Journal of Classification 2:285-287 (1985)

R. Gittens, *Canonical Analysis. A review with Applications in Ecology*. Springer-Verlag, Berlin - Heidelberg - New York - Tokyo, 1985, pp. 351.

In many of the sciences relating two sets of variables is a familiar problem. Ecologists, for example, study the relationship between soil characteristics of areas and occurrence or abundance of plants and animals. Educational researchers are interested in the relationship between background characteristics of pupils and their achievements in school. Biochemists want to relate structural characteristics of macromolecules with their biological activity. Similar examples can be found in political science, economy, demography, and so on.

All such problems have a common mathematical structure. A variable is a function on the space of units, and a set of variables defines a subspace in the linear space of such functions. If we have two sets of variables, then we have two subspaces of a common linear space, and we want to find out if and how these subspaces are related. The mathematical invariants of a pair of subspaces are the angles between them, and these angles, or rather their cosines, are known as the canonical correlation coefficients in data analysis. The canonical correlation coefficients are product moment correlations between specific elements in the subspaces, which are called canonical variates. The canonical variates are linear combinations of the original variates defining the subspaces, coefficients of the original variables in these linear combinations are called canonical weights. And finally, the canonical variates can also be described by correlating them with the original variables. The resulting correlations are called canonical loadings. Computing canonical correlation coefficients, variates, weights, and loadings is known as canonical correlation analysis (CCA).

The technique is based on geometrical insights which are both elegant and simple, and it leads to matrix calculations which are straightforward and not very expensive. Nevertheless CCA is not very popular. In most of the sciences in which it has been tried, the users have run into interpretational problems, and into problems with stability of the results. It was found, typically, that no satisfactory answers to the original research questions were possible on the basis of the CCA results. Thus we are in the situation that an elegant technique is available, but there are not many successful applications, and there is a fairly widespread belief that the technique is not really useful. Accounts of CCA are often fragmentary and superficial, and reasons for its alleged uselessness and failures are not discussed clearly.

The book by Gittens takes this unsatisfactory state of affairs as its starting point, and succeeds in making various very positive contributions to the

Reviewer's address: Jan de Leeuw, Department of Data Theory FSW/RUL, Rijksuniversiteit te Leiden, Middelstegracht 4, 2312 TW Leiden, THE NETHERLANDS.

field. On the one hand the book gives a rather exhaustive account of the mathematical and statistical aspects of CCA, but on the other hand a very important part of the book deals with the in-depth analysis of seven ecological examples using CCA. The mathematical/statistical part of the book is rather routine. Its main asset is its encyclopedic nature. Gittens has looked at a very large number of publications, also from the more applied literature. The emphasis is, of course, on ecology, but this is hardly a restriction of generality because the research problems are so very similar in the various sciences. The mathematics behind CCA is very elementary, and not much can be said about it. Whatever can be said about it, is repeated in Gittens, sometimes more than once. The same thing is true for the statistical and the computational aspects of CCA. Thus the book is enormously useful as a reference volume. Results from the literature are copied rather uncritically, however, which sometimes leads to curious statements. Two papers are mentioned, for instance, which discuss reasons why computed canonical correlations are sometimes larger than unity. Departures from normality and negative unique variances are suggested as possible reasons.

The enormous list of facts and ideas about CCA is made even longer by including a rather complete coverage of the literature dealing with various extensions. Many extensions are possible, of course. Dealing with more than two sets of variables is an obvious one, making the technique non-metric is another. Introducing asymmetry in the treatment of the two sets is also very important in many applications. Gittens tries to review all these developments, which are often quite recent. The book is very similar to an annotated bibliography in these sections, with few comments and virtually no criticism. Useful, no doubt, but rather tough going. A very good idea is to include chapters on canonical variate analysis (better known as multiple discriminant analysis) and on dual scaling (better known as correspondence analysis). The introduction to the mathematics of these special cases of CCA is satisfactory, and the two chapters end again with very impressive lists of references.

Although the encyclopedic coverage, and the more than 600 references, are certainly very useful, the most interesting parts of the book (at least for me) are the various suggestions on how to interpret the results, on how the results can be presented graphically, on auxiliary statistics, and on the study of stability and validity of the technique. These more practical concepts are applied, with great expertise, in part II of the book, which has seven real examples from ecology. Both treatment and presentation of the examples are excellent. Data are given in full, the relevant ecological content is explained, and results are always assessed in terms of ecological revenues. There is an example of correspondence analysis, of multiple discriminant analysis, and of multivariate analysis of variance. I am not an ecologist, but it seems to me that Gittens' examples clearly show that CCA can be very useful if it is applied with insight and creativity. The book neatly summarizes the strong points of the technique as they emerge from the seven analyses.

The book ends with a chapter on research issues and future developments. These developments are mainly of a statistical character, in the general sense of the word. Thus, re-expression (or transformation), robustness, stability, collinearity, residuals, and outliers are discussed, using such modern methods as Bootstrap, Jackknife, Optimal Scaling, Cross Validation, and the EM-algorithm. Again, the emphasis is on reviewing the literature, which is now extended to cover general statistical techniques that could be profitably applied in CCA. The tendency is again to cover, or at least mention, everything that is, or could be, relevant. The chapter is useful, but heavy going in various places.

Let me summarize my impression of the book. It is very useful in at least three respects. In the first place as a reference work, an annotated bibliography of canonical analysis. In the second place because of its many useful hints for interpreting and presenting CCA results. And thirdly because of its nice examples, and their exemplary analysis. The examples illustrate much more clearly than equations or methodological discussion are able to do, that canonical analysis and its various generalizations and extensions are very useful data analysis techniques, in ecology and presumably also elsewhere. The book is not a critical review of the literature, and it is rather naive from the mathematical-statistical point of view. On the whole, however, I heartedly recommend it to everyone concerned with multivariate data analysis and with the graphical representation of data matrices.

Rijksuniversiteit te Leiden Jan de Leeuw