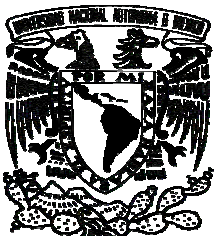


UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

POSGRADO EN CIENCIAS MATEMÁTICAS

FACULTAD DE CIENCIAS

APLICACIÓN DEL TEMPLADO SIMULADO A
UN PROBLEMA DE ESTADÍSTICA BAYESIANA

T E S I S

QUE PARA OBTENER EL GRADO ACADÉMICO DE

MAESTRO EN CIENCIAS

P R E S E N T A

MIGUEL ANGEL LÓPEZ DÍAZ

DIRECTOR DE TESIS: DR. EDUARDO GUTIÉRREZ PEÑA

MÉXICO, D.F. FEBRERO, 2008

Para mi esposa e hijos,
Ivonne, Miguel Angel, Diego y Rafael

Índice general

Prefacio **3**

1. Optimización **7**

1.1. Introducción 7

1.1.1. Optimización numérica 8

1.1.2. Optimización estocástica 9

1.2. Principios de los algoritmos metaheurísticos más exitosos 10

1.2.1. Templado simulado 10

1.2.2. Búsqueda tabú 10

1.3. El templado simulado visto más de cerca 11

1.3.1. El algoritmo 11

1.3.2. Consideraciones para su implementación 13

1.4. Ejemplos 14

1.4.1. Maximizando una función $f : \mathbb{R} \to \mathbb{R}$ 15

1.4.2. Minimizando una función $f : \mathbb{R}^2 \to \mathbb{R}$ 17

1.4.3. Cubierta de vértices mínima para un gráfica 19

1.5. Una comparación 19

2. Teoría de la Decisión en Ambiente de Incertidumbre **23**

2.1. Introducción 23

2.2. Estructura de un problema de decisión en ambiente de incertidumbre 25

2.3. Solución intuitiva a un problema de decisión 25

2.4. Axiomas de coherencia y cuantificación 27

2.5. Probabilidad como grado de creencia 29

2.6. Maximización de la utilidad esperada 31

2.7. Otros criterios de decisión en situaciones de incertidumbre 33

2.7.1. Criterio minimax 34
2.7.2. Criterio condicional 35
2.7.3. Criterio de decisión de Bayes 36

3. Estadística Bayesiana **41**
3.1. Introducción 41
3.2. Antecedentes históricos 42
3.3. Algunas objeciones al método de inferencia clásico 43
3.4. ¿Por qué la inferencia Bayesiana? 44
3.5. El enfoque Bayesiano 45

4. Selección de Modelos **47**
4.1. Introducción 47
4.2. La selección de modelos en el enfoque $\mathcal{M}$-cerrado 48
4.3. La selección de modelos en el enfoque $\mathcal{M}$-abierto 52

5. Un enfoque Bayesiano para la Comparación de Modelos de Regresión **55**
5.1. El Modelo de regresión semiparamétrico 56
5.2. Un criterio predictivo para la selección de modelos 59

6. Aplicaciones **63**
6.1. Implementación del criterio vía templado simulado 63
6.2. Ejemplo 1. Datos de Hald 64
6.3. Ejemplo 2. Reducción en el desarrollo intelectual de niños de 7 años con exposición prenatal a plomo 65

Conclusiones y Sugerencias **69**

Apéndice A. Código en R de los programas usados **71**

Apéndice B. El problema de reportar inferencias como problema de decisión **89**

Referencias

Bibliografía

Prefacio

A diferencia de los métodos estadísticos clásicos, los métodos Bayesianos no se reducen a operar con la información empíricamente conseguida sino que la combinan con los criterios *a priori* que posee el investigador, nacidos tanto de estudios previos como de reflexiones racionales y juicios razonablemente conformados. Como resultado de tal integración, que se realiza por conducto del teorema de Bayes, se obtiene una llamada visión *a posteriori* que constituye la base de las inferencias subsiguientes. A pesar de que los métodos Bayesianos conforman una nueva forma de hacer inferencias, su desarrollo se ha visto limitado por la complejidad del cálculo matemático. Para nuestra fortuna, el avance en materia de cómputo a partir de los años 80, ha venido a resolver en gran parte este problema.

Por otro lado, las técnicas de regresión se cuentan entre los métodos más utilizados en la estadística aplicada. Dada una variable de respuesta $\boldsymbol{Y}$ y un conjunto de covariables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$, es de interés estimar una relación funcional supuesta entre la variable de respuesta y las covariables. Un problema que surge al construir el modelo de regresión es decidir qué covariables incluir. Este problema, conocido como el problema de selección de variables, se puede pensar como un problema de selección de modelos, donde cada modelo considerado corresponde a los distintos subconjuntos de covariables. Claramente este problema es de interés cuando el número de covariables es grande ($p > 10$).

El problema de la selección de modelos se basa primordialmente en la maximización de una utilidad esperada; sin embargo, la complejidad de las expresiones involucradas hace imposible en muchas ocasiones que los máximos se puedan obtener de manera analítica. Las primeras técnicas numéricas desarrolladas solucionan en parte este problema; sin embargo, en ocasiones es difícil verificar las condiciones que se piden sobre la función a optimizar. Las técnicas computacionales más recientes están asociadas con un intenso cálculo computacional. Uno de estos métodos es el llamado algoritmo de templado simulado desarrollado por Kirkpatrick *et al.* (1983), y cuyo nombre se deriva del proceso físico de llevar a altas temperaturas y después enfriar lentamente una sustancia cristalina. Es un hecho físico que si el proceso de enfriamiento es lo suficientemente lento, entonces la estructura molecular será extremadamente rígida y esto corresponde a un estado de energía mínimo. El templado simulado imita este proceso en el cual la energía es sustituida por la función objetivo y las configuraciones mole-

culares de la sustancia líquida por los parámetros. Este algoritmo ha tenido innumerables aplicaciones en distintas áreas, principalmente en diseño de circuitos electrónicos, procesamiento de imágenes y transporte.

En el presente trabajo de tesis abordamos el problema de la selección de modelos desde una perspectiva Bayesiana, bajo el siguiente esquema. Consideramos una variable de interés y un conjunto de posibles covariables suficientemente grande, cuya relación funcional es muy compleja o desconocida. Partimos de una clase de modelos potenciales mucho más amplia como lo es la clase de modelos de regresión semiparamétricos, y aplicamos el criterio Bayesiano predictivo semiparamétrico propuesto por Gutiérrez-Peña (1997), para elegir el mejor modelo predictivo que ajusta a los datos observados. Bajo estas consideraciones el problema de selección de modelos se convierte en un problema de optimización no trivial, pues habrá que comparar 2^p modelos. Para lidiar con el problema de la optimización, aprovechamos el templado simulado que ha mostrado ser exitoso en otras áreas.

El esquema de trabajo desarrollado es como sigue. En el primer capítulo hacemos una revisión de optimización estocástica, analizamos sus ventajas y desventajas haciendo énfasis en el templado simulado. Finalizamos mostrando algunas de las aplicaciones más elementales del templado simulado. Dado que el criterio a usar está basado en la solución de un problema de decisión, en el segundo capítulo hacemos un breve resumen de la teoría de decisión bajo incertidumbre, se presentan y defienden determinados principios de comportamiento coherente, y de estos se deduce la necesidad de contar en primer lugar, con una medida de probabilidad que describa el conocimiento con los que cuenta el tomador de decisiones. Y en segundo lugar, contar con una función de utilidad que sea capaz de describir sus preferencias, y si finalmente aceptamos que el tomador de decisiones satisface los principios de comportamiento coherente, entonces elegirá aquella decisión que maximice su utilidad esperada. Analizamos y criticamos algunos otros de los criterios de decisión en situaciones de incertidumbre expuestos en la teoría. En el tercer capítulo hacemos una descripción muy general del enfoque Bayesiano, de manera simultánea analizamos las ventajas y desventajas que presentan con respecto a los métodos clásicos. En el cuarto capítulo hacemos una revisión breve de los métodos de selección de modelos desde el punto de vista de la teoría de la decisión, mostramos los criterios que se usan desde las perspectivas abierta y cerrada propuestas por Bernado y Smith. Y finalmente, en el último capítulo damos el criterio Bayesiano predictivo semiparamétrico propuesto por Gutiérrez-Peña y lo aplicamos al problema de selección de variables en el modelo de regresión.

Finalmente, quiero hacer un reconocimiento a la Universidad Autónoma del Estado de México por el financiamiento otorgado para poder concluir los estudios de maestría y elaborar el presente trabajo de tesis. Agradezco al Dr. Eduardo Gutiérrez Peña por su dedicación y paciencia en la dirección y revisión esta tesis. Agradezco también a la Dra. Silvia Ruíz Velasco por permitirme trabajar con los

datos del utilizados en la sección 3 del capítulo 6, al Dr. Alberto Contreras, al Dr. Ramsés Mena y al Dr. Raúl Rueda por el tiempo empeñado en la revisión y las valiosas aportaciones realizadas.

Capítulo 1

Optimización

En estadística muchos problemas tales como estimación y selección de modelos se pueden plantear en términos de un problema de optimización. En el caso particular de la estadística Bayesiana, la solución a un problema de decisión en ambiente de incertidumbre consiste en la maximización de la utilidad esperada, posiblemente sobre un espacio de acciones muy grande o con una estructura muy compleja. En este presente capítulo analizamos el problema de la optimización y algunos enfoques para resolverlo. Hacemos énfasis en los métodos estocásticos dentro de los cuales los más usados son los metaheurísticos. Este capítulo está basado primordialmente en Fouskakis & Draper (2002) así como en Dréo *et al.* (2006) y Bernardo (1992).

1.1. Introducción

Todos los días ingenieros, arquitectos, actuarios y muchos otros profesionales se enfrentan a problemas de optimización. Por ejemplo, un ingeniero en electrónica que está encargado de desarrollar un nuevo circuito electrónico de manera que las conexiones entre sus componentes sean lo más cortas posible, o un corredor de bolsa encargado de diseñar un portafolio de inversiones de manera que las utilidades sean lo más altas.

La optimización es una herramienta importante en muchas áreas, pero en particular en el área de toma de decisiones. Para usarla primero debemos identificar una *función objetivo,* cuyo valor es una cantidad numérica que mide de alguna manera el desempeño del modelo o del sistema que se está utilizando. Esta función bien puede ser una ganancia, una pérdida o cualquier combinación de cantidades que puedan ser representadas numéricamente. La función objetivo depende de ciertas características del modelo o sistema llamadas *configuraciones*.

El problema de la optimización es hallar aquella configuración que optimice la función objetivo. Así, por ejemplo, en el caso de maximización, si S denota al conjunto de todas las configuraciones del

modelo y $f : S \to \mathbb{R}$ es la función objetivo, el problema es hallar un configuración llamada *configuración óptima* c^* que cumpla:

$$f(c^*) \geq f(c) \text{ para toda configuración } c \in S.$$

A lo largo del tiempo se han desarrollado técnicas para resolver el problema de la optimización, pero muchas de ellas imponen condiciones sobre la función objetivo e incluso sobre las configuraciones que son en ocasiones difíciles de verificar o de cumplir. También, en muchas aplicaciones la función objetivo puede ser muy compleja. Por estas razones, entre otras, muchos investigadores se han dado a la tarea de desarrollar métodos numéricos.

1.1.1. Optimización numérica

Desde el desarrollo de las computadoras a partir de los años 70, se han implementado algoritmos tales como el de *Newton* y el del *Gradiente,* por mencionar sólo algunos, para dar solución numérica al problema de la optimización. Sin embargo, no es tan fácil garantizar que la función objetivo sea hasta dos veces diferenciable o más aún que nuestro problema tenga soluciones en un espacio continuo. Si, por otro lado, las configuraciones deben cumplir ciertas restricciones, entonces hay otras herramientas tales como el método de *Nelder (simplex)* o el de los *Multiplicadores de Lagrange,* pero para aplicarlos debemos garantizar que la función objetivo sea, por ejemplo, convexa. Para evitar pedir condiciones sobre la función objetivo o sobre las configuraciones se han desarrollado algunas otras herramientas tales como los algoritmos de búsqueda; sin embargo, en muchas ocasiones al igual que en muchos otros algoritmos, estos llevan a encontrar óptimos locales o soluciones que dependen de la configuración inicial. Por ejemplo, en el caso de la búsqueda descendente se empieza en una aproximación inicial a la solución la cual es llamada configuración inicial c_0 que, en principio, puede ser elegida al azar. Se prueba entonces con una nueva configuración c_1, que es una modificación elemental de la inicial (llamada a menudo *movimiento),* y que es *vecina* de ésta. Posteriormente se evalúa la función objetivo en ambas configuraciones y se comparan: si el movimiento reduce la función objetivo se acepta a c_1 como nueva configuración inicial para la siguiente iteración. En caso contrario, regresamos a nuestra configuración inicial c_0 y probamos con un nuevo movimiento; este proceso se repite de manera iterativa hasta que ningún movimiento reduzca la función objetivo. La figura 1.1 muestra que este principio *descendente* en general no lleva a un mínimo global; sin embargo, la configuración final constituye la mejor solución obtenida a partir de la configuración inicial. Para contrarrestar este problema, se puede aplicar este algoritmo varias veces partiendo de distintas configuraciones iniciales y tomando como solución aquella que sea mínima; sin embargo, este procedimiento aumenta el tiempo de cómputo y puede no hallar la configuración óptima.

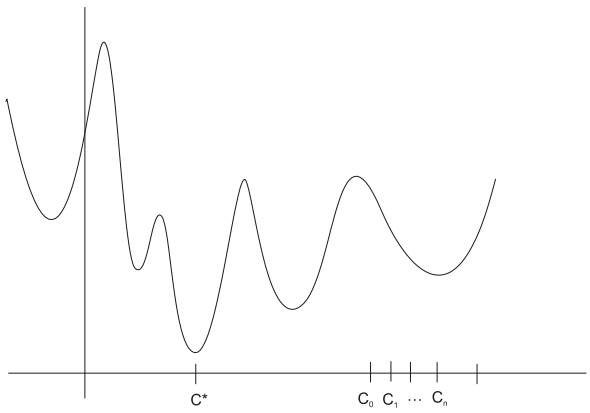


Figura 1.1: Cuando la función objetivo tiene muchos óptimos locales, un algoritmo como el de búsqueda descendente resultaría ineficiente.

1.1.2. Optimización estocástica

Otra aproximación para la solución del problema de optimización ha sido la *optimización estocástica,* en la cual la búsqueda de una solución óptima involucra herramientas *heurísticas.* Los algoritmos llamados *metaheurísticos,* disponibles desde la década de los años 80 y cuyo objetivo es resolver problemas de optimización extremadamente complejos *tan bien* como sea posible, marca una unificación a la solución de problemas del tipo discreto o continuo y tiene las siguientes características:

- Son aleatorios, lo que permite atacar la explosión combinatoria de todas las posibilidades.
- Generalmente tienen su origen en problemas discretos, pero con la ventaja en el caso discreto de ir directo, *i.e.* no resolver problemas intermedios como calcular un gradiente de la función objetivo.
- Están inspirados por analogías con otras ciencias como Física, Biología y Etología.
- Todos comparten las mismas desventajas: dificultades para el ajuste de parámetros iniciales y mucho tiempo de computación.

Contrario a otro tipo de algoritmos, los algoritmos metaheurísticos tienen la capacidad para no quedar *atrapados* en un mínimo local y esto se debe simplemente a que, de vez en cuando, *autorizan* tomar una decisión que no sea la óptima; en otras palabras, aceptan poder pasar a una situación peor que la actual. Este mecanismo utilizado para controlar las degradaciones está en la base de todos los métodos metaheurísticos y ha mostrado ser benéfico pues permite modificar la configuración actual para poder explorar algunas otras configuraciones que en principio podrían ser más prometedoras. Dentro

de los algoritmos metaheurísticos más usados y que han mostrado su eficacia en muchas aplicaciones podemos mencionar al algoritmo de búsqueda *tabú* y *templado simulado*.

1.2. Principios de los algoritmos metaheurísticos más exitosos

1.2.1. Templado simulado

En 1953 Metropolis y sus colaboradores desarrollaron un algoritmo para simular cómo se alcanzaba el equilibrio térmico en un sistema de partículas que estaban sometidas a una cierta temperatura. El algoritmo a cada paso elige de manera aleatoria una de las partículas en el sistema, la cual es *movida* de su posición original, posteriormente se calcula el cambio resultante de energía ΔE de todo el sistema. Si $\Delta E \leq 0$ se acepta el movimiento de la partícula y la configuración actual del sistema se convierte en la configuración inicial para la siguiente iteración. De otro modo si el $\Delta E > 0$ el cambio se acepta de acuerdo con una probabilidad $P(\Delta E) = \exp\left(-\frac{\Delta E}{K_B T}\right)$, donde T representa la temperatura a la que se encuentra el sistema y K_B es la constante de *Boltzman.* Repitiendo estos pasos básicos es posible simular el movimiento de las partículas antes de alcanzar el equilibrio térmico. Posteriormente Kirkpatrick, Gelatt & Vecchi (1983) desarrollaron el templado simulado, un algoritmo basado en la idea de Metropolis para resolver problemas de optimización combinatoria. La idea de este algoritmo consiste en reemplazar la función objetivo por la energía y las configuraciones por un conjunto de parámetros $\{x_i\}$, esto logra generar una población de configuraciones del problema de optimización a una temperatura, la cual no es más que un parámetro de control y que se expresa en las mismas unidades de la función objetivo. El templado simulado consiste entonces en que una vez alcanzado el equilibrio térmico a una temperatura dada (*i.e* en principio ya no es posible realizar un cambio en la configuración que haga reducir la función objetivo) se disminuye la temperatura lentamente hasta que el sistema se *enfría* y no ocurren más cambios. Este algoritmo ha tenido innumerables aplicaciones en distintas áreas, principalmente en diseño de circuitos electrónicos, procesamiento de imágenes y transporte.

1.2.2. Búsqueda tabú

El método de búsqueda tabú, fue descrito por primera vez en Glover (1986) y su principal característica es que está inspirado en el proceso de la memoria humana. El método tabú es, desde este punto de vista, opuesto al templado simulado el cual no utiliza memoria para nada y es por lo tanto incapaz de aprender de experiencias pasadas. Por otra parte, modelar la memoria introduce múltiples grados de libertad y esto hace muy complicado dar desde el punto de vista matemático un tratamiento riguroso, aún en opinión del propio autor. Al igual que el templado simulado, el método tabú empieza

con una configuración inicial (elegida al azar), la cual es actualizada a lo largo de iteraciones sucesivas. En cada iteración el proceso de pasar de una configuración c_0 a una nueva configuración c_1 consta de dos etapas:

1 Se empieza por construir un conjunto $V(c_0)$ de configuraciones vecinas de c_0, *i.e.* el conjunto de todas las configuraciones a las que se pueden acceder a través de un único movimiento elemental de c_0. Si $V(c_0)$ es muy extenso, se puede reducir seleccionado por ejemplo al azar.

2 Se calcula el valor de la función objetivo f en cada una de las configuraciones de $V(c_0)$, la configuración siguiente c_1 en la serie de soluciones será aquella en la que f tome el valor mínimo. Notemos que esta decisión es tomada aún en el caso de que $f(c_0) < f(c_1)$. Debido a esta característica el método tabú evita quedarse atrapado en un mínimo local.

El proceso descrito anteriormente corre el riesgo de ciclarse, pues en la siguiente iteración se puede elegir la configuración previa. Para evitar esta situación el método consulta y actualiza una lista de movimientos prohibidos, la *lista tabú* que contiene los m movimientos $(c_1 \to c_0)$ que son los últimos m movimientos opuestos a los ya realizados $(c_0 \to c_1)$.

Este algoritmo modela un tipo elemental de memoria de *corto plazo* de las soluciones elegidas recientemente. Dos mecanismos adicionales llamados *intensificación* y *diversificación*, son a menudo implementados para equipar al algoritmo con un tipo de memoria de largo plazo.

Para cierto tipo de problemas, el método tabú ha dado excelentes resultados; además en su forma más básica comprende menos parámetros de ajuste que el templado simulado, lo que lo hace más fácil de usar. Sin embargo, los procesos adicionales de intensificación y diversificación lo vuelven muy complejo.

1.3. El templado simulado visto más de cerca

Desde que fue propuesto por Kirkpatrick, Gelatt y Vecchi en 1983, el algoritmo de templado simulado ha mostrado su efectividad en muchos campos como: diseño de circuitos electrónicos, procesamiento de imágenes, optimización combinatoria, etc. En esta sección hacemos una presentación del algoritmo y proponemos algunas consideraciones para su implementación.

1.3.1. El algoritmo

El templado simulado es un método para aproximar el mínimo de una función. Si S denota el conjunto (en principio *finito*) de todas las posibles soluciones, la tarea es hallar una configuración

$c^* \in S$ que satisfaga

$$f(c^*) \leq f(c) \qquad \text{para toda } c \in S.$$

El algoritmo opera de manera iterativa, empezando desde una configuración inicial c_0 con valor $f(c_0)$, si c_1 es una configuración *vecina* de c_0 con valor $f(c_1)$, ésta se elige como nueva configuración inicial si $f(c_1) < f(c_0)$ y con probabilidad $\exp\left(-\frac{f(c_1)-f(c_0)}{T}\right)$ en otro caso, donde T es un parámetro -inicialmente grande- que imita la temperatura; esta es la llamada *regla de aceptación de Metropolis,* que se deriva completamente de las leyes de la termodinámica .El proceso descrito anteriormente es repetido hasta que se alcanza un equilibrio a la temperatura T, y una vez alcanzado el valor de la temperatura se reduce de acuerdo con cierto esquema llamado *esquema de enfriamiento*, y nuevamente se repiten el procesos hasta obtener un nuevo equilibrio a la nueva temperatura. La secuencia anterior es repetida hasta obtener valores pequeños de T *i.e* hasta que el sistema se enfríe. A continuación damos el pseudo-código de una versión del algoritmo.

Algoritmo 1 *(Templado simulado)*

Inicio

Elegir una configuración inicial c_0;

Seleccionar las temperaturas inicial y final T_0, $T_f > 0$;

Seleccionar el número de iteraciones a realizar n_{iter};

Seleccionar un esquema de reducción de la temperatura;

$T := T_o; iter = 0;$

repite:

repite:

Elegir una nueva configuración c_1 en una vecindad de c_0;

$iter = iter + 1;$

$\delta = f(c_1) - f(c_0)$

Si $\delta \leq 0$ **entonces** $c_0 := c_1$

En otro caso

Elegir un número aleatorio uniforme u en $(0, 1)$;

Si $\exp\left(-\frac{\delta}{T}\right) < u$ **entonces** $c_0 = c_1$;

hasta que $iter = n_{iter}$;

Disminuir la temperatura T de acuerdo con el esquema seleccionado;

hasta que $T < T_f$;

c_0 es la aproximación a la configuración óptima;

Fin.

Al principio, con valores grandes de temperatura T será más probable poderse mover a puntos en donde la función objetivo aumenta (una situación peor). Esto limita como ya hemos mencionado las probabilidades de quedar atrapado en un mínimo local.

1.3.2. Consideraciones para su implementación

Este es un algoritmo muy general y son muchas las decisiones que han de tomarse para su implementación en la solución de un problema en particular. Estas consideraciones se dividen en dos categorías:

- *Genéricas*: son decisiones sobre las entradas del algoritmo, tales como el esquema de enfriamiento, la temperatura inicial y final, el criterio de paro y el número de iteraciones que han de realizarse antes de alcanzar el equilibrio; y
- *Específicas del problema:* que son decisiones que tienen que ver con la estructura de vecindad o lo que se entiende por vecindad.

Ambos tipos de decisiones afectan el desempeño del algoritmo.

Decisiones genéricas y específicas

Las decisiones genéricas en el templado simulado básicamente involucran al esquema de enfriamiento. El más usado es el esquema de reducción geométrico de la temperatura:

$$T_{t+1} = T_t \left(1 - \epsilon\right).$$

Valores pequeños de ϵ parecen tener un mejor desempeño. Muchos autores sugieren tomar $0.01 \leq \epsilon \leq 0.2$, aunque también puede definirse en términos de la temperatura inicial, final y el número de iteraciones M

$$\epsilon = 1 - \left(\frac{T_f}{T_0}\right)^{M-1}.$$

Otra elección común, sugerida en Lundy & Mees (1986), ejecuta sólo una iteración por cada etapa de la temperatura pero se forza a un enfriamiento más lento

$$T_{t+1} = \frac{T_t}{1 + \beta T_t},$$

donde β es una constante y que se elige de manera muy parecida a ϵ.

Y finalmente otro esquema de enfriamiento muy usado es el logarítmico,

$$T_{t+1} = \frac{1}{\log\left(T_t + 1\right)}.$$

El número total de iteraciones M sirve como criterio de paro, pues permite medir el tiempo de cómputo que necesitará el algoritmo.

En la literatura se han propuesto una gran variedad de esquemas de enfriamiento. De experiencias exitosas y del trabajo teórico en la literatura la tasa de enfriamiento, entendida como la razón a la cual decrece la temperatura, parece ser más importante que el esquema utilizado. Así, cuando se usa el templado simulado en un problema nuevo, el plan sugerido es empezar con un esquema de enfriamiento común -tal como el geométrico- y concentrarse en afinar la tasa de enfriamiento. Es importante también no descartar los otros esquemas en caso de que no se obtengan soluciones satisfactorias inmediatas. Por otro lado $T_0 = 1$ y $0.01 \leq T_f \leq 0.1$ son las especificaciones más populares para la temperaturas inicial y final, véase Fouskakis (2001) para más detalles. Otro parámetro es el número de iteraciones n_{iter} en cada etapa de la temperatura, el cual estará ligado al tamaño de vecindad y puede pensarse como una función de la temperatura en el siguiente sentido: para asegurarse que el templado simulado haya explorado completamente un mínimo local parecería razonable esperar que el templado simulado realizara más iteraciones cuando la temperatura es más baja. Una manera de realizar esto es aumentar el número de iteraciones n_{iter} en forma geométrica multiplicando por una constante cada vez que la temperatura disminuye o de manera aritmética sumando una constante en cada etapa. De esta forma el templado simulado gastará menos tiempo al principio cuando la tasa de aceptación, entendida como la proporción de iteraciones que son aceptadas, es muy alta; pero al final puede ocupar mucho tiempo alcanzar el número de iteraciones n_{iter} cuando la tasa de aceptación es muy baja. Sin embargo, lo usual es realizar una cantidad fija de movimientos antes de que la temperatura disminuya.

Dentro de las decisiones específicas del problema, es difícil, igual que en el caso de las decisiones genéricas, identificar reglas que garanticen el buen desempeño del templado simulado, pero nuevamente la literatura sugiere, por lo que respecta a la estructura de vecindad, que ésta sea simétrica (*i.e.* todas las configuraciones tengan el mismo número de vecinos). Para ahorrar tiempo es importante que la estructura se elija de manera que los cálculos sean rápidos y eficientes. También se sugiere que ésta no sea larga ni que tenga una estructura compleja.

1.4. Ejemplos

Ahora mostramos el desempeño del templado simulado mediante tres ejemplos y una comparación: en el primero maximizamos una función $f : \mathbb{R} \to \mathbb{R}$, en el segundo minimizamos una función $f : \mathbb{R}^2 \to \mathbb{R}$, en el tercer ejemplo vemos una aplicación a Teoría de las gráficas y finalmente comparamos el templado simulado con un algoritmo cuasi-Newton para estimar por máxima verosimilitud el parámetro de localización de una distribución Cauchy. El código en R para la solución de estos problemas se halla en el apéndice A.1.

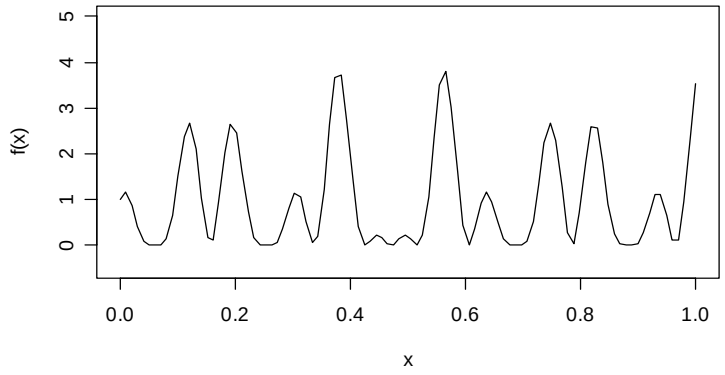


Figura 1.2: Comportamiento bimodal de $f(x) = [\cos(50x) + \sin(20x)]^2$

1.4.1. Maximizando una función $f : \mathbb{R} \to \mathbb{R}$

Se considera el problema de maximizar la función

$$f(x) = [\cos(50x) + \sin(20x)]^2 \text{ para } 0 \leq x \leq 1,$$

Robert & Casella (2004), la cual tiene un valor máximo de 3.83254 en $x = \{0.3791, 0.5633\}$. La gráfica de la función se muestra en la figura 1.2. Para este problema queda claro que la función objetivo es justamente f y que el espacio de soluciones posibles es $S = [0, 1]$. La estructura de vecindad para una configuración x_0 la definimos como sigue

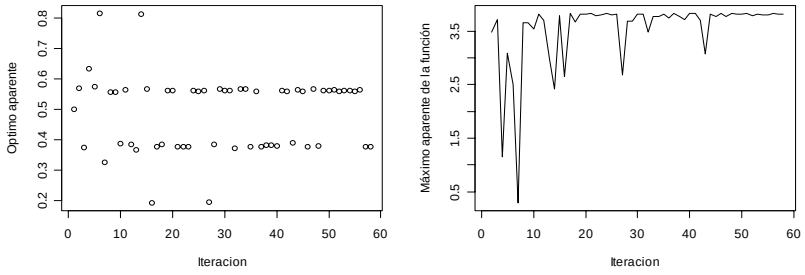
$$V(x_0) = \{x \in \mathbb{R} : \text{máx}[0, x_0 - 0.5] \leq x \leq \text{mín}[1, x_0 + 0.5]\}.$$

Por otro lado, las consideraciones genéricas son las sugeridas en las secciones previas: $T_0 = 1$ y $T_f = 0.050$ o un máximo de 1000 iteraciones. Dentro de las consideraciones específicas se eligió el radio 0.5 de la vecindad y realizar un total de 100 iteraciones intermedias antes de alcanzar el equilibrio a la temperatura dada.

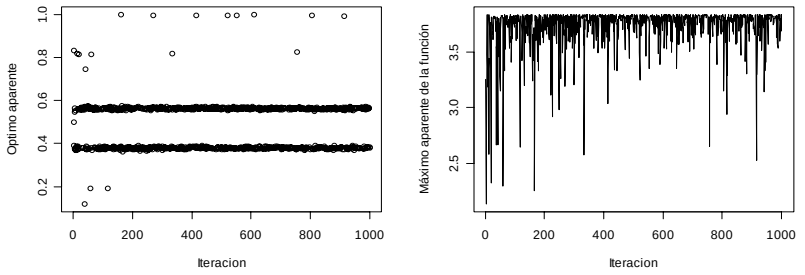
La tabla 1 resume los resultados obtenidos. En este caso, dado que es posible comparar con el óptimo verdadero podemos concluir que las soluciones obtenidas por el algoritmo son satisfactorias. Sin embargo, lo más interesante es ver cómo se da el proceso de búsqueda del óptimo. Las gráficas en el esquema 1-1 muestran este proceso.

Esquema de enfriamiento	x_{opt}	$f(x_{opt})$	Tasa de aceptación
Geométrico ($\epsilon = 0.05$)	0.5641	3.8287	0.9607
Logarítmico	0.3797	3.8306	0.9449
Lundy & Mees ($M = 1000$)	0.5631	3.8323	0.9400

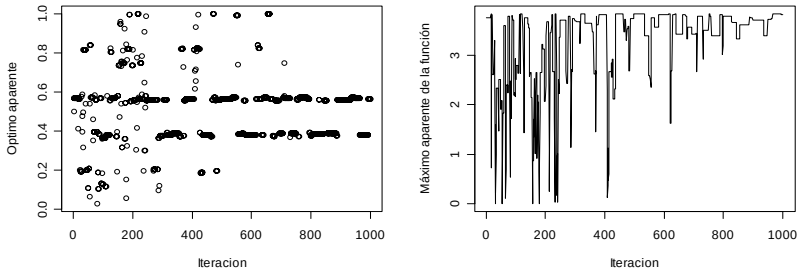
Tabla 1. Resultados del templado simulado tomando a 0.5 como aproximación inicial



Esquema Geométrico $\epsilon = 0.05$



Esquema Logarítmico



Esquema Lundy & Mees con $M = 1000$

Esquema 1-1. Proceso de convergencia al óptimo de la función utilizando templado simulado

Para el esquema geométrico se observa una convergencia rápida al máximo de f, después el algoritmo oscila entre los dos óptimos. Para los otros dos esquemas la convergencia al máximo es más lenta y

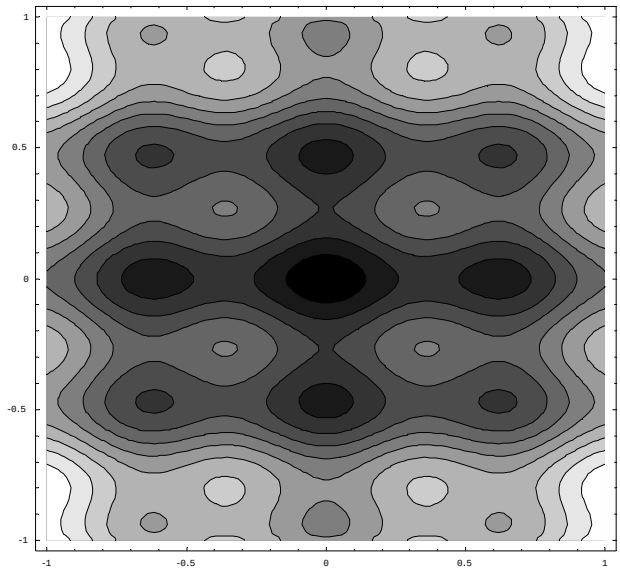


Figura 1.3: Gráfica de contornos para la función $f(x, y)$. Las regiones oscuras corresponden a valores más bajos.

después, al igual que en esquema geométrico, oscila entre los dos óptimos. Note también que bajo estos últimos dos esquemas es más notorio el proceso de elegir peores decisiones, siendo como ya lo esperábamos, más frecuente al principio.

1.4.2. Minimizando una función $f : \mathbb{R}^2 \to \mathbb{R}$

Se considera ahora el problema de minimizar la función

$$f(x, y) = x^2 + 2y^2 - 0.3\cos(3\pi x) - 0.4\cos(4\pi y) + 0.7 \text{ para } -1 \leq x, y \leq 1,$$

propuesta en Brooks & Morgan (1995), la cual tiene un tiene un óptimo en $(0, 0)$ y cuya gráfica de contornos se muestra en la figura 1.3. El problema de minimizar f lo podemos plantear como el de maximizar $-f$; queda claro que la función objetivo es $-f(x, y)$ y que el espacio de soluciones es $S = [-1, 1] \times [-1, 1]$.

La estructura de vecindad para una configuración $\mathbf{x}_0$ la definimos como

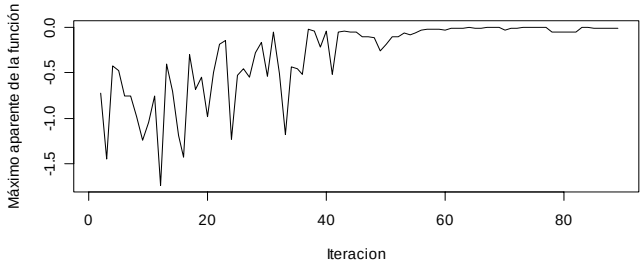
$$V(\mathbf{x}_0) = \{\mathbf{x} \in S : d(\mathbf{x}, \mathbf{x}_0) \leq 0.5\},$$

donde $d(\mathbf{x}, \mathbf{x}_0)$ representa la distancia euclidiana entre los vectores $\mathbf{x}$ y $\mathbf{x}_0$. Las consideraciones genéricas fueron tomadas de las sugeridas en la sección anterior: $T_0 = 1$ y $T_f = 0.01$ o 1000 iteraciones como

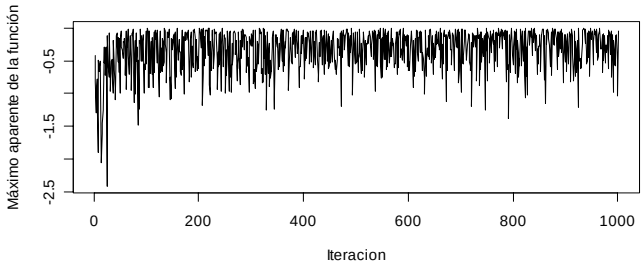
máximo. Como consideraciones específicas utilizamos una cantidad de 100 iteraciones intermedias antes de alcanzar el equilibrio y el radio de 0.5 para la vecindad. La tabla 2 resume los resultados obtenidos.

Esquema de enfriamiento	x_{opt}	$f(x_{opt})$	Tasa de aceptación
Geométrico ($\epsilon = 0.05$)	$(-0.0067, 0.0140)$	0.0072	0.8977
Logarítmico	$(-0.0392, -0.0207)$	0.0362	0.9321
Lundy & Mees ($M = 1000$)	$(-0.0222, -0.0370)$	0.0524	0.8390

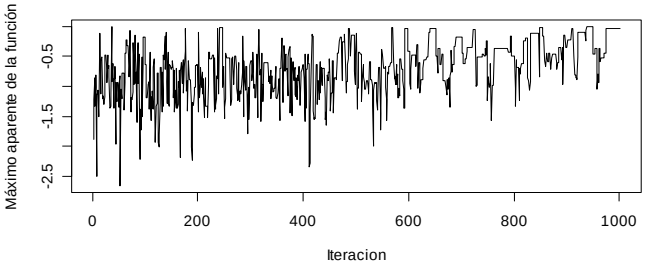
Tabla 2. Resultados del templado simulado tomando a 0.5 como aproximación inicial



Esquema Geométrico



Esquema Logarítmico



Esquema Lundy & Mees

Esquema 1-2. Valores mínimos aparentes de $f(x, y)$.

El proceso de cómo se alcanza el máximo de la función bajo distintos esquemas de enfriamiento es

mostrado en el esquema 1-2. Nuevamente se observa una rápida convergencia al mínimo de la función bajo el esquema de enfriamiento geométrico

1.4.3. Cubierta de vértices mínima para un gráfica

Para una gráfica $G = (V, E)$, se dice que $V' \subset V$ es una cubierta de vértices si toda arista tiene al menos un extremo en V'. El problema de la cubierta de vértices mínima es encontrar una cubierta de vértices con la menor cantidad de elementos. Es claro que este problema se complica cuando el conjunto de vértices es muy grande, de hecho es un problema de complejidad computacional NP[1], que puede resolverse utilizando templado simulado. Aunque en este problema ya no es tan claro cúal es la función objetivo, lo que sí queda claro es que el espacio de soluciones S tiene $2^V - 1$ elementos (el conjunto vacío no es solución). En Aarts & Korst (1989) se sugiere minimizar

$$f\left(V'\right) = \left|V'\right| - \lambda \left|E'\right|$$

donde $V' \in S$, $\lambda > 1$ y E' es el conjunto de aristas que V' no cubre. En este sentido λ se entiende como una penalización del conjunto V' por aquellas aristas que no cubre. Una configuración en este caso, no es más que una cubierta V' y se maneja usualmente como un vector de dimensión igual a la cantidad de vértices de la gráfica y cuya i-ésima entrada es igual a 1 si el vértice i pertenece a V' y 0 en otro caso. Así, la vecindad para una configuración C_0 podemos definirla como el conjunto de configuraciones tal que difieren de C_0 en exactamente una entrada.

Considere por ejemplo la gráfica mostrada en la figura 1.4 y propuesta en Fleischer (1995), donde los vértices $\{v2, v3, v5, v8, v9, v11\}$ marcados en negro corresponden a una cubierta de vértices mínima. A través del templado simulado con un esquema de enfriamiento geométrico y consideraciones genéricas $T_0 = 1\,, T_f = 0.1$ con 1000 iteraciones como máximo, encontramos cubiertas mínimas como $\{v2, v4, v7, v8, v11, v10\}$ y $\{v2, v4, v5, v7, v9, v12\}$.

1.5. Una comparación

Consideremos finalmente una muestra $x_1, x_2, \ldots, x_n$, de una población con distribución Cauchy cuya densidad está dada por

$$f(x; \alpha, \beta) = \frac{\beta}{\pi \left\{\beta^2 + (x - \alpha)^2\right\}}$$

[1] En teoría de la complejidad computacional el término NP hace referencia a los problemas de decisión cuya *solución* puede ser verificada en tiempo polinomial. La importancia de este tipo de problemas de decisión se debe a que incluye problemas de optimización y búsqueda para los que se desea saber si existe una solución en particular o sí existe una solución mejor que las conocidas.

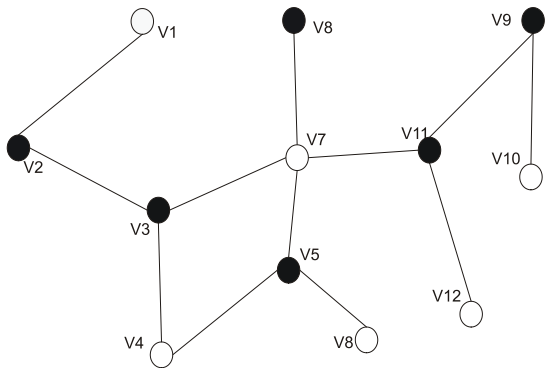


Figura 1.4: Los vértices $\{v2, v3, v5, v8, v11, v9\}$ constituyen un cubierta mínima para la gráfica

para $-\infty < x < \infty,\ \beta \geq 0$ y $-\infty < \alpha < \infty$.

El problema de calcular el estimador máximo verosímil para el parámetro de localización α, cuando β es constante conocida se vuelve un problema complejo dada la forma que puede tener la log-verosimilitud, que puede ser escrita como

$$f(\alpha; \beta, x_1, \ldots, x_n) = n \log \beta - \sum_{i=1}^{n} \log \left\{ \beta^2 + (x_i - \alpha)^2 \right\} - n \log \pi$$

para β fijo el estimador de máxima verosimilitud de α es el valor que minimiza la función

$$L(\alpha) = \sum_{i=1}^{n} \log \left\{ \beta^2 + (x_i - \alpha)^2 \right\}.$$

Si fijamos por ejemplo a $\beta = 0.1$ y consideramos la muestra aleatoria

$$\{-4.20, -2.85, -2.30, -1.02, 0.70, 0.98, 2.72, 3.50\}$$

obtenemos una log-verosimilitud dada en la figura 1.5.

Utilizamos un algoritmo del tipo cuasi-Newton implementado en R dentro de la función *optim* para probar y comparar el templado simulado de la siguiente manera. Se generó una muestra aleatoria de 1000 números aleatorios en $[-6, 6]$ mismos que se consideraron como puntos iniciales para ambos algoritmos. Para cada uno de ellos se registró la solución obtenida con un margen de error menor a 0.1 con respecto de la solución verdadera. Se observó que el método cuasi-Newton identifica en sólo 308 veces el mínimo global, mientras que el templado simulado bajo un esquema geométrico y consideraciones genéricas sugeridas, identifica al óptimo global en 720 ocasiones.

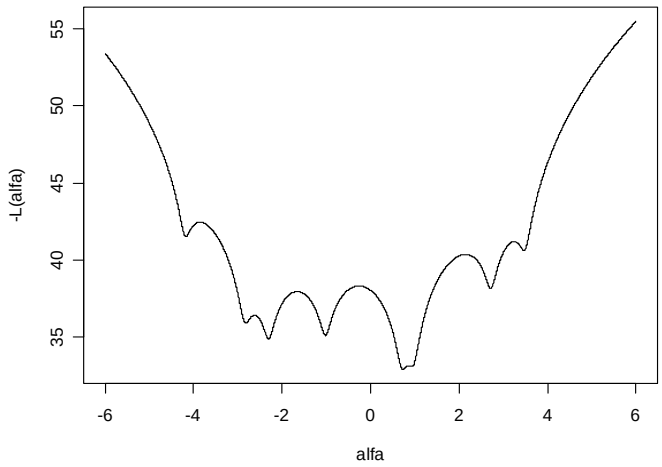


Figura 1.5: Gráfica de -Log-verosimilitud Cauchy, para los ocho datos. El mínimo se alcanza en $\alpha = 0.7327$

Capítulo 2

Teoría de la Decisión en Ambiente de Incertidumbre

La teoría de la decisión bajo incertidumbre tiene su desarrollo original durante la II Guerra Mundial. El texto clásico donde se expone partiendo de un perspectiva *frecuentista* es Wald (1950). A partir de allí ha tenido un vigoroso desarrollo. En el presente capítulo, basado principalmente en Bernardo (1981) y Bernardo & Smith (1994), se estructura un problema de decisión y se hace una presentación informal al proceso lógico que debe seguirse para tomar una decisión. Esto proporciona el fundamento sobre el que descansan el concepto de probabilidad, los métodos de inferencia y los criterios de decisión, que son globalmente conocidos como *métodos Bayesianos.*

El procedimiento seguido es *axiomático.* Se presentan y se defienden determinados principios de comportamiento *coherente* y se deduce de ellos la necesidad de asignar una medida de probabilidad que describa la información del tomador de decisiones y una función de utilidad que describa sus preferencias, y la de elegir aquella decisión capaz de maximizar la utilidad esperada.

El capítulo concluye analizando críticamente los resultados obtenidos y comparándolos con otros criterios de decisión expuestos en la literatura.

2.1. Introducción

La teoría de la decisión se ocupa de analizar cómo elige un tomador de decisiones, al que llamaremos simplemente *decisor*, aquella acción que, de entre un conjunto de acciones posibles, le conduce al mejor resultado dadas sus preferencias. Si se debe y cómo invertir en un determinado tipo de fondos, qué carrera estudiar, qué película de la cartelera ver, son problemas de decisión muy comunes a los que se enfrentan las personas de manera cotidiana.

El paradigma canónico de la teoría de la decisión se caracteriza por los siguientes elementos centrales. Contamos, para empezar, con un individuo, o grupo que actúa como individuo que ha de tomar una decisión cualquiera y de quien se dan por supuestas sus preferencias entre las distintas consecuencias de sus acciones. La teoría formal de la decisión no entra a considerar la naturaleza de las preferencias de los individuos ni por qué las personas prefieren unas cosas en lugar de otras. Desde la perspectiva formal que adopta la teoría, lo único que importa es que dichas preferencias, sean las que fueren, satisfagan ciertos criterios básicos de coherencia, entre los que cabe destacar por su importancia la *comparabilidad*, *transitividad* y *sustitución*. Si alguno de estos supuestos se viola resultará imposible saber qué es lo que el individuo prefiere; no se podrían jerarquizar sus preferencias, y la teoría de la decisión considerará que dicha persona no elige de manera racional. Por el contrario, si estos requisitos se cumplen, es posible atribuir al individuo una función de utilidad o beneficio, es decir, un índice o número a cada una de las consecuencias y que es consistente con sus preferencias de forma que las podamos ordenar de menor a mayor, de la menos preferida a la más preferida.

Para proceder al análisis de la decisión en estos términos es preciso identificar previamente un conjunto de alternativas posibles (desde la perspectiva del tomador de decisión) al que llamaremos *espacio de decisiones*. Debe ponerse especial atención en la construcción del espacio de decisiones porque el modelo que vamos a construir se limitará a elegir uno de sus elementos. Un buen decisor debe tener inventiva y el conocimiento suficiente del tema para elaborar un espacio de decisiones exhaustivo, es decir, que agote todas las posibilidades que parezcan, en principio, razonables. Asimismo es conveniente que el espacio de decisiones esté construido por un conjunto de alternativas, de forma que la elección de uno de sus elementos excluya la elección de cualquier otro. También debemos identificar un conjunto de consecuencias de cada una de las decisiones, consecuencias que se puedan anticipar y ordenar según las preferencias del individuo. Se supone que, dado el espacio de decisiones, el individuo elegirá aquella acción que tenga las mejores consecuencias, es decir, la que prefiera más. A esto habrá que agregar la cantidad de información con que cuenta el individuo para decidirse por una acción u otra de su espacio de decisiones. Si la información sobre los resultados de distintas acciones es completa -conocemos con toda seguridad las consecuencias de nuestras acciones- el individuo se hallará ante una situación de certidumbre; si por el contrario, la información es incompleta -desconocemos qué consecuencia tendrán nuestras acciones- la situación será de *incertidumbre*.

La teoría de la decisión aborda la naturaleza formal de las decisiones individuales y analiza criterios diversos de decisión bajo el contexto informativo en que se desenvuelva el individuo. Dicha naturaleza formal se puede tratar de manera normativa, prescriptiva o descriptiva. En el primer caso, se estudia qué decisión *debe* tomar un tomador de decisiones idealizado (que es racional y capaz de optimizar la búsqueda de información). La teoría prescriptiva de la decisión se ocupa, en cambio, de cómo pueden *elegir bien* individuos reales, dadas sus limitaciones cognitivas e informativas. La teoría descriptiva de

la decisión estudia cómo deciden, de hecho, las personas.

2.2. Estructura de un problema de decisión en ambiente de incertidumbre

Los elementos de un problema de decisión a ser especificados por el decisor para cada problema son los siguientes:

1. Un espacio de decisiones $\mathcal{D}$ que sea exhaustivo y excluyente.
2. Un espacio de sucesos inciertos $\mathcal{E}$ que es un álgebra de eventos relevantes al problema de decisión.
3. Un espacio de consecuencias posibles $\mathcal{C}$ que describe las consecuencias de tomar la decisión $d \in \mathcal{D}$ cuando ocurre un evento $E \in \mathcal{E}$.
4. Una relación de preferencia $\succeq$ que es una relación binaria definida sobre algunos elementos de $\mathcal{D}$.

En lo que resta de la exposición supondremos que, los espacios de decisiones $\mathcal{D}$ y de sucesos incierto $\mathcal{E}$ son finitos. Frecuentemente el conjunto de sucesos inciertos es el mismo cualquiera que sea la decisión que se tome, es decir, $E_i = \{E_{i1}, E_{i2}, \ldots, E_{im_i}\} = \{E_1, E_2, \ldots, E_m\} = \mathcal{E}$ para todo i. Si, además, sólo hay un número finito de alternativas y de sucesos inciertos, entonces el problema de decisión puede representarse esquemáticamente mediante una tabla de decisión de la forma

	E_1	$\ldots$	E_i	$\ldots$	E_m
d_1	c_{11}	$\ldots$	c_{1i}	$\ldots$	c_{1m}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
d_i	c_{i1}	$\ldots$	c_{ii}	$\ldots$	c_{im}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
d_k	c_{k1}	$\ldots$	c_{ki}	$\ldots$	c_{km}

donde $\mathcal{D} = \{d_1, d_2, \ldots, d_k\}$ es el espacio de decisiones y $E = \{E_1, E_2, \ldots, E_m\}$ es el conjunto de los sucesos inciertos, cualquiera que sea la decisión tomada y c_{ij} es la consecuencia de tomar la decisión d_i y que suceda E_j.

2.3. Solución intuitiva a un problema de decisión

Aunque los sucesos que componen cada E_i son inciertos, en el sentido de que no sabemos cuál de ellos tendrá lugar, no nos resultan, en general, igualmente verosímiles. En efecto, aunque no se disponga

de la información suficiente para determinar cuál de ellos tendrá lugar, típicamente se dispone de cierta información que hace unos elementos de cada E_i, más verosímiles que otros.

Nuestro primer objetivo será precisar de forma cuantitativa el contenido de este tipo de información incompleta. Una forma de hacerlo sería asignar a cada suceso incierto un número que midiese la verosimilitud que se le atribuye, de forma que a los sucesos más verosímiles se les haga corresponder un número mayor.

Desde hace varios siglos, la incertidumbre existente sobre la ocurrencia de algunas clases de sucesos ha venido midiéndose por un número entre 0 y 1 al que se ha llamado *probabilidad*.

Para hacerlo, la unidad de probabilidad correspondiente a un suceso cierto se distribuye entre una clase de sucesos mutuamente excluyentes de forma que el número asignado a un determinado suceso -su probabilidad- es tanto mayor cuando más verosímil se juzga la posibilidad de que tenga lugar.

Para nosotros, la probabilidad de un suceso A en una situación H, que representamos por $p(A \mid H)$, será una medida sobre una escala $[0,1]$ de la verosimilitud que se concede al suceso A en las condiciones descritas por H, esto es, una medida del grado de creencia en A que nos sugiere la información contenida en H. En un extremo, si dado H se está seguro de que A tendrá lugar, $p(A \mid H) = 1$; en el otro extremo, si dado H se está convencido de que A no sucederá, $p(A \mid H) = 0$. Otros valores en el intervalo $(0,1)$ expresan grados de creencia intermedios. La posibilidad asignada a un suceso es siempre condicional a la información que se posee sobre él: no existen *probabilidades absolutas*.

Volviendo al problema de decisión, la información que el decisor posee sobre la verosimilitud de los distintos sucesos inciertos de cada E_i en el momento de tomar la decisión podría pues ser cuantificada distribuyendo la unidad de probabilidad, para cada d_i, entre los sucesos $\{E_{i1}, E_{i2}, \ldots, E_{im_i}\}$ que comprende E_i. Puesto que por hipótesis, estos sucesos son mutuamente excluyentes, podríamos describir la información disponible sobre su verosimilitud, en las condiciones H en que debe tomarse la decisión, mediante un conjunto de números $\{p(E_{ij} \mid d_i, H),\ j = 1, 2, \ldots, m_i\}$ para cada i, tales que

$$0 \leq p(E_{ij} \mid d_i, H) \leq 1, \qquad \sum_{j=1}^{m_i} p(E_{ij} \mid d_i, H) = 1.$$

Si no hay confusión posible, omitiremos la condición H y escribiremos simplemente $p(E_{ij} \mid d_i)$ en lugar de $p(E_{ij} \mid d_i, H)$.

Siguiendo con nuestro problema de decisión en el que tanto el espacio de decisiones como el de sucesos inciertos son finitos, es claro que el decisor tendrá sus preferencias entre las distintas consecuencias. En principio, tales preferencias podrían ser cuantificadas asignando a cada una de las consecuencias c_{ij} un número $u(c_{ij})$ que mida la *utilidad* que cada una de ellas tenga para el decisor. Por ejemplo, podría asignarse el valor u_1 a la consecuencia más preferida, el valor u_0 a la menos preferida y valores intermedios $u(c_{ij})$ al resto, de forma que si una consecuencia es preferida a otra le sea asignado un número mayor. La función de utilidad así construida mide las preferencias del decisor entre las posibles

consecuencias de su decisión. Los conceptos coloquiales de probabilidad y utilidad serán formalizados respectivamente en las secciones 2.5 y 2.6.

Una vez especificadas las probabilidades $p(E_{ij} \mid d_i, H)$ que describen la verosimilitud de los sucesos inciertos y las utilidades $u(c_{ij})$ que describen las preferencias del decisor entre las posibles consecuencias, el problema planteado tiene ya una solución inmediata. En efecto, si se toma la decisión d_i en las condiciones H se puede obtener utilidad $u(c_{i1})$ con probabilidad $p(E_{i1} \mid d_i, H)$, $u(c_{i2})$ con probabilidad $p(E_{i2} \mid d_i, H), \ldots, u(c_{im_i})$ con probabilidad $p(E_{im_i} \mid d_i, H)$. Por lo tanto, la utilidad esperada de la decisión d_i que representaremos por $\overline{u}(d_i)$, vendrá dada por

$$\overline{u}(d_i) = \sum_{j=1}^{m_i} u(c_{ij})\, p(E_{ij} \mid d_i, H). \tag{2.1}$$

Resulta natural elegir como decisión más razonable aquella que maximiza la utilidad esperada, esto es aquella que maximiza la expresión (2.1) entre las k alternativas $\{d_1, d_2, \ldots, d_k\}$. Este es el *criterio de Bayes* para la toma de decisiones.

2.4. Axiomas de coherencia y cuantificación

En esta sección mostraremos que, si el decisor está dispuesto a aceptar los principios de comportamiento coherente que a continuación se exponen, entonces debe decidirse por aquella acción que maximice su utilidad esperada.

En la vida real, cuando se elige una determinada forma de actuar, no es frecuente poder elegir entre consecuencias predeterminadas; ya hemos comentado que, en general, las consecuencias de tomar una cierta decisión suelen depender de la ocurrencia de determinados sucesos inciertos. Formalmente llamaremos una *opción* y denotaremos por $l = \{c_1 \mid E_1, c_2 \mid E_2, \ldots, c_k \mid E_k\}$ a una situación en la que se obtiene la consecuencia c_1 si sucede E_1, la c_2 si sucede $E_2, \ldots,$ la c_k si sucede E_k, con la condición de que los sucesos E_i sean excluyentes y exhaustivos.

Dadas las opciones l_1 y l_2 escribiremos $l_1 \succ l_2$ si se prefiere la opción l_1 a la opción l_2, $l_1 \sim l_2$ si resultan igualmente deseables y $l_1 \succsim l_2$ si l_1 es preferible o igualmente deseable a l_2.

Note que las consecuencias son casos particulares de las opciones, puesto que si Ω es el evento seguro, una consecuencia c equivale a la opción $\{c \mid \Omega\}$. Con esto, puede verse que la relación de preferencias $\succeq$ definida sobre el espacio de opciones $\mathcal{D}$ induce una relación de preferencias sobre el espacio de consecuencias $\mathcal{C}$ definida como $c_1 \succeq c_2$ si, y sólo si $\{c_1 \mid \Omega\} \succeq \{c_2 \mid \Omega\}$. Esta identificación nos permitirá comparar consecuencias.

Dado cualquier problema de decisión real con un número finito de opciones y de eventos relevantes tiene una *mejor* consecuencia c^*, que es preferible o igualmente deseable a cualquiera de las que lo integran, y una peor consecuencia c_*, que no es estrictamente preferible a ninguna de ellas. De esta

forma, para cualquier consecuencia $c \in \mathcal{C}$ tendremos $c^* \succeq c \succeq c_*$. Para evitar el caso trivial de que todas las consecuencias sean igualmente deseables, supondremos además que $c^* \succ c_*$.

Axioma 1 ***Comparabilidad.*** *Para todo par de opciones l_1 y l_2 es cierta una de las tres relaciones $l_1 \succ l_2$, $l_2 \succ l_1$ o $l_1 \sim l_2$. Además, es posible encontrar dos consecuencias c^* y c_* tales que $c^* \succ c_*$ y que para toda consecuencia c, $c^* \succeq c \succeq c_*$*

La comparabilidad de opciones implica en particular la comparabilidad de las consecuencias y de verosimilitudes. Ya hemos visto que las consecuencias son casos particulares de opciones. Además, comparar la opción $l_1 = \left\{c^* \mid A, c_* \mid A^{'}\right\}$ con $l_2 = \left\{c^* \mid B, c_* \mid B^{'}\right\}$, donde A' y B' son, respectivamente, los eventos complementarios de A y B, equivale a comparar las verosimilitudes de A y B. Obviamente l_1 será preferible a l_2 si A es más verosímil que B.

Cualquiera de nosotros encontraría probablemente incómodo el hecho de comenzar afirmando que la opción l_1 es preferible a l_2 y la l_2 es preferible a l_3 y terminar concluyendo que l_3 es preferible a l_1. Postulamos pues que las preferencias deben ser transitivas.

Axioma 2 ***Transitividad.*** *Si $l_1 \succ l_2$ y $l_2 \succ l_3$, entonces $l_1 \succ l_3$. Análogamente, si $l_1 \sim l_2$ y $l_2 \sim l_3$, entonces $l_1 \sim l_3$.*

Claramente, la equivalencia entre opciones puede ser establecida por partes. Así por ejemplo, si la opción l_1 se juzga equivalente a la opción l_2 en los días laborales y también en los festivos, entonces las opciones l_1 y l_2 deben siempre ser juzgadas equivalentes.

Axioma 3 ***Sustitución.*** *Si $l_1 \succ l_2$ cuando sucede A y $l_1 \succ l_2$ cuando sucede A', donde A' es el evento complementario de A, entonces $l_1 \succ l_2$. Análogamente, si $l_1 \sim l_2$ cuando sucede A y $l_1 \sim l_2$ cuando sucede A', entonces $l_1 \sim l_2$.*

En virtud del principio de sustitución, en una opción puede reemplazarse una consecuencia por otra opción equivalente a ella. En efecto, si $c_1 \sim l_1$ la opción $l_2 = \{c_1 \mid A, c_2 \mid A'\}$ es equivalente a $l_3 = \{l_1 \mid A, c_2 \mid A'\}$. Basta observar que si sucede A, l_2 da lugar a c_1 y l_3 a l_1 y puesto que por hipótesis $c_1 \sim l_1$ tenemos que, si sucede A, $l_2 \sim l_3$. Por otro lado si sucede A', ambas opciones dan lugar a c_2 y por lo tanto $l_2 \sim l_3$ si sucede A'. Consecuentemente, en virtud del axioma 3, $l_2 \sim l_3$.

Estos tres axiomas de coherencia nos dan un conjunto mínimo de reglas que garantizan que las comparaciones cualitativas basadas en $\succeq$ no tengan implicaciones intuitivamente indeseables. Sin embargo, este enfoque puramente cualitativo no es adecuado para lograr una comparación sistemática de opciones. Hemos anticipado que para poder tomar decisiones de forma razonable es necesario *medir* la información y las preferencias del decisor expresándolas de forma cuantitativa. Para poder medir es

necesaria una *unidad de medida*. Con este objeto, postularemos que es posible imaginar métodos para elegir puntos al azar en el cuadrado unitario, de forma que resulte igualmente verosímil elegir cualquier punto, y los utilizaremos para construir con ellos una familia de sucesos que nos sirvan como referencia, como unidad de medida.

Axioma 4 ***Sucesos de referencia.*** *El decisor puede concebir un procedimiento de generar un punto aleatorio* $\mathbf{z}$ *en el cuadrado unitario, esto es, un número* $\mathbf{z} = (x, y)\,,\ 0 \leq x \leq 1,\ 0 \leq y \leq 1$ *tal que para cualquier par de regiones* $R_1,\ R_2$ *del cuadrado unitario el suceso* $\{\mathbf{z} \in R_1\}$ *le resulta menos verosímil que el suceso* $\{\mathbf{z} \in R_2\}$ *si, y solo si, el área de* R_1 *es menor que la de* R_2.

El axioma anterior nos permite construir un tipo de opciones con las que medir, por comparación, la deseabilidad de todas las demás. Específicamente, consideraremos opciones de la forma

$$\left\{c^* \mid R, c_* \mid R'\right\}, \qquad R \subset [0,1] \times [0,1]\,,$$

esto es, situaciones en las que se obtiene la mejor consecuencia posible c^* si sucede $\{\mathbf{z} \in R\}$ y la peor posible c_* si tiene lugar el suceso complementario $\{\mathbf{z} \notin R\}$, donde $\mathbf{z}$ es un punto aleatorio en el cuadrado unitario.

2.5. Probabilidad como grado de creencia

En la sección 2.3 mencionamos la necesidad de determinar la probabilidad asignada por el decisor a los distintos sucesos inciertos que puedan influir en las consecuencias de sus decisiones.

Históricamente, el concepto de probabilidad surgió con el estudio de los juegos de azar, en los que se dan ciertas *simetrías* que permiten concebir la probabilidad de un suceso como el cociente entre los números de casos en que puede darse y el número de casos totales, cuando todos los casos se consideran igualmente verosímiles y mutuamente excluyentes.

Más tarde, con el estudio de los problemas planteados por las compañías de seguros apareció el concepto de probabilidad de un suceso como el límite a que tendería la *frecuencia relativa* con que ese suceso se presentaría si ese suceso se repitiera indefinidamente en las mismas condiciones.

Aunque pueden servir para cuantificar la verosimilitud de determinados sucesos, los conceptos de simetría y frecuencia relativa no necesariamente sirven, para *definir* el concepto de probabilidad. Sin embargo, los axiomas de coherencia que hemos descrito permiten definir la probabilidad de un suceso cualquiera. Esto se consigue comparando el suceso en cuestión con el suceso de que un punto aleatorio se sitúe en una determinada región del cuadrado unitario y eligiendo la región de forma que, para el decisor, ambos sucesos sean igualmente verosímiles.

Definición 5 *La probabilidad de un suceso E en las condiciones H, que denotaremos por $p(E \mid H)$, es igual al área de una región R del cuadrado unitario elegida de forma que las opciones $l_1 = \{c^* \mid E, c_* \mid E'\}$ y $l_2 = \{c^* \mid R, c_* \mid R'\}$ sean igualmente deseables en las condiciones H.*

En la definición E' y R' son, naturalmente, los sucesos complementarios de E y R y las consecuencias c^*, c_* las que aparecen definidas en el axioma 1. También, es fácil ver que la probabilidad de cualquier suceso queda así bien definida, esto es, que siempre existe una región del cuadrado unitario que cumple con la condición exigida: que si dos regiones distintas la cumplen deben tener la misma área; y que la probabilidad asignada es independiente de las consecuencias de referencia c^*, c_* que se hayan elegido.

Teorema 6 *Cualquiera que sean las condiciones de referencia H, la medida de probabilidad verifica*

i) *Para todo suceso A, $0 \leq p(A) \leq 1$ y $p(H \mid H) = 1$.*

ii) Si *A y B dos sucesos incompatibles dado H, $p(A \cup B \mid H) = p(A \mid H) + p(B \mid H)$.*

iii) *Para todo par de sucesos A y B,*

$$p(A \cap B \mid H) = p(A \mid H) \cdot p(B \mid A, H) = p(B \mid H) \cdot p(A \mid B, H).$$

Para nosotros, la probabilidad $p(E \mid H)$ de un suceso E en las condiciones H es una medida, sobre la escala $[0, 1]$ de la verosimilitud que el decisor concede al suceso E en la situación descrita por H, esto es, una medida del grado de creencia en la ocurrencia de E que la información contenida en H le sugiere al decisor.

Hay dos puntos importantes que deben subrayarse. En primer lugar, la probabilidad asignada a un suceso es *siempre* condicional a la información que se posee sobre él; no existen probabilidades absolutas. En segundo lugar, estamos intentando cuantificar mediante probabilidades la información que *una persona determinada* posee sobre los sucesos inciertos que afectan a las consecuencias de sus decisiones: no existen probabilidades objetivas.

Obviamente, con la definición adoptada, la probabilidad de que un punto aleatorio se sitúe en una región $R(x)$ del cuadrado unitario de área x es precisamente x. En consecuencia, una forma alternativa de describir la situación representada por la opción

$$l(x) = \left\{c^* \mid R(x),\ c_* \mid R(x)'\right\},$$

que permite obtener c^* con probabilidad x y c_* con probabilidad $1-x$, consiste en escribir simplemente $\{c^* \mid x,\ c_* \mid 1-x\}$. En general, con la notación

$$l = \{c_1 \mid p_1, c_2 \mid p_2, \ldots, c_k \mid p_k\},$$

describiremos una opción que permite obtener la consecuencia c_1 con probabilidad p_1, c_2 con probabilidad p_2,...,c_k con probabilidad p_k.

2.6. Maximización de la utilidad esperada

Los axiomas de coherencia también permiten definir formalmente una medida de las preferencias del decisor entre las consecuencias. En efecto, definiremos la *utilidad* de un consecuencia c como un número $u(c)$ en la escala $[0,1]$ que mide la deseabilidad relativa de la consecuencia c. Es esta escala, como veremos a continuación, la utilidad de la peor consecuencia c_* será $u(c_*) = 0$ y para la mejor consecuencia c^* será $u(c^*) = 1$.

Como en el caso de las probabilidades, necesitaremos un elemento de referencia para poder medir. En aquel caso utilizábamos las regiones del cuadrado unitario. Aquí compararemos la deseabilidad de una consecuencia con la de una opción del tipo $\{c^* \mid x,\ c_* \mid 1-x\}$.

Definición 7 *La utilidad de una consecuencia c, que denotaremos por $u(c)$ es la probabilidad del evento E para el que la consecuencia c es igualmente deseable que la opción $\{c^* \mid E, c_* \mid E'\}$.*

De esta forma en términos de nuestra notación usada, para toda consecuencia c,

$$c \sim \{c^* \mid u(c), c_* \mid 1-u(c)\}.$$

En virtud de los axiomas de comparabilidad, transitividad y sustitución, siempre existirá un número $u(c)$ en $[0,1]$ que cumpla esa condición puesto que $c^* \sim \{c^* \mid 1, c_* \mid 0\}$, $c_* \sim \{c^* \mid 0, c_* \mid 1\}$ y $c^* \succeq c \succeq c_*$. Además, la utilidad de c queda así bien definida salvo una transformación lineal que, como veremos más adelante, no afecta a la elección de la decisión óptima.

Utilizando los axiomas de coherencia postulados en la sección 2.4 hemos podido asignar a los sucesos inciertos unos números (sus probabilidades) que miden la verosimilitud que les asigna el decisor en el momento, y en las condiciones, en que toma su decisión. Análogamente, hemos asignado a las consecuencias otro conjunto de números (sus utilidades) que miden las preferencias del decisor entre ellas. El paso final consiste en asignar un número a cada una de las decisiones posibles de forma que la mejor decisión sea aquella a la que se asigna el número más alto. Esto puede hacerse sin invocar nuevos principios ni hacer nuevas asignaciones numéricas. En efecto, tomar la decisión d_i es aceptar la opción

$$d_i = \{c_{i1} \mid E_{i1}, c_{i2} \mid E_{i2}, \ldots, c_{im_i} \mid E_{im_i}\},$$

donde $E_i = \{E_{i1}, E_{i2}, \ldots, E_{im_i}\}$ es el conjunto de sucesos inciertos que pueden afectar a las consecuencias de tomar la decisión d_i y c_{ij} es la consecuencia de haber tomado la decisión d_i cuando ocurre E_{ij}. De acuerdo con las definiciones 5 y 7 el decisor puede asignar, para cada E_{ij} la probabilidad $p(E_{ij} \mid d_i, H)$ de que suceda E_{ij} cuando se elige d_i en las condiciones H, y la utilidad $u(c_{ij})$ de la consecuencia a que esto da lugar.

Teorema 8 ***Criterio de decisión de Bayes****. Considérese el problema de decisión definido por* $\mathcal{D} = \{d_1, d_2, \ldots, d_k\}$ *donde*

$$d_i = \{c_{i1} \mid E_{i1}, c_{i2} \mid E_{i2}, \ldots, c_{im_i} \mid E_{im_i}\}.$$

Sea $p(E_{ij} \mid d_i, H)$ *la probabilidad de que suceda* E_{ij} *si se elige* d_i *en las condiciones* H *y sea* $u(c_{ij})$ *la utilidad de la consecuencia a que ello da lugar. Entonces, si se define la utilidad esperada de la decisión* d_i *como*

$$\overline{u}(d_i) = \sum_{j=1}^{m_i} u(c_{ij})\, p(E_{ij} \mid d_i, H),$$

la decisión óptima es aquella con máxima utilidad esperada.

Demostración. En efecto, por definición de utilidad,

$$c_{ij} \sim \{c^* \mid u(c_{ij}), c_* \mid 1 - u(c_{ij})\} \tag{2.2}$$

y por definición de probabilidad

$$d_i \sim \{c_{i1} \mid p(E_{i1} \mid d_i, H), c_{i2} \mid p(E_{i2} \mid d_i, H), \ldots, c_{im_i} \mid p(E_{im_i} \mid d_i, H)\}. \tag{2.3}$$

Introduciendo las expresiones (2.2) y (2.3) en virtud del axioma de sustitución y del teorema 6,

$$\begin{aligned} d_i \sim \quad & \{c^* \mid u(c_{i1})\, p(E_{i1} \mid d_i, H), c_* \mid [1 - u(c_{i1})]\, p(E_{i1} \mid d_i, H), \\ & c^* \mid u(c_{i2})\, p(E_{i2} \mid d_i, H), c_* \mid [1 - u(c_{i2})]\, p(E_{i2} \mid d_i, H), \ldots, \\ & c^* \mid u(c_{im_i})\, p(E_{im_i} \mid d_i, H), c_* \mid [1 - u(c_{im_i})]\, p(E_{im_i} \mid d_i, H)\} \end{aligned}$$

y, por tanto,

$$d_i \sim \{c^* \mid \overline{u}(d_i), c_* \mid [1 - \overline{u}(d_i)]\} \tag{2.4}$$

donde $\overline{u}(d_i)$ viene dado por la expresión (2.1). Ahora bien, en la forma (2.4) todas las decisiones son inmediatamente comparables y, en virtud de la transitividad, la preferible es aquella que da más probabilidad a la consecuencia óptima c^*, es decir, aquella que maximiza $\overline{u}(d_i)$ como queríamos demostrar. ■

Si la función de utilidad $u_1(c)$ se sustituye por una transformación lineal suya $u_2(c) = a \cdot u_1(c) + b$, la nueva utilidad esperada es claramente

$$\overline{u}_2(d) = a \cdot \overline{u}_1(d) + b.$$

Así pues, si $a > 0$, la decisión que maximiza $\overline{u}_1(d)$ también maximiza $\overline{u}_2(d)$ y por tanto, como habíamos anticipado, la elección de las consecuencias de referencia c_* y c^* no afectan la determinación de la decisión óptima.

Aunque la función de utilidad ha sido definida en la escala $[0,1]$, resulta a veces más natural medir la deseabilidad de las consecuencias en una escala distinta (tiempo, dinero, años de vida,...). Siempre es posible, sin embargo, reducir las utilidades, así obtenidas a la escala $[0,1]$ mediante la transformación

$$u'(c) = \frac{u(c) - u(c_*)}{u(c^*) - u(c_*)}. \tag{2.5}$$

Sin embargo, puesto que (2.5) es una transformación lineal, tal reducción no es necesaria, según hemos visto, para determinar la decisión óptima. El uso de la escala $[0,1]$ tiene, no obstante, la ventaja de permitir una interpretación probabilista de las utilidades.

2.7. Otros criterios de decisión en situaciones de incertidumbre

La fuerza del criterio de maximización de la utilidad esperada reside esencialmente en su fundamento axiomático. Es el único criterio de decisión compatible con los axiomas de coherencia expuestos en la sección 2.4. Cualquier otro criterio es equivalente a la maximización de una utilidad esperada, o es incoherente respecto de estos axiomas. En el resto de la sección comentaremos brevemente algunos otros criterios de decisión expuestos en la literatura y mostraremos la aplicación del criterio de decisión de Bayes en el contexto del siguiente ejemplo.

Ejemplo 9 ***(El vendedor de Temporada)*** *Muchas compañías están dedicadas a la venta de productos de temporada, por ejemplo: tiendas de ropa, de decoración navideñas, revistas y periódicos. Estos productos están caracterizados por una temporada relativamente corta de venta, después de la cual el valor de los productos decrece de manera sustancial. A menudo, la decisión de cuánta cantidad del producto manufacturar o comprar debe hacerse antes de que comience la temporada de ventas. Durante la temporada de ventas no hay tiempo para realizar un cambio en la cantidad, el decisor puede implementar algún otro tipo de mecanismos para alcanzar los resultados esperados, tales como cambiar el precio del producto de acuerdo con la demanda observada a lo largo de la temporada. Dado que la decisión de cuánta cantidad del producto manufacturar o comprar debe hacerse antes de conocer cuál será la demanda del producto, nos encontramos en un ambiente de incertidumbre.*

Supongamos que un administrador de una tienda ha decidido ordenar la cantidad x del producto de temporada. Esta variable de decisión x es una cantidad no negativa. El costo del producto a la tienda es c por unidad de producto. Durante la temporada de ventas el producto puede ser vendido a un precio r por unidad. Después de la temporada de ventas cualquier remanente en el producto es vendido al precio s por unidad. Si la demanda D es más grande que la cantidad ordenada x, entonces el total del producto se vende durante la temporada y no hay remanente al final de la misma por lo que el beneficio

obtenido es $rx - cx = (r - c)\, x$. Si por el contrario la demanda D es menor a la cantidad ordenada x, entonces el beneficio obtenido es $Dr + (x - D)\, s - xc = (s - c)\, x + (r - s)\, D$. Así el beneficio, bajo el supuesto de que se vende toda la mercancía y que no hay costos de oportunidad, puede escribirse como

$$\pi(x, D) = \begin{cases} (s - c)\, x + (r - s)\, D & \text{si } x \geq D \\ (r - c)\, x & \text{si } x < D \end{cases}.$$

El administrador desea elegir el valor x que maximice el beneficio $\pi(x, D)$, pero el dilema es que D es desconocido, o en otras palabras incierto al momento en que la decisión se realiza. Note que si $r \leq c$ y $s \leq c$, entonces la empresa no obtiene beneficio de comprar y vender el producto, así que la cantidad óptima a ordenar es $x^* = 0$, independientemente de cual sea el valor de la demanda. También si $s \geq c$, entonces cualquier remanente del producto puede ser vendido a un precio al menos igual al que se compró entonces la cantidad óptima a ordenar es tanto como sea posible, independientemente de cual sea el valor de la demanda. Por lo tanto, en lo que resta del ejemplo, asumiremos que $s < c < r$. Bajo este supuesto, para cualquier valor dado $D \geq 0$, el beneficio $\pi(x, D)$ es una función lineal a pedazos con pendiente positiva $r - c$ para $x < D$ y con pendiente negativa $s - c$ para $x \geq D$. Por lo que si la demanda D es conocida al momento de tomar la decisión de cuánto ordenar, entonces la mejor decisión es ordenar la cantidad $x^* = D$. Sin embargo, si el valor de D es desconocido el problema se vuelve más complicado.

2.7.1. Criterio minimax

El criterio de decisión minimax consiste en tomar la decisión que maximiza el beneficio garantizado: se determina lo peor que puede pasar para cada acción y se elige aquella decisión para la que resulta menos mala. De una manera más técnica, la decisión óptima según este criterio es aquella que maximiza el beneficio mínimo a que puede dar lugar. Pensemos en el ejemplo anterior. Supongamos que nuestro administrador cree (por experiencias de años anteriores) que la demanda D estará en un intervalo $[a, b]$ con $a < b$. Con este criterio el administrador tratará de alejarse del peor de los escenarios. Esto es, el administrador maximizará la función $p(x) = \text{mín}_{D \in [a,b]}\, \pi(x, D)$ sobre $x \geq 0$. Esto lleva al siguiente problema máx-mín

$$\max_{x \geq 0} \min_{D \in [a,b]} \pi(x, D)$$

no es difícil ver en este caso que $p(x) = \pi(x, a)$, y de aquí $x^* = a$ es la solución óptima. Claramente, en muchos casos esta sería una decisión conservadora.

En otra versión del método minimax se escoge aquella decisión que hace mínima la mayor *pérdida de oportunidad* posible, entendiendo por tal la diferencia entra la utilidad conseguida y la que hubiese podido conseguir.

Consideremos el problema de decisión definido por las tablas siguientes (utilidades a la izquierda, pérdidas de oportunidad a la derecha)

	E_1	E_2
d_1	0.8	0.0
d_2	0.2	0.4

	E_1	E_2	*máx*	
d_1	0.0	0.4	0.4	mín
d_2	0.6	0.0	0.6	

Claramente, según el criterio minimax, debe preferirse d_1 a d_2. Consideremos ahora una tercera decisión posible, d_3, de forma que las nuevas tablas de utilidades y pérdidas de oportunidad resultan ser

	E_1	E_2
d_1	0.8	0.0
d_2	0.2	0.4
d_3	0.1	0.7

	E_1	E_2	*máx*	
d_1	0.0	0.4	0.4	
d_2	0.6	0.0	0.6	mín
d_3	0.7	0.0	0.7	

En este caso, el criterio minimax sugiere la elección de d_2. Resulta así que la consecuencia de incorporar una nueva decisión d_3 al conjunto posible de alternativas es nada menos que sustituir $d_1 \succ d_2$ por $d_2 \succ d_1$.

A nivel intuitivo, la causa de la incoherencia del método minimax reside tanto en su excesivo pesimismo como en el hecho de que no utiliza la información de que se dispone sobre la verosimilitud relativa de los sucesos. Desde un punto de vista técnico, la causa de la incoherencia está en el hecho de que el criterio minimax no satisface el axioma de sustitución.

2.7.2. Criterio condicional

Otro criterio muy usado en la práctica es el criterio condicional que consiste en tomar la decisión que resultaría óptima si el suceso más probable tuviera lugar. Consideremos nuevamente el caso de nuestro administrador y supongamos que ahora cree que el valor de la demanda D puede ser alguno de los valores $\{d_1, d_2, \ldots d_n\}$ (obtenidos de registros históricos) y suponga que $p\,()$ es una distribución de probabilidad sobre el conjunto anterior. De acuerdo con este criterio si d_k con $k \in \{1, 2, \ldots, n\}$ es el evento más probable entonces la decisión óptima es $x^* = d_k$.

Sin embargo, el método resulta ser incoherente como se mostrará a continuación. Considere el problema de decisión descrito por la tabla de utilidades

	E_1	E_2	E_3
d_1	0.5	0.5	0.6
d_2	0.4	0.4	0.7

y suponga que $p\,(E_1) = 0.30,\ p\,(E_2) = 0.25,\ p\,(E_3) = 0.45$. De acuerdo con este criterio, $d_2 \succ d_1$ puesto que si el suceso más probable (E_3) tiene lugar, da una utilidad mayor. Sin embargo, observando que las

utilidades correspondientes a d_1 y d_2 coinciden indistintamente si ocurre el evento E_1 o E_2, el problema de decisión puede reformularse mediante la tabla

	$E_1 \cup E_2$	E_3
d_1	0.5	0.6
d_2	0.4	0.7

donde, obviamente, $p(E_1 \cup E_2) = 0.55$ y $p(E_3) = 0.45$. El suceso más probable es ahora $\{E_1 \cup E_2\}$ y por tanto $d_1 \succ d_2$. De nuevo, esto es incoherente. A nivel intuitivo la causa de la incoherencia del criterio condicional reside en su excesivo optimismo.

2.7.3. Criterio de decisión de Bayes

¿Qué pasaría ahora si abordamos las decisiones inciertas en términos probabilísticos, como en el criterio condicional, pero atendiendo a posibles atribuciones subjetivas de esa probabilidad? La respuesta se obtiene a través del teorema de Bayes. En lugar de atribuir una probabilidad objetiva a cada uno de los resultados cabría la posibilidad de que el decisor, por razones de registros históricos creyera que ciertos eventos son más probables que otros. Más aún, cabría la posibilidad de que por conocimiento propio o creencias propias tuviera una distribución de probabilidad *a priori* de los valores posibles. Entonces a través del teorema de Bayes puede combinar ambas fuentes de información para asignar nuevas probabilidades subjetivas a cada uno de los eventos. Pensemos nuevamente en el problema del administrador de productos de temporada y supongamos ahora que la demanda D puede ser vista como una variable aleatoria cuya distribución de probabilidad acumulada $F(w) = \Pr(D \leq w)$ es conocida o por lo menos puede ser estimada por medio de los registros históricos y la información inicial disponible para el administrador. Entonces se puede pensar en optimizar el beneficio esperado $E[\pi(x, D)] = \int_0^\infty \pi(x, w)\, dF(w)$. La decisión óptima x^* es, según el criterio de Bayes, tal que

$$x^* = \arg \max_{x \geq 0} E[\pi(x, D)]$$

Probablemente, la mayor limitación al principio de maximización de la utilidad esperada radica en el hecho de que su uso está restringido a un solo decisor y no es inmediatamente aplicable a situaciones que involucran a dos o más decisores. Así, nos hemos referido continuamente a un único decisor, dispuesto a expresar sus preferencias y a ser consistente con ellas, y hemos concluido que tal persona debe elegir la decisión que maximice su utilidad esperada; sin embargo, no hay nada en el desarrollo precedente que obligue a dos decisores a ponerse de acuerdo en las probabilidades asignadas a los sucesos inciertos o en las utilidades dadas a las posibles consecuencias.

A menudo, una decisión debe ser tomada por un organismo colegiado, por un comité, por una asamblea. En este caso puede suponerse frecuentemente que todos los componentes tienen los mismos

objetivos, y por tanto las mismas utilidades, pero que difieren en su apreciación de la realidad, es decir, en sus probabilidades. En otros casos sin embargo, las preferencias de los decisores son claramente contrapuestas entre sí; sus utilidades son diferentes y nos encontramos en una típica situación de conflicto. No existe una teoría axiomática que resuelva el problema planteado por las situaciones de comité o de conflicto de forma comparable a la solución que la teoría expuesta ofrece para el problema de decisión *unipersonal.* La consecución de tal teoría es la llamada *Teoría de Decisión de Grupos* que no tocaremos aquí.

Para finalizar el capítulo consideremos una aplicación a la teoría de inventarios, en la que se muestra el uso del templado simulado para la maximización de un beneficio esperado. El código en R para la solución del problema se encuentra en el apéndice A.2.

Ejemplo 10 ***(El problema del panadero)*** *Un fabricante de pan lo distribuye todos los días a la tiendas de abarrotes. El costo del pan para la compañía es \$0.80 por pieza. La compañía vende el pan a las tiendas a \$1.20 la pieza siempre y cuando sea pan fresco (vendido el día que se hornea). El pan que no se vende se regresa a la compañía. Esta tiene una pequeña tienda que vende pan de un día antes o más a \$0.60 la pieza. No se incurre en costo de almacenamiento significativo. El costo de la pérdida por faltantes se estima en \$0.80 por pieza. La demanda diaria tiene una distribución uniforme discreta entre 1000 y 2000 piezas. Encuentre el número óptimo de piezas que se deben producir al día.*

Solución.

Este problema puede plantearse en términos de un problema de decisión con los siguientes elementos.

El espacio de decisiones

$$\mathcal{D} = \{n \in \mathbb{Z} : 1000 \leq x \leq 2000\}$$

que consta de la cantidad de piezas de pan a elaborar.

El espacio de eventos relevantes

$$\mathcal{E} = \{E_i : i = 1000, 1001..., 2000\}$$

donde E_i representa el evento de que i es la cantidad de pan demandada.

Finalmente una función de utilidad u planteada en términos del ingreso neto de la compañía, y que es igual al rendimiento total menos el costo en que se incurre (al producir, almacenar y tener faltantes). Si se supone que no hay inventario inicial, el ingreso de la compañìa es

$$\begin{aligned} \text{ingreso neto} \quad = \quad & 1.20 \times \text{cantidad vendida de pan por la compañía} - 0.80 \times \text{cantidad} \\ & \text{de pan comprada por la tienda} + 0.60 \times \text{cantidad no vendida de pan y} \\ & \text{liquidada al valor de recuperación} - 0.80 \times \text{por piezas de pan faltantes.} \end{aligned}$$

Sean

$$n = \text{el número de piezas de pan, que se producen al día, y}$$

$$D = \text{la demanda de pan en un día cualquiera (variable aleatoria)}$$

de manera que

$$\begin{aligned} \text{mín}\,(D, n) &= \text{cantidad vendida de pan,} \\ \text{máx}\,(0, n - D) &= \text{cantidad no vendida de pan,} \end{aligned}$$

y

$$\text{máx}\,(0, D - n) = \text{cantidad de piezas faltantes.}$$

Denotando por $u\,(n, D)$ al ingreso que tiene la compañía, cuando se produce n piezas de pan y la demanda es D. Entonces

$$u\,(n, D) = 1.20\,\text{mín}\,\{D, n\} - 0.80n + 0.60\,\text{máx}\,(0, n - D) - 0.80\,\text{máx}\,(0, D - n)\,.$$

El primer término también se puede escribir como

$$1.20\,\text{mín}\,\{D, n\} = 1.20n - 1.20\,\text{máx}\,(0, n - D)\,.$$

Por tanto,

$$\pi\,(n, D) = 0.40n - 0.60\,\text{máx}\,(0, n - D) - 0.80\,\text{máx}\,(0, D - n)\,.$$

Si bien la compañía no sabe cuál será la demanda, se puede obtener una política óptima de inventario utilizando la información sobre la distribución de probabilidad de D. Sea

$$P_D\,(d) = P\,(\{D = d\})\,.$$

De aquí, la utilidad esperada está dada por

$$\begin{aligned} E\,[u\,(n, D)] &= \sum_{d=1000}^{2000} [0.40n - 0.60\,\text{máx}\,(0, n - d) - 0.80\,\text{máx}\,(0, d - n)]\,P_D\,(d) \\ &= 0.40n - 0.60 \sum_{d=1000}^{n} \text{máx}\,(0, n - d)\,P_D\,(d) - 0.80 \sum_{d=1000}^{2000} \text{máx}\,(0, d - n)\,P_D\,(d) \\ &= 0.40n - 0.60 \sum_{d=1000}^{n} (n - d)\,\frac{1}{1001} - 0.80 \sum_{d=n}^{2000} (d - n)\,\frac{1}{1001}. \end{aligned}$$

Aplicando el templado simulado, considerando a $E\,[u\,(n, D)]$ como la función objetivo, se encuentra que la cantidad óptima de pan a producir es de 1857 piezas, obteniéndose un ingreso neto máximo de \$514.20. La figura 2.1, muestra los valores óptimos y el máximo de la función objetivo a través de las iteraciones del templado simulado.

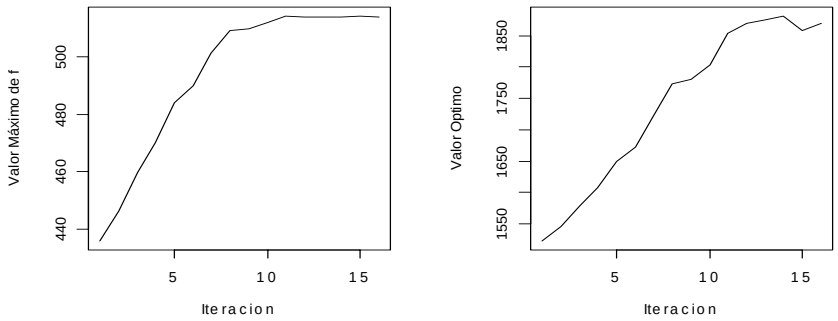


Figura 2.1: Trayectorias de los óptimos para el problema del panadero obtenidas a través de templado simulado.

Capítulo 3

Estadística Bayesiana

"La distribución final de hoy es la distribución inicial de mañana".

Luis Carlos Silva A.

Si bien tanto la bibliografía general como los programas docentes de estadística suelen eludir toda mención a las contradicciones inherentes a los métodos clásicos de Inferencia y al intenso debate desarrollado durante casi 70 años, muchos investigadores han alertado sobre sus limitaciones conceptuales y prácticas. Lo cierto es que ya no es posible desconocer las observaciones críticas que se hacen a dicho procedimiento, más aún, cuando se cuenta con otras formas de inferir, tales como la que ofrecen los métodos Bayesianos. En este capítulo, basado en Silva (2000) y Silva & Benavides (2001), hacemos una descripción del enfoque Bayesiano, y analizamos las ventajas y desventajas que presentan con respecto a los métodos clásicos.

3.1. Introducción

El investigador que está planteando una hipótesis estadística, comunica *a priori* su grado de convicción acerca de la validez de la hipótesis o, lo que es lo mismo, su grado de creencia en esa hipótesis y la medida en que no confía en ella. Esto lo plantea en términos probabilísticos, típicamente en términos de una distribución de probabilidad que llamaremos distribución de probabilidad *inicial*. Este componente es subjetivo. Después hay un componente, que depende exclusivamente de los datos, la llamada *verosimilitud*, y finalmente una conjunción de estos dos elementos vía el teorema de Bayes, que produce una distribución de probabilidad *final* de la veracidad de la hipótesis. Dicho en otras palabras, el investigador tiene inexorablemente una visión *a priori* de la realidad. Luego aparece esta realidad, que se mide objetivamente y que viene a modificar el grado de convicción que tenemos. Este es el principal elemento distintivo de la escuela Bayesiana.

Entonces, entre los rasgos claves de toda modelación Bayesiana hay que subrayar que no parte de un presunto vacío de información, sino que permite incorporar evidencias de las experiencias, de los experimentos y datos previos dentro de las conclusiones. Este enfoque, por otra parte, permite de manera natural y directa calcular probabilidades de eventos relacionados con observaciones futuras, tal posibilidad es obviamente atractiva a la hora de tomar decisiones.

3.2. Antecedentes históricos

En la década de los 20's Sir Ronald Aylmer Fisher propuso por primera vez valorar una hipótesis H_0 a través de una observación concreta d_0 así como la construcción del famoso p-valor definido como

$$p = \Pr(d \geq d_0 \mid H_0 \text{ es cierta})$$

que es la probabilidad de observar algo mayor o igual que lo que objetivamente se observó, suponiendo que sea válida la hipótesis que se valora. Fisher lo propuso como una medida de la discrepancia de los datos con la hipótesis. No tardaron en producirse críticas a este enfoque; basta con recordar la muy conocida expresión de uno de los estadístico más connotados del siglo XX, Leonard J. Savage, referida a las pruebas de hipótesis que se resume en que "cuando se sabe de antemano que la hipótesis de nulidad es falsa, el rechazo o la aceptación simplemente es reflejo del tamaño de la muestra, y las pruebas no hacen por tanto contribución alguna a la ciencia". Esta realidad es válida para cualquier enfoque que trabaje con la p propuesta por Fisher.

Para los años 30's y como reacción ante el planteamiento de Fisher, apareció por primera vez el tema de la hipótesis alternativa H_1, ausente en el planteamiento original. Fue un planteamiento nuevo: se fijan tasas de error, tipo I denotado por α y tipo II como todos conocemos. El planteamiento consistió en obtener la observación d_0 y realizar el cómputo de p, bajo el supuesto de que H_0 es cierta basándonos en los supuestos probabilísticos que procedan para poder hacerlo. Este cómputo de p es el mismo de Fisher, pero se modifica el empleo que ha de dársele. Ahora se trata de apoyarse en él para adoptar una decisión. Este es un planteamiento singular, radicalmente diferente y opuesto en cierto sentido al de Fisher. Aquí aparece el propósito de adoptar una decisión, como todos conocemos, de este modo:

$$\text{Si } p < \alpha \text{ se rechaza } H_0 \text{ en favor de } H_1; \text{ de lo contrario, se acepta } H_0$$

Y finalmente, de manera anónima e inercial, cerca del inicio de la II Guerra Mundial se unieron ambas teorías, se computa p y se empieza a decir que "la diferencia es significativa al nivel p". En este contexto se evita hablar de *aceptar* la hipótesis nula, para decir simplemente que la hipótesis nula *se rechaza* o que *no se rechaza*.

Esta combinación de ambas teorías disgustaría tanto a los creadores de una corriente como a los de la otra, ni Fisher por una parte, ni Neyman y Pearson por otra, congeniarían con la práctica contemporánea, porque ninguna de las dos escuelas está planteando nada como lo que actualmente se hace cada día.

3.3. Algunas objeciones al método de inferencia clásico

Es especialmente elocuente el siguiente texto, escrito por lo propios creadores de las pruebas de hipótesis: "Ninguna prueba basada en la teoría de probabilidad puede por sí misma generar índices válidos sobre la veracidad o falsedad de una hipótesis, las pruebas de hipótesis deben ser miradas desde otra perspectiva, siguiendo la regla de aceptar o rechazar una hipótesis no estamos diciendo nada definitivo sobre si la hipótesis es o no verdadera. Lo que se puede demostrar, es que si somos consistentes con esa regla, a la larga la rechazaremos cuando sea cierta no más de una vez de cada 100, si $p < 0.01$".

Esto traducido al lenguaje familiar, sería similar a un sistema de justicia al cual no le concierne si un individuo es inocente o culpable, sólo trata de limitar el número total de veredictos incorrectos.

Ahora, se aspira a contar con un procedimiento inferencial que este libre de las siguientes objeciones:

1. Dada la naturaleza del p-valor, la decisión de rechazar o no una hipótesis resulta ser, simplemente, un reflejo del tamaño de muestra. En este sentido es frecuente leer expresiones del tipo: puesto que "$p = 0.12$ no hay evidencia muestral de que la hipótesis sea falsa", esto si el autor no es partidario de la hipótesis alternativa. Ahora si estaba muy esperanzado en poder aceptarla (hipótesis alternativa) escribiría: "puesto que $p = 0.12$ no se puede decir formalmente que hay significancia, con otro tamaño de muestra probablemente se habría encontrado".

2. Usando la prueba de hipótesis, las decisiones se adoptan sin considerar la información externa al experimento u observación actual. Los resultados de estudios previos y la solidez de la teoría no participaron del proceso de inferencia. Esto supone un vacío de opiniones, una orfandad total de información, siempre irreal en la práctica.

3. La teoría de la pruebas de hipótesis es un instrumento para tomar decisiones dicotómicas sobre las hipótesis, en lugar de contribuir a valorar la credibilidad que estas últimas pudieran merecernos. Nuestras convicciones científicas pueden ser son más o menos sólidas, pero siempre son provisionales, y nuestras representaciones de la realidad tienen en cada momento cierto grado de credibilidad, que ha de estar abierto a cambios y perfeccionamientos en la medida que nuevos datos lo aconsejen.

3.4. ¿Por qué la inferencia Bayesiana?

Resulta clara ahora la necesidad de contar con un *nuevo* paradigma, que esté en lo posible exento de las objeciones que cuestionan al procedimiento inferencial clásico. En este sentido el enfoque Bayesiano constituye una alternativa atractiva puesto que cumple las siguientes condiciones:

Ventajas:

1. No está limitado por el tamaño muestral, en el sentido de que, si éste es pequeño, el impacto informativo también lo será.

2. Lejos de operar en un vacío de información, el modelo de análisis Bayesiano exige contemplar formal y explícitamente el conocimiento previo (información *a priori*). Esta información requiere de ser modelada en términos probabilísticos, materializada en una función de distribución de probabilidad.

3. Valora la credibilidad o verosimilitud de las hipótesis en lugar de obligarnos a adoptar decisiones dicotómicas sobre ellas.

4. La interpretación de intervalos de confianza, p-valores y predicciones, son más naturales.

Así pues la inferencia Bayesiana se presenta como una herramienta capaz de combinar de manera coherente información *a priori* con información muestral para producir información *a posteriori* en la que basar nuestras conclusiones finales.

Si la inferencia Bayesiana resulta más natural, ¿Por qué no se usa en lugar de la inferencia clásica?

Desventajas:

1. La introducción de la información *a priori*. La polémica surge por la subjetividad a la que está sujeta dicha información. Sin embargo, la subjetividad no es sinónimo de arbitrariedad ni capricho. La información *a priori* proviene, en la mayor parte de los casos de la experiencia previa del investigador, de investigaciones similares o meta-análisis. ¿Por qué renunciar a esta información *a priori* y no introducirla en el análisis?.

   Incluso en el peor de los casos en los que el investigador carezca de información *a priori*, existe la posibilidad de modelar la *no información* mediante una distribución de probabilidad inicial no informativa, basando la decisión final en la información proporcionada por los datos muestrales, tal como hace la estadística frecuentista. También es necesario recalcar que la inferencia clásica no esta libre de cierta subjetividad como por ejemplo, el nivel de significancia del 5 %, usar una o dos colas, etc.

2. Otra desventaja en el uso de la inferencia Bayesiana se concentra en la complejidad del cálculo matemático.

3.5. El enfoque Bayesiano

En la aproximación Bayesiana inicialmente se consulta con los investigadores para recabar toda la información disponible basada tanto en estudios anteriores como en creencias actuales de los especialistas. Esta información deberá ser presentada en forma de distribución de probabilidad, denominada distribución *inicial,* que es la cuantificación matemática de las creencias y los juicios iniciales sobre el parámetro de interés.

A continuación es necesario *recolectar* la información muestral que los datos aportan al estudio. Esta información muestral se genera vía la función de verosimilitud (en este aspecto coinciden ambas escuelas).

El último paso consiste en combinar la distribución inicial con la función de verosimilitud usando el teorema de Bayes que, a modo de caja negra, transforma nuestros juicios iniciales en nuevas creencias sobre el parámetro de interés, incorporando al análisis la información muestral.

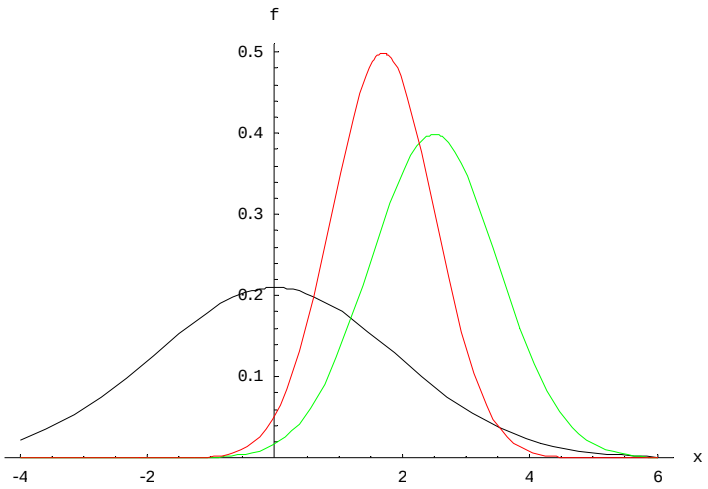
El teorema de Bayes puede expresarse de manera simple como sigue:

$$\text{La distribución final es proporcional a la verosimilitud} \times \text{distribución inicial}$$

en otras palabras

$$\Pr(\text{parámetros} \mid \text{datos}) \propto \Pr(\text{datos} \mid \text{parámetros}) \times \Pr(\text{parámetros})$$

El mecanismo de este proceso se ilustra en la figura siguiente.



Ejemplo del mecanismo del teorema de Bayes. Representamos en negro a la distribución inicial y en verde a la verosimilitud. La distribución final representada en rojo, se verá más influida por la verosimilitud al ser esta más precisa que la inicial

En este diagrama se observa como el teorema de Bayes combina las dos fuentes de información. La fuerza de cada fuente de información se indica por lo estrecho de la distribución. En este caso observamos que la verosimilitud (verde) resulta ser más informativa que la distribución inicial (negro). Dado que el teorema de Bayes reconoce la fuerza de cada curva, la distribución final (rojo) estará más influenciada por la verosimilitud que por la distribución inicial. Como una aplicación del Teorema de Bayes, consideremos el siguiente ejemplo.

Ejemplo: Un investigador desea saber si su paciente padece o no una enfermedad determinada E. El médico posee un juicio inicial sobre la probabilidad de que el individuo objeto de análisis padezca la enfermedad E. Este juicio la habrá formado por distintas fuentes de información: estudios históricos, grupo de riesgo, características socioeconómicas, etc., y tiene que ser expresado en términos de probabilidades: $P(E)$.

A continuación necesita recabar información muestral, realizando exploraciones, análisis clínicos, etc. Los datos obtenidos nos informan sobre las posibilidades de que la o las pruebas sean positivas en presencia de la enfermedad; es la llamada sensibilidad de la prueba diagnóstica. Esta información es expresada en términos de probabilidades: $P(+ \mid E)$.

El teorema de Bayes postula que la probabilidad de que el paciente esté enfermo, una vez realizada la prueba diagnóstica y observando un resultado positivo, es proporcional al producto de la probabilidad inicial de padecer la enfermedad E y la verosimilitud muestral. Por esta razón, el teorema de Bayes se interpreta como un proceso de aprendizaje y de actualización de conocimientos sobre el suceso de interés.

Capítulo 4

Selección de Modelos

En esta sección, basada en el capítulo 6 de Bernardo & Smith (1994), Key *et al.* (1999) y Chipman *et al.* (2001), abordamos el problema de la selección de modelos desde el punto de vista de la teoría de la decisión. Presentamos también, algunos de los métodos y de los criterios propuestos en la literatura para la solución del problema.

4.1. Introducción

Consideremos a $\boldsymbol{x}$ como un vector o matriz de datos y a $\mathcal{M} = \{M_i, i \in I\}$ como una colección finita de modelos distintos para el conjunto de datos. Supongamos que bajo M_i, $\boldsymbol{x}$ tiene una distribución dada por $p(\boldsymbol{x}|M_i, \theta_i)$, donde θ_i es un vector o matriz de parámetros del modelo M_i con el que buscamos explicar $\boldsymbol{x}$. Con base en los datos observados y asumiendo el modelo M_i, a través de la distribución final

$$p(\theta_i \mid \boldsymbol{x}, M_i) = \frac{p(\boldsymbol{x} \mid \theta_i, M_i)\, p(\theta_i \mid M_i)}{p(\boldsymbol{x} \mid M_i)}. \tag{4.1}$$

reflejamos todo nuestro conocimiento acerca del vector de parámetros θ_i. La lógica de la estadística Bayesiana permite a través del teorema de Bayes, hacer afirmaciones probabilistas acerca de lo que no sabemos (en este caso, saber si el modelo es correcto o no) condicionados en lo que sí sabemos (i.e. los datos), esto significa que la distribución final para cada modelo puede ser usada para medir nuestra creencia sobre si M_i es el modelo correcto, es decir,

$$p(M_i \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x}|M_i)\, p(M_i)}{p(\boldsymbol{x})}, \tag{4.2}$$

donde la distribución inicial $p(M_i)$ refleja el grado de creencia *a priori* sobre si M_i es el modelo verdadero y la verosimilitud marginal $p(\boldsymbol{x}|M_i)$ se obtiene de la ecuación (4.1), integrando ambos lados

con respecto de θ_i

$$p_i(\boldsymbol{x}) = p(\boldsymbol{x}|M_i) = \int p_i(\boldsymbol{x}|\theta_i)\, p_i(\theta_i)\, d\theta_i \tag{4.3}$$

donde $p_i(\boldsymbol{x}|\theta_i) = p(\boldsymbol{x}|\theta_i, M_i)$ y $p_i(\theta_i) = p(\theta_i \mid M_i)$.

Antes de continuar con nuestra discusión sobre la selección de modelos es importante distinguir entre dos posibles escenarios. En el primero, el modelo verdadero que genera los datos está en $\mathcal{M}$ aunque no se sabe de antemano cuál de ellos es, y en el segundo, el modelo verdadero puede no pertenecer a la clase $\mathcal{M}$. El primer caso corresponde al llamado enfoque $\mathcal{M}$-cerrado y el segundo al enfoque $\mathcal{M}$-abierto. La distinción entre ambos enfoques cobra importancia cuando se desea expresar la distribución final del modelo [ver ecuación (4.2)]. En efecto, en el enfoque $\mathcal{M}$-cerrado

$$p(\boldsymbol{x}) = \sum_{i \in I} p(M_i)\, p(\boldsymbol{x}|M_i). \tag{4.4}$$

Aunque, en el enfoque $\mathcal{M}$-abierto no hay una especificación de este tipo para $p(\boldsymbol{x})$, en la sección 4.3 veremos una alternativa para estudiar este caso.

4.2. La selección de modelos en el enfoque $\mathcal{M}$-cerrado

Denotaremos por m_i la acción de elegir a M_i como el modelo que genera los datos observados $\boldsymbol{x}$. En un primer caso consideremos el problema de elegir un modelo de $\mathcal{M}$ sin ninguna acción subsiguiente. De esta manera, la utilidad de la acción m_i tiene la forma $u(m_i, \omega)$ donde ω es un parámetro de interés (o estado de la naturaleza). El modelo óptimo m^*, que surge de maximizar la utilidad esperada (ver sección 2.6), está dado por

$$\overline{u}(m^* \mid \boldsymbol{x}) = \sup_{i \in I} \overline{u}(m_i \mid \boldsymbol{x}),$$

donde

$$\overline{u}(m_i \mid \boldsymbol{x}) = \int u(m_i, \omega)\, p(\omega \mid \boldsymbol{x})\, d\omega, \qquad i \in I.$$

En el enfoque $\mathcal{M}$-cerrado

$$p(\omega \mid \boldsymbol{x}) = \sum_{i \in I} p_i(\omega \mid \boldsymbol{x})\, p(M_i \mid \boldsymbol{x}), \tag{4.5}$$

donde la probabilidad de que M_i sea el modelo correcto condicionado en los datos observados se obtiene de las ecuaciones (4.2) y (4.4)

$$p(M_i \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid M_i)\, p(M_i)}{\sum_{i \in I} p(\boldsymbol{x} \mid M_i)\, p(M_i)},$$

y donde la verosimilitud marginal es como en la ecuación (4.3).

Desde la perspectiva $\mathcal{M}$-abierta no es posible en general decir algo sobre $p(\omega \mid \boldsymbol{x})$. Sin embargo, como veremos más adelante es posible comparar los modelos en $\mathcal{M}$ con base en sus utilidades esperadas (sección 4.3).

Consideremos ahora un esquema diferente, el cual, requiere la elección de un modelo M_i de la clase $\mathcal{M}$ y posteriormente, asumiendo que dicho modelo es el verdadero se requiere de una respuesta $a_j,\ j \in J$, relativa a un estado de la naturaleza ω que es de interés.

La aplicación de manera sistemática de la maximización de la utilidad esperada, establece que el óptimo $m^*_{\boldsymbol{x}}$ es tal que

$$\overline{u}(m^*_{\boldsymbol{x}} \mid \boldsymbol{x}) = \sup_{i \in I} \overline{u}(m_i \mid \boldsymbol{x}),$$

donde la utilidad esperada de seleccionar M_i dados los datos observados está dada por

$$\overline{u}(m_i \mid \boldsymbol{x}) = \int u(m_i, a^*_{\boldsymbol{x}}, \omega)\, p(\omega \mid \boldsymbol{x})\, d\omega,$$

y $a^*_{\boldsymbol{x}}$ se obtiene de maximizar

$$\int u(m_i, a_j, \omega)\, p_i(\omega \mid \boldsymbol{x})\, d\omega.$$

Es importante notar que diferentes definiciones de ω y diferentes formas de u siempre implican diferentes soluciones al problema de elección.

Por ejemplo, considere el problema de la selección de modelos cuando no es seguido de una acción subsiguiente y el estado de la naturaleza ω está definido como "elegir el modelo verdadero M_t", asumiendo una muestra $\boldsymbol{x}=(x_1, \ldots x_s)$ tal que $p(M_t \mid \boldsymbol{x}) \to 1$ cuando $s \to \infty$. En este caso es natural elegir la utilidad como:

$$u(m_i \mid \omega) = \begin{cases} 1 & \text{si } \omega = M_i \\ 0 & \text{si } \omega \neq M_i \end{cases}$$

y fácilmente se puede ver que

$$p_i(\omega \mid \boldsymbol{x}) = \begin{cases} 1 & \text{si } \omega = M_i \\ 0 & \text{si } \omega \neq M_i \end{cases}$$

por lo que

$$p(\omega \mid \boldsymbol{x}) = \begin{cases} p(M_i \mid \boldsymbol{x}) & \text{si } \omega = M_i \\ 0 & \text{si } \omega \neq M_i \end{cases}$$

y la utilidad esperada de la decisión m_i dado $\boldsymbol{x}$ está dada por

$$\begin{aligned} \overline{u}(m_i \mid \omega) &= \int u(m_i, \omega)\, p(\omega \mid \boldsymbol{x})\, d\omega \\ &= p(M_i \mid \boldsymbol{x}), \qquad i \in I, \end{aligned}$$

es decir, la decisión óptima es elegir el modelo que tiene *máxima probabilidad final.*

Ahora consideramos el problema de la selección de modelos dados los datos $\boldsymbol{x}$, con el fin de hacer una predicción de una observación futura y. Asumimos observaciones reales $\boldsymbol{x}$, y que m_i, denota como antes, la acción de elegir a M_i como el modelo verdadero, y finalmente que $a_j,\ j \in J$ denota la elección basada en M_i de la predicción $\widehat{y}_i$ para una observación futura y, con una utilidad cuadrática

$$u\left(m_i, \widehat{y}_i, y\right) = -\left(\widehat{y}_i - y\right)^2, \qquad i \in I.$$

Recordando que la elección óptima m^* está dada por

$$\overline{u}\left(m^* \mid \boldsymbol{x}\right) = \sup_{i \in I} \int u\left(m_i, \widehat{y}_i^*, y\right) p\left(y \mid \boldsymbol{x}\right) dy,$$

donde $\widehat{y}_i^*$ es la predicción óptima de una observación futura y dado los datos y asumiendo a M_i como el modelo verdadero, esto es, el valor $\widehat{y}$ que minimiza

$$\int \left(\widehat{y} - y\right)^2 p_i\left(y \mid \boldsymbol{x}\right) dy,$$

se sigue que

$$\widehat{y}_i^* = E\left[y \mid \boldsymbol{x}, M_i\right] = \int y p_i\left(y \mid \boldsymbol{x}\right) dy, \qquad i \in I,$$

es decir, la media de la distribución predictiva de y dado el modelo M_i.

Se sigue entonces que

$$\int u\left(m_i, \widehat{y}_i^*, y\right) p\left(y \mid \boldsymbol{x}\right) dy = -\int \left(\widehat{y}_i^* - y\right)^2 p\left(y \mid \boldsymbol{x}\right) dy$$

y, bajo el enfoque M-cerrado de la ecuación (4.5) con $\omega = y$

$$p\left(y \mid \boldsymbol{x}\right) = \sum_{i \in I} p_i\left(y \mid \boldsymbol{x}\right) p\left(M_i \mid \boldsymbol{x}\right).$$

Entonces

$$\begin{aligned} \int \left(\widehat{y}_i^* - y\right)^2 p\left(y \mid \boldsymbol{x}\right) dy &= \sum_{j \in I} p\left(M_j \mid \boldsymbol{x}\right) \int \left(\widehat{y}_i^* - \widehat{y}_j^* + \widehat{y}_j^* - y\right)^2 p_j\left(y \mid \boldsymbol{x}\right) dy \\ &= \sum_{j \in I} p\left(M_j \mid \boldsymbol{x}\right) V\left[y \mid M_j, \boldsymbol{x}\right] + \sum_{j \in I} \left(\widehat{y}_j^* - \widehat{y}^*\right)^2 p\left(M_j \mid \boldsymbol{x}\right) \\ &\quad + \left(\widehat{y}_i^* - \widehat{y}^*\right)^2 \end{aligned}$$

donde

$$\widehat{y}^* = \sum_{j \in J} \widehat{y}_j^* p\left(M_j \mid \boldsymbol{x}\right),$$

por lo que el modelo óptimo M_i es aquel para el cual $\widehat{y}_i^*$ *está más cerca de* $\widehat{y}^*$.

Más allá de los problemas de estimación puntual y predicción, consideremos el problema de comparación de modelos con el fin de reportar inferencias acerca de un estado de la naturaleza ω.

Este es un problema más general y que cae dentro del esquema de elección seguida de una acción subsiguiente. Asumiendo el modelo M_i como el modelo verdadero y una utilidad definida como

$$u\left(m_i, a_j, \omega\right) = u_i\left(p_i\left(\cdot \mid \boldsymbol{x}\right), \omega\right)$$

para alguna función de puntaje u_i y donde $p_i\left(\cdot \mid \boldsymbol{x}\right)$ denota nuestra creencia sobre el estado de la naturaleza ω implicada por m_i. Para más detalle de las propiedades de las funciones de puntaje y su relación con el problema de reportar inferencias, se sugiere revisar el material contenido en el apéndice B.

Si la función de utilidad u_i es propia entonces la acción óptima $a_j,\ j \in J$ debe ser $a_i^* = p_i\left(\cdot \mid \boldsymbol{x}\right)$ y

$$u\left(m_i, a_i^*, \omega\right) = u_i\left(p_i\left(\cdot \mid \boldsymbol{x}\right), \omega\right), \quad i \in I.$$

Si más aún es local, debe tener forma logarítmica

$$u\left(m_i, a_i^*, \omega\right) = A \log p_i\left(\omega \mid \boldsymbol{x}\right) + B\left(\omega\right), \quad i \in I,\ A > 0.$$

Por lo tanto, la utilidad de m_i es

$$\overline{u}\left(m_i \mid \boldsymbol{x}\right) = \int \left[A \log p_i\left(\omega \mid \boldsymbol{x}\right) + B\left(\omega\right)\right] p\left(\omega \mid \boldsymbol{x}\right) d\omega \tag{4.6}$$

y el modelo preferido es aquel para el cual la expresión anterior se maximiza.

Desde la perspectiva $\mathcal{M}$-cerrada recordemos de la ecuación (4.5) que

$$p\left(\omega \mid \boldsymbol{x}\right) = \sum_{i \in I} p_i\left(\omega \mid \boldsymbol{x}\right) p\left(M_i \mid \boldsymbol{x}\right).$$

Sustituyendo esta expresión en la ecuación (4.6) tenemos:

$$\begin{aligned}
\overline{u}\left(m_i \mid \boldsymbol{x}\right) &= \sum_{j \in I} Ap\left(M_j \mid \boldsymbol{x}\right) \int p_j\left(\omega \mid \boldsymbol{x}\right) \log p_i\left(\omega \mid \boldsymbol{x}\right) d\omega + \int B\left(\omega\right) p\left(\omega \mid \boldsymbol{x}\right) d\omega \\
&= \sum_{j \in I} Ap\left(M_j \mid \boldsymbol{x}\right) \int p_j\left(\omega \mid \boldsymbol{x}\right) \left[\log p_i\left(\omega \mid \boldsymbol{x}\right) + \log p_j\left(\omega \mid \boldsymbol{x}\right) - \log p_j\left(\omega \mid \boldsymbol{x}\right)\right] d\omega \\
&= -\sum_{j \in I} Ap\left(M_j \mid \boldsymbol{x}\right) \int p_j\left(\omega \mid \boldsymbol{x}\right) \log \frac{p_j\left(\omega \mid \boldsymbol{x}\right)}{p_i\left(\omega \mid \boldsymbol{x}\right)} d\omega \\
&\quad + \sum_{j \in I} Ap\left(M_j \mid \boldsymbol{x}\right) \int p_j\left(\omega \mid \boldsymbol{x}\right) \log p_j\left(\omega \mid \boldsymbol{x}\right) d\omega,
\end{aligned}$$

que equivale a minimizar

$$\sum_{j \in I} Ap\left(M_j \mid \boldsymbol{x}\right) \int p_j\left(\omega \mid \boldsymbol{x}\right) \log \frac{p_j\left(\omega \mid \boldsymbol{x}\right)}{p_i\left(\omega \mid \boldsymbol{x}\right)} d\omega,$$

la distribución final ponderada sobre los modelos de la divergencia logarítmica entre p_i y p_j, $i \neq j$.

4.3. La selección de modelos en el enfoque $\mathcal{M}$-abierto

Para el problema general de selección de modelos seguida de una acción subsiguiente, el análisis sugiere la elección óptima de la clase $\mathcal{M}$ como el modelo M_i para el cual

$$\int u\left(m_i, a_i^*, \omega\right) p\left(\omega \mid \boldsymbol{x}\right) d\omega$$

se maximiza sobre $i \in I$ y a_i^* denota el óptimo dado M_i. Ya vimos en la sección anterior que desde la perspectiva $\mathcal{M}$-cerrada la forma de $p\left(\omega \mid \boldsymbol{x}\right)$ permite dar una forma explícita a la solución. Abordaremos ahora la tarea de comparar los valores de M_i, $i \in I$ desde la perspectiva $\mathcal{M}$-abierta. ¿Qué es lo que podemos hacer para comparar los valores de M_i, $i \in I$, que se aproximan a un modelo que no está o no ha sido especificado?. Esto es, no tenemos una forma para $p\left(\omega \mid \boldsymbol{x}\right)$. Ilustraremos una posible aproximación al problema poniendo énfasis en el caso $\omega = y$ una observación futura, para la cual se requiere una estimación puntual con respecto a la pérdida cuadrática, o una distribución predictiva con respecto a la función de puntaje logarítmica o cuadrática.

Note primeramente que, en cualquier caso, la utilidad esperada de elegir el modelo M_i tiene la forma

$$\int u\left(m_i, a_i^*, y\right) p\left(y \mid \boldsymbol{x}\right) dy = \int f_i\left(y, \boldsymbol{x}\right) p\left(y \mid \boldsymbol{x}\right) dy$$

para alguna función f_i cuya forma funcional es conocida.

En segundo lugar, note que hay n posibles particiones de $\boldsymbol{x} = \boldsymbol{x}_n = (x_1, \ldots x_n)$ en $\boldsymbol{x}_n = \left[\boldsymbol{x}_{n-1}\left(j\right), x_j\right]$, $j = 1, \ldots, n$ donde $\boldsymbol{x}_{n-1}\left(j\right)$ denota a $\boldsymbol{x}_n$ sin la $j-$ésima entrada y que si n es grande y las x's son intercambiables, cada una de las particiones dada por $\boldsymbol{x}_{n-1}\left(j\right)$ se puede considerar como una aproximación de $\boldsymbol{x}$ y x_j como aproximación para y, la observación futura.

Si ahora elegimos k de las n particiones, la ley fuerte de los grandes números sugiere que cuando $n, k \to \infty$

$$\left| \int u\left(m_i, a_i^*, y\right) p\left(y \mid \boldsymbol{x}\right) dy - \frac{1}{k} \sum_{j=1}^{k} f_i\left(x_j, \boldsymbol{x}_{n-1}\left(j\right)\right) \right| \to 0.$$

Es posible entonces, comparar las utilidades esperadas de M_i, $i \in I$ con base en las cantidades

$$\frac{1}{k} \sum_{j=1}^{k} f_i\left(x_j, \boldsymbol{x}_{n-1}\left(j\right)\right), \qquad i \in I.$$

En el caso de la estimación puntual si y es una observación futura y $\widehat{y}_i^*\left(j\right)$ denota el valor de $E\left[y \mid M_i, \boldsymbol{x}\right]$ cuando $\boldsymbol{x}$ es reemplazada por $\boldsymbol{x}_{n-1}\left(j\right)$, esta aproximación implica que minimizamos sobre $i \in I$ la expresión

$$\frac{1}{k} \sum_{j=1}^{k} \left(\widehat{y}_i^*\left(j\right) - x_j\right)^2$$

que es una medida promedio de las distancias al cuadrado de qué tan bien M_i se aproxima cuando se deja fuera un elemento a la vez.

En el caso de una distribución predictiva con el puntaje logarítmico maximizamos sobre $i \in I$, la expresión

$$\frac{1}{k} \sum_{j=1}^{k} \log p\left(x_j \mid M_i, \boldsymbol{x}_{n-1}\left(j\right)\right),$$

la cual se puede ver como una medida promedio de la verosimilitud integrada bajo el modelo M_i.

El desarrollo anterior, establece claramente el rol que estas técnicas de *validación cruzada* juegan en el proceso de aproximar utilidades esperadas en problemas de decisión donde es posible comparar conjunto de modelos sin la necesidad de actuar o suponer que uno de ellos es el modelo verdadero.

Capítulo 5

Un enfoque Bayesiano para la Comparación de Modelos de Regresión

Las técnicas de regresión se cuentan entre los métodos más utilizados en la estadística aplicada. Dada una variable de respuesta $\mathbf{Y}$ y un conjunto de covariables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$, es de interés estimar una relación funcional supuesta entre la variable de respuesta y las covariables.

Un problema que surge al construir un modelo de regresión es decidir qué variables incluir en el modelo. Este problema, conocido como el problema de *selección de variables,* se puede ver como un problema de selección de modelos, donde cada modelo considerado corresponde a usar como covariables a cada uno de los distintos subconjuntos de $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$. La comparación Bayesiana de modelos no dice cuál es el modelo verdadero, sino que establece una relación de preferencia con base en los datos y alguna otra información sobre el conjunto de modelos a comparar. Esta relación de preferencia se usa entonces para elegir el "mejor" modelo.

Se ha desarrollado una variedad de criterios para comparar los 2^p modelos posibles, por ejemplo: factores de Bayes, el criterio de información de Bayes (BIC) y el criterio de información de Akaike (AIC) entre otros. Desafortunadamente, algunos criterios dependen del número de covariables por lo que hace imposible la comparación de modelos de dimensiones diferentes. Además el trabajo computacional para calcular estos criterios (en caso que se pueda) cuando p es grande es exhaustivo.

En este capítulo se presenta el enfoque semiparamétrico predictivo Bayesiano para la comparación de modelos de regresión en el enfoque $\mathcal{M}$-abierto, desarrollado por Gutierrez-Peña (1997). En la siguiente sección hacemos una descripción del modelo semiparamétrico que utilizaremos en las demás secciones.

5.1. El Modelo de regresión semiparamétrico

El modelo de regresión lineal normal

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

supone que los errores son normales, independientes e idénticamente distribuidos y que la relación funcional entre la variable de respuesta y las covariables es lineal en los parámetros. Tales supuestos son necesarios para dar forma a la función de verosimilitud, que es un componente crucial para el análisis Bayesiano. Sin embargo, en la vida real es ilógico hacer estas suposiciones acerca de la distribución y la forma funcional. Como alternativa contamos con métodos *no-paramétricos* o *semiparamétricos* cuyo objetivo es relajar dichos supuestos, en su totalidad en el caso no-paramétrico y de manera parcial en el caso semiparamétrico. Esta característica hace más flexible a la clase de modelos semiparamétricos y por lo tanto más amplia que la clase de modelos de regresión lineal normal. El objetivo que perseguiremos ahora será elegir un modelo dentro de la clase de los modelos semiparamétricos que mejor se aproxima al modelo que genera los datos.

Supongamos que los datos $\{(y_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$ están generados por el modelo

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i, \tag{5.1}$$

donde los $\{\epsilon_i\}$ son variables aleatorias independientes e idénticamente distribuidas $N\left(0, \sigma^2\right)$ y $f(\cdot)$ es una función suave en $\mathbb{R}^r$. Típicamente f es una función desconocida o con una forma funcional extremadamente compleja, por lo que es frecuentemente aproximada por una forma paramétrica simple como

$$P_q(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x})^T \boldsymbol{\beta},$$

donde $\boldsymbol{h}(\boldsymbol{x})$ es un vector q-dimensional de funciones suaves conocidas y $\boldsymbol{\beta} \in \mathbb{R}^q$ es un vector de parámetros desconocidos. Usualmente, el modelo es analizado asumiendo que $P_q(\boldsymbol{x})$ es igual a $f(\boldsymbol{x})$, por lo que el problema se reduce a hacer inferencias sobre $\boldsymbol{\beta}$.

Desde la perspectiva Bayesiana $f(\boldsymbol{x})$ puede ser tratado como un parámetro más, representando nuestro conocimiento *a priori* sobre $f(\boldsymbol{x})$ a través de un proceso Gaussiano. Por tanto el análisis del modelo (5.1) se reduce a un análisis del modelo lineal normal.

Supongamos el siguiente modelo Jerárquico inicial

Etapa I: condicional en $f(\boldsymbol{x})$ y σ^2,

$$y \sim N\left(f(\boldsymbol{x}), \sigma^2\right), \qquad \boldsymbol{x} \in \mathbf{R}^r,\ \sigma \in \mathbb{R}_+.$$

Etapa II: condicional en $\boldsymbol{\beta}$,

$$\begin{aligned} f(\boldsymbol{x}) &\sim N\left(\mu'_\beta(\boldsymbol{x}), \Sigma'(\boldsymbol{x},\boldsymbol{x})\right) && \text{(proceso Gaussiano)}, \\ \sigma^2 &\sim IGa(\alpha/2, \nu/2), && \alpha,\ \nu \in \mathbb{R}_+, \end{aligned}$$

donde $IGa(\alpha/2, \nu/2)$ denota la distribución gamma inversa con parámetros $\alpha/2$ y $\nu/2$, y

$$\begin{aligned} \mu'_{\boldsymbol{\beta}}(\boldsymbol{x}) &= \mathbf{h}(\boldsymbol{x})^T \boldsymbol{\beta}, \qquad \boldsymbol{\beta} \in \mathbb{R}^q, \\ \Sigma'(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) &= \rho^2 V(\boldsymbol{x}, \widetilde{\boldsymbol{x}}), \qquad \rho^2 \in \mathbb{R}_+. \end{aligned}$$

donde la función de covarianza $V(\cdot, \cdot)$ se define más adelante en (5.2).

Etapa III:

$$\boldsymbol{\beta} \sim N_q\left(\boldsymbol{b}_0,\ \rho^2 \boldsymbol{B}_0^{-1}\right),$$

donde $\boldsymbol{b}_0 \in \mathbb{R}^q$ y $\boldsymbol{B}_0$ es una matriz $q \times q$ simétrica y definida positiva. (Suponemos que $f(\cdot)$ y $\boldsymbol{\beta}$ son independientes de σ^2).

Para la función de covarianza $V(\cdot, \cdot)$, que controla el grado de suavidad del estimador de $f(\cdot)$, se ha utilizado la expresión

$$V_\theta(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) = \exp\left\{-\theta(\boldsymbol{x} - \widetilde{\boldsymbol{x}})^T \boldsymbol{W}(\boldsymbol{x} - \widetilde{\boldsymbol{x}})\right\}, \quad \boldsymbol{x},\ \widetilde{\boldsymbol{x}} \in \mathbb{R}^r,\ \theta \in \mathbb{R}_+, \tag{5.2}$$

en donde $\boldsymbol{W}$ es una matriz $r \times r$ de pesos simétrica, misma que ha sido usada por O'Hagan (1978, 1992) y otros; por lo que respecta a la matriz de pesos $\boldsymbol{W}$ se ha utilizado la forma particular (Gutiérrez-Peña & Smith (1998))

$$\boldsymbol{W} = Diag\left(z_1^2, \ldots, z_r^2\right),$$

donde

$$z_j = (n-1)^{-1} \sum_{l=1}^{n-1} \left| x_{(l+1)j} - x_{(l)j} \right|$$

y $x_{(1)j} \leq \cdots \leq x_{(n)j}$ denotan los estadísticos de orden de la j-ésima covariable $(j = 1, \ldots, r)$. Esta "escala" es adecuada en la mayoría de los casos, y produce inferencias que son invariantes ante transformaciones de escala en las covariables.

Dados los datos $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$, la distribución predictiva final para una nueva observación y_* en $\boldsymbol{x} = \boldsymbol{x}_*$ se obtiene fácilmente como se detalla a continuación. Sean $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ y $\varphi = \rho^2/\sigma^2$, y denotemos como $St(\cdot \mid a, b, c)$ a la densidad t de Student con a grados de libertad, b el parámetro de localización y c el de escala (así que tiene varianza $ac/(a-2)$). Finalmente, sea

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{h}(\boldsymbol{x}_1)^T \\ \vdots \\ \boldsymbol{h}(\boldsymbol{x}_n)^T \end{bmatrix}.$$

Entonces se tiene,

Etapa I:

$$p(y_* \mid \boldsymbol{y}) = St\left(y_* \mid \widehat{\alpha},\ \widehat{f}(\boldsymbol{x}_*),\ \widehat{\sigma}^2(\boldsymbol{x}_*)/\widehat{\alpha}\right),$$

donde

$$\begin{aligned}\widehat{\alpha} &= \alpha + (n-q),\\ \widehat{f}(\boldsymbol{x}) &= \mu_0''(\boldsymbol{x}),\\ \widehat{\sigma}^2(\boldsymbol{x}) &= \widehat{\nu}\left\{1 + \Sigma_0''(\boldsymbol{x},\boldsymbol{x})\right\}\end{aligned}$$

($\widehat{\nu},\ \mu_0''(\boldsymbol{x})$ y $\Sigma_0''(\boldsymbol{x},\boldsymbol{x})$) están definidas abajo en (5.4) y (5.6)).

Etapa II:

$$\pi\left(f(\boldsymbol{x}) \mid \boldsymbol{\beta},\ \sigma^2,\ \boldsymbol{y}\right) = N\left(f(\boldsymbol{x}) \mid \mu_\beta''(\boldsymbol{x}), \Sigma''(\boldsymbol{x},\widetilde{\boldsymbol{x}})\right),$$

con

$$\begin{aligned}\mu_\beta''(\boldsymbol{x}) &= \boldsymbol{h}(\boldsymbol{x})^T\boldsymbol{\beta} + \varphi \boldsymbol{t}_\theta(\boldsymbol{x})^T\left[\varphi\boldsymbol{V}_\theta + \boldsymbol{D}\right]^{-1}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{\beta}),\\ \Sigma''(\boldsymbol{x},\widetilde{\boldsymbol{x}}) &= \rho^2\left\{V_\theta(\boldsymbol{x},\widetilde{\boldsymbol{x}}) - \varphi\boldsymbol{t}_\theta(\boldsymbol{x})^T\left[\varphi\boldsymbol{V}_\theta + \boldsymbol{D}\right]^{-1}\boldsymbol{t}_\theta(\widetilde{\boldsymbol{x}})\right\},\end{aligned}$$

y donde $\boldsymbol{t}_\theta(\boldsymbol{x}) = \left(V_\theta(\boldsymbol{x},\boldsymbol{x}_1),\ldots,V_\theta(\boldsymbol{x},\boldsymbol{x}_n)\right)^T$, $\boldsymbol{V}_\theta = \left[V_\theta(\boldsymbol{x}_i,\boldsymbol{x}_j)\right]_{ij}$, y $\boldsymbol{D}$ es una matriz $n\times n$ diagonal que contienen los pesos relativos de las observaciones (usualmente $\boldsymbol{D} = \boldsymbol{I}$, donde $\boldsymbol{I}$ es la matriz identidad). Note que, dado el valor de φ, ni el valor de $\mu_\beta''(\boldsymbol{x})$ ni el de $\Sigma''(\boldsymbol{x},\boldsymbol{x})$ dependen de σ^2.

Además,

$$\pi\left(\sigma^2 \mid \boldsymbol{y}\right) = IGa\left(\sigma^2 \mid \widehat{\alpha}/2,\ \widehat{\nu}/2\right), \tag{5.3}$$

donde $\widehat{\alpha}$ se define igual que antes y

$$\widehat{\nu} = \nu + \boldsymbol{y}^T\left[\varphi\boldsymbol{V}_\theta + \boldsymbol{D}\right]^{-1}\boldsymbol{y} + \varphi^{-1}\boldsymbol{b}_0^T\boldsymbol{B}_0\boldsymbol{b}_0 - \boldsymbol{b}_1^T\boldsymbol{B}_1\boldsymbol{b}_1, \tag{5.4}$$

con

$$\begin{aligned}\boldsymbol{b}_1 &= \left(\boldsymbol{H}^T\left[\varphi\mathbf{V}_\theta + \mathbf{D}\right]^{-1}\boldsymbol{H} + \varphi^{-1}\boldsymbol{B}_0\right)^{-1}\left(\boldsymbol{H}^T\left[\varphi\boldsymbol{V}_\theta + \boldsymbol{D}\right]^{-1}\boldsymbol{y} + \varphi^{-1}\boldsymbol{B}_0\boldsymbol{b}_0\right),\\ \boldsymbol{B}_1 &= \boldsymbol{H}^T\left[\varphi\boldsymbol{V}_\theta + \boldsymbol{D}\right]^{-1}\boldsymbol{H} + \varphi^{-1}\boldsymbol{B}_0.\end{aligned}$$

Etapa III:

$$\pi\left(\boldsymbol{\beta} \mid \sigma^2,\ \boldsymbol{y}\right) = N_q\left(\boldsymbol{\beta} \mid \boldsymbol{b}_1, \sigma^2\boldsymbol{B}_1^{-1}\right).$$

De paso notamos que

$$\pi\left(f(\boldsymbol{x}) \mid \sigma^2,\ \boldsymbol{y}\right) = N\left(f(\boldsymbol{x}) \mid \mu_0''(\boldsymbol{x}),\ \sigma^2\Sigma_0''(\boldsymbol{x},\boldsymbol{x})\right), \tag{5.5}$$

donde

$$\begin{aligned}\mu_0''(\boldsymbol{x}) &= \boldsymbol{h}(\boldsymbol{x})^T\boldsymbol{b}_1 + \varphi\boldsymbol{t}_\theta(\boldsymbol{x})^T\left[\varphi\boldsymbol{V}_\theta + \boldsymbol{D}\right]^{-1}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{b}_1), \qquad (5.6)\\ \Sigma_0''(\boldsymbol{x},\widetilde{\boldsymbol{x}}) &= \varphi V_\theta(\boldsymbol{x},\widetilde{\boldsymbol{x}}) - \varphi^2\boldsymbol{t}_\theta(\boldsymbol{x})^T\left[\varphi\boldsymbol{V}_\theta + \boldsymbol{D}\right]^{-1}\boldsymbol{t}_\theta(\widetilde{\boldsymbol{x}}) + \boldsymbol{s}_{\varphi,\theta}(\boldsymbol{x})^T\boldsymbol{B}_1^{-1}\boldsymbol{s}_{\varphi,\theta}(\widetilde{\boldsymbol{x}}),\end{aligned}$$

con

$$\boldsymbol{s}_{\varphi,\theta}(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x}) - \varphi \boldsymbol{H}^T \left[\varphi \boldsymbol{V}_\theta + \boldsymbol{D}\right]^{-1} \boldsymbol{t}_\theta(\boldsymbol{x}).$$

La distribución inicial no informativa para $\left(\boldsymbol{\beta}, \sigma^2\right)$ se obtiene por el método de referencia

$$\pi\left(\boldsymbol{\beta}, \sigma^2\right) \propto \left\{\sigma^2\right\}^{-1},$$

que corresponde al caso de hacer $\alpha \to 0, \nu \to 0$ y $\boldsymbol{B}_0 \to \mathbf{O}$.

Para completar el modelo resta asignar valores a φ y θ. Para tener idea en la elección de los valores de estos hiperparámetros hay que ver el papel que juegan dentro del modelo. El valor de φ determina la flexibilidad del modelo para hacer frente a las posibles discrepancias entre la función de respuesta $f(\boldsymbol{x})$ y la forma paramétrica simple $P_q(\boldsymbol{x})$, mientras que θ es un parámetro de suavizamiento, y debe ser grande si se piensa que la respuesta es bastante variable y pequeño si se sabe que $f(\cdot)$ es una función suave.

5.2. Un criterio predictivo para la selección de modelos

Ahora estamos listos para describir nuestro criterio de selección.

Sea

$$\mathcal{M} = \{M_1, \ldots, M_k\}$$

una colección de modelos paramétricos de regresión y considere el problema de elegir uno con propósitos predictivos. Aquí

$$M_i = \left\{p_i\left(y \mid \boldsymbol{\beta}_i, \sigma_i^2\right), \pi_i\left(\boldsymbol{\beta}_i, \sigma_i^2\right)\right\}, \quad y \in \mathbb{R},\ \boldsymbol{\beta}_i \in \mathbb{R}^{q_i},\ \sigma_i^2 \in \mathbb{R}_+,$$

donde

$$p_i\left(y \mid \boldsymbol{\beta}_i, \sigma_i^2\right) = N\left(y \mid \boldsymbol{h}_i(\boldsymbol{x})^T \boldsymbol{\beta}_i,\ \sigma_i^2\right), \qquad \boldsymbol{x} \in \mathbb{R}^r,$$

y

$$\pi_i\left(\boldsymbol{\beta}_i, \sigma_i^2\right) = N_{q_i}\left(\boldsymbol{\beta}_i \mid \boldsymbol{b}_{0i},\ \ \sigma_i^2 \boldsymbol{B}_{0i}^{-1}\right) \times IGa\left(\sigma_i^2 \mid \alpha_i/2, \nu_i/2\right).$$

De este modo, el modelo M_i queda definido por la verosimilitud $p_i\left(y \mid \boldsymbol{\beta}_i, \sigma_i^2\right)$ y la distribución inicial $\pi_i\left(\boldsymbol{\beta}_i, \sigma_i^2\right)$ (no necesariamente propia).

La distribución inicial no informativa sobre $\left(\boldsymbol{\beta}_i, \sigma_i^2\right)$ es usualmente descrita vía la inicial de referencia

$$\pi_i\left(\boldsymbol{\beta}_i, \sigma_i^2\right) \propto \left\{\sigma^2\right\}^{-1}$$

que corresponde a $\alpha_i \to -q_i,\ \nu_i \to 0$ y $\boldsymbol{B}_{0i} \to \mathbf{0}$.

Sea

$$\boldsymbol{H}_i = \begin{bmatrix} \boldsymbol{h}_i(\boldsymbol{x}_1)^T \\ \vdots \\ \boldsymbol{h}_i(\boldsymbol{x}_n)^T \end{bmatrix}$$

y note que, para cada $i = 1, \ldots, k$,

$$p_i(y_* \mid \boldsymbol{y}) = St\left(y_* \mid \widehat{\alpha}_i,\ \widehat{f}_i(\boldsymbol{x}_*),\ \widehat{\sigma}_i^2(\boldsymbol{x}_*/\widehat{\alpha}_i)\right),$$

donde

$$\begin{aligned} \widehat{\alpha}_i &= \alpha_i + n, \\ \widehat{f}_i(\boldsymbol{x}) &= \boldsymbol{h}_i(\boldsymbol{x})^T \boldsymbol{b}_{1i}, \\ \widehat{\sigma}_i^2(\boldsymbol{x}) &= \widehat{\nu}_i \widehat{\tau}_i^2(\boldsymbol{x}), \end{aligned}$$

con

$$\begin{aligned} \boldsymbol{b}_{1i} &= \left[\boldsymbol{H}_i^T \boldsymbol{D}^{-1} \boldsymbol{H}_i + \boldsymbol{B}_{0i}\right]^{-1} \left(\boldsymbol{H}_i^T \boldsymbol{D}^{-1} \mathbf{y} + \boldsymbol{B}_{0i} \boldsymbol{b}_{0i}\right), \\ \widehat{\nu}_i &= \nu_i + \boldsymbol{y}^T \boldsymbol{D}^{-1} \boldsymbol{y} + \boldsymbol{b}_{0i}^T \boldsymbol{B}_{0i} \boldsymbol{b}_{0i} - \boldsymbol{b}_{1i}^T \left[\boldsymbol{H}_i^T \boldsymbol{D}^{-1} \boldsymbol{H}_i + \boldsymbol{B}_{0i}\right]^{-1} \boldsymbol{b}_{1i} \\ \widehat{\tau}_i^2(\boldsymbol{x}) &= 1 + \boldsymbol{h}_i(\boldsymbol{x})^T \left[\boldsymbol{H}_i^T \boldsymbol{D}^{-1} \boldsymbol{H}_i + \boldsymbol{B}_{0i}\right]^{-1} \mathbf{h}_i(\boldsymbol{x}). \end{aligned}$$

Sea M_t el modelo "verdadero", y suponga que $(y_1, \boldsymbol{x}_1), \ldots, (y_n, \boldsymbol{x}_n)$ es una muestra de observaciones independientes de M_t. Si M_t fuera conocido y considerando a la utilidad puntaje logarítmico

$$u_t(M_i) = \int \log p_i(y_* \mid \boldsymbol{y})\, p_t(y_* \mid \boldsymbol{y})\, dy_*, \tag{5.7}$$

donde $p_i(y_* \mid \boldsymbol{y})$ y $p_t(y_* \mid \boldsymbol{y})$ denotan la densidad predictiva final de una observación futura Y_* dado $\boldsymbol{x}{=}\boldsymbol{x}_*$ bajo el i-ésimo y el verdadero modelo, respectivamente. Entonces, un criterio predictivo razonable para la selección de un modelo sería, dados dos modelos M_i y M_j en $\mathcal{M}$, M_i será más preferido que M_j si, y sólo si $u_t(M_i) > u_t(M_j)$.

En las aplicaciones no conocemos cuál es el modelo verdadero, así que típicamente $p_t(y_* \mid \boldsymbol{y})$ no está disponible. San Martini y Spezzaferri (1984) abordan este problema desde un enfoque $\mathcal{M}$-cerrado, pero en la práctica esta suposición también es poco realista. En la sección anterior vimos cómo Bernardo y Smith discuten el problema de la selección de modelos desde un punto de vista de la teoría de la decisión, lo que involucra considerar una función de utilidad. En el enfoque predictivo elegimos la utilidad asociada con la elección de una densidad predictiva de una observación futura Y_* dado $\boldsymbol{x}{=}\boldsymbol{x}_*$. Bernardo (1979) recomienda el uso de una función de utilidad puntaje logarítmico. San Martini y

Spezzaferri (1984) adoptan esta idea para calcular la distribución predictiva final $p_t(y_* \mid \boldsymbol{y})$ usando toda la información disponible de los modelos (promedio de modelos) para lo cual necesitan calcular la probabilidad de que $P(M_i \mid \boldsymbol{y})$, es decir, la probabilidad de que M_i sea el modelo verdadero una vez observados los datos. Sin embargo esta distribución no está bien definida si la distribución inicial sobre los parámetros del modelo en consideración $\left(\boldsymbol{\beta}_i, \sigma_i^2\right)$ es impropia.

En cambio, el enfoque semiparamétrico a este problema comienza por considerarlo como un problema de decisión con los siguientes elementos.

Espacio de decisión:

$$\mathcal{D} = \mathcal{M}.$$

Espacio de estados:

$$\mathcal{F} = \left\{\left(f, \sigma^2\right) : f \text{ es una función suave en } \mathbb{R}^r, \text{ y } \sigma^2 \in \mathbb{R}_+\right\}.$$

Distribución inicial sobre $\mathcal{F}$:

Etapa I:

$$\begin{aligned} f(\boldsymbol{x}) &\sim N\left(\mu_\beta'(\boldsymbol{x}), \Sigma'(\boldsymbol{x}, \boldsymbol{x})\right) \qquad \text{Proceso Gaussiano} \\ \sigma^2 &\sim IGa(\alpha/2, \nu/2), \end{aligned}$$

donde

$$\begin{aligned} \mu'_{\boldsymbol{\beta}}(\boldsymbol{x}) &= \boldsymbol{h}(\boldsymbol{x})^T \boldsymbol{\beta} \\ \Sigma'(\boldsymbol{x}, \widetilde{\boldsymbol{x}}) &= \rho^2 \boldsymbol{V}_\theta(\boldsymbol{x}, \widetilde{\boldsymbol{x}}). \end{aligned}$$

Etapa II:

$$\boldsymbol{\beta} \sim N_q\left(\boldsymbol{b}_0,\ \rho^2 \boldsymbol{B}_0^{-1}\right)$$

Función de Utilidad sobre $\mathcal{D} \times \mathcal{F}$:

$$u\left[M_i, \left(f, \sigma^2\right)\right] = \int \log p_i(y_* \mid \mathbf{y})\, N\left(y_* \mid f(\boldsymbol{x}_*), \sigma^2\right) dy_*.$$

La solución a este problema es elegir el modelo que maximice la utilidad esperada final.

En efecto hemos encajado a $\mathcal{M}$ en una clase de modelos más rica, como lo es la formada por los modelos semiparamétricos, que ya hemos discutido en la sección anterior.

Usando (5.3) y (5.5), la utilidad esperada final del modelo M_i es

$$\begin{aligned} \overline{u}(M_i) &= E_{f,\sigma^2|y}\left\{u\left[M_i \mid \left(f, \sigma^2\right)\right]\right\} \\ &= \int \log p_i(y_* \mid \boldsymbol{y})\, \widehat{p}(y_* \mid \boldsymbol{y})\, dy_* \end{aligned} \tag{5.8}$$

donde

$$p_i\left(y_* \mid \boldsymbol{y}\right) = St\left(y_* \mid \widehat{\alpha}, \widehat{f}\left(x_*\right), \widehat{\sigma}^2/\widehat{\alpha}\right).$$

El modelo M_i es preferible al modelo M_j si, y sólo si $\overline{u}\left(M_i\right) > \overline{u}\left(M_j\right)$. Note que la utilidad esperada final depende de un valor particular $\boldsymbol{x}$ sobre el que se requiere hacer la predicción. Se propone el uso del siguiente criterio. Calcular $\overline{u}\left(M_i\right)$ para cada una de las observaciones $\boldsymbol{x}_j$ $(j = 1, \ldots, n)$ y después usar el promedio de estas utilidades esperadas como criterio de comparación.

Capítulo 6

Aplicaciones

Antes de aplicar la metodología desarrollada en el capítulo anterior, haremos algunas consideraciones acerca de la implementación del templado simulado.

6.1. Implementación del criterio vía templado simulado

Con r covariables existen 2^r posibles modelos, típicamente muchos para completar la evaluación de las utilidades esperadas. Al aplicar el templado simulado tenemos buenas probabilidades de hallar uno de los *mejores modelos.* Para ello, utilizamos un vector binario de dimensión r al que llamamos γ y que identifica al modelo de la siguiente forma, $\gamma\,[i] = 1$ si la i-ésima covariable está presente en el modelo y 0 en otro caso. Dada una configuración inicial γ_i, γ_{i+1} es una configuración vecina si puede obtenerse a partir de γ_i por medio de alguno de los siguientes movimientos:

1. *Agregar una covariable:* consiste en elegir de manera aleatoria un 0 en γ_i y cambiarlo por un 1.
2. *Eliminar una covariable:* consiste en elegir de manera aleatoria un 1 en γ_i y cambiarlo por un 0.
3. *Intercambiar dos covariables:* consiste en elegir de manera aleatoria e independiente un 1 y un 0 en γ_i, y cambiar los valores.

Cada movimiento se elige con probabilidad $\frac{1}{3}$, excepto en los casos extremos cuando están contenidas todas las covariables y cuando no hay covariables seleccionadas. En estos casos con probabilidad 1 se eligen los movimientos *eliminar una covariable* y *agregar una covariable,* respectivamente. En cada iteración se calcula la diferencia δ entre los promedios de $\overline{u}(\gamma_i)$ y $\overline{u}(\gamma_{i+1})$ para los valores observados de $\boldsymbol{x}$**,** donde $\overline{u}(\gamma_i)$ es la utilidad esperada final del modelo M_i que está representado por medio γ_i, y que se calcula usando la expresión (5.8). Si $\delta > 0$ entonces, con probabilidad 1 se acepta γ_{i+1}, de otro modo se

acepta con probabilidad $\exp(\delta/T)$, donde T es el parámetro de la temperatura. En la implementación se ha elegido un esquema de enfriamiento geométrico con tasa de 0.05. Cada n pasos se calcula la tasa de aceptación (ta), que es la proporción de los n pasos que son aceptados, y se detiene si $ta = 0$, *i.e.* cuando el sistema está tan frío que prácticamente ha dejado de moverse.

Dado γ^* el vector que representa al modelo con máxima utilidad esperada promedio, obtenido a través del método del templado simulado, el programa implementa también el *recalentado* del sistema, que consiste en ejecutar nuevamente el algoritmo partiendo de γ^* y con una menor temperatura inicial. Este procedimiento permite reducir las posibilidades de quedar atrapado en un mínimo local.

A continuación, mostramos dos aplicaciones del criterio semiparamétrico predictivo Bayesiano desarrollado en el capítulo anterior. El mejor modelo se halla vía el templado simulado. El código en *R* para la solución del problema, se halla en el apéndice A.3.

6.2. Ejemplo 1. Datos de Hald

Como una primera aplicación consideremos el conjunto de datos de Hald presentados y analizados en Draper & Smith (1988), y Gutiérrez-Peña (1997) entre otros. Este conjunto de datos se usa en particular por el hecho de que exhibe algunas de la dificultades que con mayor frecuencia aparecen en el análisis de regresión y porque muestra una de las ventajas de aplicar el templado simulado. El conjunto consta de 13 observaciones que relacionan el calor emitido en el proceso de endurecimiento del cemento (en calorías por gramo) con cuatro covariables, cada una de las cuales mide el porcentaje de la composición de cuatro ingredientes. Se asume una distribución inicial no informativa. Se establecen los valores de $\varphi = 2$, $\theta = -\log(0.75)$ sugeridos en Gutiérrez-Peña (1997), además $\mu'_\beta(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{31} x_3 + \beta_4 x_4$.

Ejecutamos el algoritmo de templado simulado con una configuración inicial saturada y después de 14 iteraciones obtenemos un vector γ^* de utilidad esperada máxima -2.4297 y 3 covariables elegidas $\{x_1, x_2, x_4\}$. Se ejecuta posteriormente un recalentamiento con temperaratura $T_0/3 = 2/3$. El recalentamiento se detiene después de 3 iteraciones más, dando en este caso el mismo modelo seleccionado durante la primera etapa. Las tres covariables seleccionadas explican el 98 % de la variación en la cantidad de calor emitida.

La figura 6.1 (izquierda) muestra la utilidad esperada a cada iteración, primero la secuencia original y después seguida del recalentamiento. Note cómo es posible seleccionar configuraciones con una utilidad esperada menor, lo que permite no quedar atrapado en un mínimo local. El número de covariables seleccionadas está dada en la figura 6.1 derecha. También partimos de distintas configuraciones, por ejemplo en la que sólo se incluye una covariable y de acuerdo a la calibración utilizada se obtiene el mismo modelo.

Los datos de Hald han sido analizados por varios autores entre los que podemos mencionar, George

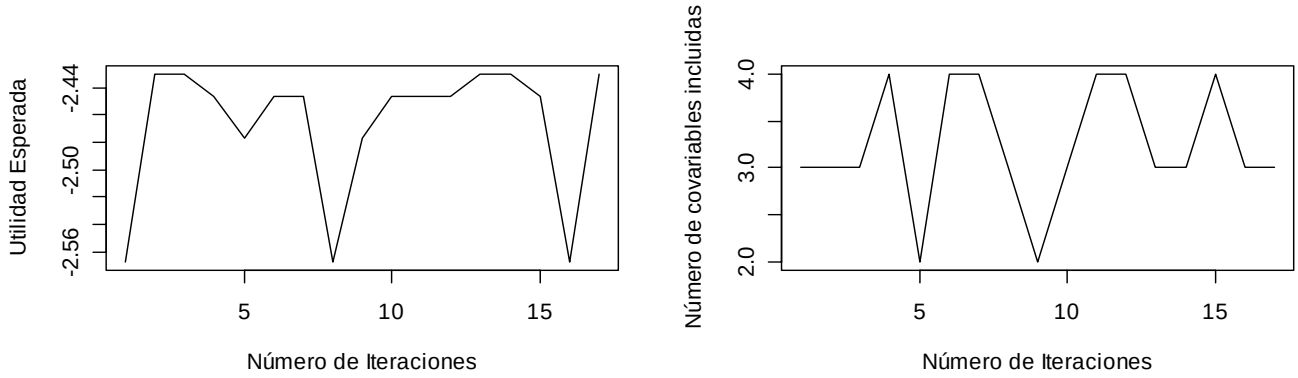


Figura 6.1: Resultados del templado simulado con recalentamiento. (izq) Utilidad Esperada en cada iteración. (der) Número de unos en el vector γ en cada iteración.

& McCulloch(1993), Kuo & Mallick (1994), Laud & Ibrahim (1995), Geweke (1996), y Berger & Pericchi (1996). Aunque en todos estos trabajos el modelo $\{x_1, x_2\}$ es reportado como uno de los mejores, el conjunto de variables $\{x_1, x_2, x_4\}$ seleccionado coincide con el seleccionado por otros criterios como el AIC y el de raíces latentes, más aún no hace uso de algunos otros criterios como el de parsimonia para penalizar la dimensión del modelo. Finalmente si nos restringimos a los modelos con a lo más dos covariables, entonces el modelo seleccionado es $\{x_1, x_2\}$ con una utilidad esperada de -2.4764. Este modelo es, como se señala en Gutiérrez-Peña (1997), de entre todos los modelos con exactamente 2 covariables el más cercano a la utilidad esperada máxima final -2.3321 alcanzada por el modelo semiparamétrico. Con esto podemos notar que el modelo semiparamétrico provee un punto de referencia desde el cual no solamente se puede valorar el mejor modelo, sino que también que tan lejos se está del modelo "verdadero".

Como una aplicación final consideremos el siguiente ejemplo con 13 covariables, en el cual una búsqueda exhaustiva ya no resulta factible.

6.3. Ejemplo 2. Reducción en el desarrollo intelectual de niños de 7 años con exposición prenatal a plomo

Con el objetivo de comprobar que la exposición prenatal a niveles bajos de plomo está asociado con un bajo desarrollo intelectual de los niños durante sus primeros 10 años de vida, Schnaas *et al.* (2006) realizaron un estudio prospectivo en niños que nacieron en el Instituto Nacional de Perinatología (Ciudad de México) entre 1987 y 1992, a los cuales se les dio seguimiento hasta el año 2002. Entre las

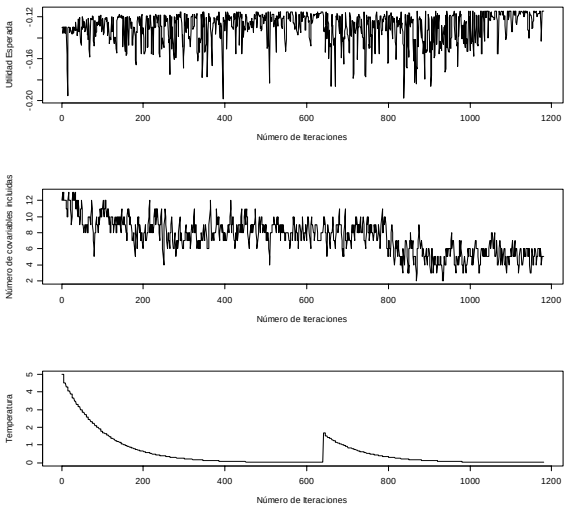


Figura 6.2: Resultados del Templado Simulado con recalentamiento para el caso de los niños.

variables que se midieron destacan, el IQ del niño (variable de respuesta y), sexo, peso al nacer (x_1), IQ de la madre (x_2), logaritmo del plomo en la sangre de la madre durante los trimestres segundo (x_3), tercero (x_4) de embarazo, logaritmo del plomo en la sangre de la madre en las semanas 20 (x_5), 28 (x_6) y 36 (x_7), el logaritmo del plomo en la sangre del niño a las edades de 1 (x_8), 2 (x_9), 3 (x_{10}), 4 (x_{11}), 5 (x_{12}) años y el promedio del logaritmo del plomo en la sangre entre los 5 y 6 años (x_{13}). En este ejemplo se consideran 62 niños de 7 años de edad, 31 hombres y 31 mujeres. Se asume una distribución inicial no informativa. Los valores de los hiperparámetros $\varphi = 10$, $\theta = -\log(0.75)$ fueron elegidos mediante una técnica Bayes-empírica, y $\mu'_\beta(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} + \beta_{12} x_{12} + \beta_{13} x_{13}$, para la variable sexo en cada uno de sus dos niveles.

En el caso del nivel para los niños, ejecutamos el algoritmo de templado simulado con una configuración inicial saturada y después de 1100 iteraciones obtenemos el vector γ^* de utilidad esperada final máxima -0.1145 y 5 covariables elegidas $\{x_2, x_3, x_4, x_5, x_7\}$. En este caso la utilidad esperada final del modelo semiparamétrico -0.087. Las 5 covariables seleccionadas explican cerca del 50 % de la variación en el IQ del niño a la edad de 7 años. Del modelo lineal encontrado, se puede concluir que exposiciones prenatales altas de plomo alrededor de la semana 20 están asociadas con un IQ bajo en los niños ($\beta_5 = -4.86$) y que un IQ alto de la madre está asociado con un IQ alto del hijo ($\beta_2 = 0.28$). Estas inferencias son parecidas a las que se obtienen en Schnaas *et al.* (2006). La figura 6.2 muestra los resultados del templado simulado para este caso. Note que en el proceso de recalentamiento se mejora el modelo. Adicionalmente a los resultados aquí mostrados, se realizaron otras corridas partiendo de

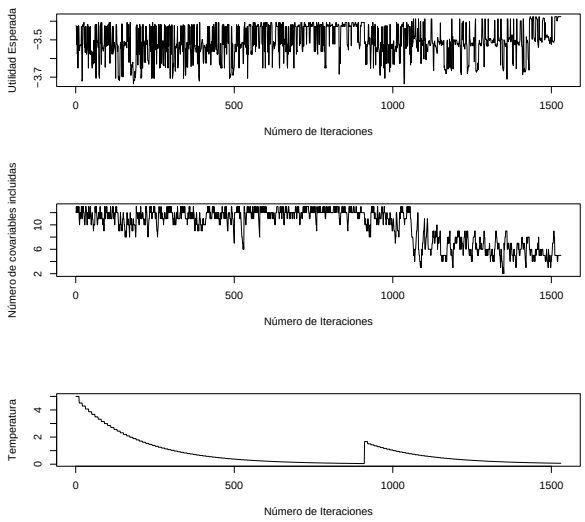


Figura 6.3: Resultados del Templado Simulado con recalentamiento para el caso de los niñas.

distintas configuraciones iniciales. Por ejemplo, partiendo de una configuración con una sola covariable, con una selección al azar, y con la obtenida de aplicar un método clásico como stepwise. En todos los casos el modelo hallado resulto ser el mismo.

Por otra parte, en el caso de las niñas, el algoritmo de templado simulado con una configuración inicial saturada después de 1530 iteraciones obtenemos el vector γ^* de utilidad esperada máxima -3.36085 y 5 covariables elegidas $\{x_1, x_2, x_6, x_8, x_9\}$. Para este caso la utilidad esperada final del modelo semiparamétrico es -3.0593. Las 5 covariables seleccionadas explican cerca del 37 % de la variación en el IQ. Del modelo lineal encontrado, se puede concluir que exposiciones prenatales altas de plomo alrededor de la semana 28 están asociadas con un IQ bajo en las niñas ($\beta_6 = -4.42$) y que un IQ alto de la madre está asociado con un IQ alto ($\beta_2 = 0.47$). Estas inferencias son parecidas a las que se obtienen en Schnaas *et al.* (2006). La figura 6.3 muestra los resultados del templado simulado para este caso. Note que en el proceso de recalentamiento se mejora el modelo. Adicionalmente a los resultados aquí mostrados, se realizaron otras corridas partiendo de distintas configuraciones iniciales. Por ejemplo, partiendo de una configuración con una sola covariable, con una selección al azar, y con la obtenida de aplicar un método clásico como stepwise. En todos los casos el modelo hallado resulto ser el mismo.

Conclusiones y sugerencias

En esta tesis hemos presentado un criterio predictivo Bayesiano para la comparación de modelos, y vimos cómo es posible aplicarlo al problema de selección de variables en modelos de regresión lineal normal. El modelo es elegido con respecto a una clase más rica como lo es la de los modelos semiparamétricos. Dado que la cantidad de modelos a comparar es relativamente grande, el algoritmo del templado simulado resultó ser una herramienta útil en la comparación de modelos. Aunque en general no podemos garantizar que el modelo final que se obtiene sea en efecto el óptimo, hemos visto a través de ejemplos sencillos cómo el algoritmo garantiza que el modelo final es uno de los mejores. La única forma de comprobarlo es a través de una búsqueda exhaustiva que no siempre resulta factible. El criterio propuesto, por un lado, resuelve el problema de establecer distribuciones iniciales sobre cada uno de los modelos a comparar, y por otro lado, el templado simulado ataca la explosión combinatoria de todas las posibilidades.

Por lo que respecta a la calibración del templado simulado sugerimos:

- Comenzar con una configuración inicial saturada y una temperatura relativamente baja, por ejemplo 2 y con un número de iteraciones de 2 veces el número covariables en el modelo. Esta última consideración, aunque no garantiza que todas las covariables sean examinadas, permite tener buenas probabilidades de mover todas las covariables. Y finalmente no utilizar el recalentado en una corrida preliminar.

- En la versión del templado simulado que se presenta, el criterio de paro es a través de la tasa de aceptación; de hecho, el programa termina cuando la tasa de aceptación (ta) es exactamente cero. En una corrida exploratoria se sugiere cambiar el criterio por ($ta < 0.20$).

- En principio, el esquema de enfriamiento geométrico resultó satisfactorio para los resultados que se esperaban obtener; de no ser el caso se sugiere un esquema de enfriamiento logarítmico. En una corrida preliminar se sugiere aumentar la tasa de enfriamiento.

- De entrada quizá los resultados obtenidos no parezcan satisfactorios, pero lo importante es ver que el programa trabaja de manera adecuada y además medir el tiempo de ejecución.

- Para una corrida final se sugiere iniciar con una temperatura inicial alta (alrededor de 15), con un número de iteraciones de 10 veces el número de covariables en el modelo y utilizar la opción de recalentado. Esto puede absorber mucho tiempo de cómputo, dependiendo del número de covariables y del número de observaciones que se tengan. Se puede reducir el número de iteraciones, pues consideramos que en la corrida final el esquema de enfriamiento es más útil que la cantidad de iteraciones que se realizan.

- Dependiendo del tiempo del que se disponga, se pueden variar los valores de la tasa de aceptación y enfriamiento, pero para mejores resultados se recomienda una tasa de enfriamiento de 0.05 y que el criterio de paro sea $ta = 0$.

- Para tener certeza de los resultados obtenidos, se sugiere también realizar varias corridas partiendo de distintas configuraciones iniciales; por ejemplo, partir de un modelo en el que esté solo alguna de las covariables. También es posible tomar como configuración inicial la que se obtenga de aplicar alguno de los métodos clásicos como el Stepwise o Forward.

Apéndice A. Código en R de los programas usados

A.1. Aplicaciones del capítulo 1

```
# # # # ESQUEMAS DE ENFRIAMIENTO # # # #
#Esquema de enfriamiento geométrico
Geom<-function(i,To,Tf,M)
{
     return(To*(1-0.05)^(i+1))
}
#Esquema de enfriamiento Logarítmico
Logar<-function(i,To,Tf,M)
{
     return(1/log(1+i))
}
#Esquema de enfriamiento Lundy & Mees
LundyMees<-function(i,To,Tf,M)
{
     To/(1+To*(i-1)*(1-(Tf/To)^(1/(M-1))))
}

# La siguiente función implementa el algoritmo del templado simulado
# para maximizar una función f, utilizando distintos esquemas de enfriamiento:
# cool. Utiliza xo como la aproximación inicial, To y Tf son la temperatura
# inicial y final respectivamente, del sistema. M el número máximo
# de iteraciones.
SimulatedAnnealing<-function(xo,To,Tf,M,n,d,cool)
```

```
{
#Variables iniciales
x<-c(); Te<-c(); maximo<-c();
#Establecer las condiciones iniciales del sistema
x[1]<-c(xo) # Aproximación inicial a la solución
Te[1]<-c(To) # Temperatura inicial de sistema
t=1 # Tiempo
iter_max=n # Iteraciones máximas: Equilibrio a la Te[i] del sistema
x_opt<-xo
#Esquema de enfriamiento
if(cool==1) ## Esquema de enfriamiento geométrico
 fcool<-Geom
else if(cool==2)
 fcool<-Logar
else if(cool==3){
 fcool<-LundyMees
 iter_max=1
 }
 repeat{
     iter_cont=0; #contador de iteraciones antes de alcanzar el equilibrio
    repeat{
         tot_iter=tot_iter+1;
         y=Neighborhood(xo,d)
         delta=f(y)-f(xo)
         if(delta>=0){
                 xo=y
                 if(f(x_opt)<f(xo)) x_opt<-xo #¿Es el óptimo hasta ahora?
                         }
         else{
             u=runif(1);
             if(u<exp(delta/Te[t]))xo=y
             }
         iter_cont=iter_cont+1;
         if(iter_cont>iter_max) break;
     }
```

```
        t=t+1; #Iteración siguiente
        x[t]<-c(xo);
        Te[t]<-c(fcool(t,To,Tf,M))
        maximo[t]<-c(f(xo));
        if((t>M)||(Te[t]<Tf))break;
    }
    #par(mfrow=c(1,2))
    #plot(x,xlab="Iteración",ylab="Óptimo aparente")
    plot.ts(maximo,xlab="Iteración",ylab="Máximo aparente de la función")
    print("óptimo: ")
    print(x_opt)
    print("Máximo de la función: ")
    print(f(x_opt))
   }

   # * LAS SIGUIENTES FUNCIONES SON DEFINIDAS POR EL USUARIO: DEPENDEN DEL PROBLEMA *
#
   #Ejemplo 1: Maximizar una función
   # Estructura de vecindad
   Neighborhood<-function(x,d)
   {
    a=max(x-d,0);
    b=min(x+d,1); ##Elegir una configuración nueva (vecina)
    y=runif(1,min=a,max=b);
    y
   }
   # Función objetivo
   f=function(z)
   {
   (cos(50*z)+sin(20*z))^2
   }
   curve(f,0,1,n=100,xlab="x", ylab="f(x)",ylim=c(-0.5,5))
   title("(cos(50*z)+sin(20*z))^2")
   SimulatedAnnealing(0.5,1,0.05,1000,100,0.5,1)
```

```
#Ejemplo 2: Minimizar una función R^2:⟶R
# Estructura de vecindad
z<-array(dim=2)
Neighborhood<-function(z,d)
{
 w<-array(dim=2)
 a=max(z[1]-d,-1);
 b=min(z[1]+d,1); ##Elegir una configuración nueva (vecina)
 w[1]<-c(runif(1,min=a,max=b));
 a=max(z[2]-d,-1);
 b=min(z[2]+d,1);
 w[2]<-c(runif(1,min=a,max=b));
 w
}

# Función objetivo
f<-function(z)
{
-(z[1]^2+2*z[2]^2-0.3*cos(3*pi*z[1])-0.4*cos(4*pi*z[2])+0.7)
}
SimulatedAnnealing(c(0.5,0.5),1,0.010,1000,100,0.5,3)


#Ejemplo 3: Hallar una cubierta de vértices mínima para una gráfica

## Leer la grafica en forma de Matriz de incidencias
S<-read.table('c:/grafica.txt',nrows=12)
G<-matrix(,nrow=12,ncol=12)
G[,1]<-S$V1    G[,2]<-S$V2        G[,3]<-S$V3        G[,4]<-S$V4
G[,5]<-S$V5    G[,6]<-S$V6        G[,7]<-S$V7        G[,8]<-S$V8
G[,9]<-S$V9    G[,10]<-S$V10    G[,11]<-S$V11    G[,12]<-S$V12


# Estructura de vecindad
z<-array(dim=12)
Neighborhood<-function(z,d)
```

```
{
 j=floor(12*runif(1)+1)
 z[j]<-(z[j]+1)%%2
 print(z)
 z
}
# Función objetivo
f<-function(z)
{
 V<-z%*%G
 EVC=0; #número de aristas no cubiertas
 for(i in 1:12)
 if(V[i]==0) EVC=EVC+1
 return (-t(z)%*%z-2*EVC)
}
SimulatedAnnealing(c(1,0,0,0,0,0,0,0,0,0,0,0),2,0.10,2000,5,0.5,1)
```

A.2. El problema del panadero

```
#################### ESQUEMAS DE ENFRIAMIENTO ##################################
#Esquema de enfriamiento geométrico
Geom<-function(i,To,Tf,M)
{
 return(To*(1-0.05)^(i+1))
}
#Esquema de enfriamiento Logarítmico
Logar<-function(i,To,Tf,M)
{
 return(1/log(1+i))
}
#Esquema de enfriamiento Lundy & Mees
LundyMees<-function(i,To,Tf,M)
{
 To/(1+To*(i-1)*(1-(Tf/To)^(1/(M-1))))
}
```

```
# La siguiente función implementa el algoritmo del templado simulado para
# maximizar una función f, utilizando distintos esquemas de enfriamiento:cool.
# Utiliza xo como la aproximación inicial, To y Tf son la temperatura inicial
# y final respectivamente, del sistema. M el número máximo
# de iteraciones.
SimulatedAnnealing<-function(xo,To,Tf,M,n,d,cool)
{
#Establecer las condiciones iniciales del sistema
t=0 # Tiempo
Te<-To # Temperatura inicial de sistema
iter_max=n # Iteraciones máximas: equilibrio a la Te[i] del sistema
tot_iter=0 # Total de iteraciones
x_opt<-xo # óptimo
f_opt<-f(x_opt) # máximo
max_apar<-c()
opt_apar<-c()
aceptados<-0;
#Esquema de enfriamiento
if(cool==1) ## Esquema de enfriamiento geométrico
     fcool<-Geom
else if(cool==2)
     fcool<-Logar
else if(cool==3){
     fcool<-LundyMees
               iter_max=1
             }
 repeat{
     iter_cont=0; #contador de iteraciones antes de alcanzar el equilibrio
     repeat{
         tot_iter=tot_iter+1
         y<-vecindad(xo,d)
         f_xo<-f(xo)
         f_y<-f(y)
         delta=f_y-f_xo
         if(delta>=0){xo=y
```

```
                        if(f_opt<f_y){x_opt=y #¿Es el óptimo hasta ahora?
                        f_opt=f_y     }
                        }else{
                            u=runif(1);
                            if(u<exp(delta/Te))xo=y
                            }
          iter_cont=iter_cont+1;
          if(iter_cont>iter_max) break;
          }
      t=t+1;
      opt_apar[t]<-x_opt #Iteración siguiente
      max_apar[t]<-f_opt;
      Te<-fcool(t,To,Tf,M)
      iter_max=iter_max+1
      if((t>M)||(Te<Tf))break;
      }

 par(mfrow=c(1,2))
 plot.ts(max_apar,xlab="Iteracion",ylab="Valor Máximo de f")
 plot.ts(opt_apar,xlab="Iteracion",ylab="Valor Óptimo")
 print('x*=')
 print(x_opt)
 print('f*=')
 print(f_opt)
}

# Estructura de vecindad
vecindad<-function(x,d)
{
 a=max(x-d,1000);
 b=min(x+d,2001); ##Elegir una configuración nueva (vecina)
 y=ceiling(runif(1,min=a,max=b));
 y
}
suma<-function(a,b)
```

```
{
 if(b<a) return (0)
 su<-0;
 for(d in a:b)
 su<-su+(b-d)
 return (su/1001)
}
suma1<-function(a,b)
{
 if(b<a) return (0)
 su<-0;
 for(d in a:b)
 su<-su+(d-a)
 return (su/1001)
}
# Función objetivo
f=function(n)
{
0.40*n-0.60*suma(1000,n)-0.80*suma1(n,2000)
}
SimulatedAnnealing(1500,1,0.1,25,10,10,1)
```

A.3 Implementación del criterio vía templado simulado

```
# El siguiente programa implementa el criterio predictivo semiparamétrico
# Bayesiano para la selección de Modelos de regresión vía el templado simulado.
# Lectura de los datos
datos<-read.delim('H:/programas/datos/iq1.dat', header=FALSE,check.names=FALSE)
# Variables
n=31 # n = número de observaciones
r=13 # r = número de covariables
theta=-log(0.75) # theta= parámetro de suavizamiento del
                        # estimador de la función de respuesta f
fi=10 # fi = parámetro de flexibilidad del modelo
q<-14 # r = numero de parámetros se incluye uno más por el
```

```
      # término independiente
# movimientos elementales descritos en la sección 6.1
# 1. Eliminar un regresor.
remover<-function(gama)
{
 dentro<-array()
 j<-1
 for(i in 2:q)
 if(gama[i]){
     dentro[j]<-i
     j<-j+1
     }
 r=floor(runif(1,1,j))
 gama[dentro[r]]<-0;
 rm(dentro)
 gama
}
# 2. Agregar un regresor.
add<-function(gama)
{
 fuera<-array()
 j<-1
 for(i in 2:q)
     if(!gama[i]){
                 fuera[j]<-i
            j<-j+1
        }
 r=floor(runif(1,1,j))
 gama[fuera[r]]<-1;
 rm(fuera)
 gama
}
# 3. Intercambiar un par de regresores
inter<-function(gama)
{
```

```
 fuera<-array()
 j<-1
 dentro<-array()
 k<-1
 for(i in 2:q)
     if(!gama[i]){
              fuera[j]<-i
         j<-j+1
         }else{
             dentro[k]<-i
             k<-k+1
             }
 r=floor(runif(1,1,j))
 gama[fuera[r]]<-1;
 r=floor(runif(1,1,k))
 gama[dentro[r]]<-0;
 gama
}
# La siguiente función regresa una configuración vecina del vector gama
# que representa al modelo actual
vecindad<-function(gama)
{
if(sum(gama)==q) {
    gama<-remover(gama)
    return(gama)
    }
if(sum(gama)==1) {
     gama<-add(gama)
    return(gama)
    }
u<-runif(1,0,1)
if(u<1/3){
    gama<-add(gama)
     return(gama)
     }
```

```
if((1/3<=u)&&(u<2/3)){
 gama<-inter(gama)
 return(gama)
 }
if((2/3<=u)&&(u<=1)){
     gama<-remover(gama)
     return(gama)
     }
gama
}
#Esquema de enfriamiento geométrico
Geom<-function(i,To)
{
 return(To*(1-0.05)^(i+1))
}

# La siguiente función implementa el algoritmo de templado simulado para
# maximizar la función f. Utiliza un esquema de enfriamiento geométrico con
# tasa de 0.05. xo denota la aproximación inicial y To la temperatura inicial
SimulatedAnnealing<-function(xo,To,n,cool,reheat)
{
 tot_iter=0 # Total de iteraciones
 num_reg<-c() # Número de regresores en el modelo
 Temp<-c() # Temperatura
 maximo<-c() # Colección de configuraciones máximas
    repeat{
         #Establecer las condiciones iniciales del sistema
         t=0 # Tiempo
         Te<-To # Temperatura inicial de sistema
         iter_max=n # Total de iteraciones antes de alcanza el equilibrio
         x_opt<-xo # optimo
        f_opt<-UE(x_opt) # máximo
             repeat{
         iter_cont=0; # contador de iteraciones antes de alcanzar el
                        # equilibrio a la temperatura Te.
```

```
    mov_propuesto<-iter_max # movimientos propuestos,
                                # sirve para calcula la tasa de aceptación.
    f_xo<-f_opt # Valor de la función en la configuración inicial
        repeat{
            tot_iter=tot_iter+1 # Contador del total de iteraciones
            y<-vecindad(xo) # Se elige una configuración vecina
            f_y<-UE(y) # Se evalúa la función objetivo en la
                        # nueva configuración
            delta=f_y-f_xo
            if(delta>0){ xo=y # Se acepta entonces, con probabilidad 1 y
                           f_xo=f_y
                              if(f_opt<f_y){ # se analiza si ha sido la mejor
                                  x_opt=y # configuración visitada
                                  f_opt=f_y
                                  }
                              }else{ # De otro modo se sigue la regla de
                              u=runif(1); # aceptación de Metropolis.
                              if(u<exp(delta/Te)) {
                              xo=y
                              f_xo=f_y
                              }else mov_propuesto<-mov_propuesto-1
                          }
            iter_cont=iter_cont+1;
            num_reg[tot_iter]<-sum(xo)-1
            Temp[tot_iter]<-Te
            maximo[tot_iter]<-f_xo
            if(iter_cont>=iter_max) break; # Si ya se alcanzo el equilibrio
                                                  # a la temperatura Te?
            }
t=t+1; # Se disminuye la temperatura
Te<-Geom(t,To) # de acuerdo al esquema elegido
print(Te)
ta<-mov_propuesto/n # Se calcula la tasa de aceptación
xo=x_opt
if(ta<0.20)break; # si ya no aceptan más cambios (AR<0.20)
```

```
    }
    if(reheat){
    To=To/3
    reheat=FALSE
    print('x*=')
    print(x_opt)
    print('f*=')
    print(f_opt)
    print('Iteraciones realizadas:')
    print(tot_iter)
    }else{
            break
    }
 }
 par(mfrow=c(3,1))
 plot.ts(maximo,xlab='Número de Iteraciones',ylab='Utilidad Esperada')
 plot.ts(num_reg,xlab='Número de Iteraciones',ylab='Número de covariables incluidas')
 plot.ts(Temp,xlab='Número de Iteraciones',ylab='Temperatura')
 print('x*=')
 print(x_opt)
 print('f*=')
 print(f_opt)
 print('total de iteraciones:')
 print(tot_iter)
}
# AQUI EMPIEZA LA IMPLEMENTACIÓN DEL CRITERIO
D<-diag(1,ncol=n,nrow=n) #Matriz Identidad
P<-array(0,dim=6);

# * * * * * * * * * z(j) * * * * * * * * * * #
z<-function(j)
{
o<-X[,j]
o<-sort(o)
suma=0
```

```
for(l in 1:(n-1))
suma=suma+abs(o[l+1]-o[l])
suma/(n-1)
}
# * * * * * * * * * V_theta * * * * * * * * * * * * #
V<-function(x1,x2)
{
 return(exp(-theta*(t(x1-x2)%*%solve(W)%*%(x1-x2))))
}
# * * * * * * * * * t_theta * * * * * * * * * * * * #
t_theta<-function(x)
{
tv<-array(dim=n)
for(j in 1:n)
{
 tv[j]<-V(x,X[j,])
 }
tv
}
# * * * * * * * * * h:R^r->R^q * * * * * * * * * * * #
h<-function(x)
{
 tv<-array(dim=q)
 tv<-c(1,x)
 diag(1,ncol=q,nrow=q)%*%tv
}
# * * * * * * * * * s_fi_theta * * * * * * * * * * * * * * * * * * #
s_fi_theta<-function(x)
{
h(x)-fi*t(H)%*%solve(fi*Vtheta+D)%*%t_theta(x)
}
# * * * * * * * ** Sigma0pp * * * * * * ** * * * * * * * * * * #
Sigma0pp<-function(x1,x2)
{
fi*V(x1,x2)-(fi^2)%*%t(t_theta(x1))%*%solve(fi*Vtheta+D)%*%t_theta(x2)
```

```
+t(s_fi_theta(x1))%*%solve(B1)%*%s_fi_theta(x2)
}
# * * * * * * * * * * Mu0pp * * * * * * * * * * * * * * * * * *#
Mu0pp<-function(x)
{
t(h(x))%*%b1+fi*t(t_theta(x))%*%solve(fi*Vtheta+D)%*%(y-H%*%b1)
}
# * * * * * * * * * f_hat * * * * * * * ** * * ** * * * #
f_hat<-function(x)
{
Mu0pp(x)
}
# * * * * * * ** ** sigma2_hat * * * * * * * * * * * * * *#
sigma2_hat<-function(x)
{
nu_hat%*%(1+Sigma0pp(x,x))
}
# * * * * * * * * * * * * * * * * * * * #
#1. Construir la matriz de datos
y<-array(dim=n);
X<-matrix(nrow=n,ncol=r);
for(i in 1:n){
     y[i]<-datos[i,1]
     for(j in 1:r)
         X[i,j]<-datos[i,j+1]; # Matriz de diseño
     }
y<-D%*%(y) # Variable dependiente
W<-matrix(0,nrow=r,ncol=r); #Matriz W de Gutiérrez-Peña y Smith
for(j in 1:r)
     W[j,j]<-(z(j)^2)

Vtheta<-matrix(,nrow=n,ncol=n)
for(i in 1:n) #Matriz Vtheta
     for(j in 1:n)
             Vtheta[i,j]<-V(X[i,],X[j,]);
```

```
H<-matrix(nrow=n,ncol=q)
for(i in 1:n)
     H[i,]<-h(X[i,])
b1<-solve(t(H)%*%solve(fi*Vtheta+D)%*%H)%*%(t(H)%*%solve(fi*Vtheta+D)%*%y)
B1<-t(H)%*%solve(fi*Vtheta+D)%*%H
nu_hat<-t(y)%*%solve(fi*Vtheta+D)%*%y-t(b1)%*%B1%*%b1 #

# Los Submodelos
 hi<-function(x,Mi)
 {
     tv<-array(dim=q)
     tv<-c(1,x)
     Mi%*%tv
 }

 ti2_hat<-function(x,Mi,Hi)
 {
     (1+t(hi(x,Mi))%*%solve(t(Hi)%*%Hi)%*%hi(x,Mi))
 }
 fi_hat<-function(x,Mi,b1i)
 {
     t(hi(x,Mi))%*%b1i
 }
 sigmai2_hat<-function(x,Mi,Hi,nui_hat)
 {
     (nui_hat*ti2_hat(x,Mi,Hi))
 }
integrand<-function(xe,P)
{
log(dt((xe-P[2])/sqrt(P[3]),P[1])/sqrt(P[3]))*(dt((xe-P[5])/sqrt(P[6]),P[4])/sqrt(P[6]))
}

parametros<-function(gama,P,x)
{
```

```
 qi=sum(gama)
 quitar<-array(0);
 e<-0;
     for(i in 1:q)
         {
         if(gama[i]==0){
         e<-e+1
         quitar[e]<-i;
         }
      }
     Mi<-diag(1,ncol=q,nrow=q)
 if(e>0)
 Mi<-Mi[-quitar,]
 Hi<-matrix(nrow=n,ncol=qi)

 for(i in 1:n)
     Hi[i,]<-t(hi(X[i,],Mi))
 b1i<-solve(t(Hi)%*%Hi)%*%(t(Hi)%*%y)

 #print(" Estimaciones (B):") #<-imprime las estimaciones
 #print(b1i)
 nui_hat<-t(y)%*%y-t(b1i)%*%(t(Hi)%*%Hi)%*%b1i #

 #Parámetros de pi
 P[1]<-(n-qi) #alfai_hat
 P[2]<-fi_hat(x,Mi,b1i) #fi_hat_x
 P[3]<-sigmai2_hat(x,Mi,Hi,nui_hat)/P[1] #sigmai2_hat_x / alfai_hat
 # Parámetros de la p_hat
 P[4]<-(n-q) #alfa_hat_x
 P[5]<-f_hat(x) #f_hat_x
 P[6]<-sigma2_hat(x)/P[4]# sigma2_hat_x /alfa_hat_x
 P
}
# Cálculo de la utilidad esperada
UE<-function(gama)
```

```
{
 esp=0;
 for(i in 1:n)
 {
 x<-X[i,]
 P<-parametros(gama,P,x)
 esp=esp+integrate(integrand,-Inf,Inf,rel.tol = .Machine$double.eps^0.35,P=P)$value

 }
 esp/n
}
#Selección de modelo inicial, en este caso el modelo saturado.
xo<-array(1,dim=q)
#Se invoca el algoritmo del templado simulado con la calibración deseada
#SimulatedAnnealing(xo,To,n,cool,recalentado)
SimulatedAnnealing(xo,20,30,1,FALSE)
```

Apéndice B. El problema de reportar inferencias como problema de decisión

El problema de reportar inferencias sobre una clase $\mathcal{E} = \{E_j : j \in J\}$ de eventos excluyentes y exhaustivos condicional a un conjunto de datos $\mathcal{D}$, puede plantearse como un problema de decisión en la forma siguiente. Consideremos a un decisor, que asigna la cantidad $p_j > 0$ a la probabilidad de que el evento $E_j \in \mathcal{E}$ sea cierto, condicional a un conjunto de datos observados. La distribución de probabilidad $\{p_j : j \in J\}$ representa entonces, el conocimiento que el decisor tiene sobre $\mathcal{E}$. Considerando al espacio de acciones $\mathcal{Q}$

$$\mathcal{Q} = \left\{ \boldsymbol{q} = (q_j,\ j \in J) : q_j \geq 0 \text{ y } \sum_{j \in J} q_j = 1 \right\}$$

como la clase de distribuciones de probabilidad sobre $\mathcal{E}$ condicionales a $\mathcal{D}$ (i.e. compatibles con la información contenida en $\mathcal{D}$), y al conjunto de consecuencias $\mathcal{C}$ como el conjunto que consta de todos los pares $(\boldsymbol{q}, E_j)$, que representan el hecho de reportar $\boldsymbol{q}$ y E_j como el evento cierto. Nos falta por definir en el problema de decisión, una función de utilidad $u(\cdot, \cdot)$ que mida el *valor* de la consecuencia $(\boldsymbol{q}, E_j)$. Esto nos lleva a la siguiente definición.

Definición 11 *Una función de puntaje u para la distribución de probabilidad* $\boldsymbol{q} = (q_j,\ j \in J)$ *definida sobre una partición* $\{E_j : j \in J\}$ *es un mapeo que asigna un número real* $u(\boldsymbol{q}, E_j)$ *a cada pareja* $(\boldsymbol{q}, E_j)$. *Se dice que esta función es suave si es continuamente diferenciable como función de cada* q_j.

Finalmente para completar el problema, elegimos a la utilidad como una función de puntaje. De acuerdo con el desarrollo de la sección 2.6 la solución al problema, es elegir aquella distribución $\boldsymbol{q} \in \mathcal{Q}$ que maximice la utilidad esperada

$$\sum_{j=1}^{m} u(\boldsymbol{q}, E_j)\, p_j.$$

Para garantizar no solamente que el decisor sea coherente sino también *honesto*, necesitamos una forma de $u(\cdot, \cdot)$ que garantice que la utilidad esperada se maximice si, y sólo si, $q_j = p_j$ para $j \in J$. Pues,

de otro modo el tomador de decisiones podría reportar cualquier otra cosa que no fueran sus creencias verdaderas. Esto motiva la siguiente definición:

Definición 12 *Una función de puntaje u es propia si, para cada distribución de probabilidad estrictamente positiva $\boldsymbol{p} = (p_j,\ j \in J)$ definida sobre una partición $\{E_j : j \in J\}$*

$$\sup_{\boldsymbol{q} \in \mathcal{Q}} \sum_{j \in J} u\left(\boldsymbol{q}, E_j\right) p_j = \sum_{j \in J} u\left(\boldsymbol{p}, E_j\right) p_j$$

donde el supremo se alcanza si, y sólo si, $\boldsymbol{q} = \boldsymbol{p}$.

Por otra parte, la calificación asignada al individuo por su "predicción" debe basarse principalmente en su juicio acerca del evento E_j que resultó ser verdadero, i.e. en q_j. Por lo que, adicionalmente, pediremos a la función de puntaje sea local.

Definición 13 *Una función de puntaje u es local si, para cada elemento $\boldsymbol{q} = (q_j,\ j \in J)$ de la clase $\mathcal{Q}$ definida sobre una partición $\{E_j : j \in J\}$, existen funciones $\{u_j(\cdot) : j \in J\}$ tales que $u\left(\boldsymbol{q}, E_j\right) = u_j\left(q_j\right)$.*

El siguiente teorema caracteriza a las funciones de puntaje suaves, locales y propias. Se da la demostración para el caso en que $J = \{1, 2, ..., m\}$.

Teorema 14 *Si $u_j(\cdot, \cdot)$ es una función de puntaje suave, local y propia para distribuciones de probabilidad $\boldsymbol{q} = (q_j,\ j = 1, 2, ..., m)$ definidas sobre $\mathcal{E} = \{E_j : j = 1, 2, ..., m\}$, entonces debe ser de la forma*

$$u\left(\boldsymbol{q}, E_j\right) = A \log q_j + B_j$$

donde $A \in \mathbb{R}_+$ y $B_j \in \mathbb{R}$ son constates arbitrarias.

Demostración. Dado que $u_j(\cdot, \cdot)$ es local y propia, se cumple que

$$\sup_{\boldsymbol{q} \in \mathcal{Q}} \sum_{j=1}^{m} u\left(\boldsymbol{q}, E_j\right) p_j = \sup_{\boldsymbol{q} \in \mathcal{Q}} \sum_{j=1}^{m} u_j\left(q_j\right) p_j = \sum_{j=1}^{m} u_j\left(q_j\right) p_j$$

para algunas funciones $\{u_j(\cdot) : j = 1, 2, ..., m\}$, donde $p_j > 0$ para $j \in J$ y $\sum_{j=1}^{m} p_j = 1$. Tomando en cuenta la restricciones $\sum_{j=1}^{m} p_j = 1$ y $\sum_{j=1}^{m} q_j = 1$, definimos

$$F\left(q_2, ..., q_m\right) = u_1\left(1 - \sum_{j=2}^{m} q_j\right)\left(1 - \sum_{j=2}^{m} p_j\right) + \sum_{j=2}^{m} u_j\left(q_j\right) p_j$$

para maximizar F con respecto a $(q_2, ..., q_m)$ calculamos

$$\frac{d}{dq_j}F = -u_1'\left(1 - \sum_{j=2}^{m} q_j\right)\left(1 - \sum_{j=2}^{m} p_j\right) + u_j'(q_j)\, p_j \qquad \text{para} \qquad j = 2, ..., m$$

Igualando estas derivadas a cero, obtenemos el siguiente sistema de ecuaciones

$$u_j'(q_j)\, p_j = u_1'\left(1 - \sum_{j=2}^{m} q_j\right)\left(1 - \sum_{j=2}^{m} p_j\right), \qquad \text{para} \qquad j = 2, ..., m$$

pero por hipótesis u es propia, por lo que se maximiza en $\boldsymbol{q} = \boldsymbol{p}$. Por lo tanto, debe tenerse que

$$u_j'(p_j)\, p_j = u_1'(p_1)\, p_1, \quad \text{para} \qquad j = 2, ..., m$$

Así, las funciones $u_2(\cdot), ..., u_m(\cdot)$ satisfacen la relación

$$u_1'(p_1)\, p_1 = u_2'(p_2)\, p_2 = \cdots = u_m'(p_m)\, p_m.$$

En otras palabras, todas las funciones $u_1(\cdot), ..., u_m(\cdot)$ satisfacen la ecuación funcional

$$u_j'(p)\, p = A \qquad 0 < p < 1$$

donde $A \in \mathbb{R}$ es una constante.

Por lo tanto,

$$u_j(p) = A \log p + B_j \quad \text{con} \qquad B_j \in \mathbb{R}$$

La condición $A > 0$ garantiza que el óptimo $\boldsymbol{q} = \boldsymbol{p}$ sea en efecto un máximo. ■

Referencias

[1] Aarts, E. & Korst, J. (1989). *Simulated Annealing and Boltzman Machines*. New York:Wiley.

[2] Berger, J.O. & Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of American Statistical Association* **91**, 109-122.

[3] Bernardo, J.M. (1981). *Bioestadística una perspectiva bayesiana*. Barcelona:Vicens-Vives.

[4] Bernardo, J.M. (1992). Simulated annealing in Bayesian decision theory. *Computational Statistics.* (Y. Dodge and J. Whittaker, eds). Heidelberg: Physica-Verlag, 547-552.

[5] Bernardo, J.M. (1979). Expected information as expected utility. *Annals of Statistics* **7**, 686-690.

[6] Bernardo, J.M. & Smith, A.F.M. (1994). *Bayesian Theory*. Chichester:Wiley.

[7] Brooks, S.P. & Morgan, B.J. (1995). Optimization using simulated annealing. *The Statistician* **44**, 241-257.

[8] Chipman, H., George, E. & McCulloch R. (2001). *The practical implementation of Bayesian model selection*. IMS Lecture Notes-Monograph Series. Vol 38.

[9] Draper, N. & Smith, H. (1998). *Applied Regression Analysis*. New York: Wiley.

[10] Dréo, J., Pétrowski, A., Siarry, P. & Taillar, E. (2006). *Metaheuristics for Hard Optimization*. New York:Springer

[11] Fleischer, M. (1995). Simulated Annealing: past, present, and future. *Proceedings of the 1995 Winter Simulation Conference.* (Alexopoulos, C. Kang, K. Lilegdon, R. and Goldsman, D. Eds). 155-161.

[12] Fouskakis, D. & Draper, D. (2002). Stochastic Optimization: a Review. *International Statistical Review* **70**, 315-349.

[13] George, E.I. & McCulloch, R.E (1993). Variable selection via Gibbs sampling. *Journal of American Statistical Association* **88**, 881-889.

[14] Geweke, J. (1996). Variable selection and model comparison in regression. In *Bayesian Statistics* 5 (J.M. Bernardo, J.O. Berger, A.P. David and A.F.M. Smith, eds.). Oxford University Press, pp. 609-620.

[15] Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. *Computers and Operations Research* **13**, 533-549.

[16] Gutiérrez-Peña, E. (1997). A Bayesian predictive semiparametric approach to variable selection and model comparison in regression. *Bulletin of the International Statistical Institute*, Tome LVII. (Proceedings of the 51st Session of the ISI, Invited Papers, Book 1.) Istanbul, Turkey, pp. 17-29.

[17] Gutiérrez-Peña, E. & Smith, A.F.M. (1998). Aspects of smoothing and model inadequacy in generalised regression. *Journal of Statistics Planning and Inference* **67**, 273-286.

[18] Hillier F. & Lieberman G. (1997). *Introducción a la Investigación de Operaciones*. México: McGraw-Hill.

[19] Key, J.T., Pericchi, L., & Smith A.F.M. (1999). Bayesian model choice: what and why?. *In Bayesian Statistics 6* (J. M. Bernardo,J. O.Berger, A. P. Dawid & A. F. M. Smith, eds.). Oxford University Press, pp. 343-370.

[20] Kirkpatrick, S., Gelatt, D. & Vecchi, M. (1983). Optimization by simulated annealing. *Science* **220**, 671-680.

[21] Kuo, L. & Mallick, B. (1994). Variable selection for regression models. *Technical Report,* **TR9426**. Departament of Statistics, University of Connecticut.

[22] Laud, P. & Ibrahim, J. (1995). Predictive model selection. *Journal of the Royal Statistical Society B* **57**, 247-262.

[23] O'Hagan, A. (1978). Curve fitting and optimal design for prediction. *Journal of the Royal Statistics Society B* **40,** 1-42.

[24] O'Hagan, A. (1992). Some Bayesian numerical analysis. *In Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith, eds.). Oxford University Press, 345-363.

[25] Robert, C. & Casella, G. (2004). *Monte Carlo Statistical Methods.* New York:Springer.

[26] San Martini, A. & Spezzaferri, F. (1984) A predictive models selection criterion. *Journal of the Royal Statistical Society B* **46**, 296-303.

[27] Schnaas, L., Rothenberg, S., Flores, M., Martínez, S., Hernández, C., Osorio, E., Ruiz, S. & Perroni, E. (2006). Reduced intellectual development in children prenatal lead exposure. *Environmental Health Perspectives* **114**, 791-797.

[28] Silva, L.C. (2000). La alternativa bayesiana. *Brotes* **1**, 1-4.

[29] Silva, L.C. & Benavides A. (2001). El enfoque bayesiano: otra forma de inferir. *Gaceta Sanitaria* **4**, 341-346.

[30] Wald, A. (1950). *Statistical Decision Functions.* New York*:*Wiley.

Bibliografía

[1] Blum, C. & Roli, A. (2003). Metaheuristics in combinatorial optimization: overview and conceptual comparison. *ACM Computing Surveys* **35**, 268-308.

[2] Bondy, J. & Murty, U. (1976). *Graph Theory with Applications*. New York: North-Holland.

[3] Brooks, S.P., Friel, N. & King R. (2003). Classical model selection via simulated annealing. *Journal of the Royal Statistical Society B* **65**, 503-520.

[4] Brown, P., Fearn, T., & Vannucci, M. (1999). The choice of multivariate regression: A non-conjugate Bayesian decision theory approach. *Biometrika* **86**, 635-648.

[5] Carson, Y. (1997). Simulation optimization: methods and applications. *Proceedings of the 1997 Winter Simulation Conference.* (Andradóttir, S., Healy, K., Withers, D. and Nelson, B. Eds). 118-126.

[6] Dimitris, B. & Tsitsiklis, J. (1993). Simulated annealing. *Statistical Science* **8**, 10-15.

[7] Fu, M. (2001). Simulation Optimization. *Proceedings of the 1995 Winter Simulation Conference.* (Peters, B., Smith, S., Medeiros, D., and Rohrer, M. Eds). 53-61.

[8] George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association* **95**, 1304-1308.

[9] Glover, F. & Laguna, M. (1997). *TABU Search*. Boston: Kluwer.

[10] Hillier F. & Lieberman G. (1997). *Introducción a la Investigación de Operaciones*. México: McGraw-Hill.

[11] Lindley, D.V. (1968). The choice of variables in multiple regression. *Journal of the Royal Statistical Society B* **30**, 31-66.

[12] Lundy, M. & Mees, A. (1986). Convergence of an annealing algorithm. *Mathematical Programming* **34**, 111-124.

[13] Marriott, J. M., Spencer, N. M. & Pettitt, A. N. (2001). A Bayesian approach to selecting covariates for prediction. *Scandinavian Journal of Statistics* **28**, 87-97.

[14] Nelder, J.A. & Mead, R. (1965). A simplex method for function minimization. *Computer Journal* **7**, 308-313.

[15] Pratt., J.W., Raiffa, H. & Schlaifer, R. (1964). The foundations of decision under uncertainty: an elementary exposition. *Journal of American Statistical Association* **59**, 353-375.

[16] Tan, S. (2001). Introduction to Bayesian Methods for Medical Research. *Annals of the Academy of Medicine* **30**, 444-446.