

TAMÁS RUDAS – WICHER BERGSMA

On applications of marginal models for categorical data

Summary - The paper considers marginal models for categorical data and after reviewing the most important theoretical results concerning the definition, estimation and testing of such models, discusses a number of common statistical problems. These examples include, among others, the analysis of repeated measurements, panel studies and missing data. Fitting marginal models in these cases has the potential of providing the researcher with substantial new insight. The examples illustrate that the marginal modeling approach may be used more widely than thought before. One of the examples shows how graphical models associated with directed acyclic graphs can be parameterized. A general algorithm is presented to compute maximum likelihood estimates under marginal models.

Key Words - missing.

1. INTRODUCTION

During the past decade, a fair number of papers applying marginal models to medical (Balagtes, Becker, Lang, (1995), Molenberghs, Lesaffre, (1999)) and sociological (Becker, (1994), Becker, Minick, Yang, (1998)) data, parallel to papers exploring components of the theory of marginal modeling (Lang, Agresti, (1994), Glonek, McCullagh, (1995), Bergsma, (1997), Lang, McDonald, Smith, (1999), Colombi, Forcina, (2001), Bartolucci, Forcina, Dardadoni, (2001), Bartolucci, Forcina, (2002), Bergsma, Rudas, (2002a), Bergsma, Rudas, (2002b)) have been published. In its most general form, a marginal model, when applied to a multivariate statistical problem, imposes structural restrictions on certain marginals (i.e., subsets) of the original variables. When the variables are categorical, the models for the marginals are usually of the log-linear or of the log-affine type. Such models are most conveniently formulated by restricting the values of appropriately defined parameters. Therefore,

Received October 2003 and revised December 2003.

the existence, flexibility and interpretability of marginal models depend largely on the parameters that are used to formulate the model.

The present paper, based on recent theoretical developments (Bergsma, Rudas, (2002a)), illustrates the applicability of a large class of marginal models to a variety of statistical problems. This class of models is based on restricting the values of certain marginal parameters of the joint distribution in a contingency table. This is a flexible class of parameters, that generalizes earlier approaches to define marginal parameters (Glonek, McCullagh, (1995), Glonek, (1996), Kauermann, (1997)), and certain combinatorial properties of the variables involved imply smoothness of the parameterization and variation independence of its components. These properties are essential in interpretation, imply the existence of a large class of models and the applicability of standard large sample theory for estimation and testing.

The present paper contains almost no proofs. Section 2 gives a somewhat informal exposition of the theory referred to above and Section 3 considers applications of marginal models to a number of common statistical problems. These include measuring the effect of a treatment, panel studies and Markov chains, data fusion, missing data, joint treatment of the sampling and statistical models, and graphical models. It is not the goal of the present paper to explore the analyses made possible by the marginal approach in any depth, rather, the aim is to illustrate that fitting marginal models and the interpretation of carefully defined parameters may yield new insight into the above problems and, often, appears to be the appropriate strategy. Finally, Section 4 describes an algorithm to fit marginal log-linear or log-affine models. Much of the theory and applications discussed in the present paper extend in a natural way to problems involving continuous data, but these generalizations will not be considered here.

2. THEORY

The class of marginal models applied in this paper is based on marginal log-linear parameters. These are obtained as ordinary log-linear parameters (Agresti, (1990)) but they are not computed from the entire contingency table, rather from a marginal of it. A marginal log-linear parameter therefore, is characterized by two subsets of the variables, one to which we first marginalize and a subset of this one, to which the parameter applies. For example, for variables A, B, C, D , λ_{i**}^{ABC} is a marginal log-linear parameter. The marginal which it pertains to is ABC , and this is shown in the superscript. Within the ABC marginal, the parameter represents the log-linear effect of category i of variable A . Note that the ordinary log-linear parameter of the variable A in category i is λ_i^A which, as a marginal log-linear parameter is denoted by λ_{i***}^{ABCD} , as it is computed from the entire $ABCD$ table, not from a marginal of it. In this

paper, only marginal log-linear parameters are considered that is, the superscript always refers to the marginal from which the parameter is computed.

The usual log-linear parameters can be interpreted (Bishop, Fienberg, Holland, (1975)) as measuring average conditional association among the variables involved, conditioned on all other variables and then average taken over all possible categories of the conditioning variables. Every marginal log-linear parameter pertains to a certain marginal and the average conditional association is measured within this marginal. For example, when all the variables are binary,

$$\lambda_{1**}^{ABC} = \frac{1}{4} \sum_j \sum_k \log(p_{1jk+}/p_{2jk+})^{1/2},$$

where p_{ijkl} is a cell probability, either theoretical or observed or estimated, and “+” is a marginalization operator. That is, λ_{1**}^{ABC} is related to the average conditional log odds of category 1 of A versus category 2, conditioned on and averaged over all categories of B and C , in the ABC marginal of the $ABCD$ table. Similarly,

$$\lambda_{11*}^{ABC} = \frac{1}{2} \sum_k \log \left(\frac{p_{11k+}p_{22k+}}{p_{12k+}p_{21k+}} \right)^{1/4},$$

that is, the marginal log-linear parameter λ_{ij*}^{ABC} of the $ABCD$ table is related to the average conditional log odds ratio between A and B , conditioned on and averaged over C , after marginalization over D . Throughout the paper, positivity of the cell frequencies is assumed.

As a common way to refer to the various possibilities, the subset in the superscript of a marginal log-linear parameter will be called the marginal and the variables whose indices appear in the subscript will be called the effect which is measured by the parameter.

The marginal parameters defined here include the ordinary log-linear parameters (those parameters which have the set of all variables as the relevant marginal), the multivariate logistic parameters of Glonek, McCullagh (1995) (those parameters for which the effect variables coincide with the marginal variables) and a mixture of these considered by Glonek (1996).

The marginal log-linear parameters can be used in several different ways to parameterize the distribution on a contingency table. The parameter selection can be done in two steps. The substantive problem at hand determines which marginals of the contingency table are of interest. In the first step, arrange these marginals in a hierarchical ordering, i.e. in such a way that no marginal contains one which comes later in the sequence. It is easy to see that such an ordering always exists. If a certain rule is followed when selecting the subsets that are effects within the marginals then, as it will be seen later, the resulting parameters will have desirable properties. This rule says that for every marginal,

only such subsets of it should be included as an effect that are not subsets of any of the previous marginals. Marginal log-linear parameters defined by this rule are called hierarchical.

For example, if for three variables A , B , C , the marginals of interest are AB and BC , a hierarchical ordering is

$$AB \ BC.$$

Then, in the second step, for the AB marginal, the marginal log-linear parameters may pertain to the effects \emptyset , A , B , AB , and for the BC marginal, the effects can be C and BC . Thus, one possible set of hierarchical marginal log-linear parameters implied by the above ordering of the relevant marginals contains the following parameters:

$$\lambda_{**}^{AB}, \lambda_{i*}^{AB}, \lambda_{ij}^{AB}, \lambda_{jk}^{BC}.$$

The following are also hierarchical marginal log-linear parameters, based on the same ordering

$$\lambda_{**}^{AB}, \lambda_{i*}^{AB} \lambda_{*j}^{AB}, \lambda_{*k}^{BC} \lambda_{jk}^{BC}.$$

Notice however, that the above parameters are not complete, i.e. they do not constitute a parameterization (see below). If the other hierarchical ordering of the relevant marginals,

$$BC \ AC,$$

is selected, the resulting parameters may be as follows:

$$\lambda_{**}^{BC}, \lambda_{j*}^{BC}, \lambda_{*k}^{BC}, \lambda_{jk}^{BC}, \lambda_{i*}^{AB} \lambda_{ij}^{AB}.$$

As is seen above, the possible choices of the hierarchical marginal log-linear parameters depend on the ordering of the relevant marginals. For example, λ_{j*}^{BC} is allowed in the second ordering but not in the first one. Note that hierarchy of these parameters refers to a property of the ordering of the marginals that determines which parameters are allowed, not to the choice of the effects within the marginals (as is the case with classical hierarchical log-linear models).

A set of hierarchical marginal log-linear parameters can be completed to a parameterization of the distribution on the contingency table. To do so, the list of marginals has to be completed by adding the entire set of variables as the last one in the hierarchical ordering and as new parameters, those have to be included that pertain to effects not present, for the marginal where it is first possible (Bergsma, Rudas, (2002a)). For example, the second set of parameters above can be completed as

$$\lambda_{**}^{AB}, \lambda_{i*}^{AB} \lambda_{*j}^{AB}, \lambda_{ij}^{AB}, \lambda_{*k}^{BC} \lambda_{jk}^{BC}, \lambda_{i*k}^{ABC}, \lambda_{ijk}^{ABC}.$$

Note that for simplicity, parameterization refers here to the parameterization of a frequency (rather than probability) distribution and for every parameter includes all linearly independent choices of the indices, i.e. for binary variables every parameter in the above list refers to one value. To obtain a parameterization of a probability distribution, the main effect (i.e., λ_{**}^{AB} above) has to be omitted.

The ordinary log-linear parameterization and the one based on the multivariate logistic transform (Glonek, McCullagh, (1995)) are both hierarchical marginal log-linear parameterizations.

Now certain desirable properties of marginal log-linear parameters and of the statistical models derived from them will be studied. Note that these properties extend to the hierarchical marginal log-linear parameterization if parameters with these properties are completed by the above procedure to yield a parameterization.

The properties studied will be smoothness of parameters, variation independence, existence of log-linear or log-affine marginal models defined by restricting a set of parameters and the applicability of standard large sample theory to these models. Proofs of the results to follow can be found in Bergsma, Rudas (2002a).

Smoothness of the parameters considered essentially means a one-to-one and differentiable correspondence between the vector of parameters and the vector of cell probabilities. Smoothness is important in the interpretation of the parameters and in studying the dimension of a model which, in turn, is crucial for testing the fit of the model.

Theorem 1. *Hierarchical marginal log-linear parameters are smooth for strictly positive frequency distributions on the contingency table.*

The above result establishes only a sufficient condition of smoothness of marginal log-linear parameters but it can be shown that if the same effect appears among marginal log-linear parameters within different marginals, then these parameters cannot be smooth.

The next property to consider is variation independence of the parameters. Variation independence means that the joint range of the parameters is the Cartesian product of the separate ranges of the parameters involved. Lack of variation independence may lead to the definition of non-existing (empty) models and makes the separate interpretation of the parameters misleading. To illustrate the importance of variation independence of the parameters, consider the following marginal log-linear parameters and their prescribed values for three binary variables:

$$\begin{aligned}\lambda_*^A &= \log 8, \lambda_1^A = 0, \lambda_1^B = 0, \lambda_1^C = 0, \\ \lambda_{11}^{AB} &= (1/4) \log(1/9), \lambda_{11}^{AC} = (1/4) \log(1/9), \lambda_{11}^{BC} = (1/4) \log(9).\end{aligned}$$

The above prescribed values are all within the ranges of the respective parameters. In spite of this, these values are not within the combined range of the parameters, that is, no distribution exists with these parameters. To see this, notice that the prescriptions imply that there are 8 observations, the one-way marginals are uniform (4, 4) and the first two two-way marginals have an odds ratio equal to 1/9, while the third two-way marginal has an odds ratio equal to 9. This completely specifies the two-way marginals of the table, but they are not compatible: there is no (non-negative) three-way table with these two-way marginals. This can be seen either by establishing a contradiction implied by the assumptions or by considering the correlation matrix and establishing that it is not positive definite. The parameters involved in this case are not variation independent and the definition above specifies a non-existing distribution or, the prescriptions define an empty model. Note, that for this example of potential contradiction, neither the specification of the value of λ_*^A nor the multipliers of (1/4) were necessary.

To see how the lack of variation independence makes the separate interpretation of the parameters invalid, consider a simple 2×2 treatment by outcome experiment in two groups say, men and women. Suppose the following data are observed:

Outcome	Treatment	Control
good	10	5
bad	40	45
Men		
Outcome	Treatment	Control
good	30	20
bad	20	30
Women		

If the measure of the effect of the treatment is the difference in proportion of positive outcome among treated and among control, then this measure is .1 for men and .2 for women. Is then the treatment twice as effective for women than for men, as the numerical values suggest? The answer is, of course, not necessarily, because, given the marginals, the maximum value of this measure is .3 for men and 1 for women. That is, the treatment is one third as useful

for men and only one fifth as useful for women as it could be. The measure of treatment effect used here is not variation independent from the marginal distributions therefore, it lacks calibration and cannot be interpreted without paying attention to the other parameters.

To assure variation independence of hierarchical marginal log-linear parameters, a generalization of the classical decomposability concept is needed. An ordering of a class of incomparable marginals is decomposable (Haberman, (1974), Lauritzen, Speed, Vijayan, (1984)) if it consists of two subsets only or every subset has the property that its intersection with the union of the previous subsets is equal to its intersection with one of the previous subsets. A hierarchical ordering of subsets consisting of t marginals is ordered decomposable if for every $3 \leq u \leq t$, the maximal ones from among the first u subsets have a decomposable ordering.

For example AB, BC, ABC is ordered decomposable but AB, BC, AC, ABC is not. Ordered decomposability in fact does depend on ordering. For variables A, B, C, D , the ordering AB, BC, ABC, ACD is ordered decomposable but the ordering AB, BC, ACD, ABC is not.

Theorem 2. *The components of a hierarchical marginal log-linear parameterization are variation independent if and only if the ordering of the marginals involved is ordered decomposable.*

Marginal log-linear parameters derived from marginals in a hierarchical and ordered decomposable order will be called hierarchical and ordered decomposable marginal log-linear parameters.

In the sequel, statistical models defined by restrictions on marginal log-linear parameters will be considered. In this context, the parameters pertain to the expectations of cell frequencies under Poisson sampling. A log-linear marginal model is defined by assuming that certain linear combinations of marginal log-linear parameters are equal to zero. Such models are never empty, as they always contain the uniform distribution. A log-affine marginal model is defined by assuming that certain linear combinations of marginal log-linear parameters are equal to given constants. Examples of such models will be considered in the next section.

The existence of log-affine marginal models is, in general, a difficult question. In fact, the example used above to illustrate the importance of variation independence is a log-affine marginal model which is empty, i.e. does not exist. The following result shows that the conclusion suggested by that example is true in general.

Theorem 3. *A log-affine marginal model defined by restrictions of variation independent parameters is not empty.*

This implies that log-affine marginal models based on ordered decomposable hierarchical marginal log-linear parameters always exist i.e., include at least one distribution.

The last desirable property we discuss here, is the applicability of standard large sample theory. This is, again, not as straightforward as everyday statistical practice may appear to suggest. For example, consider a $2 \times 2 \times 2$ table and the model assuming that $\lambda_{11}^{AB} = 0$, $\lambda_{111}^{ABC} = 0$, $\lambda_{11*}^{ABC} = 0$. The first condition specifies marginal independence of variables A and B , and the last two imply that A and B are conditionally independent given C . Dawid (1980) showed that the three assumptions imply that either A is independent of both B and C jointly or B is independent of both A and C jointly or both of these hold true. In the latter case, however large the sample is, the likelihood has, with positive probability, local maxima on both branches of the model and the likelihood ratio statistic is, asymptotically, the minimum of two chi-squared distributions rather than having asymptotic chi-squared distribution.

If however, the model is based on appropriately selected marginal log-linear parameters, the standard asymptotic theory applies.

Theorem 4. *Suppose a non-empty log-affine marginal model is based on smooth parameters. Then, under Poisson or multinomial sampling*

- a. *The probability that the maximum likelihood estimate $\hat{\pi}$ of the true probability π exists and is a stationary point of the likelihood equation tends to 1, as the sample size goes to infinity.*
- b. *The asymptotic distribution of $N^{1/2}(\hat{\pi} - \pi)$ is normal, with zero expectation, where N is the sample size.*
- c. *The likelihood ratio statistic has an asymptotic chi-squared distribution with the number of degrees of freedom being equal to the number of linearly independent restrictions.*

If the log-affine marginal model is based on hierarchical ordered decomposable marginal log-linear parameters, then the above asymptotic results hold true. In other situations, standard asymptotic theory may or may not apply, depending on the true population parameters.

3. APPLICATIONS

Once in possession of the theoretical results outlined above, one finds several important statistical problems where marginal log-linear or log-affine models may be applied. In this section, some of these situations are reviewed. Here, we formulate the relevant marginal models and investigate their properties using the results given in the previous section. Issues related to estimation of these models will be considered in the next section.

3.1. Repeated measurements

One of the most widely used experimental designs in the medical and behavioral sciences to measure the effect of a treatment is to observe the same individuals before and after it. In this design, the variables observed before and after the treatment are related because they are observed on the same individuals. If the variables are categorical, the observations before and after the treatment should be considered as marginals of the same contingency table (Hagenaars, (1990)). We now outline some potentially useful repeated measurements models and use the theory of the previous section to show that these models are well-behaved.

If the same characteristic is measured before (variable A) and after (variable D) treatment and the hypothesis is that the distributions before and after treatment are the same (no effect of the treatment), the statistical model in the $A \times D$ table is defined by

$$\lambda_i^A = \lambda_i^D, \quad \text{for all } i.$$

This is a log-linear marginal model and the marginals involved have a hierarchical and ordered decomposable ordering, e.g., A, D . It immediately follows that the model exists and that standard large sample theory applies.

In fact, this is the model of marginal homogeneity.

When two variables are measured before (A, B) and also two after (D, E) the treatment, an interesting model assumes that A and B are independent and D and E are independent. Note that this model may be meaningful whether the same or different characteristics are measured before and after treatment. This model assumes that

$$\lambda_{ij}^{AB} = 0, \lambda_{lm}^{DE} = 0, \quad \text{for all } i, j \quad \text{and } l, m.$$

Here, the marginals involved are AB, DE and they are ordered decomposable. The related hierarchical marginal log-linear parameters – that may be arbitrarily restricted – are $\lambda_{**}^{AB}, \lambda_{i*}^{AB}, \lambda_{*j}^{AB}, \lambda_{ij}^{AB}, \lambda_{l*}^{DE}, \lambda_{*m}^{DE}, \lambda_{lm}^{DE}$. Therefore, the model is a marginal log-linear model based on hierarchical and ordered decomposable parameters, and consequently standard large sample theory applies to this model. A log-affine marginal model based on the same parameters which is relevant here, is the one assuming that the ratio of the marginal odds ratios, as measures of association, between D and E and between A and B is equal to a specified constant or, equivalently, that the difference of association, as measured by log-linear parameters is equal to a specified value. This leads to the model specified by $\lambda_{11}^{AB} - \lambda_{11}^{DE} = c$. This model exists for all c and standard large sample theory applies.

If there are several variables measured before and after the treatment, arbitrary linear or affine assumptions about some of the marginal log-linear parameters pertaining to effects within the marginal of the before-treatment variables and about some of those pertaining to an effects within the marginal of after-treatment variables will obey standard asymptotic theory, if the model is not empty. The latter condition holds in the linear case and in the affine case it holds if the effects for both marginals are decomposable.

There are however, more general cases covered by the available theory. Restrictions considering the association between before and after treatment variables can also be included. For example, if there are three variables A , B and C measured before and three variables D , E and F measured after the treatment, the model with the following restrictions

$$\lambda_{ijk}^{ABC} = 0, \lambda_{lmn}^{DEF} = 0, \lambda_{i**l}^{ABCD} = 0, \lambda_{ij*l}^{ABCD} = 0, \lambda_{i*kl}^{ABCD} = 0, \lambda_{ijkl}^{ABCD} = 0, \\ \text{for all } i, j, k, l, m \text{ and } n$$

means that there is no second order association among the before-treatment variables and among the after-treatment variables and A and D are conditionally independent, given the other before treatment variables. This model can be obtained by restricting the parameters derived from the ordering of marginals ABC , $ABCD$, DEF . Since this ordering is hierarchical and ordered decomposable, the model exists and standard asymptotic theory applies.

All this extends in a natural way to designs where several measurements are taken over the same individuals, either with certain treatments applied between the measurements or time passing by between the measurements.

A further application of marginal models is needed when measurements on the same characteristic are taken repeatedly to reduce the effect of measurement error as for example, blood pressure of a person may be measured on three consecutive days, different tests of the same mental ability may be administered to the same person, or different questions measuring the same attitude may be included in a questionnaire. Suppose, A_1 and A_2 are the first and second measurements of the same characteristic, and B_1 and B_2 are the first and second measurements of another characteristic. Then, A_1 and A_2 should be essentially (disregarding error related to imprecise measurement or to temporary fluctuations of the quantity measured) identical, just like B_1 and B_2 . In addition to certain marginal homogeneity restrictions, this would also imply that the association between A_i and B_j is the same for every combination of $i, j = 1, 2$. This leads to the following model:

$$\lambda_i^{A_1} = \lambda_i^{A_2}, \lambda_j^{B_1} = \lambda_j^{B_2}, \lambda_{i_1 i_2}^{A_1 A_2} = r_{i_1 i_2}(A), \lambda_{j_1 j_2}^{B_1 B_2} = r_{j_1 j_2}(B), \\ \lambda_{ij}^{A_1 B_1} = \lambda_{ij}^{A_1 B_2} = \lambda_{ij}^{A_2 B_1} = \lambda_{ij}^{A_2 B_2},$$

for all possible values of the indices, where $r_{i_1 i_2}(A)$ and $r_{j_1 j_2}(B)$ represent the strength of association between the first and second measurements of the characteristics A and B , respectively, and should be selected based on the magnitude and distribution of error which is usually present, or acceptable, when those measurements are performed. For this model, standard asymptotic theory holds.

3.2. Panel studies and Markov chains

In this setup, again, the same individuals are observed several times on the same variables however, interest lies not so much in whether or not the distributions in the different time points are identical or different rather, the pattern of change is of interest. A frequently investigated hypothesis is that of a k -th order Markov chain that is, the conditional distribution of the variables measured at time point t depends only on the positions at the k preceding time points. This is, of course, a log-linear model but the related hypotheses discussed below are of the marginal type.

The estimation of the transition probabilities are often among the goals of the analysis of panel data. Parallel to the Markov hypothesis, one may be interested in modeling whether or not the distributions in the previous waves of the panel influence the association between the distributions in the last two waves. The pattern of association between the distributions in the t -th and $t - 1$ -st time points can be captured by the conditional odds ratios or log-linear parameters of the joint distribution of the variables measured at these two time points. If this only depends on the distribution at the $t - k - 1$ -st, \dots $t - 2$ -nd time points, the process generating the data has a k -th order memory.

Therefore, to test the hypothesis of a first order memory against that of a second order memory, one needs at least four waves of the panel. In this case, the hypothesis that one has a first order memory, given that the memory is of second order (saturated in the present case) can be formulated as

$$\lambda_{**i_3 i_4}^{A_1 A_2 A_3 A_4} = \lambda_{*i_3 i_4}^{A_2 A_3 A_4}, \lambda_{*i_2 i_3 i_4}^{A_1 A_2 A_3 A_4} = \lambda_{i_2 i_3 i_4}^{A_2 A_3 A_4}, \lambda_{i_1 i_2 i_3 i_4}^{A_1 A_2 A_3 A_4} = 0,$$

where A_i denotes the variable(s) measured at the i -th time point. This model asserts that the association between A_3 and A_4 depends only on A_2 and not on A_1 . The association is measured by the appropriate marginal log-linear parameters (or, equivalently by the appropriate marginal conditional odds ratios). The conditions imply that the marginal log-linear parameters (or, equivalently, the marginal conditional odds ratios) are the same if conditioned on A_2 only or on both A_2 and A_1 . This is a collapsibility condition (Whittemore, (1980)) and is also a marginal log-linear model. The marginal log-linear parameters in it are not contained in any hierarchical marginal log-linear parameterization

because, for example, the $\{A_3, A_4\}$ effect appears in two marginals. Therefore, the statistical properties of this model cannot be obtained from the results of this paper. In fact, it can be shown (Bergsma, Rudas, (2002a)) that the above parameters cannot be parts of a smooth marginal log-linear parameterization and the Jacobian of any marginal log-linear parameterization containing the above parameters is singular at the uniform distribution. Note, that the same applies to any similar collapsibility restriction (see also Davis, (1989)).

If the process is known to have a, say, one step memory, testing stationarity (with respect to conditional association between neighbors) requires fitting the following model:

$$\lambda_{*jk}^{A_1A_2A_3} = \lambda_{*jk}^{A_2A_3A_4}, \lambda_{jkl}^{A_1A_2A_3} = \lambda_{jkl}^{A_2A_3A_4},$$

for every j, k and l , if four waves are available. The model says that the conditional association between A_2 and A_3 when A_1 is given is the same as that between A_3 and A_4 when A_2 is given. This is a marginal log-linear model, hierarchical, ordered decomposable and standard large sample theory applies. The related log-affine marginal model, in which the above marginal log-linear parameters have prescribed values (for example, as in small area estimation), also exists and has standard large sample behavior.

3.3. Incomplete data

There are various statistical problems requiring the analysis of an incomplete set of data. Incomplete data may arise unintentionally or intentionally, in surveys or in censuses, in data collection or in secondary analysis problems.

The most common source of unintentionally missing data due to data collection is that some of the respondents in a survey or census fail to respond to certain questions in a questionnaire (item nonresponse), or to the entire questionnaire (unit nonresponse). The problem of coverage error (parts of the population being omitted from the sample frame) leads to incomplete data similar to unit nonresponse. The information collected is intentionally incomplete, when, to reduce the burden of the respondents, with respect to time and invasion of privacy, a long questionnaire is split into shorter overlapping parts and every respondent is only asked questions in one of the parts. Such split designs may be applied both in surveys and censuses. In secondary data analysis it may happen that no available data set contains all the necessary information and the researcher has to rely on several previously collected sets of data. This leads to a problem similar to analyzing data arising as a result of a split design, with the additional problem that the separate sampling procedures behind the separate sets of data make even the existence of a common underlying population distribution questionable.

When the data are categorical, the available information, in all the above cases, can be considered as being marginal distributions of a higher dimensional contingency table (the one that would contain all variables of interest). Depending on the actual circumstances, the distribution on the entire table (the complete data) would apply to the entire population or to a sample from it or may not exist at all. The first step of the analysis in all these cases is to find out whether such a joint distribution exists and if several such distributions exist, select one according to some optimality criterion. Depending on the circumstances, such a procedure may be called an extension of measures, estimation or data fusion.

Notice that this scheme also covers model-based estimation of the joint distribution, when the sufficient statistics are certain marginal distributions, like, e.g., with log-linear models. Here, the information is not incomplete in the sense that the entire table may have been observed but only certain aspects (the sufficient statistics) are relevant for further analysis.

If the distributions on an incomparable (with respect to inclusion) set of marginals are given and they are weakly compatible that is, they coincide on the intersections of the marginals, decomposability implies that there always exists an extension (in fact, usually infinitely many) and if the system is not decomposable, it depends on the actual marginals whether or not weak compatibility implies strong compatibility (Darroch, Lauritzen, Speed, (1980), Kellerer, (1964)). This classical theory however, does not cover cases when information with respect to a more complex system of marginals is available and the results of this paper are relevant. If some of the marginals for which observations are available are contained in each other, classical decomposability and the extension procedure based on it are of no help. This is the case, among others, in the common missing data situation when there are respondents who actually did respond to all questions, implying that observation not only for some marginals but also for the entire table are available. Note that our approach here to handling missing data problems is fundamentally different from the standard approach based on imputation techniques (Little, Rubin, (1987)).

In this more general case, the following procedure may be applied. Consider all marginals for which information is available and their intersections. Order these hierarchically and construct the hierarchical marginal log-linear parameterization. Determine the values of those parameters, for which this is possible using the given information. If for a certain marginal different sources of information are available, for example both A and AB are observed, consider a pooled estimate for the distribution on A . Set those marginal log-linear parameters for which no information is available to arbitrary values, for example to zero. Then, as described in Bergsma, Rudas (2002a), a generalization of the iterative proportional scaling algorithm can be used to reconstruct the entire distribution.

To illustrate the procedure, assume that for a three-way table, observations are available for the AB , AC , ABC marginals. Adding intersections and putting the marginals in a hierarchical order yields A , B , AB , AC , ABC .

The distribution on the A marginal is estimated by pooling data from all three original distributions, the B marginal is estimated by pooling data from the AB and ABC marginals. The odds ratios (Rudas, (1998a)) in the AB marginal are estimated by pooling the original AB and ABC data sets. That is, estimates for the one way marginals and the odds ratios of the AB marginal table or, equivalently, estimates of the marginal log-linear parameters λ_i^A , λ_j^B and λ_{ij}^{AB} , are obtained and combining these by the iterative scaling procedure yields our estimate for the AB marginal distribution.

Next, the distribution of the AC marginal is obtained by taking into account the already estimated A marginal distribution and the conditional distribution of C given A which is obtained by pooling data from the original AC and ABC marginals (yielding the λ_{*k}^{AC} and λ_{ik}^{AC} marginal log-linear parameters).

Finally, to estimate the ABC marginal, the already estimated AB and AC marginal distributions are combined with the conditional distribution of B , given A and C , which is taken from the original ABC marginal (i.e. the λ_{*jk}^{ABC} , λ_{ijk}^{ABC} marginal log-linear parameters are estimated).

The procedure may not yield a joint distribution but if the marginals (including intersections) have an ordered decomposable ordering, just like in the present example, there will always be a common extension to the marginals.

In the following example, a certain part of the information available needs to be discarded. Suppose that one is interested in reconstructing or estimating the joint distribution of variables A , B , C . The AB marginal was observed in a simple random sample, and the AC marginal in a sample which was stratified according to A . But the stratification in the latter data collection procedure was based on information which may not be reliable, for example outdated census data.

In this situation, one would use the AB sample to estimate the joint distribution of these two variables (disregarding the A marginal in AC) and the AC sample to estimate the C marginal and the interaction between A and C . That is, the information with respect to the distribution of A is taken entirely from the AB sample. Therefore, estimates are available for the following marginal log-linear parameters:

$$\lambda_{i*}^{AB}, \lambda_{*j}^{AB}, \lambda_{ij}^{AB}, \lambda_{*k}^{AC}, \lambda_{ik}^{AC}.$$

Because these marginal log-linear parameters are ordered decomposable, there always exists a three dimensional distribution with these parameters.

Note however, that if, additionally, observations are also available on the joint distribution of BC , no component of that, not even the association between B and C is guaranteed to be strongly compatible (i.e., yielding a joint

distribution) with the information obtained from the first two samples, because ordered decomposability is lost.

In the present example, the marginal log-linear parameters which cannot be estimated from the data pertain to the AB and ABC effects. Therefore, to be able to estimate the joint distribution, in order to have hierarchy, either λ_{ij}^{AB} or λ_{ij*}^{ABC} and λ_{ijk}^{ABC} need to be given certain values. While the most straightforward assumption is that these marginal log-linear parameters are equal to zero, this assumption will have different implications depending on the choice of parameters to which it is applied. If the parameters selected are λ_{ij}^{AB} and λ_{ijk}^{ABC} , then assuming they are equal to zero means that A and B are marginally independent, while if the same assumption is applied to λ_{ij*}^{ABC} and λ_{ijk}^{ABC} , then A and B are conditionally independent, given C . If it is only the true response to a question that decides whether or not the response is given, the observed values for A and the observed values for B both are random samples from their respective distributions in the first case. In the second case, this is only true within fixed categories of C .

3.4. Joint treatment of the sampling and statistical models

A statistical model may be viewed as a subset of the possible distributions and a statistical hypothesis assumes that the true distribution belongs to this subset. When the model is parametric, it restricts some of the parameters of the distribution. The restricted parameters need to be estimated from the data in such a way that the resulting estimates fulfill the requirements of the model and the parameters not restricted by the model are estimated from the data without further restrictions. A sampling model, on the other hand, assuming finite population size, assigns probabilities to the possible samples (subsets of the population), often in a way that it excludes certain samples from consideration. In many practical situations this implies specifying certain parameters of the observed distribution (e.g., as in stratified sampling some of the marginals are kept fixed). Then, these parameters should not be estimated from the data, even if the statistical model, without consideration of the sampling model, would call for estimating these parameters. Rather, the estimates should only be sought among distributions fulfilling both the model and the sampling restrictions.

Therefore, the resulting model is the intersection of the statistical and of the sampling model. Considering *any* model of interest as being the intersection of other (possibly simpler) models may prove useful both from a conceptual and from a computational point of view, as it was illustrated for log-linear models by Rudas (1998b, 2002).

The parameter estimates obtained under the combined restrictions are the estimates in the statistical model with the sampling model taken into account. Many of the popular sampling models restrict the values of certain marginal

log-linear parameters and if the statistical model is also a marginal model, the above combination can be carried out easily and the relationship between the two models becomes apparent, while with other approaches potential conflicts may not be easy to recognize.

As a first example, consider simple random sampling with fixed sample size N . From all possible samples (subsets of the population) only those of size N are considered and no further restriction applies (that is, these samples have equal probabilities). This restriction is equivalent to $\lambda^\emptyset = \log N$. When a marginal log-linear or log-affine statistical model is estimated with this sampling scheme, the combined model is a log-affine marginal model. If the statistical model is independence in a two-way table, the joint restrictions are

$$\lambda^\emptyset = \log N, \lambda_{ij}^{AB} = 0, \text{ fo all } i, j,$$

and this is a log-affine marginal model. It is easy to see that if the statistical model is defined by log-linear or log-affine restrictions on a hierarchical marginal log-linear parameterization and the overall effect λ^\emptyset does not appear among those restricted by the statistical model, then adding the multinomial constraint $\lambda^\emptyset = \log N$ does not affect the properties of the model.

As another example, consider a case-control study, where cases (e.g. patients to be given a certain treatment) enter the study as a result of a process not under the control of the experimenter, but the design calls for selecting one control person for every case by a certain procedure. The status of both cases and controls is measured before any treatment is applied and after the treatment was applied. Here, the design specifies that the case-control marginal is uniform, while the total, the status marginal and the association between the two variables are unrestricted. That is, if A is the case-control variable, $\lambda_i^A = 0$.

In the case of stratified sampling, the distribution of a (group of) variable(s) is fixed by sampling design. If the frequency of variable A is fixed, say $N_i > 0$ in category i , this is equivalent to $\lambda_*^A + \lambda_i^A = \log N_i$ and this restriction should be added to the restrictions imposed by the statistical model.

To illustrate the possible conflicts that may arise, assume now that the variables A , B , C are observed and the statistical model of interest prescribes the marginal distribution of C and the marginal odds ratios of AC and of BC . If the available data are obtained from a stratified sample, where stratification prescribed the AB marginal and the stratification was based on reliable information concerning the distribution of the AB marginal (for example, a recent census), then the combination of the sampling and statistical models prescribes the AB marginal distribution and the AC and BC marginal odds ratios and the further parameters of the distribution are to be estimated. This is a log-affine marginal model and as the parameters are hierarchical but not ordered decomposable therefore, depending on the actual values of the parameters, it

may be empty. But the parameters are smooth and therefore, if the model is nonempty, standard asymptotic theory applies to this model.

The advantage of the marginal modelling approach is that the combinatorial properties of the class of parameters restricted by either one of the models decides the statistical properties of the resulting model.

3.5. Graphical models

Graphical log-linear models (Darroch, Lauritzen, Speed, (1980)) use graphs to model the association structure of multivariate distributions. The nodes of the graph are identified with the variables involved, and two variables not connected by an edge are assumed to be conditionally independent, given all other variables. These are conditional independence statements involving all variables. Models pertaining to the joint distribution of variables are also associated with directed acyclic graphs (Lauritzen, (1996)). In this case, a variable is assumed to be conditionally independent from its nondescendants, given its parents, where nondescendants are those nodes into which no directed path leads from the variable and parents are the nodes from which arrows points to the variable. Graphical models based on directed acyclic graphs therefore, assume conditional independencies which do not involve all variables rather, certain marginal distributions pertaining to subsets of the variables.

Consequently, graphical models based on directed acyclic graphs are marginal log-linear models. For example, consider the directed acyclic graph in Figure 1.

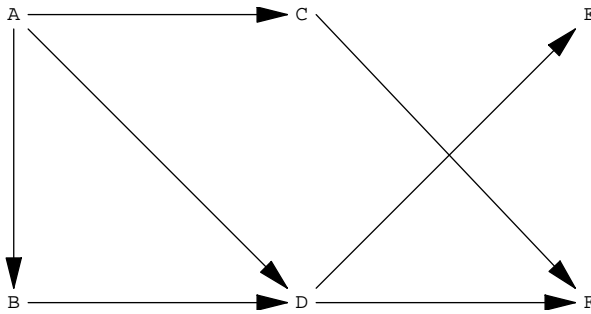


Figure 1. A directed acyclic graph.

This graph implies the following conditional independencies:

$$C \perp\!\!\!\perp BDE | A$$

$$D \perp\!\!\!\perp C | AB$$

$$E \perp\!\!\!\perp ABCF | D$$

$$F \perp\!\!\!\perp ABE | CD.$$

The model defined by the above graph or, equivalently, by the above conditional independencies is a marginal model. It will be illustrated now how the marginal log-linear parameters of this model can be used to parameterize the distributions in it. The parameters involved are associated with the arrows in the graph defining the model and their values can be given intuitively appealing interpretations.

The parameterization of graphical log-linear models defined by directed acyclic graphs is based on the factorization of the distributions in such models given in Lauritzen (1996). The factorization involves functions depending only on subsets of the variables which consist of a node and its parents:

$$p(\omega) = \prod_{\alpha \in \mathcal{V}} f_{\alpha}(\omega_{\{\alpha\} \cup pa(\alpha)}),$$

where ω is a cell of the table, \mathcal{V} is the set of variables forming the table, $pa(\alpha)$ is the set of variables that are parents of α and for $\mathcal{W} \subseteq \mathcal{V}$, $\omega_{\mathcal{W}}$ is a projection operator.

The subsets entering the factorization for the above example are, in a hierarchical order, the following:

$$A, AB, AC, ABD, DE, CDF,$$

where a node is always preceded by its parent(s). The hierarchical marginal log-linear parameters are the following:

$$\begin{aligned} & \lambda_*^A, \lambda_i^A, \lambda_{*j}^{AB}, \lambda_{ij}^{AB}, \lambda_{*k}^{AC}, \lambda_{ik}^{AC}, \lambda_{**l}^{ABD}, \lambda_{i*l}^{ABD}, \\ & \lambda_{*jl}^{ABD}, \lambda_{ijl}^{ABD}, \lambda_{*m}^{DE}, \lambda_{lm}^{DE}, \lambda_{**n}^{CDF}, \lambda_{*ln}^{CDF}, \lambda_{k*n}^{CDF}, \lambda_{kln}^{CDF}. \end{aligned}$$

It is easy to see that the distribution, assuming its positivity, has the desired conditional independence properties if and only if it has a parameterization as above, with all marginal log-linear parameters pertaining to effects not appearing in the list above set to zero. Therefore, the distributions in the graphical model are parameterized by the marginal log-linear parameters pertaining to the nodes-and-their-parent(s) type subsets of the variables.

The parameterization presented here consists of parameters with a straightforward interpretation. For example, λ_i^A is the effect of variable A , λ_{ij}^{AB} is the effect of A on B , etc. Note that λ_{ijl}^{ABD} is a measure of the joint effect of A and B on D , in addition to their separate effects, the existence of which is implied by the presence of a directed triangle containing these variables. Note that this interpretation of the meaning of λ_{ijl}^{ABD} is justified because A and B precede D and therefore the association among them may be interpreted as an effect.

The parameters are most easily interpreted when all variables are binary, as in this case they have a single numerical value. In other cases, the parameters

are vector valued and this reflects the way in which effects are measured in the log-linear tradition.

The approach to parameterize graphical models based on directed acyclic graphs presented here facilitates associating values with the arrows in the graph in a meaningful way. Many potential users of graphical modeling may find this useful and this gives a chance to graphical modeling to compete with the popular LISREL (Jöreskog, (1997)) approach that in a similar but different context provides the user with numbers assigned to the arrows, representing the strengths of effects. In LISREL, the numbers are regression coefficients in marginal regression equations but these equations are being defined by the user without the opportunity to check their consistency or implications. In the approach outlined here, the numbers are values of marginal log-linear parameters. The models are specified with respect to the entire joint distribution using graphs, all the implications can be read off from the graph and by restricting attention to directed acyclic graphs, contradicting model specifications are impossible.

4. FITTING MARGINAL MODELS

The models discussed in this paper can be specified by the constraint

$$h(\mu) = 0, \quad (1)$$

where $\mu = \log m$ is the vector of log expected cell frequencies and

$$h(\mu) = B' \log A' \exp(\mu) - v \quad (2)$$

for certain fixed matrices A and B and a vector v . Under Poisson sampling, the kernel of the unrestricted log-likelihood is given as

$$\mathcal{L}_n(\mu) = \sum n_i \mu_i - \sum \exp(\mu_i).$$

Notice that when conditioned on the sample size, the same estimates are obtained by maximizing the kernel as for multinomial sampling. Maximum likelihood estimation under a statistical model is a constrained optimization problem. If $\hat{\mu}$, the maximum likelihood estimate (MLE), exists, and if the matrices A and B possess certain regularity properties, the MLE is a saddle point of the Lagrange function

$$L(\mu, \lambda) = \mathcal{L}_n(\mu) - \lambda' h(\mu)$$

where λ is a vector of Lagrange multipliers.

Aitchinson, Silvey (1958) proposed a Fisher scoring method to find the saddle point of $L(\mu, \lambda)$, searching in the product space of the Lagrange multiplier vector and the vector of expected frequencies. A drawback of such an approach is that it does not distinguish between local maxima, local minima or saddlepoints of the likelihood function subject to the constraints, that is, the algorithm may converge to any stationary point depending on the starting point. Only in certain special cases, for example ordinary log-linear models (Haberman, (1974)), there is only one stationary point which is the maximum of the likelihood. An improved approach (Fletcher, (1970), Rapcsak, (2000)) is based on a so called exact penalty function $P_c(\mu)$, which has the MLE $\hat{\mu}$ as an unconstrained maximum. The function depends on a penalty parameter $c > 0$ which must be taken sufficiently large. The advantage is that standard optimization algorithms can be used to maximize $P_c(\mu)$, which is not possible with the Aitchison-Silvey approach. Furthermore, the search is done in the original parameter space of μ , rather than the product space of the λ and μ parameter spaces, which also simplifies the search.

The function $P_c(\mu)$ is derived from $L(\mu, \lambda)$ by (i) writing the Lagrange multiplier as a function of μ and (ii) adding a penalty term which penalizes for deviations from the model constraint $h(\mu) = 0$. The Lagrange multiplier, as a function of μ , is determined by differentiating $L(\mu, \lambda)$ with respect to μ , equating the result to zero, and solving for λ . A possible solution for λ , with the Jacobian of $h(\mu)$ given by

$$H(\mu) = \frac{\partial h(\mu)}{\partial \mu'} = B' D_{A'm}^{-1} A' D_m$$

is obtained as

$$\lambda(\mu) = (H' D_m^{-1} H)^{-1} H' D_m^{-1} (n - m)$$

where $H = H(\mu)$ and D_x is the diagonal matrix with the vector x on the main diagonal. A suitable nonnegative penalty term is the quadratic function $h(\mu)'(H' D_m^{-1} H)^{-1} h(\mu)$. Thus, an appropriate exact penalty function has the form

$$P_c(\mu) = L(\mu, \lambda(\mu)) + \frac{1}{2} c h(\mu)'(H' D_m^{-1} H)^{-1} h(\mu)$$

where c is some positive constant. Then we have (Rapcsák, (2000)):

Theorem 5. *There exists a $c^* > 0$ such that, for every $c > c^*$, $\hat{\mu}$ is an unconstrained local maximum of $P_c(\mu)$.*

Thus, standard optimization algorithms can be used to find $\hat{\mu}$. However, a large enough value of c needs to be selected. Initially, one can start with any value of c greater than one. If it is found that for the iterated estimates the penalty term does not go to zero, the penalty parameter must be increased.

When a sufficiently large penalty parameter has been found the algorithm will converge to a local maximum of the likelihood. If there is some doubt that this is not the global maximum, the procedure must be repeated with different starting values.

The standard Newton approach involves complicated derivatives making it impractical. However, a modified quasi-Newton approach which is based on simplified first and second derivatives of $P_c(\mu)$ can be used instead. It can be shown that the derivative of $P_c(\mu)$ with the derivative of $\lambda(\mu)$ replaced by its expected value is given by

$$d(\mu) = n - m - H\lambda(\mu) - (c - 1)H(H'D_m^{-1}.H)^{-1}h(\mu).$$

In spite of the simplification, $d(\mu)$ is still a valid search direction for $\hat{\mu}$. The expected value of the second derivative matrix evaluated under the model $h(\mu) = 0$ is

$$F_c(\mu) = D_m + (c - 2)H(H'D_m^{-1}.H)^{-1}H'$$

and for $c > 1$, this matrix is positive definite.

For sufficiently large $c > 1$, an algorithm then is

$$\begin{aligned}\mu^{(0)} &= \log n \\ \mu^{(k+1)} &= \mu^{(k)} - \text{step}^{(k)} F_c(\mu^{(k)})^{-1} d(\mu^{(k)})\end{aligned}$$

where $\text{step}^{(k)} \in \langle 0, 1 \rangle$ is a step size chosen such that $P_c(\mu^{(k+1)}) > P_c(\mu^{(k)})$ and if $n_i = 0$, then it is replaced by a small positive quantity, say 10^{-50} . The above algorithm will converge to $\hat{\mu}$ if the starting estimate $\log n$ is sufficiently close to it. Otherwise a different starting estimate may need to be tried.

ACKNOWLEDGMENTS

Rudas's research was supported in part by Grant No. OTKA T-032213 from the Hungarian National Science Foundation. Bergsma's research was supported by The Netherlands Organization for Scientific Research (NWO), Project Number 400-20-001.

REFERENCES

- AGRESTI, A. (1990) *Categorical Data Analysis*, Wiley, New York.
- AITCHISON, J. and SILVEY, S. D. (1958) Maximum likelihood estimation of parameters subject to restraints, *Ann. Math. Stat.*, 29, 813-828.
- BALAGTAS, C. C., BECKER, M. P., and LANG, J. B. (1995) Marginal modelling of categorical data from crossover experiments, *Applied Statistics*, 44, 63-77.

- BARTOLUCCI, F. and FORCINA, A. (2002) Extended RC association models allowing for order restrictions and marginal modeling, *J. Amer. Statist. Assoc.*, 97 (460), 1192-1199.
- BARTOLUCCI, F., FORCINA, A., and DARDANONI, V. (2001) Positive quadrant dependence and marginal modeling in two-way tables with ordered margins, *J. Amer. Statist. Assoc.*, 96 (456), 1497-1505.
- BECKER, M. P. (1994) Analysis of repeated categorical measurements using models for marginal distributions: an application to trends in attitudes on legalized abortion, In Marsden, P. V. (ed.) *Sociological Methodology*, 24, 229-265, Blackwell, Oxford.
- BECKER, M. P., MINICK, S., and YANG, I. (1998) Specifications of models for cross-classified counts: comparisons of the log-linear model and marginal model perspectives, *Sociological Methods and Research*, 26, 511-529.
- BERGSMA, W. P. (1997) *Marginal Models for Categorical Data*, Tilburg University Press, Tilburg.
- BERGSMA, W. P. and RUDAS, T. (2002a) Marginal models for categorical data, *Ann. Stat.*, 30, 140-159.
- BERGSMA, W. P. and RUDAS, T. (2002b) Modeling conditional and marginal association in contingency tables, *Ann. Fac. Sci. Toulouse Math.*, 11 (6), 443-454.
- BISHOP, Y. V. V., FIENBERG, S. E., and HOLLAND, P. W. (1975) *Discrete Multivariate Analysis*, MIT Press, Cambridge, MA.
- COLOMBI, R. and FORCINA, A. (2001) Marginal regression models for the analysis of positive association of ordinal response variables, *Biometrika*, 88 (4), 1007-1019.
- DARROCH, J. N., LAURITZEN, S. L., and SPEED, T. P. (1980) Markov fields and log-linear models for contingency tables, *Ann. Stat.*, 8, 539-552.
- DAVIS, L. J. (1989) Intersection union tests for strict collapsibility in three-dimensional contingency tables, *Ann. Stat.*, 17, 1693-1708.
- DAWID, A. P. (1980) Conditional independence for statistical operations, *Ann. Stat.*, 8, 598-617.
- FLETCHER, R. (1970) A class of methods for nonlinear programming with termination and convergence properties, In Abadie, J. Wolfe, P. (eds.) *Integer and nonlinear programming*, North Holland, Amsterdam.
- GLONEK, G. J. N. (1996) A class of regression models for multivariate responses, *Biometrika*, 83, 15-28.
- GLONEK, G. J. N. and McCULLAGH, P. (1995) Multivariate logistic models, *J. Roy. Statist. Soc., Ser. B*, 57, 533-546.
- HABERMAN, S. J. (1974) *The Analysis of Frequency Data*, University of Chicago Press, Chicago.
- HAGENAARS, J. A. (1990) *Categorical Longitudinal Data*, Sage, Newbury Park.
- JÖRESKOG, K. G. (1997) Structural equation models in the social sciences: specification, estimation, and testing, In Krishnaiah, P. R. (ed.) *Applications of Statistics*, 267-287, North-Holland, Amsterdam.
- KAUERMANN, G. (1997) A note on multivariate logistic models for contingency tables, *Austr. J. Stat.*, 39, 261-276.
- KELLERER, H. G. (1964) Verteilungsfunktionen mit gegebenen Marginalverteilungen, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 3, 247-270.
- LANG, J. B. and AGRESTI, A. (1994) Simultaneously modelling the joint and marginal distributions of multivariate categorical responses, *J. Am. Stat. Assoc.*, 89, 625-632.
- LANG, J. B., McDONALD, J. W., and SMITH, P. W. F. (1999) Association-marginal modelling of multivariate categorical responses: a maximum likelihood approach, *J. Am. Stat. Assoc.*, 94, 1161-1171.

- LAURITZEN, S. L. (1996) *Graphical Models*, Clarendon Press, Oxford.
- LAURITZEN, S. L., SPEED, T. P., and VIJAYAN, K. (1984) Decomposable graphs and hypergraphs, *J. Austr. Math. Soc., Ser. A*, 36, 12-29.
- LITTLE, R. J. and RUBIN, D. (1987) *Statistical Analysis with Missing Data*, Wiley, New York.
- MOLENBERGHS, G. and LESAFFRE, E. (1999) Marginal modelling of multivariate categorical data, *Statistics in Medicine*, 18, 2237-2255.
- RAPCSÁK, T. (2000) Global Lagrange multiplier rule and smooth exact penalty functions for equality constraints, In Di Pillo, G., Giannesi, F. (eds.) *Nonlinear Optimization and Related Topics*, Kluwer, 351-368.
- RUDAS, T. (1998a) *Odds Ratios in the Analysis of Contingency Tables*, Sage, Thousand Oaks.
- RUDAS, T. (1998b) A new algorithm for the maximum likelihood estimation of graphical log-linear models, *Computational Statistics*, 13 (9), 529-537.
- RUDAS, T. (2002) Canonical representation of log-linear models, *Communications in Statistics (Theory and Methods)*, 31 (12), 2311-2323.

TAMÁS RUDAS
Department of Statistics
Eötvös Loránd University
H-1117 Budapest
Péter sét-ny 1/A (Hungary)
rudas@tarki.hu

WICHER BERGSMA
Faculty of Social and Behavioural
Sciences Methodology and Statistics
Tilburg University
P.O. Box 90153
NL-5000 LE Tilburg (The Netherlands)
W.P.Bergsma@uvt.nl