Interaction Model

Interaction models for **categorical data** are **loglinear models** describing association among categorical variables. They are called interaction models because of the analytic equivalence of loglinear **Poisson regression** models describing the dependence of a count variable on a set of categorical explanatory variables and loglinear models for **contingency tables** based on **multinomial** or product multinomial sampling. The term is, however, somewhat misleading, because the interpretation of parameters from the two types of models are very different. *Association models* would probably be a better name.

Instead of simply referring the discussion of interaction and association models to the section on loglinear models, we will consider these models from the types of problems that one could address in connection with analysis of **association**. The first problem is a straightforward question of whether or not variables are associated. To answer this question, one must first define association and dissociation in multivariate frameworks and, secondly, define multivariate models in which these definitions are embedded. This eventually leads to a family of so-called graphical models that can be regarded as the basic type of interaction or association. The second problem concerns the properties of the identified associations. Are associations homogeneous or heterogeneous across levels of other variables? Can the strength of association be measured and in which way? To solve these problems, one must first decide upon a natural measure of association among categorical variables and, secondly, define a parametric structure for the interaction models that encapsulates this measure. Considerations along these lines eventually lead to the family of hierarchical loglinear models for nominal data and models simplifying the basic loglinear terms for **ordered categorical data**.

Graphical Interaction Models

What is meant by association between two variables? The most general response to this question is indirect. Two variables are dissociated if they are *conditionally* independent given the rest of the variables in the multivariate framework in which the two variables are embedded. Association then simply means that the two variables are not dissociated.

Association in this sense is, of course, not a very precise statement. It simply means that conditions exist under which the two variables are not independent. Analysis of association will typically have to go beyond the crude question of whether or not association is present, to find out what characterizes the conditional relationship – for instance, whether it exists only under certain conditions, whether it is homogeneous, or whether it is modified by outcomes on some or all the conditioning variables. Despite the inherent vagueness of statements in terms of unqualified association and dissociation, these statements nevertheless define elegant and useful models that may serve as the natural first step for analyses of association in multivariate frames of inference. These so-called *graphical* models are defined and described in the subsections that follow.

Definition

A graphical model is defined by a set of assumptions concerning pairwise conditional independence given the rest of the variables of the model.

Consider, for instance, a model containing six variables, A to F. The following set of assumptions concerning pairwise conditional independence defines four constraints for the joint distribution $\Pr(A, B, C, D, E, F)$. The family of probability distributions satisfying these constraints is a graphical model:

$$\begin{aligned} A \perp C|BDEF &\Leftrightarrow \Pr(A, C|BDEF) \\ &= \Pr(A|BDEF)\ \Pr(C|BDEF), \\ A \perp D|BCEF &\Leftrightarrow \Pr(A, D|BCEF) \\ &= \Pr(A|BCEF)\ \Pr(D|BCEF), \\ B \perp E|ACDF &\Leftrightarrow \Pr(B, E|ACDF) \\ &= \Pr(B|ACDF)\ \Pr(E|ACDF), \\ C \perp E|ABDF &\Leftrightarrow \Pr(C, E|ABDF) \\ &= \Pr(C|ABDF)\ \Pr(E|ABDF). \end{aligned}$$

Interaction models defined by conditional independence constraints are called “graphical interaction models”, because the structure of these models can be characterized by so-called interaction graphs, where

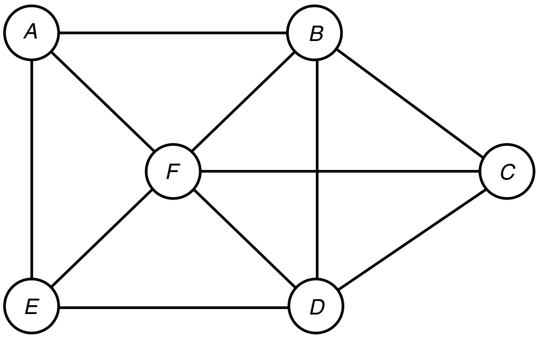

Figure 1 An interaction graph

variables are represented by nodes connected by undirected edges if and only if association is permitted between the variables. The graph shown in Figure 1 corresponds to the set of conditional independence constraints above, because there are no edges connecting A to C, A to D, B to E, and C to E.

Interaction graphs are visual representations of complex probabilistic structures. They are, however, also mathematical models of these structures, in the sense that one can describe and analyze the interaction graphs by concepts and **algorithms** from mathematical graph theory and thereby infer properties of the probabilistic model. This connection between probability theory and mathematical graph theory is special to the graphical models.

The key notion here is conditional independence, as discussed by Dawid [5]. While the above definition requires that the set of conditioning variables always includes all the other variables of the model, the results described below imply that conditional independence may sometimes be obtained if one conditions with certain subsets of variables.

Graphical models for multidimensional tables were first discussed by Darroch et al. [5]. Since then, the models have been extended both to continuous and mixed categorical and continuous data and to regression and block recursive models. Whittaker [9], Edwards [7], Cox & Wermuth [4], and Lauritzen [8] present different accounts of the theory of graphical models. The sections below summarize some of the main results from this theory.

The Separation Theorem

The first result connects the concept of graph separation to conditional independence.

First, we present a definition: a subset of nodes in an undirected graph separate two specific nodes, A and B, if all paths connecting A and B intersect the subset. In Figure 1, (B, D, F) separate A and B, as does (B, E, F). E and C are separated by both (A, D, F) and (B, D, F).

The connection between graph separation and conditional independence is given by the following result, sometimes referred to as the separation theorem.

Separation Theorem. If variables A and B are conditionally independent given the rest of the variables of a multivariate model, A and B will be conditionally independent given any subset of variables separating A and B in the interaction graph of the model.
The four assumptions on pairwise conditional independence defining the model shown in Figure 1 generate six minimal separation hypotheses:

$$A \perp C|BDF, \quad A \perp C|BEF, \quad A \perp D|BEF,$$
$$B \perp E|ADF, \quad C \perp E|ADF, \quad C \perp E|BDF.$$

Closure and Marginal Models

It follows from the separation theorem that graphical models are closed under marginalization, in the sense that some of the independence assumptions defining the model transfer to marginal models.

Collapsing, for instance, over variable C of the model shown in Figure 1 leads to a graphical model defined by conditional independence of A and D and B and E, respectively, because the marginal model contains separators for both AD and BE (Figure 2).

Loglinear Representation of Graphical Models for Categorical Data

No assumptions have been made so far requiring variables to be categorical. If all variables are categorical, however, the results may be considerably strengthened both with respect to the type of model defined by the independence assumptions of graphical models and in terms of the available information on the marginal models.

The first published results on graphical models [5] linked graphical models for categorical data to loglinear models:

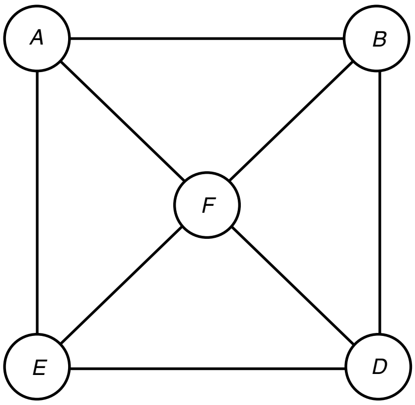

Figure 2 An interaction graph obtained by collapsing the model defined by Figure 1 over variable C

> A graphical model for a multidimensional contingency table without **structural zeros** is loglinear with generators defined by the cliques of the interaction graph.

The result is an immediate result of the fact that any model for a multidimensional contingency table has a loglinear expansion. Starting with the saturated model, one removes all loglinear terms containing two variables assumed to be conditional independent. The loglinear terms remaining after all the terms relating to one or more of the independence assumptions of the model have been deleted define a hierarchial loglinear model with parameters corresponding to each of the completely connected subsets of nodes in the graph.

The interaction graph for the model shown in Figure 1 has four cliques, $BCDF$, ABF, AEF, and DEF, corresponding to a loglinear model defined by one four-factor interaction and three three-factor interactions.

Separation and Parametric Collapsibility

While conceptually very simple, graphical models are usually complex in terms of loglinear structure. The problems arising from the complicated parametric structure are, however, to some degree to be compensated for by the properties relating to collapsibility of the models.

Parametric collapsibility refers to the situation in which model terms of a complete model are unchanged when the model is collapsed over one or more variables. Necessary conditions implying parametric collapsibility of loglinear models are described by Agresti [1, p. 151] in terms which translate into the language of graphical models:

> Suppose variables of a graphical model of a multidimensional contingency table are divided into three groups. If there are no edges connecting variables the first group with connected components of the subgraph of variables from the third group, then model terms among variables of the first group are unchanged when the model is collapsed over the third group of variables.

Parametric **collapsibility** is connected to separation in two different ways. First, parametric collapsibility gives a simple proof of the separation theorem, because a vanishing two-factor term in the complete model also vanishes in the collapsed model if the second group discussed above contains the separators for the two variables. Secondly, separation properties of the interaction graph may be used to identify marginal models permitting analysis of the relationship between two variables. If one first removes the edge between the two variables, A and B, and secondly identifies separators for A and B in the graph, then the model is seen to be parametric collapsible on to the model containing A and B and the separators with respect to all model terms relating to A and B.

The results are illustrated in Figure 3, where the model shown in Figure 3(a) is collapsed on to marginal models for $ABCD$ and $CDEF$. The separation theorem is illustrated in Figure 3(b). All terms relating to A and B vanish in the complete model. The model satisfies the condition for parametric collapsibility, implying that these parameters also vanish in the collapsed model. The second property for the association between E and F is illustrated in Figure 3(c). C and D separate E and F in the graph from which the EF edge has been removed. It follows, therefore, that E and F cannot be linked to one and the same connected component of the subgraph for the variables over which the table has been collapsed. The model is therefore parametric collapsible on to CDEF with respect to all terms pertaining to E and F.

Decomposition and Reducibility

Parametric collapsibility defines situations in which inference on certain loglinear terms may be performed in marginal tables because these parameters

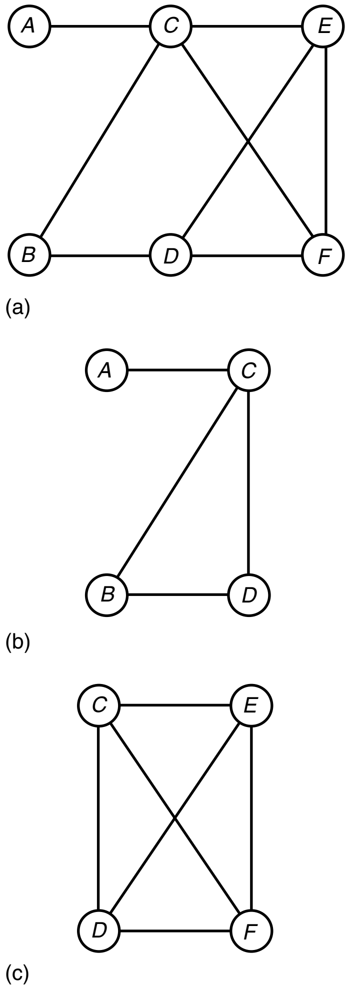

Figure 3 Collapsing the model given in (a) illustrates the separation theorem for A and B (b), and parametric collapsibility with respect to E and F (c)

are unchanged in the marginal tables. Estimates of, and test statistics for, these parameters calculated in the marginal tables will, however, in many cases differ from those obtained from the complete table. Conditions under which calculations give the same results may, however, also be stated in terms of the interaction graphs.

An undirected graph is said to be *reducible* if it partitions into three sets of nodes – X, Y, and Z – if Y separates the nodes of X from those of Z and if the nodes of Y are completely connected. If the interaction graph meets the condition of reducibility, it is said to decompose into two components, $X + Y$ and $Y + Z$. The situation is illustrated in Figure 4, which decomposes into two components, $ABCD$ and $CDEF$.

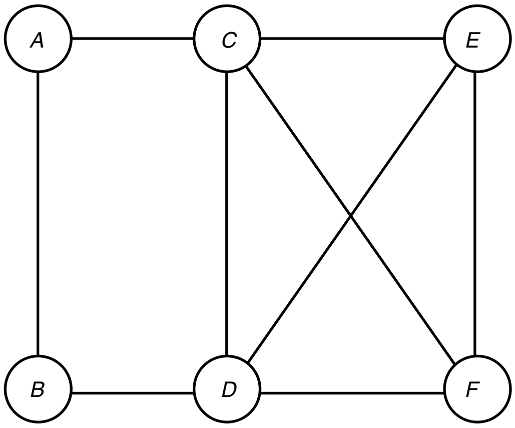

Figure 4 An interaction graph of a reducible model

It is easily seen that reducibility above implies parametric collapsibility with respect to the parameters of X and Z, respectively. It can also be shown, however, that likelihood-based estimates and test statistics obtained by analysis of the collapsed tables are exactly the same as those obtained from the complete table.

Regression Models and Recursive Models

So far, the discussion has focused on models for the joint distribution of variables. The models can, however, without any problems, be extended first to multidimensional regression models describing the conditional distribution of a vector of dependent variables given another vector of **explanatory variables** and, secondly, to block recursive systems of variables. In the first case, the model will be based on independence assumptions relating to either two dependent variables or one dependent and one independent variable. In the second case, recursive models have to be formulated as a product of separate regression models for each recursive block conditionally given variables in all prior blocks. To distinguish between symmetric and asymmetric relationships edges between variables in different recursive blocks, interaction graphs are replaced by arrows.

Parametric Structure: Homogeneous or Heterogeneous Association

The limitations of graphical models for contingency tables lie in the way in which they deal with higher-order interactions. The definition of the graphical models implies that higher-order interactions *may* exist if more than two variables are completely connected.

It is therefore obvious that an analysis of association by graphical models can never be anything but the first step of an analysis of association. The graphical model will be useful in identifying associated variables and marginal models where associations may be studied, but sooner or later one will have to address the question of whether or not these associations are homogeneous across levels defined by other variables and, if not, which variables modify the association. The answer to the question of homogeneity of associations depends on the type of measure that one uses to describe or measure associations. For categorical data, the natural measures of association are measures based on the so-called cross product ratios [2] (*see* **Odds Ratio**). The question therefore reduces to a question of whether or not cross product ratios are constant across different levels of other variables, thus identifying loglinear models as the natural framework within which these problems should be studied.

Ordinal Categorical Variables

In the not unusual case of association between ordinal categorical variables, the same types of argument apply against the hierarchical loglinear models as against the graphical models. Loglinear models are basically interaction models for nominal data; and, as such, they will give results that are too crude and too imprecise for ordinal categorical data. The question of whether or not the association between two variables is homogeneous across levels of conditioning variables can, for ordinal variables, be extended to a question of whether or not the association is homogeneous across the different levels of the associated variables. While not abandoning the basic loglinear association structure, the answer to this question depends on the further parameterization of the loglinear terms of the models. We refer to a recent discussion of these problems by Clogg & Shihadeh [3].

Discussion

The viewpoint taken here on the formulation of interaction models for categorical data first defines the family of graphical models as the basic type of models for association and interaction structure. Loglinear models are, from this viewpoint, regarded as parametric graphical models, meeting certain assumptions on the nature of associations not directly captured by the basic graphical models. Finally, different types of models for ordinal categorical data represent yet further attempts to meet assumptions relating specifically to the ordinal nature of the variables.

References

[1] Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.

[2] Altham, P.M.E. (1970). The measurement of association of rows and columns for an $r \times s$ contingency table, *Journal of the Royal Statistical Society, Series B* **32**, 63–73.

[3] Clogg, C. & Shihadeh, E.S. (1994). *Statistical Models for Ordinal Variables*. Sage, Thousand Oaks.

[4] Cox, D.R. & Wermuth, N. (1996). *Multivariate Dependencies. Models, Analysis and Interpretation*. Chapman & Hall, London.

[5] Darroch, J.N., Lauritzen, S.L. & Speed, T.P. (1980). Markov fields and log-linear models for contingency tables, *Annals of Statistics* **8**, 522–539.

[6] Dawid, A.P. (1979). Conditional independence in statistical theory, *Journal of the Royal Statistical Society, Series B* **41**, 1–15.

[7] Edwards, D. (1995). *Introduction to Graphical Modelling*. Springer-Verlag, New York.

[8] Lauritzen, S.L. (1996). *Graphical Models*. Oxford University Press, Oxford.

[9] Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester.

Svend Kreiner