The collapsibility theorem in log-linear analysis of categorical data: an application in program evaluation

PEDRO SÁNCHEZ-ALGARRA & M. TERESA ANGÜERA-ARGILAGA

1Department of Statistics Faculty of Biology, University of Barcelona, 08028 Barcelona, Spain
2Department of Behavioral Science Methodology, Faculty of Psychology, University of Barcelona, 08035 Barcelona, Spain

Abstract. The collapsibility theorem describes both the circumstances in which the effects of hierarchical models change when additional variables are introduced, as the circumstances in which the exclusion of certain variables and the analysis of specific marginal tables may lead to different conclusions.

The partial association model is here considered as a specific example of three-dimensional log-linear analysis.

Collapsibility is examined in an empirical study currently being performed in Catalonia with regard to program evaluation in penitentiary centers.

Key words: collapsibility, log-linear analysis.

Log-linear analysis has gained wide acceptance in behavioral and social sciences as the method preferably used for the analysis of categorical data. An interesting feature of log-linear models is that they may represent a solution to a long-standing problem in program evaluation research, as has been shown by Goodman in a number of articles. In addition, log-linear models provide a natural means of dealing with multivariate frequency tables.

In recent years, the use of log-linear analysis became the solution for a number of problems connected with analyzing cross-tabular data, e.g. contingency tables containing void cells. Other problems that have recently appeared include the question of fitting a model (Plackett 1969), sparse data and weighted data (Clogg & Eliason 1988), and collapsibility (Goodman 1969, 1970, 1987; Bishop 1971; Bishop et al. 1975; Fienberg 1981; Hagaenaars 1990), a problem encountered in high-dimensional data analysis. The difficulties that most frequently arise in regard with the analysis are the questions of classifying criteria and the possibility of modifying categorical variables, which would involve a change in their significance.

In principle, log-linear effects take on different values when they are
modified or when other variables are introduced. The effects in the marginal tables are usually different from the corresponding effect in the full table. The collapsibility theorem informs scientists about the way in which the behavior of hierarchical models changes when additional variables are being introduced, and also about the circumstances in which the exclusion of certain variables and the analysis of specific marginal tables lead to different conclusions.

Bishop et al. (1975: 47) formulated the collapsibility theorem as follows:

Suppose the variables in a multidimensional table are divided into three mutually exclusive groups. One group is collapsible with respect to the \( \tau \)-terms involving a second group, but not with respect to the \( \tau \)-terms involving the third group, if and only if the first two groups are independent of each other (that is, the \( \tau \)-terms linking them are 1).

It is important to establish which interactions change when the data-matrix is reduced by condensing or collapsing a variable. A variable may be defined as “collapsible”, if no interactions are changed. This notion results in the procedure of selecting those log-linear models that may be fitted directly, without interaction (Bishop 1971) and such a procedure consists of examining the effects of variables which successively collapse. Since, in the case of such models, rightness of fit can be measured without computing cell estimates, it is useful to identify those models and assist scientists in choosing strategies which will help to establish the best-fitting model.

A specific example will be considered, the partial association model. This is the most highly parametrized model that can be fitted directly. It is used currently in procedures to obtain information from two-way tables and also to compute adjusted rates (Bishop 1971).

An attempt will be made to summarize in three dimensions the data and basic notions regarding the problems which arise here, although more dimensions are involved. If only the sample size, \( N \), is fixed, the distribution is constituted by the following multinomial equation:

\[
\frac{N!}{\prod_{i,j,k} \frac{m_{ijk}^{x_{ijk}}}{N}}
\]

where \( m_{ijk} = E(x_{ijk}) \) is positive.

Following the notation of Birch (1963), we write the following equation:

\[
\log m_{ijk} = u + u_{i(1)} + u_{j(2)} + u_{k(3)} + u_{12(0)} + u_{13(0)} + u_{23(0)} + u_{123(0)}.
\]

The \( u \)-terms are deviations, as in the analysis of variance and sum to zero over each included variable. The terms with two numerical subscripts, \( u_{12}, \)