UNIVERSIDAD ANAHUAC

CON ESTUDIOS INCORPORADOS A LA UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

ESTIMACION NO PARAMETRICA DE DENSIDADES DE PROBABILIDAD

T E S I S P R O F E S I O N A L

Que para obtener el Título de :

A C T U A R I O

P R E S E N T A :

ANA MARIA ROMERO FERNANDEZ

México, D. F.

1988

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

INDICE

INTRODUCCION

Un problema fundamental en estadística es la estimación de la función de densidad de una muestra de observaciones. Por el problema de estimación de una densidad de probabilidad queremos decir que dada la muestra aleatoria $X_1,\ldots,X_n$, estimemos f, la función de densidad de probabilidad que dio lugar a esta muestra.

Pueden seguirse dos tipos de procedimientos para la estimación de una densidad de probabilidad: el paramétrico y el no paramétrico. El procedimiento paramétrico consiste en suponer la forma funcional de la densidad de probabilidad y, entonces, se estiman sólo los parámetros. El no paramétrico consiste en estimar directamente la densidad de probabilidad, sin suposiciones previas.

La primera solución al problema de estimación no paramétrica de densidades fue el histograma que es todavía uno de los estimadores más usados. Las ideas básicas del histograma fueron aplicadas primero por John Graunt en 1662.

En 1895 Karl Pearson propuso una familia de distribuciones que incluía muchas de las densidades de probabilidad univariadas que comunmente se utilizan (hipergeométrica, binomial, beta, gaussiana, etc.). Sin embargo, no fue sino hasta 1956 cuando M. Rosenblatt propuso un tipo de estimador "desplazado" de histograma.

En 1962 Parzen construyó la clase de estimadores (de kernel) de histograma y examinó sus propiedades de consistencia. Himelsdorf y Wahba, en 1970, introdujeron la aplicación de las técnicas de spline en la estadística de nuestros días. Boneva, Kendall y Stefanov (1971) y Shoenberg (1972) examinaron el uso de las técnicas de spline, para obtener, de histogramas, "estimados suaves" de una función de densidad

de probabilidad.

Un resultado importante en estadística, que se obtuvo al usar el método de momentos con la familia Pearson, apareció en 1908 en un artículo de W.S. Gosset ("Student"), la distribución conocida como la "t de Student".

En esa época se pensó que iban a darse más resultados alrededor de las generalizaciones importantes de la familia Pearson. En vez de esto, R.A. Fisher se presentó con el concepto de estimación de máxima verosimilitud y desvió el empuje de la metodología de Pearson. La controversia Pearson-Fisher representó la pugna entre el método de momentos y el de máxima verosimilitud. La victoria de Fisher fue casi completa.

Aunque la estimación de máxima verosimilitud tiene más de medio siglo de edad, es todavía la más usada dentro de todas las técnicas de estimación. R.A. Tapia y J.R. Thompson (1978) establecieron la existencia y unicidad del estimador de máxima verosimilitud en el caso finito dimensional. Good y Gaskins (1971) introdujeron estabilidad al estimador de máxima verosimilitud en el caso infinito dimensional, añadiendo un término de penalización. Scott (1976) y Scott, Tapia y Thompson (1977) llevaron al estimador de máxima verosimilitud penalizada a la práctica.

El problema de la estimación no paramétrica de densidades tiene aplicación en muchas áreas de conocimiento, incluyendo:

i) Reconocimiento de Modelos: Identificar modelos, es decir, reconocer de qué población viene un modelo dado.

ii) Estudios de Simulación: Por medio de la estimación de una función de densidad, se puede decidir si una aparente "protuberancia"

en la función de densidad forma parte genuinamente de la población, o más generalmente para decidir si cierta función de densidad es mejor estimador que cualquier otra y por cuanto. Según Grear y Cassel (1971), "la validación de protuberancias" es una de las actividades de los físicos experimentales que requiere de la ayuda de la estadística.

iii) Análisis del Discriminante. En diversas aplicaciones surge con frecuencia el problema de clasificación, el cual puede describirse de la siguiente manera:

Se observa un fenómeno que se sabe fue generado por una de k poblaciones posibles. Estas poblaciones son teóricamente distintas pero difíciles de discriminar en base a observaciones empíricas.

El análisis del discriminante presenta un grupo amplio de metodologías para lograr esta clasificación.

Hay básicamente dos propósitos que se pueden distinguir en el uso del análisis del discriminante:

1) Descripción de las diferencias entre las poblaciones con base en datos muestrales. Esto se llama análisis descriptivo del discriminante.

2) La asignación óptima de nuevos elementos de origen incierto a una de las k poblaciones, usando la información contenida en las medidas de las p variables de estos nuevos elementos. Este tipo de problemas de asignación pueden ser subdivididos en dos categorías principales:

2.a) La identificación de nuevos elementos. El término "identificación" es usado con el propósito de asignar nuevos elementos a su población de origen más probable.

2.b) La asignación de acción-orientada de nuevos elementos. La asignación de un elemento es ahora seguida por una acción particularmente inducida por la población a la que fue asignado. Esta situación es más compleja que el problema de identificación porque las elecciones de acción no adecuadas pueden tener consecuencias muy diferentes, dependiendo de la población de origen de los elementos, así como de la población a la cual son asignadas.

La finalidad del acercamiento estadístico a estos problemas de asignación puede ser enunciada así: usar la información estadística de muestras de entrenamiento (muestras con elementos cuya población de origen se conoce) para encontrar una estrategia para la asignación de nuevos elementos de origen incierto, tal que la probabilidad de asignaciones erróneas sea minimizada (problemas de identificación), o la "pérdida" incurrida por las consecuencias (desagradables) de acciones no óptimas sea minimizada (problemas de acción orientada).

Algunos ejemplos del análisis del discriminante se presentan a continuación:

1) Análisis descriptivo del discriminante.

Se toman muestras de los individuos de distintos grupos indígenas. La estatura, la estatura de los individuos sentados, la profundidad nasal y la altura nasal son las variables medidas en ellos. Se busca describir las diferencias entre estos grupos.

2) Identificación.

Cromosomas: una célula humana contiene 46 cromosomas; 23 relacionados al sexo femenino y 24 al masculino. Se hacen algunas medidas sobre características de los cromosomas en las muestras de algunos tipos. Las diferencias en la distribución de las características de los tipos

serán usadas para identificar cromosomas de nuevas células.

3) Clasificación de acción-orientada.

Una actividad importante es la de diagnóstico médico. Un ejemplo de diagnóstico diferencial es el que se tiene en la siguiente situación:

Haemophilia A: Las mujeres pueden ser portadoras obligatorias o no. La detección de las portadoras es un problema que surge en investigación genética. Dos medidas bioquímicas se hacen en muestras de entrenamiento de portadoras y no portadoras. Para una llamada "posible portadora" las mismas medidas son realizadas y comparadas con las medidas de las muestras de entrenamiento de portadoras y no portadoras para así tomar una decisión sobre la posible naturaleza portadora de la mujer en estudio.

Este trabajo de tesis tiene como objetivo estudiar algunos aspectos del problema de estimación de densidades de probabilidad. Consta de cuatro capítulos:

En el primer capítulo se presenta el método de estimación no paramétrica de densidades conocido como el método de kernel. Se habla del kernel univariado y sus propiedades y más adelante se presenta el kernel multivariado en el cual los datos que se proporcionan pueden encontrarse medidos en diferentes escalas: binarias, nominales y ordinales para el caso continuo y para el caso mixto (continuo y discreto).

En el segundo capítulo se habla del método de máxima verosimilitud en el que se marca el problema de estabilidad que se presenta en el caso infinito-dimensional por lo que se añade un término de penalización al estimador de máxima verosimilitud. A continuación se presenta el ajuste del estimador de máxima verosimilitud por medio del estimador discreto de máxima verosimilitud penalizada.

En el tercer capítulo se presentan otros métodos para la estimación no paramétrica, en concreto el de series ortogonales y el método de proyección adaptativa ("projection pursuit") para la estimación no paramétrica multivariada.

Finalmente, el cuarto capítulo presenta un ejemplo de estimación de densidades utilizando el método de kernel, aplicado al análisis del discriminante.

Apéndices al final introducen a los espacios de Hilbert y a la teoría de splines, conceptos utilizados en el desarrollo de la tesis.

CAPITULO 1

METODO DEL KERNEL

1.1. El Caso Univariado.

De los métodos usados para la estimación no paramétrica de densidades de probabilidad, el del histograma es el de mayor uso.

Supóngase que tenemos una muestra aleatoria $x_1,\dots,x_n$ de una densidad de probabilidad desconocida absolutamente continua con dominio de positividad $[a,b]$. En el caso en que la densidad desconocida, digamos $g(x)$, tenga un rango infinito, será suficiente estimar la densidad truncada

$$(1) \qquad f(x) = \frac{g(x)}{\int_a^b g(t)dt} \quad \text{para } a \le x \le b$$

$$= 0 \text{ de otra manera.}$$

En lo que sigue vamos a suponer que los puntos fuera del intervalo $[a,b]$ han sido descartados y que cada uno de las $\{x_i\}$, $i=1,\dots,n$ estan en el intervalo $[a,b]$.

Sea $a=t_0 < t_1 < \dots < t_m=b$ una partición del intervalo $[a,b]$, entonces se obtiene un estimador f_h de la forma

$$(2) \qquad f_h(t) = c_i \quad \text{para } t_i \le t < t_{i+1}, \quad i=0,\dots,m-1$$

$$f_h(b) = c_{m-1}$$

$$f_h(t) = 0 \quad \text{para } t \notin [a,b],$$

donde $f_h(t) \ge 0$ y $\int_a^b f_h(t)dt = 1$.

Si q_i es el número de observaciones que caen en el intervalo i-ésimo, entonces para $\hat{f}_h$ (el estimador de f_h) usaremos

$$(3) \qquad \hat{c}_i = \frac{q_i}{n(t_{i+1} - t_i)} \quad \text{para} \quad i=0,\ldots,m-1.$$

para estimar c_i.

El número de observaciones que caen en un intervalo es una variable multinomial. Entonces q_i/n, que es la frecuencia por intervalo, estima a $\int_{t_i}^{t_{i+1}} f(t)dt$. Dado que hemos supuesto que f es absolutamente continua, si $t_{i+1} - t_i$ es suficientemente pequeño, entonces $f(t) \sim f(t_i)$ para $t_i \leq t < t_{i+1}$. Entonces q_i/n estima a $(t_{i+1} - t_i)f(t)$ pues $\int_{t_i}^{t_{i+1}} f(t_i)dt = f(t_i)\int_{t_i}^{t_{i+1}} dt = f(t)(t_{i+1} - t_i)$ y $\frac{q_i}{n(t_{i+1} - t_i)}$ estima a $f(t)$.

Entre estimadores de la forma (2), $\hat{f}_h$ maximiza únicamente la función de verosimilitud

$$(4) \qquad L(c_0,\ldots,c_{m-1}) = \prod_{j=1}^{n} f_h(x_j) = \prod_{i=0}^{m-1} c_i^{q_i}.$$

R.... Tapia y J.R. Thompson (1978) demostraron la consistencia de este estimador bajo ciertas condiciones:

En primer lugar se supone que f tiene derivadas continuas hasta de orden 2 excepto en los puntos extremos $[a,b]$, y que f se halla acotada en $[a,b]$. Adicionalmente si se toma la partición de igual tamaño en $[a,b]$, tal que $t_{i+1,n} - t_{i,n} = 2h_n$, entonces si $n \to \infty$ y $h_n \to 0$, tal que $nh_n \to \infty$ (por lo cual $n \to \infty$ más rápidamente de lo que $h_n \to 0$) se tiene que $x \in (a,b)$ y el error cuadrático medio para $\hat{f}_h$ estaría dado por:

$$(5) \qquad ECM(\hat{f}_h(x)) = L(\hat{f}_h(x) - f(x))^2 \to 0,$$

por lo que, $\hat{f}_n(x)$ es un estimador consistente para $f(x)$.

En 1956 Murray Rosenblatt propuso otro tipo de estimador basado en una muestra aleatoria $\{x_i\}$, $i=1,\ldots,n$ de una densidad continua pero desconocida f. El estimador está dado por:

$$\hat{f}_n(x) = \frac{\text{número de puntos muestrales en } (x-h_n, x+h_n)}{2nh_n}, \tag{6}$$

donde h_n es un número real positivo constante para cada n. Entonces podemos escribir

$$\hat{f}_n(x) = \frac{F_n(x+h_n) - F_n(x-h_n)}{2h_n}, \tag{7}$$

donde

$$F_n(x) = \frac{\text{número de puntos muestrales} \leq x}{n}.$$

En otras palabras $F_n(x)$ resulta ser la función de distribución empírica para la variable aleatoria x (Mood et. al. 1974).

Para obtener la fórmula del error cuadrático medio de f_n, Rosenblatt ve que si particiona la recta de los reales en tres intervalos $\{x|x \leq x_1\}$, $\{x|x_1 < x \leq x_2\}$ y $\{x|x > x_2\}$, y sean $Y_1=F_n(x_1)$, $Y_2=F_n(x_2)-F_n(x_1)$ y $Y_3=1-F_n(x_2)$, entonces (nY_1, nY_2, nY_3) es una variable aleatoria trinomial con probabilidades $(F(x_1), F(x_2)-F(x_1), 1-F(x_2))=(p_1,p_2,p_3)$, donde F es la función de distribución acumulativa de x, o sea, $F(x)=\int_{-\infty}^{x} f(t)dt$. Entonces tenemos que

$$E\left[F_n(x)\right] = F(x) \tag{8}$$

$$\mathrm{Cov}\left[F_n(x_1), F_n(x_2)\right] = -\frac{F(x_1)F(x_2)}{n} + \frac{F(x_1)}{n}, \tag{9}$$

suponiendo sin pérdida de generalidad que $x_1 < x_2$.

Entonces, sin hacer restricciones en x_1 y x_2,

$$(10)\quad \mathrm{Cov}\left[\hat{f}_n(x_1),\hat{f}_n(x_2)\right]$$

$$= \frac{1}{4nh_n^2}\Big[-F(x_1 + h_n)F(x_2 + h_n) + F(\min(x_1 + h_n, x_2 + h_n))$$
$$+ F(x_1 + h_n)F(x_2 - h_n) - F(\min(x_1 + h_n, x_2 - h_n))$$
$$+ F(x_1 - h_n)F(x_2 + h_n) - F(\min(x_1 - h_n, x_2 + h_n))$$
$$- F(x_1 - h_n)F(x_2 - h_n) + F(\min(x_1 - h_n, x_2 - h_n))\Big].$$

Si $x_1 = x_2 = x$, entonces

$$(11)\quad \mathrm{Var}\left[\hat{f}_n(x)\right]$$
$$= \frac{1}{4nh_n^2}\Big[F(x + h_n) - F(x - h_n) - (F(x + h_n) - F(x - h_n))^2\Big].$$

Entonces si escogemos $x_1 < x_2$ y h_n suficientemente pequeño de tal manera que $x_1 + h_n < x_2 - h_n$

$$(12)\quad \mathrm{Cov}\left[\hat{f}_n(x_1),\hat{f}_n(x_2)\right]$$
$$= \frac{-1}{n} f(x_1)f(x_2) - \frac{h_n^2}{6n}\Big[F(x_1)f'(x_2) + F(x_2)f''(x_1)\Big] + U\left(\frac{h_n^3}{n}\right)$$

suponiendo que f es diferenciable tres veces en x_1 y x_2 y $U\left(\frac{h_n^3}{n}\right)$ implica que si $h_n \to 0$ cuando $n \to \infty$, $nh_n \to \infty$. Ahora,

$$(13)\quad \mathrm{ECM}(\hat{f}_n(x)) = E\left[(\hat{f}_n(x) - f(x))^2\right] = \mathrm{Var}\left[\hat{f}_n(x)\right] + \mathrm{Bias}^2\left[\hat{f}_n(x)\right]$$
$$= \frac{1}{4nh_n^2}\Big[F(x + h_n) - F(x - h_n) - (F(x + h_n) - F(x - h_n))^2\Big]$$
$$+ \left[\frac{1}{2h_n}(F(x + h_n) - F(x - h_n)) - f(x)\right]^2.$$

Pero $F(x + h_n) - F(x - h_n) = 2h_nf(x) + \frac{h_n^3}{3} f''(x) + U(h_n^4)$, (suponiendo que f es tres veces diferenciable en x).

$$(14) \qquad ECM(\hat{f}_n(x)) = \frac{f(x)}{2h_n n} + \frac{h_n^4}{36}(f''(x))^2 + o\left(\frac{1}{h_n n} + h_n^4\right).$$

teniendose entonces que si $h_n \to 0$, cuando $n \to \infty$ de tal manera que $nh_n \to \infty$, $ECM(\hat{f}_n(x)) \to 0$. Haciendo f,x y n constantes; podemos minimizar los primeros dos términos en (14) y llegamos a

$$(15) \qquad \hat{h}_n = \left(\frac{9}{2} \frac{f(x)}{(f''(x))^2} \right)^{1/5} n^{-1/5},$$

con el correspondiente error cuadrático medio asintótico (en n) de

$$(16) \qquad ECM(\hat{f}_n(x)) = \frac{5}{4} 9^{-1/5} 2^{-4/5} |f(x)|^{4/5} (|f''(x)|)^{2/5} n^{-4/5}$$

Estudios de una variedad de medidas de consistencia de estimados de densidades son dados en: Bickel et. al. (1973), Kim et. al. (1974), Nadaraya, E.A. (1965), Schuster et. al. (1970), Van Ryzin, J. (1969) y Woodroofe, M. (1970).

Se puede decir que el estimador de Rosenblatt es un histograma desplazado de tal manera que x cae en el centro de un intervalo de la partición. Cuando se evalúa la densidad en otro punto "y", el intervalo es desplazado otra vez de tal manera que "y" esté en el centro de un intervalo de la partición. La ventaja del estimador desplazado sobre el de intervalo fijo es que hay una reducción en el sesgo. Otra ventaja sobre el de intervalo fijo es que la tasa de decremento del procedimiento de Rosenblatt es de $n^{-4/5}$ en vez de $n^{-2/3}$ (más lenta).

R.A. Tapia y J.R. Thompson (1978) formulan otra representación del histograma desplazado:

(17) $$\hat{f}_n(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h_n} w\left(\frac{x - x_j}{h_n}\right),$$

donde

$$w(u) = 1/2 \qquad \text{si } |u| < 1$$
$$= 0 \qquad \text{de otra manera,}$$

y los $\{x_j\}$, $j=1,\ldots,n$ son las observaciones, y para todos los $\{x_j\}$, $j=1,\ldots,n$ nuevamente se cumple que:

$$\int \hat{f}_n(x)dx = 1 \quad \text{y} \quad \hat{f}_n(x) \geq 0.$$

Intuitivamente lo que pasa es que a los valores de x_j cercanos a x se les asigna un peso y a los lejanos a x se les descarta.

El histograma desplazado es una función de densidad de probabilidad con buen comportamiento al igual que el histograma con intervalo fijo.

Grace Wahba (1977) habla de la manera en que se escoge el parámetro h y comenta que el estadístico tiene que escoger el parámetro h que va a controlar la suavidad visual de la densidad resultante. Matemáticamente hablando, comenta Wahba, h controla el balance entre el sesgo al cuadrado y la varianza; entonces la h óptima desde el punto de vista del ECM depende del tamaño de la muestra y de la densidad desconocida.

Para escoger un valor global para h_n, en el caso de la ecuación (17), R.A. Tapia y J.R. Thompson (1978) analizan el error cuadrático medio integrado

$$\int ECM(\hat{f}_n(x))dx = ECMI.$$

que en este caso es proporcional a

$$(18) \qquad ECMI \sim \frac{1}{2h_n n} + \frac{h_n^4}{36} \int_{-\infty}^{\infty} (f''(x))^2 dx + o\left(\frac{1}{h_n n} + h_n^4\right),$$

de aquí se tiene que

$$(19) \qquad h_n = \left[\frac{9}{2 \int (f''(x))^2 dx} \right]^{1/5} n^{-1/5},$$

y entonces

$$(20) \qquad ECMI \sim n^{-4/5}\, 9^{-1/5}\, 5/4 \left[\int (f''(x))^2 dx \right]^{1/5} n^{-4/5}.$$

Aunque Rosenblatt sugirió la generalización de la forma (17) a estimadores usando bases diferentes que las funciones escalonadas, la explicación detallada de los estimadores de kernel se debe a Parzen (1962). R.A. Tapia y J.R. Thompson (1978) nos dan la forma de este estimador, demuestran su consistencia y proponen un procedimiento para obtener el valor de h.

Considérese un estimador de f(x) de la siguiente forma:

$$(21) \qquad \hat{f}_n(x) = \int_{-\infty}^{\infty} \frac{1}{h_n} K\left(\frac{x - y}{h_n}\right) dF_n(y) = \frac{1}{h_n n} \sum_{j=1}^{n} K\left(\frac{x - x_j}{h_n}\right),$$

donde

$$(22) \qquad \begin{cases} \int_{-\infty}^{\infty} |K(y)|\, dy < \infty \\ \sup_{-\infty < y < \infty} |K(y)| < \infty \\ \lim_{y \to \infty} |yK(y)| = 0 \end{cases}$$

y

$$(23) \qquad K(y) \geq 0$$

$$\int_{-\infty}^{\infty} K(y) dy = 1.$$

Este estimador sujeto a las condiciones antes descritas es asintóticamente insesgado si $h_n \to 0$ cuando $n \to \infty$, o sea,

$$\lim_n E(\hat{f}_n(x)) = f(x)$$

y es consistente, si añadimos la restricción adicional

$$\lim_{n\to\infty} nh_n \to \infty.$$

Para el histograma desplazado de Rosenblatt la tasa de decremento óptima del ECM es del orden de $n^{-4/5}$. Para la clase de estimadores de kernel de Parzen satisfaciendo (22) la tasa de decremento del ECM es del orden de $n^{-2r/(2r+1)}$. Por lo tanto en la práctica no debemos esperar un decremento más rápido del ECM para estimadores de esta clase que $n^{-4/5}$, el obtenido usando el histograma desplazado.

Un procedimiento, para generar el kernel en el caso de la ecuación (21), se obtiene al minimizar el ECMI, respecto a h_n, esto es:

$$(24) \qquad ECMI[\hat{f}_n] \sim \frac{1}{nh_n} \int_{-\infty}^{\infty} K^2(y)dy + h_n^{2r} K_r^2 \int_{-\infty}^{\infty} \left| f^{(r)}(x) \right|^2 dx$$

de aquí se tiene que:

$$(25) \qquad h_n = n^{-1/(2r+1)} \alpha(K) \beta(f),$$

donde

$$\alpha(K) = \left[\frac{\int K^2(y)dy}{2r \left(\int y^r K(y)dy/r! \right)^2} \right]^{1/(2r+1)}$$

y

$$\beta(f) = \left[\int \left| f^{(r)}(y) \right|^2 dy \right]^{-1/(2r+1)}.$$

Puesto que el método de Parzen supone que la forma funcional de K (y por lo tanto de r) está dada, la evaluación de $n^{-1/(2r+1)} \alpha(K)$ no es problema, sin embargo, para calcular $\beta(f)$ es necesario conocer f.

La h óptima teórica sirve para ver qué tan bueno es nuestro estimador. Para el cálculo de la h_n en la práctica, se recomienda empezar con valores muy grandes de h_n y secuencialmente hacer decrementos en h_n hasta que se obtienen estimados de densidad de probabilidad altamente ruidosos y entonces se hace un retroceso hasta llegar a la h_n con la que claramente se ve la forma de f.

1.7. El Caso Multivariado.

El método de kernel puede ser extendido para estimar densidades multivariadas. Un kernel se acomoda alrededor de cada observación de la muestra. Aquí, las funciones de kernel pueden tener cualquier forma, pero siempre deben satisfacer los supuestos de la función de densidad de probabilidad y deberán estar centrados en las observaciones.

El estimador no paramétrico de densidad de probabilidad resultante para una población se obtiene calculando el promedio de los kernels de una muestra correspondiente. La "suavidad" de los kernels tiene que ser fijada a priori o tiene que ser estimada de la muestra.

Una ventaja del método de kernel es que todo tipo de densidades de probabilidad subyacentes se pueden describir. Una desventaja es que algunas observaciones distantes pueden tener un gran impacto en el valor del estimado de la densidad de probabilidad en la vecindad de estas observaciones. Esto puede suceder, particularmente, cuando los tamaños de muestra son pequeños.

Supóngase entonces que tenemos que estimar la función de densidad

p-dimensional f(X) de una distribución desconocida, la información de f está dada a través de N observaciones independientes de esta distribución, o sea, $Y_i=(Y_{i1},\dots,Y_{im},\dots,Y_{ip})$ con i=1,...,N.

La idea básica del método de kernel es la de colocar una función con las propiedades de una distribución de probabilidad alrededor de cada observación Y_i y tomar el promedio de estas funciones sobre las N observaciones como el estimado de densidad.

Sea $K^{(P)}(X;\ Y_i,u)$ una distribución de probabilidad o función de kernel centrada en Y_i y sea u el parámetro de suavizamiento de la función de kernel $K^{(P)}$. Se supone aquí que u es independiente de Y_i, así que el parámetro de suavizamiento u tiene un valor fijo sobre todos sus posibles valores.

El estimador no paramétrico de f(X) está dado por

$$f(X) = \frac{1}{N}\sum_{i=1}^{N} K^{(P)}(X;Y_i,u). \tag{26}$$

Se tienen que hacer dos elecciones para llegar a un estimado único:

a) la especificación de $K^{(P)}$ y

b) la estimación de u.

Con respecto a la especificación de $K^{(P)}$ nos restringimos a las funciones de kernel $K^{(P)}$ con p coordenadas independientes. Así, entonces,

$$K^{(P)}(X;Y_i,u) = \prod_{m=1}^{P} K_{(m)}(X_m;Y_{im},u), \tag{27}$$

con $K_{(m)}$ indicando el componente de la función de kernel de la variable X_m; u es independiente de m.

Como una consecuencia de este supuesto nada más tenemos que especificar el componente de kernel $K_{(m)}$ para cada una de las p dimensiones lo cual nos permite obtener a través de (26) y (27) la función de kernel $K^{(p)}$ y el estimador no paramétrico de densidad f.

Aquí hay que enfatizar que el supuesto de la independencia de las coordenadas hecho en (27), no implica independencia en la f final.

Por otro lado, el supuesto de (27) lleva a una subestimación de la estructura de correlación de los datos. Esto ha sido examinado por Kaasenbrood (1978) tanto teóricamente como en un estudio de simulación. Su conclusión general fue que el supuesto de independencia de (27) lleva a estimación de densidades poco satisfactoria en el caso de tamaños de muestra pequeños cuando la correlación entre variables es de 0.9 o mayor. Como este caso es muy raro en aplicaciones, se espera que (27) no sea restrictivo.

Por lo que se refiere al parámetro de suavizamiento u, la siguiente modificación del método de máxima verosimilitud fue propuesta por Habbema (1974) y por Duin (1976):

$$\max L(u) = \prod_{j=1}^{N} f_{jack}(Y_j) \tag{28}$$

con

$$f_{jack}(Y_j) = \frac{1}{N-1} \sum_{\substack{i=1 \\ i \neq j}}^{N} K^{(p)}(Y_j; Y_i, u). \tag{29}$$

El estimador que se obtiene es del tipo Jacknife.

Ahora veremos los componentes de kernel para datos continuos, así como para datos mixtos (continuos y discretos).

a) Un Componente de Kernel para Datos Continuos.

Escogemos para $K_{(m)}(x_m; Y_{im}, u)$ una densidad normal con media Y_{im} y

varianza $u^2 S_m^2$ donde

$$S_m^2 = \frac{1}{N-1} \sum_{i=1}^{N} (Y_{im} - \bar{Y}_m)^2 \qquad (30)$$

$$K_{(m)} = \frac{1}{u\, S_m \sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{X_m - Y_{im}}{u\, S_m} \right)^2 \right). \qquad (31)$$

b) Componentes de Kernel para Datos Mixtos (Continuos y Discretos).

Una ecuación general que define un componente de kernel para datos discretos (ordenados y desordenados) así como para datos continuos, está dada por

$$K_{tipo(m)}(X_m; Y_{im}, u) = \frac{1}{C_{tipo(m)}(u)}\, u^{d^2_{tipo\ (m)}(X_m; Y_{im})}, \qquad (32)$$

con: $m=1,\ldots,p$; $i=1,\ldots,N$; $0 < u < 1$ independiente del tipo m y

$$tipo\ (m) = \begin{cases} b & \text{si los datos son binarios} \\ n & \text{si los datos son nominales} \\ o & \text{si los datos son ordinales} \\ c & \text{si los datos son continuos} \end{cases}$$

El factor C(u) va a depender de la escala de X_m. Para variables discretas vamos a denotar las categorías de posibles resultados por $X_m(t)$, $t=1,\ldots,T_m$, y determinaremos el factor C(u) por

$$\sum_{t=1}^{T_m} K_{tipo(m)}(X_m(t); Y_{im}, u) = 1 . \qquad (33)$$

La función d^2 es una función de distancia normalizada, o sea, su promedio entre puntos muestreles toma algun valor constante normalizado:

(34) $$\underset{\substack{i,i'\\ i=i'}}{\text{promedio}}\ d^2_{\text{tipo}(m)}(Y_{im},Y_{i'm}) = \text{constante}$$

con

(35) $$d^2_{\text{tipo}(m)}(Y_{im},Y_{i'm}) = D^2_{\text{tipo}(m)}(Y_{im},Y_{i'm}) \Big/ S^2_{\text{tipo}(m)} .$$

En el caso continuo (tipo(m)=c), la normalización tiene lugar dividiendo la distancia Euclidiana por la varianza:

$$D^2_c(X_m,Y_{im}) = (X_m - Y_{im})^2$$

(36) y

$$S^2_c = \frac{1}{N-1}\sum_{i=1}^{N}(Y_{im} - \bar{Y}_m)^2 .$$

La constante en (34) toma el valor de 2. Para los otros 3 tipos se sigue exactamente el mismo razonamiento: se especifica una D^2 y una S^2 se deriva usando la relación (34) con la constante igual a 2.

c) Componente de Kernel Binario.

La distancia D^2_b y la varianza S^2_b son las mismas que en el caso continuo, vease (36). Computacionalmente se asignan conteos a los dos resultados $X_m(1)$ y $X_m(2)$. Suponemos que los conteos son escogidos de tal manera que su diferencia sea 1; así que $D^2_b = 0$ si $X_m = Y_{im}$ y $D^2_b = 1$ si $X_m \neq Y_{im}$. El factor C(u) es entonces

(37) $$C_b(u) = 1 + u^{1/S^2_b} .$$

d) Componente de Kernel Nominal.

El componente de kernel nominal solamente puede ser establecido reconociendo cuando X_m es igual a Y_{im} y cuando no lo es, por lo tanto la especificación de la distancia es

$$(38) \qquad D_n^2(X_m, Y_{im}) = \begin{cases} 0 & \text{si} \quad X_m = Y_{im} \\ 1 & \text{si} \quad X_m \neq Y_{im} \end{cases}.$$

La varianza muestral se define otra vez de manera que sea satisfecha la normalización en (34):

$$(39) \qquad S_n^2 = \left(N^2 - \sum_{t=1}^{T_m} N_m^2(t) \right) \Big/ 2N(N-1),$$

donde $N_m(t)$ es igual al número de observaciones, provenientes de $Y_{1m}, \ldots, Y_{Nm}$, con categoría de resultado t. El factor C (u) entonces es:

$$(40) \qquad C_n(u) = \sum_{t=1}^{T_m} u^{d_n^2(X_m(t), Y_{im})} = 1 + (T_m - 1)u^{1/S_n^2}.$$

e) Componente de Kernel Ordinal.

Suponemos que se asignaron calificaciones (scores) a las categorías ordenadas de resultados T_m, o sea, $X_m(t)$ y Y_{im} son calificaciones. En base a estas calificaciones se define una distancia similar al caso continuo: $D_o^2(X_m, Y_{im}) = (X_m - Y_{im})^2 / S_o^2$, ver (36).
El factor C (u) es entonces:

$$(41) \qquad C_o(u) = \sum_{t=1}^{T_m} u^{(X_m(t) - Y_{im})/S_o^2}.$$

Una complicación para el trabajo numérico es que $C_D(u)$ es explícitamente dependiente de la calificación observada de Y_{im}; esto es contrario al caso binario y al nominal, véase (37) y (40). Se tiene que ver que la anterior especificación de la distancia requiere de la asignación de calificaciones a las categorías de resultados, por lo tanto, las variables ordinales son manejadas como variables de escala de intervalos discretos.

f) Componente de Kernel Continuo.

La distancia D_c^2 se especifica en (36). El factor $C(u)$ es, para $0 < u < 1$ el siguiente:

$$(42) \qquad C_c(u) = \int_{X_m \in R} u^{(X_m - Y_{im})^2/S_c^2} dX_m = \sqrt{-\pi S_c^2 / \ln(u)} \ .$$

El componente de kernel continuo es ahora, para $0 < u < 1$

$$(43) \qquad K_c(X_m;Y_{im},u) = \sqrt{\frac{-\ln(u)}{\pi S_c^2}} \ u^{(X_m - Y_{im})^2/S_c^2} \ .$$

De hecho, (43) se obtiene por la sustitución de $u=\exp(-1/2\,\sigma^2)$ en

$$(44) \qquad \frac{1}{\sqrt{2\pi\sigma^2 S_c^2}} \exp\left(-\frac{(X_m - Y_{im})^2}{2\sigma^2 S_c^2}\right)$$

que resulta ser el componente de kernel Gaussiano con varianza $\sigma^2 S_c^2$.

Resumen de Componentes de Kernel.

En la tabla 1 se presenta un resumen para la ecuación general de componentes de kernel:

$$\frac{1}{C_{tipo(m)}(u)} \; u^{d^2_{tipo(m)}(X_m, Y_{im})}$$

la distancia d^2, la varianza S^2 y el factor de normalización C para cada uno de los tipos de variables consideradas.

Tabla 1

Resumen de distancias, varianzas y factores C(u) en la ecuación del componente de kernel para los distintos tipos de variables.

tipo	distancia d^2	varianza S^2	factor de normalización C
binario (b)	$(X_m - Y_{im})^2/S_b^2$	$\frac{1}{N-1}\sum_{i=1}^{N}(Y_{im}-\bar{Y}_m)^2$	$1 + u^{1/S_b^2}$
nominal (n)	$\begin{cases} 0 & \text{si } X_m = Y_{im} \\ 1/S_n^2 & \text{si } X_m \neq Y_{im} \end{cases}$	$\frac{N^2 - \sum_{t=1}^{T_m} N_m^2(t)}{2N(N-1)}$	$1 + (T_m - 1)u^{1/S_n^2}$
ordinal (o)	$(X_m - Y_{im})^2/S_o^2$	$\frac{1}{N-1}\sum_{i=1}^{N}(Y_{im}-\bar{Y}_m)^2$	$\sum_{t=1}^{T_m} u^{(X_m(t)-Y_{im})^2/S_o^2}$
continuo (c)	$(X_m - Y_{im})^2/S_c^2$	$\frac{1}{N-1}\sum_{i=1}^{N}(Y_{im}-\bar{Y}_m)^2$	$\sqrt{-\pi S_c^2/\ln(u)}$

Para variables discretas el rango del parámetro de suavizamiento es $0 < u < 1$. Un extremo nos lleva a la distribución uniforme y el otro a una distribución degenerada en un punto:

$$u = 1 \; : \quad K(X_m; Y_{im}, 1) = 1/T_m$$

$$u = 0 \; : \quad K(X_m; Y_{im}, 0) = \begin{cases} 1 & \text{si } X_m = Y_{im} \\ 0 & \text{si } X_m \neq Y_{im} \end{cases}$$

Para variables continuas el rango es $0 < u < 1$ y $u=1$ y $u=0$ tienen que ser consideradas como casos límites, véase (43): cuando $u \uparrow 1$ tenemos la distribución uniforme sobre la recta de los reales, cuando $u \downarrow 0$ tenemos la función de Dirac situada en Y_{im}.

El tema se trata con mayor amplitud por J. Hermans, J.D.F. Habbema, T.K.D. Kasanmoentalib, J.W. Raatgever (1982). Estos autores desarrollaron un programa para realizar un análisis del discriminante no paramétrico vía estimación de densidades utilizando el método de kernel, conocido como ALLOC80.

CAPITULO 2

METODO DE MAXIMA VEROSIMILITUD

2.1. Introducción.

El método de estimación de máxima verosimilitud es hasta ahora la técnica de estimación más usada. La función de verosimilitud de n variables aleatorias $X_1,\ldots,X_n$ se define como la densidad conjunta de las n variables aleatorias, sean $f_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta)$, que se consideran como una función de θ. En particular, si $X_1,\ldots,X_n$ es una muestra aleatoria de la densidad $f(x;\theta)$, entonces la función de verosimilitud es $f(x_1;\theta)\cdot f(x_2;\theta)\cdots f(x_n;\theta)$.

La función de verosimilitud $L(\theta;x_1,\ldots,x_n)$ da la verosimilitud de que las variables aleatorias tomen el valor particular $x_1,\ldots,x_n$. La verosimilitud es el valor de la función de densidad, así que para variables aleatorias discretas es una probabilidad.

Si tenemos los valores observados $x_1,\ldots,x_n$ queremos saber de qué densidad es más probable que hayan venido. Queremos un valor θ^* que maximice la función de verosimilitud $L(\theta;x_1,\ldots,x_n)$, entonces $\theta^*(x_1,\ldots,x_n)$ es el estimador de máxima verosimilitud de θ. Si $X_1,\ldots,X_n$ es una muestra aleatoria de la densidad $f(x;\theta)$ tenemos entonces que la función de verosimilitud a maximizar es:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta). \tag{1}$$

El estimador de máxima verosimilitud es la solución a la ecuación:

$$\frac{d\,L(\theta)}{d\theta} = 0. \tag{2}$$

Como $L(\theta)$ y $\text{Log } L(\theta)$ tienen su máximo en el mismo valor de θ, a veces es más fácil encontrar el máximo del logaritmo de la verosimilitud.

Ahora veremos el método de estimación no paramétrica de máxima verosimilitud.

2.2. Estimadores No Paramétricos de Máxima Verosimilitud.

R.A. Tapia y J.R. Thompson (1978) muestran la existencia y unicidad de la solución para el problema de estimación por el método de máxima verosimilitud en el caso finito dimensional. Presentan al histograma como un estimador de máxima verosimilitud y que en el caso infinito dimensional el funcional de verosimilitud no va a estar acotado y el estimador de máxima verosimilitud no va a existir.

Considérese el intervalo (a,b). Dada una muestra aleatoria $x_1,\dots,x_n \in (a,b)$ de una población con función de densidad $f(x)$, nos interesa estimar la función de densidad de probabilidad desconocida $f \in L^1(a,b)$ (integrable en el sentido de Lebesgue en (a,b)).

Dada la muestra aleatoria $x_1,\dots,x_n$ el funcional de verosimilitud $v \in L^1(a,b)$ está dado por

$$L(v) = \prod_{i=1}^{n} v(x_i). \tag{3}$$

Sea H un espacio vectorial (un espacio de funciones) en $L^1(a,b)$ y considérese el siguiente problema de optimización con restricciones:

(4) maximizar $L(v)$ sujeto a

$$v \in H, \quad \int_a^b v(t)dt = 1 \quad \text{y} \quad v(t) \geq 0 \quad \forall t \in (a,b).$$

Cuando nos referimos a una estimación de máxima verosimilitud basada en la muestra aleatoria $x_1,\dots,x_n$ y correspondiente al espacio vectorial H, nos referimos a cualquier solución del problema (4).

Ahora nos restringiremos al caso en el cual H es un subespacio finito dimensional (espacio vectorial lineal) de $L^1(a,b)$. Si H es un subespacio finito dimensional de $L^1(a,b)$ entonces un estimador de máxima verosimilitud basado en $x_1,\dots,x_n$ y correspondiendo a H existe. Además de existir es único pues, suponiendo que H es un subespacio finito dimensional de $L^1(a,b)$ con la propiedad de que existe por lo menos un $\varphi \in H$ satisfaciendo que $\varphi(t) \geq 0$ para toda $t \in [a,b]$ y $\varphi(x_i) > 0$, $i=1,\dots,n$ para la muestra aleatoria $x_1,\dots,x_n$, se tiene que si φ_1 y φ_2 son estimadores de máxima verosimilitud basados en $x_1,\dots,x_n$ y correspondiendo a H, entonces

$$\varphi_1(x_i) = \varphi_2(x_i), \quad i=1,\dots,n,$$

o sea que cualesquiera dos estimadores deben coincidir en los puntos muestrales.

Se tiene entonces de lo anterior que el estimador de máxima verosimilitud existe y es único, si H es de dimensión finita.

2.3. El Histograma como un Estimador de Máxima Verosimilitud.

Considérese la partición del intervalo (a,b), se define como $a=t_1 < \cdots < t_{m+1}=b$. Sea T_i un intervalo cerrado por la izquierda y abierto por la derecha, o sea, $[t_i,t_{i+1})$ para $i=1,\dots,m$. Déjese a $I(T_i)$ denotar la función indicadora del intervalo T_i, o sea,

$I(T_i)(x)=0$ si $x \notin T_i$ y $I(T_i)(x)=1$ si $x \in T_i$.

Para la muestra aleatoria $x_1,\dots,x_n \in (a,b)$, dejamos que $M(T_i)$ denote el número de estas muestras que caen en el intervalo T_i. Se ve que $\sum_{i=1}^{m} M(T_i)=n$. La teoría clásica nos dice que si queremos construir un histograma con intervalos de clase T_i, hacemos las alturas de los rectángulos con base T_i, proporcionales a $M(T_i)$. Esto nos lleva a que el histograma va a tener la siguiente forma:

$$f^*(x) = \sum_{i=1}^{m} \frac{\alpha M(T_i)}{(t_{i+1} - t_i)} I(T_i), \tag{5}$$

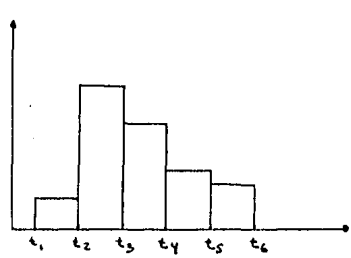
para alguna constante de proporcionalidad α. Dado que el área debajo de la función escalonada $f^*(x)$ tiene que ser igual a uno, tenemos entonces que

$$\sum_{i=1}^{m} \alpha M(T_i) = 1 \tag{6}$$

y por lo tanto $\alpha = \frac{1}{n}$ y el histograma estará dado por

$$f^*(x) = \sum_{i=1} \frac{M(T_i)}{n(t_{i+1} - t_i)} I(T_i). \tag{7}$$

La gráfica de $f^*(x)$ sería parecida a la siguiente:



Este histograma es un estimador de máxima verosimilitud único basado en la muestra aleatoria $x_1,\ldots,x_n$ y correspondiente al subespacio de $L^1(a,b)$ definido por

$$S_0(t_1,\ldots,t_m) = \left\{ \sum_{i=1}^{n} y_i \; I(T_i) \; : \; y_i \in R \right\}.$$

La notación $S_i(t_1,\ldots,t_m)$, i=1,2,... usada denota la clase de splines* (polinomios por secciones que se ajustan en alguna forma suave y los puntos en donde se ajustan los pedazos polinomiales son llamados nodos polinomiales de grado i con nodos en los puntos $t_1,\ldots,t_m$. Véase apéndice II).

2.4 El Caso Infinito Dimensional.

Se puede decir en general, que si el espacio vectorial H en el problema (4) es infinito dimensional, entonces un estimador de máxima verosimilitud no va a existir.

Para confirmar esto obsérvese que la solución está idealizada por

$$f^*(x) = \frac{1}{n} \sum_i \delta(x - x_i), \tag{8}$$

donde δ es la Delta de Dirac en el origen. Esta función va a dar el valor de ∞ si $x=x_i$ y de cero si $x \neq x_i$. La combinación lineal de Deltas de Dirac $f^*(x)$ "satisface las restricciones del problema (2)" y maximiza (hace infinito) el funcional de verosimilitud. Hemos entrecomillado porque tal comentario tiene más bien un valor de forma que de hecho, pues δ no existe como un miembro de $L^1(a,b)$. Sin embargo, en cualquier espacio vectorial infinito-dimensional $H \subset L^1(a,b)$, que tiene la propiedad de que es posible aproximar Deltas de Dirac, es decir,

* Usamos la palabra Spline porque la unica traducción que conocemos (cercha) nos parece inapropiada.

que dada una $t^* \in (a,b)$ existe $f_m \in H$ tal que f_m integra a uno, $f_m(t) \geq 0$ para toda $t \in (a,b)$ y $\lim_{m \to \infty} f_m(t^*) = +\infty$, el funcional de verosimilitud no va a estar acotado y el estimador de máxima verosimilitud no va a existir.

El hecho de que en general el estimador de máxima verosimilitud no existe para el H infinito-dimensional tiene una interpretación muy importante en términos de la H finito-dimensional. Específicamente para H de dimensión grande pero finita, el método de máxima verosimilitud va a llevar necesariamente a estimadores ásperos (que contienen picos) y a un problema numéricamente mal condicionado. Si se sabe a priori que la función de densidad de probabilidad desconocida es un miembro del espacio vectorial finito-dimensional H, entonces el metodo de máxima verosimilitud va a dar probablemente resultados satisfactorios. Sin embargo, este conocimiento solo se tiene en casos muy especiales.

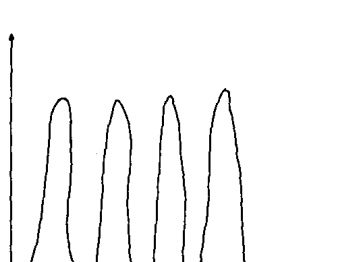
Las observaciones anteriores nos llevan a la siguiente crítica de la estimación de máxima verosimilitud en general: Para una H de dimensión pequeña no hay mucha flexibilidad y la solución va a estar muy influenciada por la elección subjetiva de H; mientras que para una H de dimensión grande la solución va a ser necesariamente áspera y el problema de optimización va a crear necesariamente problemas numéricos.

De Kontricher, RA. Tapia y JR. Thompson (1975) dieron un ejemplo para ilustrar muchos de estos puntos. Para un entero positivo n dado, particione el intervalo (a,b) en n intervalos medio-abiertos, medio-cerrados disjuntos $T_1, \ldots, T_n$ de igual longitud $h = \frac{b-a}{n}$. Déjese que $I(T_i)$ denote la función indicadora del intervalo T_i y déjese que ν_i denote el número de muestras en el intervalo T_i. El estimador de his-

tograma bien conocido está dado por

$$f^*(x) = \sum_{i=1}^{n} \frac{K}{Nh} I(T_i). \tag{9}$$

Este es un estimador de una densidad de probabilidad para la muestra aleatoria $x_1,\ldots,x_n$ correspondiente al espacio vectorial n-dimensional H, donde H es un espacio cuya base son las funciones indicadoras $\{ I(T_i) : i=1,\ldots,n \}$. Nótese que para una muestra fija cuando $n \to \infty$ el estimador $f^*(x)$ dado por (9) tiene la propiedad de que $f^*(x_i) \to \infty$, $i=1,\ldots,N$ mientras $f^*(x) \to 0$ si $x \notin \{x_1,\ldots,x_N\}$. Esto se ve en la siguiente gráfica:



Entonces para una n grande nuestro estimador de máxima verosimilitud es muy áspero e insatisfactorio y si obtenemos un estimador razonable depende completamente del delicado y engañoso arte de escoger la n bien.

Por estas razones y otras basadas en consideraciones heurísticas, Good y Gaskins (1971) sugirieron añadir un funcional de penalización al funcional de verosimilitud que castigara estimadores ásperos. Su-

girieron dos funcionales de penalización pero no probaron la existencia de estos. También sugirieron un método alterno para construir un estimador de máxima verosimilitud penalizada que evita la restricción de no-negatividad, pero tampoco aquí probaron la existencia de este método ni que éste daría el mismo resultado que el método original.

G.F. de Montricher, R.A. Tapia y J.R. Thompson en 1975 establecieron una teoría general de la existencia y unicidad para una clase grande de estimadores de máxima verosimilitud penalizada. Esta teoría se usa para demostrar que una clase bien conocida de espacios de Hilbert de kernel reproductor (espacios de Sobolev - ver apéndice I) llevan a estimadores de máxima verosimilitud penalizada que son splines polinomiales con nodos en los puntos muestrales. Demostraron la existencia y unicidad de uno de los estimadores de máxima verosimilitud penalizada de Good y Gaskins y que éste es un Spline exponencial (ver apéndice II) positivo con nodos solamente en los puntos muestrales y que el método alterno sugerido por Good y Gaskins da un estimador adecuado (que se acerca a la función verdadera con una aproximación razonable y que es consistente, o sea, que en el límite es la función verdadera). También demuestran la existencia y unicidad del otro estimador de Good y Gaskins.

2.5. Estimación de Máxima Verosimilitud Penalizada de Densidades.

Sea H un espacio vectorial en $L^1(a,b)$ y considérese un funcional $\Phi : H \to R$. Dada una muestra aleatoria $x_1,\dots,x_n \in (a,b)$ la verosimilitud penalizada Φ de $v \in H$ está definida por

$$P(v) = \prod_{i=1}^{n} v(x_i) \exp(-\Phi(v)). \tag{10}$$

Considérese el siguiente problema de optimización:

(11) Maximizar $F(v)$, sujeto a

$$v \in H, \quad \int_a^b v(t)dt = 1 \quad y \quad v(t) \geq 0, \ \forall t \in (a,b).$$

La forma general de la verosimilitud penalizada (10) se debe a Good y Gaskins (1971).

Cualquier solución al problema (11) se dice que es un estimador de máxima verosimilitud penalizada basado en la muestra aleatoria $x_1,\ldots,x_n$, correspondiente al espacio vectorial H y a la función de penalización ϕ. Vamos a estar especialmente interesados en el caso en que H sea un espacio de Hilbert infinito-dimensional.

En el caso de que H es un espacio de Hilbert, una función de penalización natural que se usa es $\phi(v) = \|v\|^2$ donde $\|\cdot\|$ denota la norma en H. Consecuentemente cuando H es un espacio de Hilbert y nos referimos al funcional de verosimilitud penalizado en H o a la estimación de máxima verosimilitud penalizada correspondiente a H con ninguna referencia a la función de penalización ϕ, estamos asumiendo que ϕ es el cuadrado de la norma en H. El producto interno del espacio de Hilbert será denotado por $\langle \cdot,\cdot \rangle$ de tal manera que $\langle x,x \rangle = \|x\|^2$.

Para que el problema (11) tenga sentido nos gustaría que H tuviera la propiedad de que para $x_1,\ldots,x_n \in (a,b)$ existiera por lo menos una $v \in H$ tal que

(12) $$\int_a^b v(t)dt = 1, \quad v(t) \geq 0 \quad \forall t \in (a,b)$$

y

$$v(x_i) \geq 0 \quad i=1,\ldots,n.$$

Supóngase que H es un espacio de Hilbert de Kernel reproductor, la integración sobre (a,b) es un funcional continuo y existe por lo

menos una $v \in H$ que satisface (12). Entonces la estimación de máxima verosimilitud penalizada correspondiente a H existe y es única.

La restricción de no-negatividad en el problema (11) lo convierte en un problema muy difícil de optimización a menos que se trabaje con algoritmos que tratan con densidades continuas que son lineales por secciones.

Por esta razón comunmente encontramos ejemplos en la literatura estadística donde un problema con una restricción de no-negatividad se resuelve trabajando con un problema equivalente planteado en términos de la raíz cuadrada de la densidad desconocida. Entonces la restricción de no-negatividad es redundante. Específicamente, dado $H \subset L^1(a,b)$ y $J:H \to R$ considere los siguientes dos problemas de maximización:

(13) Maximizar $J(v)$; sujeto a

$$v \in H, \quad \int_a^b v(t)dt = 1 \quad y \quad v(t) \geq u \quad \forall\, t \in (a,b)$$

y

(14) Maximizar $J(u^2)$; sujeto a

$$u^2 \in H \quad y \quad \int_a^b u^2(t)dt = 1 .$$

Entonces tenemos que:

i) Si v^* resuelve el problema (13), entonces $u^* = \sqrt{v^*}$ resuelve el problema (14).

ii) Si u^* resuelve el problema (14), entonces $v^* = (u^*)^2$ resuelve el problema (13).

Esto se ve claramente haciendo el cambio de variable $v=u^2$ y por lo tanto $u=\sqrt{v}$.

Tenemos como consecuencia de ya no trabajar con la restricción de no-negatividad que la restricción de la integral ahora es no-lineal.

Cuando consideremos los estimadores de máxima verosimilitud penalizada de Good y Gaskins, estaremos tratando con problemas de optimización con restricciones de la forma

(15) Maximizar J(v); sujeto a

$$\forall v \in \hat{H}, \quad \int_a^b v(t)dt=1 \quad y \quad v(t)\geq 0, \quad \forall\, t\in(a,b),$$

donde J esta definido en $H\subset L^1(a,b)$ y $\hat{H}$ tiene la propiedad de que $w \in \hat{H}$ implica que $w^2 \in H$. Para evitar trabajar con la restricción de no-negatividad, Good y Gaskins sugirieron que se trabajara con la versión análoga del problema (14), o sea,

(16) Maximizar $J(u^2)$; sujeto a

$$u\in\hat{H} \quad y \quad \int_a^b u^2(t)dt=1.$$

Aquí tenemos que los problemas (15) y (16) no siempre son equivalentes. Específicamente, tenemos la siguiente relación entre estos dos problemas: Si u^* resuelve el problema (16), entonces $v^*=(u^*)^2$ resuelve el problema (15) si y sólo si tenemos una condición adicional de que $|u^*|\in H$.

2.6. El Estimador Montricher-Tapia-Thompson.

Considérese el espacio de Sobolev:

$$H_0^s(a,b)=\left\{f: f^{(j)}\in L^2(a,b),\ j=0,\ldots,s \text{ y } f^{(j)}(a)=f^{(j)}(b)=0,\ j=0,\ldots,s-1\right\},$$

con producto interno

$$\langle f,g\rangle = \int_a^b f^{(s)}(t)\, g^{(s)}(t)\, dt\ .$$

El estimador de máxima verosimilitud penalizada correspondiente al espacio de Hilbert $H_0^s(a,b)$ y a la función de penalización $\phi(v)=\langle v,v\rangle$ existe y es único. Además, si el estimador es positivo en el interior del intervalo, entonces en ese intervalo es un spline polinomial de grado 2s y de clase de continuidad 2s-2 con nodos exactamente en los puntos muestrales.

2.7. El Primer Estimador de Good y Gaskins.

Como ya se había mencionado, por consideraciones teóricas, Good y Gaskins consideran el estimador de máxima verosimilitud penalizada correspondiente a la función de penalización

$$\phi_1(v) = \alpha \int_{-\infty}^{+\infty} \frac{v'(t)^2}{v(t)}\, dt \qquad (\alpha > 0)\ . \tag{17}$$

ellos no especifican en que espacio vectorial trabaja este estimador, pero Kontricher, Tapia y Thompson (1975) comentan que dadas las restricciones a cumplirse y el hecho de que:

$$\frac{1}{4}\phi_1(v) = \alpha \int_{-\infty}^{+\infty} \left[\frac{dv^{1/2}}{dt}\right]^2 dt \tag{18}$$

el espacio vectorial subyacente H, a saber $v^{1/2} \in H^1(-\infty,\infty)$ donde $H^1(-\infty,\infty)$ es el espacio de Sobolev

$$H^1(-\infty,\infty) = \left\{ f:R\to R : f' \text{ existe casi en todos lados y } f,f' \in L^2(-\infty,\infty)\right.,$$

con producto interno

$$\langle f,g\rangle = \int_{-\infty}^{\infty} f(t)g(t)dt + \int_{-\infty}^{\infty} f'(t)g'(t)dt\ .$$

Aquí se presenta una situación delicada porque es posible demostrar que el funcional de integración no es continuo en $H^1(-\infty,\infty)$. El problema (11) toma la siguiente forma

(19) Maximizar $P_1(v) = \prod_{i=1}^{n} v(x_i)\exp(-\phi_1(v))$, sujeto a

$$v^{1/2}\in H^1(-\infty,\infty), \quad \int_{-\infty}^{\infty} v(t)dt=1 \quad \text{y} \quad v(t)\geq 0 \quad \forall\, t\in(-\infty,\infty).$$

Para evadir la restricción de no-negatividad en el problema (16), Good y Gaskins consideraron trabajar con $v^{1/2}$ en vez de trabajar con v. Si dejamos que $u=v^{1/2}$, entonces reexpresando el problema (16) en términos de u obtendremos:

(20) Maximizar $\prod_{i=1}^{n} u(x_i)^2 \exp(-4\alpha\int_{-\infty}^{\infty} u'(t)^2\,dt)$; sujeto a

$$u\in H^1(-\infty,\infty) \quad \text{y} \quad \int_{-\infty}^{\infty} u(t)^2dt=1\,.$$

El problema (20) es resuelto para u^* y entonces $v^*=(u^*)^2$ se acepta como solución al problema (19). La transformación fue discutida en la sección 2.5 y la relación entre los problemas (15) y (16) demuestran que como estamos trabajando con $H^1(-\infty,\infty)$ entonces es válido.

El problema (20) no puede tener una solución única porque si u^* es una solución también lo es $-u^*$.

Añadiendo la restricción de no-negatividad al problema (20) y reexpresándolo al sacarle raíz cuadrada al funcional objetivo (ya que es no-negativo), llegamos al siguiente problema de optimización con restricciones:

(21) Maximizar $P(v) = \prod_{i=1}^{n} v(x_i)\exp(-\phi(v))$; sujeto a

$$v\in H^1(-\infty,\infty), \quad \int_{-\infty}^{\infty} v(t)^2dt=1 \quad \text{y} \quad v(t)\geq 0, \quad \forall\, t\in(-\infty,\infty)$$

donde

$$\Phi(v) = 2\alpha \int_{-\infty}^{\infty} v'(t)^2 dt$$

y α está dada por (17). Entonces,

i) Si v resuelve el problema (19), entonces $v^{1/2}$ resuelve el problema (20) y el (21).

ii) Si u resuelve el problema (20), entonces $|u|$ resuelve el problema (21) y u^2 resuelve el problema (19).

iii) Si v resuelve el problema (21), entonces v resuelve el problema (20) y v^2 resuelve el problema (19).

A partir de éste, Tapia y Thompson (1978) demostraron que el primer estimador de máxima verosimilitud penalizada de Good y Gaskins existe y es único; específicamente, el estimador de máxima verosimilitud penalizada correspondiente a la función penalizadora

$$\Phi(v) = \alpha \int_{-\infty}^{\infty} \frac{v'(t)^2}{v(t)} dt \qquad (\alpha > 0)$$

y al espacio vectorial

$$H = \left\{ v : v \geq 0 \quad \text{y} \quad v^{1/2} \in H^1(-\infty, \infty) \right\}$$

existe y es único, además el estimador es positivo y es un spline exponencial con nodos nada más en los puntos muestrales.

2.8. El Segundo Estimador de Good y Gaskins.

Dado un número de observaciones $x_1, \ldots, x_N$, el segundo método consiste en penalizar la esperanza del logaritmo del funcional de verosimilitud. Aquellos usados son combinaciones lineales de $\int y'^2\, dx$ y $\int y''^2\, dx$, donde $y = \sqrt{f}$. Este método parece ser consistente bajo am-

plias condiciones, aunque métodos consistentes pueden ser ásperos. Good y Gaskins demostraron que para ciertos valores de los coeficientes en esta expresion lineal, la función de densidad resulta ser muy suave aún cuando N es pequeña.

Consideren el funcional $\Phi(v)$: $H^2(-\infty,\infty)\to R$ definido por

$$\Phi(v) = \alpha\int_{-\infty}^{\infty} v'(t)^2dt + \beta\int_{-\infty}^{\infty} v''(t)^2dt \tag{22}$$

para algunas $\alpha \geq 0$ y $\beta > 0$. Por el segundo estimador de máxima verosimilitud penalizada de Good y Gaskins nos referimos a cualquier solución del siguiente problema de optimización con restricciones:

$$\text{(23) Maximizar } P_1(v) = \prod_{i=1}^{n} v(x_i)\exp(-\Phi(v^{1/2})); \text{ sujeto a}$$

$$v^{1/2}\in H^2(-\infty,\infty), \quad \int_{-\infty}^{\infty} v(t)dt=1 \quad y \quad v(t)\geq 0 \quad \forall t\in(-\infty,\infty).$$

Aquí sugirieron eliminar la restricción de no-negatividad calculando la solución del siguiente problema de optimización con restricciones:

$$\text{(24) Maximizar } \prod_{i=1}^{n} v(x_i)^2\exp(-\Phi(v)); \text{ sujeto a}$$

$$v \in H^2(-\infty,\infty) \quad y \quad \int_{-\infty}^{\infty} v(t)^2dt=1\ ,$$

donde Φ está dado por (22).

Junto con el problema (24) consideramos el problema de optimización con restricciones:

$$\text{(25) Maximizar } P(v) = \prod_{i=1}^{n} v(x_i)\exp(-1/2\,\Phi(v)); \text{ sujeto a}$$

$$v\in H^2(-\infty,\infty), \quad \int_{-\infty}^{\infty} v(t)^2dt=1 \quad y \quad v(x_i)\geq 0, \quad i=1,\ldots,n\ .$$

El problema (25) fue obtenido del problema (24) secándole la raíz

cuadrada al funcional que se va a maximizar (dado que es no-negativo) y requiriendo no-negatividad en los puntos muestrales; entonces, los dos problemas difieren en la restricción de no-negatividad en los puntos muestrales. Esta diferencia es lo que permitió establecer unicidad en la solución del problema (25), mientras que el problema (24) no puede tener una solución única. R.A. Tapia y J.R. Thompson demostraron que las soluciones del problema (24) y el problema (25) no son necesariamente no-negativas. De aquí se sigue que no podemos obtener la solución del problema (23) considerando los problemas (24) y (25). Dada la muestra aleatoria $x_1,\ldots,x_n$, si usamos v_*^2, donde v_* resuelve el problema (25), como un estimador para la función de densidad de probabilidad, entonces v_*^2 va a ser no-negativa y va a integrar a uno y por lo tanto es una densidad de probabilidad. Sin embargo, el estimador obtenido de esta manera no va a ser un estimador de máxima verosimilitud penalizada en el sentido estricto de la palabra, correspondiente al problema (23), es decir, el segundo estimador de máxima verosimilitud penalizada de Good y Gaskins. Por esta razón se refieren a este último estimador como el pseudo estimador de máxima verosimilitud penalizada.

R.A. Tapia y J.R. Thompson (1978) se encargaron de demostrar, a partir de lo anterior, que:

i) El problema (25) tiene una solución única, o sea, que el pseudo estimador de máxima verosimilitud penalizada existe y es único.

ii) El segundo estimador de máxima verosimilitud penalizada de Good y Gaskins existe y es único.

iii) El segundo estimador de máxima verosimilitud penalizada y el pseudo estimador de máxima verosimilitud penalizada de Good y Gaskins

son distintos.

B.W. Silverman (1981) se refiere a la verosimilitud penalizada de Good y Gaskins y la pone en la siguiente forma:

$$(16) \qquad U(f) = \sum \log f(x_i) - \alpha k(f)$$

donde $k(f)$ es un funcional como $\int (f'')^2$ y el parámetro α controla la cantidad por la cual los datos son suavizados para dar el estimador.

B.W. Silverman muestra que todos los métodos de estimación de densidades tienen la propiedad de que el estimador límite cuando la cantidad de suavización decrece es la suma de los picos en las observaciones, pero lo que sucede cuando la cantidad de suavización aumenta depende de exactamente qué método se está utilizando. Silverman dice que los penalizadores de asperezas con un funcional de penalización adecuado tienen una propiedad muy atractiva que él ilustra considerando un caso especial.

Supóngase que la penalización

$$(17) \qquad \Phi_N(f) = \int_{-\infty}^{\infty} \left\{ (d/dx)^2 \log f(x) \right\}^2 dx$$

es usada. Entonces, el estimador límite cuando el parámetro α tiende a infinito va a ser la densidad normal con la misma media y varianza que los datos. Entonces, según α varíe, el método va a dar un rango de estimadores desde la infinitamente áspera suma de funciones delta hasta el infinitamente suave ajuste normal de máxima verosimilitud de los datos.

Dificultades computacionales y matemáticas aparte, esta observación presenta un caso muy fuerte en el uso de estimación de densidades del método de penalización de asperezas con la penalización Φ_N. Dado que uno de los aspectos del método de estimación de densidades no para

métrico es el de investigar el efecto de los valores de parámetros que controlan la suavidad del estimador. Parece sensible que en el caso límite el estimador de densidad no paramétrico sea un estimador paramétrico natural.

Es posible definir otros penalizadores de asperezas de acuerdo con otras percepciones de la "infinitamente suave" familia exponencial de densidades de probabilidad. La propiedad escencial es que R(f) debe ser cero si y sólo si f está en la familia requerida.

2.9. Estimación Discreta de Máxima Verosimilitud Penalizada.

La estimación de máxima verosimilitud penalizada va a ser ahora llevada a la práctica. Los argumentos se deben principalmente a Scott (1976) y a Scott, Tapia y Thompson (1977).

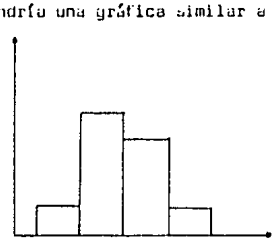
Dada la muestra aleatoria $x_1,\ldots,x_n$ queremos estimar, en un intervalo (a,b), la función de densidad desconocida f. Como ya se había visto, los estimadores necesitan integrar a uno en (a,b) y tener soporte en (a,b); entonces se sigue que si el soporte de la densidad desconocida f no está contenido en (a,b) entonces vamos a estar estimando la función de densidad $\bar{f}$, que está cerca de f en el sentido siguiente:

$$\bar{f}(x) = \begin{cases} \dfrac{f(x)}{\int_a^b f(x)\,dx} & \text{si } a < x < b \\ 0 & \text{de otra manera.} \end{cases} \tag{28}$$

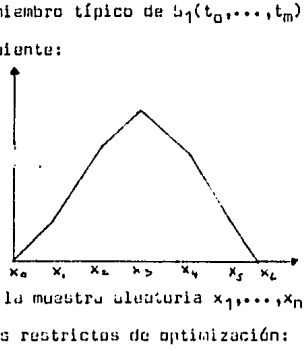
Nótese que f y $\bar{f}$ van a coincidir si y sólo si (a,b) contiene el

soporte de f. Para la muestra aleatoria $x_1,\ldots,x_n$ se asume que todas las observaciones afuera de (a,b) han sido censuradas.

Se van a construir aproximaciones al estimador de máxima verosimilitud penalizada correspondientes al espacio de Hilbert $H_0^1(a,b)$ por medio de la resolución de versiones finito dimensionales del problema (11), que son aproximaciones razonables al problema infinito dimensional. El término "discreta" se usa porque las versiones del problema finito dimensional (11) van a ser obtenidas trabajando con los espacios vectoriales lineales finitos que surgen de introducir una partición discreta en (a,b) y después considerando espacios de <u>splines</u> polinomiales de funciones que son constantes por secciones o lineales por secciones en esta partición. Específicamente, para un entero positivo dado m, considérese la partición uniforme $a=t_0<t_1<\cdots<t_m=b$, donde $t_i=a+ih$, $i=0,\ldots,m$, con $h=(b-a)/m$. Déjese que T_i denote el intervalo medio-abierto y medio-cerrado $(t_{i-1},t_i]$, para $i=1,\ldots,m$, y déjese que $I(T_i)$ denote la función indicadora del intervalo T_i, es decir, $I(T_i)(x)=0$ si $x\notin T_i$ y $I(T_i)(x)=1$ si $x\in T_i$. Déjese que $S_0(t_0,\ldots,t_m)$ denote las funciones con soporte en $[t_0,t_m]$, equivalentemente $[a,b]$, y que son constantes en los intervalos T_i. Finalmente, déjese que $S_1(t_0,\ldots,t_m)$ denote las funciones continuas con soporte en (t_0,t_m) y que son lineales en los intervalos T_i. Un miembro típico de $S_0(t_0,\ldots,t_m)$ tendría una gráfica similar a la siguiente:



mientras que un miembro típico de $S_1(t_0,\dots,t_m)$ tendría una gráfica similar a la siguiente:



Ahora, dada la muestra aleatoria $x_1,\dots,x_n$, considérense los siguientes problemas restrictos de optimización:

(29) Maximizar $L_h^0(y_0,\dots,y_{m-1}) = \prod_{i=1}^{n} S_0(x_i)\ \exp\left[-\alpha h^{-1}\sum_{i=1}^{m}(y_i-y_{i-1})^2\right]$;

sujeto a

$$h\sum_{i=0}^{m-1} y_i = 1 \qquad y \qquad y_i \geq 0, \qquad i=0,\dots,m-1,$$

donde S_0 está dada por la siguiente ecuación:

(30) $$S_0(x) = \sum_{i=1}^{m} y_{i-1} I(T_i)(x)$$

y $y_m=0$.

(31) Maximizar $L_h^1(y_1,\dots,y_{m-1}) = \prod_{i=1}^{n} S_1(x_i)\ \exp\left[-\alpha h^{-1}\sum_{i=1}^{m}(y_i-y_{i-1})^2\right]$;

sujeto a

$$h\sum_{i=1}^{m} y_i = 1 \qquad y \qquad y_i \geq 0, \qquad i=1,\dots,m-1,$$

donde $y_0 = y_m = 0$ y S_1 está dada por la siguiente ecuación:

(32) $$S_1(x) = \sum_{i=1}^{m}\left[y_{i-1} + h^{-1}(x-t_{i-1})(y_i-y_{i-1})\right] I(T_i)(x),$$

donde $y_i=S_1(t_i)$, $i=1,\dots,m-1$ y por suposición $y_0=S_1(t_0)=0$ y $y_m=S_1(t_m)=0$.

El spline constante (miembro de $S_0(t_0,\ldots,t_m)$) correspondiente a la solución del problema (29) y el spline lineal (miembro de $S_1(t_0,\ldots,t_m)$) correspondiente a la solución del problema (31) son llamados estimadores discretos de máxima verosimilitud penalizada (EDMVP).

Scott, Tapia y Thompson (1977) demostraron que el EDMVP existe y es único.

En contraste con los histogramas descritos en las secciones 2.2 y 2.3, los EDMVP son dimensionalmente estables. Desde este punto de vista, debemos pensar en los EDMVP como "histogramas estables".

Sea h el tamaño de la partición usada para obtener el EDMVP, entonces el spline constante EDMVP converge en subnorma cuando $h \to 0$ al monospline cuadrático EDMVP para $H_0^1(a,b)$, o sea, converge a la solución del problema infinito dimensional cuando el tamaño del intervalo se va a cero. Además, el spline lineal EDMVP converge en la norma de $H_0^1(a,b)$ cuando $h \to 0$ a este mismo monospline EDMVP.

Déjese que S_h denote el spline constante EDMVP dado por el problema (29) para un valor particular de h. Recuerde que $\bar{F}$ está dada por (20) y que $x_1,\ldots,x_n$ es nuestra muestra aleatoria. Considere el spline constante EDMVP, con el número de particiones dado por $m=\lfloor Cn^q \rfloor$ (de esta manera $h=1/(n^{-q})$) donde $C>0$, $0<q<1/4$ y $\lfloor a \rfloor$ denota al entero más cercano a a. Suponga que f es continua en (a,b). Entonces para $x \in (a,b)$

$$\lim_{n\to\infty} S_h(x) = \bar{F}(x) \quad \text{casi seguramente (c.s.)} \tag{33}$$

por lo que el EDMVP es consistente. Este resultado quiere decir que con la interacción apropiada entre el tamaño del intervalo y el tamaño de la muestra el EDMVP converge punto a punto c.s. a la densidad ver-

dadera.

Tapia y Thompson (1978) dieron ejemplos numéricos que calcularon usando el spline lineal SDMV que es obtenido de la verosimilitud penalizada, usando una segunda derivada discreta en el término penalizador. Hicieron esta elección principalmente porque el spline lineal da un estimador continuo y el uso de la segunda derivada en el término penalizador parece dar estimadores "más llenos".

Específicamente, dada una muestra aleatoria $x_1,\ldots,x_n$, un intervalo (a,b), un escalar positivo α y un entero positivo m, dejamos que

$$h = (b-a)/m \tag{34}$$

$$t_i = a + ih \quad i=0,\ldots,m \tag{35}$$

$$p_0 = p_m = 0 \tag{36}$$

y resolvemos el problema restricto de optimización m-1 dimensional:

$$\text{Maximizar } L(p_1,\ldots,p_{m-1}) = \sum_{i=1}^{n} \log p(x_i) - \frac{\alpha}{h}\sum_{k=1}^{m-1}\left[p_{k+1}-2p_k+p_{k-1}\right]^2 \tag{37}$$

sujeto a

$$h\sum_{k=1}^{m-1} p_k = 1 \quad \text{y} \quad p_k \geq 0, \quad k=0,\ldots,m-1$$

donde

$$p(t) = \begin{cases} p_k + \dfrac{p_{k+1}-p_k}{h}(t-t_k) & t\in[t_k,t_{k+1}) \\ 0 & t\notin(t_0,t_m). \end{cases}$$

Para concluir, podemos afirmar que los métodos de máxima verosimilitud son utilizados con frecuencia en la práctica. Existen paquetes de computación que implementan algunos de los métodos descritos.

En concreto, la librería IMSL (1986) incluye un programa para resolver el problema (36).

CAPITULO 3

OTROS METODOS

3.1. El Método de Proyección Adaptativa (Projection Pursuit).

El método de proyección adaptativa es uno más de los que pueden ser utilizados para estimar densidades. El proceso resultante es no-paramétrico y generalmente menos sesgado que los métodos de kernel.

Como ya se dijo, el objetivo de la estimación no paramétrica de densidades es el de estimar la densidad de probabilidad de un vector aleatorio p-dimensional $X \in R^p$ en base a observaciones independientes idénticamente distribuidas $x_1, \ldots, x_n$ sin hacer ninguna suposición sobre la familia paramétrica a la que puede pertenecer ésta.

En la práctica hay un objetivo muy importante que es el de tener una visión geométrica de la distribución de los datos en R^p.

La extensión de la estimación no paramétrica de funciones de densidad de probabilidad univariadas a las funciones de densidad de probabilidad multivariadas no ha sido exitosa en la práctica. Esto se debe en parte al deterioro de su realidad estadística causada por la llamada "maldición de dimensionalidad", esto es, se requieren amplias extensiones de los radios de vecindad para alcanzar suficientes conteos por clase.

Por esta razón, los estimadores resultantes son muy sesgados. Además estos métodos no proporcionan ninguna información comprensible sobre la forma en que se colocan los puntos en varias dimensiones (estructura de colocación).

Nuestra aproximación a la estimación no paramétrica de densida-

des multivariadas está basada en la noción de proyección adaptativa (projection pursuit)(Friedman y Tukey 1974; Friedman y Stuetzle 1981). Intenta dominar (sobreponerse a) la "maldición de dimensionalidad" extendiendo los métodos clásicos de estimación no paramétrica de densidades univariadas a dimensiones más grandes a través de las ideas propias de la estimación univariada. Adicionalmente, esto puede ayudar a explorar y entender mejor la distribución multivariada de los datos.

El objetivo del método de proyección adaptativa es el de estimar funciones multivariadas por medio de combinaciones de funciones univariadas seleccionando cuidadosamente combinaciones lineales de las variables.

El método construye estimadores de la forma

$$(1) \qquad f_M(x) = f_0(x) \prod_{m=1}^{M} f_m(\theta_m \cdot x),$$

donde f_M es el estimador de la densidad (o modelo actual) después de M iteraciones del procedimiento; f_0 es una función de densidad multivariada para ser usada como el modelo inicial; θ_m es un vector unitario especificando una dirección en R^p, por lo que $\theta_m \cdot x = \sum_{i=1}^{p} \theta_{mi} x_i$ es una combinación lineal de las medidas de las coordenadas originales; y f_m es una función de densidad univariada.

De (1) método de proyección adaptativa aproxima la densidad multivariada por medio de una densidad inicial propuesta f_0, multiplicada (aumentada) por el producto de las funciones univariadas f_m de las combinaciones lineales $\theta_m \cdot x$ de las coordenadas.

La elección de la densidad inicial es dejada al usuario y debe reflejar su mejor conocimiento *a priori* de los datos.

Una densidad Gaussiana con media muestral es generalmente una

elección natural. El propósito del método de proyección adaptativa es escoger las direcciones Θ_m y construir las funciones correspondientes $f_m(\Theta_m \cdot X)$. El producto de estas funciones estima el cociente entre los datos de la densidad del modelo y la densidad inicial.

De (1) obtenemos la relación recursiva

$$p_M(X) = p_{M-1}(X)\, f_M(\Theta_M \cdot X). \tag{2}$$

Dado que f_M se usa para modificar p_{M-1} para obtener p_M, nos referimos a las f_m como las funciones multiplicadoras.

La definición recursiva del modelo (2) sugiere una aproximación escalonada para su construcción. En la m-ésima iteración hay un modelo actual $p_{M-1}(X)$ construido en los pasos anteriores. (Para el primer paso, M=1 y el modelo actual es el modelo inicial $p_0(X)$ especificado por el usuario). Dado $p_{M-1}(X)$ buscamos un nuevo modelo $p_M(X)$ que genere una mejor aproximación a la densidad muestral $p(X)$. Entonces, son escogidas una dirección Θ_M y sus funciones multiplicadoras correspondientes $f_M(\Theta_M \cdot X)$ para maximizar la bondad de ajuste de $p_M(X)$. Se mide la bondad de ajuste relativa por medio del término de entropía cruzada de la distancia de Kullback-Leibler :

$$I = \int \log p_M(X)\, p(X)\, dX \tag{3}$$

cuyo objetivo es proporcionar una medida de información, en la misma, sobre la densidad que se está estimando. Por ejemplo, dadas dos densidades f y g, el número de Kullback-Leibler entre f y g se define como:

$$I(f,g) = \int f(x) \log(f(x)/g(x))dx,$$

donde $f(x) > 0$ y $g(x) \neq 0$.

$I(f,g)$ es no negativo y es igual a cero sólo cuando f es igual a g, por lo que puede ser interpretado como una medida de distancia entre f y g. Sin embargo, no es una distancia ya que no es simétrico y no satisface la desigualdad del triángulo.

De (2) puede demostrarse que $\bar{w}$ llega a su máximo en el mismo lugar que

$$\bar{w}(\Theta_M, f_M) = \int \log f_M(\Theta_M \cdot X)\, p(X)dX. \tag{4}$$

La ecuación (4) se maximiza bajo la restricción de que $p_M(X)$ esta normalizada, esto es, $\int p_M(X)dX=1$. Para una dirección Θ_M dada y $p(X)$ conocida,

$$f_M(\Theta_M \cdot X) = p^{\Theta_M}(\Theta_M \cdot X)/p_{M-1}{}^{\Theta_M}(\Theta_M \cdot X) \tag{5}$$

maximiza a (4).

Aquí p^{Θ_M} y $p_{M-1}{}^{\Theta_M}$ representan tanto a los datos como a las densidades marginales del modelo actual a lo largo del subespacio (unidimensional) extendido por Θ_M.

Usando esta f_M para un Θ_M dado, queda encontrar la dirección para la cual la ecuación (4) llega a su valor máximo. El óptimo Θ_M y su correspondiente función multiplicadora $f_M(\Theta_M \cdot X)$ definen el nuevo modelo a partir de (2).

En la práctica la densidad muestral $p(X)$ es desconocida. Tenemos, en su lugar, una muestra de N observaciones independientes e idénticamente distribuidas $x_1,\ldots,x_N$ de $p(X)$. La entropía cruzada $\bar{w}$ se estima a través del logaritmo de la verosimilitud

$$\hat{\bar{w}} = \frac{1}{N}\sum_{i=1}^{N} \log p_M(x_i). \tag{6}$$

Análogamente, $w(\Theta_M, f_M)$ es estimada por:

$$\hat{w}(\Theta_M, f_M) = \frac{1}{N}\sum_{i=1}^{N} \log f_M(\Theta_M \cdot X_i), \tag{7}$$

donde $f_M(\Theta_M \cdot X)$ es un estimador para el cociente de los datos y los modelos marginales a lo largo de Θ_M. El valor óptimo Θ_k que maximiza $\hat{w}(\Theta_M, f_M)$, y por lo tanto el logaritmo de la verosimilitud $\hat{\Omega}$ del nuevo modelo, es determinado por optimización numérica.

Procedimiento de estimación:

A continuación se presenta la estimación de $f(\Theta \cdot X)$, el cociente entre los datos y las distribuciones marginales a lo largo de la dirección Θ. Primero considere la distribución marginal $P_{M-1}^{\Theta}(\Theta \cdot X)$. Sin pérdida de generalidad, sea Θ el primer eje coordenado, o sea, $\Theta \cdot X = x_1$, entonces:

$$P_{M-1}^{\Theta}(x_1) = \int P_{M-1}(X)\,dx_2 dx_3 \cdots dx_n. \tag{8}$$

Si $P_{M-1}^{\Theta}(x_1)$ es continua, entonces

$$P_{M-1}^{\Theta}(x_1) = \lim_{h \to o} \frac{1}{2h} \int_{x_1-h}^{x_1+h} P_{M-1}^{\Theta}(z)\,dz \tag{9}$$

$$= \lim_{h \to o} \frac{1}{2h} \int_{-\infty}^{\infty} I(x_1-h \leq z \leq x_1+h) P_{M-1}^{\Theta}(z)\,dz, \tag{10}$$

donde

$$I(s) = 1 \quad \text{si } s \text{ es verdadera}$$
$$= 0 \quad \text{de otra manera.} \tag{11}$$

De (8) se tiene

$$P_{M-1}^{\Theta}(x_1) = \lim_{h \to o} \frac{1}{2h} \int I(x_1-h \leq y_1 \leq x_1+h) P_{M-1}(Y)\,dY \tag{12}$$

y $$F_{M-1}^{\theta}(x_1) = \lim_{h\to 0} \frac{1}{2h} E_{F_{M-1}} [I(x_1-h \le y_1 \le x_1+h)]$$

El estimado de $F_{M-1}^{\theta}(x_1)$ se obtiene de (7) usando un valor finito pequeño para h y empleando el método de Monte Carlo (Morrison, 1976) para estimar el valor esperado. Una muestra de Monte Carlo $y_1,\dots,y_{N_S}$, de tamaño N_S se genera con densidad $f_{M-1}(x)$ y

(13) $$\hat{F}_{M-1}^{\theta}(x_1) = \frac{1}{2hN_S} \sum_{j=1}^{N_S} I(x_1-h \le y_{j1} \le x_1+h)$$

se toma como el estimador de $F_{M-1}^{\theta}(x_1)$. Dado que la elección de x_1 así como la de la dirección θ fue arbitraria, (13) se puede escribir igualmente como

(14) $$\hat{F}_{M-1}^{\theta}(\theta\cdot x) = \frac{1}{2hN_S} \sum_{j=1}^{N_S} I(\theta\cdot x-h \le \theta\cdot y_j \le \theta\cdot x+h)$$

para cualquier θ. Nótese que la misma muestra de Monte Carlo se puede usar para todas las θ y x.

Los datos representan una muestra de $f(x)$ que puede ser usada, en analogía con (14), para estimar la marginal de los datos $F^{\theta}(\theta\cdot x)$ a través de

(15) $$\hat{F}^{\theta}(\theta\cdot x) = \frac{1}{2hN} \sum_{i=1}^{N} I(\theta\cdot x-h \le \theta\cdot x_i \le \theta\cdot x+h).$$

De (5) el estimador de las funciones multiplicadoras se convierte en

(16) $$f_{\theta}(\theta\cdot x) = \frac{f_S \sum_{i=1}^{N} I(\theta\cdot x-h \le \theta\cdot x_i \le \theta\cdot x+h)}{f_M \sum_{j=1}^{N_S} I(\theta\cdot x-h \le \theta\cdot y_j \le \theta\cdot x+h)}.$$

Esta es el cociente del conjunto de conteos muestrales al conjunto de conteos de Monte Carlo en un intervalo de longitud 2h centrado en $\theta\cdot x$. Para facilitar la estabilidad del denominador, escogemos h

de tal manera que siempre incluya exactamente a αN_θ observaciones de Monte Carlo. En este caso (16) se convierte en

$$(17) \qquad f_\theta(\theta \cdot x) = \frac{1}{\alpha N} \sum_{i=1}^{N} I(\theta \cdot x - h \leq \theta \cdot x_i \leq \theta \cdot x + h).$$

La fracción α se llama la extensión (span); es un parámetro del procedimiento. En aplicaciones actuales la extensión puede ser ajustada basándose en la inspección visual de las funciones multiplicadoras r_m y los estimadores del histograma de las marginales de los datos y modelo a lo largo de las direcciones θ_m. La longitud de la ventana del histograma (binwidths) debe escogerse pequeña de manera que estime las marginales con poco sesgo. El fin es escoger la mayor extensión posible, que hará suaves las funciones multiplicadoras r_m, sujeto a la restricción de que los histogramas de los datos y el modelo a lo largo de las direcciones no deberán diferir sistemáticamente.

3.7. Método de Series Ortogonales.

Los estimadores de series ortogonales han sido estudiados por Watson (1969), Kronmal y Tarter (1968) y recientemente por Brunk (1976) y por Crain (1976). Wahba (1977) hizo una presentación del método.

El estimador de series ortogonales para densidades con soporte en $[0,1]$ está dado por:

$$(18) \qquad \hat{f}(x) = \sum_{v=1}^{r} \hat{f}_v \phi_v(x),$$

donde $r << n$, los $\hat{f}_v$ son los coeficientes muestrales de Fourier

$$(19) \qquad \hat{f}_v = \frac{1}{n} \sum_{i=1}^{n} \phi_v(x_i)$$

y $\{\phi_v\}$ es un conjunto de $L_2[0,1]$ funciones ortonormales.

El parámetro r que tiene que ser escogido controla la "balanza" entre el sesgo cuadrado y la varianza. La r óptima depende del tamaño de muestra y de la densidad desconocida.

Wahba (1977) propuso un estimador de densidades basado en series ortogonales que tiene un algoritmo viable para la estimación del parámetro de suavizamiento óptimo.

CAPITULO 4

APLICACION DE LA ESTIMACION DE DENSIDADES EN EL ANALISIS DISCRIMINANTE

4.1. Introducción.

Se mencionó con anterioridad que el problema de análisis discriminante es básicamente el de clasificación. Usualmente el análisis discriminante se aplica de las siguientes dos maneras:

a) la identificación de nuevos elementos y

b) la asignación de acción-orientada de nuevos elementos.

A continuación se describe brevemente la forma en que se aplica el análisis discriminante. (Para cualquier lector interesado en profundizar más en el tema, se sugiere consultar a Morrison (1981) o Hermans et. al. (1982)).

Se comienza por obtener muestras de las k poblaciones de interés. La población de orígen tiene que conocerse para cada elemento. Estas muestras suelen llamarse muestras de entrenamiento.

Adicionalmente se debe proponer una regla que establezca a cuál de las k poblaciones deberá asignarse cada uno de los nuevos elementos muestreados cuya población de orígen es desconocida. Esta regla debe usar la información de los patrones de variabilidad de cada una de las k poblaciones, que es dada por las muestras de entrenamiento, y también la información a priori del posible orígen del nuevo elemento.

Una vez definida la regla de asignación para un problema en particular, es importante evaluar su comportamiento, esto es, conocer el porcentaje de asignaciones erróneas y el nivel de pérdidas esperadas en que se incurre por decisiones no óptimas. Este problema será tra-

tado más adelante con mayor detalle.

4.7. El Teorema de Bayes y la Regla de Asignación de Bayes.

En la práctica existen diferentes metodologías con las que pueden diseñarse reglas de asignación. Para los fines de este trabajo, discutiremos un procedimiento bayesiano.

El objetivo final de la regla de asignación es determinar la población de orígen de los nuevos elementos muestrales. Desde el punto de vista del problema de identificación, parece deseable hacer esta asignación en base a las probabilidades de pertenencia de estos elementos a las diferentes poblaciones bajo estudio. Un procedimiento plausible sería asignar el elemento a aquella población que genera la probabilidad más alta de pertenencia.

En el caso de no contar con medidas de ninguna variable, la asignación debería hacerse de acuerdo con el conocimiento *a priori* que se tenga del fenómeno bajo estudio. Este conocimiento *a priori* viene dando cuantitativamente la probabilidad *a priori* de pertenencia a las diferentes poblaciones.

Sea

(1) $P(A_j)$ - La probabilidad *a priori* de la población A_j.
$j=1,\ldots,k$.

En la práctica estas probabilidades no se conocen y se tienen que estimar. Cuando las muestras de entrenamiento se unen, formando un conjunto total, las probabilidades *a priori*, pueden ser estimadas de los tamaños de las muestras de entrenamiento $(N_1,\ldots,N_k)$.

Así mismo, si se juzga que cada población tiene la misma probabi-

lidad de ser escogida; todas las probabilidades a priori serían 1/k. Ocasionalmente las probabilidades a priori se calculan de otras maneras; en el ejemplo de los cromosomas, ya citado, parece adecuado fijar las probabilidades a priori para cada tipo como 2/46 ya que hay dos cromosomas de cada tipo en una célula humana (a excepción de los cromosomas del sexo masculino).

Sin embargo, cuando una observación $X=(x_1,\ldots,x_p)$ está disponible para el elemento a identificar, X nos da la información probabilística de la población de orígen. La probabilidad de ocurrencia va a ser generalmente distinta para cada una de las k poblaciones.

En este caso, la función verosimilitud o densidad de probabilidad puede denotarse por

(2) $f(X|A_j)$ - Esta densidad proporciona las probabilidades de pertenencia de la observación x a la población A_j. $j=1,\ldots,k$.

Queda claro entonces que estas densidades de probabilidad $f(X|A_1)$, $\ldots,f(X|A_k)$ deben de ser estimadas de las observaciones de las muestras de entrenamiento utilizando algún método de estimación.

La información contenida en la probabilidad a priori tiene que ser combinada con la información contenida en la verosimilitud para así obtener la probabilidad de pertenencia a la población A_j, una vez que se obtiene X.

(3) $P(A_j|X)$ - Es la probabilidad a posteriori de pertenencia de la población A_j: esta es la probabilidad de venir de la población A_j dada la observación X, $j=1,\ldots,k$.

La probabilidad a posteriori puede ser calculada a partir de la

probabilidad a priori y la verosimilitud por medio del teorema de Bayes:

$$(4) \qquad P(A_j|X) = \frac{P(A_j)\, f(X|A_j)}{\sum_{m=1}^{k} P(A_m) f(X|A_m)} \qquad j=1,\ldots,k.$$

De aquí vemos que la probabilidad a posteriori es proporcional al producto de la probabilidad a priori y la verosimilitud. El denominador $\sum_{m=1}^{k} P(A_m) f(X|A_m)$ es solamente un factor normalizante.

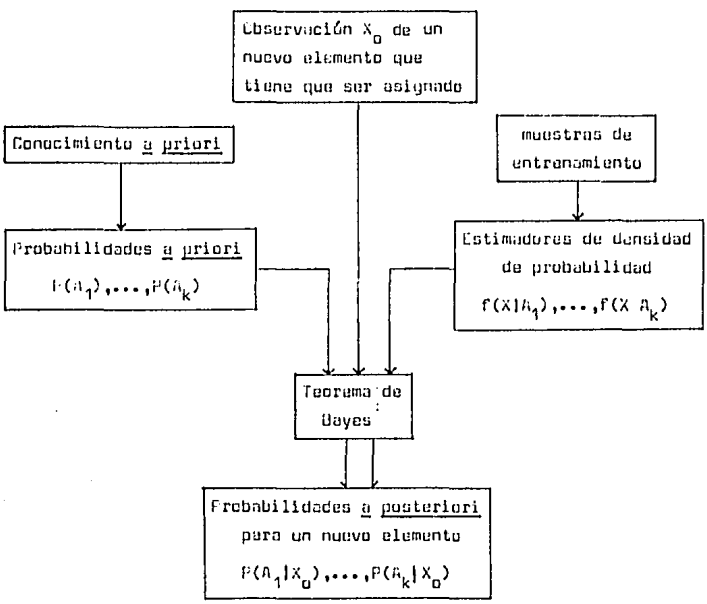


Figura 1: Una ilustración esquemática del cálculo de probabilidades a posteriori para un nuevo elemento.

Una regla común para la identificación de nuevos elementos es asignar cada elemento a la población que genera la probabilidad a posteriori más alta. Esta regla de asignación es una función de la observación X, con k posibles resultados: la asignación a una de las k poblaciones.

$$(5) \qquad a(X) = a_i \text{ si } P(A_i|X) = \max\left\{P(A_1|X),\ldots,P(A_k|X)\right\},$$

donde a(X) es la regla de asignación y a_i es el resultado "asignar a la población A_i".

En muchas aplicaciones, la identificación sólo se acepta cuando la probabilidad a posteriori más alta excede a un valor puesto de antemano, digamos δ. Por ejemplo, en el problema de diagnóstico médico donde hay dos categorías de enfermedad, la identificación (hacer un diagnóstico) puede ser apropiada sólo si la probabilidad a posteriori más alta es de por lo menos 0.90. La regla de asignación (5) se modifica a:

$$a(X) = \begin{cases} a_i & \text{si } P(A_i|X) = \max\left\{P(A_1|X),\ldots,P(A_k|X)\right\} \\ & \text{y } P(A_i|X) > \delta \\ \text{duda de otra manera.} & \end{cases}$$

Cuando asignamos a la población con la probabilidad a posteriori máxima, se minimiza la probabilidad de una asignación errónea, la cual quedaría medida a través de $1 - P(A_i|X)$ para la asignación a_i.

No se hace distinción entre los tipos de asignación erróneas, lo cual quiere decir que se juzgan como igualmente serios.

El tipo de asignación errónea que se comete es relevante en los problemas de asignación de acción orientada. Para cada combinación de

una población de orígen y la población a la cual es asignada, existe una pérdida específica: sea

(6) $L(a_i,A_j)$ - La pérdida incurrida cuando un elemento que viene de la población A_j es asignado a la población A_i.

$L(a_i,A_j)$ es una función sobre la que se imponen las siguientes restricciones:

(7) $$L(a_i,A_i) = 0$$
$$L(a_i,A_j) \geq 0.$$

Esto es, la pérdida incurrida en una decición correcta es cero, mientras que es no negativa en el caso de una incorrecta.

En este caso, el propósito será el de asignar cada elemento a la población, de manera que la pérdida esperada por una asignación incorrecta sea minimizada. La pérdida esperada condicional de la asignación de la población A_i, dada la observación X, denotada como $EL(a_i)$, es la suma de las pérdidas $L(a_i,A_j)$, $j=1,\ldots,k$, donde cada término de la suma es ponderado con la probabilidad a posteriori de pertenencia a las poblaciones de orígen correspondientes. Esto es:

(8) $$EL(a_i) = \sum_{j=1}^{k} L(a_i,A_j)\, P(A_j|X).$$

Entonces, la regla de asignación para el problema de acción-orientada sería: asignar el nuevo elemento a la población para la cual la pérdida condicional esperada sea minimizada:

(9) $$a(X) = a_i \quad \text{si } EL(a_i) = \min\{EL(a_1),\ldots,EL(a_k)\}.$$

Para el caso de una función binaria de pérdida (pérdida cero para asignación correcta y pérdida uno para incorrecta), la regla de asignación de acción orientada (9) se simplifica a la regla de identificación (5).

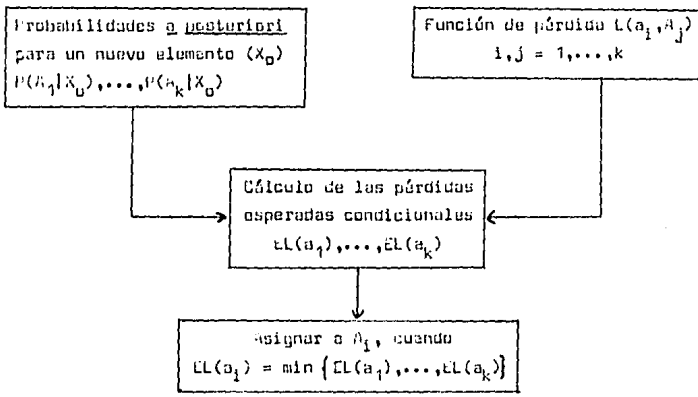


Figura 2: Una ilustración esquemática del proceso por medio del cual un elemento es asignado.

4.3. Un Ejemplo de Estimación de Densidades en el Análisis Discriminante.

Para ilustrar la utilización de la estimación de densidades de probabilidad, se presenta a continuación el uso que pueden tener dentro del análisis discriminante no paramétrico.

En la sección anterior se explicó la utilidad de conocer la función de verosimilitud de las muestras de entrenamiento y su liga con la estimación de las probabilidades a posteriori de pertenencia a cada población. La estimación de esta verosimilitud puede llevarse a cabo median-

te los procedimientos de estimación de densidades de probabilidad presentados con anterioridad.

En concreto, para este trabajo, se seleccionó el método del kernel en el caso multivariado para estimar, dentro de un problema de análisis del discriminante, las probabilidades a posteriori de clasificación. El problema puede describirse de la siguiente manera: se tienen dos formas de clasificar ciertas especies vegetales, herbáceas o arbustos. En muchas ocasiones es claro a cuál de estos dos grupos pertenece cierta especie; sin embargo, existen ocasiones en las que ésto no es así. En base a otras características de estas últimas especies, se desearía clasificarlas dentro de los dos grupos mencionados vía algún procedimiento que tomara en cuenta esta información.

En concreto la información adicional a considerarse sería: a) habitat, b) color de las flores, c) forma de la planta (hierba o arbusto), d) promedio de hojas y e) promedio de flores. Esto para cada especie. Se obtuvo una muestra de entrenamiento de tamaño 30, 11 pertenecientes a la población de herbáceas y 19 a la de arbustos; a cada elemento le fueron medidas las otras características de interés. En base a esta información, se desea clasificar a ocho nuevos sujetos.

Con la finalidad de llevar a cabo este análisis, se escribió un programa de computadora que implementara el método del kernel multivariado para tres tipos de medidas (continuas, nominales y binarias). En el apéndice IV se presentan los resultados que fueron obtenidos para el problema descrito, cuyos datos se presentan en el apéndice III.

Como puede observarse en la siguiente tabla, los resultados son muy satisfactorios para clasificar en la población 1, sin embargo esto no sucede con la población 2. Esto puede deberse a la estructura de

correlación de las variables.

MATRIZ DE ASIGNACION

GRUPO			GRUPO DE ASIGNACION	
TAMAÑO	NO.	TIPO	1	2
11	1	ENTRENAM.	11 100%	0 0%
19	2	ENTRENAM.	12 63.2%	7 36.8%

Para corroborar el problema, se decidió estimar una función discriminante normal. Para hacerlo se utilizó el paquete SAS (1985). A continuación se presenta la matriz de asignación que se obtuvo.

MATRIZ DE ASIGNACION

GRUPO			GRUPO DE ASIGNACION	
TAMAÑO	NO.	TIPO	1	2
11	1	ENTRENAM.	10 90.91%	1 9.09%
19	2	ENTRENAM.	8 42.7%	11 57.8%

Como puede observarse, los resultados son muy similares a los obtenidos vía el procedimiento no paramétrico estimando las densidades con el método del kernel.

Basados en estos resultados, se procedió a clasificar vía kernel los ocho casos de interés. Para estos casos, se obtienen las siguientes dos clasificaciones; una libre y otra restringida a probabilidades

superiores a 0.8.

MUESTRA A CLASIFICAR

ELEMENTO NO.	PROBABILIDADES 1	2	GRUPO ASIG. LIBRE	GRUPO ASIG. RESTRINGIDA
1	0.42	0.58	2	(2)
2	1.0	0.0	1	1
3	1.0	0.0	1	1
4	0.61	0.39	1	(1)
5	0.78	0.22	1	(1)
6	0.99	0.01	1	1
7	0.0	1.0	2	2
8	0.99	0.01	2	2

Es interesante mencionar algunas características de la aplicación del método del kernel:

a) No se hace ningún supuesto sobre el comportamiento probabilístico de ninguna de las variables utilizadas.

b) No solo se trabaja con variables continuas. La variable habitat es binaria y la variable color de la flor es nominal.

c) El problema es multivariado.

d) Se clasifica en términos de probabilidades y no de distancias.

e) Podría estudiarse el comportamiento que generaría el postular diferentes probabilidades a priori.

Estas características señalan las capacidades de adaptación del método a circunstancias que suelen encontrarse en la práctica y que

violan los supuestos con que se construyen otras metodologías. Pensamos entonces que resulta de gran utilidad el contar con una herramienta de este tipo.

APENDICE 1

ESPACIOS DE HILBERT

En esta sección se darán los conocimientos básicos que se necesitan como respaldo a este texto en lo que se refiere a espacios de Hilbert.

El par $(H, \langle \cdot , \cdot \rangle)$ se llama espacio de producto interno si H es un espacio vectorial y $\langle \cdot , \cdot \rangle : H \times H \to R$ satisface las siguientes propiedades:

i) $\langle x,x \rangle \geq 0$ con igualdad si y sólo si $x=0$.

ii) $\langle x,y \rangle = \langle y,x \rangle \quad \forall x,y \in H$.

iii) $\langle \alpha x + \beta y, z \rangle = \alpha \langle x,z \rangle + \beta \langle y,z \rangle \quad \forall \alpha, \beta \in R$ y $\forall x,y,z \in H$.

Por la norma del espacio de producto interno H nos referimos a $\|\cdot\| : H \to R$ definido por

$$\|x\| = \langle x,x \rangle^{1/2}$$

Por simplicidad, cuando el producto interno se sobreentiende, el espacio de producto interno $(H, \langle , \rangle)$ es denotado simplemente como H.

Las propiedades del producto interno y de la norma son los siguientes:

i) $\|x\| \geq 0$ con igualdad si y sólo si $x=0$.

ii) $\|\alpha x\| = |\alpha| \, \|x\|, \ \forall \alpha \in R$ y $\forall x \in H$.

iii) $|\langle x,y \rangle| \leq \|x\| \, \|y\| \quad \forall x,y \in H$ (Cauchy-Schwarz)

iv) $\big| \|x\| - \|y\| \big| \leq \|x + y\| \leq \|x\| + \|y\|, \ \forall x,y \in H$ (igualdad del triángulo).

Cuando H y J son espacios de producto interno, entonces $f: H \to J$ es llamado un operador, y si $J = R$, el operador f se dice que es un funcional. Además, el operador f se dice que es lineal si

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y) \ , \ \forall \alpha, \beta \in R \text{ y } \forall x, y \in H$$

La norma en H no es un funcional lineal; sin embargo, el funcional $f(\cdot) = \langle x, \cdot \rangle$ para una x fija $x \in H$ es lineal.

Sea H un espacio de producto interno. Entonces, una secuencia $\{x^m\} \subset H$ se dice que converge a $x^* \in H$ (que se denota $x^m \to x$), si dada una $\varepsilon > 0$ existe un entero N, tal que $\|x^m - x^*\| \leq \varepsilon$ cuando $m \geq N$. También, una secuencia $\{x^m\} \subset H$ se dice que es una secuencia Cauchy si dada una $\varepsilon > 0$ existe un entero N, tal que $\|x^n - x^m\| \leq \varepsilon$ cuando $n, m \geq N$.

Un espacio de producto interno H se dice que es completo si cada secuencia Cauchy en H converge a un miembro de H. Un espacio de producto interno completo es llamado un Espacio de Hilbert.

Ejemplo 1. El espacio vectorial $L^2(a,b) = \{f: (a,b) \to R: f \text{ es integrable en Lebesgue cuadrado}\}$ con producto interno

$$\langle f, g \rangle = \int_a^b f(t)g(t)dt$$

es un espacio de Hilbert infinito dimensional si identificamos a los miembros de $L^2(a,b)$ que se diferencian en un conjunto de Lebesgue con medida cero.

Ejemplo 2. (Espacios de Sobolev en la recta de los reales). Para $s = 1, 2, \ldots$ el espacio vectorial,

$$H^s(-\infty,\infty) = \left\{ f:\ f^{(j)} \in L^2(-\infty,\infty) \text{ para } j = 0,\ldots,s \right\}$$

con producto interno

$$\langle f,g \rangle = \sum_{j=0}^{s} \langle f^{(j)}, g^{(j)} \rangle_{L^2(-\infty,\infty)}$$

es un espacio de Hilbert infinito dimensional. Podemos pensar en $H^0(-\infty,\infty)$ como $L^2(-\infty,\infty)$.

Ejemplo 3. (Espacios de Sobolev restrictos). Sea (a,b) un intervalo finito. Entonces el espacio vectorial

$$H_0^s(a,b) = \left\{ f:\ f^{(j)} \in L^2(a,b),\ j = 0,\ldots,s \text{ y } f^{(j)}(a)=f^{(j)}(b)=0\ ,\right.$$

$$\left. j = 0,\ldots,s-1 \right\} \text{ con producto interno}$$

$$\langle f,g \rangle = \langle f^{(s)}, g^{(s)} \rangle_{L^2(a,b)}$$

es un espacio de Hilbert infinito dimensional.

Un espacio de Hilbert $H(T)$ de funciones definidas en el conjunto T se dice que es un Espacio de Hilbert Reproductor de Kernel (EHRK) si existe un funcional reproductor de Kernel $K(\cdot,\cdot)$ definido en $T \times T$ con las siguientes propiedades:

i) $K(\cdot,t) \in H(T)\ ,\ \forall\, t \in T$

ii) $f(t) = \langle f, K(\cdot,t) \rangle\ ,\ \forall f \in H(T)$ y $\forall\, t \in T$.

Como ejemplos de EHRK tenemos el espacio de Sobolev $H^s(-\infty,\infty)$ dado en el ejemplo 2 y el espacio de Sobolev restricto $H_0^s(a,b)$ dado en el ejemplo 3. (Para mayor información sobre el tema consultar a R.A. Tapia y J.R. Thompson (1978).)

APENDICE II

SPLINES

a) Splines Polinomiales.

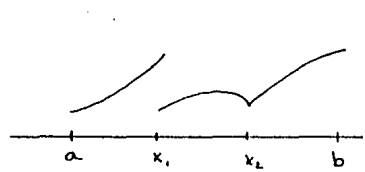
Para poder dar la definición de spline es necesario definir el concepto de polinomios por secciones o polinomios por pedazos.

Polinomios por pedazos: Sea $a=x_0<\cdots<x_{k+1}=b$ y denótese a $\Delta=\{x_i\}$ con $i=0,\ldots,k+1$. El conjunto Δ particiona el intervalo $[a,b]$ en $k+1$ subintervalos, $I_i=[x_i,x_{i+1})$, $i=0,\ldots,k-1$, y $I_k=[x_k,x_{k+1}]$. Dado un entero positivo m, sea

$$(1)\qquad \left\{ f:\ \text{existen polinomios } PP_m(\Delta)=p_0,\ldots,p_k \text{ en } P_m \text{ con } f(x)=p_i(x) \text{ para } x\in I_i,\ i=0,\ldots,k \right\}$$

llamamos a $PP_m(\Delta)$ el espacio de polinomios por pedazo de orden m con nodos en $x_1,\ldots,x_k$.

Se tiene entonces que un elemento $f\in PP_m(\Delta)$ consiste de $k+1$ pedazos polinomiales. Se puede ver un típico ejemplo de los polinomios por pedazos de orden tres con dos nodos:



Claramente hemos ganado flexibilidad al ir de polinomios a polinomios por pedazos, pero hemos perdido una propiedad importante: los polinomios por pedazos no son necesariamente suaves. Como se ve en la

figura anterior, incluso pueden ser discontinuos.

Funciones de Spline.

Para mantener la flexibilidad de los polinomios por pedazos mientras que se logra cierto grado de suavidad global, se define la clase de funciones de spline con nodos simples.

Sea Δ una partición del intervalo $[a,b]$ como en la definición de polinomios por pedazos y sea m un entero positivo. Sea

$$S_m(\Delta) = PP_m(\Delta) \cap C^{m-2}[a,b] , \quad m \geq 2 \tag{2}$$

donde $PP_m(\Delta)$ es el espacio de polinomios por pedazos definida en (1). Se le llama a $S_m(\Delta)$ el espacio de splines polinomiales de orden m con nodos en los puntos $x_1,\ldots,x_k$.

Es fácil definir varias clases de polinomios por pedazos relacionadas con distintos grados de suavidad entre los pedazos. A estos espacios de funciones los llamamos splines polinomiales. Estos tienen las siguientes características:

1. Los espacios de splines polinomiales son espacios lineales finito-dimensionales con bases muy convenientes.

2. Los splines polinomiales son funciones relativamente suaves.

3. Los splines polinomiales son fácilmente almacenados, manejados y evaluados en una computadora digital.

4. Las derivadas y antiderivadas de los splines polinomiales son, a su vez, splines polinomiales cuyas expansiones se pueden encontrar en una computadora.

5. Los splines polinomiales tienen buenas propiedades del cero, análogas a aquellas para polinomios.

6. Varias matrices que surgen naturalmente en el uso de splines en la teoría de la aproximación y en el análisis numérico, tienen convenientes propiedades de signo y de determinantes.

7. La estructura de signo y la forma de un spline polinomial puede estar relacionado a la estructura de signo de sus coeficientes.

8. Toda función continua en el intervalo [a,b] puede ser aproximada arbitrariamente bien por splines polinomiales fijando el orden m, siempre que se permita tener un número suficiente de nodos.

9. Tasas precisas de convergencia pueden darse para la aproximación de funciones suaves por splines; no sólo son aproximadas las funciones a un orden alto, sino que también sus derivadas son simultáneamente bien aproximadas.

10. Splines de orden bajo son muy flexibles y no muestran las oscilaciones que comunmente se asocian con polinomios.

b) Splines Exponenciales.

Dado cualquier $\alpha_1 < \cdots < \alpha_m$, sea U_m = extención (span)$\{e^{\alpha_1 x}, \ldots, e^{\alpha_m x}\}$ Como los elementos del espacio U_m son usualmente llamados polinomios exponenciales, es natural llamarle a $S(U_m ; \Delta)$ el espacio de splines exponenciales. Aquí nos referimos a funciones que son exponenciales por pedazos.

Para profundizar sobre este tema, consultar a Schumaker (1981).

APENDICE III

DATOS DEL EJEMPLO

MUESTRA DE ENTRENAMIENTO

ESPECIE	HABITAT	FLOR	FORMA	HOJAS	RFLOR
Chamaecrista Cham	1	2	1	4.009	3.764
Crotalaria Inca	0	2	1	3.688	3.090
Crotalaria Inca	1	2	1	3.382	3.054
Lantana Cama	0	3	1	3.745	4.090
Lantana Cama	1	3	1	3.462	4.367
Porophyllum Numm	0	1	1	3.798	3.518
Porophyllum Numm	1	1	1	3.729	3.747
Turnera Ulmi	0	2	1	3.819	5.204
Turnera Ulmi	1	2	1	4.051	4.648
Paullinia Tome	0	1	1	3.914	2.666
Serjania Mace	0	1	1	3.722	3.384
Panicum Repe	0	4	2	3.684	4.000
Panicum Repe	1	4	2	3.705	0.000
Bidens Pilo	0	1	2	3.697	5.562
Bidens Pilo	1	1	2	3.786	5.200
Iresince Celo	0	1	2	3.401	4.080
Indigofera Hart	0	5	2	3.653	5.796
Pectis Satu	0	2	2	3.589	4.178
Erigeron Myri	0	1	2	4.157	4.426
Erigeron Myri	1	1	2	4.184	3.824
Hydrocotile Bona	0	1	2	3.459	4.000
Phyla Nodi	0	7	2	3.543	3.932
Phyla Nodi	1	7	2	3.375	3.469
Cardiospermun Hali	0	1	2	3.578	3.591
Cardiospermun Hali	1	1	2	3.965	3.657
Macroptilium Atro	0	6	2	3.926	4.470
Macroptilium Atro	1	6	2	3.655	3.973
Metastelma Prin	0	1	2	3.708	3.747
Metastelma Prin	1	1	2	3.850	3.797

MUESTRA A CLASIFICAR

ESPECIE	HABITAT	FLOR	HOJAS	TALLO
Chiococca Alba	0	1	3.789	4.000
Randia Laet	0	1	3.774	5.395
Randia Laet	1	1	3.802	4.978
Cyperus Arti	0	4	3.364	3.529
Cyperus Arti	1	4	3.466	5.500
Thoradendron Tama	0	3	3.026	8.000
Thoradendron Tama	1	3	4.080	0.000
Opuntia Stri	0	2	3.719	4.181

APENDICE IV

RESULTADOS DEL EJEMPLO

1. MEDIA POR GRUPO

GRUPO		NUMERO DE LA VARIABLE			
NO.	TIPO	HABITAT	FLOR	HOJAS	PELO
1	ENTRENAM.	0.45	1.82	3.75	3.79
2	ENTRENAM.	0.42	2.74	3.72	3.99

2. DESVIACION ESTANDAR POR GRUPO

GRUPO		NUMERO DE LA VARIABLE			
NO.	TIPO	HABITAT	FLOR	HOJAS	PELO
1	ENTRENAM.	0.52	0.75	0.70	0.78
2	ENTRENAM.	0.51	2.35	0.72	1.16

3. ESTRUCTURA DE GRUPO

GRUPO NO.	TAMAÑO	PROBABILIDAD A PRIORI	PARAMETRO SUAVIZAMIENTO
1	11	0.5	0.5
2	19	0.5	0.5

4. GRUPO 1. MUESTRA DE ENTRENAMIENTO

ELEMENTO NO.	GRUPO ASIGNACION	PROBABILIDADES A POSTERIORI 1	2
1	1	0.91	0.09
2	1	0.91	0.09
3	1	0.98	0.02
4	1	0.99	0.01
5	1	1.00	0.00

ELEMENTO NO.	GRUPO ASIGNACION	PROBABILIDADES A POSTERIORI 1	2	
6	(1)	0.79	0.21	ASIGNACION CON DUDA
7	(1)	0.76	0.23	CUANDO PROB. < 0.8
8	1	0.93	0.07	
9	1	0.95	0.05	
10	1	1.00	0.00	
11	1	1.00	0.00	

5. GRUPO 2. MUESTRA DE ENTRENAMIENTO

ELEMENTO NO.	GRUPO ASIGNACION	PROBABILIDADES A POSTERIORI 1	2	
1	(2)	0.45	0.55	
2	2	0.16	0.84	
3	(2)	0.21	0.78	
4	2	0.13	0.87	
5	2	0.06	0.94	
6	(1)	0.56	0.43	
7	(2)	0.47	0.53	
8	(1)	0.59	0.41	
9	(1)	0.62	0.38	
10	(1)	0.52	0.48	ASIGNACION CON DUDA
11	(2)	0.49	0.51	CUANDO PROB. < 0.6
12	(1)	0.54	0.46	
13	(1)	0.52	0.48	
14	(1)	0.61	0.39	
15	(1)	0.66	0.32	
16	(1)	0.64	0.36	
17	(1)	0.63	0.37	
18	(1)	0.59	0.40	
19	(1)	0.64	0.35	

6. MATRIZ DE ASIGNACION

GRUPO			GRUPO DE ASIGNACION	
TAMAÑO	NO.	TIPO	1	2
11	1	ENTRENAM.	11 100%	0 0%
19	2	ENTRENAM.	12 63.2%	7 36.8%

7. MATRIZ DE ASIGNACION CON DUDA

GRUPO				GRUPO DE ASIGNACION	
TAMAÑO	NO.	TIPO	DUDA	1	2
11	1	ENTRENAM.	2 10.2%	9 81.8%	0 0%
19	2	ENTRENAM.	16 84.2%	0 0%	3 15.8%

BIBLIOGRAFIA

Boneva, L., Kendall, D. and Stefanov, I. (1971), "Spline transformations: three new diagnostic aids for the statistical data analyst", Journal of the Royal Statistical Society, Vol.33, 1-70.

Bickel, P.J., and Rosenblatt, M. (1973). "On some global measures of the deviations of density function estimates", Annals of statistics 1: 1071-95.

Brunk, H.D. (1976), "Univariate density estimation by orthogonal series", TR.M.51, Dept. of Statistics, Oregon State University, Corvallis, Oregon.

Crain, B.R. (1976), "Matrix density estimation", Commun. Statistic, Vol. A5(1), 89-96.

De Montricher, G.F., Tapia, R.A., and Thompson, J.R. (1975), "Nonparametric maximum likelihood estimation of probability densities by penalty function methods", The Annals of Statistics, Vol. 3, 6, 1329-1348.

Duin, R.P.W. (1976), "On the choice of smoothing parameters for Parzen estimators of probability density functions", IEEE Transactions on Computers, C-25, 1175-1179.

Fisher, R.A. (1922), "On the mathematical foundations of theoretical statistics", Philosophical Transactions of the Royal Society of London, Series A222, 309-368.

Friedman, J.H., and Tukey, J.W. (1974), "A projection persuit algorithm for exploratory data analysis", IEEE Transactions on Computers, C23 881-890.

Friedman, J.H., Stuetzle, W., and Schroeder, A. (1984), "Projection persuit density estimation", Journal of the American statistical Association, 79, 387, Theory and Methods Section.

Good, I.J., and Gaskins, R.A. (1971), "Nonparametric roughness penal-

ties for probability densities", Biometrika 58, 255-277.

Gosset, W.S. (1908), "The probable error of a mean", Biometrika, 6, 1-25.

Habbema, J.D.F., Hermans, J., and Van Den Broek, K. (1974), "A stepwise discriminant analysis program, using density estimation", in G. Bruckmann (Ed.), COMPSTAT, Proceedings in Computational Statistics, Physica Verlag, Wien.

Hermans, J., Habbema, J.D.F., Kasanmoentalib, T.S.H., and Raatgever, J.W. (1982), "Manual for the ALLOC 80 discriminant analysis program", Dept. of Medical Statistics, Univ. of Leiden, Netherlands.

IMSL (1983). IMSL Libraries, Edition 10, International Mathematical and Statistical Libraries, Inc., Houston, Texas.

Kim, Bock Ki, and Van Ryzin, J. (1974), "Uniform consistency of a histogram density estimator and model estimation", MRC Report 1494.

Kimeldorf, G.S. and Jahba G. (1970), "A correspondence between bayesian estimation on stochastic processes and smoothing by splines", Annals of Mathematical Statistics, 41, 495-502.

Kronmal, R. and Tarter, M. (1968), "The estimation of probability densities and cumulatives by Fourier series methods", Journal of the American Statistical Association, 63, 925-952.

Mood, A.M., Graybill, F.A., and Boes, D.C. (1974), Introduction to the Theory of Statistics, McGraw Hill International Book Company.

Morrison, D.J. (1976), Multivariate Statistical Methods, McGraw Hill Book Company.

Nadaraya, E.A. (1965), "On nonparametric estimates of density functions and regression curves", Theory of Probabilities and Its Aplications 10: 186-90.

Orear, J. and Cassel, D. (1971), "Aplication of statistical inference to physics", Foundation of Statistical Inference, eds. V.P. Godambe

and D.R. Sprott, 780-782, Toronto: Holt, Rinehart and Winston.

Parzen, E. (1962), "On estimation of a probability density function and mode", Annals of Mathematical Statistics, 33, 1065-1076.

Pearson, K. (1936), "Method of moments and method of maximum likelihood", Biometrika, 28, 34-59.

Rosenblatt, M. (1956), "Remarks on some nonparametric estimates of a density function", Annals of Mathematical Statistics, 27, 832-835.

SAS Institute Inc. (1985), SAS User's Guide: Statistics, Version 5 Edition, Cary, NC: SAS Institute Inc.

Schuster, E.F. (1970), "Note on the uniform convergence of density estimates", Annals of Mathematical Statistics 41: 1347-48.

Scott, D.W. (1976), "Nonparametric probability density estimation by optimization theoretic techniques", Doctoral Dissertation at Rice University, Houston, Texas.

Scott, D.W., Tapia, R.A., and Thompson, J.R. (1977), "Nonparametric probability density estimation by discrete maximum penalized likelihood criteria", (submitted for publication).

Schumaker, L. (1972), "Splines and histograms", Mathematics Research Center Report 1273, University of Wisconsin, Madison.

Silverman, B.W. (1981), "On the estimation of a probability density function by the maximum penalized likelihood method", Mathematics Research Center University of Wisconsin, Technical Summary Report No. 2228.

Tapia, R.A. and Thompson, J.R. (1978), Nonparametric Probability Density Estimation, The John Hopkins University Press, Baltimore, Maryland.

Van Ryzin, J. (1969), "On strong consistency of density estimates", Annals of Mathematical Statistics 40: 1765-72.

Wahba, G. (1977), "Optimal smoothing of density estimates", Classification and Clustering, Academic Press, New York, San Francisco, London.

Watson, G.S. (1969), "Density estimation by orthogonal series", Annals of Mathematical Statistics, 40, 1496-1498.

Wegman, E.J. (1972), "Nonparametric probability density estimation: a comparison of density estimation methods", Journal of Statistics Comput. Simul., 1, 225-245.

Woodroofe, M. (1970), "On choosing a delta-sequence", Annals of Mathematical Statistics 41: 1665-71.