223

DATA ANALYSIS AND INFORMATICS, IV
E. Diday et al. (editors)

THE PERMUTATIONAL LIMIT DISTRIBUTION OF GENERALIZED CANONICAL CORRELATIONS

Jan de Leeuw
Eeke van der Burg
Department of Data Theory FSW/RUL
Middelstegracht 4
2312 TW Leiden, The Netherlands

ABSTRACT

We study the permutational limit distribution of goodness-of-fit statistics computed in various forms of generalized canonical correlation analysis. In simple cases of canonical analysis the exact distribution can be computed. For somewhat more complicated forms approximations have been tabulated by Krishnaiah and others. For the most general forms of canonical analysis we must use resampling methods to generate the permutation distribution. Our approach is illustrated by examples of varying degree of complexity. For small examples we can actually study the quality of our approximations. For more complicated examples this is not possible, but the approximate permutation distributions themselves are a very valuable data analytical tool.

INTRODUCTION

Canonical correlation analysis is a familiar data analysis technique. It is closely related to other well known techniques such as multiple regression, discriminant analysis, analysis of variance, principal component analysis, correspondence analysis, and so on. If we define canonical analysis broadly enough, it includes the other techniques as special cases. This fact is used in Gifi (1981) to build a very general system of multivariate analysis methods, which are all versions of a very general form of canonical analysis. A short, but fairly complete, introduction to the Gifi system is given by De Leeuw (1984a). A detailed discussion of the general form of canonical analysis used in this paper is contained in Van der Burg, et al. (1984). The various forms of canonical analysis are typical data analysis techniques, in the sense that they are used in exploratory situations, emphasize graphical representation, and are seldom used for inferential purposes. This is sometimes presented as a disadvantage of this class of techniques, because we have no information about 'generalizability' or 'significance' of the results. It is shown in Gifi (1981), compare also De Leeuw (1984b), that under random sampling assumptions it is possible to derive confidence interval information for large classes of canonical correlation techniques.

If the random sampling assumptions are not appropriate, which will very often be the case, we can use the resampling framework provided by the Bootstrap or the Jackknife (Efron, 1979, 1982, Efron and Gong, 1983). This provides us with 'nonstochastic confidence interval estimates', to paraphrase a term of Freedman and Lane (1983).
In the same way we can try to find significance tests for some interesting hypotheses in this class of techniques. There are some proposals valid under random sampling assumptions in De Leeuw (1984b). As is also pointed out there, these tests have a 'nonstochastic' interpretation in a randomization framework based on permutations. Freedman and Lane (1983) present randomization versions of chi-square and F-test in a very similar framework. Edgington (1980) studies permutation tests as general data analytical tools. In this paper we work out some of the suggestions in De Leeuw (1984b), and study some permutation tests for generalized canonical analysis.

We shall first discuss a fairly general class of generalized canonical correlation techniques in a coordinate-free way. This does not cover all cases of interest. We shall introduce additional generality further on.

Suppose H is a vector space, and $L_1,\ldots,L_m$ are subspaces of H. Then $L_1,\ldots,L_m$ are said to have a p-meet if the dimensionality of their intersection is at least p, i.e. if and only if they have a p-dimensional subspace in common. In canonical analysis we start with m given subspaces, and we try to find the largest possible common subspace. In practical problems, of course, subspaces have no nontrivial meets, and we have to find a best-fitting subspace of predetermined dimensionality. In order to define fit, we need ways to measure distance between subspaces.

In canonical analysis we suppose that H is a real inner product space. If P_j is the orthogonal projector corresponding with L_j, and P_* is the average of the P_j, the fit is the sum of the p largest eigenvalues of P_*. This definition assumes, of course, that these eigenvalues exist, which will always be the case if the L_j are finite dimensional. For computational purposes we need bases for these finite dimensional spaces. Suppose $G_1,\ldots,G_m$ are orthonormal bases for $L_1,\ldots,L_m$. Collect them in the supermatrices $G=(G_1 \cdots G_m)$. Then the p largest eigenvalues of $P_*=\frac{1}{m}GG'$ are equal to the p largest eigenvalues of $C=\frac{1}{m}G'G$, compare De Leeuw (1984a). We can also think of C as a super-matrix, of which the elements C_{ij} are matrices equal to $\frac{1}{m}G_i'G_j$. Clearly C depends on the choice of the bases, but its eigenvalues do not. The eigenvalues of P_* (or of C) are called generalized canonical correlations.

We now make the further generalization mentioned above. Let us call the generalized canonical correlation technique introduced above linear. It is characterized among the more general class of nonlinear techniques by the fact that the subspaces L_j are completely known. In the more complicated nonlinear techniques L_j is only partially known. More precisely L_j is defined as the linear span of k_j vectors. Each of these vectors is known to lie in a given convex cone in H. Thus there are k_j convex cones for L_j, a total of $k_1+\ldots+k_m$ convex cones. We are still interested in the p largest eigenvalues of P_* or C, but these now depend on the choice of each of the vectors from their cones. Generalized canonical correlations in nonlinear problems are defined as those eigenvalues corresponding with a fit which is maximized over this choice. The linear case corresponds with the exceedingly special case in which each of the cones is a single ray through the origin. For more details we refer to Gifi (1981), De Leeuw (1984a), and Van der Burg et al. (1984).

Nonlinear generalized canonical correlation problems are no longer eigenvalue problems. Moreover they cannot be presented in a coordinate-free form. If the columns of G_j are chosen from the corresponding k_j cones, and A is arbitrary, then the columns of G_jA will not be in the same cones any more (unless A is diagonal). In order to stress the similarity with the linear case, we define a matrix $G=(G_1 \cdots G_K)$, with K the sum of the k_j, and with each G_j the spanning set of a single cone. The generalized canonical correlations in the nonlinear problem are a function of the matrix $C=\frac{1}{m}G'G$, although often a very complicated function. If all cones are rays the spanning sets have dimension one, and we are back in the linear case. Again we refer to the literature for all the additional computational and conceptual details. We merely emphasize here that we gain a great deal of generality in passing from the linear to the nonlinear case, but we loose many useful structural mathematical properties.

PERMUTATION DISTRIBUTION OF EIGENVALUES

Suppose that for a given canonical correlation analysis problem we have computed the generalized canonical correlations. We would like to know if these canonical

correlations are 'significant', i.e. if they are larger than one would expect if there was no structure at all in the data. We study the concept of 'no structure at all' in the case where the G_j are given n x k_j matrices, i.e. H is R^n.

Let us start with the linear case. The canonical correlations, collected in the vector λ, are functions of $G=(G_1 \ \ldots \ G_m)$. We write $\lambda(G_1,\ldots,G_m)$ to indicate this. Now suppose that $\underline{Q}_1,\ldots,\underline{Q}_m$ are independent random matrices, which are all uniformly distributed over the n! permutation matrices of order n. Observe that we underline random variables, a nice convention described in detail by Hemelrijk (1966). We study the distribution of the random variable $\underline{\lambda}=\lambda(\underline{Q}_1G_1,\ldots,\underline{Q}_mG_m)$, i.e. we study the permutation distribution (PD) of $\underline{\lambda}$. The idea behind using permutation tests is, of course, that they are one way of formalizing the notion of 'no structure'. The random variation is introduced conditionally on the data, which implies that we do not have to assume a particular probability model that has generated these data.

The PD can be studied in various ways. It can be computed exactly, but because the m independent permutations take on $(n!)^m$ different values, this is only feasible for small n and m. A considerable saving is possible by realizing that λ depends on $G_1,\ldots,G_m$ only through $C=\frac{1}{m}G'G$. Thus it suffices to study the different values taken on by C, each weighted with their probability under the PD.

We illustrate this with a familiar example (Lebart, 1976). Suppose m=2 and F_1 and F_2 are indicators of two nominal variables. Then $F_1'F_1$ and $F_2'F_2$ are diagonal matrices with univariate marginals. Now define $G_1=F_1(F_1'F_1)^{-\frac{1}{2}}$ and $G_2=F_2(F_2'F_2)^{-\frac{1}{2}}$. Suppose $k_1\leq k_2$, and suppose ρ_s (s=1,....,k_1) are the usual canonical correlations between G_1 and G_2. Then C has k_1 eigenvalues (or generalized canonical correlations) equal to $\frac{1}{2}(1+\rho_s)$, k_1 eigenvalues equal to $\frac{1}{2}(1-\rho_s)$, and k_2-k_1 eigenvalues equal to $\frac{1}{2}$. Studying the PD of $\underline{\lambda}$ amounts to the same thing as studying the PD of $\underline{\rho}$, which is a function of $F_1'F_2$ only. Now $\underline{Q}_1'F_1'F_2\underline{Q}_2$ takes as its values all tables with the same marginals as $F_1'F_2$, and gives each table a weight equal to the corresponding hypergeometric probability. Thus the PD we study is the same as the distribution studied in Fisher's exact test of independence in k_1xk_2 contingency tables. Clever enumeration methods to compute this distribution are reviewed and implemented by Verbeek and Kroonenberg (in press) An even more special case should perhaps also be mentioned. If $m=k_1=k_2=2$, then we have Fisher's test for a 2x2 table. In this case the canonical correlations are $\frac{1}{2}(1+\rho)$, with ρ the point-correlation or phi-coefficient. This is monotonic with the chi-square of the table, and for given marginals monotonic with the contents of cell (1,1) in the cross table. Compare Kendall and Stuart (1967, page 549-555), where references to published tables of significance probabilities can be found. If m=2 and both variables are numerical (i.e. $k_1=k_2=1$), then the eigenvalues are $\frac{1}{2}(1+\rho)$ and $\frac{1}{2}(1-\rho)$, with ρ the product moment correlation between the variables. Thus the PD we study in this case is the PD of the correlation coefficient (Kendall and Stuart, 1967, page 373-475), which very closely follows its normal theory sampling distribution. Of course enumeration as a method for the computation of the PD of $\underline{\lambda}$ can also be used in the nonlinear case. It does not seem very practical for m>2, however, and nonlinear methods are often used for large m.

APPROXIMATING THE PERMUTATION DISTRIBUTION

We study the permutation distribution of the matrix $\underline{C}$ first, where $\underline{C}=\frac{1}{m}(\underline{Q}_1G_1 \ \ldots \ \underline{Q}_mG_m)'(\underline{Q}_1G_1 \ \ldots \ \underline{Q}_mG_m)$.
We assume here that $G_i'G_j=I$, and that the columns of all G_j add up to zero. This can be done without loss of generality, at least in the linear case. The diagonal submatrices $m\underline{C}_{ij}$ are all equal to the identity, only the off-diagonal matrices $\underline{C}_{ij}$ have random variation. The asymptotic distribution of the elements of $\underline{C}$, if $n\to\infty$, follows directly from the permutational limit theorems of Wald, Wolfowitz, Hoeffding, Hajek, Motoo, and many others. Compare Puri and Sen (1971, section

3.4) for references. More precise results have been derived recently by Ho and Chen (1978).

For our purposes it suffices to observe that the elements in the off-diagonal submatrices of $m\underline{C}$ are approximately independent normal variables, with means zero, and with dispersion $(n-1)^{-1}$. Normality follows from the permutational limit theorems, while the expression for the variances follows from simple algebraic computations. We prefer to interpret this result, in conformity with Ho and Chen (1978), as a approximation theorem, and not as a limit theorem. Thus off-diagonal elements of $(n-1)^{\frac{1}{2}}m\underline{C}$ are approximately independent standard normal. This implies, by the way, the approximate chi-square distribution of the sum of squares of all elements in a single off-diagonal $\underline{C}_{ij}$, which was used earlier by Freedman and Lane (1983).

The above result is useful also in the nonlinear case, which can be formulated in terms of $\underline{C}$. We now proceed with the eigenproblem for $\underline{C}$, which is only relevant in the linear case. If we write $m\underline{C}=I+(n-1)^{\frac{1}{2}}\underline{Z}$, then $\underline{Z}$ has approximately standard normal off-diagonal blocks and exactly zero diagonal blocks. Clearly $m\lambda(\underline{C})=1+(n-1)^{\frac{1}{2}}\lambda(\underline{Z})$, or $(n-1)^{-\frac{1}{2}}(m\lambda(\underline{C})-1)=\lambda(\underline{Z})$. Now suppose $\underline{\lambda}_s$ are the eigenvalues of $\underline{C}$, in decreasing order. Because ordered eigenvalues are a continuous function of the matrix elements (Kato, 1970, section II.6.4, page 124), it follows that the approximate PD of $(n-1)^{-\frac{1}{2}}(m\underline{\lambda}_s-1)$ is the distribution of $\underline{\omega}_s$, which is the s-th eigenvalue of a matrix with off-diagonal blocks composed of independent standard normals and diagonal blocks zero. The distribution of $\underline{\omega}_s$ only depends on the numbers k_j, and it can in principle be tabulated as a function of these numbers. Of course the attractiveness of this result is that the approximate PD does not depend on the spaces L_j, but only on their dimensionalities. The result is thus not only coordinate-free, but actually 'space-free'.

It is possible to proceed a little further if m=2. There is only one off-diagonal block in this case, and the $\underline{\omega}_s$ are the singular values of this block. Singular values of this block $\underline{Z}$ are square roots of eigenvalues of $\underline{Z}'\underline{Z}$, which are approximately standard Wishart matrices. Significance points of eigenvalues of standard Wishart matrices have been tabulated by Clemm et al. (1973a, 1973b), compare also Krishnaiah (1980). The tables are computed by using formulas for the exact distribution. It would be interesting to generalize these formulas to the case m>2, and to write algorithms for evaluating the resulting distribution. Such a project seems to be quite feasible, and it would make it possible to compute the approximate PD of the generalized canonical correlation coefficients in all linear problems. For nonlinear problems the situation is inherently more complicated. Although the canonical correlations are still functions of $\underline{C}$, these functions are so complicated that we cannot proceed any further in the approximation of the PD. In the nonlinear case we have to resort to Monte Carlo methods, which we will described next.

MONTE CARLO METHODS

We give a brief summary of the results obtained so far. Let
$F_s(x)=\text{prob}\{(n-1)^{-\frac{1}{2}}(m\underline{\lambda}_s-1)<x\}$,
$G_s(x)=\text{prob}\{\underline{\omega}_s<x\}$.
Then we know that $F_s(x)\cong G_s(x)$, with the approximation becoming more precise as n increases. We have seen that $F_s(x)$ can sometimes be computed exactly (if m=2, and k_1, k_2, and n are all small). In other cases $G_s(x)$ can be computed exactly. If m=2, for instance, and $k_1=k_2=1$, then $\underline{\omega}_1$ is the modulus of a standard normal variable. For general k_1 and k_2 we need the distribution of the eigenvalues of a standard Wishart matrix, for which formulas and tables are available in the work of Clemm, Krishnaiah, et al. (1973b).

In other cases (and this includes all nonlinear generalized canonical analysis problems) we have to approximate either F_s or G_s by Monte Carlo methods. Approxi-

mating F_s is very easy and straightforward. Each sample consists of m random permutations of order n, and after permuting the individuals or objects the canonical analysis is carried out. Then a new sample is drawn, and so on. We call this the random permutation method. Approximating G_s is somewhat more complicated. We fill the non-diagonal blocks of C with independent standard normals, and then compute our statistics. Then draw another sample, and so on. This is the C-matrix method. We do not know how many Monte Carlo samples are needed for a good approximation, and which of the two methods works better. It seems to us that approximating F_s gives a more direct answer to the question we are investigating, and it is also more easy to interpret from a data analysis point of view. Approximating G_s only seems to make sense if n is already quite large, and in such conditions it may very well be more economical.
Very often our choice between the two methods is more or less determined by the fact that some canonical analysis programs use the data matrix as input, while others use the C-matrix (the Burt-table).

A final possibility, which we shall not consider in this paper, is to approximate $F_s(x)$ or $G_s(x)$ for very large m, or very large Σk_j. De Leeuw (1984b) has some tentative suggestions and some references relevant for this case.

EXAMPLES

In this section we analyze various examples to illustrate significance testing, and to compare the two approximations to the PD. The random permutation method was implemented in FORTRAN, on the basis of the generalized canonical correlation program OVERALS (Van der Burg a.o., 1984). The computations for the C-matrix method were done in APL, on the basis of an ad hoc program for the linear case. The enumeration method for computing the exact PD used programs developed by Verbeek and Kroonenberg (in press). We would like to thank Pieter Kroonenberg, who assisted us with the use of these programs.

Some general remarks about our analysis of the examples are perhaps in order. We use 'generalized correlation coefficients' and 'eigenvalues' for the same objects, i.e. the eigenvalues of P_* and C. Variables are treated as multiple nominal if they are coded as dummies i.e. as sets of binary variables. On these sets of dummies a linear analysis is performed. A variable is treated in a single nominal way if it is quantified as a single vector selected from the subspace spanned by the dummies, it is single ordinal if in addition we require that the order of the categories is maintained. Analyses with single variables are necessarily nonlinear, except in the case that p=1.
In our tables the empirical value of a generalized canonical correlation is the value which corresponds to the original data ('before permutation'). The distribution of the eigenvalues is plotted by using cumulative normal probability plots, and tabulated by giving percentiles (5,25,50,75,95). Because our Monte Carlo runs use samples of size 100, we simply tabulate the order statistics $\lambda_{(5)}$, $\lambda_{(25)}$, $\lambda_{(50)}$, $\lambda_{(75)}$, $\lambda_{(95)}$.

Example 1: Journal data, taken from Kroonenberg and Van der Veer (1980). For several Dutch journals the authors investigated crime reports, more precisely the country of origin of the criminal (Surinam, Turkey & Marocco, No reference) versus the number of the page on which the crime was reported (One, Three, or Other). We used the frequency tables of four papers (Parool, Volkskrant, NRC, Telegraaf). In table 1 we give the two largest generalized canonical correlations, and their significances (i.e. the probability or estimated probability of a value at least as large as the observed value). We use three methods: complete enumeration, the random permutation method, and the C-matrix method. Both Monte Carlo methods used 100 permutations or normal samples. They are not very precise in this case, but they do give a quite reliable idea about the exact significance. For 'Telegraaf' the approximation was very good, because the Telegraaf

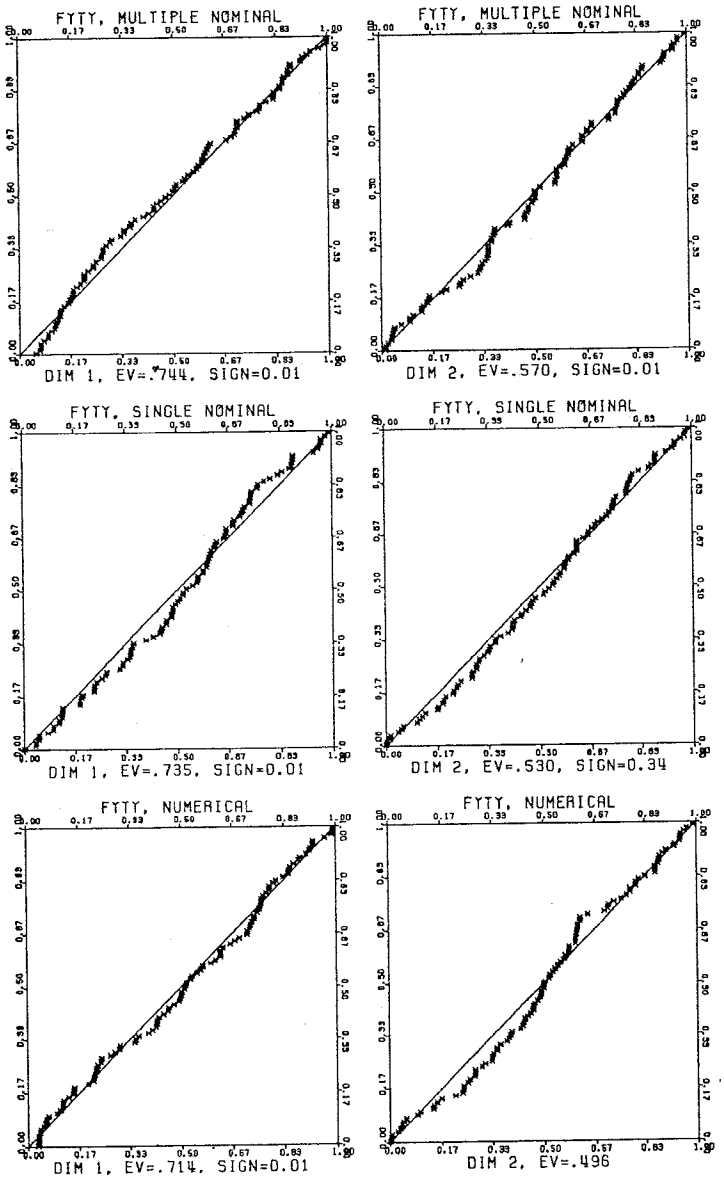


FIGURE 1: From Year to Year random permutations. Probability plot of eigenvalues.

table is based on 106 crime reports, while the other newspapers each had approximately 50. The tendency is that the C-matrix method underestimates and the permutation method overestimates the true probability.

TABLE 1: Journals: ethnic group x page number. Generalized canonical correlation (ev), significance according to all permutations (sign1), to a random sample of permutations (sign2), and to the C-matrix method (sign3).

	ev	sign1	sign2	sign3
Parool	.605	.62	.59	.64
	.571	.22	.25	.11
Volkskrant	.593	.91	.94	.81
	.544	.46	.54	.44
NRC	.691	.05	.14	.09
	.571	.11	.12	.11
Telegraaf	.596	.36	.39	.36
	.507	.83	.83	.83

Example 2: Patients, taken from Williams and Grizzle (1972). Cancer patients classified according to ulcer pain and type of medication. A 4x3 table, with 244 individuals, of which 170 are in cell (no pain, no medication). The eigenvalues are .846 and .645, they are both significant ($p<.01$) according to the random permutation method, the C-matrix method, and the tables of Krishnaiah. The exact significance probability, computed by enumeration, is smaller than 10^{-7}. Table 2 compares the two Monte Carlo methods, which give very similar estimates of the distribution.

TABLE 2: Generalized canonical correlations: empirical values (ev) and generated values at 5, 25, 75 and 95 percent. Random permutation method and C-matrix method. Patients data, multiple nominal.

	Permutation method		C-matrix method	
	dim1	dim2	dim1	dim2
ev	.846	.645	.846	.645
5	.530	.508	.535	.507
25	.555	.518	.550	.517
50	.564	.526	.567	.528
75	.577	.537	.579	.538
95	.596	.551	.608	.553

Example 3: From Year to Year (FYTY) data are part of a large schoolcareer survey (for references, see De Leeuw and Stoop, 1979). There are 1762 school children in our study. The four variables are choice of school after leaving primary school, achievement test score in sixth grade, educational level of father, and educational level of mother. We did three analyses: multiple nominal, single nominal, and numerical. The two educational level variables were used in the first set, the test score and school choice in the second set. Eigenvalues of the three analyses (all with p=2) were, respectively, .744 and .570, .735 and .530, .714 and .496. The probability plots in figure 1 show that the eigenvalue distribution does not deviate very much from normality. The first multiple nominal eigenvalue is a little bit light in the tails, but the deviations are

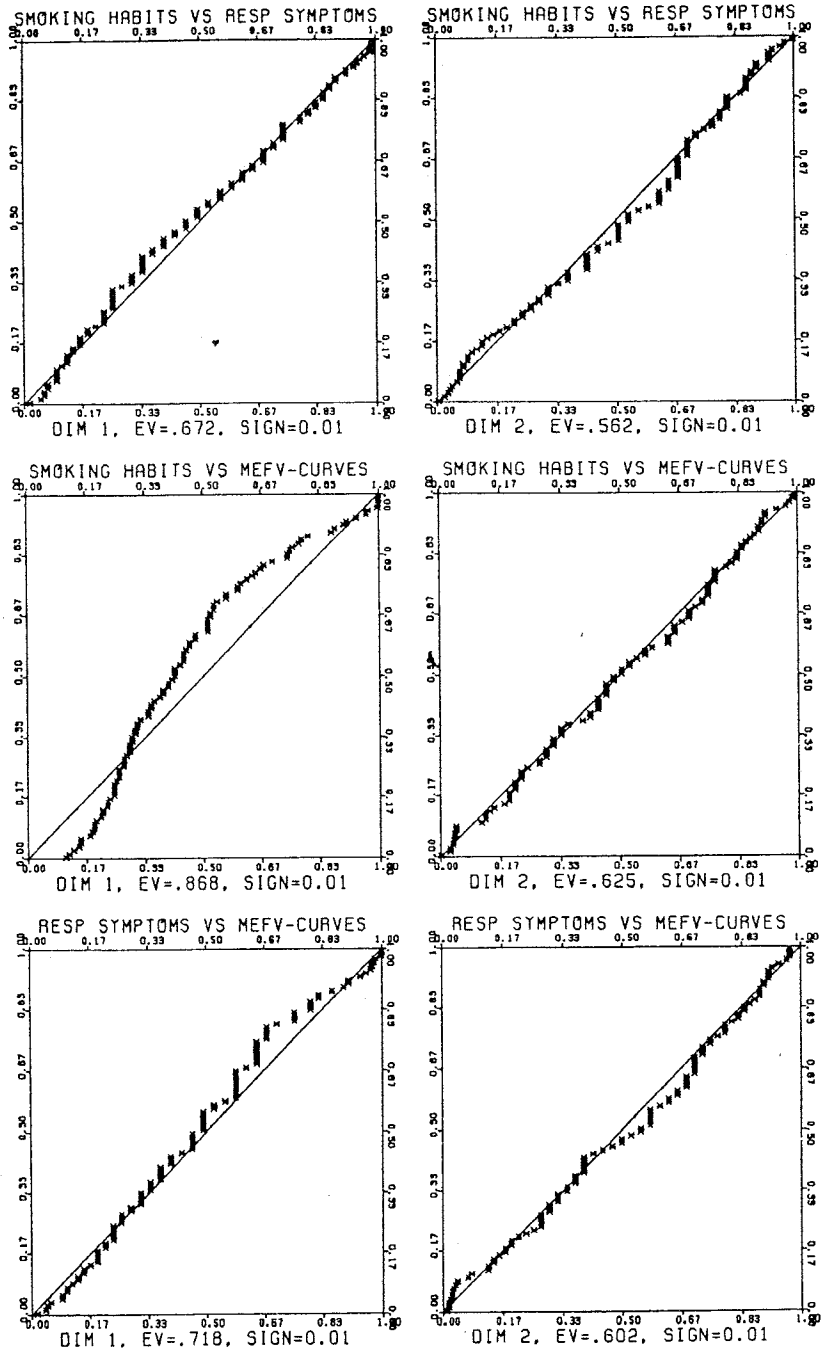


FIGURE 2: Chronic lung disease random permutations. Probability plot of eigenvalues.

rather small. Tables 3 and 4 show the percentiles obtained from the Monte Carlo method. The tables agree, although the C-matrix method gives slightly higher values, especially for lower percentages. This may be due partly to the superior precision of the APL eigenvalue algorithm, the last column of table 3 shows that OVERALS cannot be very precise (theoretically all values should be larger than .500 in the two sets case). Tables 3 and 4 tell us that the second eigenvalue has a low significance probability. Especially in the numerical case it is not significant. The same conclusion follows form the tables of Krishnaiah.

TABLE 3: Generalized canonical correlations: empirical values (ev) and generated values at 5, 25, 50, 75 and 95 percent. Random permutation method. From Year to Year data, multiple nominal, single nominal and numerical.

	mult nominal		single nom		numerical	
	dim1	dim2	dim1	dim2	dim1	dim2
ev	.744	.570	.735	.530	.714	.496
5	.549	.532	.533	.511	.501	.482
25	.554	.542	.543	.521	.511	.494
50	.559	.545	.549	.527	.518	.499
75	.564	.550	.554	.532	.525	.504
95	.570	.557	.565	.541	.532	.514

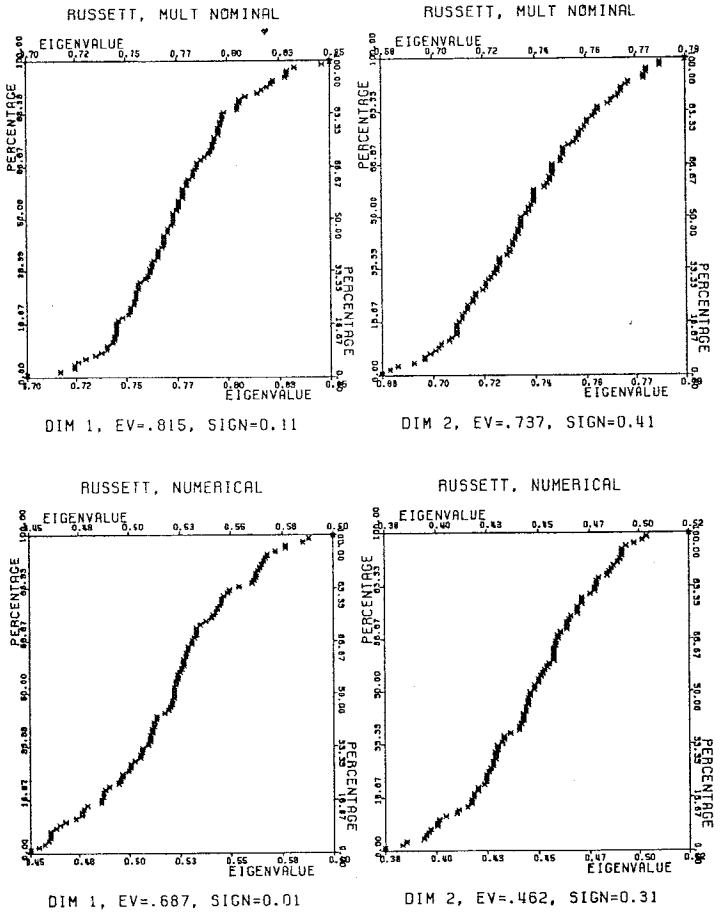
TABLE 4: Generalized canonical correlations: empirical values (ev) and generated values at 5, 25, 50, 75 and 95 percent. C-matrix method. From Year to Year data, multiple nominal and numerical.

	multiple nom		numerical	
	dim1	dim2	dim1	dim2
ev	.744	.570	.714	.496
5	.553	.545	.509	.500
25	.560	.550	.514	.503
50	.565	.554	.521	.506
75	.569	.557	.525	.504
95	.574	.562	.535	.514

Example 4: Chronic Lung Disease data, taken from Van Pelt a.o. (1985). Gerard Borsboom and Wilfrid van Pelt assisted us in analyzing these data. There are 4241 patients, and three sets of variables. The first set (smoking habits) has two variables, the second set (respiratory symptoms) has five, and the third set (flow-volume-curve parameters) has 13. Variables in the first two sets are treated as single nominal, and those in the third set as single ordinal. The three two-sets nonlinear canonical correlation problems with p=2 were computed separately. The eigenvalues are in the first row of table 5, while the rest of the table shows percentiles of the PD (all estimated by the random permutation method). In the S & M analysis probability plots [given in figure 2] show a large deviation from normality for the first eigenvalue, all eigenvalues are very clearly significant, however. All estimated PD's have small variance, we see that $\lambda_{(95)}-\lambda_{(5)}<.03$, except for dimension 1 of S & M (which has a very light tail on the right, $\lambda_{(95)}-\lambda_{(75)}=.035$).

TABLE 5: Generalized canonical correlations: empirical values (ev) Random permutation method. Chronic Lung Disease data, S = smoking habits, M = MEFV-curve indices, R = respiratory symptoms.

	S & R		S & M		R & M	
	dim1	dim2	dim1	dim2	dim1	dim2
ev	.672	.562	.868	.625	.718	.602
5	.539	.513	.589	.559	.581	.555
25	.544	.520	.598	.569	.585	.565
50	.548	.526	.608	.578	.589	.571
75	.552	.532	.619	.581	.592	.577
95	.559	.535	.654	.587	.601	.580



DIM 1, EV=.815, SIGN=0.11

DIM 2, EV=.737, SIGN=0.41

DIM 1, EV=.687, SIGN=0.01

DIM 2, EV=.462, SIGN=0.31

FIGURE 3: Russett random permutations. Eigenvalues (ordered) against percentages.

Example 5: Data from Russett (1969). Three sets of variables describing 47 countries. The first set contains two variables concerning ownership of land, the second set consists of gross national product and percentage of people working in agriculture, and the third set has four variables that are indicators for political instability. We used discretizations of the variables as in Gifi (1981). Data were analyzed either with multiple nominal or with numerical options, PD's were estimated by random permutations and by the C-matrix method. Eigenvalues and percentiles are in table 6. The estimated significance levels (using random permutations) are .11 and .41 for multiple nominal and .01 and .31 for numerical. These significances are low, but the number of countries is rather small. The small number of countries also explains the fairly large differences between the two Monte Carlo methods. The C-matrix method gives higher values for the percentiles. Figure 3 shows the empirical distribution functions of the eigenvalues, and figure 4 the cumulative normal plots. There seems to be no systematic deviation from normality.

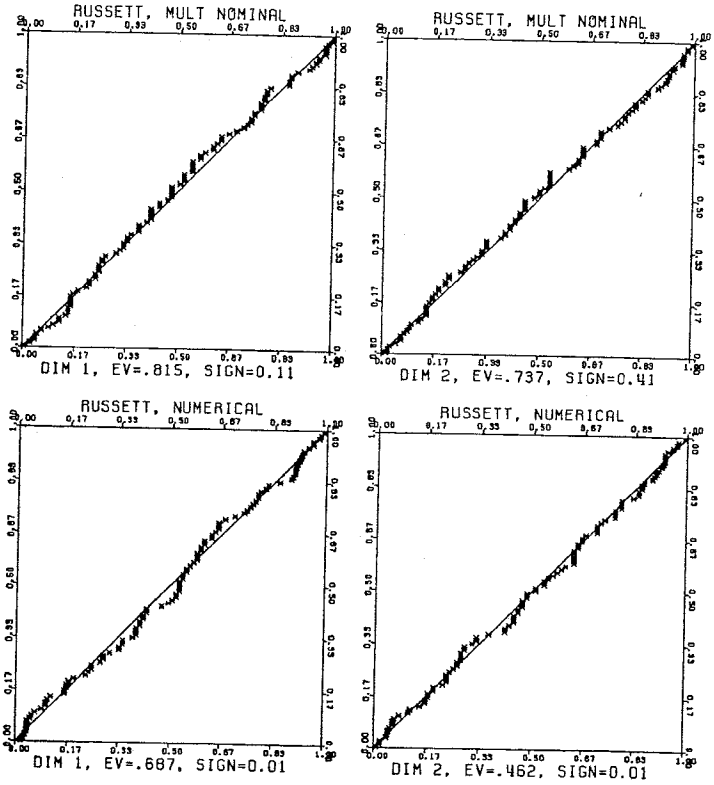


FIGURE 4: Russett random permutations. Probability plot of eigenvalues.

TABLE 6: Generalized canonical correlations: empirical values (ev) and generated values at 5, 25, 50, 75 and 95 percent. Random permutation method and C-matric method. Russett data, multiple nominal and numerical.

	Random permutation method				C-matrix method			
	mult nom		numerical		mult nom		numerical	
	dim1	dim2	dim1	dim2	dim1	dim2	dim1	dim2
ev	.815	.737	.687	.462	.815	.737	.687	.462
5	.724	.695	.462	.397	.748	.710	.457	.408
25	.754	.714	.499	.426	.775	.732	.480	.426
50	.773	.731	.525	.445	.795	.745	.500	.443
75	.794	.752	.545	.467	.809	.757	.519	.463
95	.830	.778	.576	.488	.841	.782	.562	.488

CONCLUSIONS

It appears from our examples that both the random permutation method and the C-matrix method give a fairly good approximation to significance probabilities and permutation distributions. In order to be sure, we shall have to apply enumeration somewhat more extensively. We also need to generalize the exact computation of G_s to the case of more than two sets. But from our experience so far, we can say that the Monte Carlo methods give the correct indication of the order of significance, and they show that with multiple options and not too many individuals the eigenvalues have to be very high to be significant. It is not surprising, for instance, that Gifi (1981) had great difficulty in interpreting the second dimension in the solution for the Russett data. Our approximations to the permutation distribution guard us against chance capitalization, and against trying to interpret effects which are not really there.

REFERENCES

Clemm, D.S., Chattopadhyay, A.K. and Krishnaiah, P.R. (1973a). Upper percentage points of the individual roots of the Wishart matrix. Sankhya, Ser. B 35, 325-338.

Clemm, D.S., Krishnaiah, P.R. and Waikar, V.B. (1973b). Tables for the extreme roots of the Wishart matrix. J. Statist. Comp. and Simulation 2, 65-92.

De Leeuw, J. (1984a). The Gifi system of nonlinear multivariate analysis. In: E. Diday, M. Jambu, L.Lebart, J. Pages, and R. Tomassone (eds), Data Analysis and Informatics, vol III, Amsterdam: North Holland.

De Leeuw, J. (1984b). Statistical properities of multiple correspondence analysis, RR-84-06, Department of Data Theory, The Netherlands.

De Leeuw, J. and Stoop, I. (1979). Sekundaire analyse 'Van Jaar tot Jaar' met behulp van niet-lineaire multivariate technieken. In: J.L. Peschar (ed), Van Achteren naar Voren, 's-Gravenhage: Staatsuitgeverij.

Edgington, E.S. (1980). Randomization Tests. New York: Marcel Dekker Inc..

Efron, B. (1979). Bootstrap methods: another look at the Jackknife. Annals of Statistics, 7, 1-26.

Efron, B. (1982). The JackKnife, the Bootstrap, and other Resampling Plans. CBMS-NSF Regional Conference Series in Applied Mathematics, no 38, Philadelphia: SIAM.

Efron, B. and Gong, G.A. (1983). A leisurely look at the Bootstrap, the Jackknife, and cross-validation. The American Statistician, 37, 36-48.

Freedman, D.A. and Lane, D. (1983). Significance testing in a nonstochastic setting. In: P.J. Bickel, K. Doksum, J.L. Hodges jr (eds), A Festschrift for Erich L. Lehmann. Belmont: Wadworth.

Gifi, A. (1981). Nonlinear Multivariate Analysis, Department of Data Theory, University of Leiden. New edition by DSWO-press, Leiden, 1985.

Hemelrijk, J. (1966). Underlining random variables. Statistica Neerlandica, 20, 1-8.

Ho, S.T. and Chen, L. (1978), An L_n bound for the remainder in a combinatorial central limit theorem, Annals of Probability, 6, 231-279.

Kato, T. (1976), Perturbation Theory of Linear Operators. Berlin: Springer.

Kendall, M.G. and Stuart, A. (1967), The Advanced Theory of Statistics (3 vols), London: Griffin.

Krishnaiah, P.R. (1980), Computations of Some Multivariate Distributions. In: P.R. Krishnaiah, Handbook of Statistics, Amsterdam: North Holland.

Kroonenberg, P.M. and Van der Veer, R. (1980). Rasdiscriminatie bij het opmaken van een krantebericht: een heranalyse van een onderzoek van Frans Bovenkerk. Mens en Maatschappij, 3, 301-315.

Lebart, L. (1976), The significancy of eigenvalues issued from correspondence analysis. In: Proc. Comput. Statist., 38-45. Vienna: Physica Verlag.

Puri, M.L. and Sen, P.K. (1971), Nonparametric methods in multivariate analysis. New York: Wiley.

Van der Burg, E., De Leeuw, J. and Verdegaal, R. (1984), Generalized nonlinear canonical correlation, RR-84-12, Department of Data Theory, University of Leiden.

Van Pelt, W., Quanjer, Ph.H., Wise, M.E., Van der Burg, E. and Van der Lende, R. (1985). Analysis of maximum expiratory flow-volume curves using canonical correlation analysis. Methods of Information in Medicine, vol. 24, no. 2.

Verbeek, A., Kroonenberg, P.M. (in press). A survey of algorithms for exact distributions of test statistics in R*C contingency tables with fixed margins. Computational Statistics and Data Analysis.

Williams, O.D. and Grizzle, J.E. (1972). Analysis of contingency tables having ordered response categories. J. of the American Stat. Association,67, 55-63.