14

Algoritmos EM

14.1 Introducción

En el capítulo 8, hemos hablado de métodos para maximizar la función log-verosimilitud (LL). A medida que los modelos se vuelven más complejos, la maximización a través de estos métodos se hace más difícil. Varias cuestiones contribuyen a esta dificultad. En primer lugar, la obtención de modelos más flexibles y realistas suele lograrse aumentando el número de parámetros. Sin embargo, los procedimientos descritos en el capítulo 8 utilizan el cálculo del gradiente respecto a cada parámetro, por lo que a medida que el número de parámetros se eleva, requieren más tiempo de cálculo. El hessiano, o el hessiano aproximado, deben ser calculados e invertidos; con un gran número de parámetros, la inversión puede ser numéricamente difícil. Asimismo, a medida que el número de parámetros crece, la búsqueda de los valores de maximización debe realizarse sobre un espacio de mayor dimensionalidad, de tal manera que la localización del máximo necesita más iteraciones. En definitiva, cada iteración usa más tiempo y se requieren más iteraciones.

En segundo lugar, la función LL para los modelos simples a menudo es aproximadamente cuadrática, de manera que los procedimientos del capítulo 8 operan de manera efectiva. Sin embargo, cuando el modelo se vuelve más complejo, la función LL generalmente se vuelve menos cuadrática, al menos en algunas regiones del espacio de parámetros. Este problema puede manifestarse de dos formas. El procedimiento iterativo puede quedarse "atrapado" en las áreas no cuadráticas de la función LL, dando pasos pequeños que no representan apenas mejoría en la LL. O el procedimiento puede "pasar de largo" el máximo repetidas veces, dando grandes pasos en cada iteración, pero sin poder localizar el máximo.

Por último, existe otro problema más fundamental que el número de parámetros y la forma de la función LL. Por lo general, cuando un investigador especifica un modelo más general, y por lo tanto complejo, lo hace porque quiere depender menos de los supuestos y, por el contrario, obtener más información de los datos. Sin embargo, el objetivo de obtener más información de los datos es intrínsecamente contrario a la simplicidad de la estimación.

Los algoritmos de maximización de la esperanza (*expectation-maximization, EM*) son procedimientos que permiten la maximización de una función LL cuando los procedimientos estándar son numéricamente difíciles o inviables. El procedimiento fue introducido por Dempster, Laird y Rubin (1977), como un mecanismo para manejar información perdida o ausente. Sin embargo, es aplicable de forma mucho más general y se ha utilizado con éxito en muchos campos de la estadística. McLachlan y

Krishnan (1997) ofrecen una revisión de casos prácticos. En el campo de los modelos de elección discreta, los algoritmos EM han sido utilizados por Bhat (1997a) y Train (2008a, b).

El procedimiento consiste en definir una esperanza o expectativa en particular, y luego maximizarla (de ahí el nombre). Esta expectativa está relacionada con la función LL de una forma que vamos a describir, pero difiere de tal manera que facilita la maximización. El procedimiento es iterativo, iniciándose en un cierto valor inicial de los parámetros y actualizando los valores en cada iteración. Los parámetros actualizados en cada iteración son los valores que maximizan la expectativa en esa iteración particular. Como se verá, la maximización repetida de esta función converge al máximo de la propia función LL.

En este capítulo se describe el algoritmo EM en general y se desarrollan algoritmos específicos para modelos de elección discreta con coeficientes aleatorios. Mostraremos que el algoritmo EM se puede utilizar para estimar distribuciones de preferencias muy flexibles, incluidas especificaciones no paramétricas que pueden aproximar asintóticamente cualquier distribución verdadera subyacente. Aplicaremos estos métodos en un caso de estudio relativo a la elección que los consumidores hacen entre vehículos de hidrógeno y vehículos de gas.

14.2 Procedimiento general

En esta sección describiremos el procedimiento EM de una manera muy general, a fin de elucidar sus características. En las secciones siguientes, aplicaremos el procedimiento general a modelos específicos. Denotemos colectivamente las variables dependientes observadas como y, representando las elecciones o la secuencia de elecciones de una muestra completa de decisores. Las elecciones dependen de variables explicativas observadas que, por conveniencia de notación, no indicamos de forma explícita. Las elecciones también dependen de datos que faltan (o datos ausentes, *missing data*), designados colectivamente como z. Dado que los valores de estos datos ausentes no son observados, el investigador especifica una distribución que representa los valores que estos datos ausentes podrían tomar. Por ejemplo, si no tenemos el nivel de ingresos de algunos individuos de la muestra, la distribución de los ingresos en la población puede ser una especificación útil para la distribución de los valores de los ingresos que no tenemos. La densidad de los datos ausentes se denota como $f(z|\theta)$, que en general depende de parámetros θ a estimar.

El modelo de comportamiento relaciona datos observados y ausentes con elecciones de decisores. Este modelo predice las elecciones que se producirían si los datos ausentes fueran realmente observados en lugar de estar ausentes. Este modelo de comportamiento se denota como $P(y|z,\theta)$, donde θ son parámetros que pueden solaparse o ampliar los de f. (Para mantener la notación compacta, utilizaremos θ para referirnos a todos los parámetros a estimar, incluyendo los que entran en f y los que entran en P). Sin embargo, dado que realmente los datos ausentes no están, la probabilidad de las elecciones observadas, utilizando la información que el investigador observa, es la integral de la probabilidad condicionada sobre la densidad de los datos ausentes[ix]:

$$P(y|\theta) = \int P(y|z,\theta) f(z|\theta) dz.$$

La densidad de los datos ausentes, $f(z|\theta)$, se utiliza para predecir las elecciones observadas, y por lo tanto no depende de y. Sin embargo, podemos obtener alguna información acerca de los datos ausentes mediante la observación de las elecciones que se hicieron. Por ejemplo, en la elección del vehículo, si los ingresos de una persona no están disponibles pero se observa que la persona ha

[ix] Asumimos en esta expresión que z es continua, de manera que la probabilidad no condicionada es una integral. Si z es discreta, o una mezcla de variables continuas y discretas, entonces la integración se sustituye por una suma sobre los valores discretos, o una combinación de integrales y sumas.

comprado un Mercedes, se puede inferir que es probable que los ingresos de esta persona estén por encima de la media. Definamos $h(z|y,\theta)$ como la densidad de los datos ausentes condicionada a las elecciones observadas en la muestra. Esta densidad condicionada está relacionada con la densidad no condicionada través de la identidad de Bayes:

$$h(z|y,\theta) = \frac{P(y|z,\theta)f(z|\theta)}{P(y|\theta)}.$$

Dicho sucintamente, la densidad de z condicionada a las elecciones observadas es proporcional al producto de la densidad no condicionada de z por la probabilidad de las elecciones observadas dada esta z. El denominador es simplemente la constante de normalización, igual a la integral del numerador. Este concepto de distribución condicionada debería resultar familiar a los lectores del capítulo 11.

Ahora consideremos la estimación. La función LL se basa en la información que el investigador tiene, que no incluye los datos ausentes. La función LL es por tanto

$$LL(\theta) = log\, P(y|\theta) = log\left(\int P(y|z,\theta)f(z|\theta)dz\right).$$

En principio, esta función se puede maximizar mediante el uso de los procedimientos descritos en el capítulo 8. Sin embargo, como vamos a ver, a menudo es mucho más fácil maximizar LL de forma diferente.

El procedimiento alternativo es iterativo, comenzando con un valor inicial de los parámetros y actualizándolos de una manera que se describirá a continuación. Denotemos el valor de prueba de los parámetros en una iteración dada como θ^t. Definamos una nueva función en θ^t que se relacione con LL pero que utilice la distribución condicionada h. Esta nueva función es

$$\mathcal{E}(\theta|\theta^t) = \int h(z|y,\theta^t)\, log(P(y|z,\theta)f(z|\theta))\, dz,$$

donde la densidad condicionada h se calcula utilizando el valor de prueba de los parámetros actual, θ^t. Esta función tiene un significado específico. Tenga en cuenta que la parte más a la derecha de esta expresión, $P(y|z,\theta)f(z|\theta)$, es la probabilidad conjunta de las elecciones observadas y de los datos ausentes. El logaritmo de esta probabilidad conjunta es la LL de las elecciones observadas *y* de los datos ausentes combinados. Esta LL conjunta está integrada sobre una densidad, concretamente, $h(z|y,\theta^t)$. Por tanto, nuestra función $\mathcal{E}$ es una esperanza de la LL conjunta de los datos ausentes y de las elecciones observadas. Es una esperanza específica, en concreto, la esperanza sobre la densidad de los datos ausentes condicionada a las elecciones observadas. Puesto que la densidad condicionada de z depende de los parámetros, esta densidad se calcula utilizando los valores θ^t. Dicho de manera equivalente, $\mathcal{E}$ es el promedio ponderado de la LL conjunta, utilizando $h(z|y,\theta^t)$ como pesos.

El procedimiento EM consiste en maximizar $\mathcal{E}$ repetidamente. Empezando con un cierto valor inicial, los parámetros se actualizan en cada iteración a través de la siguiente fórmula:

$$\theta^{t+1} = argmax_\theta \mathcal{E}(\theta|\theta^t). \qquad (14.1)$$

En cada iteración, los valores actuales de los parámetros, θ^t, se utilizan para calcular los pesos h, y a continuación se maximiza la LL conjunta ponderada. El nombre EM proviene del hecho de que el procedimiento utiliza una esperanza que es maximizada.

Es importante reconocer la doble función de los parámetros en $\mathcal{E}$. En primer lugar, los parámetros entran en la función log-verosimilitud conjunta de las elecciones observadas y de los datos ausentes, $log(P(y|z,\theta)f(z|\theta))$. En segundo lugar, los parámetros entran en la densidad condicionada de los datos ausentes, $h(z|y,\theta)$. La función $\mathcal{E}$ se maximiza respecto a la primera manteniendo constante la segunda. Es decir, $\mathcal{E}$ se maximiza sobre la θ que entra en $log(P(y|z,\theta)f(z|\theta))$, manteniendo el valor de θ que entra en los pesos $h(z|y,\theta)$ en su valor actual θ^t. Para indicar este doble rol, $\mathcal{E}(\theta|\theta^t)$ se expresa como una función de θ (el argumento sobre el que se realiza la maximización) dado θ^t(el valor utilizado en los pesos que se mantiene fijo durante la maximización).

En condiciones muy generales, las iteraciones definidas por la ecuación (14.1) convergen al máximo de LL. Bolyes (1983) y Wu (1983) proporcionan pruebas formales. En la siguiente sección ofrezco una explicación intuitiva. Sin embargo, los lectores que estén interesados en ver en primer lugar ejemplos del algoritmo pueden consultar directamente la sección 14.3.

14.2.1 ¿Por qué el algoritmo EM funciona?

La relación entre el algoritmo EM y la función LL se puede explicar en tres pasos. Cada paso es un poco opaco, pero los tres combinados proporcionan una comprensión sorprendentemente intuitiva.

Paso 1: Ajustamos $\mathcal{E}$ para igualarla a LL en θ^t

$\mathcal{E}(\theta|\theta^t)$ no es igual a $\mathrm{LL}(\theta)$. Para facilitar la comparación entre ambas funciones, vamos a añadir una constante a $\mathcal{E}(\theta|\theta^t)$ igual a la diferencia entre las dos funciones en θ^t:

$$\mathcal{E}^*(\theta|\theta^t) = \mathcal{E}(\theta|\theta^t) + [LL(\theta^t) - \mathcal{E}(\theta^t|\theta^t)]$$

El término entre corchetes es constante respecto a θ, así que maximizar $\mathcal{E}^*$ es equivalente a maximizar la propia $\mathcal{E}$. Observe sin embargo que, por construcción, $\mathcal{E}^*(\theta|\theta^t) = LL(\theta)$ en $\theta = \theta^t$.

Paso 2: Observe que la derivada respecto a θ es la misma para $\mathcal{E}^*$ y para LL evaluada en $\theta = \theta^t$.

Considere la derivada de $\mathcal{E}^*(\theta|\theta^t)$ respecto a su argumento θ:

$$\frac{d\mathcal{E}^*(\theta|\theta^t)}{d\theta} = \frac{d\mathcal{E}(\theta|\theta^t)}{d\theta}$$

$$= \int h(z|y,\theta^t)\left(\frac{d\, log\, P(y|z,\theta)f(z|\theta)}{d\theta}\right)dz$$

$$= \int h(z|y,\theta^t)\frac{1}{P(y|z,\theta)f(z|\theta)}\frac{dP(y|z,\theta)f(z|\theta)}{d\theta}dz.$$

Ahora evaluemos esta derivada en $\theta = \theta^t$:

$$\left.\frac{d\mathcal{E}^*(\theta|\theta^t)}{d\theta}\right|_{\theta^t}$$

$$= \int h(z|y,\theta^t)\frac{1}{P(y|z,\theta^t)f(z|\theta^t)}\left(\frac{dP(y|z,\theta)f(z|\theta)}{d\theta}\right)_{\theta^t}dz$$

$$= \int \frac{P(y|z,\theta^t)f(z|\theta^t)}{P(y|\theta^t)} \frac{1}{P(y|z,\theta^t)f(z|\theta^t)} \left(\frac{dP(y|z,\theta)f(z|\theta)}{d\theta}\right)_{\theta^t} dz$$

$$= \int \frac{1}{P(y|\theta^t)} \left(\frac{dP(y|z,\theta)f(z|\theta)}{d\theta}\right)_{\theta^t} dz$$

$$= \frac{1}{P(y|\theta^t)} \int \left(\frac{dP(y|z,\theta)f(z|\theta)}{d\theta}\right)_{\theta^t} dz$$

$$= \left(\frac{d\, log\, P(y|\theta)}{d\theta}\right)_{\theta^t}$$

$$= \left(\frac{dLL(\theta)}{d\theta}\right)_{\theta^t}.$$

En $\theta = \theta^t$, las dos funciones, $\mathcal{E}^*$ y LL, tienen la misma pendiente.

Paso 3: Observe que $\mathcal{E}^* \leq LL$ para todo θ.

Esta relación se puede demostrar de la siguiente manera:

(14.2) $$LL(\theta) = logP(y|\theta)$$

$$= log \int P(y|z,\theta)f(z|\theta)dz$$

$$= log \int \frac{P(y|z,\theta)f(z|\theta)}{h(z|y,\theta^t)} h(z|y,\theta^t)dz$$

(14.3) $$\geq \int h(z|y,\theta^t)\, log \frac{P(y|z,\theta)f(z|\theta)}{h(z|y,\theta^t)} dz$$

$$= \int h(z|y,\theta^t)\, log\big(P(y|z,\theta)f(z|\theta)\big)\, dz - \int h(z|y,\theta^t)\, log(h(z|y,\theta^t))\, dz$$

$$= \mathcal{E}(\theta|\theta^t) - \int h(z|y,\theta^t)\, log(h(z|y,\theta^t))\, dz$$

$$= \mathcal{E}(\theta|\theta^t) - \int h(z|y,\theta^t)\, log\left(h(z|y,\theta^t)\frac{P(y|\theta^t)}{P(y|\theta^t)}\right) dz$$

$$= \mathcal{E}(\theta|\theta^t) + \int h(z|y,\theta^t)\, log(P(y|\theta^t))\, dz - \int h(z|y,\theta^t)\, log(h(z|y,\theta^t)P(y|\theta^t))\, dz$$

$$= \mathcal{E}(\theta|\theta^t) + log(P(y|\theta^t)) \int h(z|y,\theta^t)dz - \int h(z|y,\theta^t)\, log(h(z|y,\theta^t)P(y|\theta^t))\, dz$$

$$= \mathcal{E}(\theta|\theta^t) + log(P(y|\theta^t)) - \int h(z|y,\theta^t)\, log(h(z|y,\theta^t)P(y|\theta^t))\, dz \quad (14.4)$$

$$= \mathcal{E}(\theta|\theta^t) + LL(\theta^t) - \int h(z|y,\theta^t)\, log\big(P(y|z,\theta^t)f(z|\theta^t)\big)\, dz \quad (14.5)$$

$$= \mathcal{E}(\theta|\theta^t) + LL(\theta^t) - \mathcal{E}(\theta^t|\theta^t)$$

$$= \mathcal{E}^*(\theta|\theta^t).$$

La desigualdad mostrada en la ecuación (14.3) se debe a la desigualdad de Jensen, que establece que $log(E(x)) > E(log(x))$. En nuestro caso, x es el estadístico $\frac{P(y|z,\theta)f(z|\theta)}{h(z|y,\theta^t)}$ y la esperanza es respecto a la densidad $h(z|y,\theta^t)$. Un ejemplo de esta desigualdad se muestra en la figura 14.1, donde los promedios son respecto a dos valores etiquetados como a y b. El promedio de $log(a)$ y $log(b)$ es el punto medio de la línea discontinua que conecta estos dos puntos de la curva logarítmica. El logaritmo evaluado en el promedio de a y b es $log((a+b)/2)$, que está por encima del punto medio de la línea discontinua. La desigualdad de Jensen es simplemente una consecuencia de la forma cóncava de la función logarítmica.

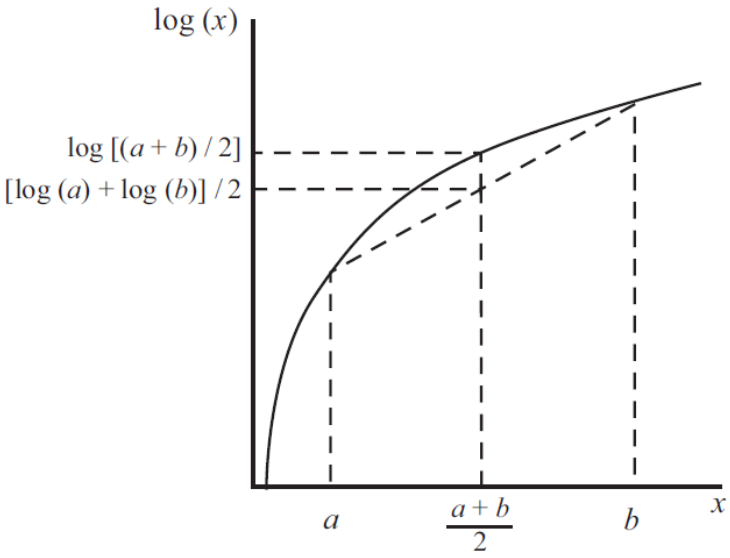

Figura 14.1. Ejemplo de la desigualdad de Jensen

La ecuación (14.4) se obtiene gracias a que la integral de la densidad h es 1. La ecuación (14.5) se obtiene sustituyendo $h(z|y,\theta^t) = P(y|z,\theta^t)f(z|\theta^t)/P(y|\theta^t)$ dentro del logaritmo y cancelando luego los términos $P(y|\theta^t)$.

Combinamos resultados para comparar $\mathcal{E}^*$ y LL.

La figura 14.2 muestra la relación entre $\mathcal{E}^*(\theta|\theta^t)$ y $LL(\theta)$. Como hemos demostrado, en $\theta = \theta^t$ estas dos funciones son iguales y tienen la misma pendiente. Estos resultados implican que las dos funciones son tangentes entre sí en $\theta = \theta^t$. También hemos demostrado que $\mathcal{E}^*(\theta|\theta^t) \leq LL(\theta)$ para todo θ. De acuerdo con esta relación, $\mathcal{E}^*$ se dibuja en el gráfico por debajo de $\mathrm{LL}(\theta)$ en todos los puntos, excepto en θ^t, donde son iguales.

El algoritmo EM maximiza $\mathcal{E}^*(\theta|\theta^t)$ en cada iteración para encontrar el siguiente valor de prueba de θ. El valor de maximización se muestra como θ^{t+1}. Como indica el gráfico, la función LL es necesariamente mayor en el valor del nuevo parámetro θ^{t+1}, que en el valor original θ^t. Mientras la derivada de la

función LL no sea cero en θ^t, maximizar $\mathcal{E}^*(\theta|\theta^t)$ incrementa $\mathrm{LL}(\theta)$[x]. Cada iteración del algoritmo EM incrementa la función LL hasta que el algoritmo converge en el máximo de la función LL.

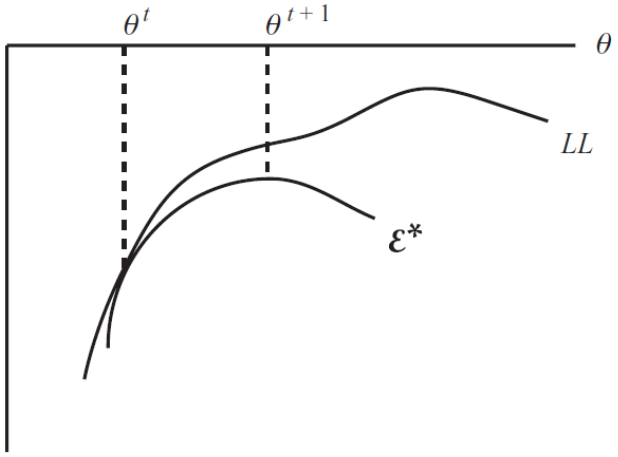

Figura 14.2. Relación entre $\mathcal{E}^*$ y LL.

14.2.2 Convergencia

La convergencia del algoritmo EM se define generalmente como un cambio suficientemente pequeño en los parámetros (por ejemplo, Levine y Casella, 2001) o en la función LL (por ejemplo, Weeks y Lange, 1989; Aitkin y Aitkin, 1996). Estos criterios deben ser utilizados con cuidado, ya que el algoritmo EM se puede mover lentamente cerca de la convergencia. Ruud (1991) muestra que el estadístico de convergencia visto en la sección 8.4 se puede utilizar con el gradiente y con el hessiano de $\mathcal{E}$, en lugar del LL. Sin embargo, el cálculo de este estadístico puede ser computacionalmente más exigente que la iteración del algoritmo EM en sí, llegando a ser inviable en algunos casos.

14.2.3 Errores Estándar

Hay 3 formas en que podemos calcular los errores estándar. En primer lugar, una vez se ha encontrado el máximo de $LL(\theta)$ con el algoritmo EM, los errores estándar se pueden calcular a partir de LL de la misma forma en que se haría si hubiésemos maximizado directamente la función LL. Los procedimientos de la sección 8.6 son aplicables: los errores estándar asintóticos se pueden calcular a partir del hessiano o a partir de la varianza de los gradientes específicos de cada observación (es decir, las puntuaciones), calculados a partir de LL evaluada en θ^t.

Una segunda opción surge del resultado que obtuvimos en el paso 2. Vimos que $\mathcal{E}$ y LL tienen el mismo gradiente en $\theta = \theta^t$. En el punto de convergencia, el valor de θ no cambia de una iteración a la siguiente, de tal manera que $\hat{\theta} = \theta^{t+1} = \theta^t$. Por lo tanto, en $\hat{\theta}$, las derivadas de estas dos funciones son iguales. Este hecho implica que las puntuaciones se pueden calcular a partir de $\mathcal{E}$ en lugar de LL. Si $\mathcal{E}$ toma una forma más conveniente que LL, como suele ser el caso cuando empleamos un algoritmo EM, esta forma de cálculo alternativo puede ser atractiva.

Una tercera opción es el *bootstrapping*, ya visto en la sección 8.6. En este caso, el algoritmo EM se aplica en numerosas ocasiones, usando una muestra diferente de las observaciones en cada ocasión. En muchos de los contextos en los que se aplican los algoritmos EM, el cálculo de los errores estándar a través de *bootstrapping* es más factible y útil que el uso de fórmulas asintóticas. El caso de estudio mostrado en la última sección proporciona un ejemplo.

[x] De hecho, cualquier incremento de $\mathcal{E}^*(\theta|\theta^t)$ conduce a un incremento de $LL(\theta)$

14.3 Ejemplos de algoritmos EM

Describiremos en esta sección varios tipos de modelos de elección discreta cuyas funciones LL son difíciles de maximizar directamente, pero son fáciles de estimar con algoritmos EM. El objetivo de la exposición es proporcionar ejemplos concretos de cómo se especifican los algoritmos EM e ilustrar al mismo tiempo el valor de este enfoque.

14.3.1 Distribución de mezcla discreta con puntos fijos

Una de las cuestiones que se plantean con los modelos logit mixtos (de hecho, con cualquier modelo mixto) es la especificación adecuada de la distribución de mezcla. Es habitual usar una distribución conveniente, como una normal o una log-normal. Sin embargo, es cuestionable que la verdadera distribución de los coeficientes tome una forma matemáticamente conveniente. Usar distribuciones más flexibles puede ser útil, donde flexibilidad significa que la distribución especificada puede tomar una variedad amplia de formas, dependiendo de los valores de sus parámetros.

Por lo general, podemos lograr mayor flexibilidad mediante la inclusión de más parámetros. En la estimación no paramétrica, se especifica una familia de distribuciones que tienen la propiedad de que la distribución se hace más flexible a medida que el número de parámetros se eleva. Al permitir que el número de parámetros aumente con el tamaño de la muestra, el estimador no paramétrico es consistente con cualquier distribución verdadera. El término "no paramétrico" es un nombre poco apropiado en este contexto: "superparamétrico" sería tal vez más apropiado, ya que el número de parámetros empleados es generalmente mayor al de especificaciones estándar, y aumenta con el tamaño de la muestra para obtener cada vez mayor flexibilidad.

El gran número de parámetros en la estimación no paramétrica hace que la maximización directa de la función LL resulte difícil. En muchos casos, sin embargo, es posible desarrollar un algoritmo EM que facilita esta estimación considerablemente. La presente sección muestra uno de estos casos.

Considere un modelo logit mixto con una distribución desconocida de coeficientes. Cualquier distribución se puede aproximar de forma arbitrariamente precisa a través de una distribución discreta con un número suficientemente grande de puntos. Podemos utilizar este hecho para desarrollar un estimador no paramétrico de la distribución de mezcla, utilizando un algoritmo EM para la estimación.

Representemos la densidad de los coeficientes mediante C puntos, siendo β_c el punto de c-ésimo. Supondremos que la ubicación de estos puntos (para este procedimiento en concreto) es fija y la masa en cada punto (es decir, la proporción o cuota de la población en cada punto) es el parámetro a estimar. Una forma de seleccionar los puntos es especificar un máximo y un mínimo de cada coeficiente, y crear una red de puntos uniformemente espaciados entre máximos y mínimos. Por ejemplo, supongamos que hay cinco coeficientes y que el rango entre el mínimo y el máximo de cada coeficiente está representado por 10 puntos uniformemente espaciados. Los 10 puntos en cada dimensión crean una cuadrícula de $10^5 = 100.000$ puntos en el espacio de cinco dimensiones. Los parámetros del modelo son la proporción o cuota de la población en cada uno de los 100.000 puntos. Como veremos, la estimación de un gran número de parámetros de este tipo es bastante asequible con un algoritmo EM. Al aumentar el número de puntos, la red se hace cada vez más fina, de tal manera que la estimación de las cuotas de población en los puntos permite aproximar cualquier distribución subyacente.

La utilidad que el agente n obtiene de la alternativa j es

$$U_{nj} = \beta_n x_{nj} + \varepsilon_{nj}$$

donde ε se distribuye valor extremo iid. Los coeficientes aleatorios tienen la distribución discreta que se ha descrito anteriormente, siendo s_c la cuota de la población en el punto β_c. La distribución se expresa a través de la siguiente función

$$f(\beta_n) = \begin{cases} s_1 & si\ \beta_n = \beta_1 \\ s_2 & si\ \beta_n = \beta_2 \\ \vdots & \\ s_C & si\ \beta_n = \beta_C \\ 0 & en\ otro\ caso, \end{cases}$$

donde las cuotas suman 1: $\sum_c s_c = 1$. Para mayor comodidad, nos referiremos al conjunto de todas las cuotas a través del vector $s = \langle s_1, \dots, s_c \rangle$[xi].

Condicionando a $\beta_n = \beta_c$ para unos valores c determinados, el modelo de elección es un logit estándar:

$$L_{ni}(\beta_c) = \frac{e^{\beta_c x_{ni}}}{\sum_j e^{\beta_c x_{ni}}}.$$

Dado que β_n no se conoce para cada persona, la probabilidad de elección es la de un modelo logit mixto, mezclando respecto a la distribución discreta de β_n:

$$P_{ni}(s) = \sum_c s_c L_{ni}(\beta_c).$$

La función LL es $LL(s) = \sum_n log\ P_{ni_n}(s)$, donde i_n es la alternativa elegida por el agente n.

Para estimar las cuotas s podemos maximizar directamente esta función LL. Con un gran número de clases, como se requiere normalmente para representar de forma flexible la distribución real, esta maximización directa puede ser difícil. Sin embargo, es posible utilizar un algoritmo EM increíblemente sencillo para este modelo, incluso con cientos de miles de puntos.

Los "datos ausentes" en este modelo son los coeficientes de cada agente. La distribución f indica la cuota de la población con cada valor del coeficiente. Sin embargo, como vimos en el capítulo 11, las elecciones que hace una persona revelan información sobre sus coeficientes. Condicionando a que la persona n elige la alternativa i_n, la probabilidad de que la persona tenga coeficientes β_c está dada por la identidad de Bayes:

$$h(\beta_c | i_n, s) = \frac{s_c L_{ni_n}(\beta_c)}{P_{ni_n}(s)}.$$

El algoritmo EM utiliza esta distribución condicionada. En particular, la esperanza para el algoritmo EM es

$$\mathcal{E}(s|s^t) = \sum_n \sum_c h(\beta_c | i_n, s^t)\ log\left(s_c L_{ni_n}(\beta_c)\right).$$

[xi] Esta especificación se puede considerar un tipo de modelo de clases latentes, donde hay C clases, los coeficientes de las personas de la clase c son β_c, y s_c es la proporción de la población en la clase c. Sin embargo, el término "modelo de clases latentes " por lo general se refiere a un modelo en el que la ubicación de los puntos son parámetros, así como las proporciones. Consideraremos esta forma más tradicional en nuestro siguiente ejemplo.

Dado que $log\left(s_c L_{ni_n}(\beta_c)\right) = log(s_c) + log\left(L_{ni_n}(\beta_c)\right)$, esta esperanza puede ser reescrita en dos partes:

$$\mathcal{E}(s|s^t) = \sum_n \sum_c h(\beta_c|i_n, s^t)\, log(s_c) + \sum_n \sum_c h(\beta_c|i_n, s^t)\, log\left(L_{ni_n}(\beta_c)\right).$$

Esta esperanza es la que debe maximizarse respecto a los parámetros s. Sin embargo, tenga en cuenta que el segundo término de esta expresión no depende de s: sólo depende de los coeficientes β_c, que en este procedimiento no paramétrico son puntos fijos. Por lo tanto, maximizar la fórmula anterior es equivalente a maximizar sólo la primera parte:

$$\mathcal{E}(s|s^t) = \sum_n \sum_c h(\beta_c|i_n, s^t)\, log(s_c).$$

Esta función es muy fácil de maximizar. En particular, el valor de maximización de s_c, teniendo en cuenta la restricción de que la suma de las cuotas debe ser 1, es

$$s_c^{t+1} = \frac{\sum_n h(\beta_c|i_n, s^t)}{\sum_n \sum_{c'} h(\beta_{c'}|i_n, s^t)}.$$

Usando la nomenclatura que hemos definido en la descripción general de los algoritmos EM, $h(\beta_c|i_n, s^t)$ son los pesos, calculados en el valor actual de las cuotas s^t. La cuota actualizada para la clase c es el porcentaje que la suma de los pesos en el punto c representa respecto a la suma de los pesos en todos los puntos.

Este algoritmo EM se implementa mediante los siguientes pasos:

1. Definimos los puntos β_c para c $= 1, \ldots,$ C.
2. Calculamos la fórmula logit para cada persona en cada punto: $L_{ni}(\beta_c)\ \forall n, c$.
3. Especificamos los valores iniciales de las cuotas en cada punto, etiquetadas colectivamente como s^0. Es conveniente que las cuotas iniciales sean $s_c = 1/C\ \forall c$.
4. Para cada persona y cada punto, se calcula la probabilidad de que la persona tenga esos coeficientes condicionando a las elecciones que ha realizado, usando las cuotas iniciales s^0 como las probabilidades no condicionadas: $h(\beta_c|i_n, s^0) = s_c^0 L_{ni}(\beta_c)/P_{ni}(s^0)$. Observe que el denominador es la suma sobre todos los puntos del numerador. Para mayor comodidad, etiquetamos este valor calculado como h_{nc}^0.
5. Actualizamos la cuota de la población en el punto c como $s_c^1 = \frac{\sum_n h_{nc}^0}{\sum_n \sum_{c'} h_{nc'}^0}$.
6. Repita los pasos 4 y 5 utilizando las cuotas actualizadas s en lugar de los valores iniciales originales. Repita estos pasos hasta lograr la convergencia.

Este procedimiento no requiere el cálculo de ningún gradiente ni la inversión de ningún hessiano, algo que los procedimientos descritos en el capítulo 8 sí utilizan para la maximización directa de LL. Por otra parte, las probabilidades logit se calculan sólo una vez (en el paso 2), y no en cada iteración. Las iteraciones consisten en volver a calibrar las cuotas en cada punto, algo que es pura aritmética. Dado que se necesita tan poco cálculo para cada punto, el procedimiento puede implementarse para un gran número de puntos. Por ejemplo, el caso real de Train (2008a) incluía más de 200.000 puntos, y sin embargo la estimación se llevó a cabo en sólo unos 30 minutos. Por el contrario, difícilmente la

maximización directa descrita en los métodos del capítulo 8 habría sido siquiera factible, puesto que implicaría una inversión de un hessiano de 200.000×200.000 valores.

14.3.2 Distribución de mezcla discreta con puntos como parámetros

Podemos modificar el modelo anterior tratando los coeficientes β_c para cada c, como parámetros a estimar en lugar de considerarlos puntos fijos. Los parámetros del modelo son, por lo tanto, la ubicación y el porcentaje o cuota de la población en cada punto. Etiquetamos estos parámetros colectivamente como $\theta = \langle s_c, \beta_c, c = 1, \dots, C \rangle$. Esta especificación a menudo recibe el nombre de *modelo de clases latentes* (*latent class model*): la población está compuesta por C clases distintas, de manera que todas las personas dentro de una clase tienen los mismos coeficientes, siendo los coeficientes diferentes para clases diferentes. Los parámetros del modelo son las proporciones o cuotas de la población en cada clase y los coeficientes de cada clase.

Los datos ausentes de este modelo son la pertenencia de las personas a cada clase. La esperanza empleada por el algoritmo EM es la misma de la especificación anterior, exceptuando que ahora que los coeficientes β_c son tratados como parámetros:

$$\mathcal{E}(\theta|\theta^t) = \sum_n \sum_c h(\beta_c|i_n, s^t)\, log\left(s_c L_{ni_n}(\beta_c)\right).$$

Observe que cada conjunto de parámetros entra solamente en un término dentro del logaritmo: el vector de cuotas s no entra en ninguno de los términos $L_{ni_n}(\beta_c)$s y cada β_c sólo entra a formar parte de $L_{ni_n}(\beta_c)$ para la clase c. Por lo tanto, la maximización de esta función es equivalente a la maximización por separado de cada una de las siguientes funciones:

$$\mathcal{E}(s|\theta^t) = \textstyle\sum_n \sum_c h(\beta_c|i_n, s^t)\, log(s_c), \tag{14.6}$$

y para cada c:

$$\mathcal{E}(\beta_c|\theta^t) = \textstyle\sum_n h(\beta_c|i_n, s^t)\, log\, L_{ni_n}(\beta_c). \tag{14.7}$$

El máximo de (14.6) se alcanza, como antes, en

$$s_c^{t+1} = \frac{\sum_n h(\beta_c|i_n, s^t)}{\sum_n \sum_{c'} h(\beta_{c'}|i_n, s^t)}.$$

Para la actualización de los coeficientes β_c, tenga en cuenta que (14.7) es la función LL para un modelo logit estándar, con cada observación ponderada por $h(\beta_c|i_n, s^t)$. Los valores actualizados de β_c se obtienen mediante la estimación de un modelo logit estándar, donde cada persona proporciona una observación que es ponderada adecuadamente. La misma estimación logit se lleva a cabo para cada clase c, pero con diferentes pesos para cada clase.

El algoritmo EM se implementa a través de los siguientes pasos:

1. Especificamos los valores iniciales de la cuota y de los coeficientes de cada clase, etiquetados como s_c y β_c $\forall c$. Es conveniente que las cuotas iniciales sean $1/C$. He observado que es posible obtener fácilmente valores de inicio para los coeficientes mediante la partición de la muestra en C grupos y la posterior ejecución de un logit en cada grupo[xii].

[xii] Tenga en cuenta que estos grupos no representan una división de la muestra en clases. Las clases son latentes por lo que dicha partición no es posible. En realidad, el objetivo es obtener C conjuntos de

2. Para cada persona y cada clase, calculamos la probabilidad de formar parte de cada clase condicionando a la elección de la persona, usando las cuotas iniciales s^0 como las probabilidades no condicionadas: $h(\beta_c^0|i_n, s^0) = s_c^0 L_{ni}(\beta_c^0)/P_{ni_n}(s^0)$. Observe que el denominador es la suma del numerador para todas las clases. Para mayor comodidad, etiquetamos este valor calculado como h_{nc}^0.
3. Actualizamos la cuota de la clase c como $s_c^1 = \frac{\sum_n h_{nc}^0}{\sum_n \sum_{c'} h_{nc'}^0}$.
4. Actualizamos los coeficientes para cada clase c mediante la estimación de un modelo logit, ponderando cada persona n por h_{nc}^0. Se estiman un total de C modelos logit, usando las mismas observaciones pero con diferentes pesos en cada una.
5. Repita los pasos 2-4 utilizando las cuotas actualizadas s y los coeficientes β_c $\forall c$ en lugar de los valores iniciales originales. Continúe repitiendo estos pasos hasta lograr la convergencia.

Una ventaja de este enfoque es que el investigador puede aplicar la estimación no paramétrica, con las cuotas y los coeficientes de cada clase tratados como parámetros, usando cualquier paquete de software estadístico que incluya una rutina de estimación logit. Para un número dado de clases, este procedimiento requiere mucho más tiempo que el anterior, ya que se deben estimar C modelos logit en cada iteración. Sin embargo, probablemente se necesiten muchas menos clases con este enfoque que con el anterior para representar adecuadamente la verdadera distribución, ya que este procedimiento estima la "mejor" ubicación de los puntos (es decir, los coeficientes), mientras que el anterior usa puntos fijos.

Bhat (1997a) desarrolló un algoritmo EM para un modelo similar a éste, salvo que las cuotas s_c de las clases no son parámetros en sí mismos sino que se especifican para que dependan de las características demográficas de la persona. El algoritmo EM que usó reemplaza nuestro paso 3 con un modelo logit de pertenencia de clase, con la probabilidad condicionada de pertenencia a una clase actuando como la variable dependiente. Su caso práctico muestra el punto fundamental de este capítulo: que los algoritmos EM se pueden desarrollar fácilmente para muchos tipos de modelos de elección complejos.

14.3.3 Distribución de mezcla normal con covarianza completa

En el capítulo 12 vimos que un logit mixto con covarianza plena entre coeficientes puede ser difícil de estimar mediante la maximización estándar de la función LL, debido tanto al gran número de parámetros de covarianza como al hecho de que la LL es altamente no cuadrática. Train (2008b) desarrolló un algoritmo EM muy simple y rápido para logits mixtos con covarianza plena. El algoritmo toma la siguiente forma, muy simple:

1. Especificamos valores iniciales para la media y la covarianza de los coeficientes en la población.
2. Para cada persona, extraemos valores al azar de la distribución de la población utilizando esta media y covarianza iniciales.
3. Ponderamos los valores extraídos de cada persona por la densidad condicionada de las extracciones de esa persona.
4. Calculamos la media y la covarianza de los valores extraídos ponderados de todas las personas. Éstos se convierten en la media y covarianza actualizadas de los coeficientes en la población.

valores iniciales para los coeficientes de las C clases. Estos valores iniciales no deben ser los mismos para todas las clases, ya que si fuesen iguales, el algoritmo realizaría los mismos cálculos para cada clase y devolvería la misma cuota y las mismas estimaciones actualizadas para todas las clases en cada iteración. Una manera fácil de obtener C conjuntos diferentes de valores iniciales es a dividir la muestra en C grupos y estimar un logit en cada grupo.

5. Repetimos los pasos 2-4 usando la media y covarianza actualizadas, y continuamos repitiendo estos pasos hasta que no haya ningún cambio adicional (con una cierta tolerancia) en la media y la covarianza.

Los valores convergentes son las estimaciones de la media y covarianza en la población. No se requieren gradientes. Todo lo que se necesita para la estimación de un modelo con coeficientes totalmente correlacionados es extraer repetidamente valores al azar utilizando la media y covarianza calculadas previamente, ponderar los valores apropiadamente, y calcular la media y covarianza de los valores ponderados.

El procedimiento es fácilmente aplicable cuando los coeficientes se distribuyen normalmente o cuando son transformaciones de términos conjuntamente normales. Siguiendo la notación empleada por Train y Sonnier (2005), la utilidad que obtiene el agente n de la alternativa j es

$$U_{nj} = \alpha_n x_{nj} + \varepsilon_{nj},$$

donde los coeficientes aleatorios son transformaciones de términos distribuidos normalmente: $\alpha_n = T(\beta_n)$, con β_n distribuido normalmente con media b y covarianza W. La transformación permite una flexibilidad considerable en la elección de la distribución. Por ejemplo, una distribución log-normal se obtiene especificando una transformación $\alpha_n = exp(\beta_n)$. Una distribución S_b que tenga un límite superior e inferior, se obtiene mediante la especificación de $\alpha_n = exp(\beta_n)/(1 + exp(\beta_n))$. Por supuesto, si el coeficiente es en sí mismo normal, entonces $\alpha_n = \beta_n$. La densidad normal se denota como $\phi(\beta_n|b, W)$.

Condicionadas a β_n, las probabilidades de elección son logit:

$$L_{ni}(\beta_n) = \frac{e^{T(\beta_n)x_{ni}}}{\sum_j e^{T(\beta_n)x_{nj}}}.$$

Dado que β_n no es conocido, la probabilidad de elección es un logit mixto, mezclado respecto a la distribución de β_n:

$$P_{ni}(b, W) = \int L_{ni}(\beta)\phi(\beta|b, W)d\beta.$$

La función LL es $LL(b, W) = \sum_n \log P_{ni_n}(b, W)$, donde i_n es la alternativa elegida por el agente n.

Como se trata en la sección 12.7, la estimación clásica de este modelo mediante la maximización estándar de LL es difícil, y esta dificultad es una de las razones para usar procedimientos bayesianos. Sin embargo, es posible aplicar un algoritmo EM que es considerablemente más fácil que la maximización estándar y que hace que la estimación clásica sea prácticamente tan conveniente para este modelo como la estimación bayesiana.

Los datos ausentes para el algoritmo EM son los parámetros β_n de cada persona. La densidad $\phi(\beta|b, W)$ es la distribución de β en la población. Para el algoritmo EM, utilizamos la distribución condicionada para cada persona. De acuerdo con la identidad de Bayes, la densidad de β condicionada a la alternativa i escogida por la persona n es $h(\beta|i, b, W) = L_{ni}(\beta)\phi(\beta|b, W)/P_{ni}(b, W)$. La esperanza para el algoritmo EM es

$$\mathcal{E}(b, W|b^t, W^t) = \sum_n \int h(\beta|i_n, b^t, W^t)\, log\big(L_{ni}(\beta)\phi(\beta|b, W)\big)\, d\beta.$$

Observe que $L_{ni}(\beta)$ no depende de los parámetros b y W. Por lo tanto, maximizar esta expectativa respecto a los parámetros es equivalente a maximizar

$$\mathcal{E}(b, W|b^t, W^t) = \sum_n \int h(\beta|i_n, b^t, W^t)\, log\big(\phi(\beta|b, W)\big)\, d\beta.$$

La integral dentro de esta esperanza no tiene una forma cerrada. Sin embargo podemos aproximar dicha integral a través de simulación. Sustituyendo la definición de $h(\cdot)$ y reordenando, tenemos

$$\mathcal{E}(b, W|b^t, W^t) = \sum_n \int \frac{L_{ni_n}(\beta)}{P_{ni_n}(b^t, W^t)} log\big(\phi(\beta|b, W)\big)\phi(\beta|b^t, W^t)\, d\beta.$$

La esperanza respecto a ϕ se simula mediante la extracción de R valores al azar de $\phi(\beta|b^t, W^t)$ para cada persona, etiquetando como β_{nr} el r–ésimo sorteo de la persona n. La esperanza simulada es

$$\tilde{\mathcal{E}}(b, W|b^t, W^t) = \sum_n \sum_r w_{nr}^t\, log\big(\phi(\beta_{nr}|b, W)\big)\,/R.$$

donde los pesos son

$$w_{nr}^t = \frac{L_{ni_n}(\beta_{nr})}{\frac{1}{R}\sum_{r'} L_{ni_n}(\beta_{nr'})}$$

Esta esperanza simulada tiene una forma familiar: es la función LL para una muestra de valores extraídos de una distribución normal, con cada valor ponderado por w_{nr}^t[xiii]. El estimador de máxima verosimilitud de la media y la covarianza de una distribución normal, dada una muestra ponderada de valores extraídos de esa distribución, no es más que la media y la covarianza ponderadas de los valores de la muestra. La media actualizada es

$$b^{t+1} = \frac{1}{NR}\sum_n \sum_r w_{nr}^t \beta_{nr}$$

y la matriz de covarianza actualizada es

$$W^{t+1} = \frac{1}{NR}\sum_n \sum_r w_{nr}^t (\beta_{nr} - b^{t+1})(\beta_{nr} - b^{t+1})'.$$

Observe que W^{t+1} es necesariamente definida positiva, como se requiere para una matriz de covarianza, dado que se construye como la covarianza de los valores extraídos al azar.

El algoritmo EM se implementa de la siguiente manera:

1. Especificamos los valores iniciales de la media y la covarianza, etiquetados b^0 y W^0.
2. Extraemos R valores al azar para cada una de las N personas en la muestra como $\beta_{nr}^0 = b^0 + chol(W^0)\eta_{nr}$, donde $chol(W^0)$ es el factor Choleski triangular inferior de W^0 y η_{nr} es un vector ajustado de valores normales estándar iid.

[xiii] La división por R puede ser ignorada dado que no afecta a la maximización, de la misma forma en que la división por el tamaño de la muestra N se omite en $\mathcal{E}$.

3. Para cada valor extraído de cada persona, calculamos la probabilidad logit de la elección observada de dicha persona: $L_{ni_n}(\beta_{nr}^0)$.
4. Para cada valor extraído de cada persona, calculamos el peso

$$w_{nr}^0 = \frac{L_{ni_n}(\beta_{nr}^0)}{\sum_{r'} L_{ni_n}\left(\beta_{nr'}^0\right)/R}.$$

5. Calculamos la media y la covarianza ponderadas de los $N * R$ valores extraídos $\beta_{nr}^0, r = 1, \dots, R, n = 1, \dots, N$, usando los pesos w_{nr}^0. La media y covarianza ponderada son los parámetros actualizados b^1 y W^1.
6. Repetimos los pasos 2-5 utilizando la media b y la varianza W actualizadas en lugar de los valores iniciales originales. Continuamos repitiendo estos pasos hasta lograr la convergencia.

Este procedimiento puede llevarse a cabo sin necesitad de usar un software de estimación, simplemente extrayendo valores al azar, calculando fórmulas logit para construir los pesos y calculando la media y covarianza ponderadas de los valores. Un investigador puede estimar un modelo logit mixto con covarianza plena y con coeficientes que posiblemente son transformaciones de normales a través de estos sencillos pasos.

Train (2008a) muestra que este procedimiento se puede generalizar para una mezcla finita de normales, donde β se extrae de cualquiera de entre C normales con diferentes medias y covarianzas. La probabilidad de extraer β de cada normal (es decir, la cuota de la población cuyos coeficientes son descritos por cada distribución normal) es un parámetro, junto con las medias y covarianzas. Cualquier distribución se puede aproximar por una mezcla finita de normales, con un número suficiente de normales subyacentes. Al permitir que el número de normales crezca con el tamaño de la muestra, la aproximación se convierte en una forma de estimación no paramétrica de la verdadera distribución de mezcla. El algoritmo EM para este tipo de estimaciones no paramétricas combina los conceptos mostrados en la presente sección para distribuciones normales con plena covarianza y los conceptos que vimos en la sección inmediatamente anterior sobre distribuciones discretas.

Una última observación útil: en los valores de convergencia, las derivadas de $\tilde{\mathcal{E}}(b, W|\hat{b}, \widehat{W})$ proporcionan puntuaciones que pueden ser usadas para estimar los errores estándar asintóticos de las estimaciones. En particular, el gradiente respecto a b y W es

$$\frac{\tilde{\mathcal{E}}(b, W|\hat{b}, \widehat{W})}{db} = \sum_n \left[\frac{1}{R}\sum_r -w_{nr}W^{-1}(\beta_{nr} - b)\right]$$

y

$$\frac{\tilde{\mathcal{E}}(b, W|\hat{b}, \widehat{W})}{dW} = \sum_n \left[\frac{1}{R}\sum_r w_{nr}\left(-\frac{1}{2}W^{-1} + \frac{1}{2}W^{-1}(\beta_{nr} - b)(\beta_{nr} - b)'W^{-1}\right)\right],$$

donde los pesos w_{nr} se calculan en los valores estimados $\hat{b}$ y $\widehat{W}$. Los términos entre corchetes son las puntuaciones de cada persona, las cuales podemos agrupar en un vector etiquetado como s_n. La varianza de las puntuaciones es $V = \sum_n s_n s_n'/N$. La covarianza asintótica del estimador se calcula por lo tanto como V^{-1}/N, como vimos en la sección 8.6.

14.4 Caso de estudio: demanda de coches impulsados por hidrógeno

Train (2008a) estudió las preferencias de los compradores de vehículos impulsados por hidrógeno, utilizando varios de los algoritmos EM que hemos visto. Vamos a describir uno de sus modelos estimados como una ilustración del procedimiento EM. Se realizó una encuesta a compradores de automóviles nuevos en el sur de California para evaluar la importancia que estos compradores otorgaban a varios aspectos relevantes en relación a los vehículos de hidrógeno, tales como la disponibilidad de estaciones de servicio. A cada encuestado se le presentó una serie de 10 experimentos de preferencia declarada. En cada experimento, se solicitó al encuestado que eligiese entre tres alternativas posibles: el vehículo de combustible convencional (*conventional-fuel vehicle*, CV) que el encuestado había comprado recientemente y dos vehículos de combustible alternativo (*alternative-fuel vehicle*, AVs) con los atributos especificados. Se pidió al encuestado que evaluase las tres opciones, indicando cuál consideraba mejor y cuál peor. Los atributos representan las características relevantes de los vehículos de hidrógeno, pero a los encuestados no se les dijo que el combustible alternativo era el hidrógeno a fin de evitar cualquier idea preconcebida que pudiesen haber desarrollado respecto a este tipo de vehículos. Los atributos que se incluyeron en los experimentos son los siguientes:

- Costo del combustible (*fuel cost*, FC), expresado como diferencia porcentual respecto a los vehículos de combustible convencional, CV. En la estimación, el atributo se definió con una escala porcentual, de tal manera que unos costos de combustible un 50 por ciento menores que los del vehículo de combustible convencional entraban en el modelo como -0.5 y los costos un 50 por ciento mayores entraban como 0.5.
- Precio de compra (*purchase price*, PP), expresado como diferencia porcentual respecto al CV, escalado al formar parte del modelo de manera análoga al costo del combustible.
- Radio de conducción (*driving radius*, DR): la distancia más lejana a la que uno puede viajar desde el hogar y posteriormente regresar, partiendo con un depósito lleno de combustible. Según la definición, el radio de conducción es la mitad de la autonomía del vehículo. En los modelos estimados, el DR fue escalado en cientos de millas.
- Destinos convenientes de media distancia (*convenient medium-distance destinations*, CMDD): porcentaje de destinos dentro del radio de conducción que "no requieren una planificación anticipada, ya que es posible abastecerse de combustible durante el camino o en el destino", en contraposición a los destinos que "requieren de abastecimiento de combustible (o al menos estimar si se dispone de suficiente combustible) antes de salir para asegurar que es posible hacer el viaje de ida y vuelta". Este atributo refleja la distribución de los posibles destinos y estaciones de servicio dentro del radio de conducción, reconociendo el hecho de que el depósito de combustible no estará siempre lleno en el momento de arrancar. En los modelos estimados, esta variable se introduce como una cuota o proporción, de tal manera que, por ejemplo, 50 por ciento entra como 0.50.
- Posibles destinos de larga distancia (p*ossible long-distance destinations*, PLDD) : el porcentaje de destinos más allá del radio de conducción que es posible alcanzar gracias a que el reabastecimiento es posible, en oposición a destinos que no se pueden alcanzar debido a la cobertura limitada de estaciones de repostaje. Este atributo refleja la disponibilidad de estaciones de servicio fuera del radio de conducción y su proximidad a potenciales destinos de conducción. Se introdujo en los modelos usando una escala análoga al CMDD.
- Tiempo adicional respecto a las estaciones locales (*extra time to local stations*, ETLS): tiempo de viaje adicional de ida, más allá del tiempo requerido normalmente para encontrar una estación de combustible convencional, que se requiere para llegar a una estación de combustible alternativo en el área local. El ETLS se definió para tener valores de 0, 3 y 10 minutos en los

experimentos; sin embargo, en el análisis preliminar, se encontró que los encuestados consideraron que 3 minutos no representaba ningún inconveniente (es decir, era equivalente a 0 minutos). En los modelos estimados, por lo tanto, se introdujo una variable indicadora para ETLS igual a 10 o diferente de 10, en lugar del ETLS en sí mismo.

En los experimentos, el CV comprado por el encuestado fue descrito como un vehículo con un radio de conducción de 200 millas, CMDD y PLDD igual al 100 por cien y, por definición, ETLS, FC y PP iguales a 0.

Como se indicó anteriormente, se pidió al entrevistado que identificase la mejor y la peor de las tres alternativas, proporcionando así un ranking de los tres vehículos. Condicionadas a los coeficientes de los entrevistados, las probabilidades del ranking se especificaron con la fórmula de "logit expandido" (tal como se describe en la sección 7.3.1). Mediante esta formulación, la probabilidad del ranking es la probabilidad logit de la primera elección entre las tres alternativas posibles del experimento, por la probabilidad logit de la segunda elección entre las dos alternativas restantes. Esta probabilidad se combina con la distribución de probabilidad de los coeficientes, cuyos parámetros han sido estimados.

Se aplicaron los tres métodos que hemos descrito anteriormente. Nos concentramos en el método de la sección 14.3.2, ya que proporciona una ilustración sucinta de la potencia del algoritmo EM. Para este método, hay C clases de compradores, y los coeficientes β_n y las cuotas s_c de la población en cada clase son tratados como parámetros.

Train (2008a) estimó el modelo con diferente número de clases, que van desde una clase (que es un logit estándar) hasta 30 clases. La tabla 14.1 muestra el valor de la LL para estos modelos. Aumentar el número de clases mejora la LL considerablemente, desde -7.884.6 con una clase hasta -5.953.4 con 30 clases. Por supuesto, un mayor número de clases implica más parámetros, lo que plantea la cuestión de si el ajuste mejorado justifica el esfuerzo de tratar parámetros adicionales. En situaciones como ésta, es habitual evaluar los modelos por el criterio de información de Akaike (a*kaike information criterion*, AIC) o mediante el criterio de información bayesiano (*bayesian information criterion*, BIC)[xiv]. Los valores de estos estadísticos también se muestran en la tabla 14.1. El AIC es más bajo (mejor) con 25 clases y el BIC, que penaliza en mayor medida el uso de parámetros adicionales que el AIC, es más bajo con 8 clases.

Tabla 14.1. Modelos logit mixtos con distribuciones discretas de coeficientes y diferente número de clases

Clases	Log-verosimilitud (LL)	Parámetros	AIC	BIC
1	−7,884.6	7	15,783.2	15,812.8
5	−6,411.5	39	12,901.0	13,066.0
6	−6,335.3	47	12,764.6	12,963.4
7	−6,294.4	55	12,698.8	12,931.5
8	−6,253.9	63	12,633.8	12,900.3
9	−6,230.4	71	12,602.8	12,903.2
10	−6,211.4	79	12,580.8	12,915.0
15	−6,124.5	119	12,487.0	12,990.4
20	−6,045.1	159	12,408.2	13,080.8
25	−5,990.7	199	12,379.4	13,221.3
30	−5,953.4	239	12,384.8	13,395.9

[xiv] 6 Véase, por ejemplo, Mittelhammer et. al. (2000, sección 18.5) para una exposición relativa a los criterios de información. El AIC (Akaike, 1974) es - 2LL + 2K, donde LL es el valor del logaritmo de la verosimilitud y K es el número de parámetros. El BIC, también llamado criterio de Schwarz (1978), es - 2LL + log (N) K, donde N es el tamaño de la muestra.

A efectos de evaluar el algoritmo EM, es útil tener en cuenta que la estimación de estos modelos requirió un tiempo de ejecución en torno a 1.5 minutos por clase, pariendo de los valores iniciales hasta lograr la convergencia. Esto significa que el modelo con 30 clases, que tiene 239 parámetros[xv], se estimó en tan sólo 45 minutos.

La tabla 14.2 presenta las estimaciones para el modelo con 8 clases, la mejor opción de acuerdo al criterio BIC. El modelo con 25 clases, el mejor modelo según el criterio AIC, proporciona aún mayor detalle pero no se proporciona en aras de la brevedad. Como se muestra en la tabla 14.2, la mayor de las 8 clases es la última, con el 25 por ciento. Esta clase tiene un coeficiente positivo grande para CV, a diferencia de todas las otras clases. Por lo tanto, esta clase aparentemente se compone de personas que prefieren su CV frente a los AV, incluso cuando los AV tienen los mismos atributos, tal vez a causa de la incertidumbre asociada a las nuevas tecnologías de combustible. Otras características distintivas de las clases son evidentes. Por ejemplo, la clase 3 se preocupa más por el PP (precio de compra) que las otras clases, mientras que la clase 1 da más importancia al FC (costo de combustible) que las otras clases.

Tabla 14.2. Modelo con 8 clases

Clase	1	2	3	4
Cuotas	0.107	0.179	0.115	0.0699
Coeficientes				
FC	−3.546	−2.576	−1.893	−1.665
PP	−2.389	−5.318	−12.13	0.480
DR	0.718	0.952	0.199	0.472
CMDD	0.662	1.156	0.327	1.332
PLPP	0.952	2.869	0.910	3.136
ETLS=10 (variable indicadora)	−1.469	−0.206	−0.113	−0.278
CV (variable indicadora)	−1.136	−0.553	−0.693	−2.961
Clase	5	6	7	8
Cuotas	0.117	0.077	0.083	0.252
Coeficientes				
FC	−1.547	−0.560	−0.309	−0.889
PP	−2.741	−1.237	−1.397	−2.385
DR	0.878	0.853	0.637	0.369
CMDD	0.514	3.400	−0.022	0.611
PLPP	0.409	3.473	0.104	1.244
ETLS=10 (variable indicadora)	0.086	−0.379	−0.298	−0.265
CV (variable indicadora)	−3.916	−2.181	−0.007	2.656

La tabla 14.3 muestra la media y la desviación estándar entre coeficientes de las 8 clases. Como Train (2008a) señala, estas medias y desviaciones estándar son similares a las obtenidas con un modelo logit mixto más estándar, con coeficientes distribuidos normalmente (coeficientes que se pueden consultar en su artículo publicado, pero que no se repiten aquí). Este resultado indica que el uso en este caso práctico de numerosas clases, algo que el algoritmo EM hace posible, proporciona mayor detalle en la explicación de las diferencias en las preferencias, manteniendo al mismo tiempo estadísticos resumidas muy similares.

[xv] Siete coeficientes y una cuota de mercado por cada una de las 30 clases, con una cuota de clase determinada por la restricción de que las cuotas deben sumar uno.

Tabla 14.3. Estadísticos resumidos de los coeficientes

	Medias		Desviaciones estándar	
	Est.	EE	Est.	EE
Coeficientes				
FC	−1.648	0.141	0.966	0.200
PP	−3.698	0.487	3388	0.568
DR	0.617	0.078	0.270	0.092
CMDD	0.882	0.140	0.811	0.126
PLPP	1575	0.240	1098	0.178
ETLS=10 (variable indicadora)	−0.338	0.102	0.411	0.089
CV (variable indicadora)	−0.463	1181	2142	0.216

Est=Estimación, EE=Error Estándar

Sería difícil calcular los errores estándar de las fórmulas asintóticas para este modelo (es decir, calcularlos mediante la inversa del hessiano estimado), debido a la gran cantidad de parámetros existentes. Además, estamos interesados en los estadísticos resumidas, como la media y la desviación estándar de los coeficientes entre todas las clases, dadas en la tabla 14.4. Obtener los errores estándar de estos estadísticos resumidos a partir de fórmulas asintóticas de la covarianza de los parámetros mismos sería computacionalmente difícil. En cambio, es posible calcular fácilmente los errores estándar mediante *bootstrapping*. Dada la velocidad del algoritmo EM en este caso práctico, usar *bootstrapping* es factible. Asimismo, *bootstrapping* proporciona automáticamente los errores estándar de nuestros estadísticos resumidos (mediante el cálculo de los estadísticos resumidos para cada estimación *bootstrap* y tomando sus desviaciones estándar).

Tabla 14.4. Errores estándar para la clase 1

	Est.	EE
Coeficientes		
FC	−3.546	2.473
PP	−2.389	6.974
DR	0.718	0.404
CMDD	0.662	1.713
PLPP	0.952	1.701
ETLS=10 (variable indicadora)	−1.469	0.956
CV (variable indicadora)	−1.136	3.294

Est=Estimación, EE=Error Estándar

Los errores estándar para los estadísticos resumidos se facilitan en la tabla 14.3, basados en 20 muestras de *bootstrapping*. Los errores estándar no se proporcionan en la tabla 14.2 para los parámetros de cada clase. En lugar de ello, la tabla 14.4 da los errores estándar para la clase 1, como un ejemplo ilustrativo de todas las clases. Tal y como se muestra en dicha tabla, los errores estándar de los parámetros de la clase 1 son elevados. Estos errores estándar elevados eran de esperar, y surgen por el hecho de que el etiquetado de clases en este modelo es arbitrario. Supongamos, como ejemplo extremo pero ilustrativo, que las dos muestras diferentes de *bootstrapping* dan las mismas estimaciones para dos clases pero con su orden cambiado (es decir, las estimaciones para la clase 1 convirtiéndose en las estimaciones para la clase 2 y viceversa). En este caso, los errores estándar obtenidos por *bootstrapping* para los parámetros de ambas clases aumentan a pesar de que el modelo para estas dos clases juntas es exactamente el mismo. Los estadísticos resumidos evitan este problema. Todas las medias excepto una son

estadísticamente significativas, siendo la variable indicadora de CV la única que obtiene una media no significativa. Todas las desviaciones estándar son significativamente diferentes de cero.

Train (2008a) también estimó otros dos modelos a partir de estos mismos datos utilizando algoritmos EM: (1) un modelo con una distribución discreta de coeficientes, donde los puntos son fijos y las cuotas de población en cada punto se estiman usando el procedimiento de la Sección 14.3.1 y (2) un modelo con una distribución de mezcla discreta, formada por dos distribuciones normales con covarianza plena, utilizando una generalización del procedimiento de la sección 14.3.3. La flexibilidad de los algoritmos EM para dar cabida a una amplia variedad de modelos complejos es la razón por la que vale la pena aprender su uso. Mejoran la capacidad del investigador para construir modelos diseñados a medida, que se ajustan estrechamente a la realidad de la situación y de los objetivos de la investigación, algo que ha sido el objetivo primordial de este libro.