Likelihood Factorizations for Mixed Discrete and Continuous Variables

D. R. COX

Department of Statistics and Nuffield College, Oxford

NANNY WERMUTH

Center of Survey Research and Methodology (ZUMA), Mannheim

ABSTRACT. Some general remarks are made about likelihood factorizations, distinguishing parameter-based factorizations and concentration-graph factorizations. Two parametric families of distributions for mixed discrete and continuous variables are discussed. Conditions on graphs are given for the circumstances under which their joint analysis can be split into separate analyses, each involving a reduced set of component variables and parameters. The result shows marked differences between the two families although both involve the same necessary condition on prime graphs. This condition is both necessary and sufficient for simplified estimation in Gaussian and for discrete log linear models.

Key words: conditional Gaussian model, conditional independence, graph, likelihood, median-dichotomized Gaussian distribution, multivariate normal distribution, partially dichotomized Gaussian model, prime graph, separation theorem

1. Introduction

For a given parametric family of models likelihood factorizations play an important role in formal studies of inference. Factorizations of the likelihood may also arise from the conditional independencies expressed in graphical Markov models. In this paper we explore the relations between these two ideas, with particular reference to distributions of mixtures of binary and Gaussian variables.

2. Parameter based factorizations

Suppose for a family of models specified by a parameter θ taking values in a parameter space Ω_θ we can write the likelihood for an observed vector x in the form

$$L(\theta; x) = L_1(\theta_1; x)L_2(\theta_2; x),$$

where

$$\theta_1 \in \Omega_1, \quad \theta_2 \in \Omega_2$$

and $\Omega_1 \times \Omega_2 = \Omega$. That is, the parameters split into variation independent components, any combination of allowable values of θ_1 and θ_2 being possible.

Then

(1) the maximum likelihood estimate $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$ is obtained by separate maximization of the factors;
(2) a profile likelihood for, say, θ_1 can be obtained solely from the factor L_1 and may often be the preferred base for inference about θ_1;
(3) under suitable regularity conditions the estimates $\hat{\theta}_1$, $\hat{\theta}_2$ are asymptotically independent and are asymptotically normally distributed around θ_1, θ_2 with asymptotic covariance matrices determined from the separate factors;

(4) if $\phi = \phi(\theta_1, \theta_2)$ is a parametric function depending on both θ_1 and θ_2, then the corresponding maximum likelihood estimate is $\hat{\phi} = \phi(\hat{\theta}_1, \hat{\theta}_2)$ and its asymptotic covariance matrix can be calculated directly from 3 above, although construction of a profile likelihood for ϕ will in general require calculations not confined to the separate factors.

We call such factorizations parameter based. In some cases stronger small-sample properties, such as "exact" confidence limits, are available. The definition extends directly to more than two factors.

For example, if (Y_1, Y_2) have a bivariate normal distribution with all parameters unknown, factorizations via marginal and conditional distributions are available with variation independent parameters in the two components. Under restrictions such as equal marginal variances no such factorizations are possible.

An important special case of a parameter-based factorization is that of a cut in a regular exponential family (Barndorff-Nielsen, 1988). Here we are able to separate a sufficient statistic T and a conditioning statistic C such that the conditional distribution of T given C depends only on a parameter of interest whereas the marginal distribution of C depends only on a variation-independent nuisance parameter. The present notion is, however, more general and has no necessary connection with exponential families.

An important example is that of the parametric analysis of survival data subject to uninformative censoring specified by a parametric distribution of censoring time. Here the full likelihood factorizes into a term from the failure-time distribution and a second term from the censoring-time distribution.

Sometimes, also it may be useful to consider partially variation independent components. If with two factors it is possible to write the likelihood as

$$L_1(\phi_1, \lambda; x)L_2(\phi_2, \lambda; x),$$

where $(\phi_1, \phi_2, \lambda)$ are variation independent, there can be some gain in computation especially if λ is of small dimension. If, further, it can be arranged that λ is orthogonal (Cox & Reid, 1987) to (ϕ_1, ϕ_2) then the maximum likelihood estimates of (ϕ_1, ϕ_2) from the separate factors are asymptotically fully efficient.

3. Concentration-graph factorizations

Factorizations of the likelihood often arise in a more general way via the conditional independencies expressed in graphical structures. We consider a graph with a set V of nodes, each node representing a variable. We suppose that there is at most one edge for each pair of nodes, each edge present being a full line. The absence of an edge between two nodes expresses conditional independence of the corresponding variable pair given all other variables in V. Such a graph has been called a concentration graph (Cox & Wermuth, 1993, 1996) because in a multivariate Gaussian distribution absence of an edge corresponds to a zero in the concentration or inverse covariance matrix. See also Wermuth (1998).

It is known (Lauritzen, 1996, prop. 3.8) that for such an undirected graph the factorization of any corresponding probability distribution implies the following separation criterion. Let a, b, c denote disjoint subsets of V with, for example, X_a denoting the set of random variables defined on a. If c separates a, b, that is if every path from a node in a to one in b has a node in c, then,

$$X_a \perp\!\!\!\perp X_b \mid X_c,$$

i.e. the set of variables defined on a is conditionally independent of those in b given those in c. If the conditional independence property holds for all individuals given data on a set

of independent individuals it follows that the likelihood within any parametric family of models based on (X_a, X_b, X_c) can be factorized in several ways, namely as

$$\begin{aligned} L(\theta; x) &= L^{a|c}(\theta; x_a|x_c)L^{b|c}(\theta; x_b|x_c)L^{c}(\theta; x_c) \\ &= L^{a|c}(\theta; x_a|x_c)L^{bc}(\theta; x_b, x_c) \\ &= L^{ac}(\theta; x_a, x_c)L^{b|c}(\theta; x_b|x_c). \end{aligned}$$

We call these concentration graph-factorizations of the likelihood.

Of course we can always use the recursive law of conditional probability to produce factorizations of a joint density and hence of likelihood, but in general these will not lead to useful simplifications. Sometimes, however, we may have a parametric family such that a given factorization of the likelihood is both parameter-based and corresponds to a given concentration graph. Then, in particular, maximum likelihood estimation is simplified.

4. A simple example

We give a simple illustration involving three variables, i.e. the sets a, b, c consist of a single node each. It is a special case of what we shall call the partially dichotomized Gaussian distribution, defined in generality in section 6. Suppose that (Y_a, U_b, Y_c) are trivariate normal with (Y_a, U_b) conditionally independent given Y_c and with U_b having zero mean and unit variance. Suppose further that U_b is dichotomized to form a binary variable I_b, i.e. $I_b = 1$ for U_b larger than some cutoff point α and $I_b = -1$, otherwise. The conditional independence property is retained, i.e. $Y_a \perp\!\!\!\perp I_b|Y_c$. In the associated concentration graph, the node c separates the nodes a, b. To simplify the results we make the inessential simplification that (Y_a, Y_c) have zero means. The likelihood from a single individual can be factorized in the form

$$\sqrt{\sigma}^{-1}_{aa.c}\phi\left(\frac{y_a - \beta_{ac}y_c}{\sqrt{\sigma_{aa.c}}}\right)\Phi\left(i_b\frac{-\alpha + \beta_{bc}y_c}{\sqrt{\sigma_{bb.c}}}\right)\sqrt{\sigma}^{-1}_{cc}\phi\left(\frac{y}{\sqrt{\sigma_{cc}}}\right).$$

Here $(\Phi(x), \phi(x))$ are respectively the standard normal distribution and density functions, and the convention for parameters is that, for example, σ_{aa} denotes the unconditional variance of Y_a, the regression coefficient of Y_c in the regression of Y_a on Y_c is denoted by β_{ac} and the conditional variance of Y_a given Y_c by $\sigma_{aa.c}$. The resulting factorization thus involves the parameters

$$\sigma_{aa.c}, \beta_{ac}; \quad \alpha/\sqrt{\sigma_{bb.c}}, \beta_{bc}/\sqrt{\sigma_{bb.c}}; \quad \sigma_{cc}.$$

These three parameter spaces are not subject to constraints and together form the full five dimensional space of the original specification. Thus the concentration graph factorization is also a parametric factorization for the appropriate choice of parameters.

5. Simplified estimation for Gaussian and log-linear models

When the variables are either all continuous with a joint Gaussian distribution or all have a discrete distribution they may restricted only by independence statements (Wermuth, 1976) which are conveniently captured by a concentration graph. Corresponding statistical models have been introduced as respectively covariance selection models (Dempster, 1972) and as graphical (log-linear) models (Darroch *et al.*, 1980). For both, the condition for sim- plified maximum-likelihood estimation has been directly specified in terms of the concentration graph as follows. If a, b, c are non-overlapping subsets of V which give all nodes of V, if

c separates a from b and if c is a complete separator, then the estimation problem for all variables simplifies into two separate problems involving $a \cup c$ and $b \cup c$ (Lauritzen, 1996, discussions of collapsibility).

The condition that a, b, c form a partition of V ensures that the estimation concerns the joint distribution of all variables and not some marginal distribution. When c is a complete separator no conditional or marginal independence statement is implied by the model for variables of c. If there were such an independency, as for instance in Fig. 1a, it would not imply such a restriction on the association of the involved variable pair in any marginal distribution obtained by marginalizing over all variables along a path outside c connecting the two nodes. As a consequence the marginals $a \cup c$ or $b \cup c$ would not contain the information about the independency in the joint distribution.

Figure 1 illustrates further the notion of a separator in a case when a, b, c does not partition V. Note that in general separation of a, b by c in the joint distribution implies also an independence statement $X_a \perp\!\!\!\perp X_b \mid X_c$ in the distribution of X_a, X_b, X_c obtained after marginalizing over the remaining variables.

Using the convention that dots indicate discrete variables and circles denote continuous variables, Fig. 2 gives an example of a specific log-linear model in nine variables for which simplified estimation is possible.

Since c is a complete separator the analysis of the graphical model corresponding to this concentration graph can be reduced to two separate analyses involving a five-dimensional contingency table for $a \cup c$ and a six-dimensional contingency table for $b \cup c$.

Then the maximum-likelihood estimates for the joint table of all variables can be obtained from the smaller marginal tables, the power of tests is increased and, most importantly, interpretation may be considerably simplified.

Since there are typically several complete separators the question arises whether there is a way of choosing those which give in some sense the best simplification. This is possible by

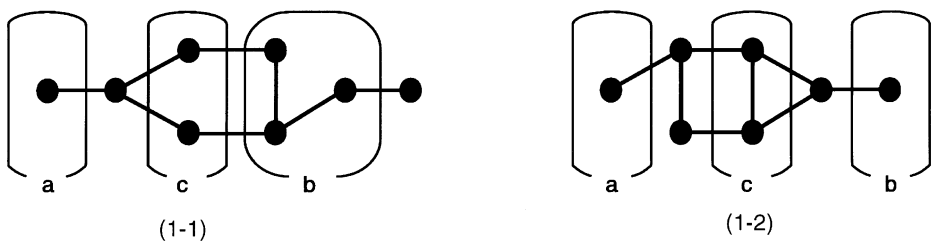

Fig. 1. Examples of separators for a, b, c (1-1) incomplete separator c; after marginalizing for instance over common adjacent node to left of c independence no longer implied for the two nodes in c; (1-2) complete separator c, no independence implied conditionally or marginally for nodes of c.

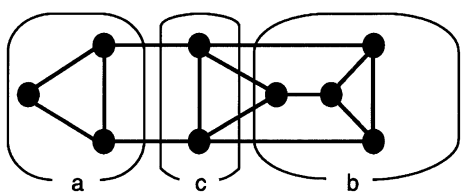

Fig. 2. Concentration graph of log-linear graphical model with nine variables. Simplified estimation possible since c complete. Possible via tables corresponding to $a \cup c$ and $b \cup c$, respectively.

using the important notion of prime graphs. Prime graphs are a direct generalization of the concept of prime numbers in the sense that prime graphs also cannot be further divided.

The prime graphs of an undirected graph are the maximal subgraphs without a complete separator. There are efficient algorithms to find all prime graphs of any undirected graph (Leimer, 1993).

Figure 3 shows the three prime graphs of the nine-node graph in Fig. 2 and Fig. 4 illustrates the different types of prime graphs: complete graphs, which have no separating set, chordless n-cycles, $n \geqslant 4$, with each pair without edge being an incomplete separator and more complex graphs containing a visually hidden cycle or several chordless n-cycles.

Results concerning simplified estimation for Gaussian and for log-linear concentration graph models may be summarized as follows.

1. Simplified maximum-likelihood estimation for the joint distribution is possible involving a reduced set of component variables and parameters if and only if the concentration graph is not a prime graph.
2. For any given concentration graph a computationally efficient simplification is obtained by

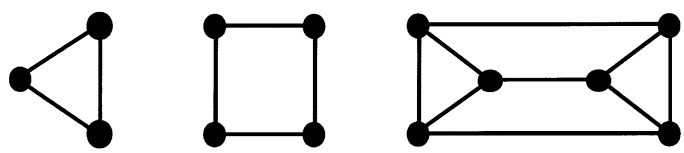

Fig. 3. Prime graphs corresponding to concentration graph of Fig. 2 pointing to tables with which computationally efficient simplifications can be achieved.

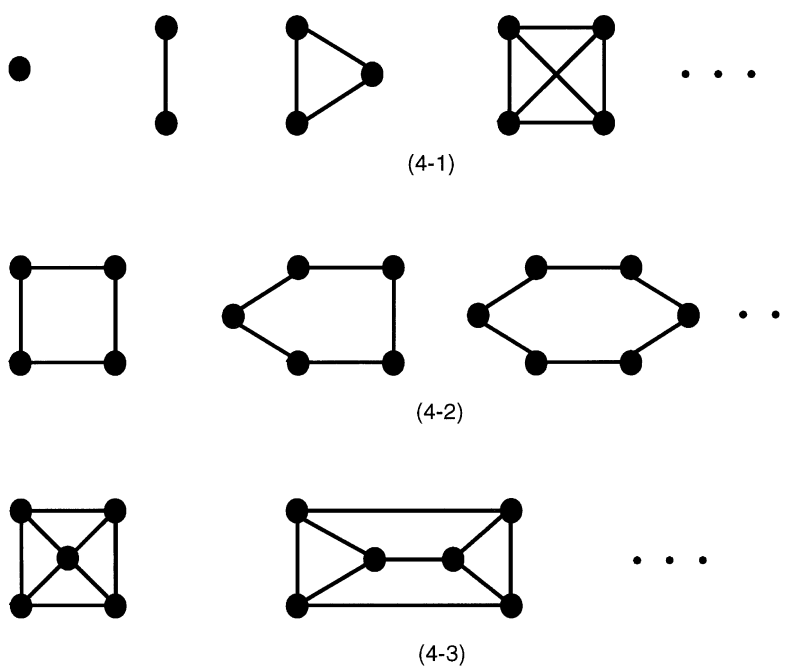

Fig. 4. Examples of different types of prime graph (4-1) complete graphs; (4-2) chordless n-cycles; (4-3) more complex graphs with visually hidden 4-cycle (left) and several such cycles (right).

splitting the estimation problem into separate analyses by taking one at a time the prime graphs from a proper full sequence of all prime graphs.

3. If all prime graphs are complete, then each separate analysis concerns a saturated marginal distribution, and hence closed-form estimation for the joint distribution is possible. Such models have been called decomposable.
4. If a model is decomposable, estimation in the joint distribution reduces to a series of univariate calculations (Wermuth, 1980; Wermuth & Lauritzen, 1983) since then single-clique nodes may be chosen one at a time and deleted from the concentration graph until all nodes are exhausted. For computational purposes each chosen node defines a response variable in a system of univariate recursive regressions with the direct adjacent nodes in the reduced graph being its explanatory variables.

The requirement that the concentration graph is not a prime graph is no longer on its own a sufficient condition for simplified estimation in distributions of mixed discrete and continuous variables.

6. Models for mixed discrete-continuous variables

Suppose that the variables are divided into two types, discrete, often to be treated as binary, and continuous, in fact having Gaussian, i.e. normal, distributions. Let the corresponding two sets of nodes in the graph be Δ and Γ. We denote the continuous components by Y, having dimension p, and the discrete components by I, having dimension q. The full observed random variable is then $X = (Y, I)$.

One important family of models for the joint distribution of such variables is the HCG (homogeneous conditional Gaussian) family (Lauritzen & Wermuth, 1989). Here the set of I of all discrete components has a multinomial distribution and given $I = i$, the set Y of all continuous components has a multivariate normal distribution of mean μ_i and covariance matrix $\Sigma_{YY.I}$. In the heterogeneous CG case, which we do not consider here, the conditional covariance matrix depends on i.

We may contrast the HCG distribution with the partially dichotomized multivariate Gaussian distribution, PDG, taken here for binary components and obtained as follows, generalizing the example of section 4. Let (U, Y) be multivariate normal with mean $(0, \mu)$ and covariance matrix Σ partitioned in the usual way and suppose that U is not directly observed but is dichotomized at α to form I, that is the components of I are defined by

$$I_s = 1 \text{ if } U_s > \alpha_s, \quad I_s = -1 \text{ if } U_s \leqslant \alpha_s.$$

Both distributional types stem from a long history in the analysis of discrete data. For early references on the CG distribution, see Lauritzen & Wermuth (1989). The PDG distribution plays an important role in the study of linear structural relations (Jöreskog, 1981; Bollen, 1989).

Let $\Phi_p(y; \Sigma)$ denote the p-dimensional cumulative normal integral corresponding to zero mean and covariance matrix Σ and $\phi_p(y; \Sigma)$ the corresponding density function. Then for an observation on a single individual we have that the joint density of a component Y_1, of dimension p_1, of the continuous variable and the probability that the component I_1, of dimension q_1 of the binary variable takes value $(1, \ldots, 1)$ given that the complementary components Y_2, I_2 of dimensions (p_2, q_2), takes values y_2 and $(1, \ldots, 1)$ is

$$\begin{aligned}
&\Phi_q(-\alpha + B_{UY_1.Y_2}(y_1 - \mu_1) + B_{UY_2.Y_1}(y_2 - \mu_2); \Sigma_{UU.Y}) \\
&\quad \times \{\Phi_{q_2}(-\alpha_2 + B_{U_2Y_2}(y_2 - \mu_2); \Sigma_{U_2U_2.Y_2})\}^{-1} \\
&\quad \times \phi_{p_1}(y_1 - \mu_1 - B_{Y_1Y_2}(y_2 - \mu_2); \Sigma_{Y_1Y_1.Y_2}),
\end{aligned}$$

where, for example, $B_{UY_1.Y_2}$ denotes the matrix of least squares regression coefficients of components Y_1 when regressing U on both Y_1 and Y_2.

If more generally the vector binary random variables take values i_1, i_2, vectors of 1s and -1s, the denominator is changed to

$$\{\Phi_{q2}\{i_2 * (-\alpha_2 + B_{U_2Y_2}(y_2 - \mu_2)\};\ i_2^{\mathrm{T}} * \Sigma_{U_2U_2.Y_2} * i_2\}^{-1},$$

where $v * w$ denotes the Hadamard product with elements $(v_1w_1, v_2w_2, \ldots)$ and there is a corresponding change in the first term in the numerator. Here U is the $q = q_1 + q_2$ dimensional random variable formed from U_1 and U_2. For arbitrary random variables X, W, Z, we have $B_{XW.Z} = \Sigma_{XW.Z}\Sigma_{WW.Z}^{-1}$, where the common notation for conditional covariance matrices is used, i.e.

$$\Sigma_{XZ.W} = \Sigma_{XZ} - \Sigma_{XW}\Sigma_{WW}^{-1}\Sigma_{WZ}.$$

Various special cases follow, such as the multivariate probit regression of I given Y, and the conditional distribution of Y given I, a special case of which is used by Azzalini & Dalla Valle (1996) as a flexible set of asymmetric multivariate distributions.

A number of special features of the partially dichotomized Gaussian distribution follow from the above form. These include the following:

(i) if the conditioning variable is continuous, $q_2 = 0$, the second term in the expression is absent and there is a parameter-based factorization;
(ii) if a conditional independence statement involves a discrete variable as conditioning variable then a stronger independence holds: $I_1 \perp\!\!\!\perp Y_1|(I_2, Y_2)$ only if either $(I_1, I_2) \perp\!\!\!\perp Y_1 \mid Y_2$ or $(Y_1, I_2) \perp\!\!\!\perp I_1 \mid Y_2$;
(iii) $Y_1 \perp\!\!\!\perp Y_2 \mid I_2$ only if either $Y_1 \perp\!\!\!\perp (Y_2, I_2)$ or $Y_2 \perp\!\!\!\perp (Y_1, I_2)$;
(iv) $Y_1 \perp\!\!\!\perp I_1 \mid I_2$ only if either $Y_1 \perp\!\!\!\perp I$ or $I_1 \perp\!\!\!\perp (Y_1, I_2)$;
(v) $I_1 \perp\!\!\!\perp I_2 \mid I_3$ only if one of I_1, I_2 is independent of the other two;
(vi) $I_1 \perp\!\!\!\perp Y_1 \mid Y_2$ if and only if $U_1 \perp\!\!\!\perp Y_1 \mid Y_2$;
(vii) $I_1 \perp\!\!\!\perp I_2 \mid Y_2$ if and only if $U_1 \perp\!\!\!\perp U_2 \mid Y_2$.

There is, however, some specialization when all variables are median dichotomized, i.e. when the cut-points are medians in the marginal distributions. Their joint distribution can be written as a log linear model with interaction terms of only even order (Edwards, 1995, app. C). Therefore $I_1 \perp\!\!\!\perp I_2 \mid I_3$ can hold without a stronger independence. The corresponding condition on the underlying Gaussian variables is then not $U_1 \perp\!\!\!\perp U_2 \mid U_3$, see Section 7. Thus, conditional independencies in the underlying Gaussian distribution are retained in the partially dichotomized distribution if and only if all involved conditioning variables are continuous.

We compare first the two joint distributions of (I, Y) without parametric restrictions, i.e. for the saturated models. The distribution-based factorizations which are also parameter-based are then

$$L^{Y|I}(\mu_i, \Sigma_{YY.I};\ y)L^I(\theta_I;\ i),$$
$$L^{I|Y}(\alpha, B_{UY}, \Sigma_{UU.Y};\ i)L^Y(\mu, \Sigma_{YY};\ y)$$

for the homogeneous conditional Gaussian and the partially dichotomized Gaussian distribution, respectively.

For the complementary factorizations the parameter spaces are, except in degenerate cases, variation dependent. The reason is that the marginal distribution of the continuous variables Y is a discrete mixture of normal distributions in the former model and the conditional distribution of Y given I involves a continuous mixture of truncated normal distributions in the latter model.

More generally the quite different behaviour of the two distributions under conditioning and marginalization is best seen by partitioning V into (a, b) so that the discrete variable I is split into (I_a, I_b) and the continuous variable into (Y_a, Y_b). Then the marginal distribution of (I_b, Y_b)

(1) for a homogeneous conditional Gaussian distribution of (I, Y) is itself homogeneous conditional Gaussian if and only if $I_a \perp\!\!\!\perp Y_b | I_b$, that is for Y_b having a Gaussian distribution with means depending only on I_b (Frydenberg, 1990);
(2) is partially dichotomized Gaussian whenever (I, Y) also is partially dichotomized Gaussian.

The second result follows by integrating over u_a and y_a in the assumed multivariate Gaussian distribution of (U, Y) underlying (I, Y). Furthermore the conditional distribution of (I_a, Y_a) given $(I_b, Y_b) = (i_b, y_b)$ is such that

(1) for a homogeneous conditional Gaussian distribution of (I, Y) it also is homogeneous conditional Gaussian (Lauritzen & Wermuth, 1989, prop. 2.3);
(2) for a partially dichotomized Gaussian distribution of (I, Y) it also is partially dichotomized Gaussian if and only if either $(I_a, Y_a) \perp\!\!\!\perp I_b | Y_b$ or $Y_a \perp\!\!\!\perp I | Y_b$.

The necessity part of the last result is proved by examining the circumstances under which the multivariate normal integral in the numerator of the conditional density can be factorized into the appropriate form.

A more general possibility is obtained by combining the two types of distribution, splitting the binary components as (I, J). Suppose that (U, Y) have a homogeneous Gaussian distribution given $J = j$, J itself having an arbitrary distribution. Suppose further that (I, Y) has a partially dichotomized Gaussian distribution given $J = j$. Then the properties listed above hold with the full set J included in the conditioning sets. The distribution of I given $(Y = y, J = j)$ has the form

$$\Phi_{qI}(-\alpha + B_{UY.J}(y - \mu) + B_{UJ.Y}j; \Sigma_{UU.YJ})$$

and Y given $J = j$ has the homogeneous Gaussian density

$$\phi_p(y - \mu - B_{YJ}j; \Sigma_{UU.J}).$$

7. Factorization for a special dichotomized Gaussian distribution

A special case of a fully dichotomized Gaussian distribution occurs for median dichotomized variables. If the concentration graph is such that the largest prime graph has no more than three nodes, that is any subgraph with more than three nodes has a complete separator, then any factorization achieved is both concentration graph and parameter-based. We may consider three binary variables, denoted by A, B, C, say, with $A \perp\!\!\!\perp B \mid C$. Because of the median dichotomy the marginal likelihood of C is constant and there is the factorization $L^{A|C}L^{B|C}$. Now the dichotomized Gaussian distribution in question is determined by just two parameters and therefore if these are taken as correlations, or equivalently odds ratios, in the (A, C) and (B, C) tables a parameter-based factorization is achieved. The explicit one-to-one relation between a marginal odds-ratio (or) and a marginal correlation coefficient ρ is

$$\text{or}\,(A, B) = (1 + \rho_{AB})^2/(1 - \rho_{AB})^2, \quad \rho_{AB} = \tanh(c), \quad c = \tfrac{1}{4}\log\{\text{or}\,(A, B)\}.$$

If it is required to relate this to the underlying distribution of the multivariate normal variable U we argue as follows. The median dichotomized Gaussian distribution in three variables is

both a linear in probabilities model and a quadratic exponential model with no three-factor terms in either representation (Cox & Wermuth, 1992; Edwards, 1996, app. C). The conditional independency can thus be expressed directly in terms of the product moment correlations of the binary variables as

$$\rho_{AB} = \rho_{AC}\rho_{BC}$$

from which the maximum likelihood estimate of ρ_{AB} can, if required, be found. Estimates of the whole table may be built up in a recursive fashion analogous to the Gaussian case (Wermuth & Cox, 1998, app. 1).

Because of Sheppard's formula relating the correlation of the binary variables and that of the underlying normal distribution, we have, for example,

$$\rho_{AB} = \sin^{-1} \rho_{U_A U_B}$$

so that the maximum likelihood estimate of the underlying correlation matrix can be computed.

Note that the hypothesis of conditional independence of the resulting binary variables is expressed via the non-linear constraint

$$\sin^{-1} \rho_{U_A U_B} = \sin^{-1} \rho_{U_A U_C} \sin^{-1} \rho_{U_B U_C}.$$

No analogous exact results hold for four variables because the dichotomized Gaussian distribution has a, possibly small, four-factor interaction.

This special family of a median-dichotomized Gaussian distribution is one example of a quasi-linear system (Wermuth & Cox, 1998) of discrete variables and as such having some properties very similar to the those of Gaussian distribution.

8. Factorization conditions for mixed distributions

We now consider a structure for a mixture of discrete and continuous variables described as above by a concentration graph, G^V_{con}, with nodes V divided into two sets $\Delta = \{1, \ldots, q\}$ and $\Gamma = \{1, \ldots, p\}$ corresponding to discrete and continuous components. When we consider a homogeneous conditional Gaussian distribution which is itself partially dichotomized we divide Δ into two parts Δ_I and Δ_J as explained at the end of section 6.

Suppose that there are three sets of nodes a, b, c which also partition V, and are such that a conditional independence is represented by c being a separator of a and b, and that the separator c is complete so that there is no such independence restriction for nodes in c.

We ask for the distributions outlined in section 6: when is a resulting concentration graph factorization also a parameter-based factorization? The answer is

(i) for a homogeneous conditional Gaussian distribution if $c \subseteq \Delta$ or if either $a \subseteq \Gamma$ or $b \subset \Gamma$;
(ii) for a partially dichotomized Gaussian distribution if $c \subseteq \Gamma$;
(iii) for a homogeneous conditional Gaussian distribution which itself is partially dichotomized if $\Delta_J \subseteq c$ and all further components of separator c are continuous, i.e. are from Γ.

The first statement is a reformulation of a result by Frydenberg & Lauritzen (1989) who consider sets a, b, c with the above properties but being ordered, in addition. The results stem from the behaviour of the distributions under marginalizing and conditioning given in section 6.

For a partially dichotomized Gaussian distribution the factorization of an arbitrary conditional distribution of, say, (I_a, Y_a) given Y_c depends only on the parameters in the conditional

distribution of (U_a, Y_a) given Y_c. When, however, the conditioning variable has binary components an essential complication is introduced via the term in the denominator of the conditional distribution.

In the present context the essence of the distinction between the two forms of distribution is that the homogeneous conditional distribution retains simple structure under conditioning whereas the partially dichotomized Gaussian does so under marginalization.

The simple concentration graphs of Fig. 5 illustrate already a main difference between the two families of distributions.

In both cases c separates a from b, but the separating set c is continuous in the first case and discrete in the second. In this and the followings graphs open circles represent again continuous components and dots represent binary components.

In Fig. 5-1 with variables A, B, X the independence $A \perp\!\!\!\perp B \mid X$ holds. This is simply represented via the partially dichotomized Gaussian distribution, because the model results if there is conditional independence given X in the underlying trivariate Gaussian distribution. In the conditional Gaussian distribution, however, this conditional independence is complicated, because it is not directly connected with the generating process, where the distribution of X arises conditionally given the marginal distribution of the two discrete components. The conditional independence of the discrete components can hold but maximum-likelihood fitting is not direct; it is iterative, involving simultaneously observations on all three variables, even though the graph separates into two complete prime graphs.

In Fig. 5-2 with variables Y, X, A the independence $Y \perp\!\!\!\perp X \mid A$ holds. This is simply represented via the homogeneous conditional Gaussian distribution by having zero correlation between Y and X at each level of A. However, the partially dichotomized Gaussian distribution represents this conditional distribution only in the degenerate situation where in addition either $A \perp\!\!\!\perp Y$ or $A \perp\!\!\!\perp X$ holds. This is proved by calculating the conditional joint distribution of X, Y given, say $A = 1$ as a function of x, y and noting that a positive function $h(x, y)$ factorizes if and only if $\partial \log h(x, y)/\partial x$ is a function of x alone.

The concentration graphs in Figs 6 and 7 with ten nodes are slightly more complicated.

The first two (Figs 6-1, 6-2) have the same set of missing edges; they differ just in the number and location of discrete and continuous nodes. They are nondecomposable graphs since not all their prime graphs are complete (Matúš, 1994).

There are four prime graphs in both Figs 6-1, 6-2, two are complete (of three nodes at the left end and of four nodes at the right end), and two are incomplete (of four and six nodes). The graph of Fig. 7-2 is a decomposable graph with the largest prime graph having four nodes.

Figures 6 and 7 show structures in which a concentration graph factorization is also parameter based: for the homogeneous conditional Gaussian (Figs 6-1, 6-2), for the partially dichotomized Gaussian distribution (Fig. 7-1) and for none of the two (Fig. 7-2). See the legend for more detailed explanation.

Fig. 5. (5-1). Continuous component separates two binary components. Corresponding parameter-based factorization for partially dichotomized Gaussian but not for homogeneous conditional Gaussian. (5-2). Binary component separates two continuous components. Corresponding parameter-based factorizations for homogeneous conditional Gaussian but for partially dichotomized Gaussian conditional independency does not hold without additional edge missing.

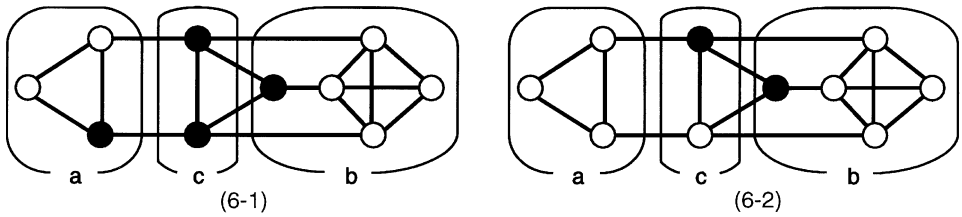

Fig. 6. (6-1). Mixed components a, b separated by c. Complete separator c discrete. $L^{a|c}L^{b|c}L^{c}$ parameter-based only for homogeneous conditional Gaussian distribution. (6-2). Continuous component a separated by c from b; b, c both mixed; $L^{a|c}L^{bc}$ parameter based only for homogeneous conditional Gaussian.

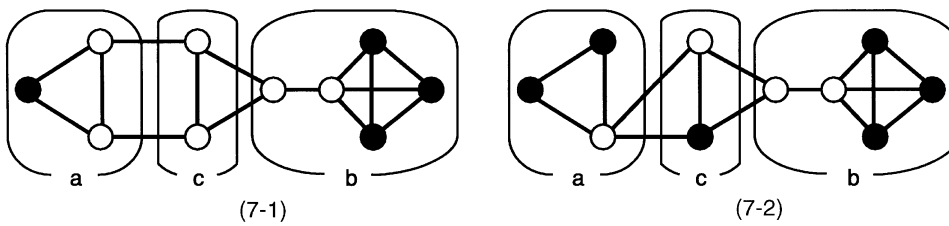

Fig. 7. (7-1). Mixed components a, b separated by c, which is complete and continuous; no node pair has minimal separator containing discrete nodes; $L^{a|b}L^{b|c}L^{c}$ parameter-based only for partially dichotomized Gaussian. (7-2). Mixed components in each of a, b, c; no parameter-based factorization for partially dichotomized Gaussian or homogeneous conditional Gaussian. If conditionally on discrete component in separator c there is underlying Gaussian distribution for all remaining variables then $L^{a|b}L^{b|c}L^{c}$ parameter-based.

If the concentration graph can be repeatedly split by a complete separator c into two parts $a \cup c$ and $b \cup c$, each containing only complete prime graphs then *the graph is decomposable.* If the resulting graph factorizations of the likelihood are also parameter-based then *the model is called decomposable*. As a consequence maximum likelihood analysis can be reduced to a series of analyses each involving just the component variables of a prime graph.

For a decomposable concentration graph the independence structure may equivalently be expressed by a directed acyclic graph for ordered nodes, that is by a recursive sequence of independence statements, each involving an individual variable X_r and subsets of X_s, $s > r$. As mentioned before a corresponding split of analysis into a sequence of univariate regressions is then possible for a joint Gaussian distribution and for an arbitrary discrete distribution but as shown here only under quite different and strong conditions for homogeneous conditional Gaussian and for partially dichotomized Gaussian distributions.

In a specific application the choice between the two different models for mixed discrete and continuous variables will often be based on some combination of consistency with an interpretable generating process, on empirical fit and on the ease of statistical analysis. For instance, if no binary variable is considered as a response to a continuous explanatory variable, an initial preference for a conditional Gaussian distribution is indicated. If the marginal distributions of all continuous variables are Gaussian with no suggestion of being mixtures of Gaussian distributions an initial preference for the partially dichotomized Gaussian distribution arises.

Often the most effective route in applications is to build up a parametric model as a recursive system of regressions. Associated with this there is always a factorization of the likelihood but the joint distribution of all variables may be of a rather complex form.

References

Azzalini, A. & Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83**, 715–726.
Barndorff-Nielsen, O. E. (1978). *Information and exponential families*. Wiley, Chichester.
Bollen, K. A. (1989). *Structural equations with latent variables.* Wiley, New York.
Cox, D. R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. Ser. B* **49**, 1–39.
Cox, D. R. & Wermuth, N. (1992). Response models for mixed binary and quantitative variables. *Biometrika* **79**, 441–461.
Cox, D. R. & Wermuth, N. (1993). Linear dependencies represented by chain graphs (with discussion). *Statist. Sci.* **8**, 204–283.
Cox, D. R. & Wermuth, N. (1996). *Multivariate dependencies: models, analysis and interpretation.* Chapman & Hall, London.
Darroch, J. N., Lauritzen, S. L. & Speed, T. P. (1980). Markov fields and log-linear models for contingency tables. *Ann. Statist.* **8**, 522–539.
Dempster, A. P. (1972). Covariance selection. *Biometrics* **28**, 157–175.
Edwards, D. (1995). *Introduction to graphical modelling.* Oxford University Press, Oxford.
Frydenberg, M. (1990). Marginalization and collapsibility in graphical interaction models. *Ann. Statist.* **18**, 790–805.
Frydenberg, M. & Lauritzen, S. L. (1989). Decomposition of maximum likelihood in mixed graphical interaction models. *Biometrika* **76**, 539–555.
Jöreskog, K. G. (1981). Analysis of covariance structures. *Scand. J. Statist* **8**, 65–92.
Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press, Oxford.
Lauritzen, S. L. & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17**, 31–54.
Leimer, H.-G. (1993). Optimal decomposition by clique separators. *Discrete Math.* **113**, 99–123.
Matúš, F. (1994). On the maximum-entropy extensions of probability measures over undirected graphs. *Proc. WUPES'94*, pp. 181–198. Institute of Information Theory and Automation, Trest.
Wermuth, N. (1976). Analogies between multiplicative models in contingency tables and covariance selection. *Biometrics* **32**, 95–108.
Wermuth, N. (1980). Linear recursive equations, covariance selection, and path analysis. *J. Amer. Statist. Assoc.* **75**, 963–997.
Wermuth, N. (1998). Graphical Markov models. In *Encyclopedia of statistical sciences* (eds S. Kotz, C. Read & D. Banks) update vol. 2, 284–300. Wiley, New York.
Wermuth, N. & Cox, D. R. (1998). On association models defined over independence graphs. *Bernoulli* (to appear).
Wermuth, N. & Lauritzen, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika* **70**, 537–552.

Received September 1997, in final form April 1998

D. R. Cox, Nuffield College, Oxford OX1 1NF, UK.