

# Stat 9370: Analysis of Multivariate Data

Stanislav Kolenikov\*

January 20, 2006

## 1 Basics

**Class time:** Winter 2006, MFW 1-1:50pm, Middlebush 308

**Instructor:** Stanislav (Stas) Kolenikov, kolenikovs@missouri.edu, 307A Middlebush. Office hours: Mon 2-3 pm, and by appointment.

**Website:** Blackboard (<http://courses.missouri.edu>)

**Information:** This course covers the main ideas in multivariate statistical analysis at the graduate level. Most of the theory is based on the multivariate normal distribution, and samples from that distribution. The main multivariate results and tests will be illustrated, and a number of application oriented techniques overviewed.

**Prerequisites:** The students will need to have received credit for STAT 4760/7760 (statistical inference) and MATH 4140/7140 (matrix theory) to be enrolled in this class. In other words, you will have understanding of statistical inference concepts (we will be deriving all sorts of tests about means, variances, covariances, and covariance matrices), and understanding of matrix operations (multiplication, inversion, eigenvalue problems, etc.).

**Tests and grade structure:** There will be a midterm exam in class on March 17, Friday. The final exam is on Wednesday, May 10, 1:00 – 3:00pm, in the regular classroom. There will be about five home assignments during the semester. Students will submit a term paper and make a presentation in class on a topic in multivariate statistics not covered in the course. (If a student does not have independent research interests in the area, a paper

---

\*Department of Statistics, 146 Middlebush Hall, University of Missouri, Columbia, MO 65211-6100. Email: kolenikovs@missouri.edu.

or a topic can be recommended by the instructor.) The total grade will be determined as follows: home works = 25%, midterm = 25%, student term paper/presentation = 20%, final = 30%.

**Other info:** Academic integrity is fundamental to the activities and principles of a university. All members of the academic community must be confident that each person's work has been responsibly and honorably acquired, developed, and presented. Any effort to gain an advantage not given to all students is dishonest whether or not the effort is successful. The academic community regards breaches of the academic integrity rules as extremely serious matters. Sanctions for such a breach may include academic sanctions from the instructor, including failing the course for any violation, to disciplinary sanctions ranging from probation to expulsion. When in doubt about plagiarism, paraphrasing, quoting, collaboration, or any other form of cheating, consult the course instructor.

If you have special needs as addressed by the Americans with Disabilities Act (ADA) and need assistance, please notify the Office of Disability Services, A038 Brady Commons, 882-4696 or course instructor immediately. Reasonable efforts will be made to accommodate your special needs.

## 2 Readings

The required textbook for the class is Johnson & Wichern (2002) (further referred to as JW). It achieves about the right balance between the topics we want to cover, and the level of technicality, although the quality of presentation is rather patchy.

There are numerous other books that can be recommended to supplement the main book. The standard statistical references on multivariate statistics are Anderson (2003) and Mardia, Kent & Bibby (1980). They are more technical, somewhat less applied, and do not cover structural equation models. Another good applied book that unfortunately does not cover structural equation models, either, but is more graphics intensive than JW is Rencher (2002), with a technical supplement given by Rencher (1998). Any of those books will work as an alternative to JW; if you already have one, and you can find supplementary readings for structural equation models, then your need in JW is mostly limited to problem sets. More theoretical readings are Muirhead (2005) and Fang & Zhang (1990). Gifi (1990) stands somewhat alone as the only statistics originated book on the analysis of discrete multivariate

data.

Further down to applications in social sciences is Bartholomew, Steele, Moustaki & Galbraith (2002). There is a number of specialized books on the methods we will be covering; see Jolliffe (2002) for principal component analysis, Everitt, Landau & Leese (2001) for cluster analysis, Bollen (1989), Bollen & Long (1993), Kaplan (2000) on structural equation models. Hastie, Tibshirani & Friedman (2001) is an invaluable source for model-free statistical learning, including clustering and discrimination. The newest methods, such as functional data analysis (Ramsey & Silverman 1997), robust diagnostics (Atkinson, Riani & Cerioli 2003), independent component analysis (Hyvärinen, Karhunen & Oja 2001), interface with multilevel models and mixed response models (Skrondal & Rabe-Hesketh 2004), have not yet it made to the standard textbooks.

Books on matrix theory include Meyer (2001), Bhatia (1996), Horn & Johnson (1990), Harville (1997), Magnus & Neudecker (1999). The latter book covers also important techniques of matrix calculus that we may have to use once or twice during the course.

Additional readings will be based on articles; see Mardia (1970), Davis (1977), Olsson (1979), Browne (1984), Hastie & Stuetzle (1989), Azzalini & Valle (1996), Johnstone (2001). I will be adding items to this list and suggest other readings.

### **3 Software**

A number of exercises will require work on computer. The students can use any of the available software packages (SAS, R, SPSS, ...) to do their homeworks and work on term papers. I shall mostly be using Stata for my demonstrations.

## 4 Topics and coverage

Week 1: Jan 17–20	Introduction and necessary matrix theory results
Week 2: Jan 23–27	Multivariate samples and display of multivariate data
Week 3: Jan 30–Feb 3	The multivariate normal distribution; distributions of $\bar{X}$ and $S$ ; assessing multivariate normality
Week 4: Feb 6–9	Inference about the multivariate mean vector; $T^2$ , MANOVA
Week 5: Feb 13–17	Inference about the multivariate mean vector; $T^2$ , MANOVA
Week 6: Feb 20–24	Distribution of the sample covariance; tests on the covariance matrix; sphericity.
Feb 21 is the last day to drop	
Week 7: Feb 27–Mar 3	Principal component analysis
Week 8: Mar 6–10	Distributions of the characteristic roots and vectors; non-normal data
Week 9: Mar 13–17	Factor analysis: EFA, CFA.
Midterm exam: March 17	
Week 10: Mar 19–24	Latent variable models
Week 11: Apr 3–7	Canonical correlation analysis
Week 12: Apr 10–14	Discriminant analysis
Week 13: Apr 17–21	Cluster analysis: hierarchical and non-hierarchical procedures; model-based clustering
Week 14: Apr 24–28	MDS, MCA and other graphical techniques
Week 15: May 1–5	Leftovers, student paper presentations

## References

- Anderson, T. W. (2003), *An Introduction to Multivariate Statistical Analysis*, Wiley Series in Probability and Statistics, 3rd edn, John Wiley and Sons.
- Atkinson, A. C., Riani, M. & Cerioli, A. (2003), *Exploring Multivariate Data with Forward Search*, Springer-Verlag Inc, New York.
- Azzalini, A. & Valle, A. (1996), ‘The multivariate skew-normal distribution’, *Biometrika* **83**(4), 715–726. doi:10.1093/biomet/83.4.715.

- Bartholomew, D. J., Steele, F., Moustaki, I. & Galbraith, J. I. (2002), *The Analysis and Interpretation of Multivariate Data for Social Scientists*, Texts in Statistical Science, Chapman and Hall/CRC, Boca Raton, FL.
- Bhatia, R. (1996), *Matrix Analysis*, Graduate Texts in Mathematics, Springer, New York.
- Bollen, K. (1989), *Structural Equations with Latent Variables*, Wiley, New York.
- Bollen, K. A. & Long, J. S., eds (1993), *Testing structural equation models*, Sage, Thousand Oaks, CA.
- Browne, M. W. (1984), ‘Asymptotically distribution-free methods for the analysis of the covariance structures’, *British Journal of Mathematical and Statistical Psychology* **37**, 62–83.
- Davis, A. W. (1977), ‘Asymptotic theory for principal component analysis: Non-normal case’, *Australian Journal of Statistics* **19**, 206–212.
- Everitt, B. S., Landau, S. & Leese, M. (2001), *Cluster Analysis*, 4th edn, Arnold Publishers.
- Fang, K. T. & Zhang, Y. T. (1990), *Generalized Multivariate Analysis*, Springer, New York.
- Gifi, A. (1990), *Nonlinear multivariate analysis*, John Wiley & Sons.
- Harville, D. A. (1997), *Matrix Algebra from a Statistician’s Perspective*, Springer, New York.
- Hastie, T. & Stuetzle, W. (1989), ‘Principal curves’, *Journal of the American Statistical Association* **84**(406), 502–516.
- Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning*, Springer-Verlag, New York.
- Horn, R. A. & Johnson, C. R. (1990), *Matrix Analysis*, reprint edn, Cambridge University Press, New York.
- Hyvärinen, A., Karhunen, J. & Oja, E. (2001), *Independent Component Analysis*, Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control, Wiley-Interscience, New York.
- Johnson, R. A. & Wichern, D. W. (2002), *Applied Multivariate Statistical Analysis*, 5th edn, Prentice Hall, Englewood Cliffs, NJ.

- Johnstone, I. M. (2001), ‘On the distribution of the largest eigenvalue in principal component analysis’, *Annals of Statistics* **29**, 295–327.
- Jolliffe, I. (2002), *Principal Component Analysis*, 2nd edn, Springer, New York.
- Kaplan, D. (2000), *Structural Equation Modeling: Foundations and Extensions*, number 10 in ‘Advanced Quantitative Techniques in Social Sciences’, SAGE, Thousand Oaks, CA.
- Magnus, J. R. & Neudecker, H. (1999), *Matrix differential calculus with applications in statistics and econometrics*, 2nd edn, John Wiley & Sons.
- Mardia, K. V. (1970), ‘Measures of multivariate skewness and kurtosis with applications’, *Biometrika* **57**, 519–530.
- Mardia, K. V., Kent, J. T. & Bibby, J. M. (1980), *Multivariate Analysis*, Academic Press, London.
- Meyer, C. D. (2001), *Matrix Analysis and Applied Linear Algebra*, SIAM.
- Muirhead, R. J. (2005), *Aspects of Multivariate Statistical Theory*, Wiley Series in Probability and Statistics, 2nd revised edn, Wiley-Interscience, New York.
- Olsson, U. (1979), ‘Maximum likelihood estimation of the polychoric correlation’, *Psychometrika* **44**, 443–460.
- Ramsey, J. O. & Silverman, B. W. (1997), *Functional Data Analysis*, Springer Series in Statistics, Springer, New York.
- Rencher, A. C. (1998), *Multivariate Statistical Inference and Applications*, Wiley, New York.
- Rencher, A. C. (2002), *Methods of Multivariate Analysis*, Wiley, New York.
- Skrondal, A. & Rabe-Hesketh, S. (2004), *Generalized Latent Variable Modeling*, Chapman and Hall/CRC, Boca Raton, Florida.