eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

TAREA 5

EJERCICIOS

1. Este es un ejercicio de aplicación del algoritmo EM a un caso simple (distribución normal bivariante). Hay detalles abundantes al final (para Esther, que aún no ha tomado la asignatura de Multivariante). Pueden verse también detalles en [10], Sección 11.2.3.

   a) Genera 100 vectores bivariantes con distribución $N(\mu, \Sigma)$. Toma μ arbitrario y Σ que implique correlación 0.9 entre las dos componentes de los vectores. Tienes así una muestra artificial procedente de una distribución completamente especificada.

   b) Ordena las 100 observaciones de acuerdo al valor la la segunda variable, X_2.

   c) Elimina las 50 últimas observaciones de X_1 (la segunda componente de los vectores). Tienes ahora una muestra incompleta de la misma distribución anterior.

   d) Estima haciendo uso de las observaciones completas la media y matriz de covarianzas. Compara con μ y Σ (que conoces por haberlas fijado previamente).

   e) Estima ahora μ y Σ haciendo uso del algoritmo EM. Deberás obtener una estimación mucho mejor. Explica por qué.

2. Queremos contrastar $H_0 : X \sim N(0, 3)$ frente a $H_a : X \sim N(0, 1)$, para lo que contamos con una m.a.s. $X_1, \ldots, X_n$.

   a) Encuentra la forma de la región crítica más potente para realizar dicho contraste.

   b) Fija la o las constantes que sean precisas de modo que el contraste obtenido en el apartado anterior tenga $\alpha = 0{,}05$.

   c) Calcula la potencia del contraste obtenido en el apartado anterior.

3. Sabemos que una v.a. sigue distribución exponencial (con densidad $f_X(x, \theta) = \theta^{-1} e^{-x/\theta}$ para $x > 0$). Queremos contrastar $H_0 : \theta = \theta_0$ frente a $H_a : \theta = \theta_a > \theta_0$, para lo que contamos con una m.a.s. $X_1, \ldots, X_n$.

   a) Obtén el contraste más potente de tamaño $\alpha = 0{,}05$ y calcula su potencia cuando $n = 20$.

4. Supón una hipótesis nula H_0 sobre la que deseas hacer un contraste de significación puro, empleando el estadístico de contraste $T = T(\vec{X})$. Realizas la toma de la muestra, y obtienes un valor $t_{\text{obs}} = T(\vec{x})$ del estadístico. La distribución $F_T(t)$ bajo H_0 es conocida.

   a) ¿Cuál es la distribución de $P = F_T(T)$?

   b) ¿Cuál es la distribución de $-2\log_e P$?

   c) Supón que puedes hacer no uno sino k contrastes, sobre muestras del mismo tamaño procedentes de la misma población. Las muestras están formadas por observaciones independientes, entre sí y de las demás muestras, de modo que puedes obtener k estadísticos $T_1, \ldots, T_k$ independientes. Haz uso de (4*b*) para construir un contraste de significación que haga uso de toda la información. ¿Es crucial que los estadísticos $T_1, \ldots, T_k$ tengan la misma distribución bajo H_0?

5. *(sencillo, pero largo; pretende ilustrar lo idiota que resulta contrastar hipótesis "à la Neyman-Pearson" con niveles de significación convencionales, sin reflexionar sobre el problema de fondo.)* Te enfrentas al problema de controlar la calidad de remesas de fruta. Recibes naranjas, en envases cerrados. Las especificaciones técnicas y contractuales son las siguientes.

   - Cada partida es de diez mil cajas. Los envíos se hacen por mar y, sea por mala estiba, sea por problemas sobrevenidos en la navegación, en el pasado se han presentado "malas" remesas en algunas ocasiones. Por "malas" se entiende que un 40 % de las cajas llegaban con su contenido mohoso e invendible.
   - Cuando las remesas no son "malas", son "buenas": ello no quiere decir que todas las cajas sean perfectas. Se considera buena una remesa con el 5 % de cajas en malas condiciones. No hay remesas "regulares": o son "malas" o son "buenas".
   - La experiencia precedente muestra que en aproximadamente un 10 % de los casos las remesas son malas, y en el 90 % de las ocasiones buenas.
   - No sabes cómo ni porqué, pero la empresa para la que trabajas pactó con el proveedor de la fruta y el consignatario del buque que las remesas se aceptarían o rechazarían en el acto, sobre el muelle; y que el comprador (o sea, tu empresa) estaría facultado para abrir diez cajas antes de decidir aceptar o rechazar la remesa.
   - El coste de aceptar una remesa "mala" asciende a 15 millones de pesetas, entre fruta que hay que tirar, fletes, abonos a nuestros clientes, etc.
   - El coste de rechazar una remesa "buena" es de 20 millones: tanto proveedor como consignatario no tienen a quien venderla, y ante la alternativa de tirarla al mar, tenemos la certeza de que abriran todas las cajas de una remesa rechazada. Si contando una por una hay 500 ó menos cajas defectuosas (el 5 % tolerado), habremos de indemnizarles con la citada cantidad.

   Con la anterior especificación del problema,

   a) Diseña un contraste para la hipótesis nula H_0 : "Remesa buena" frente a la alternativa H_a : "Remesa mala" que sea el más potente de tamaño $\alpha \leq 0{,}05$.

   b) Calcula el riesgo de Bayes derivado de emplear dicho contraste como procedimiento de decisión.

   c) ¿Cual sería el procedimiento de Bayes?

Ayuda: Hay once posibles resultados de abrir las diez cajas: de cero a diez defectuosas. Parece sensato considerar once únicos procedimientos (¿cuáles?).

d) Compara el procedimiento de Bayes con el que se obtendría de la aplicación rutinaria de un contraste con $\alpha = 0{,}05$ y escribe el comentario que tu profesor quiere leer.

e) *(mucho más arduo; no es preciso que lo hagas —requerirías un ordenador—, pero sí que lo pienses, y bosquejes el proceso que seguirías.)* Supón que el abrir las cajas e inspeccionar su contenido es un procedimiento destructivo: ya no se puede vender la caja, y se pierde su valor de 3000 ptas. Este coste es a cuenta del comprador. Si no estuvieras constreñido a muestrear precisamente diez cajas, sino las que desearas ¿que harías?

Observaciones, ayudas, comentarios

1. Es muy fácil generar con ayuda de R vectores aleatorios normales con una distribución cualquiera. Basta,

   a) Generar los vectores con vector de medias **0** y matriz de covarianzas **I**. Para ello, en vuestro caso basta algo como:

   ```
   > X <- matrix(rnorm(200), 100, 2)
   > X[1:3, ]
              [,1]       [,2]
   [1,] -1.2070657  0.4145235
   [2,]  0.2774292 -0.4747185
   [3,]  1.0844412  0.0659935
   ```

   b) Introducir la correlación y media deseadas. Para ello, dado que la matriz de covarianzas de $\mathbf{v} = \mathbf{Au}$ es $\mathbf{\Sigma} = \mathbf{AA'}$, todo lo que hace falta es multiplicar *cada fila* de `X` por **A'**. La factorización de Cholesky es uno de los posibles métodos para obtener **A'**.

   ```
   > Sigma <- matrix(c(1, 0.9, 0.9, 1), 2, 2)
   > Sigma
        [,1] [,2]
   [1,]  1.0  0.9
   [2,]  0.9  1.0
   > Ap <- chol(Sigma)
   ```

   Comprobemos que funciona:

   ```
   > A <- t(Ap)
   > A %*% Ap
        [,1] [,2]
   [1,]  1.0  0.9
   [2,]  0.9  1.0
   ```

   $\sum_i \boldsymbol{x_i x_i'}$ dados $\boldsymbol{\mu^{(0)}}$, $\mathbf{\Sigma^{(0)}}$. Ahora podemos incorporar la correlación y un vector de medias como $\boldsymbol{\mu'} = (5 \quad 7)$ así:

   ```
   > Y <- X %*% Ap
   > Y <- Y + matrix(c(5, 7), 100, 2, byrow = TRUE)
   > cov.wt(Y)
   ```

```
$cov
          [,1]      [,2]
[1,] 1.0088300 0.8964764
[2,] 0.8964764 0.9989333

$center
[1] 4.843238 6.876892

$n.obs
[1] 100
```

(un modo más rápido y elegante de sumar el vector de medias sería `Y <- sweep(Y,2,-c(5,4))`; puedes ver la ayuda de la función `sweep`).

2. Se puede demostrar[1] que la distribución condicional de X_1 dado $X_2 = x_2$ cuando $\mathbf{X'} = (X_1 \quad X_2)$ sigue una distribución normal multivariante, es:

$$X_1|X_2 = x_2 \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \tag{1}$$

3. Hemos visto que en una normal multivariante, estadísticos conjuntamente suficientes para $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ son: $\sum_i \boldsymbol{x_i}$ y $\sum_i \boldsymbol{x_i x_i'}$. Las estimaciones máximo verosímiles del vector de medias y matriz de covarianzas con datos completos son entonces:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x_i} \tag{2}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x_i x_i'} - \overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}' \tag{3}$$

4. Si los datos son completos, las expresiones anteriores dan sin más los estimadores máximo-verosímiles. Si los datos son incompletos, *hemos de sustituir los estadísticos suficientes anteriroes por sus valores medios*, calculados con ayuda de la mejor aproximación disponible que temgamos de los parámetros.

   a) No hay problema con $\frac{1}{N}\sum_{i=1}^{N}\sum_i \boldsymbol{x_i}$ y $\sum_i \boldsymbol{x_i x_i'}$ dados $\boldsymbol{\mu^{(0)}}$, $\boldsymbol{\Sigma^{(0)}}$. Basta sustituir los valores de X_1 no observados por sus valores medios condicionados de acuerdo con la expresión (1).

   b) Hay un ligero problema con $\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x_i x_i'}$ porque cuando un valor X_{i1} falta y lo reemplazamos por $\hat{X}_{i1} = E[X_{i1}|X_{i2}]$, entonces $E[\hat{X}_{i1}^2] \neq [E[\hat{X}_{i1}]]^2$ (¡otra vez la desigualdad de Jensen!). Hemos de añadir la estimación de la varianza residual $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

5. Una vez tenido en cuenta el detalle del párrafo anterior, el algoritmo EM procede así:

   a) Obtener estimaciones iniciales $\boldsymbol{\mu^{(0)}}$, $\boldsymbol{\Sigma^{(0)}}$ (por ejemplo, a partir de los datos completoxs).

   b) Obtener los valores medios de $\sum_i \boldsymbol{x_i}$ y $\sum_i \boldsymbol{x_i x_i'}$ dados $\boldsymbol{\mu^{(0)}}$, $\boldsymbol{\Sigma^{(0)}}$. Para el primero, basta reemplazar los valores perdidos por sus estimaciones. Para el segundo hay que tener en cuenta además la corrección en $4b$.

[1] Véase cualquier manual de Análisis Multivariante, como por ejemplo [10] o [4].

c) Con los valores medios de los estadísticos suficientes así obtenidos, estimar $\boldsymbol{\mu}^{(1)}$ y $\boldsymbol{\Sigma}^{(1)}$, haciendo uso de (2)–(3).

d) Volver a calcular valores medios de $\sum_i \boldsymbol{x_i}$ y $\sum_i \boldsymbol{x_i x_i'}$ dados $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\Sigma}^{(1)}$.

e) ...y así iterar hasta convergencia.

Tenéis un ejemplo completo desarrollado en [10], Ejemplo 11.2, si lo anterior todavía no resulta suficientemente claro.

Lectura recomendada. Todos los manuales que se citan a continuación son de interés: [1], [2], [5], [7], [8], [12]. [9] es una referencia ya clásica, pero de nivel superior. Hay problemas resueltos en todos ellos y en [6], [11] y [3].

Referencias

[1] P. J. Bickel and K. A. Doksum. *Mathematical Statistics*. Holden-Day, Inc., San Francisco, 1977.

[2] D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman and Hall, London, 1979 edition, 1974.

[3] D. R. Cox and D. V. Hinkley. *Problems and Solutions in Theoretical Statistics*. Chapman and Hall, London, 1980 edition, 1980.

[4] C.M. Cuadras. *Métodos de Análisis Multivariante*. Eunibar, Barcelona, 1981.

[5] E.J. Dudewicz and S.N. Mishra. *Modern Mathematical Statistics*. Wiley, 1988.

[6] A. Garín and F. Tusell. *Problemas de Probabilidad e Inferencia Estadística*. Ed. Tébar-Flores, Madrid, 1991.

[7] P.H. Garthwaite, I.T. Jolliffe, and B. Jones. *Statistical Inference*. Prentice Hall, London, 1995.

[8] E. L. Lehmann. *Theory of Point Estimation*. Wiley, New York, 1983.

[9] E. L. Lehmann. *Testing Statistical Hypothesis*. Chapman & Hall, 2 edition, 1986.

[10] D. Peña. *Análisis de Datos Multivariantes*. McGraw-Hill, 2002.

[11] J. P. Romano and A. F. Siegel. *Counterexamples in Probability and Statistics*. Wadsworth and Brooks/Cole, Monterrey, California, 1986.

[12] G.A. Young and R.L. Smith. *Essentials of Statistical Inference*. Cambridge Univ. Press, 2005.