



Unibersidad
del Pa s Vasco

Euskal Herriko
Unibertsitatea

Programa de la asignatura

Estad stica: An lisis Multivariante (15763)

Curso 2007–2008

Profesor: Fernando TUSELL

Descripción

Objetivos. Proporcionar una panorámica teórica y una base práctica que permita al alumno(a) servirse de los métodos de análisis multivariante de mayor uso en el ejercicio profesional de un economista.

Orientación bibliográfica. Como manuales generales, cualquiera de los siguientes es una buena elección: [10], [24], [11], [19], [28]; seguiremos preferentemente [24], pero con frecuente referencia a otros libros. Buenas monografías sobre teoría de distribuciones multivariantes (que va mucho más allá de lo que vemos en el curso) son: [3], [22] y el manual ya citado [28], entre otros. En castellano, además de las citadas, se dispone también de [18], pero es relativamente superficial y orientado al empleo de paquetes enlatados.

Hay además bibliografía muy abundante para cada tema, mucha de ella en la Facultad. En la biblioteca, los libros sobre Análisis Multivariante en general están en 519.237. Los específicos sobre componentes principales, análisis factorial y, en general, técnicas de reducción de la dimensionalidad están en 519.237.7, y en 519.237.8 los que hacen referencia a reconocimiento de pautas, análisis de agrupamientos y cuestiones conexas. Hay una pequeña división (519.237.35) con libros de análisis discriminante.

Los libros sobre redes neuronales están mayoritariamente en 681.83.

Puede utilizarse para algunas cosas [31] (esquemático e incompleto), pero la disponibilidad de notas del curso no debiera disuadir de consultar la bibliografía. Son una ayuda, no un sustituto.

Evaluación y desarrollo del curso. En un curso normal se realizan en torno a diez tareas semanales o decenales, que se corrigen en todo o en parte (dependiendo del número de alumnos matriculados) y se devuelven y comentan en clase. Hay además un examen final. La nota es un promedio de todo ello.

No es posible, por lo general, acomodar todos los temas en el tiempo disponible. Algunos de los que aparecen marcados con un asterisco se omiten, variando la selección de un año a otro. (Este curso 2007-2008, sin embargo, hay un inusual número de clases, al no coincidir en Lunes, Martes o Miércoles ninguna de las festividades del cuatrimestre, por lo que la pauta habitual puede variar.)

Las prácticas se realizan con R. Los alumnos disponiendo de ordenadores personales reciben, si lo desean, una copia y algún software adicional. Es más que suficiente para realizar todas las prácticas.

Actualizaciones. La versión más moderna de este programa, de las notas empleadas en el curso, de los ficheros de datos, de algún software de libre uso (como R) y de los enunciados de las tareas está disponible en E-KASI.

Temario

1. REPASO DE ALGEBRA LINEAL

(A cargo del alumno.) Repaso de cuestiones en su mayoría ya conocidas de la asignatura ESTADÍSTICA: MODELOS LINEALES. Incidiremos en clase en algunas cuestiones sobre diagonalización y descomposición en valores singulares.

BIBLIOGRAFÍA: Un libro de consulta clásico es [27]. [1], bajo la forma de libro de problemas, sirve como un completo manual.

DISTRIBUCIÓN NORMAL MULTIVARIANTE Y ASOCIADAS A ELLA.

Distribución normal singular y no singular. Distribuciones marginales de la normal. Distribuciones normales condicionadas. Coeficiente de correlación y coeficiente de correlación parcial. Contrastes sobre el coeficiente de correlación. Distribuciones de Wishart y Wilks.

[31]. [28] Sec. 2.2, 2.3. [10] Sec. 2.1 a 2.4. [19] Cap.4. [24] Cap. 9.

2. CONTRASTES SOBRE EL VECTOR DE MEDIAS Y MATRIZ DE COVARIANZAS EN LA DISTRIBUCIÓN NORMAL MULTIVARIANTE.

Distribución del estadístico T^2 de Hotelling. Contrastes sobre el vector de medias con matriz de covarianza conocida. Contrastes de igualdad de medias cuando la matriz de covarianzas es conocida. Id. cuando la matriz de covarianzas es desconocida. Contraste de hipótesis lineales en general.

BIBLIOGRAFÍA: [31]. [28] Sec. 2.4, 3.1, 3.2, 3.3 (parte), 3.4 (parte), y 3.5 (parte). [24] Cap. 10.

3. ANÁLISIS DE CORRELACIÓN CANÓNICA.

Planteamiento del problema. Contraste de la hipótesis de independencia lineal entre dos grupos de variables. Variables canónicas y correlaciones canónicas. Interpretación geométrica. Computo de correlaciones canónicas. Contraste de hipótesis de dimensionalidad. Ejemplos.

BIBLIOGRAFÍA: [31]. [10] Cap. 22. [19] Cap. 10. [24] Cap. 16. [29].

4. REGRESIÓN MULTIVARIANTE.

El modelo de regresión lineal multivariante. Estimación. Contraste de hipótesis. MANOVA. Ejemplos.

BIBLIOGRAFÍA: [17], vol. II, 7.5.

5. ANÁLISIS DISCRIMINANTE (I).

Naturaleza del problema, y distintas aproximaciones a su solución. Funciones discriminantes: el enfoque de Fisher. Interpretación gráfica. Minimización de la probabilidad total de error. Minimización del coste total de error.

BIBLIOGRAFÍA: [10] Cap. 19. [28] Cap. 6 (parte). [24] Cap. 12.

6. ANÁLISIS DISCRIMINANTE (II).

Discriminación Minimax. Funciones de discriminación cuadráticas. Discriminación en el caso de varios grupos: análisis discriminante canónico. Formas de evaluar un procedimiento discriminante: sesgo de la tasa de error aparente. Jackknifing. Bbreve referencia a procedimientos alternativos de discriminación: árboles, máquinas de vectores soporte, regresión logística.

BIBLIOGRAFÍA: [10] Cap. 19. [28] Cap. 6(parte). [23].

7. COMPONENTES PRINCIPALES. Componentes principales. Interpretación. Ventajas e inconvenientes. Ejemplos.

Clase práctica: Visualización de datos. Uso de R como paquete gráfico (I).

BIBLIOGRAFÍA: [31]. [28] Sec.5.2, y 4.5.4 (contrastes de esfericidad). [10] Cap. 12. [19]

8. ANÁLISIS FACTORIAL (I).

El modelo factorial. Teorema de Thurstone. El problema de las comunalidades. Determinación teórica de la comunalidad. Modelos factoriales simples. Caso de Heywood. Métodos para obtener soluciones factoriales. Análisis de factores principales. El método de máxima verosimilitud*.

BIBLIOGRAFÍA: [10]Cap. 4, 6 y 7 (parte). [19] Cap. 9. [24] Cap. 12. La “biblia” del Análisis Factorial es [13].

9. ANÁLISIS FACTORIAL (II).

Indeterminación de las soluciones factoriales. El objetivo de estructura ortogonal simple. Soluciones factoriales derivadas mediante rotación ortogonal. Métodos quartimax y varimax. Comparación de ambos. Aplicaciones y ejemplos.

Clase práctica: Visualización de datos. Uso de R como paquete gráfico (II).

BIBLIOGRAFÍA: [10] Cap. 8. [13].

10. BILOTS. ANÁLISIS DE CORRESPONDENCIAS.

Introducción. Representación simultánea de variables e individuos. La descomposición en valores singulares (SVD). Biplots. Análisis de correspondencias. Ejemplos.

BIBLIOGRAFÍA: [10] Cap. 14. [24] Cap. 7.

11. ANALISIS DE TABLAS DE CONTINGENCIA

Tablas de contingencia n-variantes. Contrastes de ajuste (estadísticos X^2 y G^2 . Modelos logarítmico-lineales. Ejemplos de aplicación.

BIBLIOGRAFÍA: [2],[5].

12. REESCALADO MULTIDIMENSIONAL MÉTRICO Y NO MÉTRICO*

Naturaleza del problema. Matrices de distancias semidefinidas positivas y representación métrica. Ejemplos y aplicaciones.

BIBLIOGRAFÍA: [17], vol. II, cap. 10. Una buena monografía es [6].

13. ANALISIS CLUSTER.

Naturaleza del problema. Medidas de similaridad y distancias. Procedimientos de agrupamiento jerárquico. Procedimientos de agrupamiento basados en la optimización de un criterio. Problemas computacionales. Algoritmos mas usuales: single linkage, complete linkage y k-means. Procedimientos que requieren intervención humana: rostros de Chernoff, estrellas, etc.

BIBLIOGRAFÍA: [10] Cap. 18. Späth(1980). [14] (muy antiguo, pero útil). Algunos nuevos procedimientos se describen en [20].

14. ARBOLES DE REGRESIÓN Y CLASIFICACIÓN

Arboles binarios. La metodología CART: construcción de árboles, podado, validación cruzada. Arboles para clasificación. Arboles de regresión.

BIBLIOGRAFÍA: [7], primeros capítulos. [26], Cap. 7.

15. REDES NEURONALES*

Redes neuronales como extensión no lineal de algunos métodos de análisis multivariante. Perceptrones. Redes uni- y multicapa. Estimación de los coeficientes: propagación hacia atrás.

BIBLIOGRAFÍA: [4] (punto de vista estadístico), [16] (más cercana a la literatura sobre AI). [26], Cap. 5 introduce las redes *feed-forward*.

16. ANÁLISIS DE DATOS MASIVOS*

Data mining. Los nuevos retos. ¿Es esto análisis descriptivo estadístico? La perspectiva estadística.

BIBLIOGRAFÍA: [15] es un magnífico libro, con capítulos sobre casi todas las técnicas relevantes en *data mining* con un enfoque estadístico sólido. [32] trata específicamente sobre métodos gráficos de uso con datos masivos.

17. ANÁLISIS GRÁFICO*

Datos univariantes, bivariantes, n -variantes. Gráficos dinámicos. *Projection pursuit* y el *grand tour*. Gráficos ligados. Facilidades de visualización ofrecidas por R (III).

BIBLIOGRAFÍA: [9], [8], [30]

Bilbao, 11 de julio de 2007

Calendario previsto

2007

LUNES	MARTES	MIÉRCOLES
Sep 24 1 Repaso Algebra Lineal y Matricial <i>Entrega Tarea 1</i>	25 2 Normal multivariante. Primeras propiedades. Marginales y condicionadas.	26 3 Normal multivariante y asociadas. Distribuciones de Wishart y Wilks.
Oct 1 4 Normal multivariante y asociadas. La distribución T^2 de Hotelling. <i>Entrega Tarea 2</i>	2 5 Contrastes sobre el vector de medias en una población.	3 6 Contrastes sobre los vectores de medias en dos poblaciones normales. Vence Tarea 1.
8 7 Contrastes sobre la matriz de covarianzas en población normal. <i>Entrega Tarea 3</i>	9 8 Contrastes sobre las matrices de covarianzas en varias poblaciones normales. Vence Tarea 2.	10th 9 Análisis de correlación canónica.
15th 10 Análisis de correlación canónica. <i>Entrega Tarea 4</i>	16th 11 Regresión multivariante y MANOVA. Vence Tarea 3.	17th 12 Contrastes en MANOVA. <i>Entrega Tarea 5</i>
22 13 Análisis discriminante.	23 14 Análisis discriminante. Vence Tarea 4.	24 15 Análisis discriminante. <i>Entrega Tarea 6</i>
29 16 Análisis discriminante.	30 17 Componentes principales. Vence Tarea 5.	31 18 Componentes principales. <i>Entrega Tarea 7</i>
Nov 5 19 Componentes principales.	6 20 Análisis Factorial. Vence Tarea 6.	7 21 Análisis Factorial. <i>Entrega Tarea 8.</i>

LUNES	MARTES	MIÉRCOLES
12th 22 Reescalado multidimensional.	13th 23 Reescalado multidimensional.	14th 24 Biplot. Vence Tarea 7.
19th 25 Análisis de correspondencias. <i>Entrega Tarea 9.</i>	20 26 Análisis de correspondencias.	21 27 Tablas de contingencia multidimensionales. Vence Tarea 8.
26 28 Tablas de contingencia multidimensionales.	27 29 Tablas de contingencia multidimensionales. Vence Tarea 9.	28 30 Análisis cluster. <i>Entrega Tarea 10.</i>
Dic 3 31 Análisis cluster.	4 32 Análisis cluster.	5 33 Árboles de regresión y clasificación. <i>Entrega Tarea 11.</i>
10th 34 Árboles de regresión y clasificación.	11th 35 Árboles de regresión y clasificación. Vence Tarea 10.	12th 36 Árboles de regresión y clasificación.
17th 37 Redes neuronales. <i>Entrega Tarea 12.</i>	18th 38 Redes neuronales. Vence Tarea 11.	19th 39 Redes neuronales.

2008

LUNES	MARTES	MIÉRCOLES
Ene 8 36 Redes neuronales.	9 37 Redes neuronales. Vence Tarea 12.	10th 38 Análisis gráfico. Facilidades gráficas de R. Gráficos condicionados.

LUNES	MARTES	MIÉRCOLES
16th 39 Análisis gráfico. <i>Projection pursuit.</i>	17th 40 Análisis gráfico. Gráficos dinámicos. El <i>grand tour.</i>	18th 41 Análisis gráfico. Gráficos ligados.

Bibliografía

- [1] K.M. Abadir and J.R. Magnus. *Matrix Algebra*. Cambridge Univ. Press, 2005.
- [2] A. Agresti. *Categorical Data Analysis*. Wiley, 1990.
- [3] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 1984.
- [4] C.M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1996.
- [5] Y.M.M. Bishop, S.E. Fienberg, and P.W. Holland. *Discrete Multivariate Analysis. Theory and Practice*. MIT Press, Cambridge, Mass., 1975.
- [6] I. Borg and P. Groenen. *Modern Multidimensional Scaling. Theory and Applications*. Springer-Verlag, New York, 1997.
- [7] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, California, 1984.
- [8] J.M. Chambers and T.J. Hastie. *Statistical Models in S*. Wadsworth & Brooks/Cole, Pacific Grove, Ca., 1992.
- [9] W.S. Cleveland. *Visualizing Data*. Hobart Press, NJ, 1993.
- [10] C.M. Cuadras. *Métodos de Análisis Multivariante*. Eunibar, Barcelona, 1981.
- [11] W.R. Dillon and M. Goldstein. *Multivariate Analysis: Methods and Applications*. Wiley, New York, 1984.
- [12] J.J. Hair, R.E. Anderson, R.L. Tatham, and W.C. Black. *Multivariate Data Analysis*. Maxwell MacMillan, New York, 1992.
- [13] H.H. Harman. *Análisis Factorial Moderno*. Saltés, 1980.
- [14] J.A. Hartigan. *Clustering Algorithms*. Wiley, New York, 1975.
- [15] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer-Verlag, 2001. Signatura: 519.237.8 HAS.

- [16] S. Haykin. *Neural Networks. A comprehensive Foundation*. Prentice Hall, second edition, 1998.
- [17] J.D. Jobson. *Applied Multivariate Data Analysis, vol. II*. Springer Verlag, New York, 1991. Signatura: 519.237 JOB.
- [18] D.E. Johnson. *Métodos multivariados aplicados al análisis de datos*. Thomson, 1998.
- [19] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 1992.
- [20] L. Kaufman and P.J. Rousseeuw. *Finding groups in data : an introduction to cluster analysis*. Wiley, 1990. Signatura: 519.237.8 KAU.
- [21] W.J. Krzanowski. *Principles of Multivariate Analysis: A User's Perspective*. Oxford, 1988. Signatura: 519.23 KRZ.
- [22] A. Kshirsagar. *Multivariate Analysis*. Marcel Dekker, 1978.
- [23] P.A. Lachenbruch. *Discriminant Analysis*. Hafner Press, New York, 1975.
- [24] D. Peña. *Análisis de Datos Multivariantes*. McGraw-Hill, 2002.
- [25] A.C. Rencher. *Methods of Multivariate Analysis*. Wiley, 1995.
- [26] B.D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996. 519.237.8 RIP.
- [27] S.R. Searle. *Matrix Algebra useful for Statistics*. Wiley, 1982.
- [28] G.A.F. Seber. *Multivariate Observations*. Wiley, New York, 1984.
- [29] B. Thompson. *Canonical Correlation Analysis*. SAGE, 1984.
- [30] E.R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 1983. Signatura: 519.255.
- [31] F. Tusell. Análisis multivariante. Notas de clase, Octubre 2003.
- [32] A. Unwin, M. Theus, and H. Hofmann. *Graphics of Large Datasets: Visualizing a Million (Statistics and Computing)*. Springer, 2006. Signatura: 519.255 UNW.