Context-specific independence in graphical log-linear models

Henrik Nyman\textsuperscript{1} · Johan Pensar\textsuperscript{1} · Timo Koski\textsuperscript{2} · Jukka Corander\textsuperscript{3}

Received: 10 September 2014 / Accepted: 2 July 2015 / Published online: 15 July 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract Log-linear models are the popular workhorses of analyzing contingency tables. A log-linear parameterization of an interaction model can be more expressive than a direct parameterization based on probabilities, leading to a powerful way of defining restrictions derived from marginal, conditional and context-specific independence. However, parameter estimation is often simpler under a direct parameterization, provided that the model enjoys certain decomposability properties. Here we introduce a cyclical projection algorithm for obtaining maximum likelihood estimates of log-linear parameters under an arbitrary context-specific graphical log-linear model, which needs not satisfy criteria of decomposability. We illustrate that lifting the restriction of decomposability makes the models more expressive, such that additional context-specific independencies embedded in real data can be identified. It is also shown how a context-specific graphical model can correspond to a non-hierarchical log-linear parameterization with a concise interpretation. This observation can pave way to further development of non-hierarchical log-linear models, which have been largely neglected due to their believed lack of interpretability.

Keywords Graphical model · Context-specific interaction model · Log-linear model · Parameter estimation

\textsuperscript{1} Henrik Nyman
hennyman@abo.fi

\textsuperscript{1} Department of Mathematics and Statistics, Åbo Akademi University, Turku, Finland
\textsuperscript{2} Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden
\textsuperscript{3} Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland
1 Introduction

Log-linear models for contingency tables have enjoyed a wide popularity since their introduction in the 1970’s, enabling a comprehensive approach to testing hypotheses of marginal and conditional independence, as well as more detailed global scrutiny of inter-dependencies within a set of discrete variables (Whittaker 1990; Lauritzen 1996). Graphical models have received most of the attention within the class of log-linear models, which is unsurprising given their interpretability and relative ease of model fitting. However, several other dependency structures with a log-linear representation have also been considered, such as hierarchical (Lauritzen 1996), pairwise interaction (Whittaker 1990), split (Højsgaard 2003), labeled (Corander 2003), and context-specific interaction models (Eriksen 1999; Højsgaard 2004).

Recently, Nyman et al. (2014) introduced a class of stratified graphical models (SGMs), where strata are defined locally in the outcome space such that a specific pair of variables are independent in the context defined by a combination of values of the joint neighbors of the two variables. This is in contrast to ordinary graphical models, where a pair of variables is always considered either conditionally independent or completely dependent given their joint neighbors.

The work of Nyman et al. (2014) generalizes the results on labeled graphical models, introduced in Corander (2003). To be able to obtain an analytical expression for Bayesian model scoring of SGMs, Nyman et al. (2014) restricted their attention to a class of decomposable models under a direct parameterization of the probabilities (rather than log-linear parameterization), similar to the class of graphical models where the majority of model learning approaches have been devised under the assumption of excluding non-chordal graphs from the search space. Despite of the assumption of decomposability, the resulting model class was shown to be expressive for real data and Nyman et al. (2015) additionally illustrated that SGMs can lead to more accurate probabilistic classifiers than those based on standard graphical models.

Since the assumption of decomposability is generally made for computational convenience, rather than being motivated by data met in real applications, it is desirable to develop theory which enables fitting of context-specific graphical log-linear models irrespectively of them being decomposable or non-decomposable. Using the general estimation theory from Csiszár (1975) and Rudas (1998), we introduce a cyclical projection algorithm which can be used to obtain the maximum likelihood estimate for any context-specific graphical log-linear model. This result is of interest on its own, however, to also illustrate the increased expressiveness of unrestricted context-specific graphical log-linear models for real data, we combine the maximum likelihood estimation with approximate Bayesian model scoring to define a search algorithm for the optimal model for a given data set. We additionally briefly illustrate the fact that some context-specific graphical models are non-hierarchical log-linear models. This is particularly illuminating, since non-hierarchical log-linear models have generally been avoided due to a believed lack of apparent interpretation of the parameter restrictions.

The remainder of this paper is structured as follows. In the next section, we consider graphical log-linear models and the restrictions imposed by a graphical model to the log-linear parameters. We also introduce stratified graphical log-linear models, a class of models that allow for dependence structures containing context-specific
independencies. In Sect. 3, we introduce a projection algorithm based on the work of Csiszár (1975) and Rudas (1998). This algorithm enables the calculation of the maximal likelihood estimate of the log-linear parameters satisfying the restrictions imposed by a stratified graphical log-linear model. In Sect. 4, we devise an approximate Bayesian model optimization algorithm based on the Bayesian information criterion. The algorithm is illustrated by application to real data in Sect. 5. The final section provides some remarks and possibilities for future work.

2 Stratified graphical log-linear models

For a comprehensive account of the statistical and computational theory of probabilistic graphical models, see Whittaker (1990), Lauritzen (1996), and Koller and Friedman (2009). For an introduction to stratified graphical models and a synopsis of the notations and definitions used in this paper see Nyman et al. (2014). It is assumed throughout this article that all considered variables are binary. However, the introduced theory can readily be extended to general categorical variables.

An ordinary graphical model (GM) consists of an undirected graph and a probability distribution. The graph consists of a set of nodes $\Delta$ and a set of edges $E \subseteq \{\Delta \times \Delta\}$. The nodes in the graph correspond to the set of stochastic variables $X_{\Delta}$ governed by the probability distribution. An absent edge $\{\delta, \gamma\}$ in the graph corresponds to a conditional independence $X_\delta \perp X_\gamma \mid X_{\Delta \setminus \{\delta, \gamma\}}$, while $X_\delta$ and $X_\gamma$ are marginally independent if there exists no path between $\delta$ and $\gamma$. As an example we can consider the two graphs in Fig. 1a, b. The graph in Fig. 1a is complete, which means that there exists no marginal or conditional independencies among the three variables $X_1$, $X_2$, and $X_3$. In Fig. 1b the edge $\{X_2, X_3\}$ is absent, meaning that $X_2 \perp X_3 \mid X_1$.

Throughout this paper we will use the notation $X_A$ for the outcome space of the variables $X_A$, where $A \subseteq \Delta$. An element in this outcome space will be denoted by $x_A$. We will also use two different parameterizations for a probability distribution. Firstly, the standard parameterization used for a categorical distribution, where each parameter $\theta_i$ in a parameter vector $\theta$ denotes the probability of a specific outcome $x^{(i)}_{\Delta} \in X_{\Delta}$, i.e. $P(X_{\Delta} = x^{(i)}_{\Delta}) = \theta_i$. And secondly, the log-linear parameterization (Whittaker 1990; Lauritzen 1996) defined by the parameter vector $\phi$. For this parameterization, the joint distribution of the variables $X_{\Delta}$ is defined by

$$\log P(X_{\Delta} = x_{\Delta}) = \sum_{A \subseteq \Delta} \phi_{A(x_A)},$$

with the restriction that if $x_j = 0$ for some $j \in A$, then $\phi_{A(x_A)} = 0$.

![Fig. 1](image-url) In a and b two undirected graphs, in c a stratified graph
The main reason for using the log-linear parameters is that marginal and conditional independencies impose clear and simple restrictions on $\phi$. For example, consider the graph in Fig. 1b and the induced conditional independence $X_2 \perp X_3 \mid X_1$. Under the assumption that the considered variables are binary, the conditional independence is equivalent with

$$P(X_3 = x_3 \mid X_2 = 0, X_1 = x_1) = P(X_3 = x_3 \mid X_2 = 1, X_1 = x_1).$$

In turn, this condition is equivalent with the equality

$$\frac{P(X_3 = 0 \mid X_2 = 0, X_1 = x_1)}{P(X_3 = 1 \mid X_2 = 0, X_1 = x_1)} = \frac{P(X_3 = 0 \mid X_2 = 1, X_1 = x_1)}{P(X_3 = 1 \mid X_2 = 1, X_1 = x_1)} \quad (1)$$

Re-writing the conditional probabilities

$$P(X_3 = x_3 \mid X_2 = x_2, X_1 = x_1) = \frac{P(X_3 = x_3, X_2 = x_2, X_1 = x_1)}{P(X_2 = x_2, X_1 = x_1)}$$

and observing that the denominators in (1) cancel each other out results in

$$\frac{P(X_3 = 0, X_2 = 0, X_1 = x_1)}{P(X_3 = 1, X_2 = 0, X_1 = x_1)} = \frac{P(X_3 = 0, X_2 = 1, X_1 = x_1)}{P(X_3 = 1, X_2 = 1, X_1 = x_1)} \quad (2)$$

Next, we take the logarithm of both sides of Eq. (2) and express the probabilities in terms of the log-linear parameters. This results in the equation

$$(\phi_{\emptyset} + \phi_1(x_1)) - (\phi_{\emptyset} + \phi_1(x_1) + \phi_3(1) + \phi_{1,3}(x_1,1))$$

$$= (\phi_{\emptyset} + \phi_1(x_1) + \phi_2(1) + \phi_{1,2}(x_1,1))$$

$$- (\phi_{\emptyset} + \phi_1(x_1) + \phi_2(1) + \phi_3(1) + \phi_{1,2}(x_1,1) + \phi_{1,3}(x_1,1) + \phi_{2,3}(1,1))$$

$$+ \phi_{1,2,3}(x_1,1,1))$$

$$\iff \phi_{2,3}(1,1) + \phi_{1,2,3}(x_1,1,1) = 0$$

The conditional independence must hold for all possible values of $x_1$. If $x_1 = 0$ we automatically get $\phi_{1,2,3}(0,1,1) = 0$, resulting in the restriction $\phi_{2,3}(1,1) = 0$. For $x_1 = 1$ we get the restriction $\phi_{2,3}(1,1) + \phi_{1,2,3}(1,1,1) = 0$, meaning that $\phi_{1,2,3}(1,1,1) = 0$. This example can be expanded into a general statement. In a graphical log-linear model, if the edge $\{\delta, \gamma\}$ is not present in the graph, all parameters $\phi_{A(x)}$, where $\{\delta, \gamma\} \subseteq A$, are equal to zero (Whittaker 1990). As we in this paper only consider binary variables a log-linear parameter will, from now on, be denoted using the convention $\phi_{A(x)} = \phi_A$, when no ambiguity exists.

Consider now a GM with the complete graph spanning three nodes $\{1, 2, 3\}$, specifying that there are no conditional independencies among the variables $X_1$, $X_2$, and $X_3$. However, if the probability $P(X_1 = 1, X_2 = x_2, X_3 = x_3)$ factorizes into the product $P(X_1 = 1)P(X_2 = x_2 \mid X_1 = 1)P(X_3 = x_3 \mid X_1 = 1)$ for all outcomes $x_2 \in \{0, 1\}, x_3 \in \{0, 1\}$, then a simplification of the joint distribution is hiding.
beneath the graph. This simplification can be included in the graph by adding a condition, or \textit{stratum}, to the edge \{2, 3\} specifying where the context-specific independence \(X_2 \perp X_3 \mid X_1 = 1\) of the two variables holds, as illustrated in Fig. 1c. The following is the formal definition of a stratum and a stratified graphical model (Nyman et al. 2014).

\textbf{Definition 1} \textit{Stratum} Let the pair \((G, P_\Delta)\) be a graphical model, where \(G\) is a chordal graph and \(P_\Delta\) is a probability distribution. For all \(\{\delta, \gamma\} \in E\), let \(L_{\{\delta, \gamma\}}\) denote the set of nodes adjacent to both \(\delta\) and \(\gamma\). For a non-empty \(L_{\{\delta, \gamma\}}\), let \(L_{\{\delta, \gamma\}}\) denote the set of outcomes \(x_{L_{\{\delta, \gamma\}}} \in X_{\{\delta, \gamma\}}\) for which \(X_\delta\) and \(X_\gamma\) are independent given \(X_{L_{\{\delta, \gamma\}}} = x_{L_{\{\delta, \gamma\}}}\), i.e. \(L_{\{\delta, \gamma\}} = \{x_{L_{\{\delta, \gamma\}}} \in X_{L_{\{\delta, \gamma\}}} : X_\delta \perp X_\gamma \mid X_{L_{\{\delta, \gamma\}}} = x_{L_{\{\delta, \gamma\}}}\}\).

\textbf{Definition 2} \textit{Stratified graphical model (SGM)} A stratified graphical model is defined by the triple \((G, L, P_\Delta)\), where \(G\) is a chordal graph termed as the underlying graph, \(L\) equals the joint collection of all strata \(L_{\{\delta, \gamma\}}\) for the edges of \(G\), and \(P_\Delta\) is a joint distribution over the variables \(X_\Delta\) which factorizes according to the restrictions imposed by \(G\) and \(L\).

The pair \((G, L)\) consisting of the graph \(G\) with the stratified edges (edges associated with a stratum) determined by \(L\) will be referred to as a stratified graph (SG), usually denoted by \(G_L\).

The requirement that \(G\) is chordal is necessary for the definition of a stratum to be generally applicable. Consider the graph in Fig. 2, note that the graph is not chordal as it contains the chord-less cycle \((1, 3, 4, 2, 1)\). The intended context-specific independence \(X_3 \perp X_4 \mid X_5 = 1\) does not hold as nodes 3 and 4 are connected via the path \((3, 1, 2, 4)\). By definition, no such paths are possible for chordal graphs, ensuring that given \(x_{L_{\{\delta, \gamma\}}} \in L_{\{\delta, \gamma\}}\) it will hold that \(X_\delta \perp X_\gamma \mid X_{L_{\{\delta, \gamma\}}} = x_{L_{\{\delta, \gamma\}}}\). The definition of a decomposable SG (Nyman et al. 2014) is given in “Appendix 1”. An SG that is not necessarily decomposable is referred to as an unrestricted stratified graph.

As we mentioned earlier, the use of log-linear parameters is motivated by the clear restrictions imposed by a graph to the parameters. We will now show that the restrictions imposed by the introduction of strata are also clearly defined.

\textbf{Theorem 1} Define the operator \(\mathcal{D}(A; B)\) on the two sets \(A\) and \(B\) as \(\mathcal{D}(A; B) = \{A \cup C : C \subseteq B\}\). Consider the context-specific independence \(X_\delta \perp X_\gamma \mid X_{L_{\{\delta, \gamma\}}} = x_{L_{\{\delta, \gamma\}}}\). Let \(L_Z \subseteq L_{\{\delta, \gamma\}}\) denote the set of nodes with non-zero values in \(x_{L_{\{\delta, \gamma\}}}\). The restrictions to the log-linear parameters induced by the context-specific independence \(X_\delta \perp X_\gamma \mid X_{L_{\{\delta, \gamma\}}} = x_{L_{\{\delta, \gamma\}}}\) are then given by the condition

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{example.png}
\caption{Non-chordal graph resulting in the intended context-specific independence \(X_3 \perp X_4 \mid X_5 = 1\) not holding}
\end{figure}
\[ \sum_{a \subseteq \mathcal{D}(\delta, \gamma); L_Z} \phi_a = 0, \]

**Proof** This proof follows the same outline as the example where we illustrated the restrictions caused by absent edges in a graph. Let \( \Omega = \Delta \setminus \{ L_{(\delta, \gamma)} \cup \{ \delta, \gamma \} \} \) denote the set of nodes not in \( L_{(\delta, \gamma)} \) or \( \{ \delta, \gamma \} \). The context-specific independence statement \( X_\delta \perp X_\gamma \mid X_{L_{(\delta, \gamma)}} = x_{L_{(\delta, \gamma)}} \) is equivalent with the equality

\[
P(X_\delta = 0 \mid X_\gamma = 0, X_{L_{(\delta, \gamma)}} = x_{L_{(\delta, \gamma)}}, X_\Omega = x_\Omega) = P(X_\delta = 1 \mid X_\gamma = 0, X_{L_{(\delta, \gamma)}} = x_{L_{(\delta, \gamma)}}, X_\Omega = x_\Omega)
\]

\[
\iff
\]

\[
P(X_\delta = 0, X_\gamma = 0, X_{L_{(\delta, \gamma)}} = x_{L_{(\delta, \gamma)}}, X_\Omega = x_\Omega)
= P(X_\delta = 1, X_\gamma = 0, X_{L_{(\delta, \gamma)}} = x_{L_{(\delta, \gamma)}}, X_\Omega = x_\Omega)
= P(X_\delta = 0, X_\gamma = 1, X_{L_{(\delta, \gamma)}} = x_{L_{(\delta, \gamma)}}, X_\Omega = x_\Omega)
= P(X_\delta = 1, X_\gamma = 1, X_{L_{(\delta, \gamma)}} = x_{L_{(\delta, \gamma)}}, X_\Omega = x_\Omega).
\]

Let \( Z \) denote the set of nodes corresponding to variables with non-zero outcome in \( x_{L_{(\delta, \gamma)}} \) or \( x_\Omega \). Taking the logarithm of both sides of this equation and writing the probabilities using log-linear parameters results in

\[
\sum_{a \subseteq \mathcal{D}(\emptyset; Z)} \phi_a = \sum_{a \subseteq \mathcal{D}(\emptyset; \delta \cup Z)} \phi_a = \sum_{a \subseteq \mathcal{D}(\emptyset; \gamma \cup Z)} \phi_a = \sum_{a \subseteq \mathcal{D}(\emptyset; \delta, \gamma \cup Z)} \phi_a.
\]

\[
\Leftrightarrow \sum_{a \subseteq \mathcal{D}(\emptyset; Z)} \phi_a = \sum_{a \subseteq \mathcal{D}(\emptyset; \delta \cup Z)} \phi_a \Leftrightarrow \sum_{a \subseteq \mathcal{D}(\emptyset; \delta, \gamma \cup Z)} \phi_a = 0.
\]

However, if a node \( \zeta \in \Omega \) it is not adjacent to both \( \delta \) and \( \gamma \). Consequently, any parameter \( \phi_A \) such that \( \{ \delta, \gamma, \zeta \} \subseteq A \) is restricted to zero. Therefore, if \( L_Z \) denotes the nodes corresponding to the variables with non-zero outcome in \( x_{L_{(\delta, \gamma)}} \), the restriction induced by the considered stratum can be written

\[
\sum_{a \subseteq \mathcal{D}(\emptyset; \delta, \gamma); L_Z} \phi_a = 0.
\]

\( \square \)

As an example consider again the SG in Fig. 1c. The context-specific independence \( X_2 \perp X_3 \mid X_1 = 1 \) induces the log-linear parameter restriction

\[
\phi_{\emptyset} + \phi_1 - \phi_{\emptyset} - \phi_1 - \phi_2 - \phi_{1,2}
= \phi_{\emptyset} + \phi_1 + \phi_3 + \phi_{1,3} - \phi_{\emptyset} - \phi_1 - \phi_2 - \phi_3 - \phi_{1,2} - \phi_{1,3} - \phi_{2,3} - \phi_{1,2,3}
\Leftrightarrow \phi_{2,3} + \phi_{1,2,3} = 0.
\]
In the definition of a stratum on the edge \(\{\delta, \gamma\}\), the variables that determine the stratum correspond to the nodes that are adjacent to both \(\delta\) and \(\gamma\). This is a natural definition rather than an invented restriction.

**Theorem 2** Only the variables corresponding to nodes adjacent to both \(\delta\) and \(\gamma\) may define a context-specific independence between \(X_{\delta}\) and \(X_{\gamma}\).

**Proof** The proof of this theorem follows from Theorem 1. We assume that a variable \(X_\zeta\), such that the node \(\zeta\) is not adjacent to both \(\delta\) and \(\gamma\), is included when defining the context-specific independence \(X_{\delta} \perp X_{\gamma} \mid X_L(\{\delta, \gamma\}) = x_L(\{\delta, \gamma\})\), \(X_\zeta = x_\zeta\). If \(x_\zeta = 0\), we would get the same restriction as by not including \(X_\zeta\) in the conditioning set

\[
\sum_{a \subseteq D(\{\delta, \gamma\}; L; \zeta)} \phi_a = 0.
\]

If \(x_\zeta \neq 0\) we get the restriction

\[
\sum_{a \subseteq D(\{\delta, \gamma\}; L; \zeta)} \phi_a = 0,
\]

again we know from the underlying graph that any parameter \(\phi_A\) such that \(\{\delta, \gamma, \zeta\} \subseteq A\) is restricted to zero, resulting in (3) and (4) being equivalent restrictions. \(\Box\)

As an example consider the graphs in Fig. 3a, b. Here \(X_3\) determines the stratum of the edge \(\{1, 2\}\), although node 2 and node 3 are non-adjacent. The underlying graph establishes that \(X_2 \perp X_3 \mid X_1\). However, given the proposed stratum \(X_3\) can indirectly affect \(X_2\) by determining whether or not \(X_1\) and \(X_2\) are dependent, which is an obvious contradiction. The underlying graph induces the parameter restrictions \(\phi_{2,3} = \phi_{1,2,3} = 0\). The stratum included in Fig. 3a results in the restriction \(\phi_{1,2} = 0\), while the stratum included in Fig. 3b results in the restriction \(\phi_{1,2} + \phi_{1,2,3} = 0\), leading to \(\phi_{1,2} = 0\). This means that all the graphs in Fig. 3 would induce the same restrictions, \(\phi_{1,2} = \phi_{2,3} = \phi_{1,2,3} = 0\). Note that this example is a special case of Theorem 2 as \(L_{\{1,2\}} = \emptyset\).

We conclude this section on stratified graphical log-linear models with an interesting observation. Whittaker (1990) termed a log-linear model as hierarchical if, whenever a parameter \(\phi_a = 0\) then \(\phi_t = 0\) for all \(a \subseteq t\). Whittaker (1990, p. 209) further states that “A non-hierarchical model is not necessarily uninteresting; it is just that the focus of interest is something other than independence”. This statement does not

![Fig. 3](image-url)
apply to SGMs which are, in fact, a class of non-hierarchical models. Consider, for instance, the SG attained by replacing the stratum $X_2 \perp X_3 \mid X_1 = 1$ in Fig. 1c with $X_2 \perp X_3 \mid X_1 = 0$. This leads to the single parameter restriction $\phi_{2,3} = 0$. As the parameter $\phi_{1,2,3}$ is unrestricted the model is clearly non-hierarchical.

### 3 Parameter estimation for stratified graphical log-linear models

In this section, we will derive a method that transforms an arbitrary distribution to the distribution that minimizes the Kullback–Leibler (KL) divergence, with respect to the original distribution, while satisfying the restrictions imposed by an arbitrary SG. Given that the original distribution corresponds to the maximum likelihood estimate of the parameters attained from an observed dataset, this transformation corresponds to finding the maximum likelihood estimate that satisfies the restrictions induced by the SG.

Let $\Theta_G$ denote the set of distributions satisfying the restrictions imposed by the chordal graph $G$. Lauritzen (1996, p. 91) showed that given an observed distribution $P$ the maximum likelihood (ML) projection to $\Theta_G$, resulting in the distribution $\hat{P}$, is obtained by setting

$$\hat{\theta}_i = \hat{P}(x^{(i)}_{\Delta}) = \frac{\prod_{C \in \mathcal{C}(G)} P(x^{(i)}_{C})}{\prod_{S \in \mathcal{S}(G)} P(x^{(i)}_{S})}, \quad i = 1, \ldots, |\mathcal{X}_\Delta|. \quad (5)$$

The sets $\mathcal{C}(G)$ and $\mathcal{S}(G)$ denote the maximal cliques and separators, respectively, of the chordal graph $G$ (Golumbic 2004). Given the following definition of the Kullback–Leibler divergence

$$D_{KL}(P, \hat{P}) = \sum_{x_\Delta \in \mathcal{X}_\Delta} \log \left( \frac{P(x_\Delta)}{\hat{P}(x_\Delta)} \right) P(x_\Delta),$$

Lauritzen (1996, p. 238) also showed that the ML projection corresponds to finding the distribution that minimizes $D_{KL}$ in the second argument, i.e.

$$\hat{P} = \arg \min_{Q \in \Theta_G} D_{KL}(P, Q).$$

We shall later refer to the minimum discrimination information (MDI) projection, resulting in a distribution $\hat{R}$ given a distribution $R$. The MDI projection is also defined through the KL divergence but in this case as the distribution that minimizes $D_{KL}$ in the first argument, i.e.

$$\hat{R} = \arg \min_{Q \in \Theta_G} D_{KL}(Q, R).$$

The ML projection for imposing a single context-specific independence on a distribution can also be written in closed form. Consider an outcome $x_{L[\delta, \gamma]} \in \mathcal{L}[\delta, \gamma],$
which implies the context-specific independence \( X_\delta \perp X_\gamma \mid X_{L(\delta, \gamma)} = x_{L(\delta, \gamma)} \). If we by \( \Omega = \Delta \setminus \{L(\delta, \gamma) \cup \{\delta, \gamma\}\} \) denote all nodes not in \( L(\delta, \gamma) \) or \( \{\delta, \gamma\} \), the probability

\[
P(X_\Delta = x_\Delta) = P(X_{L(\delta, \gamma)} = x_{L(\delta, \gamma)}, X_\Omega = x_\Omega, X_\delta = x_\delta, X_\gamma = x_\gamma)
\]

for any \( x_\Omega \in X_\Omega \) can be factorized as

\[
P(X_{L(\delta, \gamma)} = x_{L(\delta, \gamma)}, X_\Omega = x_\Omega)P(X_\delta = x_\delta \mid X_{L(\delta, \gamma)} = x_{L(\delta, \gamma)}, X_\Omega = x_\Omega)
\]

\[
P(X_\gamma = x_\gamma \mid X_{L(\delta, \gamma)} = x_{L(\delta, \gamma)}, X_\Omega = x_\Omega).
\]

Using the following abbreviated notation for \( \theta \) (and correspondingly for \( \hat{\theta} \))

\[
\begin{align*}
\theta_{0,0} &= P(X_{L(\delta, \gamma)} = x_{L(\delta, \gamma)}, X_\Omega = x_\Omega, X_\delta = 0, X_\gamma = 0), \\
\theta_{0,1} &= P(X_{L(\delta, \gamma)} = x_{L(\delta, \gamma)}, X_\Omega = x_\Omega, X_\delta = 0, X_\gamma = 1), \\
\theta_{1,0} &= P(X_{L(\delta, \gamma)} = x_{L(\delta, \gamma)}, X_\Omega = x_\Omega, X_\delta = 1, X_\gamma = 0), \\
\theta_{1,1} &= P(X_{L(\delta, \gamma)} = x_{L(\delta, \gamma)}, X_\Omega = x_\Omega, X_\delta = 1, X_\gamma = 1),
\end{align*}
\]

we determine the values \( \hat{\theta}_{0,0}, \hat{\theta}_{0,1}, \hat{\theta}_{1,0}, \) and \( \hat{\theta}_{1,1} \) according to

\[
\begin{align*}
\hat{\theta}_{0,0} &= (\theta_{0,0} + \theta_{0,1}) \cdot (\theta_{0,0} + \theta_{1,0}) / (\theta_{0,0} + \theta_{0,1} + \theta_{1,0} + \theta_{1,1}), \\
\hat{\theta}_{0,1} &= (\theta_{0,0} + \theta_{0,1}) \cdot (\theta_{0,1} + \theta_{1,1}) / (\theta_{0,0} + \theta_{0,1} + \theta_{1,0} + \theta_{1,1}), \\
\hat{\theta}_{1,0} &= (\theta_{1,0} + \theta_{1,1}) \cdot (\theta_{0,0} + \theta_{0,1}) / (\theta_{0,0} + \theta_{0,1} + \theta_{1,0} + \theta_{1,1}), \\
\hat{\theta}_{1,1} &= (\theta_{1,0} + \theta_{1,1}) \cdot (\theta_{0,1} + \theta_{1,1}) / (\theta_{0,0} + \theta_{0,1} + \theta_{1,0} + \theta_{1,1}).
\end{align*}
\]

A detailed derivation of the projection defined above is given in “Appendix 2”. Repeating the procedure defined in (6) for all \( x_\Omega \in X_\Omega \) will result in the ML projection of \( P \) satisfying the context-specific independence \( X_\delta \perp X_\gamma \mid X_{L(\delta, \gamma)} = x_{L(\delta, \gamma)} \).

By cyclically repeating the projections according to (5) and (6) for all instances found in the set of strata \( L \) until convergence is achieved, the resulting parameter vector will be the maximum likelihood estimate that simultaneously satisfies all the restrictions imposed by \( G_L \). In order to prove this we first need to define the following family of probability distributions.

**Definition 3** Let \( X_\delta \) and \( X_\gamma \) be two variables in \( X_\Delta \), \( X_A \) a subset of \( X_{\Delta \setminus \{\delta, \gamma\}} \) and \( X_\Omega = X_{\Delta \setminus \{A \cup \{\delta, \gamma\}\}} \). \( \mathcal{F}_{\delta, \gamma}(X_A = x_A, Q) \), where \( Q \) is an arbitrary probability distribution, is defined as the set of probability distributions for which the following properties hold for all possible values \( x_\delta, x_\gamma \) and \( x_\Omega \).

\[
\mathcal{F}_{\delta, \gamma}(X_A = x_A, Q) = \{ P : P(X_A = x_A, X_\Omega = x_\Omega) = Q(X_A = x_A, X_\Omega = x_\Omega) \}
\]

\[
\cap \{ P : P(X_\delta = x_\delta \mid X_A = x_A, X_\Omega = x_\Omega) = Q(X_\delta = x_\delta \mid X_A = x_A, X_\Omega = x_\Omega) \}
\]

\[
\cap \{ P : P(X_\gamma = x_\gamma \mid X_A = x_A, X_\Omega = x_\Omega) = Q(X_\gamma = x_\gamma \mid X_A = x_A, X_\Omega = x_\Omega) \}
\]

\[
\cap \{ P : P(X_\Delta = y_\Delta) = Q(X_\Delta = y_\Delta) \text{, when } y_\Delta \text{ is a outcome where } x_A \neq y_A \}.
\]
A set of probability distributions $\mathcal{C}$ is defined as a linear set if $P_1 \in \mathcal{C}$ and $P_2 \in \mathcal{C}$ results in $\alpha P_1 + (1 - \alpha) P_2$ also belonging to $\mathcal{C}$ for every real $\alpha$ for which it is a probability distribution (Csiszár 1975).

**Lemma 1** $\mathcal{F}_{\delta, \gamma}(X_A = x_A, \mathcal{Q})$ constitutes a linear set.

**Proof** Let $P_1$ and $P_2$ be two probability distributions in $\mathcal{F}_{\delta, \gamma}(X_A = x_A, \mathcal{Q})$, we then need to prove that $P^* = \alpha P_1 + (1 - \alpha) P_2$ also belongs to $\mathcal{F}_{\delta, \gamma}(X_A = x_A, \mathcal{Q})$. It is trivial to show that $P^*(X_A = x_A, X_\Omega = x_\Omega) = Q(X_A = x_A, X_\Omega = x_\Omega)$ and that $P^*(X_\Delta = y_\Delta) = Q(X_\Delta = y_\Delta)$, when $y_\Delta$ is a outcome where $x_A \neq y_A$. The non-trivial part consists of showing that $P^*(X_\delta = x_\delta \mid X_A = x_A, X_\Omega = x_\Omega) = Q(X_\delta = x_\delta \mid X_A = x_A, X_\Omega = x_\Omega)$.

We start from the fact that

$$Q(X_\delta = x_\delta \mid X_A = x_A, X_\Omega = x_\Omega) = P_1(X_\delta = x_\delta \mid X_A = x_A, X_\Omega = x_\Omega) = P_2(X_\delta = x_\delta \mid X_A = x_A, X_\Omega = x_\Omega),$$

which implicates that

$$\frac{P_1(X_\delta = x_\delta, X_A = x_A, X_\Omega = x_\Omega)}{P_1(X_A = x_A, X_\Omega = x_\Omega)} = \frac{P_2(X_\delta = x_\delta, X_A = x_A, X_\Omega = x_\Omega)}{P_2(X_A = x_A, X_\Omega = x_\Omega)}.$$

From the definition of $\mathcal{F}_{\delta, \gamma}(X_A = x_A, \mathcal{Q})$ we know that

$$P_1(X_A = x_A, X_\Omega = x_\Omega) = P_2(X_A = x_A, X_\Omega = x_\Omega),$$

and can therefore deduce that

$$Q(X_\delta = x_\delta, X_A = x_A, X_\Omega = x_\Omega) = P_1(X_\delta = x_\delta, X_A = x_A, X_\Omega = x_\Omega) = P_2(X_\delta = x_\delta, X_A = x_A, X_\Omega = x_\Omega).$$

For $P^*$ this means that

$$P^*(X_\delta = x_\delta \mid X_A = x_A, X_\Omega = x_\Omega) = \frac{\alpha P_1(X_\delta = x_\delta, X_A = x_A, X_\Omega = x_\Omega) + (1 - \alpha) P_2(X_\delta = x_\delta, X_A = x_A, X_\Omega = x_\Omega)}{\alpha P_1(X_A = x_A, X_\Omega = x_\Omega) + (1 - \alpha) P_2(X_A = x_A, X_\Omega = x_\Omega)} = \frac{\alpha Q(X_\delta = x_\delta, X_A = x_A, X_\Omega = x_\Omega) + (1 - \alpha) Q(X_\delta = x_\delta, X_A = x_A, X_\Omega = x_\Omega)}{Q(X_A = x_A, X_\Omega = x_\Omega)} = Q(X_\delta = x_\delta \mid X_A = x_A, X_\Omega = x_\Omega).$$

Of course the same reasoning can be used to show that $P^*(X_\gamma = x_\gamma \mid X_A = x_A, X_\Omega = x_\Omega) = Q(X_\gamma = x_\gamma \mid X_A = x_A, X_\Omega = x_\Omega)$, which concludes the proof.

**Definition 4** The log-linear model $LL_{\delta, \gamma}(X_A = x_A)$ is defined as the set of probability distributions which satisfy the condition $X_\delta \perp X_\gamma \mid X_A = x_A$.  

\[ \text{Springer} \]
In order to do this we turn to Csiszár, Matúš (2003, Theorem 1). This theorem states any further comment, but as it is not self-evident we have chosen to include a proof.

It is easy to see that for any probability distribution \( Q \), the sets \( LL_{\delta,\gamma}(X_A = x_A) \) and \( F_{\delta,\gamma}(X_A = x_A, Q) \) can have at most one common distribution, denoted by \( R \). It is also evident, using the same reasoning as in “Appendix 2”, that \( R \) is the result of the ML projection of any distribution in \( F_{\delta,\gamma}(X_A = x_A, Q) \) to \( LL_{\delta,\gamma}(X_A = x_A) \). We are now ready to prove the main theorem.

**Theorem 3** Cyclically projecting the observed distribution, \( P_0 \), in accordance with the procedures defined in (5) and (6) until convergence is achieved, will result in the maximum likelihood estimate, \( \hat{P} \), which simultaneously satisfies all the restrictions imposed by a given \( SG, G_L = (G, L) \).

**Proof** This proof uses the results found in Rudas (1998), with Theorem 2 of that paper being of paramount importance. An essential part of the proof is the so called Pythagorean identity for discrimination information, see for instance Rudas (1998), which states that if \( S \) belongs to a linear set and \( R \) is the MDI projection of a distribution \( T \) onto this set, then \( D_{KL}(S, T) = D_{KL}(S, R) + D_{KL}(R, T) \).

Let \( m \) denote the number of context-specific independencies in \( L \), i.e. the total number of instances in all strata included in \( L \). Further, let \( P_l \) be the distribution attained when projecting the distribution \( P_{l-1} \) according to the \( l \)th context-specific independence, say \( X_{\delta} \perp X_{\gamma} \mid X_{L(\delta,\gamma)} = x_{L(\delta,\gamma)} \), in \( L \). It then holds that

\[
P_l = LL_{\delta,\gamma}(X_L(\delta,\gamma) = x_{L(\delta,\gamma)}) \cap F_{\delta,\gamma}(X_L(\delta,\gamma) = x_{L(\delta,\gamma)}, P_{l-1}).
\]

\( P_l \) is also the MDI projection of any distribution in \( LL_{\delta,\gamma}(X_L(\delta,\gamma) = x_{L(\delta,\gamma)}) \) to \( F_{\delta,\gamma}(X_L(\delta,\gamma) = x_{L(\delta,\gamma)}, P_{l-1}) \). Rudas (1998) makes this statement without providing any further comment, but as it is not self-evident we have chosen to include a proof. In order to do this we turn to Csiszár, Matúš (2003, Theorem 1). This theorem states that for a log-convex set \( T \), which \( LL_{\delta,\gamma}(X_L(\delta,\gamma) = x_{L(\delta,\gamma)}) \) constitutes as it defines an exponential family, the ML projection, denoted by \( R \), of an arbitrary distribution \( S \) to \( T \) is the unique distribution that satisfies

\[
D_{KL}(S, T) \geq \min_{A \in T} D_{KL}(S, A) + D_{KL}(R, T), \quad T \in T.
\]

In our case, as \( \hat{P} \in LL_{\delta,\gamma}(X_L(\delta,\gamma) = x_{L(\delta,\gamma)}) \) and \( P_l \) is the ML projection of any distribution in \( F_{\delta,\gamma}(X_L(\delta,\gamma) = x_{L(\delta,\gamma)}, P_{l-1}) \) to \( LL_{\delta,\gamma}(X_L(\delta,\gamma) = x_{L(\delta,\gamma)}) \) it holds that

\[
D_{KL}(S, \hat{P}) \geq D_{KL}(S, P_l) + D_{KL}(P_l, \hat{P}), \quad S \in F_{\delta,\gamma}(X_L(\delta,\gamma) = x_{L(\delta,\gamma)}, P_{l-1}).
\]

Which implies that \( D_{KL}(S, \hat{P}) \geq D_{KL}(P_l, \hat{P}) \) holds for every \( S \) in \( F_{\delta,\gamma}(X_L(\delta,\gamma) = x_{L(\delta,\gamma)}, P_{l-1}) \) and \( P_l \) is the MDI projection of any distribution in \( LL_{\delta,\gamma}(X_L(\delta,\gamma) = x_{L(\delta,\gamma)}) \) to \( F_{\delta,\gamma}(X_L(\delta,\gamma) = x_{L(\delta,\gamma)}, P_{l-1}) \). Therefore the Pythagorean identity is applicable and we can conclude that

\[
D_{KL}(P_{l-1}, \hat{P}) = D_{KL}(P_{l-1}, P_l) + D_{KL}(P_l, \hat{P}). \tag{7}
\]
Rudas (1998) showed that the Pythagorean identity is also applicable when projecting a distribution onto the set of distributions satisfying the restrictions imposed by a chordal graph. I.e. if we by $P_{m+1}$ denote the distribution that results from projecting $P_m$ to $\Theta_G$ according to (5) we get that

$$DKL(P_m, \hat{P}) = DKL(P_m, P_{m+1}) + DKL(P_{m+1}, \hat{P}).$$

(8)

Combining (7) and (8) and letting the projection $n + i$ be the same projection as $i$ if $n = k(m + 1)$ for some value $k = 0, 1, \ldots$ results in

$$DKL(P_0, \hat{P}) = \sum_{l=1}^{n} DKL(P_{l-1}, P_l) + DKL(P_n, \hat{P}).$$

for every $n$. The existence of $\hat{P}$ implies that for any $n$

$$\sum_{l=1}^{n} DKL(P_{l-1}, P_l) < \infty,$$

which, in turn, implies that $DKL(P_{l-1}, P_l) \to 0$ as $l \to \infty$. Just as Rudas (1998) we can now refer to the compactness argument found in Csiszár (1975, Theorem 3.2) to complete the proof.

In practice we need a criterion to determine whether or not the cyclical projections have converged to $\hat{P}$. The criterion that we use terminates the projections once an entire cycle consisting of $m + 1$ projections has been completed with the total sum of changes made to $\theta$ being less than a predetermined constant $\epsilon$. Using $\theta_i = (\theta_{i1}, \ldots, \theta_{ik})$ to denote the parameter after the $i$:th projection in the cycle, with $\theta_0$ denoting the starting value. The cyclical projections are terminated when

$$\sum_{i=1}^{m+1} \sum_{j=1}^{k} |\theta_{ij} - \theta_{(i-1)j}| < \epsilon.$$

4 Bayesian learning of SGMs

Ideally, when determining how well a graph $G$ captures the dependence structure of a set of variables found in a dataset $X$ one calculates the marginal likelihood of the dataset given the graph, $P(X \mid G)$. To be able to calculate the marginal likelihood and perform model inference for SGs, Nyman et al. (2014) limited the model space to decomposable SGs. In this paper we introduce a method which allows us to remove these restrictions while retaining the ability to perform model inference.

Using the Bayesian information criterion (BIC) (Schwarz 1978) an approximation of the marginal likelihood of a dataset $X$, consisting of $n$ observations, can be attained for an arbitrary SG. This approximation can be written as
\[ \log P(X \mid G_L) \approx l(X \mid \hat{\theta}, G_L) - \frac{\dim(\Theta_{G_L})}{2} \log n. \]  

Here, \( \hat{\theta} \) is the maximum likelihood estimate of the model parameters under the restrictions imposed by \( G_L \), as calculated in Sect. 3. The logarithm of the likelihood function \( l(X \mid \hat{\theta}, G_L) \) is calculated as

\[ l(X \mid \hat{\theta}, G_L) = \sum_{x^{(i)} \in X} n_i \log \hat{\theta}_i, \]

where \( n_i \) is the number of observations of \( x^{(i)} \) found in \( X \). Finally, \( \dim(\Theta_{G_L}) \) is the maximum number of free parameters in a distribution with the parameter restrictions induced by \( G_L \). This number can readily be calculated using the log-linear parameterization, as discussed in Sect. 2. We denote the right hand side of (9) by \( \log S(G_L \mid X) \), i.e. \( P(X \mid G_L) \approx S(G_L \mid X) \).

When using context-specific or local dependence structures the graphs with optimal score, be it defined using the marginal likelihood or BIC, have a tendency to be very dense. This can in some cases have a negative effect on the interpretability of the dependence structure conveyed by the graph. In attempt to combat this negative development several ideas of how to penalize dense graphs have been considered (Friedman and Goldszmidt 1996; Nyman et al. 2014; Pensar et al. 2015). To this end, for the experiments conducted in the next section the following non-uniform prior (Nyman et al. 2014) over the model space is used

\[ P(G_L) \propto 2^{-\dim(\Theta_G)}. \]  

Here, \( \dim(\Theta_G) \) denotes the maximum number of free parameter in a distribution satisfying the restrictions imposed by the underlying graph \( G \).

In order to identify the stratified graph that optimizes the score function, \( S(G_L \mid X) P(G_L) \), a Markov chain that traverses the space of all considered SGs is implemented. The Markov chain follows the principles of the non-reversible Metropolis-Hastings algorithm (Corander et al. 2006, 2008; Nyman et al. 2014). The specific proposal functions used to identify the graphs presented in the next section are described in “Appendix 3”.

5 Illustration of SGM learning from data

In this section, we look at two different datasets and identify different types of models that optimize the score function \( S(G_L \mid X) P(G_L) \). In order to save space, when displaying an SG we instead of writing a stratum as \((X_1 = 1, X_2 = 0)\) only write \((1, 0)\). This is possible since, given the graph, it is clear which variables define the context-specific independence when the variables are ordered by their integer labels.

The first dataset that we have investigated includes prognostic factors for coronary heart disease and can be found in Edwards and Havránek (1985). The data consists of 1841 observations of the six variables listed in Table 1. In Fig. 4 two different
Table 1  Variables in coronary heart disease data

| Variable | Meaning                                      | Range                      |
|----------|----------------------------------------------|----------------------------|
| X_1      | Smoking                                      | No = 0, Yes = 1            |
| X_2      | Strenuous mental work                        | No = 0, Yes = 1            |
| X_3      | Strenuous physical work                      | No = 0, Yes = 1            |
| X_4      | Systolic blood pressure > 140                | No = 0, Yes = 1            |
| X_5      | Ratio of beta and alpha lipoproteins > 3     | No = 0, Yes = 1            |
| X_6      | Family anamnesis of coronary heart disease    | No = 0, Yes = 1            |

Fig. 4  Optimal SGs for heart data. In a the optimal ordinary graph is amended with optimal strata. In b the optimal unrestricted SG

SGs are displayed. The SG in Fig. 4a is obtained by first conducting a search for the optimal ordinary chordal graph and then identifying the optimal set of strata for that graph. The underlying graph has the score $-6732.84$, while the SG has the score $-6721.67$. Figure 4b contains the estimated optimal unrestricted SG, which has the score $-6713.24$. The underlying graph for this SG has the score $-6764.14$. Comparing these two graphs we can immediately see that graph in Fig. 4b is denser and induces more context-specific independencies. What is not apparent from a quick glance at the graphs is that the denser graph induces a distribution with fewer free parameters. The SG in Fig. 4a induces a distribution with 10 free parameters compared to 8 for the SG in Fig. 4b.

The following context-specific independencies are included in the SG in Fig. 4a. Smoking and systolic blood pressure are independent given that the ratio of beta and alpha lipoproteins is larger than 3. Systolic blood pressure and the ratio of beta and alpha lipoproteins are independent given that a person does not smoke. And finally, smoking and the ratio of beta and alpha lipoproteins are independent given that the systolic blood pressure is less than 140 and it is known whether or not a person has a job that requires strenuous physical work. While it is also possible to determine the context-specific independencies directly from the SG in Fig. 4b, the dense structure of the graph makes this task more difficult. In order to attain more sparse graphs one could consider using a prior that further penalizes dense graphs.

In a cross-validation experiment designed to evaluate the validity of different models we compare the complete graph with the ordinary graph, decomposable SG, and the unrestricted SG (Fig. 4b) which optimize the score function. To perform the cross-validation the data is split into ten equally large sets, $X^j$, $j = 1, \ldots, 10$. Each set
is in turn used as test data while the remaining sets are used as training data. Using only the training data the maximum likelihood estimate of the model parameters, \( \hat{\theta}^j \), is calculated using the method described in Sect. 3. Given \( \hat{\theta}^j \) the sum of the log probabilities of the test data sets is calculated as

\[
p = \sum_{j=1}^{10} \sum_{x^{(i)} \in X^j} n_i^j \log \hat{\theta}_i^j,
\]

where \( n_i^j \) denotes the number of observations of \( x^{(i)} \) found in \( X^j \). In order to get reliable results, the partitioning of the data and subsequent calculations are repeated 1000 times and the mean of the resulting values of \( p \) determined according to

\[
\text{CVScore} = \frac{1}{1000} \sum_{k=1}^{1000} p_k.
\]

Unlike the BIC score, the CVScore is not equivalent with the marginal likelihood as the number of observations tends to infinity. It can, however, be used as a tool to evaluate the predictive properties of a model given a limited amount of data. The results of the cross-validation experiment is given in Table 2. As the results show, the distribution based on the unrestricted SG has the best predictive properties compared to the other models while having the least number of free parameters. This illustrates that stratified graphs can capture the dependence structure of a set of variables given a set of observations in a very satisfying manner.

The second dataset that we consider is derived from the answers given by 1806 candidates in the Finnish parliament elections of 2011, in a questionnaire issued by the newspaper Helsingin Sanomat (Helsingin 2011). The eight questions considered, represented by eight variables, are given in “Appendix 4”. As in the previous section we present in Fig. 5 two different SGs, the SG resulting from first determining the optimal ordinary graph and finding the optimal set of strata for that graph and the optimal unrestricted SG. For the optimal unrestricted SG we, instead of displaying the exact strata, give the total number of instances included in the stratum associated with each edge. The score for the underlying graph of Fig. 5a is \(-7177.69\) and for the SG \(-7162.78\). The corresponding scores for the graph in Fig. 5b are \(-7245.11\) and \(-7139.13\).
Fig. 5  Optimal SGs for parliament election data. In a the optimal ordinary graph is amended with optimal strata. In b the optimal unrestricted SG with number of instances in each strata listed beside the corresponding edge.

Table 3  Cross-validation results of parliament election data

|                  | Complete graph | Ordinary graph | Locally optimal SG | Unrestricted SG |
|------------------|----------------|----------------|--------------------|-----------------|
| Edges            | 28             | 10             | 10                 | 16              |
| Free parameters  | 255            | 21             | 16                 | 12              |
| BIC score        | −8089.21       | −7177.69       | −7162.78           | −7139.13        |
| CVScore          | −712.14         | −711.18        | −711.22            | −708.60         |

Again, to test the predictive accuracy of the obtained model we turn to the cross-validation test defined above. In this case we compare the complete graph to the optimal ordinary graph, the graph in Fig. 5a (locally optimal SG) and the optimal unrestricted stratified graph (Fig. 5b). The results are given in Table 3 and offer the same conclusions as the heart disease dataset results. The optimal unrestricted stratified graphical model induces a distribution with fewer free parameters while offering better predictive properties compared to the other graphical models.

These examples demonstrate that when using ordinary graphical models, variables that would be considered conditionally dependent may in fact be independent in certain contexts. The examples also show that the optimal unrestricted SG contains more edges than the optimal ordinary graph. This can be accredited to the fact that when using dense graphs the set of available parameter restrictions grows, while adding strata to a dense graph can still result in models that induce distributions with few free parameters. To avoid this it would be possible to apply a stronger prior over the model space, further penalizing dense graphs or graphs with many strata as done in Pensar et al. (2015). In conclusion, these experimental results show that context-specific independencies seem to occur naturally in various datasets therefore it can be very useful to consider graphical models that are able to capture such dependence structures.
6 Discussion

Graphical models, and log-linear models more generally, are useful for many types of multivariate analysis due to their interpretability. The context-specific graphical log-linear models discussed here extend the expressiveness of the stratified models considered earlier in Nyman et al. (2014) by removing the restriction concerning overlap of strata. By applying the general estimation theory developed in Rudas (1998) and Csiszár (1975), we were able to derive a consistent procedure for estimating the parameters of a context-specific graphical log-linear model based on cyclical projections each corresponding to a specific independence restriction. Two examples with real data illustrated how the relaxation of the model class properties enables additional discovery of context-specific independencies. In future research, it would be interesting to attempt to identify further classes of non-hierarchical restrictions to log-linear parameters, such that interpretability is maintained in the same fashion as for the current context-specific models.

Acknowledgments The authors would like to thank the editor and the anonymous reviewers for their constructive comments and suggestions on the original version of this paper. H. N. and J. P. were supported by the Foundation of Åbo Akademi University, as part of the Grant for the Center of Excellence in Optimization and Systems Engineering. J. P. was also supported by the Magnus Ehrnrooth foundation. J. C. was supported by the ERC Grant No. 239784 and academy of Finland Grant No. 251170. T. K. was supported by a grant from the Swedish research council VR/NT.

Appendix 1

Definition 5 Decomposable SG Let \((G, L)\) constitute an SG with \(G\) being chordal. Further, let \(E_L\) denote the set of all stratified edges, \(E_C\) the set of all edges in the maximal clique \(C\), and \(E_S\) the set of all edges in the separators of \(G\). The SG is defined as decomposable if

\[ E_L \cap E_S = \emptyset, \]

and

\[ E_L \cap E_C = \emptyset \quad \text{or} \quad \bigcap_{\{\delta, \gamma\} \in E_L \cap E_C} \{\delta, \gamma\} \neq \emptyset \quad \text{for all} \quad C \in \mathcal{C}(G). \]

An SGM where \((G, L)\) constitutes a decomposable SG is termed a decomposable SGM.

Appendix 2

Derivation of the parameters in Eq. (6).

We will here give a more detailed explanation of how \(\hat{\theta}_{0,0} = \hat{P}(X_L|\delta, \gamma) = x_L|\delta, \gamma, X_\Omega = x_\Omega, X_\delta = 0, X_\gamma = 0)\) is derived. It is generally possible to use the factorization
\[ P(X_{L[\delta, \gamma]} = x_{L[\delta, \gamma]}, X_{\Omega} = x_{\Omega}, X_{\delta} = 0, X_{\gamma} = 0) \]
\[ = P(X_{L[\delta, \gamma]} = x_{L[\delta, \gamma]}, X_{\Omega} = x_{\Omega}) P(X_{\delta} = 0, X_{\gamma} = 0 \mid X_{L[\delta, \gamma]} = x_{L[\delta, \gamma]}, X_{\Omega} = x_{\Omega}) \]
\[ = x_{L[\delta, \gamma]}, X_{\Omega} = x_{\Omega}). \]

When considering a probability distribution where \( \delta \) and \( \gamma \) can be dependent, it is generally not true that \( P(X_{\delta}, X_{\gamma}) = P(X_{\delta})P(X_{\gamma}) \). A standard result, see e.g. Whitaker (1990), states that for a distribution where two variables are dependent the ML projection to the set of distributions where the variables are independent is obtained by calculating the product of the marginal probabilities of the two variables. This implies, in our case, creating a new distribution \( \hat{P} \) according to

\[ \hat{P}(X_{L[\delta, \gamma]} = x_{L[\delta, \gamma]}, X_{\Omega} = x_{\Omega}, X_{\delta} = 0, X_{\gamma} = 0) \]
\[ = P(X_{L[\delta, \gamma]} = x_{L[\delta, \gamma]}, X_{\Omega} = x_{\Omega}) P(X_{\delta} = 0 \mid X_{L[\delta, \gamma]} = x_{L[\delta, \gamma]}, X_{\Omega} = x_{\Omega}) \]
\[ P(X_{\gamma} = 0 \mid X_{L[\delta, \gamma]} = x_{L[\delta, \gamma]}, X_{\Omega} = x_{\Omega}). \]

Using the earlier introduced notations this corresponds to setting

\[ \hat{\theta}_{0,0} = \hat{P}(X_{L[\delta, \gamma]} = x_{L[\delta, \gamma]}, X_{\Omega} = x_{\Omega}, X_{\delta} = 0, X_{\gamma} = 0) \]
\[ = P(X_{L[\delta, \gamma]} = x_{L[\delta, \gamma]}, X_{\Omega} = x_{\Omega}) P(X_{\delta} = 0 \mid X_{L[\delta, \gamma]} = x_{L[\delta, \gamma]}, X_{\Omega} = x_{\Omega}) \]
\[ P(X_{\gamma} = 0 \mid X_{L[\delta, \gamma]} = x_{L[\delta, \gamma]}, X_{\Omega} = x_{\Omega}) \]
\[ = (\theta_{0,0} + \theta_{0,1} + \theta_{1,0} + \theta_{1,1}) \cdot (\theta_{0,0} + \theta_{0,1} + \theta_{1,0} + \theta_{1,1}) / (\theta_{0,0} + \theta_{0,1} + \theta_{1,0} + \theta_{1,1}) \]
\[ \cdot (\theta_{0,0} + \theta_{1,0}) / (\theta_{0,0} + \theta_{0,1} + \theta_{1,0} + \theta_{1,1}) \]
\[ = (\theta_{0,0} + \theta_{0,1}) \cdot (\theta_{0,0} + \theta_{1,0}) / (\theta_{0,0} + \theta_{0,1} + \theta_{1,0} + \theta_{1,1}). \]

The other parameters \( \hat{\theta}_{0,1}, \hat{\theta}_{1,0}, \) and \( \hat{\theta}_{1,1} \) can be derived in a similar fashion.

**Appendix 3**

Proposal functions used for model optimization.

The search for the optimal stratified graph is conducted using two separate Markov chains. One Markov chain is used to traverse different underlying graphs. A second chain is used to identify the optimal set of strata given the underlying graph. Combining these two searches will ultimately result in the discovery of the optimal SG.

Using the proposal function defined in Algorithm 1, running a sufficient amount of iterations, we can be assured to find the optimal set of strata for any chordal graph.

**Algorithm 1** Proposal function for finding optimal strata for a chordal graph.

Let \( G \) denote the underlying graph. By \( L_{A} \) we denote all possible instances that can be added to any stratum of \( G \). If \( L_{A} \) is empty no strata may be added to \( G \) and the algorithm is terminated. \( L \) denotes the current state with \( L \) being empty in the starting state.

1. Set the candidate state \( L^{*} = L \).
2. Perform one of the following steps.
   2.1. If \( L \) is empty add a randomly chosen instance from \( L_A \) to \( L^* \).
   2.2. Else if \( \{L_A \setminus L\} \) is empty remove a randomly chosen instance from \( L^* \).
   2.3. Else with probability 0.5 add a randomly chosen instance from \( \{L_A \setminus L\} \) to \( L^* \).
   2.4. Else remove a randomly chosen instance from \( L^* \).

Using this proposal function the optimal set of strata can be found for any underlying graph and we can proceed to the search for the best underlying graph. The proposal function in Algorithm 2 is used for this task.

**Algorithm 2** Proposal function used to find the optimal underlying chordal graph.

The starting state is set to be the graph containing no edges. Let \( G \) denote the current graph with \( G_L = (G, L) \) being the stratified graph with underlying graph \( G \) and optimal set of strata \( L \).

1. Set the candidate state \( G^* = G \).
2. Randomly choose a pair of nodes \( \delta \) and \( \gamma \). If the edge \( \{\delta, \gamma\} \) is present in \( G^* \) remove it, otherwise add the edge \( \{\delta, \gamma\} \) to \( G^* \).
3. While \( G^* \) is non-chordal repeat steps 1 and 2.

The resulting candidate state \( G^* \) is used along with the corresponding optimal set of strata \( L^* \) to form the stratified graph \( G_L^* = (G^*, L^*) \) which is used when calculating the acceptance probability.

**Appendix 4**

Questions considered in parliament election data.

1. Since the mid-1990’s the income differences have grown rapidly in Finland. How should we react to this?
   0—The income differences do not need to be narrowed.
   1—The income differences need to be narrowed.
2. Should homosexual couples have the same rights to adopt children as heterosexual couples?
   0—Yes.
   1—No.
3. Child benefits are paid for each child under the age of 18 living in Finland, independent of the parents’ income. What should be done about child benefits?
   0—The income of the parents should not affect the child benefits.
   1—Child benefits should be dependent on parents’ income.
4. In Finland military service is mandatory for all men. What is your opinion on this?
   0—The current practice should be kept or expanded to also include women.
   1—The military service should be more selective or abandoned altogether.
5. Should Finland in its affairs with China and Russia more actively debate issues regarding human rights and the state of democracy in these countries?
   0—Yes.
   1—No.
6. Russia has prohibited foreigners from owning land close to the borders. In recent years, Russians have bought thousands of properties in Finland. How should Finland react to this?
   0—Finland should not restrict foreigners from buying property in Finland.
   1—Finland should restrict foreigners’ rights to buy property and land in Finland.

7. During recent years municipalities have outsourced many services to privately owned companies. What is your opinion on this?
   0—Outsourcing should be used to an even higher extent.
   1—Outsourcing should be limited to the current extent or decreased.

8. Currently, a system is in place where tax income from more wealthy municipalities is transferred to less wealthy municipalities. In practice this means that municipalities in the Helsinki region transfer money to the other parts of the country. What is your opinion of this system?
   0—The current system is good, or even more money should be transferred.
   1—The Helsinki region should be allowed to keep more of its tax income.

References

Corander J (2003) Labelled graphical models. Scand J Stat 30:493–508
Corander J, Gyllenberg M, Koski T (2006) Bayesian model learning based on a parallel MCMC strategy. Stat Comput 16:355–362
Corander J, Ekdahl M, Koski T (2008) Parallel interacting MCMC for learning of topologies of graphical models. Data Min Knowl Disc 17:431–456
Csiszár I (1975) I-divergence geometry of probability distributions and minimization problems. Ann Probab 3(1):146–158
Csiszár I, Matúš F (2003) Information projections revisited. IEEE Trans Inf Theory 49(6):1474–1490
Edwards D, Havránek T (1985) A fast procedure for model search in multidimensional contingency tables. Biometrika 72(2):339–351
Eriksen PS (1999) Context specific interaction models. Technical report, Department of Mathematical Sciences, Aalborg University, Aalborg
Friedman N, Goldszmidt M (1996) Learning Bayesian networks with local structure. In: Proceedings of the twelfth annual conference on uncertainty in artificial intelligence, pp 252–262
Golumbic MC (2004) Algorithmic graph theory and perfect graphs, 2nd edn. Elsevier, Amsterdam
Helsingin Sanomat (2011) HS:n vaalikone 2011. http://www2.hs.fi/extrat/hsnext/HS-vaalikone2011.xls. Visited 19 Aug 2014
Højsgaard S (2003) Split models for contingency tables. Comput Stat Data Anal 42:621–645
Højsgaard S (2004) Statistical inference in context specific interaction models for contingency tables. Scand J Stat 31:143–158
Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT Press, London
Lauritzen SL (1996) Graphical models. Oxford University Press, Oxford
Nyman H, Pensar J, Koski T, Corander J (2014) Stratified graphical models—context-specific independence in graphical models. Bayesian Anal 9(4):883–908
Nyman H, Xiong J, Pensar J, Corander J (2015) Marginal and simultaneous predictive classification using stratified graphical models. Adv Data Anal Classif. doi:10.1007/s11634-015-0199-5
Pensar J, Nyman H, Koski T, Corander J (2015) Labeled directed acyclic graphs: a generalization of context-specific independence in directed graphical models. Data Min Knowl Disc 29(2):503–533
Rudas T (1998) A new algorithm for the maximum likelihood estimation of graphical log-linear models. Comput Stat 13:529–537
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464
Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley, Chichester