero también se pueden encontrar ejemplos en los que la convergencia sea lenta, por ejemplo que se cumpla la relación

$$r_n \|\mathbf{c}_i - \mathbf{x}_{op}\| < \|\mathbf{c}_{i+1} - \mathbf{x}_{op}\| < \|\mathbf{c}_i - \mathbf{x}_{op}\|$$

para  $r_n = \frac{n-1}{n}$ .

Además, la ruta de ascenso depende de la estandarización. Ante estas razones es aceptable considerar que la ruta que mejor nos acerca al vector  $\mathbf{x}_{op}$ , es la que va directo a él. Por lo tanto se propone que se utilice como ruta de ascenso la dada por el vector:

$$\mathbf{v} = \frac{\hat{\mathbf{x}}_{op}}{|\hat{\mathbf{x}}_{op}|}.$$

# Modelo Lineal Difuso

(Una Aplicación)

José C. Romero Cortés y Arturo Aguilar Vázquez

*UAM-Departamento de Sistemas*

## 1 Introducción

Los modelos estadísticos lineales son importantes no sólo por los desarrollos estadísticos-matemáticos alcanzados sino también por su aplicación a situaciones reales. Sin embargo, estas relaciones no son de utilidad para explicar las estimaciones humanas, sólo se concretan en trabajar con datos generados mediante la planeación y diseño de experimentos o a lo más con datos que la naturaleza pone a nuestra disposición para explicarlos en función de variables independientes.

La experiencia humana, cuando es inteligente conviene modelarla, esto es, explicarla en términos de variables independientes vía algún modelo. Inicialmente se propone un modelo lineal en los parámetros, donde la variable dependiente es borrosa a diferencia del enfoque convencional donde éstas corresponden a variables aleatorias observables. Esto implica que los parámetros sean de naturaleza difusa y el problema de su estimación no puede enfocarse aplicando los métodos de estimación puntual convencionales, como mínimos cuadrados, máxima verosimilitud, etc. La programación lineal se utiliza para estimar el modelo difuso, específicamente en este artículo se analiza el problema dual y sus implicaciones en reducción de la borrosidad. Se discute una aplicación del modelo difuso, con la estimación del precio de venta de casas habitación, considerando como variables independientes la superficie de terreno, de construcción, su ubicación geográfica, etc.

## 2 Modelo de Regresión Lineal Múltiple Difuso (MRLMD)

Es de la forma:

$$\underline{Y}_i = \underline{\beta}_0 + \underline{\beta}_1 X_{i1} + \underline{\beta}_2 X_{i2} + \dots + \underline{\beta}_k X_{ik}, \quad i = 1, 2, \dots, n. \quad (1)$$

donde:

$\underline{Y}_i$  = Valores estimados por un experto o que provienen de alguna fuente donde esta estimación estuvo influída por el humano,  $i=1,2,3,\dots,n$ .

$\beta_j$  = Parámetros borrosos,  $j=0, 1, 2, \dots, k$ .

El símbolo  $\sim$  asociado a las  $Y_i$ ,  $\beta_j$  indica que éstos son borrosos o difusos, es decir, que tienen asociadas funciones de pertenencia o distribuciones de posibilidad.

$X_{ij}$  = Valor fijado para la variable independiente  $j$ , en la  $i$ -ésima muestra, con  $X_{io} = 1$  para  $i = 1, 2, \dots, n$  y  $j = 0, 1, 2, \dots, k$ .

Lo anterior queda expresado mediante:

**TABLA 1**

Muestra	Respuesta difusa	Valores de variables independientes
1	$(y_1, e_1)$	$x_{11}, \dots, x_{1k}$
2	$(y_2, e_2)$	$x_{21}, \dots, x_{2k}$
3	$(y_3, e_3)$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
$n$	$(y_n, e_n)$	$x_{n1}, \dots, x_{nk}$

Las  $y_i$  asociadas a las estimaciones humanas,  $i = 1, 2, \dots, n$ , se asumen borrosas con funciones de pertenencia triangulares con dispersión  $e_i$ . Análogamente se asume que los  $\beta_j$  tienen asociadas funciones triangulares con centros  $\alpha_j$  y dispersión  $c_j$ .

Entonces estimar las  $\alpha_j$  y  $c_j$ , corresponde a resolver el siguiente problema de programación lineal:

$$\underset{\alpha c}{\text{Min}} Z = c_0 + c_1 + c_2 + \dots + c_k$$

s.a.

$$\begin{aligned} \alpha_0 x_{i0} + \alpha_1 x_{i1} + \dots + \alpha_k x_{ik} + (1-h) \sum_{j=0}^k c_j |x_{ij}| &\geq y_i + (1-h)e_i, \quad i = 1, 2, \dots, n \\ &\vdots \\ -\alpha_0 x_{i0} - \alpha_1 x_{i1} - \dots - \alpha_k x_{ik} + (1-h) \sum_{j=0}^k c_j |x_{ij}| &\geq -y_i + (1-h)e_i, \quad i = 1, 2, \dots, n \\ c_j &\geq 0, \quad j = 0, 1, 2, \dots, k. \end{aligned} \tag{2}$$

$$\alpha_j \in \Re, \quad j = 0, 1, 2, \dots, k.$$

Este programa primal tiene asociado el dual:

$$\underset{WD}{\text{Max}} Z = \sum_{i=1}^n W_i \left[ y_i + (1/h)e_i - d \sum_{j=0}^k x_{ij} \right] - \sum_{i=1}^n D_i \left[ y_i - (1-h)e_i - d \sum_{j=0}^k x_{ij} \right]$$

s.a.

$$\begin{aligned}
 & (1-h) \sum_{i=1}^n W_i |x_{ij}| - (1-h) \sum_{j=0}^n D_i |x_{ij}| \leq 1, j = 0, 1, 2, \dots, k. \\
 & \quad \vdots \\
 & \sum_{i=1}^n W_i x_{ij} - \sum_{i=1}^N d_I X_{IJ} \leq 0, j = 0, 1, 2, \dots, k.
 \end{aligned} \tag{3}$$

$$w_i \geq 0, \quad i = 1, 2, \dots, n.$$

$$D_i \geq 0, \quad i = 1, 2, \dots, n.$$

El problema anterior ha sido planteado por varios autores (ver Tanaka, 1982 y Tanaka y Asai, 1981), el aporte de este artículo es hacer análisis de sensibilidad para investigar cómo los valores  $\alpha_j^o$ ,  $c_j^o$  y  $z^o$  varían al cambiar los valores de los parámetros, al añadir observaciones y al agregar o remover variables independientes al MRLMD. Esto indicará en cuánto se corrige la borrosidad o imprecisión de las estimaciones humanas mediante las distribuciones de posibilidad asociadas a  $\underline{Y}_i = (y_i, e_i)$  y las  $\beta_j = (\alpha_j, c_j)$ , para  $i = 1, 2, \dots, n; j = 0, 1, 2, \dots, k$ .

Teniendo formulado el problema dual es fácil realizar sensibilidad y parametrización. A continuación, se presentan los casos clásicos de sensibilidad, pero sobre todo su significado asociado al análisis de Regresión Lineal Múltiple Difuso.

**i) Cambios en  $\pm y_i + (1-h)e_i$** : Estos miembros derechos del primal se les denota en la literatura de la programación lineal como  $b_i$ . Las variables duales óptimas  $W_i^o$ ,  $D_i^o$  corresponden a los precios sombra, cuya interpretación económica en este contexto debe traducirse en una interpretación de ganancia en conocimiento del sistema o reducción de borrosidad, pudiendo investigar el máximo (mínimo) incremento admisible a estos miembros derechos, manteniendo las mismas variables en la solución óptima, permitiendo corregir las percepciones borrosas expresadas en los datos.

**ii) Introducción o remoción de una nueva variable independiente**: Si se introduce una nueva variable independiente  $x_{i,k+1}$ , al MRLMD que corresponde a  $\beta_{k+1}$ , desde el punto de vista de la programación lineal sólo basta investigar si la correspondiente restricción añadida al dual se satisface también para la solución óptima, en caso contrario esta variable es básica, y se debe reoptimizar. Algo análogo ocurre cuando se remueve una variable original. Lo anterior implica mantener las  $\alpha_j^o$  y  $c_j^o$  básicas óptimas cuando las  $\alpha_{k+1}$  y  $c_{k+1}$  no sean básicas. En caso contrario habrá que reoptimizar con posibles cambios en la solución óptima, lo que puede mover algunas distribuciones de posibilidad de  $\beta_j^o$  y puede implicar una reducción en borrosidad, aunque también puede indicar lo contrario.

El caso de remoción nuevamente contempla 2 casos, si alguna de las  $\alpha_j^o$ ,  $c_j^o$  básicas óptimas es removida, a menos de que pudiera darse degeneración, la solución cambia con

incremento en  $z^o$ . Pero si la remoción corresponde a una  $\beta_j^o$  cuyas  $\alpha_j^o, c_j^o$  no sean básicas entonces la solución se mantiene.

**iii) Introducción de una nueva restricción** Esto se refiere a incluir en el modelo una nueva  $Y_{n+1}$ , asociada a  $\{x_{n+1,1}, x_{n+1,2}, \dots, x_{n+1,k}\}$ . Si esta es satisfecha por la solución óptima entonces se conserva la solución, si no habrá que iterar hasta obtener el nuevo óptimo. Este punto es importante porque sabemos que, a mayor tamaño de muestra se espera menor borrosidad.

**iv) Cambios en coeficientes de variables básicas y no básicas.** Esto no es importante en el contexto de la Regresión Lineal Múltiple Difusa porque los  $x_{ij}$  las suponemos fijas y corresponden al diseño del experimento. Respecto a la programación paramétrica sólo es de interés cambiar los valores  $\pm y_i + (1 - h)e_i$  y significa valorar la sensibilidad de la solución  $z^o$ ,  $\alpha_j^o$ ,  $c_j^o$ , cuando estos miembros derechos se incrementan en hasta  $\delta_i\theta$ , esto es, analizar sobre todo cuando se reduce  $z^o$  al mover alguno o algunos  $\pm y_i + (1 - h)e_i$  en  $\delta_i\theta$ , donde  $\theta > 0$  y  $\delta_i \in \mathbb{R}$ .

### 3 Aplicación

A continuación se da un ejemplo para ilustrar al MRLMD. Considere el desarrollo de un modelo del precio de venta de casas habitación en la Cd. de México y su área metropolitana sobre la base de la superficie del terreno, superficie construida, edad de la construcción y su ubicación o colonia a la que pertenece. Se obtuvo una muestra de expertos en esta actividad como son corredores de bienes inmuebles y hasta los propios dueños de los inmuebles para facilitar los datos que a continuación se dan.

**TABLA 2**

Casa no.	(Precio de venta, $e_i$ ) (en miles de \$)	Superficie Terreno ( $m^2$ )	Superficie Construida ( $m^2$ )	Edad (años)	Localización
$c_1$	(390, 10)	136	203	30	tacuba
$c_2$	(590, 25)	132	185	11	satélite
$c_3$	(800, 40)	120	150	9	san jerónimo
$c_4$	(280, 20)	80	130	10	ajusco
$c_5$	(480, 15)	200	100	25	agrícola oriental
$c_6$	(750, 30)	240	220	13	azcapotzalco
$c_7$	(950, 50)	215	240	35	campestre churubusco
$c_8$	(500, 30)	122	244	13	estrella tepeyac
$c_9$	(1 250, 100)	300	360	10	las águilas
$c_{10}$	(725, 40)	162	184	10	colinas del sur
$c_{11}$	(725, 30)	150	145	9	san jerónimo
$c_{12}$	(690, 30)	200	200	12	tepepan

Fuente: Segundamano, Sección inmuebles. 25 de febrero de 1997

Ya que la variable dependiente corresponde a estimaciones humanas, esta deberá mo-

delarse usando MRLMD, quedando de la forma:

$$\tilde{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6}, \quad i = 1, 2, \dots, 12 \quad (4)$$

Las variables independientes 4°, 5° y 6° son variables “dummy” (asumen valores 0, 1). La estimación de las distribuciones de posibilidad de la  $\beta_j$ , se obtiene resolviendo el correspondiente problema de programación lineal, para esto se consideró  $h = 0.5$ .

**TABLA 3**

Parámetros	$\beta_0^o$	$\beta_1^o$	$\beta_2^o$	$\beta_3^o$	$\beta_4^o$	$\beta_5^o$	$\beta_6^o$
Centro $\alpha_j^o$	0	2.703	1.024	- 0.238	-212.775	4.889	249.474
Borrosidad $c_j^o$	0	0	1.260	0	0	0	0

$$Z^o = \text{borrosidad total} = 1.26$$

Digamos que para la casa no. 9 se tiene que la estimación de su precio y borrosidad, está dada por la pareja: (1 426 976, 453 952) contra el real (1 250 000, 150 000), lo cual es correcto, ya que  $Y_i^h \subset Y_i^{*h}$ , con el  $h = 0.5$  que se utilizó.

El modelo anterior hace las veces de un experto o de un catálogo que contenga los precios de casas. Así por ejemplo el precio de una casa habitación ubicada en Sn. Jerónimo con una superficie de 120  $m^2$  con 200  $m^2$  construidos y una antigüedad de 12 años ofrecida en \$750 000, el modelo valúa esta en \$775 923 con una borrosidad de \$252 195. Si se realiza postoptimalidad decrementando los miembros derechos en lo permisible, sobre las restricciones con precios sombra no cero, se tiene el siguiente modelo:

**TABLA 4**

Parámetro Borroso	$\beta_0^o$	$\beta_1^o$	$\beta_2^o$	$\beta_3^o$	$\beta_4^o$	$\beta_5^o$	$\beta_6^o$
Centro $\alpha_j^o$	0	3.143	0.592	- 8.228	0	109.861	319.964
borrosidad $c_j^o$	0	0	1.029	0	0	0	0

$$Z^o = \text{borrosidad total} = 1.029$$

Como puede observarse en base al análisis de sensibilidad sobre los miembros derechos se reduce la borrosidad en un 18.3%. Esto logra que el análisis de regresión sea menos difuso, lo que equivale con el enfoque convencional a que fuera más preciso.

Sólo por brevedad en lo referente a los demás puntos de sensibilidad, se podría introducir otra variable al modelo; por ejemplo, el número de niveles de la casa y lo único que habría que investigar si la nueva solución básica óptima contiene los  $\alpha_7, c_7$  a nivel cero, en cuyo caso se mantiene la solución pero si esto no sucede habrá que encontrar la nueva solución, y tendríamos reducción en la borrosidad.

El introducir una nueva observación equivale a involucrar otra casa, digamos la número 14, con  $\underline{Y}_{14} = (450\ 000, 5\ 000)$  con  $X_{1,14} = 160, X_{2,14} = 120, X_{3,14} = 35, X_{4,14} = 1, X_{5,14} = 0, X_{6,14} = 0$ , está ubicada en Villa de Guadalupe, como estos datos satisfacen la solución al problema de programación lineal considerado, entonces la distribución de posibilidad del precio según el modelo es  $\underline{Y}_{14}^* = (395891, 123505)$ .

Como se señaló, si ésta satisface la solución óptima entonces también es solución óptima para el problema ampliado, en caso contrario habría que encontrar la nueva solución. Al aumentar el número de datos se esperaría reducción en borrosidad en muchas ocasiones.

## 4 Conclusiones

Cuando en la regresión, las variables dependientes más que observaciones o variables aleatorias, provienen de una fuente de estimación humana, existe entonces borrosidad en ésta y por tanto, se puede modelar considerando los parámetros del Modelo de Regresión Lineal como borrosos, esto es, expresada esta difusividad en términos de distribuciones de posibilidad. Una de las variables independientes en la aplicación es la localización geográfica de la casa habitación, existen intentos paralelos aplicando geoestadística o datos espaciales para modelar el valor de la tierra o de casas habitación tomando en cuenta esta variable de ubicación, y también se ha enfocado este problema en el contexto de la estadística bayesiana. El enfoque difuso resulta interesante porque la experiencia es la que se modela y está influida por la superficie del terreno y de la construcción, además de la edad de ésta y por supuesto su ubicación geográfica mediante variables “dummy”, así el arreglo 1 0 0 correspondió a colonias populares; 0 1 0 a colonias medias y 0 0 1 a colonias exclusivas.

## Referencias

- Levine, D. and Berenson, M. (1992). *Basic Business Statistics Concepts and Applications*. Englewood: Prentice-Hall.
- Tanaka, H. (1982). Linear Regression with Fuzzy Model. *IEEE Transaction on Systems Man and Cybernetics*, 12:6.
- Tanaka, H. and Asai, K. (1981). Fuzzy Linear Programming with Fuzzy Parameters. *Int. Cont. on Policy Analysis and Information Systems*, Taiwan, 19/22.

# Uso de Histogramas Desplazados Promedio y Estimadores de Densidad por Kernel para el Análisis de la Frecuencia de Tallas de Datos Biológico-Pesquero

Isaías H. Salgado Ugarte y Ma. José Marques dos Santos

*FES - Zaragoza, UNAM*

## 1 Introducción

Tradicionalmente, para analizar datos de frecuencia de tallas en estudios biológico-pesqueros se han utilizado al histograma y/o el polígono de frecuencia. Los histogramas y polígonos de frecuencia son estimadores no-paramétricos de la distribución. Por lo general la escala utilizada es la frecuencia o el porcentaje. Menos común es el uso de fracciones o de densidad. Esta última escala es tan sólo una transformación lineal que garantiza que el área bajo la curva sea igual a la unidad y se pueda hacer uso directo de las leyes probabilísticas. A pesar de su uso generalizado, desde hace algunos años varios autores (por ejemplo Tarter & Kronmal, 1976) han señalado que los histogramas y polígonos de frecuencia pueden resultar demasiado burdos en estudios detallados de la distribución de datos. Existen cuatro problemas en el empleo de histogramas (Fox, 1990):

1. **Dependencia al origen:** El investigador debe escoger la posición del origen de los intervalos de clase. Por lo general se utilizan por conveniencia números redondeados. Esta subjetividad puede conducir a estimaciones engañosas debido a que un cambio en el origen puede cambiar el número de modas en la distribución (Silverman, 1986; Fox, 1990; Härdle, 1991; Scott, 1992). La Figura 3 ejemplifica este inconveniente con los datos de longitud de la trucha coralina *Plectropomus leopardus*. Cada histograma usa el mismo intervalo ( $h = 38$ ) pero con diferente origen. Pueden observarse histogramas bimodales, trimodales y tetramodales. Es posible elegir alguno de ellos para representar a la distribución de los datos pero, la elección sería arbitraria. Este procedimiento puede conducir al analista a seleccionar (intencionalmente o no) aquel histograma que mejor se adecue a sus propósitos.
2. **Dependencia en la amplitud y número de intervalos:** La amplitud del intervalo (o su número) es el parámetro que determina el grado de suavidad de la estimación.

El utilizar pocos intervalos elimina detalles de la distribución; muchos intervalos producen estimaciones muy variables (ruidosas). A pesar de la importancia de esta elección, con frecuencia se hace arbitrariamente. Como ejemplo considere las Figuras 1 y 2 para los datos de la trucha coralina. El primer histograma con 5 intervalos muestra una distribución suave (quizás gaussiana), mientras que el segundo con 50 despliega una distribución con al menos cuatro modas.

3. **Discontinuidad:** Las discontinuidades del histograma son una función de la localización arbitraria de los intervalos y los valores discretos de los datos más que una característica de la población muestreada. La densidad local sólo se calcula en el centro de cada intervalo y las barras se dibujan suponiendo un valor constante a lo largo de cada intervalo (Chambers *et al.*, 1983).

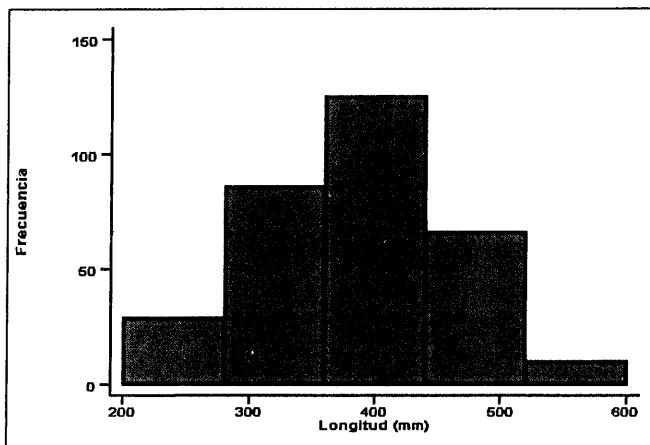


Figura 1: Histograma con cinco intervalos y origen en 200

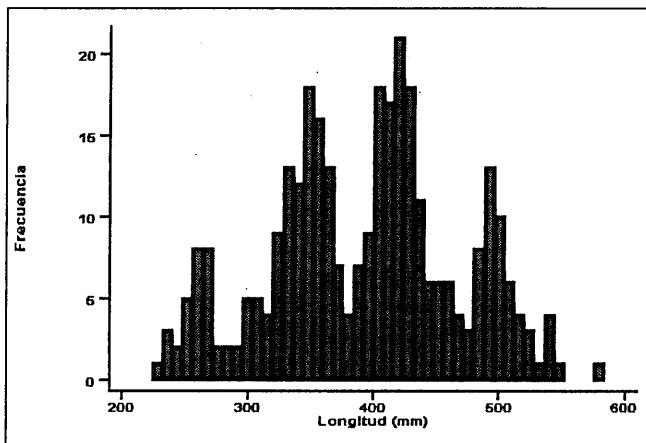


Figura 2: Histograma con 50 intervalos y origen en 200

4. **Intervalos de amplitud fija:** Si los intervalos son lo suficientemente angostos para capturar detalle donde las observaciones abundan, se produce ruido donde escasean. Este problema a menudo se trata de solucionar haciendo variar la amplitud de los intervalos extremos, pero en este caso la altura de la barra ya no es proporcional a su área, dando origen a confusión.

Para solucionar estos inconvenientes se han propuesto varios métodos. Los problemas del origen y la discontinuidad se atacan al calcular la densidad local en cada punto de los datos. En esencia, esto se realiza construyendo un intervalo de amplitud fija alrededor de cada observación. Formalmente, este intervalo o ventana es una función ponderal que asigna un peso positivo a cada observación incluida y un peso de cero a las que no. La discontinuidad es adicionalmente atacada considerando funciones ponderales de cambio gradual (y no sólo aquellas rectangulares). Estas ideas conducen a **los estimadores de densidad por kernel (EDK)**, propuestos por vez primera en 1956 por Rosenblat y cuya definición es:

$$\hat{f}(x) = \frac{1}{hn} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \quad (1)$$

donde  $\hat{f}(x)$  es la estimación de densidad de la variable  $x$ ,  $n$  es el número de observaciones,  $h$  es la amplitud de banda o parámetro de suavización y  $K(\bullet)$  es la función kernel suave y simétrica cuya integral es la unidad.

El cuadro 1, adaptado de Härdle (1991) contiene algunas de las funciones ponderales (kerneles) más comunes:

#### CUADRO 1

Algunas funciones kernel comunes

Kernel	$K(z)$
Uniforme	$1/2I( z  \leq 1)$
Triangular (ASH)	$(1 -  z )I( z  \leq 1)$
Epanechnikov	$3/4(1 - z^2)I( z  \leq 1)$
Quártica	$(15/16)(1 - z^2)^2I( z  \leq 1)$
Triponderada	$(35/32)(1 - z^2)^3I( z  \leq 1)$
Coseno	$(\pi/4) \cos((\pi/2)z)I( z  \leq 1)$
Gaussiana	$(1/\sqrt{2\pi}) \exp((-1/2)z^2)$

Todas las funciones kernel listadas en el Cuadro 1 tienen una eficiencia muy cercana a la óptima de Epanechnikov (1969). Como consecuencia, una función kernel puede escogerse por su esfuerzo computacional (Silverman, 1986). Estos EDK's emplean amplitudes fijas de banda lo que hace a las estimaciones sensibles a ruido en las colas o en cualquier otro intervalo de baja densidad. Para enfrentar este problema se ha sugerido reducir la

amplitud de banda en áreas con densidad alta y ampliarla donde las observaciones escasean. Este **EDK** variable (Jones, 1990) retiene detalle donde las observaciones se concentran y eliminan fluctuaciones ruidosas donde hay pocos datos.

Un enfoque para escoger al parámetro de suavización (amplitud de banda) es el variar  $h$  hasta que resulta una figura “adecuada” (Tarter y Kronmal, 1976). Este procedimiento recae en la valoración subjetiva del investigador, aunque puede funcionar bien para fines exploratorios (Silverman, 1986) puesto que las características en la densidad “aparecen” y “desaparecen” al cambiar la banda (Silverman, 1981a).

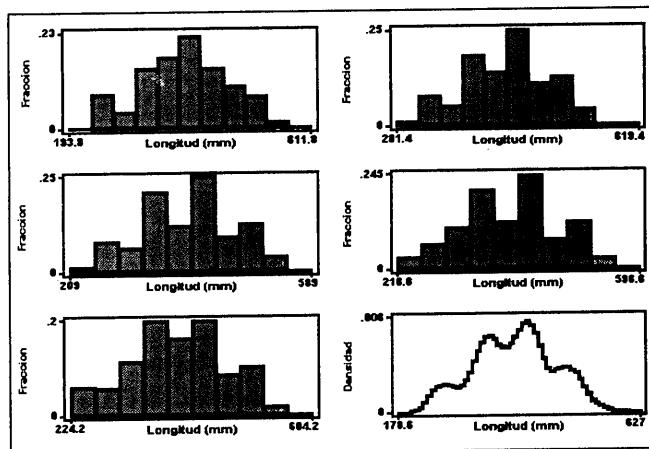


Figura 3: Cinco histogramas desplazados y su promedio.

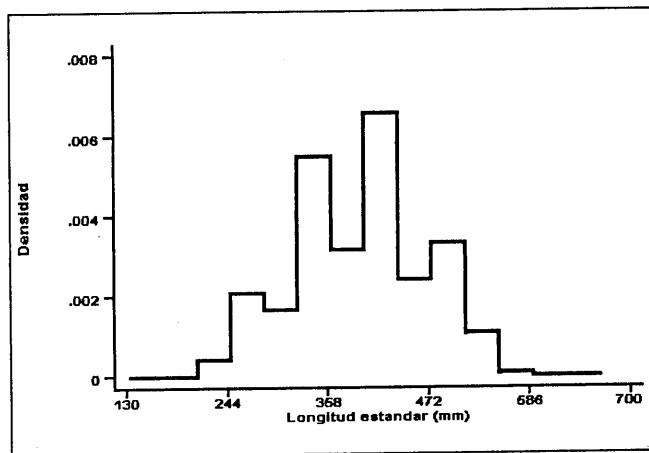


Figura 4: Histograma con la amplitud de banda óptima gaussiana ( $h = 38$ ) y origen en 133

La teoría estadística proporciona ciertas guías en la selección de la amplitud óptima de banda. Desafortunadamente, por lo general no es posible optimizar la amplitud de

banda sin el conocimiento previo de la forma verdadera de la densidad. Siguiendo las ideas de Tukey (1977), Scott (1979) y Silverman (1978, 1986), la distribución gaussiana puede utilizarse como un estándar de referencia en la elección de  $h$ . Aplicando un kernel gaussiano y minimizando el error cuadrado integrado medio (MISE por sus siglas en Inglés), se puede calcular el siguiente valor de escala (variabilidad):

$$s = \min \left[ \left( \frac{\sum(x_i - \bar{x})^2}{n - 1} \right)^{1/2}, \frac{H \text{ dispersión}}{1.349} \right]. \quad (2)$$

Entonces  $h$  puede escogerse como:

$$h = \frac{0.9s}{n^{1/5}}, \quad (3)$$

donde  $s$  es la menor de dos estimaciones del parámetro de variabilidad (escala) de la distribución gaussiana; la desviación típica y la robusta Pseudosigma, basada en la dispersión de los cuartos (Hoaglin, 1983; Fox, 1990). Este ajuste proporciona resistencia a colas potentes y trabaja bien para una gama amplia de densidades pero tiende a sobresuavizar distribuciones fuertemente sesgadas o multimodales (Silverman, 1986). Si tal es el caso, la amplitud de banda “óptima” (3) puede considerarse como un punto de partida para afinación posterior.

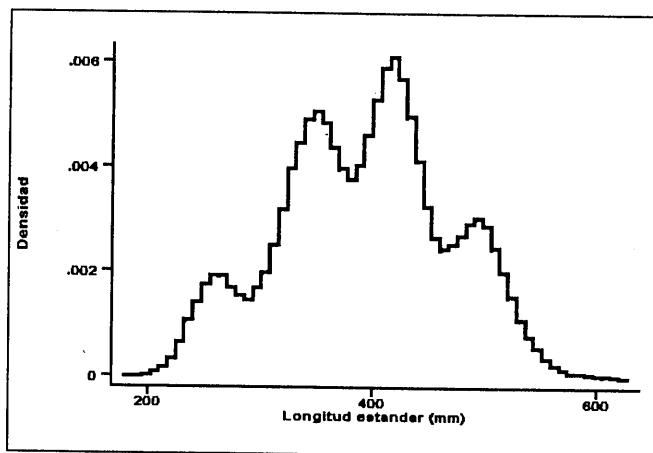


Figura 5: Promedio de cinco histogramas desplazados con intervalo óptimo gaussiano ( $h = 38$ )

Un inconveniente presentado por los EDK's es el gran número de operaciones requeridas para su cálculo. Silverman (1986) propuso el uso de la transformación de Fourier para discretizar los cálculos. Scott (1985) sugirió un procedimiento alternativo para resolver este problema: al tratar de eliminar la influencia del origen, propuso promediar varios histogramas con diferentes orígenes en lugar de elegir uno de ellos. Este es el histograma desplazado promedio ASH (Averaged Shifted Histograms) el cual, asintóticamente es equivalente a un EDK cuando el número de histogramas es muy grande. Posteriormente, Härdle y

Scott (1988) desarrollaron la idea general del promedio ponderado de puntos redondeados **WARP** (**Weighted Averaging of Rounded Points**).

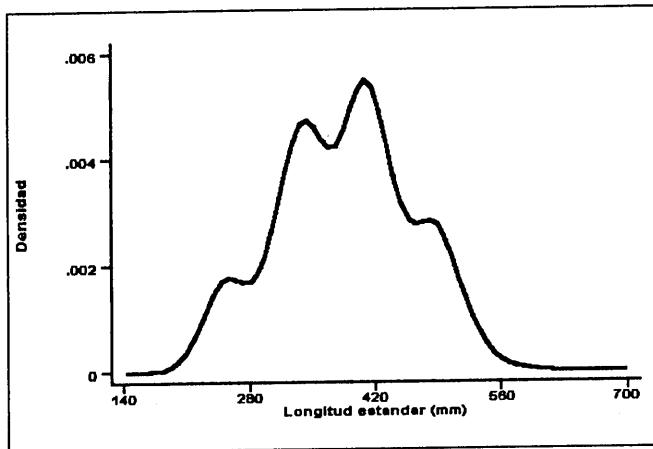


Figura 6: Estimador de densidad por kernel gaussiano con banda óptima gaussiana ( $h = 20$ )

El cálculo de una estimación **ASH-WARP** requiere de tres pasos: 1) agrupación de datos; 2) cálculo de pesos de acuerdo a una función kernel y 3) ponderación de intervalos. La última gráfica de la Figura 3 y la Figura 5 despliegan el resultado de promediar los cinco histogramas desplazados mostrados; la existencia de cuatro modas en los datos es evidente como resultado de la mejora significativa en la relación señal-ruido obtenida por el promedio del origen. El método WARP puede utilizarse para aproximar un **EDK** particular por medio de la selección de la función ponderal adecuada. El procedimiento **WARP** se aproxima a la estimación de densidad por kernel al aumentar el número de histogramas desplazados y utilizar la interpolación de puntos en lugar de una función de pasos (Härdle, 1991). Para el cálculo de **EDK**'s se pondrá la amplitud con pesos dependientes de una estimación preliminar de la densidad (para detalles consultar Salgado-Ugarte *et al.*, 1993).

## 2 Material y Métodos

Para ilustrar el uso de los **EDK**'s presentados anteriormente se utilizaron los datos de longitud de la trucha coralina *Plectropomus leopardus* adaptados del reporte de Goeden (1978). Un enfoque no paramétrico más elaborado para evaluación de multimodalidad utilizando un conjunto de datos diferente se presenta en Salgado-Ugarte *et al.* (1997). Para calcular los **EDK**'s y las reglas de amplitud de banda se emplearon los programas presentados en Salgado-Ugarte *et al.* (1993, 1995a, 1995b). El **EDK** variable se utilizó en conjunto con el método de Bhattacharya para ejemplificar la utilidad de las estimaciones suaves de densidad en la identificación y caracterización de componentes en distribuciones mezcladas. Para ejecutar estos cálculos se emplearon los programas introducidos en Salgado-Ugarte

*et al.* (1994). Utilizamos como punto de partida la amplitud de banda óptima para kerne gaussiano (20), valor que se fue disminuyendo hasta escoger finalmente una  $h = 5$ .

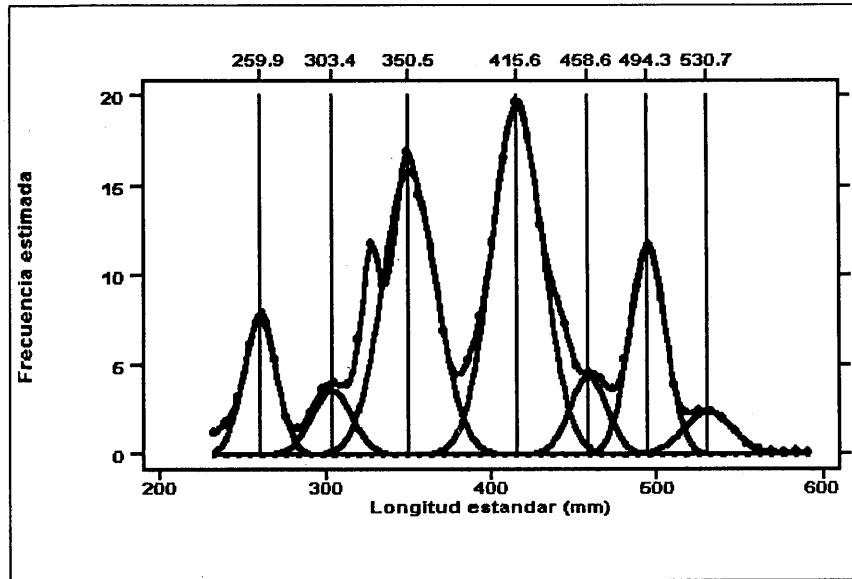


Figura 7: Estimador de densidad por kernel gaussiano de amplitud variable ( $h = 5$ )

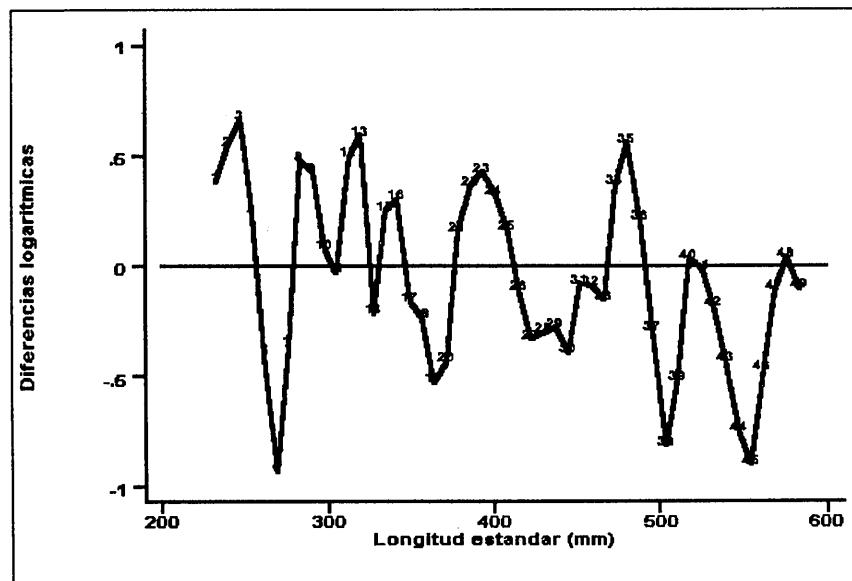


Figura 8: Gráfico de Bhattacharya para frecuencia calculada con EDK de amplitud variable

### 3 Resultados

El histograma empleando la amplitud óptima para distribución gaussiana (Scott, 1979)  $h = 38$  y origen en 133 se presenta en la Figura 4. La estimación sugiere la multimodalidad de los datos pero no proporciona suficiente detalle. La estimación resultante del promedio de cinco histogramas desplazados revela claramente la existencia de al menos cuatro modas aun cuando se utilizó la misma amplitud de intervalo y el origen de los mismos deja de ser un parámetro.

La estimación de densidad por kernel gaussiano utilizando la amplitud de banda propuesta por Silverman (1986)  $h = 20$  produce una distribución suave que muestra cuatro modas un tanto sobresuavizadas. La estimación de densidad por kernel de amplitud variable utilizando un valor medio geométrico de 5 se incluye en la Figura 7 que incluyen los componentes gaussianos estimados por el método de Bhattacharya.

### 4 Discusión y Conclusiones

Los estimadores de densidad por kernel resuelven algunos de los problemas del histograma y son un procedimiento adecuado para el análisis de datos de frecuencia de longitud. La labor excesiva para el cálculo de los EDK's puede evitarse por medio del empleo de procedimientos computacionalmente eficientes como el **ASH-WARP**. Al utilizar EDK's el problema de la elección del parámetro de suavización (amplitud de banda) persiste. Sin embargo, existen varias guías. La regla para la selección de la amplitud de banda óptima para el EDK gaussiano introducida se ha desarrollado para el caso de una sola distribución unimodal y simétrica, por lo que sobresuaviza a datos con más de una moda (Figura 6). En el caso de distribuciones mezcladas, existen varios componentes (gaussianos o de otro tipo) cada uno con parámetros diferentes (media y desviación estándar si son distribuciones gaussianas). La amplitud óptima puede ser diferente para cada componente. Los grupos dominantes (aquellos con mayor frecuencia) permiten el uso de un gran número de intervalos de amplitud pequeña; componentes con un número escaso de individuos pueden soportar sólo unos cuantos intervalos relativamente amplios. El histograma clásico utiliza una amplitud fija de intervalo; por tanto, puede desempeñar un papel pobre en representar tanto a los grupos dominantes como a los escasos. El uso del estimador de densidad por kernel de amplitud variable además de no depender de la posición del origen, ajusta la amplitud de la banda de acuerdo al numero de observaciones y por tanto revela mayor detalle en la separación de las modas al compararse con el EDK de amplitud fija. El gráfico de Bhattacharya correspondiente a esta estimación (Figura 8) expone segmentos lineales distintivos de pendiente negativa, los cuales no se distinguen al emplear el método con histogramas de agrupación mínima (Figura 9). Los parámetros estimados para cada componente pueden ser utilizados como valores iniciales en una estimación de máxima verosimilitud posterior (Akamine, 1985; Macdonald y Green, 1988; Fournier *et al.*, 1990).

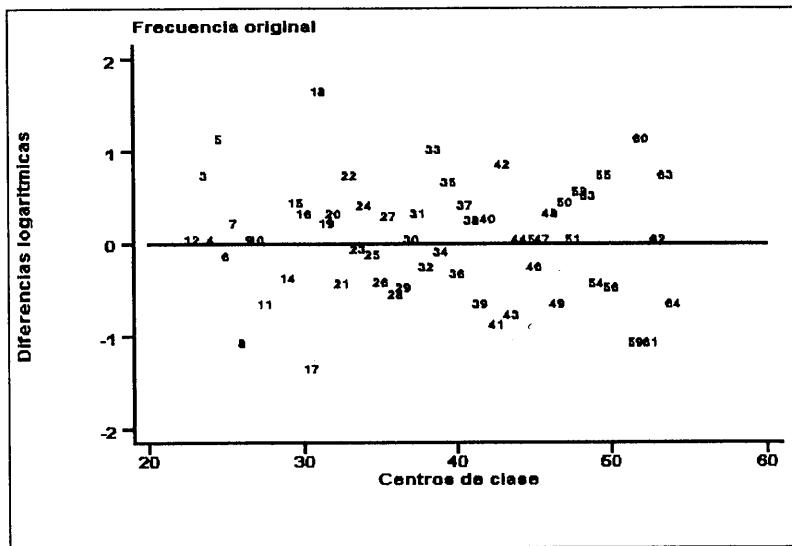


Figura 9: Gráfico de Bhattacharya para frecuencia original

Las estimaciones de densidad por kernel proporcionan varias formas para probar y evaluar la multimodalidad (detalles se incluyen en Silverman 1981b, 1983 y 1986). Un ejemplo de utilización de estos métodos más elaborados se presenta en Salgado-Ugarte *et al.* (1997) que hace uso de una versión automatizada. Otros enfoques han sido sugeridos por Cox (1966), Good y Gaskins (1980) y Wong (1985).

## Referencias

- Chambers, J.M., Cleveland, W.S., Kleiner, B. y Tukey, P.A. (1983). *Graphical methods for data analysis*. Belmont: Wadsworth.
- Cox, D.R. (1996). Notes on the analysis of mixed frequency distributions. *The British Journal of Mathematical and Statistical Psychology*, 19, 39-47.
- Epanechnikov, V.A. (1969). Nonparametric estimation of a multidimensional probability density. *Theor. Probab. Appl.*, 14, 153-158.
- Fox, J. (1990). Describing univariate distributions. In: *Modern Methods of Data Analysis*, (eds. J. Fox y J.S. Long) 58-125. Newbury Park: Sage Publications.
- Goeden, G.B. (1978). A monograph of the coral trout, *Plectropomus leopardus* (Lacépde). *Res. Bull. Fish. Serv. Queensl*, 1, 42.
- Good, I.J. y Gaskins, R.A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *Journal of the*

- American Statistical Association*, 75, 42-73.
- Härdle, W. (1991). *Smoothing Techniques. With Implementations in S*. New York: Springer-Verlag.
- Härdle, W. y Scott, D.W. (1988). Smoothing in low and high dimensions by weighted averaging using rounded points. *Technical report 88-16*, Rice University.
- Jones, M.C. (1990). Variable kernel density estimates and variable kernel density estimates. *Australian Journal of Statistics*, 32, 3.
- Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* 27, 832-837.
- Salgado-Ugarte, I.H., Shimizu, M. y Taniuchi, T. (1993). Exploring the shape of univariate data using kernel density estimators. *Stata Technical Bulletin* 16, 8-19.
- Salgado-Ugarte, I.H., Shimizu, M. y Taniuchi, T. (1994). Semi-graphical determination of Gaussian components in mixed distributions. *Stata Technical Bulletin* 18, 15-27.
- Salgado-Ugarte, I.H., Shimizu, M. y Taniuchi, T. (1995a). ASH, WARPing, and kernel density estimation for univariate data. *Stata Technical Bulletin* 26, 2-10.
- Salgado-Ugarte, I.H., Shimizu, M. y Taniuchi, T. (1995b). Practical rules for bandwidth selection in univariate density estimation. *Stata Technical Bulletin* 27, 5-19.
- Salgado-Ugarte, I.H., Shimizu, M. y Taniuchi, T. (1997). Nonparametric assessment of multimodality for univariate data. *Stata Technical Bulletin* 38, 27-35.
- Scott, D.W. (1979). On optimal and data-based histograms. *Biometrika*, 66, 605-610.
- Scott, D.W. (1985). Averaged shifted histograms: effective nonparametric density estimators in several dimensions. *Annals of Statistics*, 13, 1024-1040.
- Scott, D.W. (1992). *Multivariate density estimation: Theory, Practice, and Visualization*. New York: Wiley.
- Silverman, B.W. (1978). Chosing the windth when estimating a density. *Biometrika*, 65, 1-11.
- Silverman, B.W. (1981a). Density estimation for univariate an bivariate data. In *Interpreting Multivariate Data*, (ed. V. Barnett) 37-53, Chichester: Wiley.
- Silverman, B.W. (1981b). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society, B*, 43, 97-99.

Silverman, B.W. (1986). Density estimation for statistics and data analysis. London: Chapman & Hall.

Tarter, M.E. y Kronmal, R.A. (1976). An introduction to the implementation and theory of nonparametric density estimation. *The American Statistician*, 30, 105-112.

Tukey, J.W. (1977). *Exploratory data analysis*. Reading: Addison-Wesley.

Wong, M.A. (1985). A bootstrap testing procedure for investigating the number of sub-populations. *Journal of Statistical Computation and Simulation*. 22, 99-112.

# Sobre la Identificación de Observaciones Influyentes en el Análisis de Supervivencia

Belem Trejo Valdivia

*IIMAS-UNAM*

## 1 Modelo de Riesgos Proporcionales

Este modelo establece que el efecto de un conjunto de covariables sobre el comportamiento del tiempo de supervivencia  $T$  puede describirse a través de la función de riesgo de la siguiente manera:

$$\lambda(t|z) = \lambda_0(t) \exp(\beta' z),$$

en donde  $\lambda_0(t)$  es la función de riesgo básica (asociada a los individuos con  $z = 0$ ) y  $\beta$  es un vector de parámetros.

Dada una muestra (posiblemente censurada) de  $n$  tiempos de supervivencia, la estimación de  $\beta$  se lleva a cabo por máxima-verosimilitud parcial, es decir a partir de

$$I^*(\beta) = \log L^*(\beta) = \sum_{i=1}^n \delta_i \cdot (\beta' z_i - \log(\sum_{j \in R_i} \exp(\beta' z_j)))$$

en donde  $\delta_i$  y  $z_i$  son respectivamente el indicador de no-censura y el vector de covariables para el  $i$ -ésimo individuo;  $R_i$  es el conjunto en riesgo justo antes de  $t_i$ . Dado el estimador  $\hat{\beta}$ , para estimar  $\lambda_0(t)$  comúnmente se utiliza el estimador de Nelson-Aalen.

## 2 Métodos para el Diagnóstico de Observaciones Influyentes sobre $\beta$

En cualquier ajuste estadístico es de interés el analizar la influencia de las observaciones sobre el comportamiento de las estimaciones realizadas. En el caso de un modelo de riesgos proporcionales bastaría llevar a cabo dicho análisis para el vector  $\beta$  ya que la estimación de  $\lambda_0(t)$  y de cualquier otra función importante (riesgo relativo, vida media residual, etc.) son funciones del anterior.

Para evaluar el peso de una o de varias observaciones (en conjunto) sobre el estimador bajo estudio, es común tomar como punto de referencia el cambio que sufriría dicho estimador cuando se ignora esa o esas observaciones en un análisis similar, esto es, se analiza la magnitud del cambio  $\hat{\beta}$  (muestra completa) -  $\hat{\beta}$  (muestra reducida).

Un fuerte problema en el caso del modelo de riesgos proporcionales es que no existe una forma cerrada del estimador de  $\beta$ , por lo que no se tiene una forma general y directa de calcular  $\hat{\beta} - \hat{\beta}_{-1}$ , es decir, el peso de la  $i$ -ésima observación (censurada o no). Una forma de resolver este problema es el dar buenas aproximaciones para dicha diferencia. En la literatura se han propuesto entre otros los siguientes métodos de aproximación.

## 2.1 Método de Cain-Lange

Este método da esencialmente una aproximación en series de Taylor a primer orden bajo el siguiente contexto. Consideremos que cada una de las observaciones de la muestra tiene un peso igual a 1 excepto la  $i$ -ésima cuyo peso es  $W_i$  por lo que el estimador puede ser visto como una función de  $W_i$ . De lo anterior se tiene que  $\hat{\beta} = \hat{\beta}(W_i = 1)$  y  $\hat{\beta}_{-1} = \hat{\beta}(W_i = 0)$ . Entonces

$$\hat{\beta} - \hat{\beta}_{-1} \approx \frac{\partial \hat{\beta}}{\partial W_i} = \left( -\frac{\partial U(\beta)}{\partial \partial \beta} \right)^{-1} \times \frac{\partial U(\beta)}{\partial W_i} = I^{-1}(\beta) \times \frac{\partial U(\beta)}{\partial W_i} \text{ en } \beta = \hat{\beta} \text{ y } W_i = 1,$$

en donde  $I(\hat{\beta})$  es la matriz de información observada y el vector de puntajes está dado por

$$U(\beta) = \sum_{j=1}^n \delta_j \cdot W_j \{z_j - a_j(W_i)\} \text{ con } a_j(W_i) = \frac{\sum_{k \in R_j} W_k z_k \exp(\beta' z_k)}{\sum_{k \in R_j} W_k \exp(\beta' z_k)}.$$

Hay que notar que con esta aproximación es posible evaluar por separado el efecto de *cada observación* sobre *cada uno* de los componentes del vector  $\hat{\beta}$ .

## 2.2 Método de Rei-Crépeau

Este método se basa directamente en el concepto de función de influencia de un estimador y considera que en este caso se trabaja con la función de verosimilitud parcial. La función de influencia empírica asociada a la  $i$ -ésima observación está dada por el vector,

$$IE(i) = I^{-1}(\hat{\beta}) \cdot \left( \delta_i \cdot \left[ z_i - \frac{\sum_{j \in R_i} z_j \exp(\hat{\beta}' z_j)}{\sum_{j \in R_i} \exp(\hat{\beta}' z_j)} \right] + C_i(\hat{\beta}) \right),$$

$$\text{con } C_i(\hat{\beta}) = \left[ \sum_{t_j \leq t_i} \delta_j \cdot \frac{\sum_{k \in R_j} z_k \exp(\hat{\beta}' z_k)}{\left( \sum_{k \in R_j} \exp(\hat{\beta}' z_k) \right)^2} - z_i \sum_{k \in R_i} \exp(\hat{\beta}' z_k) \right] \exp(\hat{\beta}' z_i).$$

Como en el método anterior, esta expresión permite obtener una medida de la influencia de *cada observación* sobre *cada uno* de los componentes del estimador  $\hat{\beta}$ . En la práctica se ha encontrado que el valor de las estadísticas de diagnóstico para estos dos primeros métodos resultan muy similares numéricamente, por lo que en general, basta hacer el análisis para una sola de estas estadísticas.

### 2.3 Método de Pettitt-Bin Daud

Este método tiene como objetivo dar una medida del impacto que tiene una particular observación sobre el ajuste global realizado. Se propone tomar como referencia la diferencia  $DL_i = 2\{l^*(\hat{\beta}) - l^*(\hat{\beta}_{-i})\}$ , la cual puede a su vez ser aproximada por la forma cuadrática,

$$d'_i \cdot I^{-1}(\hat{\beta}) \cdot d_i, \text{ en donde } d_i = \delta_i \cdot (z_i - a_i(1)) + \exp(\hat{\beta}' z_i) \sum_{j|t_j < t_i} \left( \frac{\delta_j(a_j(1) - z_j)}{\sum_{k \in R_i} \exp(\hat{\beta}' z_k)} \right).$$

En cualquier método, observaciones que tengan asociadas valores relativamente grandes de la estadística de diagnóstico son considerados influyentes.

### 2.4 Ejemplo. Infección en pacientes con diálisis

Un problema que puede ocurrir en pacientes con padecimientos renales que reciben un tratamiento de diálisis es la ocurrencia de una infección en el sitio en donde el catéter es insertado. Para estudiar la incidencia de este tipo de infecciones y su posible relación con edad y sexo, se tomó una muestra de pacientes con problemas renales y en cada uno de ellos se midió el tiempo desde la inserción del catéter hasta la detección de la infección y remoción del mismo. Por simplicidad se presenta el análisis para el subgrupo de 13 individuos con padecimientos renales tipo C. Los datos son:

**TABLA 1**  
Tiempo (en días) de remoción de catéter

Paciente	Tiempo	Status	Edad	Sexo	Paciente	Tiempo	Status	Edad	Sexo
1	8	1(no cens.)	28	1(M)	8	141	1	34	2
2	15	1	44	2(F)	9	185	1	60	2
3	22	1	32	1	10	292	1	43	2
4	24	1	16	2	11	402	1	30	2
5	30	1	10	1	12	447	1	31	2
6	54	0(cens.)	42	2	13	536	1	17	2
7	119	1	22	2					

El modelo de riesgos proporcionales ajustado resulta ser:

$$\hat{\lambda}(t|z) = \hat{\lambda}_0(t) \cdot \exp(0.0304edad - 2.7108sexo), \text{ con } l^*(\hat{\beta}) = 1.8687.$$

En la Tabla 2 se presentan las diferencias exactas (calculadas a partir de los 13 ajustes adicionales que resultan de eliminar una observación a la vez) y las diferencias aproximadas (calculadas por el método de Cain-Lange) entre las estimaciones de cada una de las dos componentes del vector  $\beta$ . Como puede verse, existe una gran similitud entre los valores, por lo que se tiene un buen método de aproximación.

Para la variable edad, la mayor influencia se registra para la observación 13, su omisión producirá entre otras cosas una sobreestimación del riesgo relativo aunque resulta poco significativa. Para la variable sexo, las 2 observaciones con mayor influencia son la 2 y la 4. La omisión de cualquiera de estas observaciones producirá un incremento en la estimación del riesgo relativo levemente significativo. Se puede observar la gran similitud entre los valores aproximados y exactos de las diferencias, aunque existe una tendencia a la subestimación.

**TABLA 2**

Diferencias exactas (exact.) y aproximadas (aprox.) para  $\beta$  ( $\beta_1$  : edad,  $\beta_2$  : sexo)

Obs.	exact- $\beta_1$	aprox- $\beta_1$	exact- $\beta_2$	aprox- $\beta_2$
1	0.0031	0.0020	-0.3252	-0.1977
2	-0.0007	0.0004	0.8196	<b>0.5433</b>
3	-0.0016	-0.0011	0.1135	0.0741
4	-0.0143	-0.0119	0.8183	<b>0.5943</b>
5	0.0084	0.0049	0.1236	0.0139
6	-0.0006	-0.0005	-0.1276	-0.1192
7	-0.0126	-0.0095	0.2172	0.1270
8	-0.0042	-0.0032	-0.0152	-0.0346
9	-0.0174	-0.0073	0.0092	-0.0734
10	0.0051	0.0032	-0.2454	-0.2023
11	0.0072	0.0060	-0.2490	-0.2158
12	0.0069	0.0048	-0.2445	-0.1939
13	0.0195	<b>0.0122</b>	-0.4771	-0.3157

La Tabla 3 muestra los valores exactos y aproximados de  $DL_i$  para cada uno de los pacientes. Con estos valores podemos analizar el impacto global que tiene cada observación en el ajuste del modelo de riesgos proporcionales con edad y sexo conjuntamente.

**TABLA 3**

Valores exactos y aproximados de  $DL_i$

Paciente	Exacto	Aprox.	Paciente	Exacto	Aprox.
1	0.097	0.033	8	0.078	0.027
2	<b>0.480</b>	<b>0.339</b>	9	0.056	0.133
3	0.015	0.005	10	0.028	0.035
4	<b>0.404</b>	<b>0.338</b>	11	0.034	0.061
5	0.041	0.050	12	0.033	0.043
6	0.012	0.019	13	<b>0.608</b>	<b>0.219</b>
7	0.127	0.136			

Las observaciones que más afectan el valor de la log-verosimilitud cuando se omiten son aquéllas correspondientes a los pacientes 2 y 4. También para el paciente 13 se registra un valor grande de esta estadística de diagnóstico. Hay que notar que estos resultados son compatibles con los obtenidos de la Tabla 2.

En resumen, las observaciones de los pacientes 2,4 y 13 afectan la forma de la función de riesgo significativamente. Al omitir estas observaciones (una a la vez), se obtienen los siguientes estimadores del componente lineal en las funciones de riesgo:

Omitiendo paciente 2:	$0.0311 \text{ edad} - 3.5304\text{sexo}$
Omitiendo paciente 4:	$0.0447 \text{ edad} - 3.5291\text{sexo}$
Omitiendo paciente 13:	$0.0109 \text{ edad} - 2.2337\text{sexo}$

Para ilustrar el impacto en el cociente de riesgos, consideremos el riesgo relativo de infección al tiempo  $t$  para un paciente de edad 50 relativo a uno de edad 40 del mismo sexo. Para el conjunto completo de pacientes, el riesgo relativo es  $\exp(0.304)=1.355$ . Cuando se omiten los pacientes 2 y 4, este valor se incrementa a  $1.365$  y  $1.564$  respectivamente y decrece a  $1.14$  cuando se omite el paciente 13. El efecto sobre la función de riesgo al remover estos pacientes del análisis no es particularmente marcado. De manera similar, el riesgo de infección al tiempo  $t$  para un paciente masculino relativo a un paciente femenino de la misma edad es  $\exp(2.711)=5.041$  para el conjunto completo de datos. Cuando las observaciones 2,4 y 13 se omiten una a la vez, el riesgo relativo es  $4.138$ ,  $4.097$  y  $9.334$  respectivamente. La omisión del paciente 13 parece tener un gran efecto sobre la estimación del cociente de riesgos.

### 3 Tratamiento de Observaciones Influyentes

Es difícil dar una solución general al que hacer con una observación altamente influyente. Depende fuertemente del marco científico del estudio. Cuando sea posible, es recomendable verificar el origen de estas observaciones. En muchas situaciones no será posible confirmar que los análisis correspondientes a una observación influyente son válidos, de hecho no es recomendable su omisión de manera inmediata. En estas circunstancias, lo más apropiado sería establecer su efecto real sobre las inferencias que se obtengan. Por ejemplo, si el riesgo relativo o la vida mediana son utilizados para cuantificar el efecto de un tratamiento, los valores de estas estadísticas con y sin los valores influyentes deben compararse. Si la diferencia entre ellos no es pequeña desde un punto de vista práctico, dichas observaciones pueden incluirse. Por otro lado, si el efecto al omitirlas es grande, deben de considerarse de manera conjunta los resultados de los análisis con y sin ellas.

## Referencias

- Cain, K.C. and Kange, N.T. (1984). Approximate Case Influence for the Proportional Hazards Regression Model with Censored Data. *Biometrics*, 40, 493-499.
- McGilchrist, C.A. and Aisbert, C.W. (1991). Regression with Frailty in Survival Analysis. *Biometrics*, 47, 461-466.
- Pettitt, A.N. and Bin Daud, I. (1989). Case-weighted Measures of Influence for Proportional Hazards Regression. *Applied Statistics*, 38, 51-67.
- Reid, N. and Crépreau, H. (1985). Influence Functions for Proportional Hazards Regression. *Biometrika*, 72, 1-9.
- Storer, B.E. and Crowley, J. (1985). A Diagnostic for Cox Regression and General Conditional Likelihoods. *Journal of the American Statistical Association*, 80, 139-147.
- Weissfeld, L.A. (1990). Influence Diagnostics for the Proportional Hazards Model. *Statistics and Probability Letters*, 10, 411-417.

# Co-Integración en Series de Tiempo

Alfredo Troncoso V. y Alejandro Alegría H.

*Dept. de Estadística      Dept. de Estadística  
ACNielsen, México            ITAM*

## 1 Co-Integración

Actualmente el estudio del concepto de co-integración y el problema de raíces unitarias constituyen una gran parte de la agenda de la investigación econométrica. La literatura sobre co-integración ha modificado significativamente la forma en que los economistas modelan relaciones económicas dinámicas.

La búsqueda de relaciones de equilibrio entre series de tiempo, utilizando el modelo de regresión clásico, puede dar lugar a lo que Granger y Newbold en 1974 denominaron como regresión espuria. Al analizar series integradas de orden (d) no se mantienen las propiedades estadísticas usuales para el primer y segundo momento muestrales.

La idea entonces, es describir relaciones económicas de equilibrio mediante combinaciones lineales que resultan más estables que las variables originales. Esta relación de equilibrio es equivalente a la estacionariedad de la serie resultante de la combinación lineal de series que no son estacionarias o son integradas de orden (d).

**Definición 1. CO-INTEGRACION (Engle y Granger, 1987).**

Los componentes de un vector  $X_t$  se dicen co-integrados de orden (d, b), lo cual se denota  $X_t \sim CI(d,b)$ , si

- Todos los componentes de  $X_t$  son  $I(d)$ , es decir, integrados de orden(d).
- Existe un vector distinto de cero,  $\alpha$ , tal que  $z_t = \alpha^t X_t \sim I(d-b)$ ,  $b>0$ . El vector  $\alpha$  es llamado el *vector co-integrante*. A la combinación lineal resultante entre las series de  $X_t$ , se le denomina *regresión co-integrante*.

Supondremos que existen exactamente r vectores co-integrantes linealmente independientes con  $r \leq N - 1$ , los cuales están en una matriz  $\Gamma$ . Entonces por construcción el rango de  $c\Gamma$ , donde  $c$  es una constante, será r, el cual se denomina *rango co-integrante* de  $X_t$ .

La idea de que en una relación de equilibrio las tendencias de las series no se pueden separar mucho en el largo plazo, se ha traducido en un enunciado más preciso: *su diferencia será  $I(d-b)$* .

Analizando el modelo de regresión, para que exista co-integración se requiere solamente que los movimientos en tendencias en la variable dependiente sean iguales a combinaciones lineales de movimientos similares en las variables independientes; no es necesario que los residuos sean puramente aleatorios, basta con que sean un proceso estacionario en general.

Las propiedades de los procesos co-integrados se resumen en el Teorema de Representación de Granger, el cual también da la pauta para determinar las pruebas y métodos de estimación de relaciones co-integrantes.

**Teorema 1.** *Teorema de Representación de Granger.*

Si el vector  $X_t$  de Nx1 es co-integrado (1,1) y con rango co-integrante r, entonces

- $C(1)$  es de rango  $N-r$ , donde  $C$  es la matriz resultante de la representación multivariada de Wold para  $X_t$ ,  $(1 - L)X_t = C(L)\varepsilon_t$ .
- Existe una representación de corrección de errores, con  $z_t$  estacionario, de la forma  $A^*(L)(1 - L)X_t = -\gamma z_{t-1} + d(L)\varepsilon_t$ .

Del teorema de representación de Granger se obtiene que una condición para que exista co-integración es que  $C(1)$  tenga rango reducido.

## 2 Pruebas de Co-Integración

Básicamente existen dos tipos de pruebas de co-integración, unas basadas en los residuos resultantes de la regresión co-integrante; y otras cuyo fundamento radica en el rango de la matriz  $C(1)$ .

Las pruebas basadas en los residuos de la regresión co-integrante, surgen directamente del Teorema de Representación de Granger, y son pruebas de raíces unitarias aplicadas a las series de residuos obtenidas con la regresión co-integrante  $y_t = \beta X_t + z_t$ .

Engle y Granger (1987) hacen una revisión de varias pruebas de raíces unitarias (CRDW, DF, ADF, RVAR, ARVAR, UVAR y AUVAR), y al final proponen a los estadísticos DF y ADF como los más recomendables para pruebas de co-integración.

### 2.1 Prueba Dickey-Fuller (DF).

Se estima por mínimos cuadrados ordinarios (MCO)  $\nabla \hat{z}_t = \gamma \hat{z}_{t-1} + u_t$ , donde  $\gamma = (\rho - 1)$ .

Si el valor del estadístico t para el valor estimado de  $\gamma$  es mayor que el valor de tablas (Mackinnon, 1991), existirá evidencia significativa para rechazar la *hipótesis nula de no co-integración* ( $\gamma = 0$ ).

## 2.2 Augmented Dickey-Fuller (ADF).

Esta prueba permite mayor dinámica en la prueba DF. Se puede sobreparametrizar en el caso de primer orden, pero se especifica correctamente en los casos de mayor orden.

En este caso se estima por MCO  $\nabla \hat{z}_t = \gamma \hat{z}_{t-1} + \sum \beta_i \nabla \hat{z}_{t-i} + u_t$ ,  $\gamma = \rho - 1$ .

Una característica importante es que el estadístico es invariante con respecto al orden de la autorregresión.

## 2.3 Pruebas de rango co-integrante

Dentro de las pruebas de rango co-integrante, la más conocida y utilizada es la prueba de Johansen (1988). Esta prueba se puede ver como la generalización multivariada de la prueba ADF,

$$\begin{aligned} y_t &= A_1 y_{t-1} + \varepsilon_t, \\ \implies \nabla y_t &= A_1 y_{t-1} - y_{t-1} + \varepsilon_t, \\ \implies \nabla y_t &= (A_1 - I) y_{t-1} + \varepsilon_t, \\ \implies \nabla y_t &= \Pi y_{t-1} + \varepsilon_t. \end{aligned}$$

El punto clave para esta prueba es que la matriz  $\Pi$ , o matriz de impacto, tiene rango  $r \leq n$ . Así la hipótesis nula para la prueba de *a lo más r vectores co-integrantes* es,  $H_0 : \Pi = \gamma \alpha^t$ . Debido a que la hipótesis de co-integración es equivalente a la hipótesis de rango reducido de la matriz  $\Pi$ , es entonces razonable calcular los eigenvalores de  $\widehat{\Pi}$  y verificar si éstos son cercanos a cero (Fountis y Dickey, 1989).

El estadístico de prueba basado en el cociente de verosimilitudes es el siguiente

$$Q_r = -N \sum \ln(1 - \widehat{\lambda}_i) \sim \chi^2.$$

## 3 Estimación de Relaciones Co-Integrantes

Actualmente se han desarrollado varios métodos de estimación para relaciones co-integrantes. A continuación se mencionan los más importantes dentro de la literatura de co-integración.

Engle y Granger (1987) propusieron un método de estimación en dos etapas. En la primera de ellas, los parámetros del vector co-integrante son estimados mediante una regresión en las variables. En la segunda etapa, se utilizan estas estimaciones para construir un modelo de corrección de errores. Una de las ventajas de este método es que ambas etapas solamente requieren de MCO y los resultados son consistentes para todos los parámetros; esto último se garantiza con el Teorema de Superconsistencia de Stock (1987).

Johansen en 1988 propone un estimador máximo verosímil que sustituye el procedimiento de dos etapas, y que además se puede aplicar en el caso de vectores co-integrantes múltiples.

Cuando existe co-integración se puede expresar a la matriz de impacto como  $\Pi = \gamma\alpha^t$ . El espacio generado por los renglones de  $\Pi$  es llamado el *espacio co-integrante*. La estimación de este espacio co-integrante se obtiene a partir del siguiente teorema

**Teorema 2.** (Johansen, 1991).

El estimador máximo verosímil del espacio generado por  $\alpha$ , es el espacio generado por los  $r$  componentes canónicos correspondientes a los valores más grandes de las correlaciones canónicas al cuadrado entre los residuos de  $X_{t-k}$  y  $\nabla X_t$ .

## 4 Aplicación: Eficiencia del Mercado Cambiario en Francia Durante los Años 20's

A continuación se presenta un análisis de co-integración en series de tiempo en el que se pretende probar la eficiencia del mercado cambiario Francés durante los años veinte, época en la que se cambia del patrón oro al esquema de libre flotación. Los datos fueron recopilados por Patrick C. McMahon de ediciones pasadas del *Manchester Guardian* y consisten en observaciones diarias de las tasas de cambio *spot* y *forward* a un mes en un periodo que comprende del 1º de mayo de 1922 al 30 de mayo de 1925, expresadas en términos de la libra esterlina. Las observaciones incluyen sábados, y por lo tanto cubren 6 días de la semana, dando un total de 966 observaciones de las series en total.

Para que el mercado cambiario sea eficiente se requiere que la tasa cambiaria *forward* pueda ser considerada como un estimador insesgado de la futura tasa *spot*. Es importante tomar en cuenta que las tasas *forward* observadas en el tiempo  $t$  son comparables con las tasas *spot* en el tiempo  $t + 26$  debido a que las tasas *forward* son a un mes, y como nuestras observaciones incluyen sábados, solamente le quitamos cuatro domingos al mes que en promedio lo tomaremos como de 30 días. En las figuras 1 y 2 se pueden ver las gráficas de las series de tasas *forward* y *spot* respectivamente.

La regresión co-integrante que se estima es  $S_{t+26} = C + \beta_1 F_{t,26}$ , las estimaciones de los parámetros son

$$C = 0.338549, \quad \beta_1 = 0.823221,$$

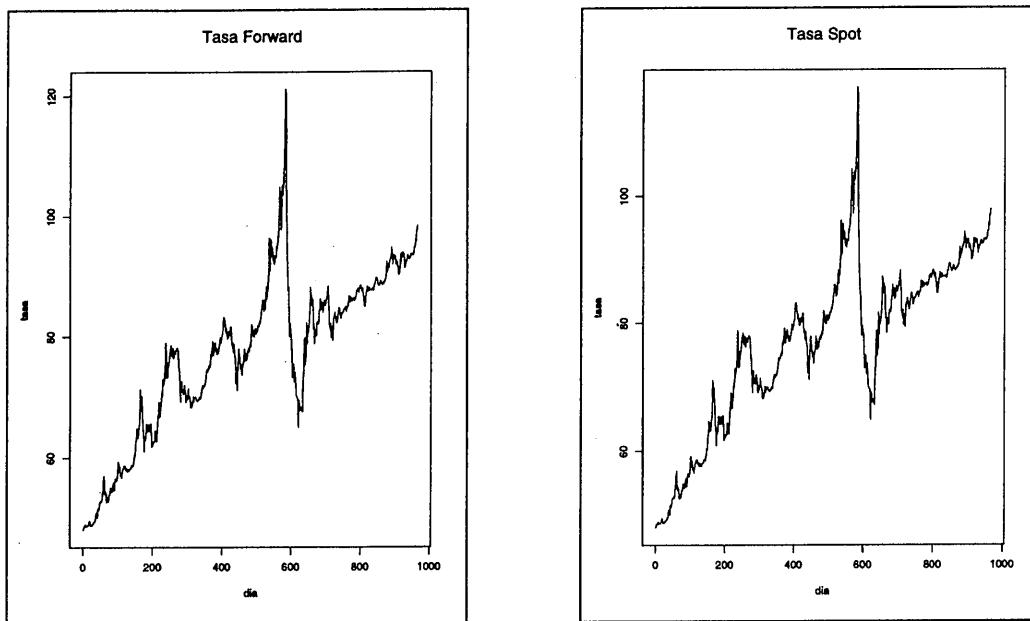
La prueba ADF sobre los residuos de la regresión co-integrante en los paquetes GAUSS y EViews da como resultado

$$\text{GAUSS: } -4.3673149, \text{ EVIEWS: } -4.367315,$$

valores para los cuales se rechaza la hipótesis nula de no estacionariedad. Por lo tanto las series de las tasas *forward* y *spot* de Francia están co-integradas y el vector co-integrante es

$(1, -0.823221)$ . De esto se deduce que la combinación lineal estacionaria entre las variables es  $Spot - 0.823221 * Forward = \varepsilon$ .

Esto quiere decir que en el largo plazo existe un equilibrio entre estas dos tasas, probando así la hipótesis de eficiencia del mercado cambiario Francés.



## 5 Conclusiones

Se ha presentado un concepto de mucha utilidad sobre todo en el campo de la econometría, ya que permite la identificación y estimación de combinaciones lineales de variables que describen relaciones de equilibrio en el largo plazo.

Con la teoría de co-integración muchos de los resultados econométricos serán más precisos al incorporar restricciones sobre raíces unitarias. La toma de decisiones en sectores económico-financieros tendrá mayor sustento y además se le permitirá al investigador buscar y plantear teorías sobre la relación y causalidad entre las variables co-integradas.

La no estacionariedad de la mayoría de las series de tiempo económicas y financieras tiene fuertes implicaciones en el modelaje, inferencia y pronóstico econométrico. Los econometristas por lo general trabajan con transformaciones de las variables originales, especialmente cuando se trata de procesos no estacionarios. El concepto de co-integración permite modelar, realizar inferencia y pronosticar con series de tiempo no estacionarias sin tener que transformar las series. Esta nueva teoría, co-integración, se ha desarrollado rápidamente y en la actualidad se cuenta con una gran variedad de pruebas de hipótesis y métodos de estimación que permiten una mejor especificación de modelos econométricos.

## Referencias

- Engle, R. F. y Granger, C.W.J. (1987). Co-integration and Error Correction: Representation, Estimation and Testing. *Econometrica*, 55, 251-276.
- Fountis, N. y Dickey, D.A. (1989). Testing for a Unit Root Nonstationary in the Multivariate Autoregressive Time Series. *Ann. Statist.*, 67, 419-426.
- Granger, C.W.J. y Newbold, P. (1974). Spurious Regressions in Econometrics. *J. Econometrics*, 2, 11-120.
- Johansen, S. (1991). Maximum Likelihood Estimation and Inference on Cointegration with application to the Demand of Money. *Oxford Bull. Econom. Statist.*, 52, 169-210.
- Johansen, S. (1998). Statistical Ananlysis of Cointegration Vectors. *J. Economic Dynamics and Control*, 12, 231-254.
- Mackinnon, J.G. (1991). Critical Values for Co-integration Tests. En: *Long Run Economic Relationships*. Oxford: University Press.
- Stock, J. H. (1987). Asymptotic Properties of the Least Squares Estimators of Cointegration Vectors. *Econometrica*, 55, 1035-1056.

# Estudio de Encuesta Sobre La Producción Industrial de Cereales

H. J. Vázquez

*Departamento de Sistemas,  
UAM-Azcapotzalco*

## 1 Introducción y Objetivos

A lo largo del proceso de integración de la Unión Europea (UE), desde sus tratados fundadores de 1957, se ha ido haciendo más importante la necesidad de centralizar la información estadística producida por los países integrantes. La internacionalización creciente y los acuerdos de intercambio comercial firmados por varios países de la Unión en 1994 obliga a la UE a alinear las estadísticas de producción con las estadísticas del comercio exterior. Esto último no ha hecho más que incrementar la urgencia por lograr este objetivo de centralización de información.

Sin embargo lograr esto no es una tarea sencilla ya que cada país cuenta con sistemas y métodos muy propios para medir y administrar su información, por lo que en 1991 la UE impuso a todos los países miembros la adopción del sistema europeo de estadísticas armonizadas de producción industrial (PRODCOM). PRODCOM, que significa producción comunitaria, solicita a cada país de realizar una encuesta anual para obtener estadísticas, sobre la producción industrial de las industrias de transformación y comercialización. Los resultados agregados deberán ser enviados a Eurostat dentro de los primeros seis meses después del año de referencia.

Este estudio presenta a grandes rasgos la experiencia de la implementación del proyecto de encuesta realizado en el ONIC (Oficina Nacional e Interprofesional de la Industria de Cereales) Francia. Presentar este trabajo en México es interesante ya que, además de formar parte de procesos de armonización de sistemas de información, trata algunos problemas claves que surgen en la industria de cereales: acceso a la información proveniente de diferentes fuentes, estructuración de las industrias (filiales, subcontratación, fusión), tratamiento e integración de resultados en sistemas estadísticos y de información heterogéneos. Dado que una gran parte de esta información es confidencial este trabajo se limita a una presentación de las principales características del proyecto de encuesta (reglamento CEE No 3924/91) y de los principios de tratamiento efectuados para resolver el tratamiento no respuestas.

## 2 Características de la Encuesta

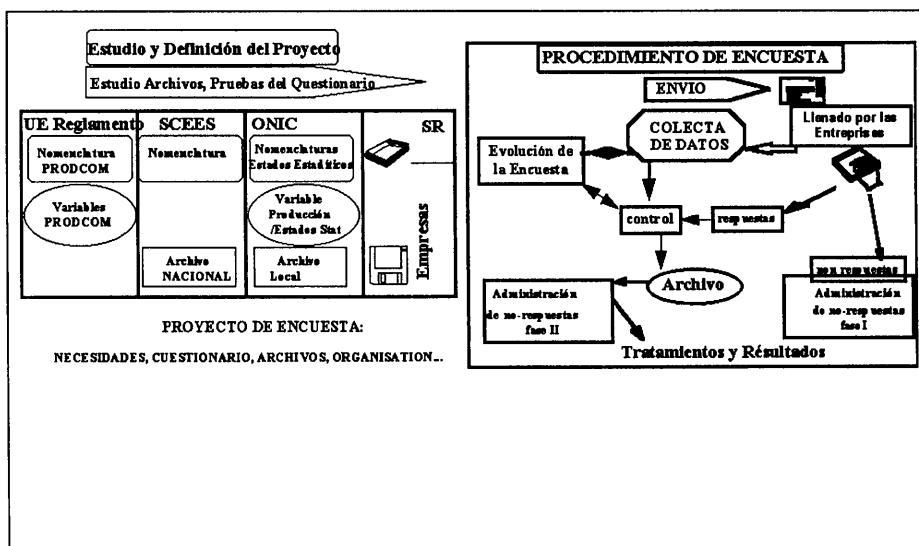
El reglamento CEE No 3924/91 establece las modalidades de aplicación y características de PRODCOM. Dentro de las principales características se indica que para que la comparación de las informaciones sea posible, los productos incluidos deben ser aquellos descritos al nivel de los primeros seis dígitos de la Nomenclatura Combinada (o sistema armonizado) basada en la NACE (clasificación nacional de actividades económicas, según las siglas en inglés). La información solicitada, (colectada mediante el uso de un cuestionario deberá representar al menos el 90% de la producción nacional de empresas con más de 20 empleados), es: volumen y valor de la producción nacional durante el año.

Los principales productos incluidos con estas características son: Cereales (maíz, trigo, arroz), harinas y productos derivados.

Sin embargo la aplicación estricta del reglamento propuesto por PRODCOM en esta rama no fue posible principalmente por las siguientes razones:

- El 30% de la producción de esta industria es realizado por empresas con un número de empleados inferior a 20.
- Dado el alto nivel de descentralización del sector no existe una institución con archivos completos de las empresas y con nomenclaturas precisas de los productos de esas ramas.
- Pocas empresas utilizan una nomenclatura completa y fácilmente comparable.
- Las instituciones con mayor información no contaban con sistemas de nomenclatura homogéneo para identificar a los productos.
- La obtención de resultados por muestreo con al menos el 90% de la producción necesitaba la implantación de un procedimiento de encuesta secuencial, que pareció costoso en dinero y en tiempo.

Figura 1



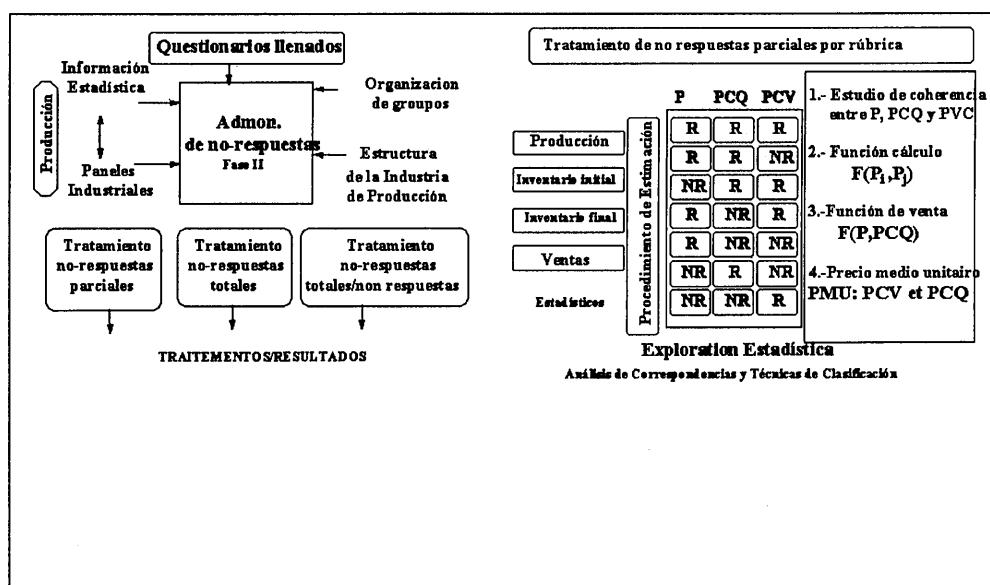
Por esto una vez tomada la decisión de aplicar PRODCOM se decidió integrar la información en un solo archivo integrando las diferentes nomenclaturas usadas por los organismos participantes) y encuestar al 100% de las industrias catalogadas.

La preparación de la encuesta con las siguientes características:

- Campo: Industria de Transformación de Cereales.
- Población: 100 % de la industria del ramo.
- Instrumento de observación :Cuestionario con guía de llenado enviado por vía postal, se inicia con la definición de los objetivos, la identificación de la unidad a encuestar, con la homogeneización de la nomenclatura, con la preparación del cuestionario y con la preparación de los archivos. Un esquema de las actividades realizadas en esta primera etapa se muestra en la Figura 1, izquierda. La institución encargada centralizar la información a partir de los principales archivos disponibles y de realizar la encuesta fue el ONIC. Los organismos participantes en forma conjunta aceptaron el plan presentado en la figura 1, derecha.

La ejecución de la encuesta, una vez informadas las empresas, inició con el envío de los cuestionarios y con la construcción de un sistema informático para el seguimiento de su evolución. Un archivo permite registrar el número de encuestas colectadas, encuestas sin respuesta, encuestas mal llenadas, empresas solicitando ayuda, etc. Una primera fase antes del registro en la computadora consiste en corregir las no respuestas. Maximizar el rendimiento de esta actividad de llenado es importante ya que se contaban, en esta primera implantación, con dificultades para evaluar la calidad de la información.

Figura 2



### 3 Tratamientos y Obtención de Resultados

Una primera fase de tratamientos permitió obtener las informaciones siguientes:

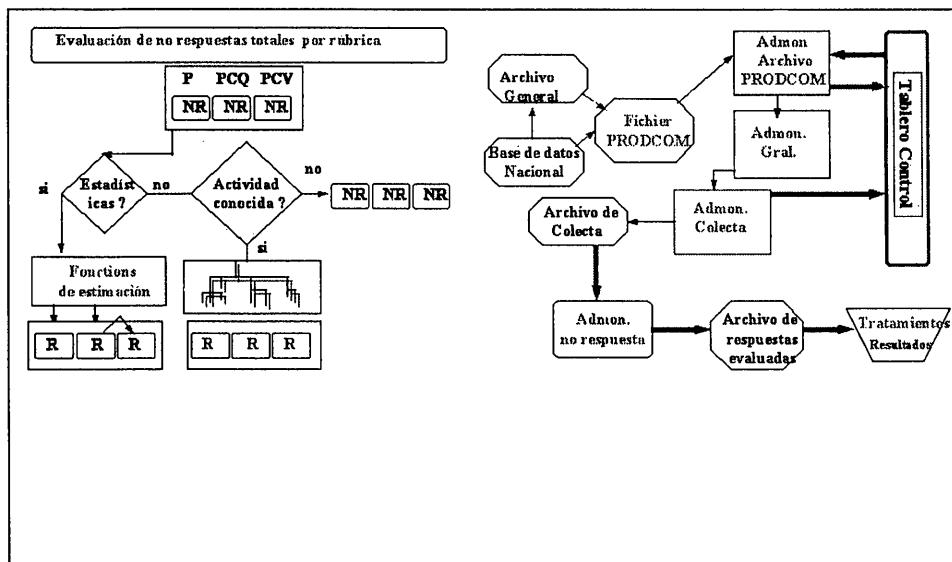
- número de respuestas (cuestionarios),
- número de respuestas por rúbrica de cuestionario,
- tipo (ausentes, parciales, ilegibles) y número de no respuestas por rúbrica
- características de las empresas
- volumen y valor de la producción total y por rúbrica.

Antes de proceder a una segunda fase de tratamientos, se procedió a tratar el problema clave de este estudio es decir el problema de no respuestas. Problema delicado dadas las restricciones establecidas por el reglamento.

Para su tratamiento fue necesario separar los diferentes tipos de no respuestas de acuerdo al tipo y proponer métodos de estimación para cada uno (Figura 2).

Por ejemplo para el caso de no respuestas parciales por rúbrica (codificadas NR), se construyeron funciones de estimación con base en los grupos que dieron respuesta (codificadas R). Grupos que compartían características comunes en otras rúbricas y que se obtuvieron mediante la aplicación de un estudio exploratorio usando Análisis de Correspondencias Múltiples y Técnicas de Clasificación (Lebart, *et al.*, 1984). Otro método de estimación de no respuestas se aplicó tomando en cuenta información estadística disponible por los organismos encargados del proyecto así como resultados de estadísticas industriales de comercio y de consumo. Este método se usó en casos de no respuestas totales (Figura 3, izquierda). El esquema general del tratamiento informático se presenta en la misma figura, a la derecha.

Figura 3



Una vez estimadas al máximo las no respuestas se procedió a la corrección de los resultados obtenidos en la primera fase, al cálculo de índices de producción y a la obtención de resultados por agrupamientos.

## 4 Conclusiones

Este proyecto es uno de los muchos proyectos estadísticos realizados en la UE en su etapa de armonización. Esta actividad es capital para la existencia de este tipo de asociaciones en una economía mundial ya que permiten la obtención de un gran número de informaciones necesarias para una mejor administración de los recursos.

Además la obtención de informaciones estadísticas mediante proyectos del tipo PROD-COM es muy importante ya que además de permitir la alineación de estadísticas, propone una clasificación clara de los productos y métodos comunes para la colecta de la información. Todo esto permite estudios más detallados de análisis de mercados y la obtención de indicadores precisos. Claro está, el inconveniente es que al inicio la cantidad y calidad de la información es pobre y necesita el paso de varios años para su mejoramiento.

A este respecto, para esta primera aplicación el esquema de tratamiento de no respuestas pareció adecuado y ha permitido estimar un buen número de no respuestas. Sin embargo es necesario mejorar este dispositivo con la integración de informaciones de encuestas posteriores que permitan construir algunos estimadores de tendencia.

El autor agradece la confianza y el apoyo otorgados por el ONIC, Francia para la realización de este proyecto.

## Referencias

Lebart, L. *et al.* (1984). *Multivariate Descriptive Analysis*. New York: Wiley.

Reglamento CEE No 3924/91, Eurostat, Luxemburgo.

# On Information Functionals and Priors

Francisco Venegas-Martínez

*Centro de Investigación y Docencia Económicas*

## 1 Introduction

Good (1968), Zellner (1971), and Bernardo (1979) have proposed several procedures to produce non-informative and informative priors. These methods, based on the maximization of a specific functional, have comparative and absolute advantages in some respects:

- (i) While Zellner's method is based on an exact finite sample criterion functional, Good's approach uses a limiting criterion, and Bernardo's procedure lies in asymptotic results.
- (ii) The criterion functional used by Bernardo is a cross-entropy which, in particular, is invariant with respect to one-to-one transformations of the parameters. In contrast, the total information functional employed by Zellner is invariant only for the location-scale family and under linear transformations of the parameters. To generate invariance under more general transformations, side conditions are needed.
- (iii) The way in which these methods have been tested is by seeing how well they perform in particular examples; the evaluation is often based on contrasting the derived priors with Jeffrey's (1961), usually improper, priors which are somewhat arbitrary and inconsistent.

In this work, we present within a unifying approach all inferential methods which by maximizing a criterion functional produce *non-informative* and *informative* priors. In our general framework, Minimax Evidence Priors (Good 1968), Maximal Data Information Priors (Zellner 1971), and Reference Priors (Bernardo 1979) are seen as special cases of maximizing a more general indexed criterion functional. In such a case, properties of the derived priors will depend on the choice of indexes from a wide range of possibilities, instead of on a few personal points of view with *ad hoc* modifications. In the spirit of Akaike (1978), this will look more like Mathematics than Psychology. This unifying approach will enable us to explore a vast range of possibilities for constructing priors.

## 2 A Family of Information Functionals

Suppose that we wish to make inferences about an unknown parameter  $\theta \in \Theta \subseteq \mathbb{R}$  of a distribution  $P_\theta$ , and there is available an observation  $X$ . Assume that  $P_\theta$  has density  $f(x|\theta)$

w.r.t. some  $\sigma$ -finite measure  $\lambda$  on  $\mathbb{R}$  for all  $\theta \in \Theta \subseteq \mathbb{R}$ , i.e.,  $dP_\theta/d\lambda = f(x|\theta)$  for all  $\theta \in \Theta \subseteq \mathbb{R}$ . We will also suppose that  $\pi(\theta)$  is a density w.r.t. some  $\sigma$ -finite measure  $\mu$  on  $\mathbb{R}$ . Once a prior  $\pi(\theta)$  has been prescribed, the information provided by the data,  $x$ , about the parameter is used to modify the initial knowledge, as expressed in  $\pi(\theta)$ , via Bayes' rule to obtain a posterior  $f(\theta|x) \propto f(x|\theta)\pi(\theta)$  (using  $f$  generically to represent densities).

Let us define an infinite system of nesting functionals (see Venegas-Martínez, 1998):

$$\mathcal{V}_{\gamma,\alpha,\delta}(\pi) = \frac{1}{1-\gamma} \int \pi(\theta) G(\mathcal{I}(\theta), \mathcal{F}(\theta), \gamma, \alpha, \delta) d\mu(\theta), \quad (2.1)$$

where

$$G(\mathcal{I}(\theta), \mathcal{F}(\theta), \gamma, \alpha, \delta) = \log \left\{ \frac{\exp\{[\mathcal{F}(\theta)/\mathcal{I}(\theta)]^{1-\delta} [\mathcal{I}(\theta)]^{\frac{1-\gamma}{1+\alpha}} - \delta[\mathcal{I}(\theta)]^{1-\alpha}\}}{\pi(\theta)^{1-\gamma}} \right\},$$

$0 \leq \gamma < 1$ ,  $\alpha \in \{0, 1\}$ ,  $\delta \in \{0, 1\}$ ,  $\mathcal{I}(\theta) = \int (\partial/\partial\theta \log f(x|\theta))^2 f(x|\theta)d\lambda(x)$  is Fisher's information about  $\theta$ , and  $\mathcal{F}(\theta) = \int f(x|\theta) \log f(x|\theta)d\lambda(x)$  is the negative Shannon's information of  $f(x|\theta)$ . Throughout this paper, we will be concerned with the following indexed family:

$$\mathcal{A} = \text{conv}[\overline{\{\mathcal{V}_{\gamma,\alpha,\delta}(\pi)\}}] = \text{convex hull of the closure of the family } \{\mathcal{V}_{\gamma,\alpha,\delta}(\pi)\}.$$

We readily identify a number of distinguished members of  $\mathcal{A}$ :

(i) Criterion for Maximum Entropy Priors (MAXENTP):

$$\mathcal{V}_{0,0,1}(\pi) = - \int \pi(\theta) \log \pi(\theta) d\mu(\theta).$$

(ii) Criterion for Minimax Evidence Priors (MEP):

$$\mathcal{V}_{1,1,1}(\pi) \stackrel{\text{def}}{=} \lim_{\gamma \rightarrow 1} \mathcal{V}_{\gamma,1,1}(\pi) = - \int \pi(\theta) \log \frac{\pi(\theta)}{p(\theta)} d\mu(\theta) - \log C, \quad (2.2)$$

which is Good's invariantized negative cross-entropy with initial density  $p(\theta) = C[\mathcal{I}(\theta)]^{\frac{1}{2}}$  with  $C = \{\int [\mathcal{I}(\theta)]^{\frac{1}{2}} d\mu(\theta)\}^{-1}$ , provided that  $\int [\mathcal{I}(\theta)]^{\frac{1}{2}} d\mu(\theta) < \infty$ . We can also write (2.2) as

$$\mathcal{V}_{1,1,1}(\pi) - \mathcal{V}_{0,0,1}(\pi) = \int \pi(\theta) \log [\mathcal{I}(\theta)]^{\frac{1}{2}} d\mu(\theta). \quad (2.3)$$

(iii) Criterion for Maximal Data Information Priors (MDIP):

$$\mathcal{V}_{0,0,0}(\pi) = \int \int f(x) f(\theta|x) \log \frac{\ell(\theta|x)}{\pi(\theta)} d\mu(\theta) d\lambda(x) \quad (2.4)$$

which is Zellner's criterion functional. Here, as usual,  $f(\theta|x) = f(x|\theta)\pi(\theta)/f(x)$ ,  $f(x) = \int f(x|\theta)\pi(\theta)d\mu(\theta)$ , and  $\ell(\theta|x) = f(x|\theta)$  is the likelihood function.

(iv) Criterion for Maximal Modified Data Information Priors (MMDIP):

$$\mathcal{V}_{0,1,0}(\pi) = \int \int f(x)f(\theta|x) \log \frac{[\ell(\theta|x)]^{[\mathcal{I}(\theta)]^{\frac{1}{2}}}}{\pi(\theta)} d\mu(\theta)d\lambda(x) \quad (2.5)$$

which is the prior average information in the data *modified* by Fisher's information minus the information in the prior. Note that when  $\mathcal{I}(\theta)$  is constant, (2.5) reduces to Zellner's criterion functional (up to a constant factor).

(v) Criterion for Maximal Fisher Information Priors (MFIP):

$$\mathcal{V}_{0,1,1}(\pi) = - \int \pi(\theta) \log \frac{\pi(\theta)}{\exp\{[\mathcal{I}(\theta)]^{\frac{1}{2}}\}} d\mu(\theta) - 1 \quad (2.6)$$

which is the prior average Fisher's information minus the information in the prior. .

### 3 Revisiting Reference Priors

The maximization of Bernardo's (1979) criterion is usually a difficult problem. In order to get a simpler alternative procedure under specific conditions, we will state a useful asymptotic approximation between Bernardo's criterion functional and some members of the class  $\mathcal{A}$ . To keep the analysis tractable, we will restrict ourselves to the continuous one-dimensional parameter case.

Suppose that there are available  $n$  independent observations, say,  $(X_1, X_2, \dots, X_n)$ , of a distribution  $P_\theta$ ,  $\theta \in \Theta \subseteq \text{IR}$ . Hence, the random vector  $(X_1, X_2, \dots, X_n)$  has density  $dP_\theta/d\nu = f(\xi|\theta) = \prod_{k=1}^n f(x_k|\theta)$  for all  $\xi = (x_1, x_2, \dots, x_n)$  and all  $\theta \in \Theta$ , where  $P_\theta$  and  $\nu$  stand for product measures with identical factors  $P_\theta$  and  $\lambda$  respectively. Following Lindley (1956), a measure of the expected information about  $\theta$  of a sampling model  $f(x|\theta)$  provided by a random sample of size  $n$  when the prior is  $\pi(\theta)$ , is defined to be

$$\mathcal{L}^{(n)}(\pi) = \int f(\xi) \int f(\theta|\xi) \log \frac{f(\theta|\xi)}{\pi(\theta)} d\mu(\theta) d\nu(\xi). \quad (3.1)$$

Under specific regularity, and bounded variance conditions, which rule out the possibility that the *essentials* of the statistical model  $f(\xi|\theta)$  change when samples grow in size, it can be shown that as  $n \rightarrow \infty$ ,  $\mathcal{L}^{(n)}(\pi) - \mathcal{V}_{1,1,1}(\pi) = -\mathcal{V}_{0,0,1}(\varphi) + \log C\sqrt{n} + o(1)$ , where  $\varphi(z)$  is the density of  $Z \sim \mathcal{N}(0, 1)$ , and  $C$  is as in (2.2). Thus, instead of maximizing  $\mathcal{L}^{(\infty)}(\pi)$ , we have as an alternative procedure maximizing  $\mathcal{V}_{1,1,1}(\pi)$ , which is independent of  $n$ .

## 4 Good-Bernardo-Zellner Priors: Main Results

In this section we introduce Good-Bernardo-Zellner's priors as solutions of convex combination of relevant members of the class  $\mathcal{A}$ . Let  $\mathcal{M}_\phi(\pi) \stackrel{\text{def}}{=} \phi \mathcal{V}_{1,1,1}(\pi) + (1 - \phi) \mathcal{V}_{0,0,0}(\pi)$ ,  $0 \leq \phi \leq 1$ . Plainly,  $\mathcal{M}_\phi(\pi) \in \mathcal{A}$  and is concave w.r.t.  $\pi$ .

Usually, in the absence of data, *supplementary* information, in terms of expectations about the parameter, comes from additional knowledge of the experiment, say,  $\int a_k(\theta) \pi(\theta) d\mu(\theta) = \bar{a}_k$ ,  $k = 1, 2, \dots, s$ . The functions  $a_k$  and the constants  $\bar{a}_k$  are known.

**Proposition 4.1** Consider the *Good-Bernardo-Zellner* problem:

$$\text{Maximize } \mathcal{M}_\phi(\pi), \quad \text{subject to } \mathcal{C} : \int a_k(\theta) \pi(\theta) d\mu(\theta) = \bar{a}_k, \quad k = 0, 1, 2, \dots, s, \quad a_0 \equiv 1 = \bar{a}_0.$$

Then a necessary condition for a maximum is

$$\pi_\phi^*(\theta) \propto [\mathcal{I}(\theta)]^{\frac{\phi}{2}} \exp\{(1 - \phi)\mathcal{F}(\theta) + \sum_{k=0}^s \lambda_k a_k(\theta)\}, \quad (4.1)$$

where  $\lambda_k$ ,  $k = 0, 1, \dots, s$ , are the Lagrange multipliers associated with the constraints  $\mathcal{C}$ . In particular,  $\pi_1^*(\theta)$  is Good-Bernardo's prior, and  $\pi_0^*(\theta)$  is Zellner's prior.

**Corollary 4.1** Consider the location and scale parameter families,  $f(x|\theta) = f(x - \theta)$ ,  $\theta \in \mathbb{R}$ , and  $f(x|\theta) = (1/\theta)f(x/\theta)$ ,  $\theta > 0$ , respectively, both satisfying  $\int [f'(x)]^2/f(x) d\lambda(x) < \infty$  and  $\int f(x) \log f(x) d\lambda(x) < \infty$ . Then, Good-Bernardo's and Zellner's priors agree regardless of the value of  $\phi \in (0, 1)$ .

To deal with the (local) uniqueness of the solution of the problem stated in Proposition 4.1, we rewrite the constraints,  $\mathcal{C}$ , as a function of the multipliers in the form

$$A(\Lambda) = [\int a_k(\theta) \pi_\phi^*(\theta) d\mu(\theta)]_{k=0}^s = \bar{A}$$

, where  $\bar{A}^T = (\bar{a}_0, \bar{a}_1, \dots, \bar{a}_s)$ , and  $\Lambda^T = (\lambda_0, \lambda_1, \dots, \lambda_s)$ .

**Proposition 4.2** Let  $\pi_\phi^*(\theta)$  be as in (4.1), and suppose that  $a_k$ ,  $k = 0, 1, \dots, s$ , are linearly independent continuous functions in  $L^2[\Theta, \pi_\phi^* d\mu]$  (the space of all  $\pi_\phi^* d\mu$ -measurable functions  $a(\theta)$  defined on  $\Theta$  such that  $|a(\theta)|^2$  is  $\pi_\phi^* d\mu$ -integrable). Suppose that  $A(\Lambda)$  is defined on an open set  $\Delta \subset \mathbb{R}^{s+1}$ , and let  $\Lambda_o$  be a solution of  $A(\Lambda) = \bar{A}$  for a fixed value of  $\bar{A} = \bar{A}_o$ . Then, there exists a neighborhood of  $\Lambda_o$ ,  $N(\Lambda_o)$ , in which  $\Lambda_o$  is the unique solution of  $A(\Lambda) = \bar{A}_o$  in  $N(\Lambda_o)$ .

**Proposition 4.3** The multipliers  $\Lambda^T = (\lambda_0, \lambda_1, \dots, \lambda_s)$  appearing in (4.1) satisfy the following non-linear system of  $s + 1$  equations:

$$1 = \lambda_0 + \log \left\{ \int [\mathcal{I}(\theta)]^{\frac{\phi}{2}} e^{(1-\phi)\mathcal{F}(\theta)} \prod_{k=1}^s e^{\lambda_k a_k(\theta)} d\mu(\theta) \right\},$$

$$1 = \lambda_0 - \log \bar{a}_k + \log \left\{ \int a_k(\theta) [\mathcal{I}(\theta)]^{\frac{\phi}{2}} e^{(1-\phi)\mathcal{F}(\theta)} \prod_{u=1}^s e^{\lambda_u a_u(\theta)} d\mu(\theta) \right\}, \quad k = 1, 2, \dots, s.$$

Moreover, (i) if the integral in the first equality has a closed-form solution, then the rest of the multipliers can be found from the relations:  $\partial \lambda_0 / \partial \lambda_k = \bar{a}_k$ ,  $k = 1, 2, \dots, s$ , and (ii)  $\phi \mathcal{V}_{1,1,1}(\pi_\phi^*) + (1-\phi)[\mathcal{V}_{0,0,0}(\pi_\phi^*) - 2\mathcal{V}_{0,0,1}(\pi_\phi^*)] = 1 - \sum_{k=0}^s \lambda_k \bar{a}_k$ , holds for all  $0 \leq \phi \leq 1$ .

The following proposition extends Good-Bernardo-Zellner's priors to a richer family by using the MMDIP and MFIP criteria:

**Proposition 4.5** Let

$$\mathcal{N}_{\phi,\psi}(\pi) \stackrel{\text{def}}{=} \phi \mathcal{V}_{1,1,1}(\pi) + (1-\phi)(1-\psi)\mathcal{V}_{0,0,0}(\pi) + (\psi(1-\phi)/2)[\mathcal{V}_{0,1,1} + \mathcal{V}_{0,1,0}],$$

$0 \leq \phi, \psi \leq 1$ . Then: (i)  $\mathcal{N}_{\phi,\psi}(\pi) \in \mathcal{A}$  and is concave w.r.t.  $\pi$ , and (ii) A necessary condition for  $\pi$  to be a maximum of the problem:

$$\text{Maximize } \mathcal{N}_{\phi,\psi}(\pi), \text{ subject to } \mathcal{C} : \int a_k(\theta) \pi(\theta) d\mu(\theta) = \bar{a}_k, \quad k = 0, 1, 2, \dots, s, \quad a_0 \equiv 1 = \bar{a}_0,$$

is given by

$$\pi_{\phi,\psi}^*(\theta) \propto [\mathcal{I}(\theta)]^{\frac{\phi}{2}} \exp \left\{ (1-\phi)(1-\psi)\mathcal{F}(\theta) + \frac{\psi(1-\phi)}{2} \left[ [\mathcal{I}(\theta)]^{\frac{1}{2}} + \frac{\mathcal{F}(\theta)}{[\mathcal{I}(\theta)]^{\frac{1}{2}}} \right] + \sum_{k=0}^s \lambda_k a_k(\theta) \right\},$$

where  $\lambda_k$ ,  $k = 0, 1, \dots, s$ , are the Lagrange multipliers associated with the constraints  $\mathcal{C}$ .

The second term inside the exponential of the above expression is the average between Fisher's information and the negative relative Shannon-Fisher's information. Notice that  $\pi_{\phi,0}^*(\theta)$  is Good-Bernardo-Zellner's prior.

## 5 Summary

We have presented, in a unifying framework, a number of well-known methods that maximize a criterion functional to obtain priors. Our general procedure is, by itself, capable of dealing with a range of interesting issues in Bayesian analysis. However, in this paper, we have limited our attention to Good-Bernardo-Zellner's priors as well as their application to some Bayesian inference problems. We have also emphasized the existence and uniqueness of the solutions of the corresponding variational problems. There are, of course, many other members of the class  $\mathcal{A}$  that deserve much more attention than that we have attempted here.

## References

- Akaike, H. (1978). A New Look at the Bayes Procedure, *Biometrika*, 65, 53-59.
- Bernardo, J. M. (1979). Reference Posterior Distributions for Bayesian Inference, *J. Royal Statist. Soc.*, B41, 113-147.
- Good, I. J. (1968). Utility of a Distribution, *Nature*, 219, 1392.
- Jeffreys, H. (1961). *Theory of Probability*, 3rd. edition, Oxford: University Press.
- Lindley, D. V. (1956). On a Measure of Information Provided by an Experiment, *Annals of Math. Statist.*, 27, 986-1005.
- Venegas-Martínez, F. (1998). On Information and Priors: A Synthesis, Tech. Rep. 100. Center for Research and Teaching of Economics, CIDE.
- Zellner, A., (1971). *An Introduction to Bayesian Inference in Econometrics*, New York: Wiley.



Esta publicación consta de 410 ejemplares y se terminó de imprimir en el mes de septiembre de 1998 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**



