Bayesian Analysis (2014) **9**, Number 4, pp. 883–908

Stratified Graphical Models - Context-Specific Independence in Graphical Models

Henrik Nyman * and Johan Pensar † and Timo Koski ‡ and Jukka Corander §

Abstract. Theory of graphical models has matured over more than three decades to provide the backbone for several classes of models that are used in a myriad of applications such as genetic mapping of diseases, credit risk evaluation, reliability and computer security. Despite their generic applicability and wide adoption, the constraints imposed by undirected graphical models and Bayesian networks have also been recognized to be unnecessarily stringent under certain circumstances. This observation has led to the proposal of several generalizations that aim at more relaxed constraints by which the models can impose local or context-specific dependence structures. Here we consider an additional class of such models, termed stratified graphical models. We develop a method for Bayesian learning of these models by deriving an analytical expression for the marginal likelihood of data under a specific subclass of decomposable stratified models. A non-reversible Markov chain Monte Carlo approach is further used to identify models that are highly supported by the posterior distribution over the model space. Our method is illustrated and compared with ordinary graphical models through application to several real and synthetic datasets.

Keywords: Graphical Model, Context-Specific Interaction Model, Markov Chain Monte Carlo, Bayesian Model Learning, Multivariate Discrete Distribution

1 Introduction

Along the path of development of the statistical theory of graphical models (GMs) largely set by the classic works of Darroch et al. (1980) and Lauritzen and Wermuth (1989), multifaceted generalizations of the original Markov dependence concepts have flourished as the field gained momentum. Despite the versatility of graphical models to encode the dependence structure over a set of discrete variables, there are several alternative model classes that are motivated by the failure of GMs to capture some forms of dependence or independence. For instance, hierarchical log-linear models that lack a direct graphical model representation were considered extensively already before the theory of graphical models took a concrete form, see for instance Haberman (1974), Bishop et al. (2007), and more recently Hara et al. (2012). A particular challenge related to such models is the burdensome interpretation, which is one of the core advantages of graphical models.

*Department of Mathematics, Åbo Akademi University, Turku, Finland, hennyman@abo.fi
†Department of Mathematics, Åbo Akademi University, Turku, Finland, jopensar@abo.fi
‡Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden, tjtkoski@kth.se
§Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland, jukka.corander@helsinki.fi

 DOI:10.1214/14-BA882

An observation independently made in several enhancements of the theory of graphical models for discrete multivariate distributions is that the use of the basic concept of conditional independence may casually hide the existence of multiple local or context-dependent independencies. Using the theory of log-linear models for contingency tables as their basis, Corander (2003b), Eriksen (1999, 2005), and Højsgaard (2003, 2004) introduced a variety of ways to generalize graphical models. The common notion in these models is that the conditional independence is replaced by an independence that holds only in a subspace of the outcome space of the variables included in a particular condition. Such restrictions may for instance be imposed in a recursive fashion as in Højsgaard (2004), in which case a variable that has been included in a context to split contingency tables into subsets where distinct dependence structures are imposed, can no longer itself be a subject to a local independence statement conditional on other variables. Completely independently of these developments found in the statistical literature, the machine learning community has witnessed the development of context-dependent Bayesian networks in Boutilier et al. (1996), Friedman and Goldszmidt (1996), and Koller and Friedman (2009). The recursive approach has been considered also in this setting, as Boutilier et al. (1996) introduced trees of conditional probability tables which form the backbone of Bayesian networks.

The above cited methods for obtaining a context-specific dependence structure for a set of variables impose rather extensive restrictions. In order to simplify statistical inference about the model parameters and learning of the model structure, we here aim at loosening the restrictions by introducing a larger and more general class of *stratified graphical models* (SGMs), expanding the results of Corander (2003b). The notion of stratification referred to here is distinct from that used in Geiger et al. (2001), who considered stratified exponential families for graphical models with hidden variables. In our framework stratification refers instead solely to observed variables. SGMs offer the advantage that context-specific independencies can be read directly off the graphs, promoting the comprehension of the dependence structure. We consider Bayesian inference for the class of SGMs and show that marginal likelihoods can be calculated analytically for a subclass of decomposable models. Learning of model structures associated with high posterior probabilities is performed using the non-reversible Markov chain Monte Carlo (MCMC) algorithm introduced in Corander et al. (2008).

This paper is organized as follows. SGMs are formally introduced in Section 2. An analytical expression for the marginal likelihood given a decomposable SGM is derived in Section 3. In Section 4 we present an MCMC-based search algorithm which is used to discover models associated with high posterior probabilities. Several synthetic and real datasets are used to illustrate the potential of SGMs in Section 5. The final section provides some concluding remarks along with some ideas for future research related to these models.

2 Stratified Graphical Models

To enable the presentation of SGMs, some of the central concepts from the theory of graphical models are first introduced. For a comprehensive account of the statistical and computational theory of probabilistic graphical models, see Whittaker (1990), Lauritzen (1996), and Koller and Friedman (2009). While the terms node and variable are closely related when considering graphical models, we will in this article strive to use the notation X_δ when referring to the variable associated to node δ. Let $G = G(\Delta, E)$, be an undirected graph, consisting of a set of nodes Δ, which represent a set of random variables, and of a set of undirected edges $E \subseteq \{\Delta \times \Delta\}$. It is assumed throughout this article that all considered variables are binary. However, the introduced theory can readily be extended to finite discrete variables.

For a subset of nodes $A \subseteq \Delta$, $G_A = G(A, E_A)$ is a subgraph of G, such that the nodes in G_A are equal to A and the edge set comprises those edges of the original graph for which both nodes are in A, i.e. $E_A = \{A \times A\} \cap E$. The outcome space for the variables X_A, where $A \subseteq \Delta$, is denoted by $\mathcal{X}_A$ and an element in this space by $x_A \in \mathcal{X}_A$. Given our restriction to binary variables, the cardinality $|\mathcal{X}_A|$ of $\mathcal{X}_A$ equals $2^{|A|}$. Two nodes γ and δ are *adjacent* in a graph if $\{\gamma, \delta\} \in E$, that is an edge exists between them. A *path* in a graph is a sequence of nodes such that for each successive pair within the sequence the nodes are adjacent. A *cycle* is a path that starts and ends with the same node. A *chord* in a cycle is an edge between two non-consecutive nodes in the cycle. Two sets of nodes A and B are said to be *separated* by a third set of nodes S if every path between nodes in A and nodes in B contains at least one node in S. A graph is defined as *complete* when all pairs of nodes in the graph are adjacent.

A graph is defined as *decomposable* if all cycles found in the graph containing four or more unique nodes contain at least one chord. A *clique* in a graph is a set of nodes A such that the subgraph G_A is complete. A *maximal clique* C is a clique for which there exists no set of nodes C^* such that $C \subset C^*$ and G_{C^*} is also complete. The set of maximal cliques in the graph G will be denoted by $\mathcal{C}(G)$. The set of *separators*, $\mathcal{S}(G)$, in the decomposable graph G can be obtained through intersections of the maximal cliques of G ordered in terms of a junction tree, see e.g. Golumbic (2004). A graphical model can be defined as the pair $G = G(\Delta, E)$ and the joint distribution P_Δ on the variables X_Δ, such that P_Δ factorizes according to G (see equation (1) for decomposable graphs). Given only the graph of a GM it is possible to ascertain if two sets of random variables X_A and X_B are conditionally independent given another set of variables X_S, due to the global Markov property

$$X_A \perp X_B \,|\, X_S, \text{ if } S \text{ separates } A \text{ from } B \text{ in } G.$$

A statement of conditional independence of two variables X_δ and X_γ given X_S imposes fairly strong restrictions to the joint distribution since the condition $P(X_\delta, X_\gamma \,|\, X_S) = P(X_\delta \,|\, X_S)P(X_\gamma \,|\, X_S)$ must hold for any joint outcome of the variables X_S. The idea common to context-specific independence models is to lift some of those restrictions to achieve more flexibility in terms of model structure. Exactly which restrictions are allowed to be simultaneously lifted varies considerably over the proposed model classes.

Consider a GM with the complete graph spanning three nodes $(1, 2, 3)$, which specifies 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 $P(X_1 = 1)P(X_2 = x_2 \,|\, X_1 = 1)P(X_3 = x_3 \,|\, X_1 = 1)$ 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 labeling the edge $\{2, 3\}$ with the *stratum* where the context-specific independence $X_2 \perp X_3 \,|\, X_1 = 1$ of the two variables holds, as illustrated in Figure 1a.

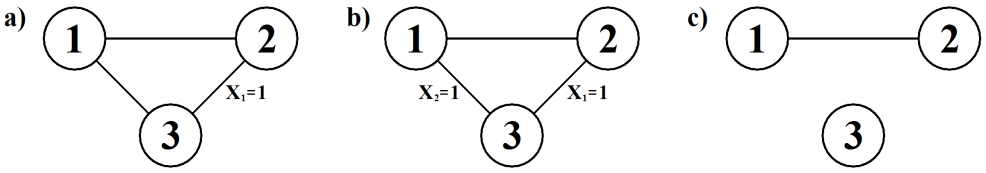

Figure 1: Graphical representation of the dependence structures of three variables. In (a) the stratum $X_1 = 1$ is shown as a label on the edge $\{2, 3\}$, in (b) strata $X_1 = 1$ and $X_2 = 1$ are shown as labels on the edges $\{2, 3\}$ and $\{1, 3\}$, respectively, in (c) an undirected graph with the maximal cliques $\{1, 2\}$ and $\{3\}$ is shown.

The following is a formal definition of what is intended by a stratum.

Definition 1. *Stratum. Let the pair (G, P_Δ) be a GM for Δ. For all $\{\delta, \gamma\} \in E$, let $L_{\{\delta,\gamma\}}$ denote the set of nodes adjacent to both δ and γ. For a non-empty $L_{\{\delta,\gamma\}}$, define the stratum of the edge $\{\delta, \gamma\}$ as the subset $\mathcal{L}_{\{\delta,\gamma\}}$ of outcomes $x_{L_{\{\delta,\gamma\}}} \in \mathcal{X}_{L_{\{\delta,\gamma\}}}$ for which X_δ and X_γ are independent given $X_{L_{\{\delta,\gamma\}}} = x_{L_{\{\delta,\gamma\}}}$, i.e. $\mathcal{L}_{\{\delta,\gamma\}} = \{x_{L_{\{\delta,\gamma\}}} \in \mathcal{X}_{L_{\{\delta,\gamma\}}} : X_\delta \perp X_\gamma \,|\, X_{L_{\{\delta,\gamma\}}} = x_{L_{\{\delta,\gamma\}}}\}$.*

A label on an edge in a graph is a graphical representation of a corresponding stratum. The idea of context-specific independence generalizes readily to a situation where multiple strata for distinct pairs of variables are considered. Figure 1b displays the complete graph for three nodes with the edges $\{2, 3\}$ and $\{1, 3\}$ labeled with the strata $X_1 = 1$ and $X_2 = 1$, respectively. In addition to the context-specific independence statement present in Figure 1a, here we have the simultaneous restriction that $X_1 \perp X_3 \,|\, X_2 = 1$, such that $P(X_1 = x_1, X_2 = 1, X_3 = x_3) = P(X_2 = 1)P(X_1 = x_1 \,|\, X_2 = 1)P(X_3 = x_3 \,|\, X_2 = 1)$ for all outcomes $x_1 \in \{0, 1\}, x_3 \in \{0, 1\}$. This pair of restrictions does not imply that $P(X_3 = x_3) = P(X_3 = x_3 \,|\, X_1 = 1, X_2 = 1)$ as would be the case given the graph in Figure 1c. It does, however, imply that the information contained about X_3 in the knowledge that $X_1 = 1$ and $X_2 = 1$ must be the same, i.e. $P(X_3 = x_3 \,|\, X_1 = 1) = P(X_3 = x_3 \,|\, X_2 = 1) = P(X_3 = x_3 \,|\, X_1 = 1, X_2 = 1)$.

The following definition is a slight modification from Corander (2003b, p. 496) and formalizes an extension to ordinary graphical models. The defined class of models allows for simultaneous context-specific independence to be represented using a set of strata, partitioning the outcome space of the variables X_Δ.

Definition 2. *Stratified graphical model (SGM). A stratified graphical model is defined by the triple* (G, L, P_Δ)*, where* G *is the underlying graph,* L *equals the joint collection of all strata* $\mathcal{L}_{\{\delta,\gamma\}}$ *for the edges of* G*, and* P_Δ *is a joint distribution on* Δ *which factorizes according to the restrictions imposed by* G *and* L*.*

The pair (G, L) consisting of the graph G with the labeled edges determined by L will be referred to as a stratified graph (SG), usually denoted by G_L. When the collection of strata L is empty, G_L equals G. Figure 1a illustrates an SG with G equal to the complete graph and L including the single stratum $\mathcal{L}_{\{2,3\}} = \{X_1 = 1\}$. Correspondingly, the SG shown in Figure 1b has the same underlying graph with the two strata $L = \{\mathcal{L}_{\{2,3\}} = \{X_1 = 1\}, \mathcal{L}_{\{1,3\}} = \{X_2 = 1\}\}$. A stratified graph induces a specific *dependence structure*.

Definition 3. *Dependence structure. The dependence structure induced by an SG is defined as the collection of all marginal, conditional and context-specific independencies that are imposed as restrictions on the probability distribution by the SG.*

Definition 4. *Faithful distribution. A distribution* P_Δ *is defined as faithful to an SG if it contains exactly the same marginal, conditional and context-specific independencies that are present in the dependence structure induced by the SG.*

Consider a distribution P_Δ that is faithful to the SG in Figure 1b. While P_Δ also factorizes according to the restrictions imposed by the SG in Figure 1a it is not faithful to it as it contains the context-specific independence $X_1 \perp X_3 \,|\, X_2 = 1$ which cannot be induced from the SG in Figure 1a.

The remainder of this section will be used to determine a framework that will allow for the derivation of an analytical expression of the marginal likelihood of a dataset given a stratified graph. Unfortunately, this will involve introducing a set of restrictions to the SG, resulting in a subclass of *decomposable SGMs*. The restrictions imposed here are, however, far less extensive then those imposed in Corander (2003b). In addition, Corander (2003b) did not consider structural learning of the context-specific graphs by using posterior probabilities, instead a simpler approach with penalized predictive entropies was adopted for such inference.

Consider a stratified graph with a decomposable underlying graph G having the maximal cliques $\mathcal{C}(G)$ and separators $\mathcal{S}(G)$. The SG is defined as decomposable if no labels are assigned to edges in any separator and in every maximal clique all labeled edges have at least one node in common.

Definition 5. *Decomposable SG. Let* (G, L) *constitute an SG with* G *being decomposable. Further, let* E_L *denote the set of all labeled edges,* E_C *the set of all edges in the maximal clique* C*, and* $E_\mathcal{S}$ *the set of all edges in the separators of* G*. The SG is defined as decomposable if*

$$E_L \cap E_\mathcal{S} = \varnothing,$$

and

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

An SGM where (G, L) constitutes a decomposable SG is termed a decomposable SGM. The graphs depicted in Figures 1a and 1b are examples of decomposable SGs. Figure 2 displays three SGs with identical underlying graphs, where it is assumed that the nodes are ordered topologically. This entails that it is not necessary to include the variables in the graphical representation of a stratum. Instead of writing a label as $(X_1 = 0, X_2 = 0)$, it is sufficient to write $(0, 0)$, as it is uniquely determined which nodes are adjacent to both nodes in the labeled edge. The SG in Figure 2a is decomposable, the SGs in Figures 2b and 2c are not. The graph in Figure 2b is not decomposable since the maximal clique $\{1, 2, 3, 4\}$ contains the labeled edges $\{1, 2\}$ and $\{3, 4\}$ which have no nodes in common. The graph in Figure 2c contains the labeled edge $\{1, 4\}$ which also constitutes the separator of maximal cliques $\{1, 2, 3, 4\}$ and $\{1, 4, 5\}$.

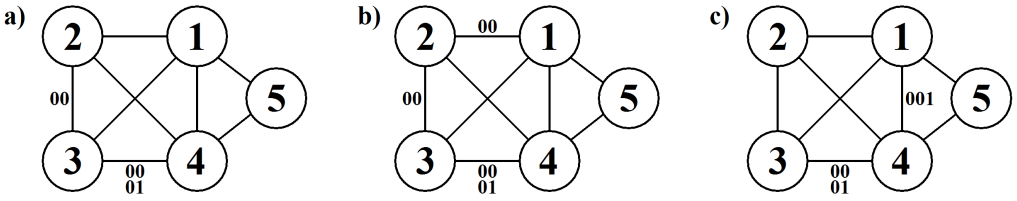

Figure 2: Three SGs with the same underlying graph. The graph in (a) is decomposable, the graphs in (b) and (c) are not.

As shown in the next section, for a decomposable stratified graph it is possible to calculate the marginal likelihood of a dataset analytically using a modification of the procedure introduced by Cooper and Herskovits (1992). This is due to the fact that a joint distribution faithful to a decomposable SG possesses a *minimal factorization.*

Definition 6. *Minimal factorization. Consider an outcome $x_\Delta \in \mathcal{X}_\Delta$ of the variables in X_Δ, and a subset $X_\Omega \subseteq X_\Delta$. Given an ordering of the variables in X_Ω, let $\prod_{\delta \in \Omega} P(X_\delta = x_\delta \,|\, X_{A(\delta)} = x_{A(\delta)})$ be a factorization of the probability $P(X_\Omega = x_\Omega)$. Each set $A(\delta)$ is a subset of B_δ, where B_δ denotes the complete set of nodes for which all variables in X_{B_δ} precede X_δ in the ordering. The set B_δ is defined as empty if X_δ is the first variable in the ordering. If a factor $P(X_\delta = x_\delta \,|\, X_{A(\delta)} = x_{A(\delta)})$ is such that there exists a non-empty subset $D \subseteq A(\delta)$ for which $X_\delta \perp X_D \,|\, X_{\Delta \setminus \{D \cup \delta\}} = x_{\Delta \setminus \{D \cup \delta\}}$, the factor contains a false dependency. A factorization that contains no false dependencies is defined as a minimal factorization. A distribution such that the probability of any outcome x_Δ can be factorized using minimal factorizations is said to possess a minimal factorization.*

If we for instance look at the graph in Figure 1c we can deduce that $P(X_1 = x_1, X_2 = x_2, X_3 = x_3)$ can be factorized both as

$$P(X_1 = x_1)P(X_2 = x_2 \,|\, X_1 = x_1)P(X_3 = x_3 \,|\, X_1 = x_1, X_2 = x_2),$$

and

$$P(X_1 = x_1)P(X_2 = x_2 \,|\, X_1 = x_1)P(X_3 = x_3).$$

However, the first factorization does not constitute a minimal factorization since $X_3 \perp (X_1, X_2)$.

Theorem 1. *A distribution that is faithful to a decomposable stratified graph possesses a minimal factorization.*

Proof. See Appendix.

There do exist distributions, faithful to non-decomposable SGs, that possess minimal factorizations. However, these distributions are induced by SGs where not only the edges on which labels may be placed, but also the labels themselves follow very strict restrictions. To avoid adding such constraints, while ensuring that the marginal likelihood of a dataset can be analytically calculated, the restrictions imposed by decomposable SGs are required.

3 Calculating the Marginal Likelihood for Decomposable Stratified Graphs

Let $\mathbf{X}$ denote a data matrix of n binary vectors, each containing $d = |\Delta|$ elements. We use X_A to denote the set of variables $\{X_\delta : \delta \in A\}$ and correspondingly $\mathbf{X}_A$ to denote the subset of $\mathbf{X}$ for the variables X_A. A probability distribution over the outcome space $\mathcal{X}_A$ is determined by a parameter vector $\theta \in \Theta$ where every element θ_i specifies the probability of a specific outcome $x_A^{(i)} \in \mathcal{X}_A$. The number of such outcomes will subsequently be denoted by k. Bayesian inference about undirected graphs and stratified graphs is derived using the posterior distribution over the model space. Given a prior distribution $P(G)$ (or $P(G_L)$), over a model space, the posterior equals $P(G \mid \mathbf{X}) = P(\mathbf{X} \mid G)P(G)/\sum_{G \in \mathcal{G}} P(\mathbf{X} \mid G)P(G)$, where $P(\mathbf{X} \mid G)$ is the marginal likelihood of the data given a graph, and $\mathcal{G}$ is the model space.

For an arbitrary decomposable graph G, the joint distribution $P_\Delta(X_\Delta)$ factorizes as

$$P_\Delta(X_\Delta) = \frac{\prod_{C \in \mathcal{C}(G)} P_C(X_C)}{\prod_{S \in \mathcal{S}(G)} P_S(X_S)}, \tag{1}$$

where $\mathcal{C}(G)$ and $\mathcal{S}(G)$ denote the maximal cliques and separators, respectively, of G. Using a prior distribution for the model parameters that also enjoys the Markov properties with respect to G, the marginal likelihood of the data $\mathbf{X}$ factorizes accordingly (Dawid and Lauritzen 1993)

$$P(\mathbf{X} \mid G) = \frac{\prod_{C \in \mathcal{C}(G)} P_C(\mathbf{X}_C)}{\prod_{S \in \mathcal{S}(G)} P_S(\mathbf{X}_S)}, \tag{2}$$

where for any subset $A \subseteq \Delta$ of nodes $P_A(\mathbf{X}_A)$ denotes the corresponding marginal likelihood of the subset $\mathbf{X}_A$ of data. By a suitable choice of prior distribution, these marginal likelihoods can be calculated analytically as follows. Let $n_A^{(i)}$ be the number

of occurrences of the outcome $x_A^{(i)}$ in the dataset $\mathbf{X}_A$ and let the probabilities determining the corresponding distribution have the Dirichlet$(\alpha_{A_1}, \ldots, \alpha_{A_k})$ distribution as the prior. Then, the marginal likelihood of $\mathbf{X}_A$ equals

$$P_A(\mathbf{X}_A) = \int_\Theta \prod_{i=1}^{k} (\theta_i)^{n_A^{(i)}} \cdot \pi_A(\theta) d\theta,$$

where $\pi_A(\theta)$ is the density function of the Dirichlet prior distribution. By the standard properties of the Dirichlet integral, the marginal likelihood can be further written as

$$P_A(\mathbf{X}_A) = \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{i=1}^{k} \frac{\Gamma(n_A^{(i)} + \alpha_{A_i})}{\Gamma(\alpha_{A_i})}, \tag{3}$$

where Γ denotes the gamma function and

$$\alpha = \sum_{i=1}^{k} \alpha_{A_i}.$$

The above result can be utilized as the basis when developing a corresponding expression for decomposable SGs. For these graphs each maximal clique and separator can be considered separately, and the factorization in (1) used. This is due to the fact that for a decomposable SG the nodes in a labeled edge $\{\delta, \gamma\}$ and the nodes in $L_{\{\delta,\gamma\}}$ all belong to the same maximal clique, as labels may not be placed on separators. Hence, a label on an edge in one maximal clique cannot imply changes to the dependence between variables corresponding to any other maximal clique. Given a maximal clique and its associated labels defined in G_L, it is necessary to define a factorization of the distribution $P_C(X_C)$ using a sequence of conditional distributions. To achieve this we introduce, in accordance with the proof of Theorem 1, a particular ordering of the variables in the maximal clique such that the variable corresponding to the node which all labeled edges have in common is last in the ordering. In the case where all labeled edges have two nodes in common, the last variable in the ordering can be chosen arbitrarily between them, and in the case with no labeled edges the ordering of the variables is arbitrary. When the factorization is based on such an ordering, it becomes clear which dependencies can be excluded from the last conditional distribution and it can be guaranteed that no false dependencies are employed.

An alternative way of formulating the factorization process for a maximal clique is to consider the variables that precede the variable X_δ in the ordering as parents of X_δ, denoted by X_{Π_δ}. Hence, except for the last variable in the ordering, all variables depend in their conditional distribution on each of the values of their parents. For the last variable some outcomes of its parents will have the same effect and these values can consequently be grouped together, as is done using default tables or conditional probability tables by Friedman and Goldszmidt (1996) and Boutilier et al. (1996). As an example, consider the SG in Figure 1a, where the parent outcomes $(X_1 = 1, X_2 = 0)$ and $(X_1 = 1, X_2 = 1)$ for X_3 are grouped together. Correspondingly, for the SG in Figure 1b, the parent outcomes $(X_1 = 1, X_2 = 0)$, $(X_1 = 1, X_2 = 1)$, and $(X_1 = 0, X_2 = 1)$

are grouped together. This means that there are effectively only two *distinguishable* parent combinations for X_3, comprising of $\{(X_1 = 1, X_2 = 0), (X_1 = 1, X_2 = 1), (X_1 = 0, X_2 = 1)\}$ and $\{(X_1 = 0, X_2 = 0)\}$.

Using the ordering of variables discussed above the marginal likelihood $P_C(\mathbf{X}_C)$ for a maximal clique of a decomposable SG can be calculated using a modified version of the formula introduced by Cooper and Herskovits (1992) for the marginal likelihood of a Bayesian network. The use of this formula is made possible by the fact that distributions faithful to decomposable SGs possess a minimal factorization, as stated in Theorem 1. Our modification is defined as

$$P_C(\mathbf{X}_C) = \prod_{j=1}^{d} \prod_{l=1}^{q_j} \frac{\Gamma(\sum_{i=1}^{k_j} \alpha_{jil})}{\Gamma(n(\pi_j^l) + \sum_{i=1}^{k_j} \alpha_{jil})} \prod_{i=1}^{k_j} \frac{\Gamma(n(x_j^i \mid \pi_j^l) + \alpha_{jil})}{\Gamma(\alpha_{jil})}, \tag{4}$$

where d equals the number of variables in the maximal clique C, q_j is the number of *distinguishable* parent combinations for variable X_j (i.e. there are q_j distinct conditional distributions for variable X_j), k_j is the number of possible outcomes for variable X_j, α_{jil} is the hyperparameter corresponding to the outcome i of variable X_j given that the parental combination of X_j equals l, $n(\pi_j^l)$ is the number of observations of the combination l for the parents of variable X_j, and finally, $n(x_j^i \mid \pi_j^l)$ is the number of observations where the outcome of variable X_j is i given that the observed outcome of the parents of X_j equals l. Note that in this context a parent configuration l is not necessarily comprised of a single outcome of the parents of variable X_j, but rather a *group* of outcomes with an equivalent effect on X_j.

The hyperparameters of the Dirichlet distribution are determined by imposing the following two requirements:

1. The resulting value of $P_C(\mathbf{X}_C)$ is independent of the initial ordering of the variables in the clique.
2. In the absence of strata the value of $P_C(\mathbf{X}_C)$ is identical to that given by (3) when the hyperparameters in the corresponding prior distribution are set equal to 1.

These two requirements can be satisfied by the following choice of hyperparameters,

$$\alpha_{jil} = \alpha_{jl} = \frac{k \cdot \lambda_{jl}}{\pi_j \cdot k_j},$$

where k again equals $|\mathcal{X}_C|$, π_j is the total number of possible outcomes for the parents of variable X_j and k_j is the number of possible outcomes for variable X_j. Further, λ_{jl} equals the number of outcomes for the parents of variable X_j in *group* l with an equivalent effect on X_j, if X_j is the last variable in the ordering. Otherwise, λ_{jl} equals one. Equation 4 can also be used to calculate $P_C(X_C)$ where C is a non-maximal clique containing no labeled edges. As a result, $P_S(X_S)$ can be calculated using (4), as any separator $S \in \mathcal{S}(G)$ comprises a clique. Using the hyperparameters defined above, the

values $P_S(X_S)$ and $P_C(X_C)$, where C is a maximal clique containing no labeled edges, can be calculated using either (3) or (4) as these will yield the same results.

An essential additional element of learning decomposable SGs given a dataset is to ensure model identifiability. There may exist several different decomposable SGs that all induce the same dependence structure. As an example we can consider the SG in Figure 2a. Here, adding the label $(0,1)$ to the edge $\{2,3\}$ would merge the parent outcomes $(X_1 = 0, X_2 = 0, X_4 = 1)$ and $(X_1 = 0, X_2 = 1, X_4 = 1)$ of X_3. However, these two outcomes have already indirectly been merged by the existing labels. The label $(0,0)$ on the edge $\{2,3\}$ merges parent outcomes $(X_1 = 0, X_2 = 0, X_4 = 0)$ and $(X_1 = 0, X_2 = 1, X_4 = 0)$. The label $(0,0)$ on the edge $\{3,4\}$ merges parent outcomes $(X_1 = 0, X_2 = 0, X_4 = 0)$ and $(X_1 = 0, X_2 = 0, X_4 = 1)$. And the label $(0,1)$ on the edge $\{3,4\}$ merges parent outcomes $(X_1 = 0, X_2 = 1, X_4 = 0)$ and $(X_1 = 0, X_2 = 1, X_4 = 1)$. Meaning that the outcomes $\{(X_1 = 0, X_2 = 0, X_4 = 0), (X_1 = 0, X_2 = 0, X_4 = 1), (X_1 = 0, X_2 = 1, X_4 = 0), (X_1 = 0, X_2 = 1, X_4 = 1)\}$ already form a group with all outcomes having identical effect on X_3, thus the inclusion of the label $(0,1)$ to the edge $\{2,3\}$ does not alter the dependence structure. To exclude this possibility, we introduce the concept of *maximal regular* SGs.

Definition 7. *Maximal regular SG. A decomposable SG is defined as maximal regular if no elements may be added to L without altering the dependence structure. Further, the stratum associated with each edge $\{\delta,\gamma\}$ must be a proper subset of $\mathcal{X}_{L_{\{\delta,\gamma\}}}$.*

An SGM where (G, L) constitutes a maximal regular SG is termed a maximal regular SGM. The regularity refers to the condition that an edge cannot be excluded completely from the graph as in an ordinary GM, by setting $\mathcal{L}_{\{\delta,\gamma\}} = \mathcal{X}_{L_{\{\delta,\gamma\}}}$. Maximal regular SGs constitute a subset of maximal SGs. In contrast to the class of all SGs, maximal regular SGs always induce different dependence structures.

Theorem 2. *The dependence structure induced by two maximal regular SGs, $G_L^1 = (G_1, L_1)$ and $G_L^2 = (G_2, L_2)$, are identical if and only if $G_1 = G_2$ and $L_1 = L_2$.*

Proof. Assume that there exists an edge $\{\delta,\gamma\}$ that is present in G_1 but absent in G_2, this implies that $X_\delta \perp X_\gamma \,|\, X_\Delta \backslash (X_\delta \cup X_\gamma)$. Considering that G_1 is decomposable, this means that $X_\delta \perp X_\gamma \,|\, X_{L_{\{\delta,\gamma\}}}$, where $L_{\{\delta,\gamma\}}$ is the set of nodes adjacent to both δ and γ in G_1, which is equivalent to $X_\delta \perp X_\gamma \,|\, X_{L_{\{\delta,\gamma\}}} \in \mathcal{X}_{L_{\{\delta,\gamma\}}}$. Hence $\mathcal{L}_{\{\delta,\gamma\}} = \mathcal{X}_{L_{\{\delta,\gamma\}}}$, contradicting the assumption that G_L^1 is maximal regular. Thus, it must hold that $G_1 = G_2$. Assume next that the stratum $\mathcal{L}^1_{\{\delta,\gamma\}} \in L_1$ contains an element $x_{L_{\{\delta,\gamma\}}}$ which is not present in $\mathcal{L}^2_{\{\delta,\gamma\}} \in L_2$. This implies that $X_\delta \perp X_\gamma \,|\, X_{L_{\{\delta,\gamma\}}} = x_{L_{\{\delta,\gamma\}}}$ and that the element $x_{L_{\{\delta,\gamma\}}}$ can be added to $\mathcal{L}^2_{\{\delta,\gamma\}}$ without altering the dependence structure induced by G_L^2, leading to a contradiction as G_L^2 cannot be a maximal regular SG. □

When learning SGs from data, we will restrict the attention to the class of maximal regular SGs. This means that we can avoid confusion over models having different appearances while leading to the same marginal likelihood due to identical dependence structures induced by them.

4 Algorithms for Bayesian Learning of SGs

Bayesian learning of graphical models has attained a considerable interest, both in the statistical and computer science literature, see, e.g. Madigan and Raftery (1994), Dellaportas and Forster (1999), Giudici and Green (1999), Corander (2003a), Giudici and Castelo (2003), Koivisto and Sood (2004), and Corander et al. (2008). Our learning algorithms described below belong to the class of non-reversible Metropolis-Hastings algorithms, introduced by Corander et al. (2006) and later further generalized and applied to learning of graphical models in Corander et al. (2008).

Let $\mathcal{M}$ denote the finite space of states over which the aim is to approximate the posterior distribution. In this paper we will run two separate types of searches. In one search the state space $\mathcal{M}$ will consist of all possible sets of labels, satisfying the restrictions of maximal regular SGs, for a given maximal clique. In the second search the state space will be the set of decomposable undirected graphs combined with the optimal set of labels for that graph. For $M \in \mathcal{M}$, let $Q(\cdot \,|\, M)$ denote the proposal function used to generate a new candidate state given the current state M. Under the generic conditions stated in Corander et al. (2008), the probability with which any particular candidate is picked by $Q(\cdot \,|\, M)$ need not be explicitly calculated or known, as long as it remains unchanged over all the iterations and the resulting chain satisfies the condition that all states can be reached from any other state in a finite number of steps. To initialize the algorithm, a starting state M_0 is determined. At iteration $t = 1, 2, \ldots$ of the non-reversible algorithm, $Q(\cdot \,|\, M_{t-1})$ is used to generate a candidate state M^*, which is accepted with the probability

$$\min\left(1, \frac{P(\mathbf{X} \,|\, M^*)P(M^*)}{P(\mathbf{X} \,|\, M_{t-1})P(M_{t-1})}\right), \tag{5}$$

where $P(M)$ is the prior probability assigned to M. The term $P(\mathbf{X} \,|\, M)$ denotes the marginal likelihood of the dataset $\mathbf{X}$ given M. If M^* is accepted, we set $M_t = M^*$, otherwise we set $M_t = M_{t-1}$.

In contrast to the standard reversible Metropolis-Hastings algorithm, for this non-reversible algorithm the posterior probability $P(M \,|\, \mathbf{X})$ does not equal the stationary distribution of the Markov chain. Instead, a consistent approximation of $P(M \,|\, \mathbf{X})$ is obtained by considering the space of distinct states $\mathcal{M}_t$ visited by time t such that

$$\hat{P}_t(M \,|\, \mathbf{X}) = \frac{P(\mathbf{X} \,|\, M)P(M)}{\sum_{M' \in \mathcal{M}_t} P(\mathbf{X} \,|\, M')P(M')}.$$

Corander et al. (2008) proved, under rather weak conditions, that this estimator is consistent, i.e.

$$\hat{P}_t(M \,|\, \mathbf{X}) \overset{a.s.}{\to} P(M \,|\, \mathbf{X}),$$

as $t \to \infty$. As our main interest will lie in finding the posterior optimal state, i.e.

$$\arg\max_{M \in \mathcal{M}} P(M \,|\, \mathbf{X})$$

it will suffice to identify

$$\arg\max_{M\in\mathcal{M}} P(\mathbf{X}\,|\,M)P(M).$$

The main goal of our search algorithm is to identify the stratified graph G_L^{opt} optimizing $P(\mathbf{X}\,|\,G_L)P(G_L)$. The search is broken down into two parts. Under the assumption that the optimal set of labels is known for each underlying graph a Markov chain traversing the set of possible underlying graphs will eventually identify G_L^{opt}. Another search may be used in order to identify the optimal set of labels given the underlying graph. It was earlier concluded that the marginal likelihood for a decomposable SG can be factorized according to (2). Due to this the search for the optimal set of labels can be conducted separately for each maximal clique.

Given a decomposable underlying graph G with the set of maximal cliques $\mathcal{C}(G)$, a search is conducted to find the optimal set of labels for each clique $C \in \mathcal{C}(G)$. The sets of labels are assigned uniform priors and cancel each other out in the acceptance probability (5). Using the proposal function defined in Algorithm 1, running a sufficient amount of iterations, we can be assured to find the optimal set of labels for each maximal clique. Combining the sets of labels for each maximal clique will result in an optimal labeling of the underlying graph.

Algorithm 1. *Proposal function used to find optimal set of labels for a maximal clique* $C \in \mathcal{C}(G)$.

The starting state is defined as the empty set containing no labels. Let L denote the current set of labels, and LP the set of labels that can be added to L without violating the restrictions of decomposable stratified graphs.

1. Set the candidate state $L^* = L$.
2. If LP is empty and L is non-empty, delete a randomly chosen label in L^*.
3. If L is empty and LP is non-empty, add a randomly chosen label from LP to L^*.
4. If both L and LP are non-empty, with probability 0.5 delete a randomly chosen label in L^*, otherwise add a randomly chosen label from LP to L^*.
5. If L^* satisfies the maximal regular restrictions set it as the candidate state, otherwise repeat steps 1-5.

Using this procedure we can assume that the optimal labeling can be found for any underlying graph and we can proceed to the search for the best underlying graph with optimal labeling. In this search, instead of using a uniform prior, we use a prior that penalizes dense graphs

$$P(G_L) \propto 2^{d-f},$$

where d is the number of nodes in the underlying graph G and f is the number of free parameters in a distribution P_Δ faithful to G. This choice of prior is motivated by the

fact that adding a label to a sparse graph often induces a context-specific independence in a larger stratum than adding a label to a dense graph. The value 2^{f-d} is a numerically convenient approximation of the number of unique dependence structures that can be derived by adding labels to an undirected graph. By looking at the conditional distributions for a variable X_δ with parents X_{Π_δ} in G, one can see that each parent outcome can be merged with a set of other outcomes by adding a label, removing a free parameter from P_Δ in the process. By adding different labels all but d of the original f free parameters in P_Δ can be removed, resulting in 2^{f-d} different dependence structures. This is, however, just an approximation as it is not possible to simultaneously remove any subset of the $f-d$ parameters by including labels. Using the proposal function in Algorithm 2 we conduct the search for the best underlying graph with optimal labeling.

Algorithm 2. *Proposal function used to find the best underlying graph with optimal set of labels.*

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 labels L.

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

The resulting candidate state G^* is used along with the corresponding optimal set of labels L^* to form the stratified graph $G^*_L = (G^*, L^*)$ which is used when calculating the acceptance probability according to (5).

In the next section we will use the search operator defined here on a set of synthetic datasets in order to illustrate its efficiency. We will also apply the search operator to a set of real datasets. Our results strongly support the use of models that enable the inclusion of context-specific independencies.

5 Illustration of SG Learning from Data

An SG including seven nodes is shown in Figure 3a. A probability distribution faithful to this SG is used to generate several sets of data of varying size, this distribution is available in the form of MATLAB code in the supplementary material.

When the size of the dataset exceeds 1,000 observations the graphs with the highest posterior probability found by the search method defined in the previous section usually coincide with the generating model. However, as the number of observations drops the optimal graphs start to deviate from the generating graph. The SG in Figure 3b is representative of the optimal graphs found for datasets including 500 observations. We

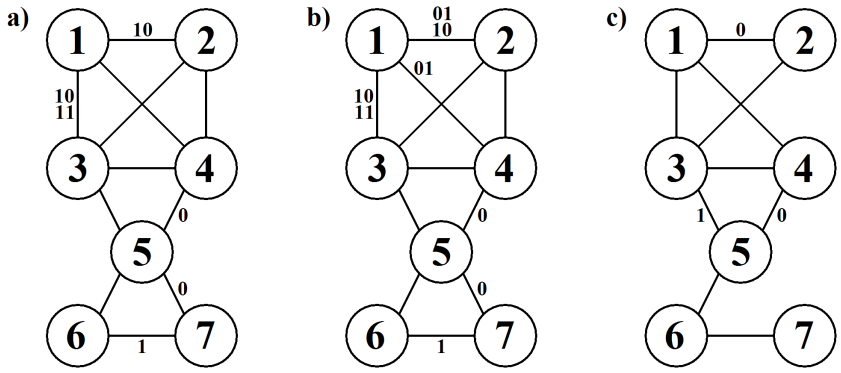

Figure 3: Dependence structures for synthetic dataset. a) Generating SG b) SG when data contains 500 observations c) SG when data contains 100 observations.

can see that the underlying graph is still the correct one, but a number of extra labels have been added to clique $\{1, 2, 3, 4\}$. The SG in Figure 3c is representative of the optimal graphs for datasets containing only 100 observations. Here we can see that not only do the labels differ strongly from those of the generating graph but also that the underlying graph is missing a couple of edges.

The experiments based on synthetic data confirm that the search algorithms are performing as expected when the data generating structure is known. However, for small datasets the observed posterior modes usually differ from the generating model. This gives rise to an important question, namely, when trying to learn the graph structure what is the required size of the dataset? For decomposable SGs, as we try to determine which labels to include, the size of the maximal cliques will be of relevance. The larger the maximal clique, the larger the set of possible labels, implying that more data is needed to have high probability of discovering the correct labels. In an effort to deduce the required number of observations in the dataset for maximal cliques with three, four, and five nodes we generate multiple datasets of varying size following the dependence structure induced by the stratified graphs in Figure 4.

Given a dataset the method defined in Section 4 is used to determine the optimal SG. The Bayes factor is then used to determine how strong the evidence is for the generating SG in comparison to the optimal SG found in the search. Presumably, as the number of observations in the data grows the optimal SG will more often coincide with the generating model, resulting in the Bayes factor equaling 1. In order to get reliable results the experiment is repeated several times and the mean of the Bayes factor is reported.

Figures 5a-c contain the values of the Bayes factor for the cases with three to five variables, respectively. Clearly, the Bayes factors increase as the number of observations grows and simulations show that they converge to 1 as the number of observations goes to infinity. However, the convergence seems to be quite slow, especially in the case with

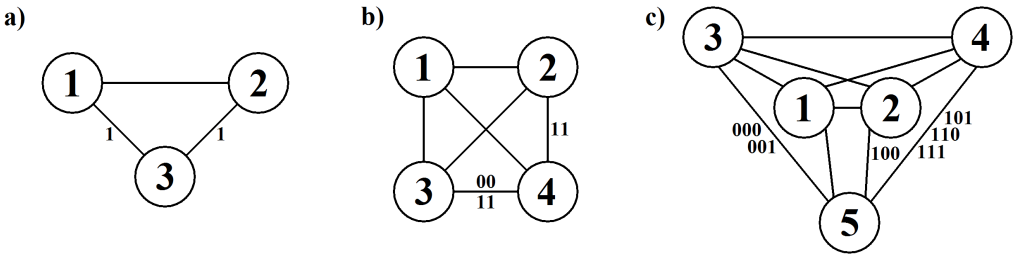

Figure 4: Dependence structures of distributions used to study the required size of datasets.

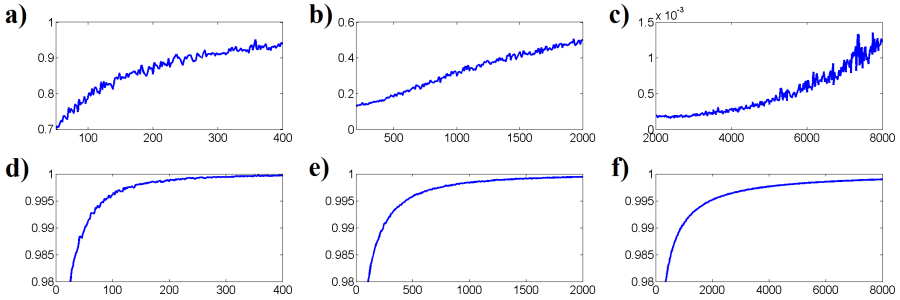

Figure 5: Plots a-c show the Bayes factor for the generating SG compared to the optimal SG for datasets generated using distributions faithful to the SGs in Figure 4. Plots d-f show the corresponding "normalized" Bayes factor.

five variables. This is a fairly obvious result as we are, in a sense, comparing the fit of the generating SG to the fit of all other possible SGs. In order to get more comparable values we calculate the n:th root of the Bayes factor, where n is the number of observations in the dataset. This "normalized" Bayes factor has the intuitive interpretation of being the amount of evidence produced by a single observation for the generating model in comparison to the optimal model. The resulting values for the three cases are shown in Figures 5d-f. We can see that all three curves tend to 1 as the number of observations grows. For instance, the number of observations needed for the three curves to reach 0.999 are 210, 1400, and 8000. This is a clear indicator of how rapidly the need for large datasets grows when the size of the maximal clique grows. For the clique with three variables the generating labels coincided with the optimal labels at a fraction of 0.9, using a dataset containing 400 observations. The corresponding fraction for four variables and data size 2000 is 0.28. In order to correctly identify the set of generating labels for five variables at a fraction of 0.05 a dataset consisting of roughly 50000 observations is needed, using 100000 observations the fraction rises to 0.5. Considering that the number of unique label sets, when considering a maximal clique with five variables, is in the vicinity of 10^{10} the need for large datasets is not surprising.

Interestingly, while the considered Bayes factor will exclusively converge to 1 for the generating SG the same does not hold for the "normalized" Bayes factor. To demonstrate this consider the case with four variables and the generating SG along with the graph where the labels have been removed, resulting in an ordinary complete graph with four nodes, as well as the graph where the labeled edges have been removed, resulting in an ordinary graph with maximal cliques $\{1, 2, 3\}$ and $\{1, 4\}$.

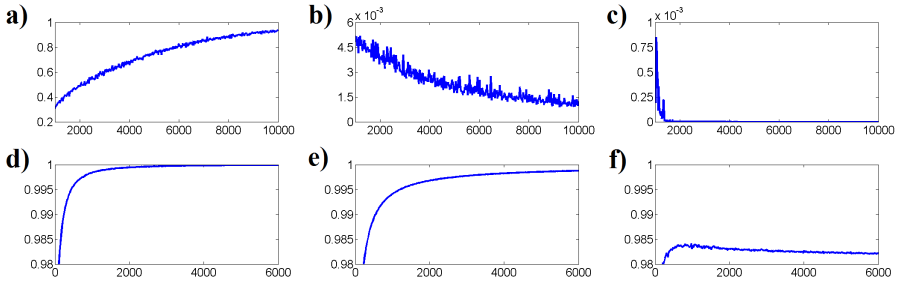

Figure 6: Bayes factor (a-c) and "normalized" Bayes factor (d-f) for SG in Figure 4 (a,d), ordinary complete graph (b,e), and ordinary graph with maximal cliques $\{1, 2, 3\}$ and $\{1, 4\}$ (c,f).

As we can see from Figure 6a the Bayes factor for the generating SG tends to 1, while the curves for the other two models steadily decrease, Figures 6b-c. Figures 6d-f show the "normalized" Bayes factor for the three models. While simulations show that for the ordinary complete graph the curve in Figure 6e converges to 1, clearly the same does not hold for the curve corresponding to the non-complete graph. While the generating SG can be considered a hybrid of these two ordinary graphs, this can be viewed as an indicator that the marginal likelihood, calculated using (2) and (4), penalizes under-parametrization harder than it does over-parametrization when the number of observations is large. However, other simulations show that for small datasets graphs that induce distributions with fewer parameters often result in higher marginal likelihoods.

While the results given here clearly show that large datasets are needed, especially for graphs containing larger maximal cliques, in order to perform stable inference of stratified graphs, any general recommendation regarding the number of observations needed is very difficult to state. Of course, the same holds for ordinary graphical models as much depends on the considered distribution, if a dependence is very strong it can be identified from a handful of observations while if the dependence is weak more observations are required.

Next we will conduct searches for context-specific independencies in real data. The first dataset includes prognostic factors for coronary heart disease and can be found in Edwards and Havránek (1985). The dataset contains 1841 observations on the six variables described in Table 1. Using the same setup as for the synthetic data, the two best decomposable SGs are depicted in Figure 7. They have the log-unnormalized

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

Table 1: Variables in coronary heart disease data.

posterior values of -6715.90 and -6716.66, respectively. The underlying graph in the optimal decomposable SG coincides with the optimal undirected graph as found by Corander et al. (2008) and is one of the two graphs suggested by Edwards and Havránek (1985). The discussion in Whittaker (1990) also suggests the possible inclusion of the edges $\{2, 5\}$ and $\{1, 2\}$. Compared to these sources our models are highly similar. However, in addition to the global independencies, our framework suggests for instance the context specific independencies $X_1 \perp X_4 \,|\, X_5 = 1$ and $X_4 \perp X_5 \,|\, X_1 = 0$.

The interpretation of the identified SGs is that the knowledge that a person smokes and has a ratio of beta and alpha lipoproteins less than or equal to 3 will affect the systolic blood pressure in one way and all the other variations for smoking and ratio of beta and alpha lipoproteins in another way. A simplified version would be to say that given that a person has a ratio of beta and alpha lipoproteins larger than 3, whether or not he smokes is unlikely to affect his systolic blood pressure. Interestingly the labeled edges are those that some sources suggest should be included in the model whilst other sources omit them from the model.

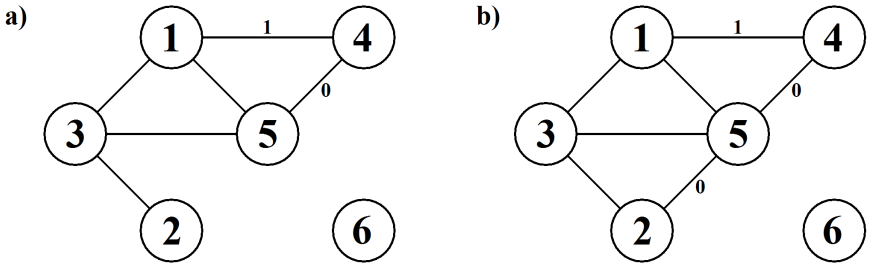

Figure 7: SGs with highest posterior probabilities for the heart disease data.

Next we consider a dataset involving 25 variables. This dataset 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 Sanomat 2011). The questionnaire contains a total of 30 questions, of these 25 are on a ordinal scale. The answers given to these 25 questions by the candidates are transformed to the binary variables listed in the supplementary material, the resulting dataset is also available as

supplementary material.

The SG with highest posterior probability is shown in Figure 8. The labels are not explicitly given in the graph, due to limited space, instead the labeled edges are colored red. This maximal regular SG contains 72 edges of which 36 are labeled. The graph contains a total of 87 labels and has a log-unnormalized posterior value of -21949.13. Conducting a search for the best ordinary graphical model, keeping in mind that undirected graphs are special cases of stratified graphs and therefore the same equations can be used to attain the log-unnormalized posterior, results in a graph with 70 edges and a log-unnormalized posterior value of -22043.15. These two graphs share 62 edges, implying that the induced dependence structures resemble each other to a considerable degree.

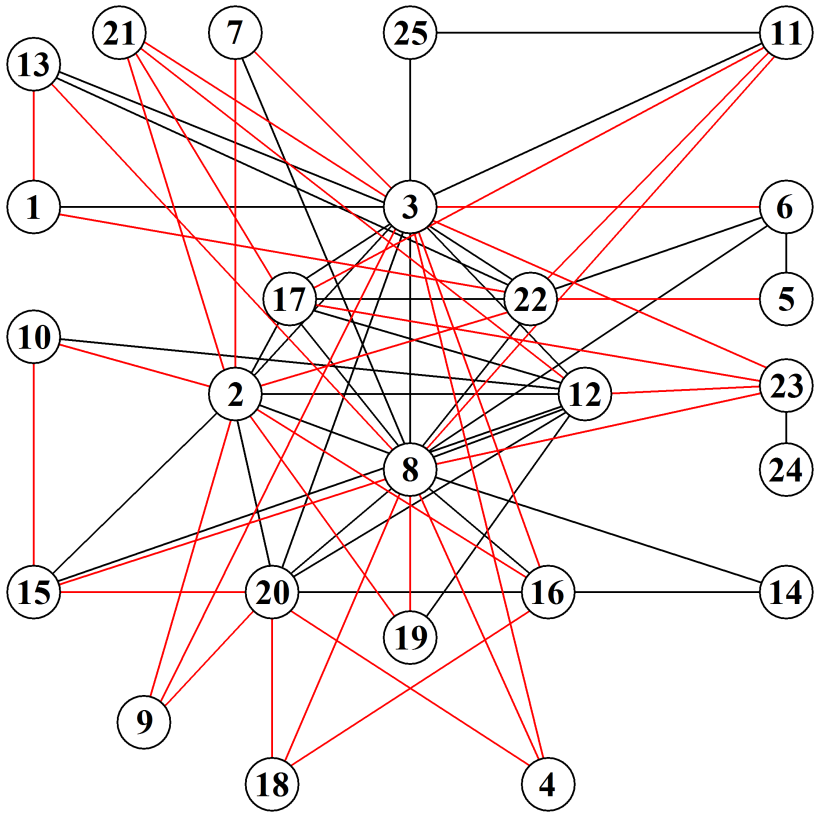

Figure 8: Optimal SG for the parliament election data, labeled edges are colored red.

Figure 9 displays two maximal cliques, found in both the optimal SG and optimal ordinary graph, with the labels associated with the SG.

The SG in Figure 9a induces a fairly straightforward context-specific dependence

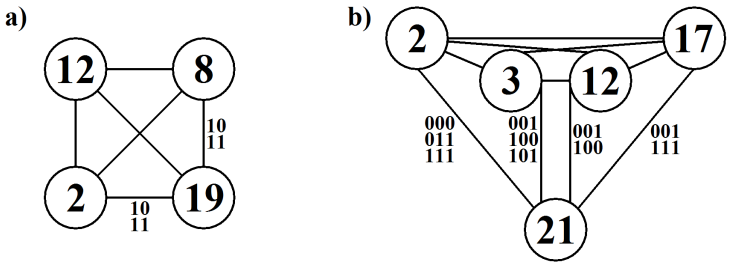

Figure 9: Two maximal cliques found both in the optimal SG and ordinary graph.

structure. Given that we know a candidate's opinion on mandatory military service (variable 12), knowing that the candidate is against equal rights for homosexuals to adopt children (variable 2) is likely to have the same effect on a candidate's view on singing Christian hymns in school (variable 19) as knowing that the candidate is in favor of economic help packages for struggling Euro countries (variable 8). The context-specific dependence structure induced by the SG in Figure 9b is much more intricate. However, a simple fact is that a probability distribution faithful to this component of the SG includes 21 free parameters, whereas the corresponding ordinary maximal clique would include 31 free parameters. In total, the optimal SG in Figure 8 induces a distribution with 324 free parameters while the underlying graph and optimal ordinary graph induce distributions with 407 and 368 free parameters, respectively. This means that the SG induces a more elaborate dependence structure using a substantially smaller number of parameters.

Most maximal cliques found in the parliament elections data contain four or five nodes, with the largest maximal cliques containing five nodes. As the dataset only contains 1806 observations the discussion above regarding the required sample size becomes relevant. To begin with it is important to notice that the restrictions imposed by decomposable SGs means that not all edges can be labeled. In the maximal clique $\{2, 8, 12, 19\}$ only the edges leading to node 19 may be labeled, similarly for the maximal clique $\{2, 3, 12, 17, 21\}$ only the edges leading to node 21 may be labeled. This heavily reduces the set of possible label combinations. In order to ascertain the robustness of the inferred labels we randomly remove observations from the dataset and see if the optimal labels change.

Figure 10 shows the average fraction with which the optimal label set given the entire data coincides with the optimal label set when a varying amount of observations has been removed from the data. Clearly, the labels for the maximal clique containing four nodes are much more robust than those for the clique containing five nodes. However, taking into account the total number of possible label sets the inferred labels seem fairly robust.

The final dataset considered lists the occurrences of different plants in different states and territories in North America (Bache and Lichman 2013). The data contains the

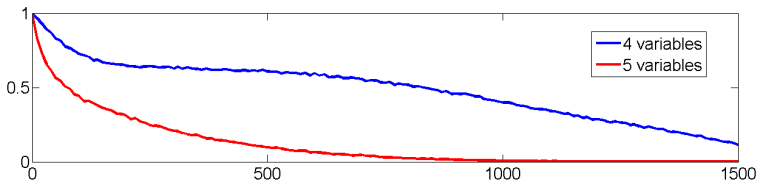

Figure 10: Fraction of the optimal label set estimated from the entire data that coincides with the optimal label set when an indicated amount of observations have been removed from the parliament election data (x-axis).

occurrences of 34781 different plants in the 69 states/territories listed in the supplementary material. The states/territories are ordered such that those that are geographically close to each other appear close to each other in the ordering. An adjacency matrix is used to display the resulting SG. Intuitively, the elements close to the diagonal in the adjacency matrix should more often indicate the presence of an edge than elements farther from the diagonal, since plants are more likely to occur simultaneously in states in close proximity. Figure 11 contains a plot of the 69-by-69 adjacency matrix representing the optimal SG found for this data.

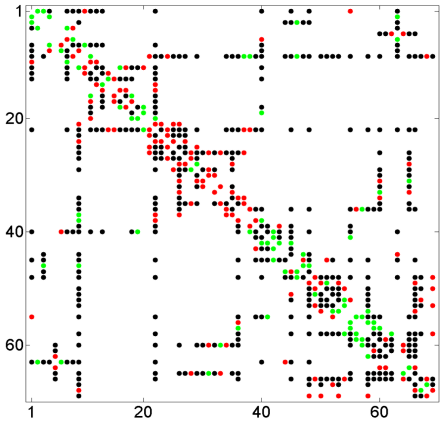

Figure 11: Adjacency matrix representing the optimal SG for the plants dataset.

For instance, the black dot on row 1, column 4 indicates that nodes 1 and 4 are connected by an unlabeled edge. A red dot corresponds to a labeled edge while a green dot corresponds to an unlabeled edge not in a separator, i.e. an unlabeled edge that could be labeled without breaking the restriction of decomposable SGs that edges in separators may not be labeled.

The resulting SG is fairly dense, containing 336 edges and maximal clique sizes ranging from 4 to 8. The underlying graph also contains some nodes that are adjacent

to a large portion of the other nodes, seen as vertical and horizontal "lines" in the plot of the adjacency matrix. While the underlying graph induces a distribution with 4774 free parameters the distribution induced by the SG only contains 4402 free parameters. An interesting observation that can be made from the adjacency matrix is that node 5 (Alberta) is adjacent to nodes 60, 61, 62, 64, 65, and 66, while node 63 (Alabama) is adjacent to nodes 1, 2, 3, 4, 6, 7, 8, and 9. This lead to the suspicion that an error had been made by the creators of the dataset, Alabama and Alberta having swapped places, further investigations indeed showed this to be the case.

One advantage with having a large number of observations is that the underlying graph of the optimal SG tends to be very similar to the optimal ordinary graph. For this dataset these graphs actually coincide. The score for the optimal ordinary graph and optimal SG are -368099.84 and -367756.68, respectively. Generally, using the optimal ordinary graph as the starting point for Algorithm 2 will make the search process much more effective. For even larger systems, spanning more than 100 variables, some restrictions may need to be made to the search process. One possible method is to first identify the optimal ordinary graph and then apply Algorithm 1 exclusively to that graph, this would also help to guard against over-parametrization. While the resulting SG may not be the global optimum, this method may still be useful in different applications, such as predictive classification (Nyman et al. 2014).

As previously shown, the complexity involved with inferring the labels grows rapidly with the increase in the size of the maximal cliques. For this data, the largest maximal clique includes 8 nodes. To assess the robustness of the inferred labels the same method as for the parliament election data is applied. For cliques comprising 4-8 nodes the optimal label sets for the entire data is compared to the optimal label sets when a varying amount of observations is removed from the data. Due to the underlying graph being quite dense it is only possible to place labels on a limited set of edges, as labels may not be placed on edges in separators. For the considered cliques it is only possible to place labels on two different edges.

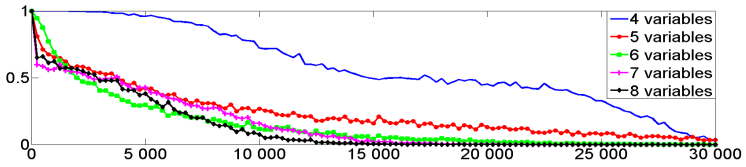

Figure 12: Fraction of the optimal label set estimated from the entire data that coincides with the optimal label set when an indicated amount of observations have been removed from the plants data (x-axis).

Figure 12 shows the average fraction with which the two considered label sets coincide. Clearly, the fraction tends to decline faster for large cliques, as the set of different label configurations are much larger for larger cliques, the clique containing six nodes deviating somewhat from this rule indicating that the inferred labels for that clique are a bit less robust in comparison. Generally, the results show that the inferred labels are

fairly robust when taking into account the huge set of applicable label sets.

While it is always possible to find the optimal set of labels given a dataset the difference between different sets of labels might in some cases be quite small and the robustness of the inferred labels questionable. One solution to this problem is to consider not only the optimal label set but a range of different label sets. Another solution is to add restrictions to the set of labels that are allowed. For instance, the number of variables used to determine if a label is fulfilled can be restricted. Considering a maximal clique $\{1, 2, 3, 4, 5, 6\}$ and the edge $\{5, 6\}$, instead of having a label $(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$ we could only allow sets of labels where X_3 and X_4 are unrestricted, effectively reducing the number of possible labels from 16 to 4.

6 Conclusions

The versatility of probabilistic graphical models has become clear through their popularity over a wide variety of application areas. On the other hand, their fairly restrictive global form of dependence structures has inspired the development of many generalizations of graphical models where independence can be a function of a context in the outcome space. In fact, similar developments have taken place for the class of Markov chain models, where the variable-order and variable-length Markov chains aim at a generalization of ordinary higher-order Markov chains where the dependence on the history of the process is context-specific (Rissanen 1983; Weinberger et al. 1995; Bühlmann and Wyner 1999; Bacallado 2011). Our formulation of the simultaneous context-specific independence restrictions allows for the derivation of an analytical Bayesian scoring function for decomposable SGs, which is particularly useful for fast learning purposes and leads to a more expressive model class than those considered by Boutilier et al. (1996), Friedman and Goldszmidt (1996), Corander (2003b), and Koller and Friedman (2009). In the future it would be fruitful to develop inference methods also for non-decomposable SGs, which do not enjoy an analytically tractable expression for the marginal likelihood. This would further reduce the constraints imposed on the dependence structure and allow for an even more expressive range of context-specific independencies to be explored.

Acknowledgments

The authors would like to thank the editors and the two 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.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.

References

Bacallado, S. (2011). "Bayesian analysis of variable-order, reversible Markov chains." *The Annals of Statistics*, 39(2): 838–864. 904

Bache, K. and Lichman, M. (2013). "UCI Machine Learning Repository." URL http://archive.ics.uci.edu/ml 901

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (2007). *Discrete multivariate analysis: theory and practice*. Springer. 883

Boutilier, C., Friedman, N., Goldszmidt, M., and Koller, D. (1996). "Context-Specific Independence in Bayesian Networks." In *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence*, 115–123. 884, 890, 904

Bühlmann, P. and Wyner, A. J. (1999). "Variable length Markov chains." *The Annals of Statistics*, 27(2): 480–513. 904

Cooper, G. F. and Herskovits, E. (1992). "A Bayesian method for the induction of probabilistic networks from data." *Machine learning*, 9(4): 309–347. 888, 891

Corander, J. (2003a). "Bayesian graphical model determination using decision theory." *Journal of multivariate analysis*, 85(2): 253–266. 893

— (2003b). "Labelled Graphical Models." *Scandinavian Journal of Statistics*, 30: 493–508. 884, 886, 887, 904

Corander, J., Ekdahl, M., and Koski, T. (2008). "Parallel interacting MCMC for learning of topologies of graphical models." *Data Mining and Knowledge Discovery*, 17: 431–456. 884, 893, 899

Corander, J., Gyllenberg, M., and Koski, T. (2006). "Bayesian model learning based on a parallel MCMC strategy." *Statistics and Computing*, 16: 355–362. 893

Darroch, J. N., Lauritzen, S. L., and Speed, T. (1980). "Markov fields and log-linear interaction models for contingency tables." *The Annals of Statistics*, 8: 522–539. 883

Dawid, A. and Lauritzen, S. (1993). "Hyper-Markov laws in the statistical analysis of decomposable graphical models." *The Annals of Statistics*, 21: 1272–1317. 889

Dellaportas, P. and Forster, J. J. (1999). "Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models." *Biometrika*, 86(3): 615–633. 893

Edwards, D. and Havránek, T. (1985). "A fast procedure for model search in multidimensional contingency tables." *Biometrika*, 72(2): 339–351. 898, 899

Eriksen, P. S. (1999). "Context specific interaction models." Technical report, Department of Mathematical Sciences, Aalborg University, Aalborg. 884

— (2005). "Decomposable Log-Linear Models." Technical report, Department of Mathematical Sciences, Aalborg University, Aalborg. 884

Friedman, N. and Goldszmidt, M. (1996). "Learning Bayesian Networks with Local Structure." In *Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence*, 252–262. 884, 890, 904

Geiger, D., Heckerman, D., King, H., and Meek, C. (2001). "Stratified exponential families: graphical models and model selection." *The Annals of Statistics*, 29(2): 505–529. 884

Giudici, P. and Castelo, R. (2003). "Improving Markov chain Monte Carlo model search for data mining." *Machine learning*, 50(1-2): 127–158. 893

Giudici, P. and Green, P. (1999). "Decomposable graphical Gaussian model determination." *Biometrika*, 86: 785–801. 893

Golumbic, M. C. (2004). *Algorithmic graph theory and perfect graphs (2nd ed.)*. Amsterdam: Elsevier. 885

Haberman, S. J. (1974). *The analysis of frequency data*, volume 194. Chicago, London: University of Chicago Press. 883

Hara, H., Sei, T., and Takemura, A. (2012). "Hierarchical subspace models for contingency tables." *Journal of Multivariate Analysis*, 103(1): 19–34. 883

Helsingin Sanomat (2011). "HS:n Vaalikone 2011." Visited 2013-10-15. URL http://www2.hs.fi/extrat/hsnext/HS-vaalikone2011.xls 899

Højsgaard, S. (2003). "Split Models for Contingency Tables." *Computational Statistics & Data Analysis*, 42: 621–645. 884

— (2004). "Statistical Inference in Context Specific Interaction Models for Contingency Tables." *Scandinavian Journal of Statistics*, 31: 143–158. 884

Koivisto, M. and Sood, K. (2004). "Exact Bayesian structure discovery in Bayesian networks." *The Journal of Machine Learning Research*, 5: 549–573. 893

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. London: The MIT Press. 884, 885, 904

Lauritzen, S. L. (1996). *Graphical models*. Oxford: Oxford University Press. 885, 908

Lauritzen, S. L. and Wermuth, N. (1989). "Graphical models for associations between variables, some of which are qualitative and some quantitative." *The Annals of Statistics*, 17(1): 31–57. 883

Madigan, D. and Raftery, A. E. (1994). "Model selection and accounting for model uncertainty in graphical models using Occam's window." *Journal of the American Statistical Association*, 89(428): 1535–1546. 893

Nyman, H., Xiong, J., Pensar, J., and Corander, J. (2014). "Marginal and simultaneous predictive classification using stratified graphical models." *arXiv:1401.8078 [stat.ML]*. 903

Rissanen, J. (1983). "A universal data compression system." *Information Theory, IEEE Transactions on*, 29(5): 656–664. 904

Weinberger, M. J., Rissanen, J. J., and Feder, M. (1995). "A universal finite memory source." *Information Theory, IEEE Transactions on*, 41(3): 643–652. 904

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Chichester: Wiley. 885, 899

Supplementary Material

The supplementary material includes the MATLAB code used to generate datasets with a dependence structure following the stratified graph in Figure 3a, a list of the questions presented to the candidates of the Finnish parliament elections of 2011 by Helsingin Sanomat along with the resulting dataset, and a list of the states/territories included in the plants dataset.

Appendix

Proof of Theorem 1.

Consider a joint distribution, over the variables $X_\Delta = (X_1, \ldots, X_d)$, faithful to a decomposable SG, $G_L = (G, L)$, and an arbitrary outcome $x_\Delta \in \mathcal{X}_\Delta$. Since G is a decomposable undirected graph, $P(X_\Delta = x_\Delta)$ can be factorized (Lauritzen 1996) as

$$P(X_\Delta = x_\Delta) = \frac{\prod_{C \in \mathcal{C}(G)} P(X_C = x_C)}{\prod_{S \in \mathcal{S}(G)} P(X_S = x_S)}.$$

For any separator $S \in \mathcal{S}(G)$ it holds that any variable $X_\delta \in X_S$ is dependent on each of the other variables in the set $X_S \backslash X_\delta$, regardless of any context-specific independencies, since decomposable SGs cannot have any labeled edges within separators. This implies that any factorization of $P(X_S = x_S)$ using a sequence of conditional probabilities $P(X_1 = x_1)P(X_2 = x_2 \,|\, X_1 = x_1) \ldots P(X_{d_S} = x_{d_S} \,|\, X_1 = x_1, \ldots, X_{d_S-1} = x_{d_S-1})$, where $d_S = |S|$, will be void of false dependencies. The probability $P(X_C = x_C)$, with $C \in \mathcal{C}(G)$ corresponding to the variables $(X_1, \ldots, X_{d_C})$ with $d_C = |C|$, can also be factorized using a minimal factorization. Assume that, in accordance with the definition of decomposable SGs, all the labeled edges have at least one node in common. The corresponding variable is chosen to be the last variable in the ordering, i.e. the variable X_{d_C}. In the case where the maximal clique only contains one labeled edge $\{\delta, \gamma\}$ the choice of X_{d_C} is ambiguous, either we choose $X_{d_C} = X_\delta$ or $X_{d_C} = X_\gamma$. It now follows that the factorization $P(X_1 = x_1, \ldots, X_{d_C-1} = x_{d_C-1}) = P(X_1 = x_1)P(X_2 = x_2 \,|\, X_1 = x_1) \ldots P(X_{d_C-1} = x_{d_C-1} \,|\, X_1 = x_1, \ldots, X_{d_C-2} = x_{d_C-2})$ contains no false dependencies. This can be seen from the stratified graph as all pairs of nodes corresponding to the variables in the set $(X_1, \ldots, X_{d_C-1})$ will be connected by an unlabeled edge. The final situation to investigate is the conditional probability $P(X_{d_C} = x_{d_C} \,|\, X_1 = x_1, \ldots, X_{d_C-1} = x_{d_C-1})$. This factor could potentially contain false dependencies as the edges leading to the node corresponding to X_{d_C} are allowed to be labeled. However, given the information that $X_1 = x_1, \ldots, X_{d_C-1} = x_{d_C-1}$, it is known which dependencies can be excluded in $P(X_{d_C} = x_{d_C} \,|\, X_1 = x_1, \ldots, X_{d_C-1} = x_{d_C-1})$, as the variables $X_1, \ldots, X_{d_C-1}$ are the variables that determine whether or not a context established by the labels in question is satisfied. Hence, it is always possible to avoid introducing false dependencies for such a clique. This proves that a distribution faithful to a decomposable SG always possesses a minimal factorization. □