Syllabus for Advanced Multivariate Analysis
(26:630:672)

Rutgers Business School Newark• PHD in Management Program

Professor J. Douglas Carroll

Spring 2009

- PHONE: (973) 353-5814 (Office)
- E-MAIL: dcarroll@rci.rutgers.edu
- OFFICE HOURS: 4-5PM Tuesdays or by appointment. E-mail communication is strongly encouraged.
- OFFICE: 125 Management Education Center
- Class meets: Tuesdays 5.30pm-8.20pm
- Text: Analyzing Multivariate Data by James M. Lattin, J. Douglas Carroll, and Paul E. Green (Belmont, CA: Duxbury Press).
- Optional Supplementary Text: Carroll, J. D., & Green, P. E. (1997). *Mathematical Tools for Applied Multivariate Analysis*. San Diego, CA: Academic Press (with contributions by A. D. Chaturvedi).

Teaching Assistant: Shuojia(Nancy) Guo
Office Hours: By appointment
E-mail: nancyguo@pegasus.rutgers.edu

Classes 1-2

Chapter 9: Canonical Correlation Analysis (CCA).

Canonical Correlation Analysis (CCA) will be presented as a very general technique for interrelating two (or more, in some generalizations, see Class 9 description) matrices of variables defined on the same objects by finding linear combinations of each having maximum correlation. These separate linear combinations maximizing the correlation are called "canonical variates." Statistical theory of CCA and applications will be emphasized. A measure called "Wilks's Λ" will be discussed as a statistic for testing the statistical significance of the relationship between the two sets of variables. Also, Stewart and Love's "redundancy" index will be discussed as a measure of the overall amount of variance accounted for by one variable set in the other.

Class 3

Chapter 12 Multiple Discriminant Analysis.

Multiple Discriminant Analysis (MDA) is used when it's desired to find linear combinations of a set of variables that best discriminate among two or more groups of objects, people or other entities. It will be shown that two group discriminant analysis can be formulated as a special case of multiple linear regression analysis (with a "dummy" dependent variable comprising a binary encoding of membership in the two groups.) Analogously, MDA with three or more groups can be formulated as a special case of CCA (with n-1 "dummy" variables redundantly encoding the n distinct groups in this case), as will be demonstrated.

Class 4

Chapter 6 Confirmatory Factor Analysis.

In Confirmatory Factor Analysis (CFA), as contrasted with Exploratory Factor Analysis (EFA), specific hypotheses are formulated regarding the structure of the factor solution. These hypotheses are explicitly tested in a confirmatory manner. This is usually done in the context of maximum likelihood fitting of the factor models, although other criteria of fit can be used. As in other cases of hypothesis testing, the best model fit with constraints implied by the stated hypothesis is compared with a full, unconstrained, model, and chi square measures of the difference in fit of the two models are used to decide whether the null hypothesis (corresponding to the hypothesized model to be confirmed or disconfirmed) can be rejected in favor of the alternative hypothesis (corresponding to the full unconstrained, model.) If the null hypothesis cannot be rejected this is taken as evidence *tending* to confirm the hypothesized model.

Classes 5-6

Chapter 10 Structural Equation Models with Latent Variables.

This chapter deals with a general approach to analysis of dependence allowing the user to account for measurement errors in observed variables while studying the dependence relationships among the latent variables. Structural equation models with latent variables can be shown to be more general than models based on canonical correlation, which can be shown to be a special case.

NOTE: Midterm take home exam will be distributed in this class. Exam should be returned by beginning of Class 9.

Classes 7-8

Chapter 13 Logit Choice Models.

A logistic, or logit, choice model assumes the probabilities of binary (or n-ary) choices by an individual are logistically distributed. In the binary case the logistic distribution has the form:

$$P_1 = \frac{e(v_1)}{e(v_1) + e(v_2)}$$

while, by symmetry

$$P_2 = \frac{e(v_2)}{e(v_1) + e(v_2)}$$

where v_1 and v_2 represent the value or *utility* of alternatives 1 and 2 respectively. In the n-ary case (in which a choice is to be made among *n* alternatives rather than only two as in the binary case) the distribution has a similar form for each of the *n* alternative choices, except that the denominators comprise the sum of the terms

$e(v_i)$for $\mathrm{i} = 1,\ 2,\cdots \mathrm{n}$. The utilities, v_i are generally assumed to be linear combinations of a set of measured variables of which these utilities are assumed to be functions.

Another distribution often assumed for these choice probabilities is the probit, in which case the choice probabilities for the binary case are defined as the normal CDF (Cumulative Distribution Function) applied to the differences in the two utilities. (In the n-ary case the probit is defined as the distribution comprising the normal CDF applied to the difference between the utility for alternative *i* and the sum of utilities for the remaining *n – 1* alternatives.) While the probit has many desirable properties, it does not yield a closed functional form for the choice probabilities whereas the logistic does. In part for this reason (as well as based on an elegant theoretical justification of the logistic derived by R. D Luce in a famous 1959 book) the logistic (which in practice is virtually indistinguishable from the probit anyway) is usually preferred.

The logistic (or logit) distribution has one property, however, that can sometimes be troublesome. This is the property of "Independence of Irrelevant Alternatives" (IIA). IIA states that adding an additional "irrelevant" alternative to a set of *n* alternatives currently available leaves the *ratios* of the choice probabilities of pairs of those initial *n* alternatives unchanged. As can easily be demonstrated, there are certain choice situations in which this condition (IIA) is clearly violated. This apparent anomaly can be corrected, however, by generalizing the logistic to the *nested* logistic model, in which the *n* alternatives are nested within a hierarchical tree structure, as will be described and discussed.

Class 9

Chapter 13 Generalized Canonical Correlation Analysis.

There are a number of generalizations of (two set) canonical correlation analyses to three or more sets. A number of these will be discussed and applications of them described. A set of papers (Carroll, J. D., 1968; Carroll, J. D. 1973; Horst, P., 1961a, b; Kettenring, J. R., 1972; McKeon, J. J., 1965 on these topics will be made available to the class dealing with this topic.

Classes 10-11: Overlapping clustering, including generalizations to the "three-way" (or individual differences) case.

A number of models and methods for overlapping clustering will be discussed. These include:

- The ADCLUS (Additive CLUStering) model (Shepard and Arabie, 1979)
- The MAPCLUS (Mathematical Programming CLUStering) method for fitting the ADCLUS model (Arabie and Carroll, 1980)
- The INDCLUS (Individual Differences CLUStering) model and method (Carroll and Arabie, 1983), which generalizes the ADCLUS/MAPCLUS approach to the three-way or individual differences case, in which data are provided for two or more subjects or other sources of data, and a joint overlapping clustering is sought, with differential weights for these clusters for each subject/source, indicating the effect of each overlapping cluster on the proximity (similarity or dissimilarity) data of that subject/source, very much analogous to the effect of subject/source weights in the INDSCAL model for individual differences multidimensional scaling.
- The SINDCLUS ("Speedy" INDCLUS) method (of Chaturvedi and Carroll, 1994), which enables fitting both the three-way INDCLUS model and the two-way ADCLUS model via an algorithm closely analogous to that used in fitting the INDSCAL

model. This algorithm is considerably faster and more efficient than the earlier INDCLUS method, and so is generally to be preferred for fitting the INDCLUS model

- Overlapping K-centroids clustering (Chaturvedi, Carroll, Green and Rotondo, 1997). This approach generalizes such well-known clustering techniques as K-means and K-medians to the overlapping case. It also allows an entire class of overlapping clustering models based on different values of p in the L_p metric. (Overlapping K-mean corresponds to p = 2 while overlapping K-medians corresponds to $p = 1$.) Two extreme cases of interest include a method called “K-midranges clustering” (corresponding to $p = \infty$) and K-modes clustering ($p = 0$).

Classes 12- 13: General three-way and multiway models and methods of data analysis.

A general three-way and/or N-way (N > 3) model originally devised by Carroll and Chang (1970) [as a generalization of the three-way INDSCAL model and method developed by these authors in 1970] called CANDECOMP (for CANonical DEComposition of N-way tables) will be discussed and a number of applications described. See Carroll and Pruzansky for a general overview of CANDECOMP (Carroll and Pruzansky, 1984) model. Especially important applications include four-way or higher-way generalizations of the INDSCAL model and method and an approach to three-way or multiway factor or components analysis, closely related to the method independently devised by Harshman and Lundy (1984a,b) and others called PARAFAC (for PARallel FACtor analysis.

A modified version of CANDECOMP, called CANDELINC (CANonical Decomposition with LINear Constraints) [Carroll, Pruzansky and Kruskal (1980)] will also be covered, and applications described. CANDELINC allows fitting the CANDECOMP model with specified linear constraints, so that dimensions emerging in a CANDECOMP analysis might be constrained, for example, to be linear combinations of a specific set of exogenous variables. (This can greatly enhance interpretability of CANDECOMP dimensions, among other advantages.) Among applications of CANDELINC to be discussed are fitting linearly constrained versions of the MDPREF vector model for individual differences analysis of preferential choice data, fitting linearly constrained versions of the INDSCAL models, and an approach allowing fitting the INDSCAL model to much larger sets of proximity data than might otherwise be possible.

Another development consists of a family of *hybrid* three-way and multiway models (Carroll and Chaturvedi, 1995, 1998) given the generic name CANDCLUS (for CANonical Decomposition CLUStering) which will be discussed, including the general model and a number of special cases. CANDCLUS combines the CANDECOMP model and method described earlier with a general three- or higher-way overlapping clustering approach (which can be viewed as a higher-way generalization of the SINDCLUS approach described earlier.

Every three- or higher-way model discussed earlier can be viewed as a special case of CANDCLUS (e.g., CANDECOMP, INDSCAL, SINDCLUS and various special cases of these), while a number of other models and methods have been devised which are additional special cases of CANDCLUS. One of the more interesting of these is a model and method called CLUSCALE (Chaturvedi and Carroll, submitted for publication, 2004) which comprises a three-way (or potentially higher-way) hybrid model combining aspects of INDSCAL with aspects of SINDCLUS, so it results in a hybrid representation combining continuous dimensional structure with discrete cluster structure. CLUSCALE will be illustrated with an application to some data on perceptions of cars.

Finally, other hybrid models, combining continuous dimensional structure with discrete structure will be discussed. One of these is an approach combining a Euclidean spatial model with one or more tree structures, the latter comprising discrete geometric models. See Carroll and Pruzansky (1975). While the approach to be discussed is limited to the two-way case, it would be straightforward to extend this to the three- or higher-way case.

Class 14 Parametric Mapping (PARAMAP).

Parametric Mapping (abbreviated as "PARAMAP") was first proposed by Carroll in a 1966 paper co-authored with R. N. Shepard ("Parametric representation of nonlinear data structures") [Shepard and Carroll, 1966] published in the first of several volumes edited by P. R. Krishnaiah simply entitled, "Multivariate Analysis." In this paper Shepard proposed a method he called "locally monotone analysis of proximities" (a version of nonmetric MDS based entirely on "small" distances.) This approach worked quite well for nonlinear data structures that were not too highly curved, but did not work with a highly nonlinear manifold constituting a closed surface, such as data points on a complete circle or sphere. Carroll then proposed Parametric Mapping (PARAMAP) in the second part of the paper. PARAMAP has been shown to be able to deal with such closed manifolds as the complete circle, sphere, torus, (defining a "flat" manifold whose linear dimensionality equals the intrinsic or topological dimensionality of the closed nonlinear manifold on which the data points provided as input to the procedure are embedded). In the case of points on a circle embedded in a two-dimensional Euclidean space PARAMAP produces as output a one dimensional linear continuum in which the local structure of the points on the circle is very well preserved, with the notable exception of one point on the circle where the circle must be cut in order to open it into this one dimensional linear space topologically equivalent to the nonlinear manifold (the circle)—everywhere *except* at that point. Analogously, in the case of the sphere, a map very similar to one of the standard maps of the earth's surface worked out by cartographers (specifically, an azimuthal equidistant projection) is obtained. In this case the functions relating data points on the sphere to those on the flat two-dimensional map obtained by PARAMAP is continuous *almost* everywhere—the notable exception corresponding to the point on the sphere where the sphere must be "punctured" in order to map the entire sphere onto the "flat" map whose dimensionality (two) corresponds to that of the intrinsic topological dimensionality of the sphere, which is locally *flat* (or *two-dimensional*). The functions relating this flat two-dimensional map to the points (analogous to cities on the earth) on the surface of the sphere will be continuous *except* for a *severe* discontinuity occurring at the point where the sphere had to be "punctured" in order to produce the appropriate reduced dimensional map obtained by PARAMAP. In maps of the earth's surface the “puncture” usually is positioned in a region (say in the Artic ocean or middle of the Atlantic or Pacific) that is uninhabited so that there are no cities or obvious geographic features, so that the extreme distortion is not nearly so obvious as if it occurred, say, in the middle of the U.S.A., Europe or Asia!

PARAMAP is based on optimizing a measure (called “Kappa”), developed by Carroll, of “continuity” or smoothness of the mapping from one space (or set of variables) to another. The definition of this measure and its justification are described in considerable detail in the 1966 paper discussed earlier. Once the dimensionality of the representation or “map” to be determined by PARAMAP is determined by the user, a gradient based optimization technique is used to seek the configuration of points on that map optimizing Kappa.

The PARAMAP procedure developed in 1966 worked well with relatively small sets of points that were regularly spaced and errorless (so the points on the sphere were precisely located on the sphere, without deviating at all from that surface, or deviating from perfectly regular spacing). Unfortunately, once this regular spacing and errorlessness condition are violated, the algorithm tended to break down badly, due to a serious “local minimum” problem.

Instead of obtaining the global optimum a series of merely local optima were obtained—none of which corresponded to (or even approximated) the correct solution corresponding to the global optimum.

Ulas Akkucuk, a Ph.D. student working with Carroll, who recently got his Ph.D. conducting a dissertation based on research with Carroll on this “PARAMAP” problem, has largely solved this problem. Akkucuk solved the “PARAMAP” problem in large part by taking advantage of the immensely greater speed and power of modern computers as compared with those of the late 1960’s. Akkucuk has improved the algorithm in a number of important ways, but, most significantly, this greater speed and power enables running the optimization algorithm from vastly more different starting points, thus obtaining a very large number of different locally optimal solutions. The best of these local optima can be plausibly assumed to be, or at least to be very close to, the desired global optimum. After obtaining this estimate of the global optimum, Akkucuk ran a large number of additional iterations of the gradient-based optimization (minimization) algorithm, using that current estimate of the globally optimal solution as a starting point, so as to push it to as complete convergence as possible. This combined numerical strategy seems to work quite well.

Akkucuk also devised a very powerful measure of “preservation of local structure” that measures the extent the local structure in the original data is preserved in the lower dimensional PARAMAP (or ISOMAP) representations.

Akkucuk also worked out procedures for transforming the obtained solutions resulting from this fitting procedure to a form that best matches the “true” configuration (known, since these studies were all “Monte Carlo” studies in which the “true” solution was given). These matching procedures involved linear transformations (rotation, translation and uniform dilation) as well as nonlinear ones based on the nature of the data structures involved. For example, in the case of the points on a sphere, since the sphere can be punctured at any point to “open” it up into a flat two-dimensional plane, a certain class of nonlinear transformations are applied that depend on the exact point at which this “puncture” in fact occurred.

A measure of agreement between the true and obtained configuration is thereby obtained. These measures were exceedingly good—even for solutions that had been seriously perturbed by the violation of the regular spacing condition and by addition of significant amounts of random error. This was confirmed statistically by use of a randomization procedure.

This dissertation work by Akkucuk (done under the general supervision of Carroll, who served as his dissertation advisor) involved two types of nonlinear manifolds. One was the sphere already discussed extensively, while the other was a very regular torus embedded in four dimensions. This four dimensional torus, or donut-like manifold, had the interesting regularity property that the diameter of the hole of the donut, that of the donut itself, and that of a cross-section of the donut taken by cutting it at any point and opening it out into a tube (or approximate cylinder) are *all* equal. This, of course, is physically impossible in the case of a three-dimensional torus, but is quite easily attained with the kind of four dimensional torus

Akkucuk and Carroll were dealing with. PARAMAP enabled remarkably good recoveries of both these types of configurations, even with considerable perturbations of the type discussed earlier.

Akkucuk also compared PARAMAP with a procedure called ISOMAP developed by Tenenbaum and colleagues. ISOMAP approximates geodetic distances among points on a manifold via graph theoretic methods, which will not be described further here. These approximate geodetic distances are then analyzed by "classical MDS" methodology, resulting in a lower dimensional spatial representation that's topologically equivalent to the original manifold—but with an important caveat. The caveat is that ISOMAP works only for manifolds that are *not* closed (as the circle, sphere, and torus configurations described above are) and, in fact, will break down if the manifold gets "close" to being closed. For example, it will work on points on a segment of a sphere comprising about three-quarters of the sphere (so that roughly the top one quarter of the sphere has been cut off, as one might remove the top quarter of, say an orange) but if more of the sphere is present the procedure breaks down badly. In fact, using the same measure of agreement described above, ISOMAP generally recovers the "true" configuration much less adequately than does PARAMAP.

One additional very important feature was added by Akkucuk. Since the earlier version of PARAMAP was seriously restricted by the number of data points it could handle, a procedure was added allowing analysis of much larger sets of points. This is accomplished by simply using a much larger set of points (say 1,000) but, in a random or systematic way, sampling a small set (on the order of 50 or 60) from this larger set, (called "landmark points")for which a PARAMAP solution is obtained. Then, fixing the smaller set of points, the remaining large set of "holdout" points excluded from the original set are mapped in, using a conditional version of PARAMAP in which only the excluded points are fit (the smaller set of "landmark" points remaining fixed). There are many details of this not explicated here, but suffice it to say that this composite procedure works quite well, and the PARAMAP solutions thus obtained conform quite well with the "true" configurations as measured by the index of agreement discussed earlier; thus enabling fitting of much larger data sets than would otherwise be possible!

This appears to be a very promising approach for extending PARAMAP and numerous other data analysis and/or nonlinear mapping techniques to *much* larger data sets.

Class 15 Final Exam

Grading: Grades will be based on Midterm Exam (40%), Final Exam (40%) and homework and class participation (20%).

Reading List for Advanced Multivariate Analysis Course

Arabie, P., & Carroll, J. D. (1980). MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika, 45,* 211–235.

Carroll, J. D. (1968). Generalization of canonical correlation analysis to three or more sets of variables. In *Proceedings of the 76th Convention of the American Psychological Association, 3,* 227-228.

Carroll, J. D. (1973). Models and algorithms for multidimensional scaling, conjoint measurement and related techniques (Appendix B). In P. E. Green & Y. Wind, with contribution by J. D. Carroll, *Multiattribute Decisions in Marketing* (pp. 299–387). Hinsdale, IL: Dryden Press. [Partially reprinted (1989) in P. E. Green, F. J. Carmone & S. M. Smith, *Multidimensional Scaling: Concepts and Applications* (pp. 332–337). Newton, MA: Allyn and Bacon.]

Carroll, J. D., & Arabie, P. (1983). INDCLUS: An individual differences generalization of the ADCLUS model and the MAPCLUS algorithm. *Psychometrika, 48,* 157–169. [Reprinted (1984) in H. G. Law, W. Snyder, J. Hattie & R. P. McDonald (Eds.), *Research Methods for Multimode Data Analysis* (pp. 372–402). New York: Praeger.]

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika, 35,* 283–319. [Reprinted (1984) in P. Davies & A. P. M. Coxon (Eds.), *Key Texts in Multidimensional Scaling* (pp. 229–252). Portsmouth, NH: Heinemann.]

Carroll, J. D., & Chaturvedi, A. (1995). A general approach to clustering and multidimensional scaling of two-way, three-way, or higher-way data. In R. D. Luce, M. D'Zmura, D. D. Hoffman, G. Iverson & A. K. Romney (Eds.), *Geometric Representations of Perceptual Phenomena* (pp. 295–318). Mahwah, NJ: Erlbaum.

Carroll, J. D., & Chaturvedi, A. (1998). Fitting the CANDCLUS/MUMCLUS models with partitioning and other constraints. In C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. -H. Bock & Y. Baba (Eds.), *Data Science, Classification, and Related Methods* (pp. 496–505). Tokyo: Springer-Verlag.

Carroll, J. D., & Pruzansky, S. (1975). Fitting of hierarchical tree structure (HTS) models, mixtures of HTS models and hybrid models, via mathematical programming and alternating least squares. *Proceedings of the U.S.-Japan Seminar on Multidimensional Scaling and Related Techniques*, 9–19.

Carroll, J. D., & Pruzansky, S. (1984). The CANDECOMP-CANDELINC family of models and methods for multidimensional data analysis. In H. G. Law, C. W. Snyder, J. A. Hattie & R. P. McDonald (Eds.), *Research Methods for Multimode Data Analysis* (pp. 372–402). New York: Praeger.

Carroll, J. D., Pruzansky, S., & Kruskal, J.B. (1980). CANDELINC: A general approach to multidimensional analysis of many-way arrays with linear constraints on parameters. *Psychometrika*, *45*, 3–24.

Chaturvedi, A., & Carroll, J. D. (1994). An alternating combinatorial optimization approach to fitting the INDCLUS and generalized INDCLUS models. *Journal of Classification*, *11*, 155-170.

Chaturvedi, A., Carroll, J. D., Green, P. E., & Rotondo, J. A. (1997). A feature based approach to market segmentation via overlapping K-centroids clustering. *Journal of Marketing Research*, 34, 370–377

Chaturvedi, A., & Carroll, J. D. (2004). CLUSCALE (CLUstering and multidimensional SCAL[E]ing): A three-way hybrid model incorporating clustering and multidimensional scaling structure. Manuscript submitted for publication.

Harshman, R. A., & Lundy, M. E. (1984a). The PARAFAC model for three-way factor analysis and multidimensional scaling. In H. G. Law, C. W. Snyder, Jr., J. Hattie, & R. P. McDonald (Eds.), *Research methods for multimode data analysis* (pp. 122-215). New York: Praeger.

Harshman, R. A., & Lundy, M. E. (1984b). Data preprocessing and the extended PARAFAC model. In H. G. Law, C. W. Snyder, Jr., J. Hattie, & R. P. McDonald (Eds.), *Research methods for multimode data analysis* (pp.216-284). New York: Praeger.

Horst, P. (1961a). Relations among m sets of measures. *Psychometrika*, *26*, 129-149.

Horst, P. (1961b). Generalized canonical correlations and their application to experimental data. *Journal of Clinical Psychology*, *17*, 331-347.

Kettenring, J. R. (1972). Canonical analysis of several sets of variables. *Biometrika*, *58*, 443-451.

Luce, R. (1959). Individual choice behavior: A theoretical analysis. New York: J. Wiley and Sons.

McKeon, J.J. (1965). Canonical analysis: Some relations between canonical correlation, factor analysis, discriminant function analysis, and scaling theory. *Psychometric Monographs*, No. 13.

Shepard, R. N., & Carroll, J. D. (1966). Parametric representation of nonlinear data structures. In P. R. Krishnaiah (Ed.), *Multivariate Analysis* (pp. 561–592). New York: Academic Press.

Shepard, R. N. & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review, 86, 87-123.*