eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

TAREA 1

EJERCICIOS

Por la teoría vista en clase sabes que cuando $\boldsymbol{X}$ es normal multivariante, la distribución de $\boldsymbol{X}_1|\boldsymbol{X}_2 = \boldsymbol{x}_2$ viene dada por

$$N(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}). \tag{1}$$

Los ejercicios que siguen son de manipulación y tienen por objeto que compruebes empíricamente algunas de las cosas estudiadas en teoría.

1. Genera 500 observaciones normales multivariantes con vector de medias $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 \quad \boldsymbol{\mu}_2)^T = (3 \quad 4)^T$ y matriz de covarianzas
$$\Sigma = \begin{pmatrix} 1 & 0{,}8 \\ 0{,}8 & 1 \end{pmatrix}. \tag{2}$$

2. Estima máximo-verosímilmente el vector de medias y la matriz de covarianzas. Con un tamaño de muestra $N = 500$ debes obtener estimaciones bastante ajustadas.

3. Como has generado artificialmente las observaciones, sabes cual es su función de densidad teórica. Calcúlala para cada observación, divide el rango de valores que obtengas en tres intervalos (los cien mayores, los doscientos siguientes y los doscientos menores) y representa las observaciones correspondientes. ¿Qué obtienes?

4. ¿Que orientación (paralela a los ejes, SW-NE, NW-SE, ...) tiene cualquiera de las nubes de puntos representadas en el apartado anterior?

5. Selecciona un valor de $\boldsymbol{X}_2$ (por ejemplo, 0.8) y calcula la media y varianza teóricas de $\boldsymbol{X}_1$ para observaciones con $\boldsymbol{X}_2 = 0{,}8$.

6. Siendo la distribución de $\boldsymbol{X}_2$ continua, sólo por autentica casualidad obtendrías alguna de tus 500 observaciones con $\boldsymbol{x}_2 = 0{,}8$. Por tanto, no puedes estimar la media obtenida en el apartado anterior de modo empírico. Pero si puedes calcular la media aritmética de valores de $\boldsymbol{X}_1$ en observaciones con $\boldsymbol{X}_2 \approx 0{,}8$ (por ejemplo, entre 0.7 y 0.9); no debería separarse mucho del valor teórico obtenido en el apartado anterior.

7. Idem. con la varianza de $\boldsymbol{X}_1|\boldsymbol{X}_2 = 0{,}8$.

8. Genera ahora 500 observaciones normales trivariantes, $\boldsymbol{X} = (X_1, X_2, X_3)^T$ con vector de medias $\mathbf{0}$ y matriz de covarianzas unidad. Obtén un nuevo vector de variables normales multivariantes (Y_1, Y_2, Y_3) a partir de $\boldsymbol{X}$ del siguiente modo:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} 1 & -1 & 3 \\ 1 & 1 & 3 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}. \tag{3}$$

   Por simple inspección ¿cuál dirías que es la correlación de Y_1 con Y_2? ¿Y la correlación *parcial* controlado el efecto de X_3?

9. Verifica tu intuición en el apartado anterior calculando los valores teóricos de las correlaciones indicadas.

10. Comprueba, a continuación, que las estimaciones de las respectivas correlaciones son razonablemente aproximadas a sus valores teóricos.

AYUDAS, SUGERENCIAS Y COMPLEMENTOS

1. Para generar observaciones normales multivariantes con vector de medias y matriz de covarianzas arbitraria Σ, sólo tienes que generar normales multivariantes con vector de medias $\mathbf{0}$ y matriz de covarianzas unidad, y hacer la oportuna transformación lineal

$$\boldsymbol{Y} = \Sigma^{1/2}\boldsymbol{X} + \boldsymbol{\mu}_1.$$

   Es fácil ver tomando valores medios y matrices de covarianzas en la ecuación anterior que $E[\boldsymbol{Y}] = \boldsymbol{\mu}_1$ y $\Sigma_{\boldsymbol{Y}} = \Sigma^{1/2} I \Sigma^{1/2} = \Sigma$, como se desea.

   Para implementar lo anterior necesitas la "raíz cuadrada" de Σ. Hay infinidad de posibilidades; alguna se sugirió en la tarea precedente (factorización de Cholesky). Puedes también recurrir a diagonalizar la matriz objetivo Σ, tomar las raíces cuadradas de los valores propios (¿por qué tenemos la certeza de que dichas raíces existen y son reales?) y "des-diagonalizar".

2. Si decides emplear la factorización de Cholesky, puedes emplear la función `chol` en R. Observa:

```
> Sigma    <- matrix(c(1,0.8,0.8,1),nrow=2)
> Sigma0.5 <- chol(Sigma)
> Sigma0.5

     [,1] [,2]
[1,]    1  0.8
[2,]    0  0.6

> t(Sigma0.5) %*% Sigma0.5

     [,1] [,2]
[1,]  1.0  0.8
[2,]  0.8  1.0
```

```
> crossprod(Sigma0.5)

     [,1] [,2]
[1,]  1.0  0.8
[2,]  0.8  1.0
```

3. Cualquiera de los manuales citados en la bibliografía del programa te servirá: Rencher (1998), Jobson (1991), Rencher (1995), Johnson y Wichern (1992), Peña (2002) Cuadras (1981), Dillon y Goldstein (1984), etc.

   Sobre cuestiones relacionadas con cálculo matricial (como el modo de calcular el factor de Cholesky) cualquier texto de análisis numérico, incluyendo Lange (1998), Golub y Loan (1989), J. E. Gentle, Härdle, y Mori (2004) o J. Gentle (2007).

Referencias

Cuadras, C. M. (1981). *Métodos de análisis multivariante*. Barcelona: Eunibar.

Dillon, W., y Goldstein, M. (1984). *Multivariate analysis: Methods and applications*. New York: Wiley.

Gentle, J. (2007). *Matrix algebra: Theory, computations, and applications in statistics*. Springer.

Gentle, J. E., Härdle, W., y Mori, Y. (Eds.). (2004). *Handbook of computational statistics. concepts and methods*. Springer-Verlag.

Golub, G. H., y Loan, C. F. van. (1989). *Matrix computations*. Baltimore: John Hopkins Univ. Press.

Jobson, J. D. (1991). *Applied multivariate data analysis, vol. II*. New York: Springer Verlag. (Signatura: 519.237 JOB)

Johnson, R. A., y Wichern, D. W. (1992). *Applied multivariate statistical analysis*. Prentice-Hall International.

Lange, K. (1998). *Numerical analysis for statisticians*. Springer. (Signatura: 519.6 LAN)

Peña, D. (2002). *Análisis de datos multivariantes*. McGraw-Hill.

Rencher, A. C. (1995). *Methods of multivariate analysis*. Wiley.

Rencher, A. C. (1998). *Multivariate statistical inference and applications*. Wiley.