Stratified Gaussian Graphical Models

Henrik Nyman[1,*], Johan Pensar[1], Jukka Corander[2]

[1]Department of Mathematics and Statistics, Åbo Akademi University, Finland
[2]Department of Mathematics and Statistics, University of Helsinki, Finland
[*]Corresponding author, Email: hennyman@abo.fi

Abstract

Gaussian graphical models represent the backbone of the statistical toolbox for analyzing continuous multivariate systems. However, due to the intrinsic properties of the multivariate normal distribution, use of this model family may hide certain forms of context-specific independence that are natural to consider from an applied perspective. Such independencies have been earlier introduced to generalize discrete graphical models and Bayesian networks into more flexible model families. Here we adapt the idea of context-specific independence to Gaussian graphical models by introducing a stratification of the Euclidean space such that a conditional independence may hold in certain segments but be absent elsewhere. It is shown that the stratified models define a curved exponential family, which retains considerable tractability for parameter estimation and model selection.

1 Introduction

Since their original introduction in the 1970's, Gaussian graphical models (GGMs) are by now ubiquitous in statistical analysis of multivariate systems, given their beneficial characteristics regarding modularity and tractability of statistical inference, see Dempster (1972), Whittaker (1990), Lauritzen (1996), Giudici and Green (1999), Wong et al. (2003), Atay-Kayis and Massam (2005), Jones and West (2005), Li and Gui (2006), Yuan and Lin (2007), Carvalho and Scott (2009), Sun and Li (2012). However, unlike their discrete counterparts, log-linear graphical models, GGMs do not allow for very flexible representation of marginal and conditional dependence between variables, since their characteristics are determined by the properties of the multivariate normal distribution. A particularly attractive generalization of log-linear graphical models and Bayesian networks is to allow the dependence structure to be context-specific, such that an independence between a pair of variables may hold only when their neighbours attain certain values, i.e. a full conditional independence not being present. Such models have been considered both for directed graphs in Boutilier et al. (1996), Geiger and Heckerman (1996), Chickering et al. (1997), and for undirected graphs in Corander (2003), Højsgaard (2003, 2004).

Here we adapt the concept of discrete stratified graphical models (Nyman et al., 2014) to the multivariate Gaussian family by introducing a stratification of the Euclidean space that specifies where context-specific independencies between variables are present and where absent. It is shown that this definition leads to a plausible characterization of the local influence of the neighbours without leading to complex mixture-type models which would be seriously challenging

from the inference perspective. As demonstrated earlier by the discrete model families, context-specific independencies lend themselves easily to applications since it is fundamentally natural to consider the dependence between variables to be absent in a given context while being present elsewhere. We establish formally that the stratification leads to a curved exponential family, which retains considerable tractability in terms of parameter estimation and model selection. The remainder of the article is structured as follows. In Section 2 we introduce stratified Gaussian graphical models (SGGMs) and examine their statistical properties in detail. In Section 3 we consider inference for SGGMs and the last two sections provide illustrations and some concluding remarks, respectively.

2 Stratified Gaussian graphical models

2.1 Motivating the introduction of stratified Gaussian models

To provide an informal and intuitive introduction to the core ideas behind SGGMs, we start by considering the classic dataset concerning mathematics marks introduced by Mardia et al. (1979), see also Whittaker (1990) and Edwards (2000). For the five variables, listed in Table 1 representing marks of students in different areas of mathematics, Whittaker (1990) presented the GGM with the dependence structure defined by the graph in Figure 1. Whittaker (1990) noted the central role of algebra in this correlation structure, it being connected to all the remaining variables. However, in analogy to discrete models, what if algebra were independent of mechanics provided that the mark for vectors is above a certain threshold, i.e. $X_1 \perp X_3 \mid X_2 > a$? Alternatively, these two variables might also be independent if the third one belongs to a certain interval: $X_1 \perp X_3 \mid a < X_2 < b$. Such context-specific independencies would appear reasonable in many applications, for instance, a certain signal may activate the dependence of other variables only once it reaches high enough a value. In Section 4 we show that the likelihood for the mathematics marks data in fact supports a local dependence structure of the kind hypothesized above.

Variable	Label
Mechanics	1
Vectors	2
Algebra	3
Analysis	4
Statistics	5

Table 1: Variables in mathematics mark dataset.

The standard restrictions on the inverse covariance structure defined by GGMs lead to an exponential family of distributions. Despite of its versatility, such families lack the flexibility to encode context-specific independencies of the type considered above. In contrast, when distinct inverse covariance structures are specified over particular segments of the Euclidean space, one arrives at a more flexible family of distributions which still remains tractable in terms of inference, as shown later in more detail.
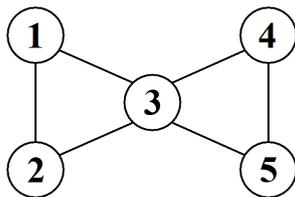
2

Figure 1: Graph representing the dependence structure of the five variables in mathematics mark data.

2.2 Notation and preliminaries

In order to define SGGMs we need the following basic concepts pertaining to Gaussian graphical models. For a more comprehensive treatment, see, for instance, Lauritzen (1996) or Whittaker (1990). A d-dimensional random vector $X_\Delta = (X_1, \ldots, X_d)$ follows a multivariate Gaussian distribution $N(\mu, \Sigma)$ if there exists a mean vector $\mu \in \mathbb{R}^d$ and a positive definite covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ such that X_Δ has the probability density function

$$f_{\mu,\Sigma}(x) = (2\pi)^{-d/2} |K|^{1/2} e^{-1/2(x-\mu)^T K(x-\mu)},$$

where $K = \Sigma^{-1}$ is the precision matrix of the distribution. This will be denoted $X_\Delta \sim N(\mu, \Sigma)$. Independence between two variables or alternatively two sets of variables X_A and X_B can be readily deduced from Σ, since $\sigma_{\delta,\gamma} = 0$ for every $X_\delta \in X_A$ and $X_\gamma \in X_B$ if and only if the two sets of variables X_A and X_B are marginally independent. Similarly, conditional independence between two variables X_δ and X_γ can be deduced from the precision matrix, since

$$X_\delta \perp X_\gamma \mid X_\Delta \backslash \{X_\delta, X_\gamma\} \Leftrightarrow k_{\delta,\gamma} = 0. \tag{1}$$

Next we define the necessary concepts from graph theory relevant for SGGMs. An undirected graph $G(\Delta, E)$ consists of a set of nodes Δ and of a set of undirected edges $E \subseteq \{\Delta \times \Delta\}$. 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 equal A and the edge set comprises those edges of the original graph for which both nodes are in A, i.e. $E_A = \{A \times A\} \cap E$. 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 two consecutive 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. A graph is defined as decomposable if all cycles found in the graph containing four or more unique nodes contain at least one chord. 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 clique in a graph is a set of nodes C such that the subgraph G_C is complete and there exists no other set C^*, such that $C \subset C^*$ and G_{C^*} is also complete. The set of cliques in a graph are denoted by $\mathcal{C}(G)$. For a decomposable graph the set of separators $\mathcal{S}(G)$ can be obtained through intersections of the cliques of G ordered in terms of a junction tree, see e.g. Golumbic (2004).

Associating each node $\delta \in \Delta$ with a variable X_δ, a graphical model is defined by a pair (G, P_Δ), where P_Δ is a probability distribution over the variables in X_Δ satisfying a set of restrictions induced by G. Given a graphical model it is possible to ascertain if two sets of random variables X_A and X_B are conditionally independent given a third set of variables X_S, due to the global Markov property

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

3

From this property it immediately follows that if there exists no path between the nodes in A and the nodes in B then the two sets of variables X_A and X_B are independent of each other. If the graph G is decomposable P_Δ factorizes as

$$P_\Delta(X_\Delta) = \frac{\prod_{C \in \mathcal{C}(G)} P_C(X_C)}{\prod_{S \in \mathcal{S}(G)} P_S(X_S)} \tag{2}$$

If P_Δ is a multivariate Gaussian distribution, (G, P_Δ) is a Gaussian graphical model. The
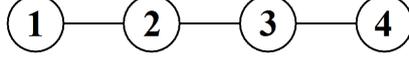

Figure 2: Graph containing four nodes and three edges.

restrictions induced by G are manifested as restrictions on the covariance matrix. Consider the graph in Figure 2 and the conditional independencies of the form described in (1) that it induces. These conditional independencies are

$$X_1 \perp X_3 \mid \{X_2, X_4\} \qquad X_1 \perp X_4 \mid \{X_2, X_3\} \qquad X_2 \perp X_4 \mid \{X_1, X_3\},$$

which results in a precision matrix of the form

$$K = \begin{pmatrix} k_{11} & k_{12} & 0 & 0 \\ k_{12} & k_{22} & k_{23} & 0 \\ 0 & k_{23} & k_{33} & k_{34} \\ 0 & 0 & k_{34} & k_{44} \end{pmatrix}.$$

A statement of conditional independence is such that it holds throughout the outcome space. Consider now the situation with $\Delta = \{1, 2, 3\}$ where the two variables X_2 and X_3 are independent only within a specific segment (or stratum) of the outcome space, for instance, given that the third variable X_1 is strictly positive, i.e.

$$X_2 \perp X_3 \mid X_1 > 0.$$

Such a local independence restriction cannot be captured by standard GGMs, which provides the motivation to develop a class of more general models termed as stratified Gaussian graphical models.

2.3 Stratified Gaussian graphical models

SGGMs belong to the class of context-specific models which allow for particular conditional independencies to be present only in a subset, or context, of the outcome space. The appearance and interpretation of an SGGM is quite similar to that of a GGM. For instance, in both types of models an edge between two nodes represents marginal dependence between two variables, conditional dependence is also modeled identically in GGMs and SGGMs. However, for SGGMs a context-specific independence can be introduced by assigning a specific condition to an edge in the graph. For example, in the previous section, where $\Delta = \{1, 2, 3\}$, the context-specific independence $X_2 \perp X_3 \mid X_1 > 0$ can be captured using the graph in Figure 3. The condition $X_1 > 0$ assigned to the edge $\{2, 3\}$ is referred to as a stratum.
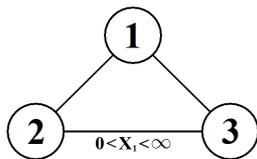
4

Figure 3: Stratified graph over three variables inducing the context-specific independence $X_2 \perp X_3 \mid X_1 > 0$.

Definition 1 (*Stratum*). *Let the pair (G, P_Δ) be a graphical model. 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 the outcome space of the variables in $X_{L_{\{\delta,\gamma\}}}$ such that $X_\delta \perp X_\gamma \mid X_{L_{\{\delta,\gamma\}}} = x_{L_{\{\delta,\gamma\}}}$, whenever $x_{L_{\{\delta,\gamma\}}} \in \mathcal{L}_{\{\delta,\gamma\}}$. Furthermore, $\mathcal{L}_{\{\delta,\gamma\}}$ must be definable by a union of sets that can be written*

$$\{x_{L_{\{\delta,\gamma\}}} : \bigcap_{\zeta \in L_{\{\delta,\gamma\}}} a_\zeta < x_\zeta < b_\zeta \text{ for some constants } a_\zeta, b_\zeta \in \mathbb{R} \cup \{\infty, -\infty\}\}. \tag{3}$$

Given the definition of a stratum an SGGM is defined as follows.

Definition 2 (*Stratified Gaussian graphical model*). *A stratified Gaussian graphical model is defined by the triple (G, L, P_Δ), where G is the underlying graph, L is the joint collection of all strata $\mathcal{L}_{\{\delta,\gamma\}}$ for the edges of G, and P_Δ is a piecewise Gaussian distribution satisfying the restrictions induced by (G, L). The pair (G, L) is termed a stratified graph (SG).*

In order to illustrate the difference between densities relating to GGMs and SGGMs consider two covariance matrices

$$\Sigma^{(1)} = \begin{pmatrix} 1 & 0.75 & 0.75 \\ 0.75 & 1 & 0.75 \\ 0.75 & 0.75 & 1 \end{pmatrix} \qquad \Sigma^{(2)} = \begin{pmatrix} 1 & 0.75 & 0.75 \\ 0.75 & 1 & 0.5625 \\ 0.75 & 0.5625 & 1 \end{pmatrix},$$

where $\Sigma^{(2)}$ is identical to $\Sigma^{(1)}$ except for the elements, $\sigma_{23}^{(2)} = \sigma_{32}^{(2)}$, which are altered to satisfy the condition $k_{23}^{(2)} = k_{32}^{(2)} = 0$. Let $\mu = (0, 0, 0)^T$ and let X_Δ be a random vector such that $X_\Delta \sim N(\mu, \Sigma^{(1)})$. Further, let Y_Δ be a random vector with the density

$$f_Y(y) = \begin{cases} (2\pi)^{-3/2} |K^{(1)}|^{1/2} e^{-1/2(y-\mu)^T K^{(1)}(y-\mu)}, & \text{if } y_1 \leq 0, \\ (2\pi)^{-3/2} |K^{(2)}|^{1/2} e^{-1/2(y-\mu)^T K^{(2)}(y-\mu)}, & \text{if } y_1 > 0, \end{cases}$$

where $K^{(1)}$ and $K^{(2)}$ are the inverse matrices of $\Sigma^{(1)}$ and $\Sigma^{(2)}$, respectively. It is obvious that f_Y defines a proper probability distribution since it is strictly positive and integrates to one due to the fact that using either $\Sigma^{(1)}$ or $\Sigma^{(2)}$ the probabilities $P(Y_1 > 0) = P(Y_1 < 0) = 0.5$. Additionally, this distribution follows the dependence structure induced by the SG in Figure 3. Figure 4 illustrates the noticeable difference between the conditional distributions of $X_1 \mid \{X_2, X_3\}$ and $Y_1 \mid \{Y_2, Y_3\}$, and shows the potential of a stratum to modify the shape of the density function.

For an SGGM containing a single stratum the interpretation of the context-specific dependence structure is straightforward. However, for SGGMs containing several strata, the dependence structure can prove less intuitive. To demonstrate this, consider the stratified graph in Figure 5a. This SG induces the two separate context-specific independencies: $X_2 \perp X_3 \mid X_1 > 0$
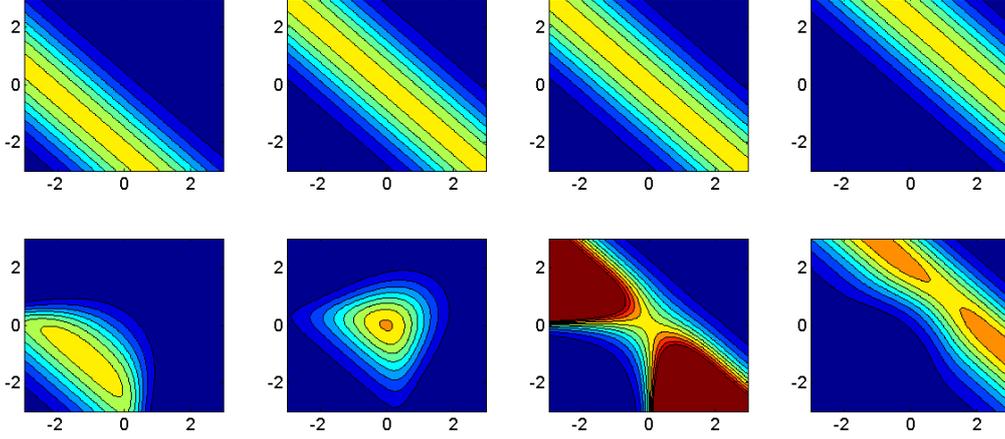
Figure 4: Top row: Conditional density function for X_1 given X_2 and X_3 (x- and y-axis), for $x_1 = -1$, $x_1 = 0$, $x_1 = 0.01$, and $x_1 = 1$. Bottom row: Conditional density function for Y_1 given Y_2 and Y_3 (x- and y-axis), for $y_1 = -1$, $y_1 = 0$, $y_1 = 0.01$, and $y_1 = 1$.
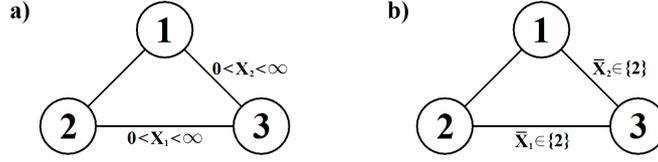

Figure 5: In (a) a stratified graph with two strata and in (b) its discretized counterpart.

and $X_1 \perp X_3 \mid X_2 > 0$. Intuitively, this would be interpreted to correspond to the case where the edge $\{1,3\}$ is removed in the context $X_2 > 0$, and the edge $\{2,3\}$ in the context $X_1 > 0$, consequently excluding both edges when the conditions $X_1 > 0$ and $X_2 > 0$ are simultaneously fulfilled. However, a closer scrutiny reveals that this is not a correct interpretation of the induced dependence structure. To show this, let $f(X_3 \mid \cdot)$ denote the conditional probability density function of X_3 given X_1 and/or X_2. Assume that a_1, a_2, and b_1 are all constants > 0. The following equalities will then hold for a density function following the dependence structure indicated by the SG in Figure 5a.

$$f(X_3 \mid X_1 = a_1) = f(X_3 \mid X_1 = a_1, X_2 = b_1) = f(X_3 \mid X_2 = b_1) = \tag{4}$$
$$f(X_3 \mid X_1 = a_2, X_2 = b_1) = f(X_3 \mid X_1 = a_2).$$

The equality $f(X_3 \mid X_1 = a_1) = f(X_3 \mid X_1 = a_2)$, whenever a_1 and $a_2 > 0$ will be denoted as $X_3 \perp X_1 \mid X_1 > 0$ or as $f(X_3 \mid X_1 = a_1) = f(X_3 \mid X_1 > 0)$. Analogous calculations will result in the independence restriction $X_3 \perp X_2 \mid X_2 > 0$. Recalling the assumption that a_1 and b_1 are positive constants, the following observation can be made

$$f(X_3 \mid X_1 = a_1, X_2 = b_1) = f(X_3 \mid X_1 = a_1) = f(X_3 \mid X_2 = b_1), \tag{5}$$

i.e. $f(X_3 \mid X_1 > 0) = f(X_3 \mid X_2 > 0)$. Now let $a > 0$ and $b \le 0$, using the results from (4) and (5) we obtain the following equalities

$$f(X_3 \mid X_1 = a, X_2 = b) = f(X_3 \mid X_1 = a) = f(X_3 \mid X_1 > 0) = f(X_3 \mid X_2 > 0).$$

6

A similar result is achieved when $b > 0$ and $a \leq 0$. Consequently, given that at least one of the conditions $X_1 > 0$ or $X_2 > 0$ is fulfilled, X_3 is independent of both X_1 and X_2. This leads to a model representation with two separate dependence structures, one for the context $X_1 \leq 0$ and $X_2 \leq 0$ depicted in Figure 6a, and another for the context $X_1 > 0$ or $X_2 > 0$ depicted in Figure 6b. This simple example illustrates the principle by which the complexity of the dependence
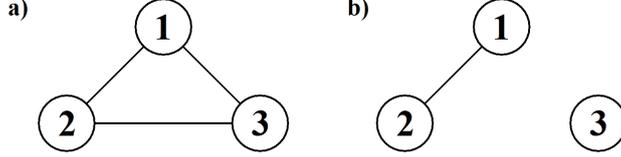

Figure 6: Different dependence structures for (a) the context $X_1 \leq 0$ and $X_2 \leq 0$ (b) the context $X_1 > 0$ or $X_2 > 0$.

structure increases as the number of variables and strata grow. In order to create a coherent modeling framework we therefore introduce the concept of a decomposable stratified graph. This will enable a discretization of the graph, translating a continuous SG into a discrete SG for which the induced dependence structure is thoroughly investigated in Nyman et al. (2014).

Definition 3 (*Decomposable SG*). *Let G be a decomposable graph and $G_L = (G, L)$ an SG with G as its underlying graph. In G_L, let E_L denote the set of all stratified edges (edges associated with a stratum), E_C the set of all edges in clique $C \in \mathcal{C}(G)$, and $E_\mathcal{S}$ the set of all edges in the separators $\mathcal{S}(G)$ of G. The SG is defined as decomposable if*

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

and

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

An SG is defined as decomposable if the underlying graph is decomposable, no strata are associated with edges in any separator, and in every clique all stratified edges have at least one node in common. An SGGM where (G, L) constitutes a decomposable SG is termed a decomposable SGGM. Restricting the underlying graph to be decomposable in combination with not allowing strata to be associated with edges in separators permits a factorization of the density function according to (2). This is due to the fact that for a decomposable SG the nodes in a stratified edge $\{\delta, \gamma\}$ and the nodes in $L_{\{\delta, \gamma\}}$ all belong to the same clique. Hence, the strata on an edge in one clique cannot imply changes to the dependence structure between variables associated to nodes in any other clique.

As the separators contain no stratified edges, the dependence structure between variables associated to nodes in a separator is trivial, the same holds for cliques containing no stratified edges. Therefore, further analysis of the dependence structure of decomposable SGGMs can be restricted to cliques of the underlying graph containing one or more stratified edges. This analysis is simplified by first transforming the continuous SG to a discrete SG, for which such an analysis is readily performed, at which point we can revert back to the continuous setting.

We start by defining a discretization process. Let C be the clique under consideration and let X_ζ be a variable such that $\zeta \in C$. Set Ω_ζ to be the set containing $-\infty$ and ∞, along with all

7

the endpoints of the intervals associated to X_ζ when defining the strata on the edges in C, i.e.

$$\Omega_\zeta = \{-\infty, \infty\} \cup \bigcup_{\substack{\delta,\gamma \in C \\ \zeta \in L_{\{\delta,\gamma\}}}} \{\omega : \omega = a_\zeta \text{ or } \omega = b_\zeta \text{ in a condition } a_\zeta < X_\zeta < b_\zeta \text{ in } \mathcal{L}_{\{\delta,\gamma\}}\}.$$

In addition, if an element ω features as a lower limit in one condition and as an upper limit in another condition it appears twice in Ω_ζ. Next the elements in $\Omega_\zeta = (\omega_1, \ldots, \omega_t)$ are sorted such that $\omega_i \leq \omega_{i+1}$. The elements in Ω_ζ are then used as endpoints when defining a set of intervals, $\{(\omega_1, \omega_2), (\omega_2, \omega_3), \ldots, (\omega_{t-1}, \omega_t)\}$. The interval endpoints can either be included or excluded, the first interval, (ω_1, ω_2), is always left-open. An interval, (ω_i, ω_{i+1}), is right-open if $\omega_i \neq \omega_{i+1}$ and ω_{i+1} is the upper limit in some stratum condition associated to ζ, i.e $\omega_{i+1} = b_\zeta$ in some condition $a_\zeta < X_\zeta < b_\zeta$ in a stratum, otherwise the interval is right-closed. If the interval (ω_{i-1}, ω_i) is right-open the interval (ω_i, ω_{i+1}) is left-closed and correspondingly, if (ω_{i-1}, ω_i) is right-closed the interval (ω_i, ω_{i+1}) is left-open. The last interval, (ω_{t-i}, ω_t), is always right-open. Following this method none of the intervals will overlap and the union of all intervals will equal $(-\infty, \infty)$. If a variable X_η does not appear in any condition in a stratum, or if all conditions are of the form $-\infty < X_\eta < \infty$, Ω_η will equal $\{-\infty, \infty\}$ resulting in the single interval $(-\infty, \infty)$. Each interval is then assigned an integer value, such that $(\omega_i, \omega_{i+1}) \to i$. This will allow for the translation of a condition in a stratum into a discrete form. If we by $\bar{X}_\zeta$ denote the discrete counterpart of X_ζ the condition $a_\zeta < X_\zeta < b_\zeta$, which is equivalent to $\omega_i < X_\zeta < \omega_{j+1}$ for some values i and j, can be written as $\bar{X}_\zeta \in \Lambda_\zeta = \{i, i+1, \ldots, j\}$. Once all the conditions used to define a set according to (3) have been converted to there discrete counterparts, the discrete version of the set can be written as

$$\bar{X}_{L_{\{\delta,\gamma\}}} \in \underset{\zeta \in L_{\{\delta,\gamma\}}}{\times} \Lambda_\zeta.$$

Here $\bar{X}_{L_{\{\delta,\gamma\}}}$ denotes the discretized versions of the variables in $X_{L_{\{\delta,\gamma\}}}$. This means that a single set will be converted into $\prod_{\zeta \in L_{\{\delta,\gamma\}}} |\Lambda_\zeta|$ discrete outcomes. Transforming all the sets composing a stratum to their discretized versions discretizes the stratum. Once all the strata in an continuous SG have been discretized the result is a discrete SG.

Next we make use of the trait that all stratified edges in a clique of a decomposable SG, and its discrete counterpart, have at least one node in common, this allows us to introduce an ordering of the d variables corresponding to clique C such that the last variable $\bar{X}_d$ in the ordering corresponds to the node found in all stratified edges. We will define the variables $(\bar{X}_1, \ldots, \bar{X}_{d-1})$, which are considered pairwise dependent in the entire outcome space, since the edges connecting the corresponding nodes are not stratified, as the parents of $\bar{X}_d$ and denote them by $\bar{\Pi}_d$. All the changes induced to the dependence structure by the introduction of strata can be seen in the conditional dependence of $\bar{X}_d$ given the set of variables $\bar{\Pi}_d$.

Context-specific independencies are readily illustrated using conditional probability tables, which assign a specific distribution to $\bar{X}_d$ for each outcome of $\bar{\Pi}_d$. Instead of $\bar{X}_d$ being assigned a unique distribution for each outcome of $\bar{\Pi}_d$, a partition of the outcome space of $\bar{\Pi}_d$ is devised, such that any two outcomes in the same block induce the same distribution of $\bar{X}_d$. Given our discretized SG we can utilize conditional probability tables, as a discrete stratum conveniently merges parent outcomes creating a partition of the outcome space of $\bar{\Pi}_d$.

Consider the edge $\{(d-1), d\}$ with the associated discrete stratum $(\bar{X}_1 = \lambda_1, \ldots, \bar{X}_{d-2} = \lambda_{d-2})$. This induces the context-specific independence $\bar{X}_{d-1} \perp \bar{X}_d \mid \{\bar{X}_1 = \lambda_1, \ldots, \bar{X}_{d-2} = \lambda_{d-2}\}$, which in terms of a conditional probability table corresponds to merging all outcomes of $\bar{\Pi}_d$ where $\bar{X}_1 = \lambda_1, \ldots, \bar{X}_{d-2} = \lambda_{d-2}$. Completing this procedure for all discrete strata will result in the desired partition of the outcome space of $\bar{\Pi}_d$. Each block of the partition is associated with

a specific dependence structure which can be ascertained from the conditional probability table. Given a block of outcomes an edge in the SG is deleted if any outcome in the block satisfies any condition in the discrete stratum associated to the edge, any stratum which is not satisfied is also deleted, resulting in each block being associated with a dependence structure determined by an ordinary graph.

Once this entire procedure is completed for all of the cliques found in the underlying graph the results can be combined and translated back to the continuous setting. This yields a set of conditions on the variables X_Δ that form a partition of the outcome space, each condition associated with a specific dependence structure in the form of an ordinary graph. The above described method offers a consistent approach to resolving the dependence structure induced by any decomposable SG, and will later be used when performing inference for SGGMs. We end this section with an example. Consider again the SG in Figure 5a, which corresponds to the sets $\Omega_1 = \{-\infty, 0, \infty\}$, $\Omega_2 = \{-\infty, 0, \infty\}$, and $\Omega_3 = \{-\infty, \infty\}$. The resulting intervals along with their discretized values are listed in Table 2. Using these discrete values we can form the

X_1	X_2	X_3
$(-\infty, 0] \to 1$	$(-\infty, 0] \to 1$	$(-\infty, \infty) \to 1$
$(0, \infty) \to 2$	$(0, \infty) \to 2$	

Table 2: Discretization of variables in the SG in Figure 5a.

discrete SG in Figure 5b, from which the conditional probability table for $\bar{X}_3$, found in Table 3, can be derived. The stratum $\bar{X}_1 \in \{2\}$ on the edge $\{2, 3\}$ will merge outcomes (3) and (4), while the stratum $\bar{X}_2 \in \{2\}$ on the edge $\{1, 3\}$ will merge outcomes (2) and (4). If $\bar{X}_1 \in \{2\}$ or

Outcome	X_1	X_2	Partition
(1)	1	1	p_1
(2)	1	2	p_2
(3)	2	1	p_2
(4)	2	2	p_2

Table 3: Conditional probability table resulting from the SGM in Figure 5b.

$\bar{X}_2 \in \{2\}$, corresponding to $X_1 \in (0, \infty)$ or $X_2 \in (0, \infty)$, both the edges $\{1, 2\}$ and $\{2, 3\}$ are deleted in that context resulting in the conditions and dependence structures found in Figure 6.

2.4 Identifiability of SGGMs

Identifiability of models with context-specific independence restrictions is of concern since one wishes to avoid situations where two distinct sets of restrictions lead to the same parametric model to retain interpretability and tractability of inference. Also the class of SGGMs necessitates a careful analysis of model identifiability. We illustrate that two distinct decomposable SGs may induce exactly the same dependence structure, as exemplified by the SGs in Figure 7a and Figure 7c. It is far from trivial to deduce that these two graphs induce the same dependence structure. To prove that this indeed is the case we apply the method from the previous section to the SG in Figure 7a. We start by defining the sets $\Omega_1 = \{-\infty, \infty\}$, $\Omega_2 = \{-\infty, 0, 2, \infty\}$, $\Omega_3 = \Omega_2$, and $\Omega_4 = \Omega_1$, which leads to the discretization in Table 4. Using this discretization we can form the discrete SG in Figure 7b and the corresponding conditional probability table for $\bar{X}_4$ in Table 5. The stratum on the edge $\{3, 4\}$ merges outcomes (1), (2), and (3) and (4), (5), and (6), while the
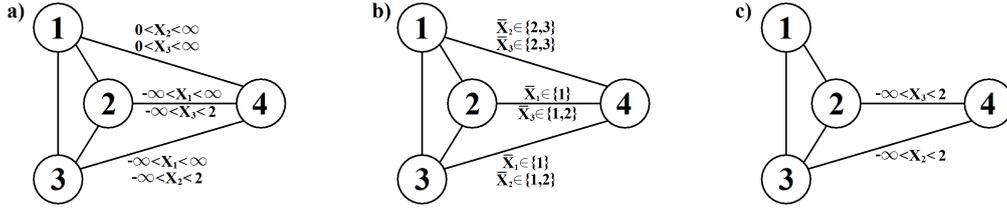
Figure 7: Two distinct SGs, (a) and (c), that look completely different but induce the same dependence structure. The graph in (b) is the discretized version of the graph in (a).

X_1	X_2	X_3	X_4
$(-\infty, \infty) \to 1$	$(-\infty, 0] \to 1$	$(-\infty, 0] \to 1$	$(-\infty, \infty) \to 1$
	$(0, 2) \to 2$	$(0, 2) \to 2$	
	$[2, \infty) \to 3$	$[2, \infty) \to 3$	

Table 4: Discretization of variables in the SG in Figure 7a.

stratum on the edge $\{2, 4\}$ merges the outcomes (1), (4), and (7) and (2), (5), and (8), resulting in a partition with only two blocks. Both blocks contain outcomes that satisfies the stratum on the edge $\{1, 4\}$, meaning that it will in all cases be deleted and can therefore be deleted from the underlying graph as well. Deleting the edge will result in X_1 being removed from $L_{\{2,4\}}$ and $L_{\{3,4\}}$. This, however, is irrelevant as the condition placed on X_1 in the strata $\mathcal{L}_{\{2,4\}}$ and $\mathcal{L}_{\{3,4\}}$ is of the form $-\infty < X_1 < \infty$ and can therefore be removed from the strata. As we can now see the SG in Figure 7a can be transformed into the SG in Figure 7c without altering the dependence structure. In order to construct a class of SGs in which no two graphs induce the

Outcome	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	Partition
(1)	1	1	1	p_1
(2)	1	1	2	p_1
(3)	1	1	3	p_1
(4)	1	2	1	p_1
(5)	1	2	2	p_1
(6)	1	2	3	p_1
(7)	1	3	1	p_1
(8)	1	3	2	p_1
(9)	1	3	3	p_2

Table 5: Conditional probability table for $\bar{X}_4$ resulting from the discrete SG in Figure 7b.

same dependence structure, the concept of maximal regular SG is introduced.

Definition 4 (*Maximal regular SG*). *A decomposable SG is defined as maximal regular if for no edge $\{\delta, \gamma\}$ in G does the set $\mathcal{L}_{\{\delta,\gamma\}}$ encompass the entire outcome space of the variables $X_{L_{\{\delta,\gamma\}}}$, nor can the set $\mathcal{L}_{\{\delta,\gamma\}}$ be expanded without altering the dependence structure.*

An SGGM where (G, L) constitutes a maximal regular SG is termed a maximal regular SGGM.

10

Theorem 1. *Two maximal regular SGs induce the same dependence structure if and only if they are identical.*

Proof of Theorem 1. Assume that $G_L^1 = (G^1, L^1)$ and $G_L^2 = (G^2, L^2)$ are two distinct maximal regular SGs that induce the same dependence structure. Further, assume that the underlying graphs G^1 and G^2 differ in at least one edge $\{\delta, \gamma\}$ which is present in G^1 but not in G^2. For G_L^2 it then holds that $X_\delta \perp X_\gamma \mid X_\Delta \backslash (X_\delta, X_\gamma)$. For the same to be true for G_L^1 the stratum $\mathcal{L}_{\{\delta, \gamma\}}^1$ has to consist of the entire outcome space of the variables $X_{L_{\{\delta, \gamma\}}}$. This contradicts the assumption that G_L^1 is maximal regular. Therefore, it can be concluded that $G^1 = G^2$. In order to prove that L^1 and L^2 are identical we start by assuming that there exists a set of outcomes $\omega \in \mathcal{L}_{\{\delta, \gamma\}}^2$ such that $\omega \notin \mathcal{L}_{\{\delta, \gamma\}}^1$. This means that the dependence structure for G_L^1 and G_L^2 encompasses the context-specific independence $X_\delta \perp X_\gamma \mid X_{L_{\{\delta, \gamma\}}} \in \omega$, and that ω can be added to $\mathcal{L}_{\{\delta, \gamma\}}^1$ without changing the dependence structure. Again, this contradicts the assumption that G_L^1 is maximal regular and proves that $L^1 = L^2$. $\qquad \square$

When performing inference restricting the model space to maximal regular SGs will decrease the size of the model space as well as at the same time avoiding the problem of different models being assigned the same likelihood due to fact that they induce identical dependence structures. In the next section we will prove that the family of distributions induced by a decomposable SG is a part of the curved exponential family.

2.5 SGGMs and curved exponential families

Distributions in GGMs belong to the exponential family, as shown for instance in Lauritzen (1996). A particularly useful characteristic of the exponential family is the consistency of the model selection criterion introduced by Schwarz (1978), often referred to as the Bayesian information criterion (BIC). Haughton (1988) extended the consistency result to the curved exponential family, which we will utilize for model selection among SGGMs. For an introduction to the statistical theory for the exponential and curved exponential family, see, for instance, DasGupta (2011).

The probability density function of a distribution in a decomposable SGGM is a function depending on the parameters corresponding to those of a multivariate normal distribution, i.e. the covariance Σ and mean μ. However, as is in general assumed for GGMs, we restrict the mean to zero and ignore it in the remainder of the article. In section 2.3 it was established that an SG induces a set of conditions on the variables X_Δ, that result in a partition of the outcome space where each block is associated with a distinct dependence structure represented by an ordinary graph. These conditions, which are functions of x, are denoted as $c^{(1)}, \ldots, c^{(\rho)}$. Using iterative proportional fitting, see, for instance Whittaker (1990), the covariance matrix Σ can be manipulated to reflect the dependence structure associated with any of the conditions. As such, each condition $c^{(r)}$ gives rise to a specific covariance matrix $\Sigma^{(r)}$ and corresponding precision matrix $K^{(r)}$ which are completely determined by the covariance matrix Σ and the dependence structure associated with $c^{(r)}$. Using this notation, the density function can be written as

$$g_\Sigma(x) = \frac{1}{Z} \sum_{r=1}^{\rho} f_{\Sigma^{(r)}}(x) I_{c^{(r)}}(x) = \frac{1}{Z} \sum_{r=1}^{\rho} (2\pi)^{-d/2} |K^{(r)}|^{1/2} e^{-\frac{1}{2} x^T K^{(r)} x} I_{c^{(r)}}(x), \qquad (6)$$

where $f_{\Sigma^{(r)}}(x)$ is the density function of the multivariate normal distribution with covariance matrix $\Sigma^{(r)}$ and Z is a normalizing constant. The terms $I_{c^{(r)}}(x)$ are indicator functions equaling 1, if x satisfies the condition $c^{(r)}$ and 0 otherwise. As each term in the sum is constituted by a density function of a multivariate normal distribution multiplied by an indicator function, each

term will be strictly positive in the part of the outcome space where the corresponding condition is fulfilled. Since the conditions induce a partition of the entire outcome space, guaranteeing that exactly one of the conditions will be fulfilled for every x, it follows that $g_\Sigma(x)$ is strictly positive for every x. The inclusion of the normalizing constant Z ensures that the integral of $g_\Sigma(x)$ over the entire outcome space is equal to one. Determining the value of Z is computationally straightforward as each block of the partition of the outcome space corresponding to a condition $c^{(r)}$ is of the form

$$\{x : \bigcap_\zeta a_\zeta^{(r)} < x_\zeta < b_\zeta^{(r)}, \text{ for all } \zeta \in \{1, \ldots, d\} \text{ for some constants } a_\zeta^{(r)} \text{ and } b_\zeta^{(r)}\},$$

which implies that Z can be calculated as

$$Z = \sum_{r=1}^\rho \int_{a_1^{(r)}}^{b_1^{(r)}} \ldots \int_{a_d^{(r)}}^{b_d^{(r)}} f_{\Sigma^{(r)}}(x) dx_d \ldots dx_1.$$

The following criterion is used to determine whether or not a continuous distribution belongs to the exponential family/curved exponential family.

Definition 5 (*Exponential family*). *A continuous distribution belongs to the exponential family if the probability density function can be written in the form*

$$f_\theta(x) = e^{\sum_{i=1}^k \eta_i(\theta)T_i(x) - \psi(\theta)} h(x),$$

where k, which is the length of vectors η and T, equals the dimension of the parameter θ. In the case when k exceeds the dimension of θ the distribution belongs to the curved exponential family.

Theorem 2. *The distribution in a decomposable SGGM belongs to the curved exponential family.*

Proof of Theorem 2. See Appendix A. □

3 Inference for SGGMs

3.1 Score function for SGGMs

To perform inference and model selection in the class of SGGMs, we adopt an approximate Bayesian approach based on the model scoring criterion introduced by Schwarz (1978) combined with a stochastic search for optimal models. We will apply a non-reversible Markov chain Monte Carlo algorithm, introduced by Corander et al. (2006) and further developed in Corander et al. (2008), to identify the model with the optimal score, which consistently approximates the mode of the posterior distribution over the space of the considered models.

Let $\mathbf{X} = (x_{ij})_{i=1,j=1}^{d,n}$ be a matrix consisting of n exchangeable observations of a d-dimensional random vector, assuming no missing data. In general, when performing inference in order to ascertain an optimal dependence structure the score function would equal the posterior probability, defined as

$$P(G_L \mid \mathbf{X}) = \frac{P(G_L, \mathbf{X})}{P(\mathbf{X})} = \frac{P(\mathbf{X} \mid G_L)P(G_L)}{\sum_{G_L \in \mathcal{G}} P(\mathbf{X} \mid G_L)P(G_L)},$$

where $P(G_L)$ is a prior distribution on the space of SGs, denoted by $\mathcal{G}$, and $P(\mathbf{X} \mid G_L)$ is the marginal likelihood calculated as the expectation $\int_{\Theta_{G_L}} P(\mathbf{X} \mid \theta_{G_L}, G_L)P(\theta_{G_L})d\theta_{G_L}$ of the likelihood with respect to the prior distribution of the parameters of G_L. Since analytical calculation

12

of the marginal likelihood appears intractable for stratified graphs, in contrast to ordinary graphs under conjugate priors (Dawid and Lauritzen, 1993), we use a consistent approximation of the log marginal likelihood based on the maximum likelihood function under the restrictions imposed by G_L combined with the BIC penalty function:

$$\log P(\mathbf{X} \mid G_L) \approx \log l(\mathbf{X} \mid G_L) - \frac{k(G_L)}{2} \log n = S(G_L \mid \mathbf{X}), \tag{7}$$

here $l(\mathbf{X} \mid G_L)$ is the maximized value of the likelihood function induced by G_L and $k(G_L)$ is the cardinality of the parameter space induced by G_L. The likelihood function can be expressed in the form

$$\prod_{j=1}^{n} g_\Sigma(x_j) = \prod_{j=1}^{n} \frac{1}{Z} \sum_{r=1}^{\rho} f_{\Sigma^{(r)}}(x_j) I_{c^{(r)}}(x_j),$$

here x_j is the jth column of $\mathbf{X}$. Using the Bayesian information criterion approximation, the problem reduces to finding the maximum likelihood estimate of the model parameters for any given candidate of the dependence structure. For ordinary decomposable graphs the maximum likelihood estimate $\hat{\Sigma}$ of Σ is analytically tractable, and it is relatively simple to obtain an estimate even for non-decomposable graphs. A straightforward approach is to first calculate the maximum likelihood estimate of the covariance without imposing any constraints and then enforce the constraints of the graph by using iterative proportional fitting. However, this method is not directly applicable to SGs, since the dataset is partitioned, with each partition associated with its own dependence structure.

Currently we are not aware of any method for analytically calculating $\hat{\Sigma}$ for SGs. Instead, we exploit a method that cyclically optimizes one element σ_{ij} at a time until sufficient numerical convergence is reached for the whole structure Σ. Importantly, not all elements in the covariance structure need to be optimized in this manner, since they will in each instance $\Sigma^{(1)}, \ldots, \Sigma^{(\rho)}$ be determined by certain other elements in Σ. To illustrate this, consider the example with the SG in Figure 7c for which the corresponding covariance matrix is shown below.

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{pmatrix}$$

The elements inside the blue rectangles are those that will be identical for all covariance matrices $\Sigma, \Sigma^{(1)}, \ldots, \Sigma^{(\rho)}$. This follows from the fact that iterative proportional fitting never changes elements on the diagonal, i.e. σ_{ii}, nor does it change an element σ_{ij} if there is an edge between nodes i and j in the graph. Since the elements inside the blue rectangles correspond to pairs of nodes connected by edges with no associated strata, these elements will be identical in $\Sigma^{(1)}, \ldots, \Sigma^{(\rho)}$.

The elements inside the green rectangles corresponds to pairs of nodes connected by stratified edges, meaning that these elements will be changed in some of the covariance matrices $\Sigma^{(1)}, \ldots, \Sigma^{(\rho)}$. The elements in the red rectangles correspond to pairs of nodes that are not connected by an edge, which means that they can take different values in all of the covariance matrices. The value of these elements in $\Sigma^{(1)}, \ldots, \Sigma^{(\rho)}$ will be completely determined by the other elements and it is therefore unnecessary to include them in the optimization process. The numerical optimization of each element σ_{ij} is carried out as follows.

1. Choose the value of two scalars $\epsilon < \delta$.

13

2. Evaluate the likelihood function using both values $\sigma_{ij} \pm \epsilon/2$ as candidates of σ_{ij}. If $\sigma_{ij} + \epsilon/2$ yields a higher likelihood than σ_{ij}, set $d = 1$, otherwise if $\sigma_{ij} - \epsilon/2$ yields higher likelihood, set $d = -1$. If both $\sigma_{ij} \pm \epsilon/2$ produce inferior results or results in Σ being negative semi-definite, stop the optimization process for σ_{ij}.

3. Evaluate the likelihood function using $\sigma_{ij}^* = \sigma_{ij} + d \times \delta$ as a candidate of σ_{ij}. If this improves the likelihood function, set $\sigma_{ij} = \sigma_{ij}^*$, otherwise set $\delta = \max(\epsilon/2, \delta/2)$ and repeat step 3.

4. Repeat steps 2 - 4.

It was determined above that some covariance elements will be completely determined by the other elements and need not be included in the optimization process. These elements will, however, affect whether or not Σ is positive definite or not. Therefore, it will be necessary to, at the end of each cycle, transform Σ to comply with the restrictions induced by the underlying graph.

Additionally, a criterion to determine whether or not Σ has converged needs to be defined. One possible approach would be to terminate the procedure when a whole cycle has been completed without changing any of the elements in Σ. Another, more pragmatic definition which we apply in our illustrations, depends on the improvement in the likelihood function after each cycle. When the resulting improvement in the likelihood function during a complete cycle is less than a predefined tolerance value, the estimation procedure is terminated. Given an appropriate starting value, for example, the sample covariance matrix, the time needed for convergence is tractable for a moderate number of variables.

The parameter space for multivariate normal distributions corresponding to the complete graph spanning d variables contains $(d^2 + d)/2 + d$ free parameters, equaling the sum of parameters found in Σ and μ. Although we have restricted μ to zero, this corresponds to a preprocessing of the data and therefore the contribution of μ to the number of free parameters is included. Removing an edge from the graph corresponds to removing a free parameter from the parameter space. This can be seen from the precision matrix as each absent edge in the graph corresponds to a conditional independence of the type found in (1), which in turn corresponds to forcing an element to equal zero in the precision matrix. Using this method we can deduce the number of free parameters induced by the underlying graph of the SG, however, we also need to take into account the number of parameters needed to define the set of strata included in the SG. The number of these parameters can be specified by studying the structure of a stratum. Each set of the form in (3) requires the introduction of $2 \times |L_{\{\delta,\gamma\}}|$ new parameters, where $|L_{\{\delta,\gamma\}}|$ denotes the number of nodes adjacent to both δ and γ. Let $|\mathcal{L}_{\{\delta,\gamma\}}|$ denote the number of sets of form (3) used to define $\mathcal{L}_{\{\delta,\gamma\}}$ and let $|E^c|$ denote the difference between the number of edges in G and the corresponding complete graph. The cardinality of the parameter space induced by an SG then equals

$$k(G_L) = \frac{d^2 + d}{2} + d - |E^c| + \sum_{\delta=1}^{d-1} \sum_{\gamma=\delta+1}^{d} |\mathcal{L}_{\{\delta,\gamma\}}| \times 2 \times |L_{\{\delta,\gamma\}}|.$$

As an example, for the SG in Figure 7c the cardinality of the parameter space equals

$$k(G_L) = (4^2 + 4)/2 + 4 - 1 + 1 \times 2 \times 1 + 1 \times 2 \times 1 = 17.$$

Given the obtained maximum likelihood estimate $\hat{\Sigma}$ and the cardinality of the parameter space induced by the SG, equation (7) can be used to approximate the posterior probability for SGGMs.

3.2 Non-reversible Markov chain Monte Carlo search for SGGMs

The learning algorithm described below belongs 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 models over which the aim is to identify the model with the optimal score. For $M \in \mathcal{M}$, let $Q(\cdot \mid M)$ denote the proposal function used to generate a new candidate model given any model M. Under the generic conditions stated in Corander et al. (2008), the probability assigned to any particular candidate by $Q(\cdot \mid M)$ need not be explicitly calculated or known, as long as it remains unchanged over the iterations of the algorithm and the resulting chain satisfies the condition that all states can be reached from any other state in a finite number of steps. Assume that the model learning is initialized by a model M_0. At iteration $t = 1, 2, ...$ of the algorithm, $Q(\cdot \mid M_{t-1})$ is used to generate a candidate model M^*, which is accepted with the probability

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

where $P(M)$ is a prior probability assigned to M and $P(\mathbf{X} \mid M)$ is the marginal likelihood of the dataset $\mathbf{X}$ given M.

Contrary to reversible Markov Chains, for non-reversible Markov chains the posterior probability $P(M \mid \mathbf{X})$ is not approximated by the stationary distribution. Instead, a consistent approximation of $P(M \mid \mathbf{X})$ is obtained by considering the space of distinct models $\mathcal{M}_t$ visited by time t such that

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

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

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

as $t \to \infty$. Since our main interest lies in finding the posterior optimal model, i.e.

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

it will suffice to identify

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

Throughout this article we use a uniform prior distribution over the model space, which further simplifies the search algorithm as the prior then cancels out in all the formulas. As an approximation, we will replace the marginal likelihood with the score function $S(G_L \mid \mathbf{X})$, resulting in a stochastic search for the model with optimal score. The proposal function used in our algorithm is available in Appendix B.

As the model space consisting of all maximal regular SGs grows extremely fast in relation to the number of nodes in the system, identifying a good initial state M_0 for the non-reversible Markov chain is of importance. A viable initial state is found by first conducting a search for the optimal undirected graph and then performing a search for strata separately for each edge included in this graph. Combining the graph with the resulting strata, or a subset of the strata in case the entire set results in a non maximal regular SG, provides a reasonable initial state using less computationally demanding operations compared to an algorithm that can traverse the entire space of maximal regular SGs.

4 Illustrations

We start by re-visiting the mathematic marks dataset mentioned in Section 2. Conducting a search for the optimal GGM using the framework described in the previous section, but modified to only consider traditional GGMs, results in a model with the graph displayed in Figure 1. This model has the score of -1731.33. However, by enlarging the model space to also include SGGMs, the optimal model identified has the score -1730.21, the corresponding stratified graph is shown in Figure 8. This SGGM, while having the same conditional dependence structure as the optimal
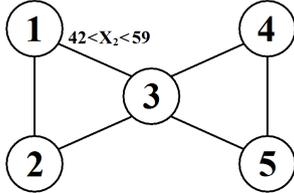

Figure 8: Optimal SG for mathematic marks data.

GGM, also incorporates the context-specific independence $X_1 \perp X_3 \mid X_2 \in (42, 59)$. The marks for X_2 range from 9 to 82 with the interval $(42, 59)$ composing 39 observations. Considering the clique $\{1, 2, 3\}$ along with the entire dataset the partial correlation between variables X_1 and X_3 equals 0.3171. However, if we only consider the subset of data where $X_2 \in (42, 59)$ the corresponding value equals -0.0017. Clearly, the included stratum manages to identify a subset of data where X_1 and X_3 are, in practice, conditionally independent given X_2.

In order to show that our scoring and search methods perform as intended, we generate a synthetic dataset following a distribution where the dependence structure can be represented by the SG in Figure 9. The specific distribution is available in Appendix C. The model space for
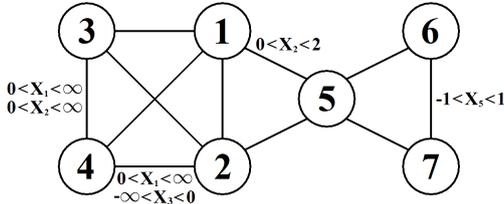

Figure 9: Dependence structure of the distribution used to generate synthetic data.

SGs is in theory infinitely large as the endpoints of a stratum are defined by continuous variables. However, when trying to fit an SG to a dataset, the model space can be considered finite as, for instance, a condition $0 < X_\zeta < a$ will in practice be the same as $0 < X_\zeta < b$ if $a \leq b$ and there exist no observations for which $X_\zeta \in [a, b)$. Nevertheless, the model space for SGs is still astronomically large when compared to the model space for ordinary graphs. Consequently, in order to preform solid inference for SGs we generally expect that larger datasets are required than for ordinary graphs.

In our experiments, a dataset containing more than $1,000$ observations would generally yield an inferred model very close to the generating model. The most challenging part is to correctly identify the endpoints for the interacting strata associated to the edges $\{2, 4\}$ and $\{3, 4\}$, since minor changes in them only lead to subtle changes in the dependence structure. Several synthetic
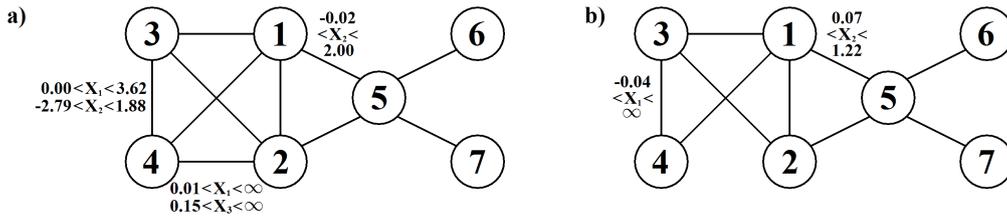
Figure 10: SGs with optimal score for datasets containing 1,000 observations (a) and 100 observations (b).

datasets of $1,000$ and 100 observations were generated. From these a single representative dataset of each size was chosen to demonstrate how a small number of data points limits the power to infer the generating graph. The dependence structure of the SGGMs with optimal score are displayed in Figure 10. The scores of the generating model for the two datasets are -6149.06 and -674.11, respectively. The scores for the optimal models are -6134.03 and -656.88, respectively. Comparing these models with the generating model one can see considerable similarities, however, some discrepancies are also present.

The edge $\{6,7\}$, which in the generating model is a stratified edge, is missing in both of the optimal models. This can be explained by the amount of observations that belong to the outcome space where the stratum is satisfied. The variance of X_5, which is the variable defining the stratum, is slightly larger than 1, which implies that roughly 70% of the observations on X_5 will reside within the interval $(-1, 1)$ satisfying the stratum. This can be compared to the approximately 50% of observations belonging to the stratum associated to the edge $\{1,5\}$, which is present in both of the optimal models. In other words, for a large majority of observations X_6 and X_7 are conditionally independent given X_5. This in combination with the fact that in the context where X_6 and X_7 are conditionally dependent given X_5, the absolute value of X_5 is large and will thus have a larger impact on X_7 than will X_6. In combination these circumstances will lead to the edge $\{6,7\}$ being relatively weak, requiring a large amount of data in order to be conclusively supported since the scoring criterion will penalize and attempt to filter out weak associations from the models.

Even for the larger dataset the optimal endpoints in the strata $\mathcal{L}_{\{2,4\}}$ and $\mathcal{L}_{\{3,4\}}$ can differ from those specified in the generating model. However, the actual implications on the dependence structure are not that extreme. For the smaller dataset the edge $\{2,4\}$ is not present in the optimal model, this can again be explained by the relative weakness of the edge, often resulting in either the edge $\{2,4\}$ or $\{3,4\}$ being omitted from the optimal model in the multiple realizations analyzed. The exclusion of one of these edges greatly simplifies the task of finding a suitable stratum for the remaining edge. In summary, even for relatively small datasets, the search method produces a fairly accurate approximation to the generating model. It is, however, evident that in order to capture more subtle dependencies a larger amount of data will be required.

Next we will consider two real datasets involving gene expression data and protein expression data. However, in order to do this it will first be necessary to slightly modify the score function in (7). When the number of variables considered in a model is large, Gaussian graphical models have a tendency to be quite dense, often overestimating the number of included edges (Foygel and Drton, 2010). To compensate for this issue Foygel and Drton (2010) suggested the inclusion of an additional penalty term in the BIC resulting in the *extended BIC*. This corresponds to

modifying our score function according to

$$S_{\exp}(G_L \mid \mathbf{X}) = \log l(\mathbf{X} \mid G_L) - \frac{k(G_L)}{2} \log n - \frac{|E|}{2} \log \kappa |\Delta|, \qquad (8)$$

where $|E|$ denotes the number of edges found in G_L, $|\Delta|$ the number of nodes, and κ is a tuning parameter. Choosing a suitable value of κ will result in a graph with optimal comprehensibility, as it is very difficult to determine strong dependencies from graphs that are very dense and, contrary, graphs with very few edges might leave out fairly strong dependencies. Foygel and Drton (2010) showed that the extended BIC is asymptotically equivalent to BIC as the number of observations, n, goes to infinity. When considering SGGMs operating with fairly sparse graphs facilitates the inference of strata as it is easier to identify viable strata in smaller cliques than in larger cliques. Therefore, the use of the extended BIC can also be extremely useful when inferring SGGMs for large systems.

For the gene expression data 15 variables were randomly chosen from the dataset found in Hiissa et al. (2009) containing 335 microarray observations that were quantile normalized prior to our analysis as described in the data source. The considered genes are available in Appendix D. For this dataset the extended BIC score using $\kappa = 2.5$ was used. The resulting optimal SG, shown in Figure 11, has a score of -16967.38. Instead of giving the strata explicitly the stratified
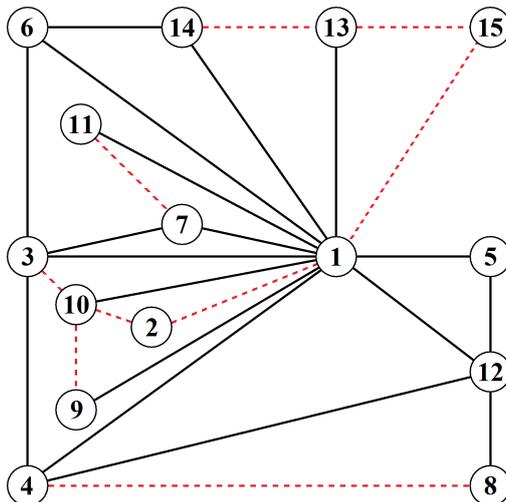

Figure 11: Optimal SG for gene expression data.

edges are shown using dashed lines. The SG contains 27 edges and nine stratified edges. The underlying graph of the SG is very similar to the graph used in the optimal GGM with the only difference being that the edge $\{2, 10\}$ is included in the SG but not in the ordinary graph. The graph of the optimal GGM has a score of -17027.92. If we again consider a single clique, for instance $\{1, 7, 11\}$ containing the context-specific independence $X_7 \perp X_{11}|X_1 \in (96.37, 118.49)$, the partial correlation between X_7 and X_{11} is 0.4021 for the entire dataset. Considering only the data where $X_1 \in (96.37, 118.49)$ the corresponding value is 0.0925, i.e. considerably closer to 0 compared to the value for the entire data.

The protein data is taken from Kornblau et al. (2009) and contains 256 observations on 51 variables. The optimal SG is displayed using the adjacency matrix in Figure 12, the circles represent stratified edges, the triangles edges to which strata could be added while still retaining

18

a decomposable SG, and the squares edges that cannot be stratified in a decomposable SG. For
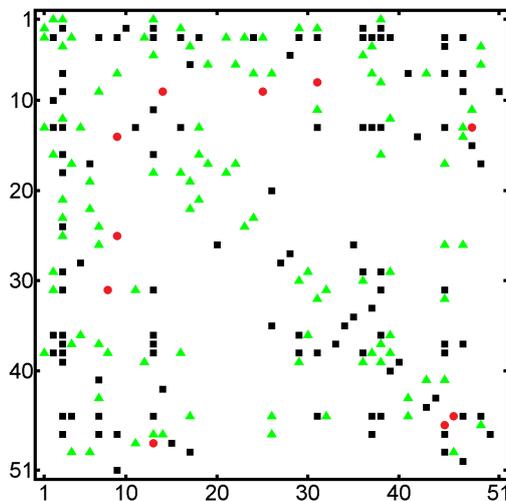

Figure 12: Optimal SG for protein expression data displayed using an adjacency matrix.

this dataset the extended BIC score using $\kappa = 0.5$ was used. The underlying graph of the inferred SG coincides with the graph of the optimal GGM. However, the SG also contains five stratified edges improving the score from -12548.47 for the ordinary graph to -12532.40.

5 Discussion

Gaussian graphical models have gained popularity for a multitude of applications involving analysis of multivariate continuous systems, in analogy with their counterparts for discrete variables. For discrete graphical models several approaches have been proposed for generalizing the dependence structure by local or context-specific independencies such that more flexible model classes are obtained which can reveal additional hidden independencies in data. To the best of our knowledge, such an approach has not been previously adopted for Gaussian graphical models. Using the concept of stratification of the joint outcome space, introduced by Nyman et al. (2014) for discrete graphical models, we showed here that context-specific independence generalizes well to the multivariate Gaussian distribution, such that both interpretability and inferential tractability are preserved. Despite of our stratification of the outcome space, the resulting models are not typical mixture-type latent class models, for which inference is notoriously challenging, especially in the multivariate setting considered here. An interesting further generalization of the SGGM class would be to consider an adaptation to directed Gaussian graphical models, for which Bayesian learning has been recently considered in Consonni and Rocca (2012). A potential solution to obtaining such a generalization would be to employ the concept of labeled directed acyclic graphs, introduced for discrete-valued systems by Pensar et al. (2014). Valuable additional insight to the applicability of SGGMs could also be obtained by developing faster inference tools suitable for the analysis of large continuous systems.

Acknowledgement

Appendix A

Proof of Theorem 2. It is a well established fact that multivariate normal distributions and distributions in GGMs belong to the exponential family. To show that the distribution in a decomposable SGGM belongs to the curved exponential family, we consider first a simple example and then the general case to provide clearer intuition for the reader. The density function of a multivariate normal distribution with zero mean can, using x_i to denote the ith element of the column vector x, be written as

$$f_\Sigma(x) = (2\pi)^{-d/2}|K|^{1/2}e^{-\frac{1}{2}x^T K x} = \frac{e^{-\frac{1}{2}\sum_{i=1}^{d}x_i^2 k_{ii} - \sum\sum_{i<j}x_i x_j k_{ij}}}{(2\pi)^{d/2}|K|^{-1/2}},$$

which is in the exponential family form with $h(x) = (2\pi)^{-d/2}$, $\psi(\Sigma) = \log(|K|^{-1/2})$,

$$\eta(\Sigma) = (k_{11}, \ldots, k_{dd}, k_{12}, \ldots, k_{(d-1)d}),$$

and

$$T(x) = (-\frac{1}{2}x_1^2, \ldots, -\frac{1}{2}x_d^2, -x_1 x_2, \ldots, -x_{d-1}x_d).$$

Consider now the SG depicted in Figure 3. Following from (6) a density function following the dependence structure defined by this SG factorizes as

$$g_\Sigma(x) = \frac{1}{Z}\left((2\pi)^{-d/2}|K^{(1)}|^{1/2}e^{-\frac{1}{2}x^T K^{(1)}x}I_{x_1\le 0} + (2\pi)^{-d/2}|K^{(2)}|^{1/2}e^{-\frac{1}{2}x^T K^{(2)}x}I_{x_1>0}\right). \quad (9)$$

The covariance matrices $\Sigma^{(1)}$ and $\Sigma^{(2)}$ generating the precision matrices $K^{(1)}$ and $K^{(2)}$, respectively, are identical to Σ except for the element $\sigma_{2,3}^{(2)}$ (and $\sigma_{3,2}^{(2)}$) which is modified such that the corresponding value in the precision matrix $k_{2,3}^{(2)}$ equals zero. The equation in (9) can be re-written as

$$g_\Sigma(x) = \frac{1}{Z}\left(\frac{e^{-\frac{1}{2}\sum_{i=1}^{d}x_i^2 k_{ii}^{(1)} - \sum\sum_{i<j}x_i x_j k_{ij}^{(1)}}}{(2\pi)^{d/2}|K^{(1)}|^{-1/2}}I_{x_1\le 0} + \frac{e^{-\frac{1}{2}\sum_{i=1}^{d}x_i^2 k_{ii}^{(2)} - \sum\sum_{i<j}x_i x_j k_{ij}^{(2)}}}{(2\pi)^{d/2}|K^{(2)}|^{-1/2}}I_{x_1>0}\right).$$

Noting that Z is determined by Σ we can see that this density defines a distribution in the curved exponential family by setting $h(x) = (2\pi)^{-d/2}$, $\psi(\Sigma) = \log(Z)$,

$$\eta(\Sigma) = (k_{11}^{(1)}, \ldots, k_{dd}^{(1)}, k_{12}^{(1)}, \ldots, k_{(d-1)d}^{(1)}, \log(|K^{(1)}|^{1/2}),$$
$$k_{11}^{(2)}, \ldots, k_{dd}^{(2)}, k_{12}^{(2)}, \ldots, k_{(d-1)d}^{(2)}, \log(|K^{(2)}|^{1/2})),$$

and

$$T(x) = (-\frac{1}{2}f_1^{(1)}(x)^2, \ldots, -\frac{1}{2}f_d^{(1)}(x)^2, -f_1^{(1)}(x)f_2^{(1)}(x), \ldots, -f_{d-1}^{(1)}(x)f_d^{(1)}(x), I_{x_1\le 0},$$
$$-\frac{1}{2}f_1^{(2)}(x)^2, \ldots, -\frac{1}{2}f_d^{(2)}(x)^2, -f_1^{(2)}(x)f_2^{(2)}(x), \ldots, -f_{d-1}^{(2)}(x)f_d^{(2)}(x), I_{x_1>0}).$$

The functions $f_i^{(1)}(x)$ and $f_i^{(2)}(x)$ are defined as

$$f_i^{(1)}(x) = \begin{cases} x_i, & \text{if } x_1 \leq 0 \\ 0, & \text{if } x_1 > 0 \end{cases} \qquad f_i^{(2)}(x) = \begin{cases} x_i, & \text{if } x_1 > 0 \\ 0, & \text{if } x_1 \leq 0 \end{cases}$$

The general case is proven following the same approach as in the example above. The density function is written as

$$g_\Sigma(x) = \frac{1}{Z} \sum_{r=1}^{\rho} \frac{e^{-\frac{1}{2} \sum_{i=1}^{d} x_i^2 k_{ii}^{(r)} - \sum \sum_{i<j} x_i x_j k_{ij}^{(r)}}}{(2\pi)^{d/2} |K^{(r)}|^{-1/2}} I_{c^{(r)}}(x),$$

which again can be identified as a member of the curved exponential family by setting $h(x) = (2\pi)^{-d/2}$, $\psi(\Sigma) = \log(Z)$,

$$\eta(\Sigma) = (k_{11}^{(1)}, \ldots, k_{dd}^{(1)}, k_{12}^{(1)}, \ldots, k_{(d-1)d}^{(1)}, \log(|K^{(1)}|^{1/2}), \ldots,$$
$$k_{11}^{(\rho)}, \ldots, k_{dd}^{(\rho)}, k_{12}^{(\rho)}, \ldots, k_{(d-1)d}^{(\rho)}, \log(|K^{(\rho)}|^{1/2})),$$

and

$$T(x) = \left(-\frac{1}{2} f_1^{(1)}(x)^2, \ldots, -\frac{1}{2} f_d^{(1)}(x)^2, -f_1^{(1)}(x) f_2^{(1)}(x), \ldots, -f_{d-1}^{(1)}(x) f_d^{(1)}(x), I_{c^{(1)}}(x), \ldots, \right.$$
$$\left. -\frac{1}{2} f_1^{(\rho)}(x)^2, \ldots, -\frac{1}{2} f_d^{(\rho)}(x)^2, -f_1^{(\rho)}(x) f_2^{(\rho)}(x), \ldots, -f_{d-1}^{(\rho)}(x) f_d^{(\rho)}(x), I_{c^{(\rho)}}(x) \right).$$

The functions $f_i^{(r)}(x)$ are defined similarly to above as

$$f_i^{(r)}(x) = \begin{cases} x_i, & \text{if } c^{(r)} \text{ is satisfied,} \\ 0, & \text{otherwise.} \end{cases}$$

This establishes that the distribution in a decomposable SGGM belongs to the curved exponential family. $\qquad \square$

Appendix B

The following proposal function is used to generate a candidate graph G_L^* given the current graph G_L. The function is composed of five operators one of which is randomly chosen at each iteration:

1. Add or delete a randomly chosen edge in the underlying graph G, while ensuring that this operation does not result in a non-decomposable G. This operation may violate the compatibility between G and some of the strata in L. As a result some strata may need to be altered or removed. A stratum $\mathcal{L}_{\{\delta,\gamma\}}$ is removed if the edge $\{\delta, \gamma\}$ is included in a separator, if it is included in a clique containing less then three nodes, or if the edge is not present in the new underlying graph. Let $L_{\{\delta,\gamma\}}^O$ and $L_{\{\delta,\gamma\}}^N$ denote the set of nodes adjacent to both δ and γ in the old and new underlying graph, respectively. The stratum $\mathcal{L}_{\{\delta,\gamma\}}$ is altered if $L_{\{\delta,\gamma\}}^O \neq L_{\{\delta,\gamma\}}^N$. If a node ζ is present in $L_{\{\delta,\gamma\}}^O$ but not present in $L_{\{\delta,\gamma\}}^N$ the conditions $a_\zeta < X_\zeta < b_\zeta$ are removed from $\mathcal{L}_{\{\delta,\gamma\}}$. Alternatively, if ζ is present in $L_{\{\delta,\gamma\}}^N$ but not in $L_{\{\delta,\gamma\}}^O$ the condition $-\infty < X_\zeta < \infty$ is added to $\mathcal{L}_{\{\delta,\gamma\}}$. Merging two cliques by adding an edge can lead to the negation of the property that all stratified edges in a clique have at least one node in common, which is an essential property of decomposable

SGs. If this occurs the least possible amount of strata is removed in order to restore the property. Additionally, if an edge is added, a random stratum may be appended to the edge in accordance with the procedure in step 2.

2. Add a random stratum to a randomly chosen eligible edge in G. An edge is defined as eligible if adding a stratum to the edge does not result in a non-decomposable SG. This operation is performed by first randomly choosing an eligible edge $\{\delta, \gamma\}$. For each node $\zeta \in L_{\{\delta, \gamma\}}$ draw two values uniformly from the interval $(-2\sigma_\zeta, 2\sigma_\zeta)$, where σ_ζ is the standard deviation of X_ζ calculated from the data $\mathbf{X}_\zeta$. Use a_ζ and b_ζ to denote the smaller and larger value, respectively. If $a_\zeta < \min(\mathbf{X}_\zeta)$ set $a_\zeta = -\infty$ and analogously, if $b_\zeta > \max(\mathbf{X}_\zeta)$ set $b_\zeta = \infty$. Using the attained values the new stratum can be written in the form of equation (3) of the main manuscript.

3. Remove a randomly chosen stratum from L.

4. Change the lower and upper limits of all strata associated to edges in a randomly chosen clique of G. This is done by randomly choosing a clique in G and for each stratum associated to an edge belonging to this clique change the stratum's upper and lower limits. Let a_ζ and b_ζ be the lower and upper limits in such a stratum for the variable X_ζ. If $a_\zeta = -\infty$ set $a_\zeta = -3\sigma_\zeta$, draw a value q uniformly from the interval $(-0.5\sigma_\zeta, 0.5\sigma_\zeta)$ and set $a_\zeta^* = a_\zeta + q$, repeat this procedure for b_ζ. If $a_\zeta^* > b_\zeta^*$ switch the values, insuring that $a_\zeta^* < b_\zeta^*$. If $a_\zeta^* < \min(\mathbf{X}_\zeta)$ set $a_\zeta^* = -\infty$ and analogously, if $b_\zeta^* > \max(\mathbf{X}_\zeta)$ set $b_\zeta^* = \infty$. Set the new lower and upper limits for X_ζ in the stratum to be a_ζ^* and b_ζ^*, respectively.

5. Remove a randomly chosen stratum from L, then add a random stratum to a randomly chosen eligible edge following the procedure defined in step 2.

The probabilities with which the operations are selected can be chosen quite freely. However, in order to guarantee that each state can be reached from any other state, operations 1-3 need to be selected with a strictly positive probability. The resulting candidate is transformed to its maximal regular counterpart, if no such counterpart exists a new candidate is generated. This transformation may lead to a change in the number of free parameters and subsequently a change in the score. In other cases the transformation can remove ambiguity concerning graphs with different appearance inducing identical dependence structures.

Appendix C

The following conditional distributions, for variables X_1-X_7 given in Table 6, are used to define a probability distribution following the dependence structure induced by the SG in Figure 9. The stochastic variables Y_1-Y_9 follow the standard normal distribution.

Appendix D

The following genes and corresponding indexes are used in the gene expression data example.

References

Atay-Kayis, A. and Massam, H. A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. Biometrika, 92:317–335 (2005).

Variable	Condition	Conditional Distribution
X_1		$X_1 \sim N(0,1)$
X_2		$X_2 = (X_1 + Y_1)/\sqrt{2}$
X_3		$X_3 = (X_1 + X_2 + Y_2)/\sqrt{3}$
X_4	$X_1 \in (0,\infty) \wedge (X_2 \in (0,\infty) \vee X_3 \in (-\infty,0))$	$X_4 = (X_1 + Y_3)/\sqrt{2}$
	$X_1 \in (-\infty,0] \vee (X_2 \in (-\infty,0] \wedge X_3 \in [0,\infty))$	$X_4 = (X_1 + X_2 + X_3 + Y_4)/\sqrt{4}$
X_5	$X_2 \in (0,2)$	$X_5 = (X_2 + Y_5)/\sqrt{2}$
	$X_2 \in (-\infty,0] \vee X_2 \in [2,\infty)$	$X_5 = (X_1 + X_2 + Y_6)/\sqrt{3}$
X_6		$X_6 = (X_5 + Y_7)/\sqrt{2}$
X_7	$X_5 \in (-1,1)$	$X_7 = (X_5 + Y_8)/\sqrt{2}$
	$X_5 \in (-\infty,-1] \vee X_5 \in [1,\infty)$	$X_7 = (X_5 + X_6 + Y_9)/\sqrt{3}$

Table 6: Conditional distributions used to specify a probability distribution following the dependence structure of the SG in Figure 9.

Index	Gene	Index	Gene	Index	Gene
1	AS3MT	6	DCN	11	HS.387405
2	C10orf4	7	DENND2D	12	HS.403212
3	C20orf144	8	NSUN7	13	HS.500666
4	CHAD	9	FLJ27255	14	HS.520628
5	CSTL1	10	FRMD5	15	HS.537675

Table 7: Genes included in gene expression data example.

Boutilier, C., Friedman, N., Goldszmidt, M., and Koller, D. Context-specific independence in Bayesian networks. In Proceedings of the Twelfth Annual Conference on Uncertainty in Artificial Intelligence, pages 115–123 (1996).

Carvalho, C. M. and Scott, J. G. Objective Bayesian model selection in Gaussian graphical models. Biometrika, 96:497–512 (2009).

Chickering, D. M., Heckerman, D., and Meek, C. A Bayesian approach to learning Bayesian networks with local structure. In Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence, pages 80–89 (1997).

Consonni, G. and Rocca, L. L. Objective Bayes factors for Gaussian directed acyclic graphical models. Scand. J. Stat., 39:743–756 (2012).

Corander, J. Labelled graphical models. Scand. J. Stat., 30:493–508 (2003).

Corander, J., Ekdahl, M., and Koski, T. Parallell interacting MCMC for learning of topologies of graphical models. Data Min. Knowl. Discov., 17:431–456 (2008).

Corander, J., Gyllenberg, M., and Koski, T. Bayesian model learning based on a parallel MCMC strategy. Stat. Comput., 16:355–362 (2006).

DasGupta, A. Probability for Statistics and Machine Learning. Springer, New York (2011).

Dawid, A. and Lauritzen, S. Hyper-Markov laws in the statistical analysis of decomposable graphical models. Ann. Statist., 21:1272–1317 (1993).

Dempster, A. Covariance selection. Biometrics, 28:157–175 (1972).

Edwards, D. Introduction to Graphical Modelling. Springer-Verlag, New York, 2nd edition (2000).

Foygel, R. and Drton, M. Extended bayesian information criteria for Gaussian graphical models. In Advances in Neural Information Processing Systems 23, pages 604–612 (2010).

Geiger, D. and Heckerman, D. Knowledge representation and inference in similarity networks and Bayesian multinets. Artificial Intelligence, 82:45–74 (1996).

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

Golumbic, M. C. Algorithmic graph theory and perfect graphs. Elsevier, Amsterdam, 2nd edition (2004).

Haughton, D. On the choice of a model to fit data from an exponential family. Ann. Statist., 16:342–355 (1988).

Hiissa, J., Elo, L. L., Huhtinen, K., Perheentupa, A., Poutanen, M., and Aittokallio, T. Resampling reveals sample-level differential expression in clinical genome-wide studies. OMICS A Journal of Integrative Biology, 13:381–396 (2009).

Højsgaard, S. Split models for contingency tables. Comput. Statist. Data Anal., 42:621–645 (2003).

Højsgaard, S. Statistical inference in context specific interaction models for contingency tables. Scand. J. Stat., 31:143–158 (2004).

Jones, B. and West, M. Covariance decomposition in undirected Gaussian graphical models. Biometrika, 92:779–786 (2005).

Kornblau, S. M., Tibes, R., Qiu, Y. H., Chen, W., Kantarjian, H. M., Andreeff, M., Coombes, K. R., and Mills, G. B. Functional proteomic profiling of AML predicts response and survival. Blood, 113:154–164 (2009).

Lauritzen, S. L. Graphical models. Oxford University Press, Oxford (1996).

Li, H. and Gui, J. Gradient directed regularization for sparse Gaussian concentration graphs with applications to inference of genetic networks. Biostatistics, 7:302–317 (2006).

Mardia, K. V., Kent, J. T., and Bibby, J. M. Multivariate Analysis. Academic Press, London (1979).

Nyman, H., Pensar, J., Koski, T., and Corander, J. Stratified graphical models - context-specific independence in graphical models. Bayesian Anal.: in press (2014).

Pensar, J., Nyman, H., Koski, T., and Corander, J. Labeled directed acyclic graphs: a generalization of context-specific independence in directed graphical models. Data Min. Knowl. Discov.: in press (2014).

Schwarz, G. Estimating the dimension of a model. Ann. Statist., 6:461–464 (1978).

Sun, H. and Li, H. Robust Gaussian graphical modeling via l_1 penalization. Biometrics, 68:1197–1206 (2012).

Whittaker, J. Graphical models in applied multivariate statistics. Wiley, Chichester (1990).

Wong, F., Carter, C. K., and Kohn, R. Efficient estimation of covariance selection models. Biometrika, 90:809–830 (2003).

Yuan, M. and Lin, Y. Model selection and estimation in the Gaussian graphical model. Biometrika, 94:19–35 (2007).