CORRESPONDENCE ANALYSIS AND STABILITY

Jacqueline Meulman
Department of Data Theory FWS/RUL
Middelste gracht 4
2312 TW Leiden
The Netherlands

Paper prepared for the Table Ronde 'Analyse des Données'
Toulouse, January 9-10, 1984

Correspondence analysis and stability

This paper discusses two applications of correspondence analysis that are quite different from a statistical point of view. This will be illustrated by performing stability analyses, both by the delta method and by the bootstrap method. In some cases the different methods give similar results in estimating the variances, while in others the bootstrap method is much more appropriate; its application is not as limited as the employment of the delta method. For bias correction the bootstrap method is also more convenient.

1. The analysis of contingency tables

A very common application of correspondence analysis is the analysis of two dimensional contingency tables, displaying frequencies of observations on two categorical variables. Correspondence analysis will give quantifications for both the rows and the columns of this table, which are optimal in a number of ways. Since the optimality properties are very well documentated elsewhere, we shall not amplify upon them here.
It is convenient to realize that the contingency table can be rewritten as a matrix with n rows, the total number of observations, and 2 columns, the number of variables. Applying Multiple Correspondence or Homogeneity analysis to this table will give results that are equivalent to correspondence analysis of the contingency table. In this paper we concentrate on the stability of the quantifications (and their corresponding singular values), which will be expressed in their variances.

1.1. The bootstrap and delta method for computation of variances.

Each cell of the contingency table represents a profile associated with a a frequency. To apply the delta method we have to make assumptions on their distribution. It is assumed that the n observations are a random sample from an infinite population, that they are independently distributed and that the frequencies of the profiles, the bivariate marginals, have a multinomial distribution that is asymptotically multinormal.
Variances of the estimates can be obtained by computing the partial derivatives of the singular vectors (Gifi 1981). These variances can be employed to construct confidence ellipses around the category points in two dimen-

sional space by using χ^2 values to obtain 95% confidence regions.
Another way to acquire variances of the estimates is to apply resampling methods. Among them the Jackknife and the Bootstrap are best known (Efron 1982). In this paper we will apply the bootstrap method. This method does not assume an apriori specified joint distribution, as the delta method, but uses the empirical distribution itself: we resample with replacement from the n observations to get a data matrix with slightly different bivariate marginals while n is kept constant. Next the analysis is performed on this table, and this procedure is repeated a number, say B, times. From the B different quantifications for each category we compute the variances and covariances. Confidence ellipses can be drawn in the same way as mentioned above, using the mean of the B coordinates as the location of the category point and correcting the variance with $\frac{n}{n-1}$.

The bootstrap and the delta method have been applied to data from Fisher (1940), concerning the cross classification of individuals with respect to eye color and hair color. The table is given below.

Table 1. Tocher's data for Caithness compiled by K. Maung.

Eye color	Hair color					
	5 Fair	6 Red	7 Medium	8 Dark	9 Black	Total
1 Blue	326	38	241	110	3	718
2 Light	688	116	584	188	4	1580
3 Medium	343	84	909	412	26	1774
4 Dark	98	48	403	681	85	1315
Total	1455	286	2137	1391	118	5387

Choosing the number of bootstrap samples B is still a largely unexplored problem for the kind of techniques we are using. In this study we decided to vary B from 10, 20, 30, 40, 50, 100, 150, 200, 250, 300 to 500 samples. The sum of the variances across the two dimensions is given in table 2 and graphically depicted in figure 1.
It is clear that categories with large univariate marginals are very stable, n'importe quoi the size of B (categories 5,7,4,3,8 and 4, in the order of their stability), the relative instability of category 1 is itself stable for different sizes of B, and the instability of categories 6 and 9 is either

underestimated or overestimated for small values of B. The singular values are very stable.
The results for the delta method are quite similar; the estimates for the categories 6 and 9, of which the variances are fluctuating for different sizes of B, are similar to the results for small values of B.
Nevertheless, figure 1 might draw our attention too much to details. When we plot the confidence regions for both methods, with B=300, in a joint figure (fig. 2) we see that the discrepancies described above are visible, but not substantive. The graphical display smoothes away small numerical differences in the tables 2 and 3 and shows strikingly similar ellipses for the categories 5,2,7,3,4 and 8.

1.2. The bootstrap method and the delta method for bias correction

Both the delta and the bootstrap method can be used for bias correction, but here the bootstrap method is more clearly favored for its simplicity. The delta method gives the estimates for the population by computing the second derivatives (see the paper of de Leeuw, this meeting), which makes the method rather unwieldly. The bootstrap results can be used straight forward which can be seen from the following formula

$$X = \alpha X_o + (1-\alpha)\ \bar{X}_b$$

When $\alpha = 1$ we obtain the solution for the original table X_o, $\alpha = 0$ gives the mean of the B bootstrap quantifications $\bar{X}_b$, and $\alpha = 2$ will give the population estimate corrected for bias.
This has been done for B = 300 in table 4, and it is clear that the bias is not very large in this example, due to large n. Original quantifications, the mean of the bootstrap values and the quantifications corrected for bias are also given in figure 3.
Smaller sizes for B do not diverge much in estimating the bias. These estimates are plotted in figure 4 for 10 different sizes of B. Again most variation is to be found for categories 6 and 9, but it is obvious that already small sizes of B sufficiently recover the hardly existing bias.

2. The analysis of similarity tables.

Our second exapmple concerns data from a completely different nature, but still very eligible for correspondence analysis. This application employs correspondence analysis as a Multidimensional Scaling technique, which tries to map (dis)similarities between objects as distances between points in low dimensional space. The close relation between correspondence analy-

sis and multidimensional scaling à la Torgerson (1958), also called Classical MDS, has been discussed in Heiser & Meulman (1983). Here it suffices to mention that we can compute the χ^2 - distances between the rows or the columns of a table which will serve as derived dissimilarities. These dissimilarities are approximated from below by correspondence analysis; full dimensionality will represent them perfectly.

The data to be analyzed consist of 20 replications of a dissimilarity matrix. They are due to Ramsay (1968) and the cells contain ratio judgements of dissimilarity between 21 colors from the Munsell system. The dissimilarities have been transformed for correspondence analysis into similarities by a negative exponential function, i.e. $s_{ij} = e^{-d_{ij}}$.

The fact that we have the disposal of 20 replications makes a robust estimation possible by analyzing the *average* data matrix across replications. Moreover it gives us the opportunity to apply the bootstrap method in a quite dissimilar fashion compared to the first example. In this case we consider the set of similarity matrices as the distribution function to resample from; each complete similarity matrix is treated as an observation. In this way resampling with replacement will give a new set of 20 matrices of which we analyze the average.

The delta method has the same interpretation in this framework as in the first example. The cells are considered as weights associated with a profile that represents two colors being grouped together. The assumption that the ratio judgements can be thought of as frequencies of joint occurrence, based on independent trials, does not make much sense, though the fact that the cells are averages across replications makes this application somewhat less awkward. Thus the bootstrap method defines the distribution function *across replications*, the delta method defines it *across the mean values themselves*, ignoring the existing information of the variation across replications.

The bootstrap has been performed with B = 20; confidence ellipses for both methods can be seen in figure 5. For each color point the ellipses from the delta method are much larger than from the bootstrap, but they are clearly 'nested'. From previous analyses it is known that most of the instability, represented by overlapping ellipses can be explained by the fact that the singular values of the first two dimensions are not very separated. This is caused by the properties of the data: the colors, except for colors 6,7, and 8, are defined on a circle in the Munsell system. The delta method results are sensitive for this situation and the bootstrap will display configurations in a slightly different principal plane. For the latter method

the undesirable representation can be anticipated by rotating the B configurations separately towards the solution for the original data. Orthogonal procrustes rotations have been performed, since this procedure does not change the distances that are approximations of the χ^2 distances from the similarity table.

The increase in stability for the bootstrap is clearly visible in figure 6. Most ellipses are now neatly separated, except for colors 10,11 and 12. This result is very similar to the outcome of previous analyses with nonmetric MDS. Finally we want to mention the close resemblance with the performance of classical MDS (Ramsay (1968, see also the paper by Heiser & Meulman for this meeting). This situation can be explained by the fact that the marginals of the similarity table are not too different, thus the differential weighting does not have a large influence.

This second example shows the flexibility of the bootstrap method: in cases that the principal axes orientation is not very informative we still can obtain meaningful information about the stability of the coordinates and in cases that rigid assumptions are unwarranted we do not have to refrain from stability analysis completely.

References

Fisher, R.A. The precision of discriminant functions. Ann. Eugenics, 10, 1940, 422-429

Gifi, A. Nonlinear Multivariate analysis. Department of Data Theory, Leiden 1981

Ramsay, J.O. Economical method of analyzing perceived color differences. Journal of the Optical Society of America, 1968, 58, 19-22

Torgerson, W.S. Theory and methods of scaling. Wiley, New York 1958.

Efron, B. The Jackknife, the Bootstrap and other Resampling plans. CBMS NSF Regional Conference Series in Applied Mathematics no. 38. Philadelphia, SIAM 1982.

Heiser, W.J. & Meulman, J. Analyzing rectangular tables by joint and constrained Multidimensional Scaling. Journal of Econometrics, 1983, 22, 139-167.

De Leeuw, J. Jackknife and Bootstrap in multinomial situations, 1984.*

Heiser, W.J. & Meulman, J. Cross validation of multidimensional representations, 1984.*

* Papers prepared for the Table Ronde 'Analyse des Donnees'. Toulouse.

Table 2. Sum of variances across dimensions

	B = 10	B = 20	B = 30	B = 40	B = 50	B = 100
SV	.269E-3	.208E-3	.234E-3	.239E-3	.232E-3	.302E-3
1	.766E-2	.782E-2	.804E-2	.821E-2	.732E-2	.784E-2
2	.148E-2	.255E-2	.247E-2	.266E-2	.247E-2	.224E-2
3	.145E-2	.163E-2	.187E-2	.160E-2	.145E-2	.189E-2
4	.123E-2	.117E-2	.145E-2	.133E-2	.120E-2	.154E-2
5	.986E-3	.139E-2	.174E-2	.152E-2	.139E-2	.160E-2
6	.149E-1	.238E-1	.277E-1	.306E-1	.278E-1	.267E-1
7	.122E-2	.151E-2	.181E-2	.162E-2	.153E-2	.149E-2
8	.150E-2	.191E-2	.254E-2	.211E-2	.193E-2	.182E-2
9	.498E-1	.492E-1	.505E-1	.436E-1	.404E-1	.396E-1
Sum	.802E-1	.910E-1	.981E-1	.933E-1	.855E-1	.847E-1

	B = 150	B = 200	B = 250	B = 300	B = 500	Delta
SV	.318E-3	.309E-3	.300E-3	.303E-3	.315E-3	.317E-3
1	.838E-2	.845E-2	.845E-2	.859E-2	.873E-2	.817E-2
2	.216E-2	.222E-2	.219E-2	.227E-2	.233E-2	.238E-2
3	.187E-2	.190E-2	.204E-2	.205E-2	.200E-2	.188E-2
4	.151E-2	.158E-2	.165E-2	.166E-2	.161E-2	.154E-2
5	.175E-2	.177E-2	.177E-2	.172E-2	.171E-2	.154E-2
6	.265E-1	.273E-1	.270E-1	.269E-1	.268E-1	.251E-1
7	.147E-2	.142E-2	.145E-2	.146E-2	.149E-2	.145E-2
8	.189E-2	.190E-2.	.190E-2	.201E-2	.203E-2	.217E-2
9	.426E-1	.413E-1	.413E-1	.429E-1	.430E-1	.465E-1
Sum	.881E-1	.879E-1	.877E-1	.896E-1	.897E-1	.907E-1

Table 3. Variances and covariances

	dim 1	B = 300	dim 2	dim 1	Delta	dim 2
SV	.124E-3	.491E-4	.178E-3	.136E-3	.431E-4	.181E-3
1	.215E-2	-.838E-3	.644E-2	.224E-2	-.857E-3	.593E-2
2	.596E-3	-.121E-3	.168E-2	.619E-3	-.106E-3	.176E-2
3	.139E-2	.105E-3	.657E-3	.134E-2	.785E-4	.542E-3
4	.783E-3	-.367E-3	.880E-3	.726E-3	-.407E-3	.810E-3
5	.748E-3	.313E-3	.975E-3	.696E-3	.317E-3	.848E-3
6	.559E-2	-.116E-2	.213E-1	.641E-2	-.166E-2	.187E-1
7	.970E-3	-.464E-4	.494E-3	.100E-2	-.660E-4	.452E-3
8	.817E-3	.128E-4	.119E-2	.804E-3	-.783E-4	.137E-2
9	.123E-1	.108E-1	.306E-1	.123E-1	.107E-1	.342E-1

Table 4. Coordinates

	B = 300		Original		Bias correction	
1	-.600	.387	-.599	.397	-.598	.408
2	-.659	.217	-.660	.212	-.661	.208
3	.051	-.588	.050	-.588	.049	-.588
4	1.051	.324	1.052	.322	1.053	.319
5	-.811	.415	-.814	.417	-.818	.420
6	-.358	.134	-.349	.116	-.340	.098
7	-.063	-.500	-.063	-.500	-.063	-.500
8	.881	.250	.881	.250	.881	.249
9	1.631	.671	1.639	.688	1.645	.705

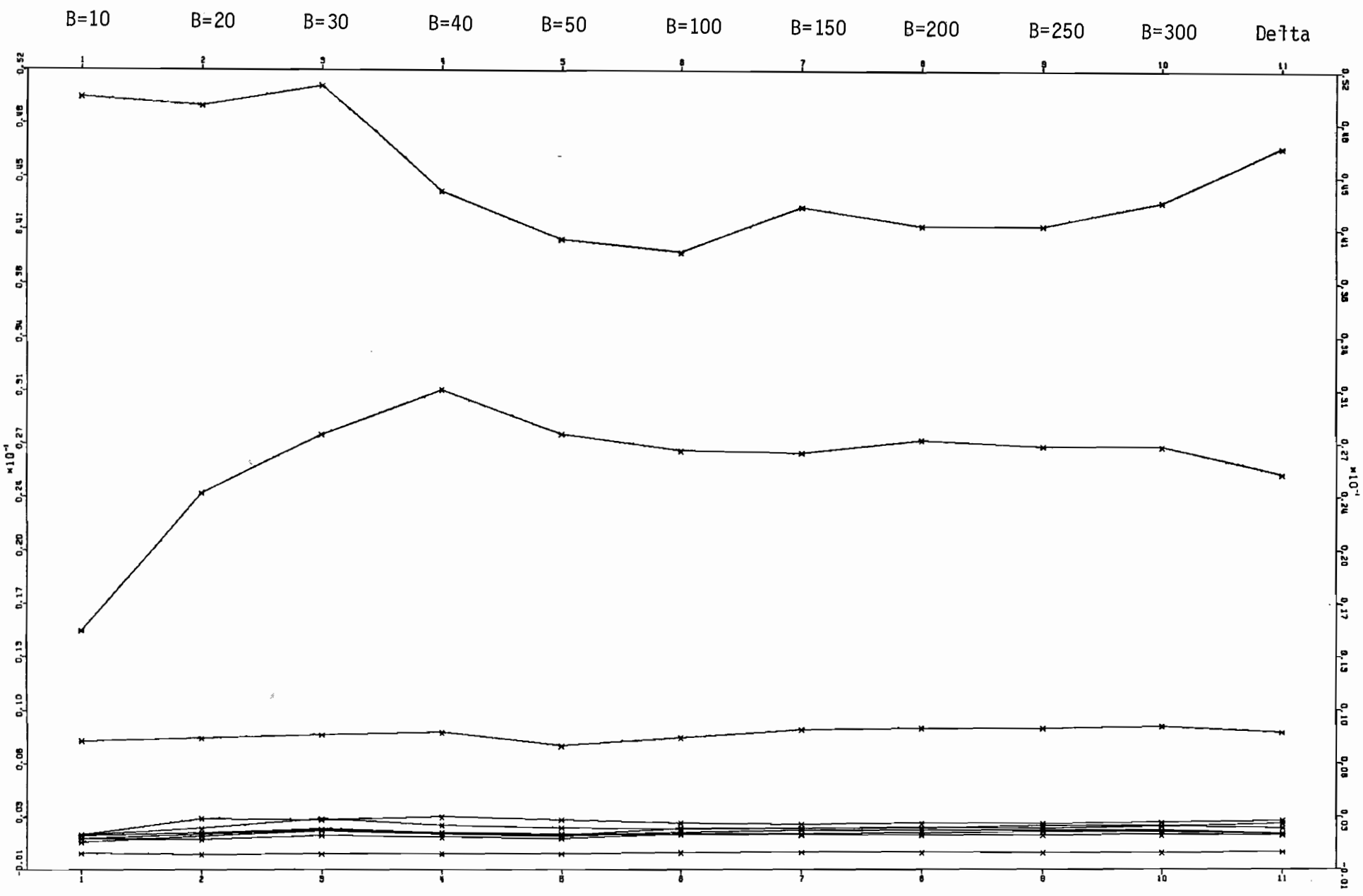

Figure 1. Variances for each category for the delta and the bootstrap method with different numbers of samples

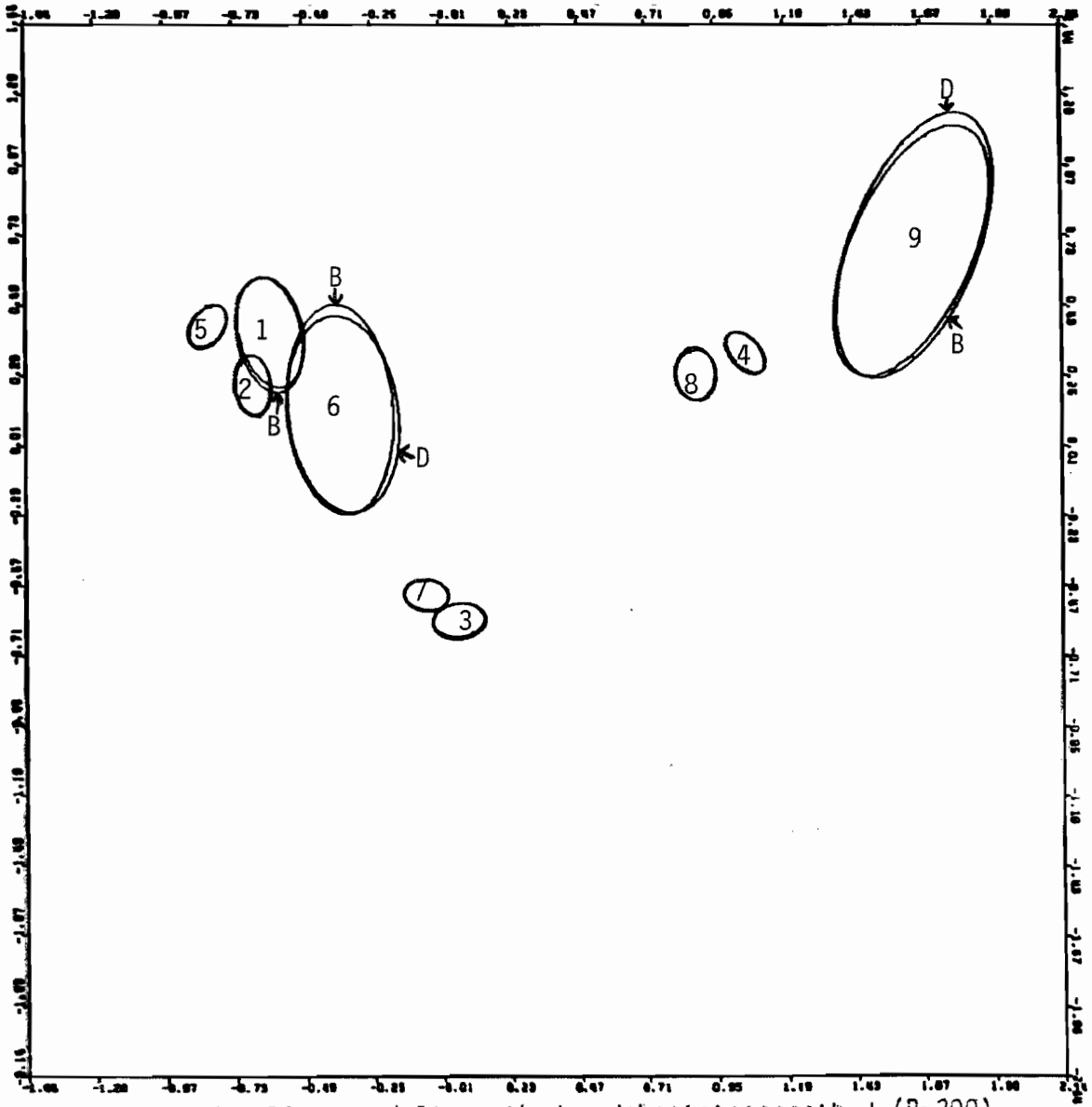

Figure 3. 95% ellipses: delta method and bootstrap method (B=300)

1 - BLUE EYES	2 - LIGHT EYES	3 - MEDIUM EYES
4 - DARK EYES	5 - FAIR HAIR	6 - RED HAIR
7 - MEDIUM HAIR	8 - DARK HAIR	9 - BLACK HAIR

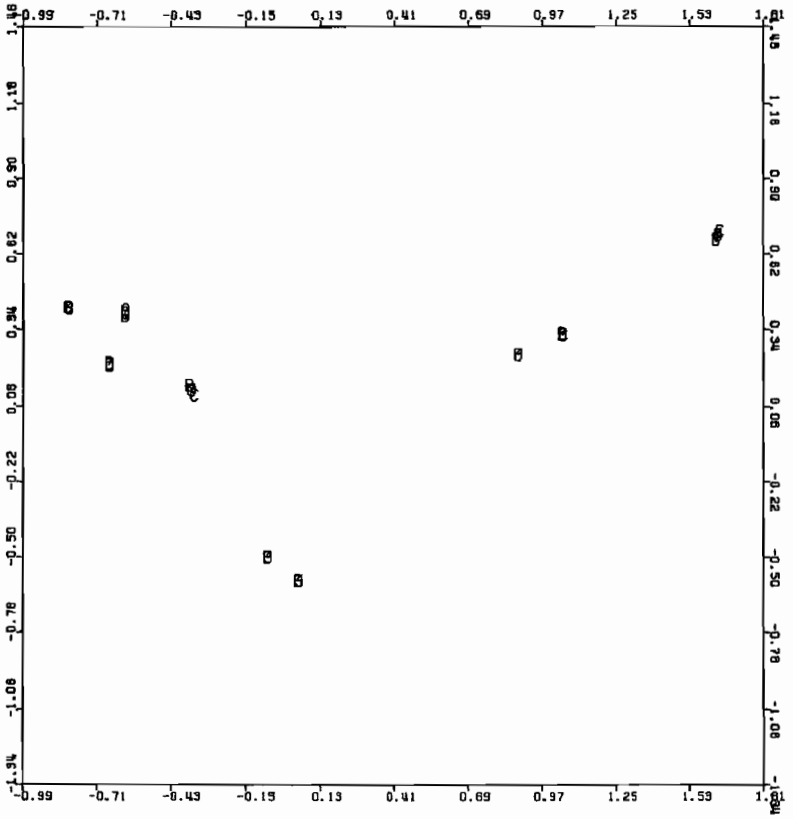

Figure 3. Bias correction. O=original, B=mean bootstrap, C=correction

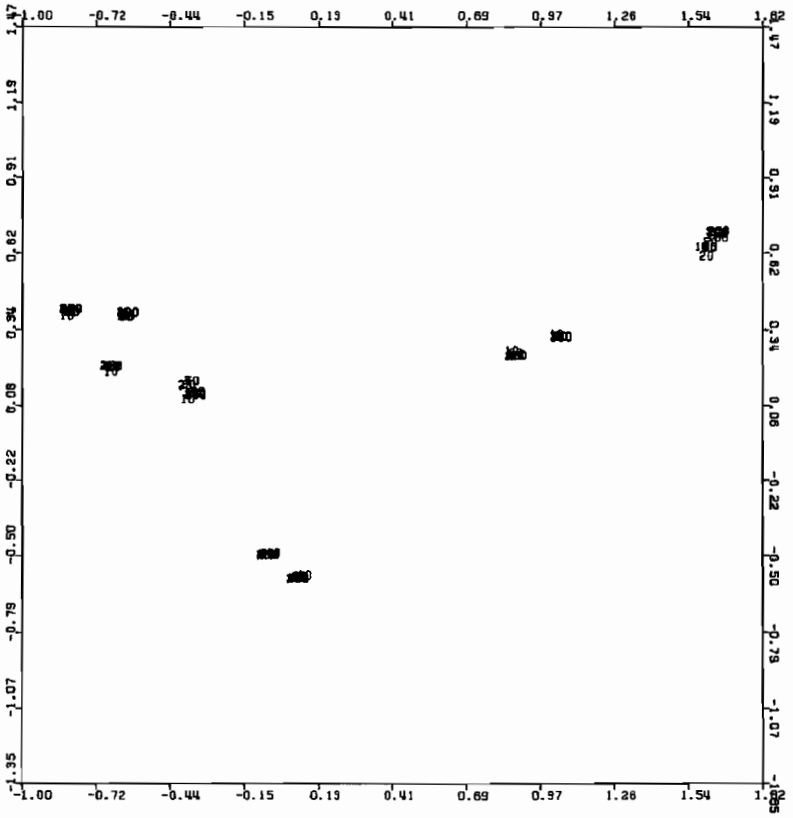

Figure 4. Bias correction for B=10, 20, 30, 40, 50, 100, 150, 200, 250 and 300

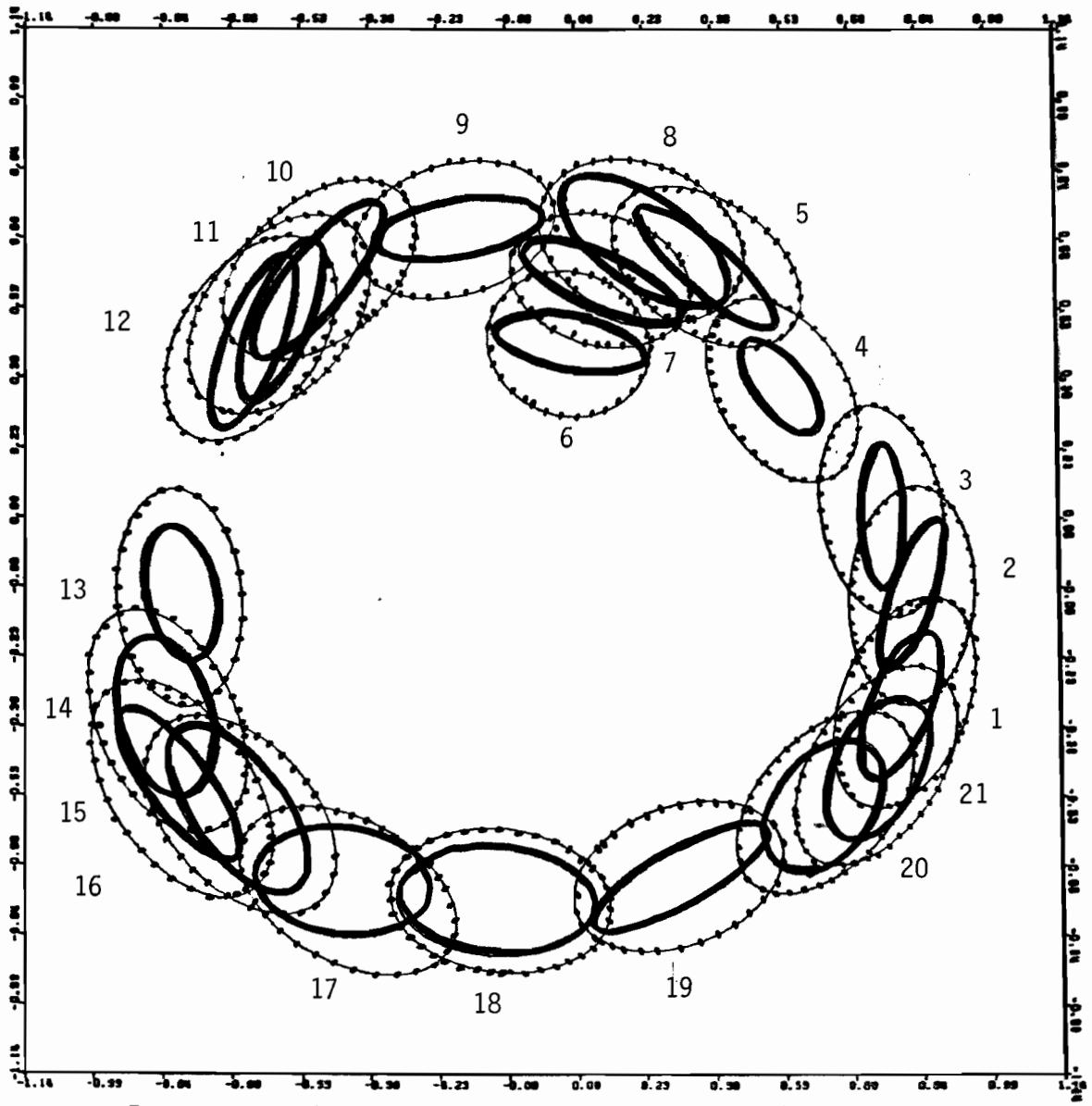

Figure 5. 95% confidence ellipses: bootstrap (solid) & delta method (dotted).

1-5R	2-10R	3-5YR	4-10YR	5-5Y	6-10Y2	7-10Y4
8-10Y8	9-5GY	10-10GY	11-5G	12-10G	13-10BG	14-5B
15-10B	16-5PB	17-10PB	18-5P	19-10P	20-5RP	21-10RP

R = Red G = Green Y = Yellow B = Blue P = Purple

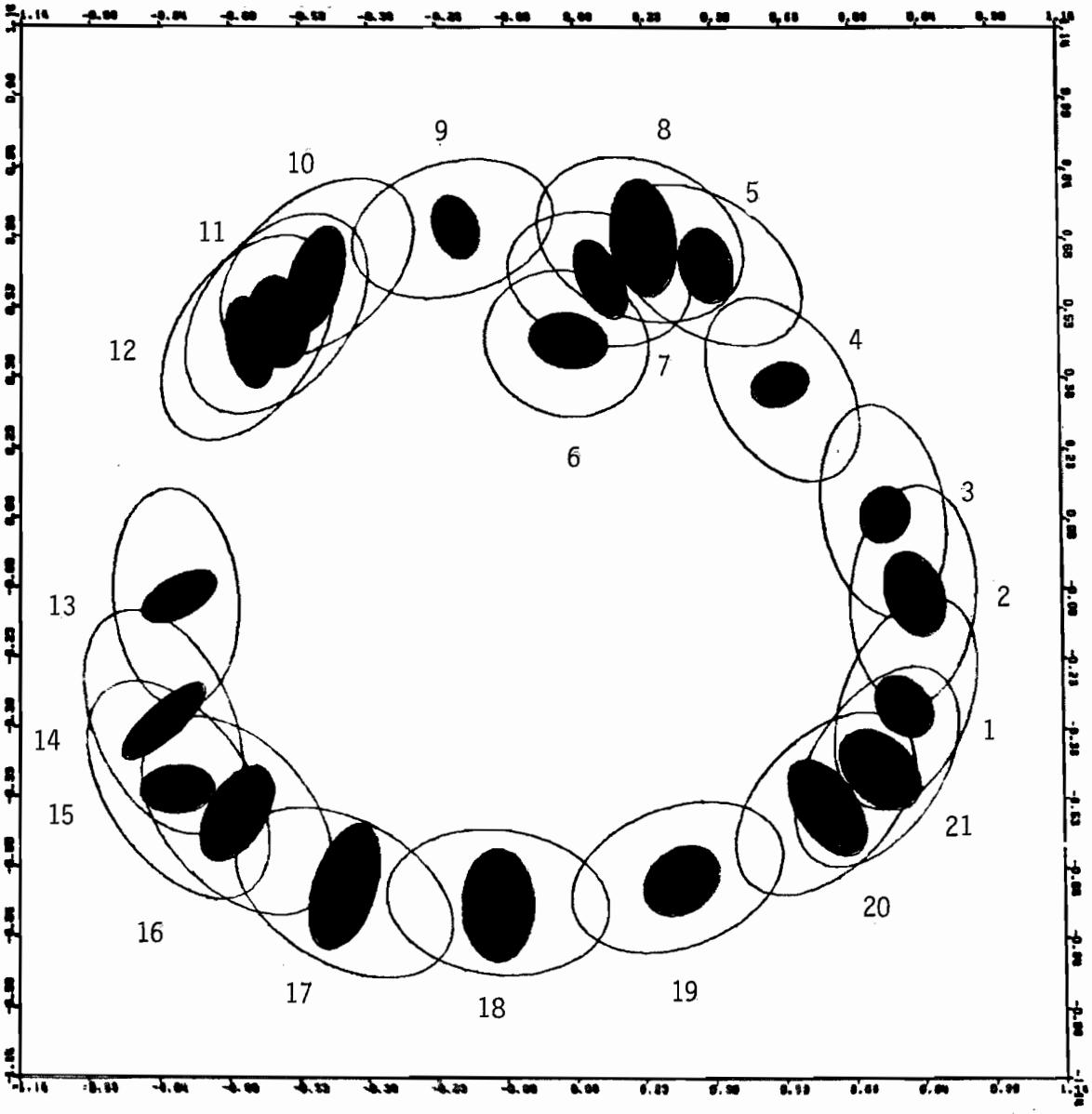

Figure 6. 95% confidence ellipses: delta method & bootstrap after orthogonal procrustes rotation (filled up).