Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

SCIENZE STATISTICHE

Ciclo 33

Settore Concorsuale: 13/D1 - STATISTICA

Settore Scientifico Disciplinare: SECS-S/01 – STATISTICA

PROFILE GRAPHICAL MODELS

Presentata da: Andrea Lazzerini

Coordinatore Dottorato

Monica Chiogna

Supervisore

Monia Lupparelli

Co-supervisore

Francesco C. Stingo

Esame finale anno 2021

To my parents and girlfriend

Abstract

This thesis concerns the theory and the inference of a new class of independence models based on a graphical representation that we name profile graphs. Multiple graph models are special cases in this class and the compatibility in terms of independence structure is derived with respect to chain graph models of different types. Inference and model selection based on both Lasso methodology and Bayesian theory are studied and implemented. The latter is specifically used for the selection of multiple Ising graph models. The thesis is composed of four chapters.

In the first chapter, we present a literature review of multiple and chain graphs. Markov properties are illustrated for undirected, bidirected, LWF chain and regression graphs. Parameterization and inference are also reviewed for data coming from both multivariate Gaussian and Bernoulli sampling schemes.

In the second chapter, a class of profile graphs is introduced for modelling the effect of an external factor on the independence structure of a multivariate set of variables. This class is quite general and includes multiple graphs and chain graphs as special cases. Conditional and marginal independence structures are explored by using profile undirected and bi-directed graphical models, respectively. These two families of graphical models are formally defined with their corresponding Markov properties. Furthermore, necessary conditions are derived to induce, for any profile undirected and bi-directed graph model, a compatible class of chain graph models of different type known as LWF chain graph and regression graph, respectively. An application on protein networks in various subtypes of acute myeloid leukemia is discussed.

In the third chapter, we propose two Bayesian approaches for the selection of Ising models associated to multiple undirected graphs. We devise a Bayesian exact-likelihood inference for low-dimensional binary response data, based on conjugate priors for log-linear parameters, where we implement a computational strategy that uses Laplace approximations and a Metropolis-Hastings algorithm that allows us to perform a stochastic model search. We also propose a quasi-likelihood Bayesian approach for fitting high-dimensional multiple Ising graphs, where the normalization constant results computationally intractable, with spike-and-slab priors to encode sparsity and MCMC algorithms for sampling from the quasi-posterior distribution which enables variable selection and estimation simultaneously. In both methods, we define a Markov Random Field prior on the graph structures, which encourages the selection of the same edges in related graphs. We finally perform simulation studies to compare the proposed approaches with competing methods.

Finally, in the fourth chapter we present some final remarks on Chapters 2 and 3.

Contents

Chapter 1

Literature review

Contents

1 Introduction

Graphical models are statistical models associated to a graph, a structure consisting of a set of *vertices* or *nodes* and a set of *edges*. The vertices of the graph represent observed random variables and the independence structure of the model is represented by missing edges. An interesting characteristic of graphical models is that many features and properties of the model can be simply read from the corresponding graph. The graph greatly simplifies the interpretation of the model, making its independence structure more immediate and intuitive. In addition, graphical models constitute a very versatile methodology that has proved useful in a wide range of applications, for example, genetics and image analysis. An historical overview of graphical models can be found in Cox and Wermuth (1996) and Lauritzen (1996). A wide range of families of graphical models are available; see Sadeghi and Lauritzen (2014) and references therein. The different families of models can be distinguished for the kind of graph associated to the probability distribution. There are models with symmetric relationships between variables represented by the so called undirected and bidirected graphs. *Undirected graphs*, also known as *Markov random fields*, are characterized by collections of conditional independence relationships (Lauritzen, 1996) where vertices in the graph can be joined by only undirected edges (——), whereas *bidirected graphs* are characterized by collections of marginal independencies (Kauermann, 1996) and edges in the graph can be only bidirected ($\longleftrightarrow$). Asymmetric relationships between variables are considered by the family of models associated with *directed acyclic graphs* (DAGs), also known as *Bayesian networks* (Bishop, 2006) where the graphs present only directed edges or arrows ($\longrightarrow$). Undirected graphs, bi-directed graphs and DAGs are special cases of *chain graphs* whose edge set may contain both not oriented and oriented edges (Drton, 2009). In this manuscipt we will focus on chain graphs and *multiple graphs*, i.e. a collection of graphs with same vertex set but different sets of edges. (Guo et al., 2011).

2 Multiple graphical models

2.1 Multiple graphs

In many applications it is more realistic to consider a collection of graphical models, due to the heterogeneity of the data involved, where the dependence structure of the variables

may differ with respect to one or more factors (Guo et al., 2011; Peterson et al., 2015). An example can be found in gene networks describing different subtypes of the same cancer: there are some shared pathways across different subtypes, and there are also links that are unique to a particular subtype. In these situations, the presence of edges may change in some graphs while in others not. Multiple graphs are two or more graphs with the same vertex set where in each graph the edge set can differ according to a factor. More formally let Y_V be a random vector with elements indexed by V, a set of p response variables, and X be a random variable corresponding to a categorical factor external to V, taking value $x \in \mathcal{X}$, with $|\mathcal{X}| = q$. A collection of *multiple graphs* is denoted by $G_{V|\mathcal{X}} = [G(x)]_{x\in\mathcal{X}}$, where each graph $G(x) = (V, E(x))$ is associated to the random vector $Y_V|\{X = x\}$, where V is the node set and $E(x)$ is the edge set which depends on x, $x \in \mathcal{X}$. For any couple $r, j \in V$ and $x \in \mathcal{X}$, if $(r, j) \in E(x)$ then we have an edge between r and j in the corresponding graph $G(x)$ while if $(r, j) \notin E(x)$ then the two nodes are disjoined in $G(x)$. If all the edges in the graphs are undirected, then we call them *multiple undirected graphs* denoted by $U_{V|\mathcal{X}} = [U(x)]_{x\in\mathcal{X}}$, where each graph $U(x) = (V, E_U(x))$ is an undirected graph while if all the edges are bidirected, then we call them *multiple bidirected graphs* denoted by $B_{V|\mathcal{X}} = [B(x)]_{x\in\mathcal{X}}$ and each graph $B(x) = (V, E_U(x))$ is a bidirected graph.

2.2 Markov properties

We firstly introduce some useful technical definitions. Let $G = (V, E)$ be a graph with vertex set V and edge set E. For any couple of vertices $i, j \in V$, if i and j are joined by a not oriented edge we say that they are *neighbors*. The set of neighbors of a vertex i in G is denoted as $nb(i)$. A *path* of length k from i to j is a sequence of distinct vertices $i = v_0, \ldots, v_{r-1}, v_r, \ldots, v_k = j$ such that the vertices v_{r-1} and v_r are neighbors for all $r = 1, \ldots, k$. Let A, B and C be three subsets of V; The subset C is said to *separate* the subsets A and B if all the paths from any vertex $i \in A$ to any vertex $j \in B$ intersects C. We denote with G_A the *induced subgraph* of G by the subset $A \subseteq V$, i.e. the graph with vertex set A and all those edges which join two vertices that are both in A. A graph $G = (V, E)$ is *connected* when every pair of distinct vertices in V is joined by a path. A nonempty subset $A \subseteq V$ is a connected set in G if the induced subgraph G_A is connected, and it is *disconnected* otherwise. Every nonempty subset $A \subseteq V$ can be partitioned uniquely into maximal connected sets, $A = K_1 \cup K_2 \cup \cdots \cup K_r$. The sets $K_1 \cup K_2 \cup \cdots \cup K_r$ are called the

connected components of A. As already mentioned above, graphical model uses a graph to represent conditional or marginal independence relations holding among a set of variables according to their joint probability ditribution. The rules that translate properties of the graph into conditional or marginal independence statements are called *Markov properties*. There are three Markov properties named *pairwise*, *local* and *global*, which vary according to the type of graph considered.

2.2.1 Markov properties for undirected graphs

Consider the undirected graph $U = (V, E_U)$. The probability distribution $P(Y_V)$ associated to U is said to obey

a) the *undirected pairwise Markov property* (UPMP) if for every $i, j \in V$ such that $(i, j) \notin E_U$ it holds that

$$Y_i \perp\!\!\!\perp Y_j | Y_{V \setminus \{i,j\}}, \tag{2.1}$$

b) the *undirected local Markov property* (ULMP) if for every $i \in V$ it holds that

$$Y_i \perp\!\!\!\perp Y_{V \setminus \{nb(i) \cup i\}} | Y_{nb(i)}, \tag{2.2}$$

c) the *undirected global Markov property* (UGMP) if for any triple of pairwise disjoint subsets $A, B, C \in V$ such that C separates A from B in U it holds that

$$Y_A \perp\!\!\!\perp Y_B | Y_C, . \tag{2.3}$$

It can be easily shown that the implications $UPMP \Longleftarrow ULMP \Longleftarrow UGMP$ always hold (Lauritzen, 1996). Unfortunately the reverse implications are not true in general; however, a sufficient condition for the reverse implications to hold is that the joint probability distribution is strictly positive (Pearl and Paz, 1987; Lauritzen, 1996). If $U = U(x)$, i.e. it belongs to a collection of multiple undirected graphs $[U(x)]_{x \in \mathcal{X}}$ with probability distribution $P(Y_V | \{X = x\})$, all the above statements and implications hold wrt $U(x)$, for all $x \in X$, if we extend the conditioning set with $\{X = x\}$.

2.2.2 Markov properties for bidirected graphs

Consider the bidirected graph $B = (V, E_B)$. The probability distribution $P(Y_V)$ associated to B is said to obey

d) the *bidirected pairwise Markov property* (BPMP) if for every $i, j \in V$ such that $(i, j) \notin E_B$ it holds that

$$Y_i \perp\!\!\!\perp Y_j, \tag{2.4}$$

e) the *bidirected local Markov property* (BLMP) if for every $i \in V$ it holds that

$$Y_i \perp\!\!\!\perp Y_{V \setminus \{nb(i) \cup i\}}, \tag{2.5}$$

f) the *bidirected global Markov property* (BGMP) if for any triple of pairwise disjoint subsets $A, B, C \in V$ such that C separates A from B in B it holds that

$$Y_A \perp\!\!\!\perp Y_B | Y_{V \setminus \{A \cup B \cup C\}}, \tag{2.6}$$

g) the *connected set Markov property* (CSMP) if for every disconnected set D of B it holds that

$$Y_{K_1} \perp\!\!\!\perp Y_{K_2} \perp\!\!\!\perp \ldots \perp\!\!\!\perp Y_{K_r}, \tag{2.7}$$

where $K_1, K_2, \ldots, K_r$ are the connected components of D.

The connected set Markov property is equivalent to the bidirected global Markov property as proved in Richardson (2003); see also Drton and Richardson (2008). The implications $BPMP \Longleftarrow BLMP \Longleftarrow BGMP \Longleftrightarrow CSMP$ are always true. Nevertheless, in this case the BPMP does not imply the BGMP, even for strictly positive distributions, so the reverse implications never hold (Richardson, 2003; Kauermann, 1996). Also in this case, if $B = B(x)$, i.e. it belongs to a collection of multiple bidirected graphs $[B(x)]_{x \in \mathcal{X}}$ with probability distribution $P(Y_V | \{X = x\})$, all the above statements and implications hold wrt $B(x)$, for all $x \in \mathcal{X}$, if we extend the conditioning set with $\{X = x\}$.

2.3 Parametrization and inference

2.3.1 Gaussian data

Let us consider a continuous random vector Y_V. For all $x \in \mathcal{X}$, we assume $Y_V | \{X = x\} \sim N(\alpha + \beta_x, \Sigma(x))$ where $[\alpha_i + \beta_{ix}]_{i \in V} = \mathbb{E}[Y_i | \{X = x\}]_{i \in V}$ is the *marginal mean vector* and $\Sigma(x)$ is the *covariance matrix* with entries $\omega_{ij}(x)$, all conditional to $X = x$, $x \in \mathcal{X}$. Let also γ_{ix}, $i \in V$, be the linear effect of X on the *conditional mean vector* $\mathbb{E}[Y_i | \{Y_{V \setminus i}, X = x\}]_{i \in V}$ and let $\Lambda(x) = \Sigma^{-1}(x)$ be the *precision matrix* with entries $\lambda_{ij}(x)$,

$x \in \mathcal{X}$; note that $\gamma_x = \Lambda(x)\beta_x$ where $\gamma_x = [\gamma_{ix}]_{i \in V}$ and $\beta_x = [\beta_{ix}]_{i \in V}$ (Andersson et al., 2001). Zero-constraints over the off-diagonal entries of the precision matrix $\Sigma(x)$ define a *Gaussian undirected graphical model* while zero-constraints over the off-diagonal entries of the covariance matrix $\Lambda(x)$ define a *Gaussian bidirected graphical model*, for any $x \in \mathcal{X}$. The Gaussian undirected graphical model for $Y_V|\{X = x\}$ denoted by $U(x) = (V, E_U(x))$ is such that, for any $x \in \mathcal{X}$ and any couple $i, j \in V$

h) if $(i, j) \notin E_U(x)$ then $\lambda_{ij}(x) = 0$.

The Gaussian bidirected graphical model for $Y_V|\{X = x\}$ denoted by $B(x) = (V, E_B(x))$ is such that, for any $x \in \mathcal{X}$ and any couple $i, j \in V$

i) if $(i, j) \notin E_U(x)$ then $\omega_{ij}(x) = 0$.

The problem of estimating a Gaussian graphical model is equivalent to estimating a covariance matrix. A first approach was proposed by Dempster (1972), who advocated the estimation of a sparse dependence structure, i.e., setting some elements of the inverse covariance matrix to zero. More recently, the focus has shifted to using regularization for sparse estimation of the covariance matrix. For instance, Meinshausen and Bühlmann (2006) proposed to select edges for each node in the graph by regressing each variable on all other variables using ℓ_1-penalized regression. Some approaches for inferring multiple graphical models have been proposed in recent years. Guo et al. (2011) proposed to infer multiple undirected graphs by expressing the elements of the precision matrix associated to each graph as a product of common and group-specific factors. From a Bayesian point of view, Peterson et al. (2015) infer multiple Gaussian undirected graphs by linking the estimation of the graph structures via a Markov random field (MRF) prior, which encourages common edges.

2.3.2 Binary data

Assume that we have observed $n^{(x)}$ realizations of $Y_V|\{X = x\}$ following a Multivariate-Bernoulli distribution with parameter $\pi^{(x)}$ for all $x \in \mathcal{X}$, where $\pi^{(x)} = [\pi_D^{(x)}]_{D \subseteq V}$ with $\pi_D^{(x)} = P(Y_D = 1_D, Y_{V \setminus D} = 0_{V \setminus D}|X = x)$, where 1_D is a vector of 1s of size $|D|$ and $0_{D \setminus V}$ is a vector of 0s of size $|V \setminus D|$. The *log-linear parameter* $\theta^{(x)} = [\theta_D^{(x)}]_{D \subseteq V}$ is an alternative parametrization obtained through the *Zeta matrix* ($\mathbb{Z}$) and the *Möbius matrix*

($\mathbb{M}$) (Roverato et al., 2013), $\mathbb{M} = \mathbb{Z}^{-1}$, i.e.

$$\theta^{(x)} = \mathbb{M}^T \log \pi^{(x)} \iff \pi^{(x)} = \exp(\mathbb{Z}^T \theta^{(x)}). \tag{2.8}$$

The log-linear parameter allow to express pairwise independencies on $Y_V|\{X = x\}$ as zero-constraints on $\theta^{(x)}$ with $x \in \mathcal{X}$. Let $U(x) = (V, E_U(x))$ be the undirected graph associated to $Y_V|\{X = x\}$, for any $i, j \in V$ if $(i, j) \notin E_U(x)$ it holds that $\theta^{(x)}_D = 0$ for all $D \supseteq \{i, j\}$. In case of binary response data it is possible to assume $Y_V|\{X = x\} \sim \text{Ising}(\theta^{(x)})$ (Besag, 1974) where $\theta^{(x)} = [\theta^{(x)}_{ij}]_{i,j\in V} \in \mathbb{R}^{p+(p\times(p-1))/2}$ is the loglinear parameter conditional to $X = x$, $x \in \mathcal{X}$ (Lauritzen, 1996). Note that for all $i, j \in V$ and any $x \in \mathcal{X}$, if $i = j$ then $\theta^{(x)}_{ij}$ represents the *main effects* while if $i \neq j$ then $\theta^{(x)}_{ij}$ represents the *two-way interaction* between variables i and j, in particular the logarithm of the *conditional odds ratio* of i and j given $X = x$. Zero-constraints over the two-way loglinear parameter $[\theta^{(x)}_{ij}]_{i,j\in V,i\neq j}$ define an *Ising undirected graphical model*. The Ising undirected graphical model for $Y_V|\{X = x\}$ denoted by $U_I(x) = (V, E_U(x))$ is such that for any $x \in \mathcal{X}$ and any couple $i, j \in C_1$ with $i \neq j$,

j) if $(i, j) \notin E_U(x)$ then $\theta^{(x)}_{ij} = 0$.

Let $\zeta^{(x)} = [\zeta^{(x)}_D]_{D\subseteq V}$ be the *log-mean linear parameter* (Roverato et al., 2013), a different parametrization obtained as

$$\zeta^{(x)} = \mathbb{M}^T \log \mu = \mathbb{M}^T \log \mathbb{Z}\pi, \tag{2.9}$$

where $\mu = [\mu_D]_{D\subseteq V}$ with $\mu_D = P(Y_D = 1_D)$. The *binary graphical model* for $Y_V|\{X = x\}$, $x \in \mathcal{X}$, denoted by $B(x) = (V, E_B(x))$ is such that for any disconnected set $D \subseteq V$ of $B(x)$ it holds that $\zeta^{(x)}_D = 0$. In this case the Ising model does not implies a simplification of the model. A crucial aspect of these models compared to the undirected ones, is that they are defined by many parameters and so the models are less parsimonious. When p is large, inference under the Ising model is difficult because of the intractability of the normalization constant. In particular, maximum likelihood estimation can generally not be performed. Various solutions have arisen in the literature. Ravikumar et al. (2010) proposed to use multiple ℓ_1-penalized logistic regressions, extending the approach developed by Meinshausen and Bühlmann (2006) in the Gaussian case. Several authors have studied the estimation of regression models on stratified data, suitable for multiple graphs, to take advantage of the

potential homogeneity among the corresponding strata. Tibshirani et al. (2005) proposed the fused lasso, a generalization that encourages sparsity of the coefficients and also sparsity of their differences. Ollier and Viallon (2017) developed an approach referred to as data shared lasso that bypasses the arbitrary choice of the reference stratum.

3 Chain graphical models

3.1 Chain graphs

Chain graphs are generally used for modelling the effect of background variables on joint response variables. Under suitable rules, chain graphs provide a different multivariate regression framework for modelling the independence structure of the outcomes given the explanatory variables. In a chain graph the vertex set can be partitioned into an ordered sequence of pairwise disjoint blocks. Directed edges are not allowed within blocks, and all the edges joining vertices belonging to different blocks are arrows pointing from the lower to the higher of the two blocks with respect to the ordering. There are several types of chain graph models (Drton, 2009), in this paper we will focus on two types, (i) LWF chain graph models (Frydenberg, 1990) and (ii) regression graph models (Wermuth and Sadeghi, 2012). More formally, a *LWF chain graph* $C_C = (V, E_U, E_D)$ is defined by a set of vertices V partitioned in chain components $C_0, C_1, \ldots, C_k$, a set of undirected edges E_U and a set of directed edges E_D. Vertices within any chain component can be joined by undirected edges and vertices between chain components can be joined by arrows preserving the same direction such that cycles are not allowed. A *regression graph* $C_R = (V, E_U, E_B, E_D)$ is defined by a set of vertices V partitioned in chain components $C_0, C_1, \ldots, C_k$, a set of undirected edges E_U, a set of bidirected edges E_B and a set of directed edges E_D. Vertices within the chain component C_0 can be joined by undirected edges while vertices within the remaining chain components $C_1, \ldots, C_k$, can be joined by bidirected edges. As in LWF chain graphs, vertices between chain components are joined by arrows preserving the same direction such that cycles are not allowed.

3.2 Markov properties

As above we firstly introduce other useful technical definitions. Let G be a chain or regression graph with blocks $C_0, C_1, \ldots, C_k$. If an arrow points from i to j we call i a *parent* of

j. The set of parents of a vertex i in G is denoted as $pa(i)$. Furthermore, the set of parents of a subset $A \subseteq V$ is defined as $pa(A) = \cup_{i \in A}\, pa(i)$. We define the set of *predecessors* of a block C_r wrt G as $pr(C_r) = C_0 \cup C_1 \cup \ldots, C_{r-1}$, for $r = 1, \ldots, k$. An important property is the factorization of the probability distribution of Y_V over the chain components of a chain graph. Let $Y_V = (Y_{C_r}) : \tau \in T(G)$, with $T(G) = (C_0, C_1, \ldots, C_k)$ be the random vector corresponding to the chain graph $G = (V, E)$; under several assumptions (Frydenberg, 1990; Drton, 2009) and if the probability distribution $P(Y_V)$ is strictly positive, then it holds the following factorization

$$P(Y_V) = \prod_{\tau \in T(G)} P(Y_\tau \mid Y_{pa(\tau)}). \tag{3.1}$$

3.2.1 Markov properties for LWF chain graphs

Consider the LWF chain graph $C_C = (V, E_U, E_D)$ with chain components $C_0, C_1, \ldots, C_k$. The probability distribution $P(Y_V)$ associated to C_C is said to obey the *LWF Markov property* when

k) for any $A \subseteq C_r$ and every $r = 1, \ldots, k$ it holds that

$$Y_A \perp\!\!\!\perp Y_{\{pr(C_r) \setminus pa(A)\}} | Y_{\{pa(A) \cup C_r \setminus A\}}, \tag{3.2}$$

l) the probability distribution of $Y_{C_r}|Y_{pr(C_r)}$ obeys the UGMP with respect to the induced subgraph G_{C_r}, for every $r = 0, 1, \ldots, k$; in particular when $r = 0$ then conditioning set is empty.

3.2.2 Markov properties for regression graphs

Consider the regression graph $C_R = (V, E_U, E_B, E_D)$ with blocks $C_0, C_1, \ldots, C_k$. The probability distribution $P(Y_V)$ associated to C_R is said to obey the *regression Markov property* when

m) for any $A \subseteq C_r$ and every $r = 1, \ldots, k$ it holds that

$$Y_A \perp\!\!\!\perp Y_{\{pr(C_r) \setminus pa(A)\}} | Y_{pa(A)}, \tag{3.3}$$

n) the probability distribution of $Y_{C_r}|Y_{pr(C_r)}$ obeys the BGMP with respect to the induced subgraph G_{C_r}, for every $r = 1, \ldots, k$.

o) the probability distribution of Y_{C_0} obeys the UGMP with respect to the induced subgraph G_{C_0}.

3.3 Parametrization and inference

Let $X_R = [X_r]_{r\in R}$ be a random vector with elements indexed by R, a set of u variables. Here we focus on a two-blocks LWF chain and regression graph with chain components $C_0 = R$ and $C_1 = V$.

3.3.1 Gaussian data

We present the case of Gaussian data for both the blocks. Consider the random vector (Y_V, X_R) with $(p+u)\times 1$ zero-mean vector and $(p+u)\times(p+u)$ positive definite covariance matrix Σ with entries σ_{ij} for all $ij \in \{V \cup R\}$. Therefore Σ_{ij}^{-1} is the corresponding $(p+u)\times(p+u)$ precision matrix with entries σ_{ij}^{-1} for all $ij \in \{V \cup R\}$. The conditional distribution of Y_V given X_R is multivariate Normal given by $Y_V|X_R \sim N_p(\beta X_R, \Sigma_{V|R})$ where $\Sigma_{V|R}$ is the $(p\times p)$ conditional covariance matrix with entries ω_{ij} for all $i,j \in V$, and β is the $(p\times u)$ matrix of regression coefficients of Y_V on X_R, with entries β_{ir} for any $i \in V$ and any $r \in R$. We denote with $\Sigma_{V|R}^{-1}$ the $(p\times p)$ conditional precision matrix with entries ω_{ij}^{-1} for all $i,j \in V$. Let also $\Gamma = \Sigma_{V|R}^{-1}\times\beta$ be the $(p\times u)$ matrix of parameters occurring in the conditional distribution of Y_V given Y_R, with entries γ_{ir} for any $i \in V$ and any $r \in R$.

The *LWF chain Gaussian graphical model* for (Y_V, X_R) denoted by $C_C = (\{V\cup R\}, E_U, E_D)$ is such that for any $i,j \in V$ and any $r,t \in R$

p) if $(i,r) \notin E_D$ then $\gamma_{ir} = 0$,

q) if $(i,j) \notin E_U$ then $\omega_{ij}^{-1} = 0$,

r) if $(r,t) \notin E_U$ then $\sigma_{rt}^{-1} = 0$.

The *regression Gaussian graphical model* for (Y_V, X_R) denoted by $C_R = (\{V\cup R\}, E_U, E_B, E_D)$ is such that for any $i,j \in V$ and any $r,t \in R$, if it hold r) and

s) if $(i,r) \notin E_D$ then $\beta_{ir} = 0$,

t) if $(i,j) \notin E_B$ then $\omega_{ij} = 0$,

From Equation (3.1) derives a decomposition of the likelihood associated to a chain graph. Indeed, it can be shown (Drton, 2009) that the likelihood of (Y_V, X_R) associated to a chain graph, can be maximized by maximizing the likelihood of Y_τ for every $\tau \in T(G)$ separately, and then by combining the optima according to (3.1). For each chain component, maximum likelihood estimates can be obtained by fitting an undirected or bidirected graph model to the residuals computed using the regression coefficients (linear effects) estimates (Speed and Kiiveri, 1986; Edwards, 2000).

3.3.2 Binary data

We present the case where both Y_V and X_R follow a multivariate Bernoulli distribution. We now introduce a further parameterization associated with the blocks (R, V) and consists of probabilities involving all the variables in X_R but only certain subvectors of Y_V. Formally, the *hybrid parameter* for (X_R, Y_V), with respect to the partition (R, V) is the vector $\pi^{(R,V)} = [\pi_D^{(R,V)}]_{D \subseteq \{R \cup V\}}$ where $\pi_D^{(R,V)} = P(Y_D = 1_D, Y_{R \setminus D} = 0_{R \setminus D})$. We obtain the obtain the *log-hybrid linear parameter* (La Rocca and Roverato, 2017) as

$$\psi^{(R,V)} = \mathbb{M}^T \log \pi^{(R,V)}. \tag{3.4}$$

The model specification depends on the partition adopted for $Q = (R \cup V)$. If we consider the partition $(Q, \emptyset)$, the log-hybrid parameterization $\psi^{(Q,\emptyset)}$ corresponds to the log-linear parameterization for the joint distribution of (X_R, Y_V). If we consider the partition $(R, \emptyset)$ for the set R, the log-linear parameterization $\psi^{(R,\emptyset)}$ results for the marginal distribution of X_R. So, LWF chain graph models can be specified by zero constrants on the parameters $\psi^{(Q,\emptyset)}$ and on $\psi^{(R,\emptyset)}$ for the joint distribution of (X_R, Y_V) and the marginal distribution of X_R, respectively.

The LWF chain graphical model for (X_R, Y_V) denoted by $C_C = (Q, E_U, E_D)$ is such that for any $i, j \in V$ and any $r, t \in R$,

u) if $(r, i) \notin E_D$ then $\psi^{(Q,\emptyset)}_{\{r,i\} \cup A} = 0$ for all $A \subseteq \{Q \setminus \{r, i\}\}$,

v) if $(i, j) \notin E_U$ then $\psi^{(Q,\emptyset)}_{\{i,j\} \cup A'} = 0$ for all $A' \subseteq \{Q \setminus \{i, j\}\}$,

w) if $(r, t) \notin E_U$ then $\psi^{(R,\emptyset)}_{\{r,t\} \cup A''} = 0$ for all $A'' \subseteq \{R \setminus \{r, t\}\}$.

The regression binary graphical model for (X_R, Y_V) denoted by $C_R = (Q, E_U, E_B, E_D)$ is such that for any $i, j \in V$ and any $r, t \in R$, if it hold w) and

x) if $(r,i) \notin E_D$ then $\psi^{(R,V)}_{\{r,i\}\cup A} = 0$ for all $A \subseteq R \setminus \{r\}$,

y) if $(i,j) \notin E_B$ then $\psi^{(R,V)}_{\{i,j\}\cup A'} = 0$ for all $A' \subseteq R$.

The advantage of log-hybrid is that it allows the specification of both binary LWF and regression chain graph models. Nevertheless, further parameterization are avalaible in particular for regression graph models, based on different transformation function of the probability parameter; (Drton, 2009; Marchetti and Lupparelli, 2011). Maximum likelihood estimates of the parameters in case of binary data can be obtained using a general iterative algorithm for constrained likelihood maximization provided by Lang (1996). Drton (2008) extends the iterative conditional fitting (ICF) algorithm for bidirected graphs to fit regression graphs.

3.4 Main objectives

In chapter 2, we aim to model the effect of a categorical factor on the dependence structure of a set of random response variables. In particular our interest focus on the effect of the categorical factor on the interactions among the response variables. We propose two novel classes of graphical models, termed *profile undirected and bi-directed graphical models*, which preserve the convenient aspects of a graphical approach and enhance, at the same time, the modelling prospects given by chain graphs and multiple graphs. A crucial profit in using profile graphs is that they encode in a single graph all the independencies that can be read off on both a collection of multiple graphs and a chain graph. For both the proposed graphs, we derive the Markov properties based on the same connected set rule for modelling all the *profile outcome distributions*, that is the set of all conditional probability distributions of the response variables given any level of an external risk factor. We formally establish the compatibility, in terms of independence models, of the proposed profile graphical models and certain types of chain graphs. The proposed approach is compatible with different types of chain graph models, which provide different regression frameworks for data analysis. The class of graphs we propose can be used for modelling the distributions of both continuous and discrete outcomes. We illustrate in details suitable parameterizations for these classes of models for the Gaussian case such that Markov properties can be satisfied by zero parameter constraints. We conclude the manuscript with a cancer genomics application aimed at the reconstruction of a profile graphical model, a protein network that changes with respect

to the disease subtype. In this manuscript some aspects still require further investigations. Firstly, at this stage, inference and model selection are performed by means of independent Lasso Sep-Logit approach (Meinshausen and Bühlmann, 2006) which does not account for the dependence between sub-group models. In this context, more efforts are needed for the implementation of a joint selection procedure based, for instance, of a Data-shared Lasso strategy (Ollier and Viallon, 2017). From a modelling perspective, an interesting development could be given by the generalization of the profile approach to chain graph models to explore profile dependence structures among variables grouped in chain components. This generalization is not trivial in terms of Markov property specification, since we need to consider the effect of an external factor on variables collected both within and between chain components. However, we conjecture that profile chain graph models would provide useful insights to investigate data generating processes for data which, in principle, might be different in each sub-group.

In chapter 3, we propose two Bayesian approaches for the selection of log-linear models associated to multiple Ising graphs. Following Massam et al. (2009), we devise a Bayesian exact-likelihood inference for low-dimensional binary response data, based on conjugate priors for log-linear parameters, where we implement a computational strategy that uses Laplace approximations and a Metropolis-Hastings algorithm that allows us to perform a stochastic model search. We also propose a quasi-likelihood Bayesian approach, extending the work of Bhattacharyya and Atchade (2019), for fitting high-dimensional Ising multiple graphs, where the normalization constant results computationally intractable, with spike-and-slab priors to encode sparsity and MCMC algorithms for sampling from the quasi-posterior distribution which enables variable selection and estimation simultaneously. In both methods, we define a Markov Random Field prior on the graph structures, which encourages the selection of the same edges in related graphs (Peterson et al., 2015). Simulation studies show that our methods perform better than the same ones using identical and independent Bernoulli distributions for the prior distribution of the model, as in Bhattacharyya and Atchade (2019). Performances of our methods are comparatevely better than the competing frequentist approaches Indep-Seplogit (Meinshausen and Bühlmann, 2006) and DataShared-SepLogit (Ollier and Viallon, 2017).

References

Andersson, S. A., D. Madigan, and M. D. Perlman (2001). Alternative markov properties for chain graphs. *Scandinavian journal of statistics 28*(1), 33–85.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B 36*(2), 192–225.

Bhattacharyya, A. and Y. Atchade (2019). A bayesian analysis of large discrete graphical models. *ArXiv*.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

Cox, D. R. and N. Wermuth (1996). *Multivariate Dependencies*. London: Chapman & Hall.

Dempster, A. P. (1972). Covariance selection. *Biometrics 28*(1), 157–175.

Drton, M. (2008). Iterative conditional fitting for discrete chain graph models. In *Brito P. (eds) COMPSTAT 2008*. Physica-Verlag HD.

Drton, M. (2009). Discrete chain graph models. *Bernoulli 15*(3), 736–753.

Drton, M. and T. Richardson (2008). Binary models for marginal independence. *Journal of the Royal Statistical Society: Series B 70*(2), 287–309.

Edwards, D. (2000). *Introduction to Graphical Modelling* (2nd ed.). Springer.

Frydenberg, M. (1990). The chain graph markov property. *Scandinavian Journal of Statistics*, 333–353.

Guo, J., E. Levina, G. Michailidis, and J. Zhu (2011). Joint estimation of multiple graphical models. *Biometrika 98*(1), 1–15.

Kauermann, G. (1996). On a dualization of graphical gaussian models. *Scandinavian Journal of Statistics 23*(1), 105–116.

La Rocca, L. and A. Roverato (2017). Discrete graphical models. In *M. Drton, S. L. Lauritzen, M. Maathuis & M. Wainwright, eds, Handbook of Graphical Models*, Handbooks of Modern Statistical Methods. Chapman and Hall/CRC.

Lang, J. B. (1996, 04). Maximum likelihood methods for a generalized class of log-linear models. *Ann. Statist. 24*(2), 726–752.

Lauritzen, S. L. (1996). *Graphical Models*. New York: Oxford Univ. Press.

Marchetti, G. M. and M. Lupparelli (2011, 08). Chain graph models of multivariate regression type for categorical data. *Bernoulli 17*(3), 827–844.

Massam, H., J. Liu, and A. Dobra (2009, 12). A conjugate prior for discrete hierarchical log-linear models. *Ann. Statist. 37*(6A), 3431–3467.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*(3), 1436–1462.

Ollier, E. and V. Viallon (2017, 01). Regression modelling on stratified data with the lasso. *Biometrika 104*(1), 83–96.

Pearl, J. and A. Paz (1987). *Graphoids: a graph-based logic for reasoning about relevancy relations. In B. D. Boulary, D. Hogg & L. Steel, eds,.*

Peterson, C. B., F. Stingo, and M. Vannucci (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association 110*(509), 159–174.

Ravikumar, P., M. J. Wainwright, and J. D. Lafferty (2010). High-dimensional ising model selection using l1-regularized logistic regression. *Ann. Statist. 38*(3), 1287–1319.

Richardson, T. S. (2003). Markov property for acyclic directed mixed graphs. *Scandinavian Journal of Statistics 30*(1), 145–157.

Roverato, A., M. Lupparelli, and L. La Rocca (2013). Log-mean linear models for binary data. *Biometrika 100*(2), 485–494.

Sadeghi, K. and S. Lauritzen (2014). Markov properties for mixed graphs. *Bernoulli 20*(2), 676–696.

Speed, T. P. and H. T. Kiiveri (1986). Gaussian markov distributions over finite graphs. *The Annals of Statistics 14*(1), 138–150.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B 67*(1), 91–108.

Wermuth, N. and K. Sadeghi (2012). Sequences of regressions and their independences. *TEST 21*(2), 215–252.

Chapter 2

Profile graphical models

Contents

List of Figures

List of Tables

Abstract

A class of profile graphs is introduced for modelling the effect of an external factor on the independence structure of a multivariate set of variables. This class is quite general and includes multiple graphs and chain graphs as special case. Conditional and marginal independence structures are explored by using profile undirected and bi-directed graphical models, respectively. These two families of graphical models are formally defined with their corresponding Markov properties. Furthermore, necessary conditions are derived to induce, for any profile undirected and bi-directed graph model, a compatible class of chain graph models of different type known as LWF chain graph and regression graph, respectively. An application on protein networks in various subtypes of acute myeloid leukemia is discussed.

1 Introduction

Multivariate regression models can be represented by chain graphs (Lauritzen and Wermuth, 1989; Frydenberg, 1990; Andersson et al., 2001); this representation sheds light on the conditional independence structure between a set of multiple response variables and a set of explanatory variables, all represented by vertices of the chain graph. In its simplest form, response and explanatory variables are grouped in two different chain components or blocks. Missing edges between vertices correspond to conditional independencies for the joint distribution of variables under suitable Markov properties specified for a class of chain graphs. There are several types of chain graph models (Drton, 2009), in this paper we will focus on two types which correspond to smooth statistical models, (i) LWF chain graph (Frydenberg, 1990) and (ii) regression graph models (Wermuth and Sadeghi, 2012). Under the corresponding Markov properties, these two classes of chain graphs provide a different multivariate regression framework for modelling the independence structure of the outcomes given the explanatory variables. Model selection of chain graphs is an active area of research, see recent approaches based on penalized likelihood by Rothman et al. (2010), Yin and Li (2011) and Lee and Liu (2012), two-step approaches by Cai et al. (2012) and Chen et al. (2016), and Bayesian approaches by Bhadra and Mallick (2013) and Consonni et al. (2017). Despite recent advancements, the use of chain graph models might be limited in some contexts since relevant aspects could not be totally addressed only through conditional independencies. Principally, our interest is on the effect that an explanatory variable may have on the joint probability distribution of the outcomes rather on each single outcome. The matter is that the independence model given by any chain graph under its own Markov properties does not provide a comprehensive information about the role that an explanatory variable, hereafter termed *external factor*, has on the pairwise independence between response variables and on their joint independence structure. Basically, we think chain graphs will not suffice whenever, beyond the conditional independence model represented through a set of missing edges, we want to say something more about the not missing edges and, in this context, about the effect the factor has on the edges in the response chain component. This issue has been widely discussed in the literature and there are some extreme cases which show how the interaction between variables may considerably change under different levels of an external factor. Notable instances are given by the effect reversal

(Cox and Wermuth, 2003) and also by the well-known Simpson Paradox (Simpson, 1951). This same problem can be also tackled from a different perspective, as external factors can be used to define subgroups and subpopulations. In recent years this approach has lead to the development of methods for multiple graphical models for Gaussian random variables (Guo et al., 2011; Danaher et al., 2014; Peterson et al., 2015). Similarly, in the context of multinomial sampling models, there have been few proposals of graphical models for context-specific independencies; these approaches allow conditional independences to hold only for a subset of the sample space of one, or more, variables we condition upon. (Hojsgaard, 2003; Corander, 2003; Nyman et al., 2014, 2016). Approaches for multiple graphical models do not directly include the external factor into the model, and, more importantly, approaches for multiple and context-specific graphical models do not fully account for the effects of external or internal factors on the variables included in the graphical model – for example, in Corander (2003) and Nyman et al. (2014) context-specific independencies vary with respect only to adjacent veetices. Given the well-established use of chain graphs for modelling multivariate regression framework and the recent developments of multiple and context-specific graphical models, our goal is four-fold: first, we propose two novel classes of graphical models, termed *profile undirected and bi-directed graphs*, which preserve the convenient aspects of a graphical approach and enhance, at the same time, the modelling prospects given by chain graphs and multiple graphs; second, we derive the Markov properties for both families of profile graphs based on the same connected set rule for modelling all the *profile outcome distributions*, that is the set of all conditional probability distributions of the response variables given any level of an external risk factor; third, we formally establish the compatibility, in terms of independence models, of the proposed profile graphical models and certain types of chain graphs; finally, we illustrate suitable parameterizations for these classes of models for the Gaussian case such that Markov properties can be satisfied by zero parameter constraints. We conclude the manuscript with a cancer genomics application aimed at the reconstruction of a profile graphical model, a protein network that changes with respect to the disease subtype.

2 Theoretical framework

2.1 Basic setup

Let $G = (V, E)$ be a graph defined by a set of vertices $a \in V$ and a set of edges $(a, b) \in E$ joining pairs of vertices $a, b \in V$, and let $Y_V = (Y_a)_{a \in V}$ be a random vector of variables indexed by the finite set V with $p = |V|$. A graph, associated to a random vector Y_V, is generally used to represent conditional independence structures under suitable Markov properties. Typically, missing edges in G correspond to conditional independencies for the joint distribution of Y_V. Also, let us consider the random categorical variable X representing an external factor with respect to (in the sequel, wrt) the random vector Y_V of outcomes/response variables. The variable X takes level $x \in \mathcal{X}$, with $|\mathcal{X}| = q$. Our interest lies in the effect of X on the joint independence structure of Y_V and, in particular, the interest is exploring via a graphical modelling approach how this structure may change under different levels $x \in \mathcal{X}$, which we call *profiles*. Chain graphs are generally used for modelling the effect of background variables on joint response variables. In the simplest form, a two-block chain graph $C = [\{C_1, C_2\}, E]$ is defined by a set of vertices partitioned in chain components C_1 and C_2, and a set of edges E. Depending on the set of Markov properties specified for the chain graph we may have different independence models for the joint distribution of random vectors $(Y_{C_t})_{t \in \{1,2\}}$, associated to the chain components $\{C_t\}_{t \in \{1,2\}}$. In particular we focus on (i) the class of LWF chain graph models (Frydenberg, 1990) and on (ii) the class of regression graph models (Wermuth and Cox, 2004). Both models correspond to multivariate regression models with suitable independence constraints corresponding by missing edges, both within and between chain components. Any pair of vertices $a, b \in C_t$ within the same chain component with $t = 1, 2$ and $a \neq b$, can be joined by undirected or by bi-directed edges, respectively for chain graph of type (i) and (ii); vertices between chain components, $a \in C_1$ and $b \in C_2$, are joined by directed edges preserving the same direction such that cycles are not allowed. For our purpose, the set of vertices C_1 and C_2 are associated, respectively, to the random vector Y_V of response variables and to the background variable X, so that $C_1 = V$ and $|C_2| = 1$. In principle, the chain component C_2 may include a multiple categorical random vector; in this case X represents a random variabile with state space given by the combination of a multiple factor levels. For any $x \in \mathcal{X}$, let $Y_V(x)$ be a *x-profile outcome vector*, that is the random vector $Y_V|\{X = x\}$

conditioned on a specific profile x of the factor X, and let $P(Y_V(x))$ be the corresponding *x-profile probability distribution* of $Y_V(x)$, that is the conditional probability distribution $P(Y_V|\{X = x\})$. Note that $P(\cdot)$ can be a probability density function or a probability mass function, depending on the continuous or discrete nature of the multivariate random variable $Y_V(x)$, with $x \in \mathcal{X}$. For the sake of simplicity, in the sequel we omit the prefix x to denote both the profile outcome vector and the profile outcome distribution. Then, for a given multivariate random vector Y_V and an external factor X, let $Y_{V|\mathcal{X}} = [Y_V(x)]_{x\in\mathcal{X}}$ be the finite set of all profile outcome vectors and let $P(Y_{V|\mathcal{X}}) = [P(Y_V(x))]_{x\in\mathcal{X}}$ be the corresponding set of all profile outcome distributions. For any $A \subseteq V$, $Y_{A|\mathcal{X}} = [Y_A(x)]_{x\in\mathcal{X}}$ is set of marginal profile outcome vectors with corresponding profile probability distributions $P(Y_{A|\mathcal{X}}) = [P(Y_A(x))]_{x\in\mathcal{X}}$. Given a partition $A, B, C \subseteq V$, the *profile conditional independence* $Y_A(x) \perp\!\!\!\perp Y_B(x)|Y_C(x)$ corresponds to the factorization

$$P[Y_A(x), Y_B(x)|Y_C(x)] = P[Y_A(x)|Y_C(x)] \times P[Y_B(x)|Y_C(x)], \quad x \in \mathcal{X}, \tag{2.1}$$

of the joint profile distribution $Y_V(x)$. Similarly, we say that the *profile marginal independence* $Y_A(x) \perp\!\!\!\perp Y_B(x)$ corresponds to the factorization

$$P[Y_A(x), Y_B(x)] = P[Y_A(x)] \times P[Y_B(x)], \quad x \in \mathcal{X}. \tag{2.2}$$

If the profile independence statements in Equations (2.1) and (2.2) hold for any level $x \in \mathcal{X}$, then these equations imply that $Y_A \perp\!\!\!\perp Y_B|\{Y_C, X\}$ and $Y_A \perp\!\!\!\perp Y_B|X$, respectively. Finally, let us consider a collection of *multiple graphs* $G_{V|\mathcal{X}} = \{G(x) = (V, E(x))\}_{x\in\mathcal{X}}$ associated to the profile outcome distributions $P(Y_{V|\mathcal{X}})$. Under suitable Markov properties, any graph $G(x)$ represents an independence model for the profile outcome vector $Y_V(x)$, for any $x \in \mathcal{X}$. In particular, missing edges wrt $G(x)$ correspond to profile conditional independencies for the joint distribution of $Y_V(x)$, with $x \in \mathcal{X}$. Graphs $G(x) \in G_{V|\mathcal{X}}$ may have different skeletons.

We remark that chain graph models do not allow to explore how the independence structure of Y_V may considerably vary for any profile $x \in \mathcal{X}$. Multiple graphs do not allow to model the effect of X on each outcome $Y_a \in Y_V$. In essence, the idea is to provide a single graph able to embed, at the same time, information about the profile independence structure for any $Y_V(x) \in Y_{V|\mathcal{X}}$ and about the conditional independence between X and any outcome $Y_a \in Y_V$.

2.2 Profile graphs

We introduce the class of profile graphs. A profile graph $\mathcal{G} = (V, \mathcal{E})$ is defined by the set V of vertices and a set of $\mathcal{Z}$-labelled edges $\mathcal{E}$ which are labelled according to a subset $\mathcal{Z} \subseteq \mathcal{X}$. Let $(a, b)^{\mathcal{Z}}$ be the generic element of $\mathcal{E}$ associated to any pair $a, b \in V$, where the presence or absence of the edge between a and b is determined by the subset $\mathcal{Z}$ of the state space $\mathcal{X}$. For each pair $a, b \in V$, the corresponding edge $(a, b)^{\mathcal{Z}} \in \mathcal{E}$ will belong to one of the following three categories: (i) if $\mathcal{Z} = \mathcal{X}$, vertices a and b are not joined by any edge, (ii) if $\mathcal{Z}$ is a nonempty proper subset of $\mathcal{X}$, $\mathcal{Z} \subset \mathcal{X}$ and $\mathcal{Z} \neq \emptyset$, vertices a and b are joined by a dotted $\mathcal{Z}$-labelled edge; (iii) if $\mathcal{Z} = \emptyset$, vertices a and b are joined by a full edge and, for sake of simplicity, the $\emptyset$-label is not displayed in the graph. Under suitable Markov properties, the profile graph $\mathcal{G}$ provides an independence model for the joint distributions of a random vector $Y_{V|\mathcal{X}}$ of profile outcomes. In particular, a missing edge in $\mathcal{G}$ corresponds to a profile conditional independence for each profile $x \in \mathcal{X}$. A $\mathcal{Z}$-labelled dotted edge in $\mathcal{G}$ corresponds to profile conditional independencies holding only for the profiles $x \in \mathcal{Z}$, with $\mathcal{Z} \subset \mathcal{X}$ and $\mathcal{Z} \neq \emptyset$.

Further technical definitions are given. For any couple of vertices $a, b \in V$, we say that b is an *x-neighbour* of a and *vice versa*, if they are joined by a $\mathcal{Z}$-labelled edge such that $x \notin \mathcal{Z}$, with $\mathcal{Z} \subset \mathcal{X}$. Let $nb_x(a)$ be the set of all x-neighbours of a, with $a \in V$ and $x \in \mathcal{X}$. For any pair $a, b \in V$ and $x \in \mathcal{X}$, an *x-path* between a and b is given by a sequence of $(a, b)^{\mathcal{Z}}$ edges, for any $\mathcal{Z} \subset \mathcal{X}$, such that $x \notin \mathcal{Z}$ for all edges in the sequence. Given any nonempty subset C of V, C is said to be *x-connected* if any pair $a, b \in C$ is joined by a x-path, with $x \in \mathcal{X}$. Any nonempty subset D of V is said to be *x-disconnected* if it is not x-connected, with $x \in \mathcal{X}$ and let $K_1, \ldots, K_r$ be the x-connected components of D. For any triple A, B, C of disjoint subsets of V and $x \in \mathcal{X}$, we say that C *x-separates* A from B if every x-path from any vertex $a \in A$ to any vertex $b \in B$ intersects C. Technical x-definitions above can be simply extended to $\mathcal{Z}$-definitions for any subset $\mathcal{Z}$ of $\mathcal{X}$ if they hold for all $x \in \mathcal{Z}$.

3 Profile graphical models: Markov properties

3.1 Profile undirected graphical models

In this section we consider a special case of profile graph, named profile undirected graph $\mathcal{G}_{\mathcal{U}} = (V, \mathcal{E}_{\mathcal{U}})$: for any $a, b \in V$, edges $(a, b)^{\mathcal{Z}} \in \mathcal{E}_{\mathcal{U}}$ are drawn either as $\mathcal{Z}$-labelled dotted

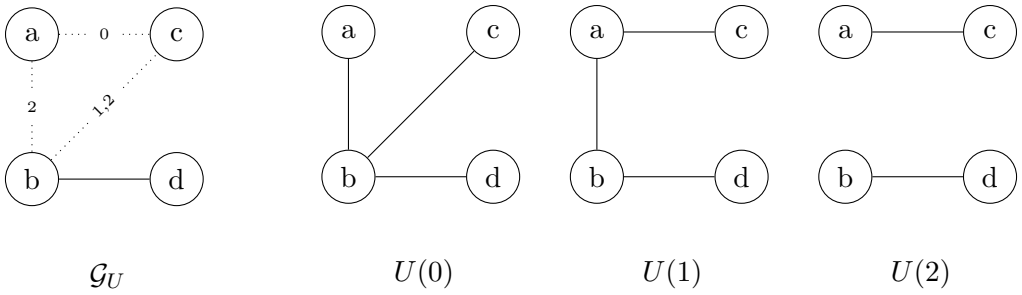


Figure 1: Given $V = \{a, b, c, d\}$, $\mathcal{G}_{\mathcal{U}}$ is a profile undirected graph for the profile outcome vectors $Y_{V|\mathcal{X}} = [(Y_V(x))_{x\in\mathcal{X}}]$ with $\mathcal{X} = \{0, 1, 2\}$. Any $U(x)$ is the induced undirected graph for the profile outcome vector $Y_V(x)$, with $x \in \mathcal{X}$.

undirected edges if $\mathcal{Z} \subset \mathcal{X}$ and $\mathcal{Z} \neq \emptyset$, or as full undirected edges if $\mathcal{Z} = \emptyset$; if $\mathcal{Z} = \mathcal{X}$, vertices a and b are disjoined. Consider for instance the profile undirected graph $\mathcal{G}_{\mathcal{U}}$ in the left panel of Figure 1: there are four vertices $V = (a, b, c, d)$, and the graph is defined by three $\mathcal{Z}$-labelled dotted edges $\{(a, b)^2, (a, c)^0, (b, c)^{\{1,2\}}\}$, one full edge $(b, d)^\emptyset$ and two missing edges $\{(a, d)^{\mathcal{X}}, (c, d)^{\mathcal{X}}\}$. Technical definitions given in Section 2.2 are illustrated in the following example.

Example 3.1. *Consider the profile undirected graph $\mathcal{G}_U$ in the left panel of Figure 1. Vertices a and c are both $\{1, 2\}$-neighbours, because they are joined by a dotted edge with label $\mathcal{Z} = \{0\}$ that does not contain neither 1 or 2. Vertices b and d are $\mathcal{X}$-neighbours because they are joined by a full edge. The sequence of edges $\{(a, c)^{\{0\}}, (a, b)^{\{2\}}, (b, d)^\emptyset\}$ is a $\{1\}$-path, since 1 is not included in any label of the edges in the sequence. The same sequence is not a $\{2\}$-path since the label of the couple (a, b) contains 2. The set V is $\{1\}$-connected, because every pair of vertices in V are joined by a $\{1\}$-path. The same set is $\{2\}$-disconnected with $\{2\}$-connected components $\{a, c\}$ and $\{b, d\}$, because there is no $\{2\}$-path between a and b. Vertices c and d are $\{1\}$-separated by a because the only $\{1\}$-path $\{(a, c)^{\{0\}}, (a, b)^{\{2\}}, (b, d)\}$ between c and d intersects a; vertex a does not $\{0\}$-separates c and d because there exists the $\{0\}$-path $\{(b, c)^{\{1,2\}}, (b, d)^\emptyset\}$ between them that does not intersect a.*

The probability distributions $P[Y_{V|\mathcal{X}}]$ of the profile outcome vectors $Y_{V|\mathcal{X}}$ satisfy the *profile undirected Pairwise Markov Property* ($\mathcal{U}$-PMP) wrt the graph $\mathcal{G}_{\mathcal{U}} = (V, \mathcal{E}_{\mathcal{U}})$ if, for any $(a, b)^{\mathcal{Z}} \in \mathcal{E}_{\mathcal{U}}$ with $\mathcal{Z} \subseteq \mathcal{X}$,

$$Y_a(x) \perp\!\!\!\perp Y_b(x)|Y_{V\setminus\{a,b\}}(x), \qquad x \in \mathcal{Z}. \tag{3.1}$$

The probability distributions $P[Y_{V|\mathcal{X}}]$ of the profile outcome vectors $Y_{V|\mathcal{X}}$ satisfy the *profile undirected Global Markov Property* ($\mathcal{U}$-GMP) wrt the graph $\mathcal{G}_{\mathcal{U}} = (V, \mathcal{E}_{\mathcal{U}})$ if, for any triple A, B, C of disjoint subsets of V such that C x-separates A from B in $\mathcal{G}_{\mathcal{U}}$,

$$Y_A(x) \perp\!\!\!\perp Y_B(x)|Y_C(x), \qquad x \in \mathcal{X}. \tag{3.2}$$

The probability distributions $P[Y_{V|\mathcal{X}}]$ of the profile outcome vectors $Y_{V|\mathcal{X}}$ satisfy the *profile undirected Connected Set Markov Property* ($\mathcal{U}$-CSMP) wrt the graph $\mathcal{G}_{\mathcal{U}} = (V, \mathcal{E}_{\mathcal{U}})$ if, for any x-disconnected set D of V, with $K_1, \ldots, K_r$ x-connected components of D,

$$Y_{K_1}(x) \perp\!\!\!\perp \ldots \perp\!\!\!\perp Y_{K_r}(x)|Y_{V\setminus D}(x), \qquad x \in \mathcal{X}. \tag{3.3}$$

Example 3.2. *Consider the left panel including the graph $\mathcal{G}_{\mathcal{U}}$ in Figure 1. If $P[Y_{V|\mathcal{X}}]$ satisfy the $\mathcal{U}$-PMP wrt $\mathcal{G}_{\mathcal{U}}$ then $Y_b(x) \perp\!\!\!\perp Y_c(x)|\{Y_a(x), Y_d(x)\}$ for $x \in \{1, 2\}$, since $(b, c)^{\{1,2\}} \in \mathcal{E}_{\mathcal{U}}$. $P[Y_{V|\mathcal{X}}]$ satisfy the $\mathcal{U}$-GMP wrt $\mathcal{G}_{\mathcal{U}}$ if $Y_c(1) \perp\!\!\!\perp \{Y_b(1), Y_d(1)\}|Y_a(1)$ because a $\{1\}$-separates c from $\{b, d\}$. Consider the subset $D = \{a, b, c\}$ of V; $P[Y_{V|\mathcal{X}}]$ satisfy the $\mathcal{U}$-CSMP wrt $\mathcal{G}_{\mathcal{U}}$ if $\{Y_a(2), Y_c(2)\} \perp\!\!\!\perp Y_b(2)|Y_d(2)$ because D is $\{2\}$-disconnected set with two $\{2\}$-connected components $\{a, c\}$ and b.*

We prove that all the independence statements encoded in a profile undirected graph under the global Markov property can be derived by applying the connected set rule.

Theorem 3.1. *Let $\mathcal{G}_{\mathcal{U}} = (V, \mathcal{E}_{\mathcal{U}})$ be a profile undirected graph model associated to the profile outcome vectors $Y_{V|\mathcal{X}}$ with probability distributions $P[Y_{V|\mathcal{X}}]$. The $\mathcal{U}$-GMP is satisfied if and only if the $\mathcal{U}$-CSMP is satisfied wrt $\mathcal{G}_{\mathcal{U}}$.*

The proof of Theorem 1 is given in the Appendix, along with all other proofs. The local Markov property for profile undirected graph is also included in the Appendix.

Given a profile undirected graph $\mathcal{G}_{\mathcal{U}} = (V, \mathcal{E}_{\mathcal{U}})$ for the profile outcome vectors $Y_{V|\mathcal{X}}$, the corresponding class of multiple undirected graphs associated to each random vector $Y_V(x) \in Y_{V|\mathcal{X}}$ can be defined.

Definition 3.1. *Given a profile undirected graph $\mathcal{G}_{\mathcal{U}} = (V, \mathcal{E}_{\mathcal{U}})$ for the profile outcome vectors $Y_{V|\mathcal{X}}$, let $U_{V|\mathcal{X}} = \{U(x) = (V, E_U(x))\}_{x\in\mathcal{X}}$ be the induced class of multiple undirected graphs, where, for any $U(x) \in U_{V|\mathcal{X}}$, the couple $a, b \in V$ is joined by an undirected edge if $x \notin \mathcal{Z}$ in the corresponding edge $(a, b)^{\mathcal{Z}} \in \mathcal{E}_{\mathcal{U}}$, with $\mathcal{Z} \subseteq \mathcal{X}$.*

Then, a missing edge in $\mathcal{G}_{\mathcal{U}}$ corresponds to a missing edge in $U(x)$, for any $x \in \mathcal{X}$; a $\mathcal{Z}$-labelled dotted edge in $\mathcal{G}_{\mathcal{U}}$ corresponds to a missing edge in $U(x)$ if $x \in \mathcal{Z}$, and to a full edge in $U(x)$ if $x \notin \mathcal{Z}$; a full edge in $\mathcal{G}_{\mathcal{U}}$ corresponds to a full edge in $U(x)$, for any $x \in \mathcal{X}$.

Example 3.3. *Consider Figure 1. Given the profile undirected graph $\mathcal{G}_{\mathcal{U}}$, let $U_{V|\mathcal{X}} = \{U(0), U(1), U(2)\}$ be the induced class of multiple undirected graphs. The couple a, d is disjoined in $\mathcal{G}_{\mathcal{U}}$ and in any $U(x) \in U_{V|\mathcal{X}}$. The couple b, c is joined by a $\{1, 2\}$-labelled dotted edge in $\mathcal{G}_{\mathcal{U}}$ then is joined by a full edge in $U(0)$ and is disjoined in $U(1), U(2)$. The couple b, d is joined by a full edge in $\mathcal{G}_{\mathcal{U}}$ and in any $U(x) \in U_{V|\mathcal{X}}$.*

Pairwise, local, and global Markov property of probability distributions associated to undirected graphs are well known (Lauritzen, 1996). The following corollary, derived directly from Theorem 3.1, shows that the full set of conditional independencies implied by the global Markov property for any undirected graph can be also derived by applying the connected set rule.

Corollary 3.1. *Given an undirected graph model $U(x) = (V, E(x))$ associated to the profile outcome vectors $Y_{V|\mathcal{X}}$, the probability distributions $P[Y_V(x)]$ satisfy the global Markov property wrt $U(x)$ if and only if the connected set Markov property is satisfied for every x-disconnected set $D \subseteq V$, with $x \in \mathcal{X}$.*

The following proposition shows that the full set of independencies encoded in the induced undirected graph model for any $Y_V(x) \in Y_{V|\mathcal{X}}$ can be derived from the profile undirected graph model for the joint distributions of $Y_{V|\mathcal{X}}$.

Proposition 3.1. *Consider a profile undirected graph $\mathcal{G}_{\mathcal{U}} = (V, \mathcal{E}_{\mathcal{U}})$ associated to the profile outcome vectors $Y_{V|\mathcal{X}}$ and the induced class of multiple undirected graphs $U_{V|\mathcal{X}}$. If the probability distributions $P[Y_{V|\mathcal{X}}]$ satisfy the $\mathcal{U}$-CSMP wrt $\mathcal{G}_{\mathcal{U}}$, the probability distribution $P[Y_V(x)]$ of each profile vector $Y_V(x) \in Y_{V|\mathcal{X}}$ satisfies the global Markov property wrt the induced undirected graph $U(x) \in U_{V|\mathcal{X}}$.*

In the following proposition we show that $\mathcal{U}$-GMP, $\mathcal{U}$-CSMP and $\mathcal{U}$-PMP are equivalent for the class of profile undirected graph models in case of strictly positive probability distributions. This result directly derives from Proposition 3.1.

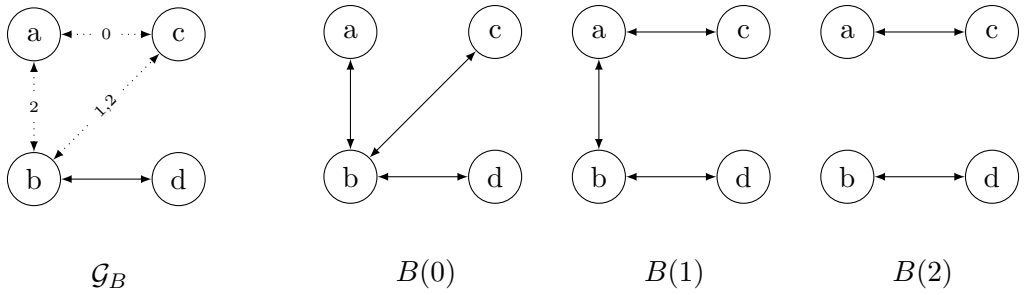


Figure 2: Given $V = \{a, b, c, d\}$, $\mathcal{G}_{\mathcal{B}}$ is a profile bi-directed graph for the profile outcome vectors $Y_{V|\mathcal{X}} = (Y_V(x))_{x\in\mathcal{X}}$ with $\mathcal{X} = \{0, 1, 2\}$. Any $B(x)$ is the induced bi-directed graph for the profile outcome vector $Y_V(x)$, with $x \in \mathcal{X}$.

Proposition 3.2. *Let $\mathcal{G}_{\mathcal{U}} = (V, \mathcal{E}_{\mathcal{U}})$ be a profile undirected graph associated to the profile outcome vectors $Y_{V|\mathcal{X}}$ with strictly positive probability distributions $P[Y_{V|\mathcal{X}}]$. The $\mathcal{U}$-GMP is satisfied if and only if the $\mathcal{U}$-PMP is satisfied wrt $\mathcal{G}_{\mathcal{U}}$.*

3.2 Profile bi-directed graphical models

In this section we consider a class of graphical models for profile marginal independencies based on the family of profile bi-directed graphs. A profile bi-directed graph $\mathcal{G}_{\mathcal{B}} = (V, \mathcal{E}_{\mathcal{B}})$ is defined by a set V of vertices and by a set of edges $\mathcal{E}_{\mathcal{B}}$ with generic element $(a, b)^{\mathcal{Z}}$, for any pair $a, b \in V$, with $\mathcal{Z} \subseteq \mathcal{X}$. For a given $(a, b)^{\mathcal{Z}} \in \mathcal{E}_{\mathcal{B}}$, if $\mathcal{Z} = \mathcal{X}$, we have that $a, b \in V$ are disjoint vertices; vertices $a, b \in V$ are joined by a bi-directed edge, drawn as $\mathcal{Z}$-labelled dotted bi-directed edge if $\mathcal{Z} \subset \mathcal{X}$ and $\mathcal{Z} \neq \emptyset$, and as full bi-directed edge if $\mathcal{Z} = \emptyset$ where the $\emptyset$-label is not displayed on the edge of the graph. For instance, consider the profile bi-directed graph in the left panel of Figure 2: the graph includes four vertices $V = (a, b, c, d)$ and it is defined by three $\mathcal{Z}$-labelled dotted edges $\{(a, b)^2, (a, c)^0, (b, c)^{\{1,2\}}\}$, one full bi-directed edge $(b, d)^{\emptyset}$ and two missing edges $\{(a, d)^{\mathcal{X}}, (c, d)^{\mathcal{X}}\}$.

The probability distributions $P[Y_{V|\mathcal{X}}]$ of the profile outcome vectors $Y_{V|\mathcal{X}}$ satisfy the *profile bi-directed Pairwise Markov Property* ($\mathcal{B}$-PMP) wrt the graph $\mathcal{G}_{\mathcal{B}} = (V, \mathcal{E}_{\mathcal{B}})$ if, for any $(a, b)^{\mathcal{Z}} \in \mathcal{E}_{\mathcal{B}}$ with $\mathcal{Z} \subseteq \mathcal{X}$,

$$Y_a(x) \perp\!\!\!\perp Y_b(x), \qquad x \in \mathcal{Z}. \tag{3.4}$$

The probability distributions $P[Y_{V|\mathcal{X}}]$ of the profile outcome vectors $Y_{V|\mathcal{X}}$ satisfy the *profile bi-directed Global Markov Property* ($\mathcal{B}$-GMP) wrt the graph $\mathcal{G}_{\mathcal{B}} = (V, \mathcal{E}_{\mathcal{B}})$ if, for any

triple A, B, C of disjoint subsets of V such that C x-separates A from B in $\mathcal{G}_{\mathcal{B}}$,

$$Y_A(x) \perp\!\!\!\perp Y_B(x) | Y_{V \setminus \{A \cup B \cup C\}}(x), \qquad x \in \mathcal{X}. \tag{3.5}$$

The probability distributions $P[Y_{V|\mathcal{X}}]$ of the profile outcome vectors $Y_{V|\mathcal{X}}$ satisfy the *profile bi-directed Connected Set Markov Property* ($\mathcal{B}$-CSMP) wrt the graph $\mathcal{G}_{\mathcal{B}} = (V, \mathcal{E}_{\mathcal{B}})$ if, for any x-disconnected set D of V, with $K_1, \ldots, K_r$ x-connected components of D,

$$Y_{K_1}(x) \perp\!\!\!\perp \ldots \perp\!\!\!\perp Y_{K_r}(x), \qquad x \in \mathcal{X}. \tag{3.6}$$

Example 3.4. *Consider the left panel profile bi-directed graph model $\mathcal{G}_{\mathcal{B}}$ in Figure 2. If the profile probability distributions in $P[Y_{V|\mathcal{X}}]$ satisfy the $\mathcal{B}$-PMP wrt $\mathcal{G}_{\mathcal{B}}$ then $Y_b(x) \perp\!\!\!\perp Y_c(x)$ for $x \in \{1, 2\}$. If $P[Y_{V|\mathcal{X}}]$ satisfy the $\mathcal{B}$-GMP wrt $\mathcal{G}_{\mathcal{B}}$ then $Y_c(x) \perp\!\!\!\perp \{Y_b(x), Y_d(x)\}$, with $x \in \{1, 2\}$, since a $\{1, 2\}$-separates c from $\{b, d\}$. Consider the subset $D = \{a, b, c\}$ of V; if $P[Y_{V|\mathcal{X}}]$ satisfy the $\mathcal{B}$-CSMP wrt $\mathcal{G}_{\mathcal{B}}$ then $\{Y_a(2), Y_c(2)\} \perp\!\!\!\perp Y_b(2)$ because D is $\{2\}$-disconnected set with $\{2\}$-connected components $\{a, c\}$ and b.*

Building upon Drton and Richardson (2008), the following proposition shows that the global and connected set Markov properties for profile bi-directed graphs are equivalent.

Proposition 3.3. *Let $\mathcal{G}_{\mathcal{B}} = (V, \mathcal{E}_{\mathcal{B}})$ be a profile bi-directed graph model associated to the profile outcome vectors $Y_{V|\mathcal{X}}$ with probability distributions $P[Y_{V|\mathcal{X}}]$. The $\mathcal{B}$-GMP is satisfied if and only if the $\mathcal{B}$-CSMP is satisfied wrt $\mathcal{G}_{\mathcal{B}}$.*

Unlike profile undirected graphs, we remark that for profile bi-directed graph the $\mathcal{B}$-PMP does not necessarily imply the $\mathcal{B}$-GMP.

Proposition 3.4. *Let $\mathcal{G}_{\mathcal{B}} = (V, \mathcal{E}_{\mathcal{B}})$ be a profile bi-directed graph model associated to the profile outcome vectors $Y_{V|\mathcal{X}}$ with probability distributions $P[Y_{V|\mathcal{X}}]$. The $\mathcal{B}$-PMP is satisfied if the $\mathcal{B}$-GMP is satisfied wrt $\mathcal{G}_{\mathcal{B}}$.*

Given a profile bi-directed graph $\mathcal{G}_{\mathcal{B}} = (V, \mathcal{E}_{\mathcal{B}})$ for the profile outcome vectors $Y_{V|\mathcal{X}}$, we define the class of multiple bi-directed graphs associated to each $Y_V(x) \in Y_{V|\mathcal{X}}$.

Definition 3.2. *Given a profile bi-directed graph $\mathcal{G}_{\mathcal{B}} = (V, \mathcal{E}_{\mathcal{B}})$ for the profile outcome vectors $Y_{V|\mathcal{X}}$, let $B_{V|\mathcal{X}} = \{B(x) = (V, E_B(x))\}_{x \in \mathcal{X}}$ be the induced class of multiple bi-directed graphs, where, for any $B(x) \in B_{V|\mathcal{X}}$, the couple $a, b \in V$ is joined by a bi-directed edge if $x \notin \mathcal{Z}$ in the corresponding $(a, b)^{\mathcal{Z}} \in \mathcal{E}_{\mathcal{B}}$, with $\mathcal{Z} \subseteq \mathcal{X}$.*

Then, a missing edge in $\mathcal{G}_{\mathcal{B}}$ corresponds to a missing edge in $B(x)$, for any $x \in \mathcal{X}$; a $\mathcal{Z}$-labelled dotted edge in $\mathcal{G}_{\mathcal{B}}$ corresponds to a bi-directed edge in $B(x)$ if $x \notin \mathcal{Z}$ or to a missing edge in $B(x)$ if $x \in \mathcal{Z}$; a full bi-directed edge in $\mathcal{G}_{\mathcal{B}}$ corresponds to a bi-directed edge in any $B(x) \in B_{V|\mathcal{X}}$.

Example 3.5. *Consider in Figure 2 the induced class $B_{V|\mathcal{X}} = \{B(0), B(1), B(2)\}$ of multiple bi-directed graphs. There is no edge between a and d in $\mathcal{G}_{\mathcal{B}}$, then $(a, d) \notin E_B(x)$ for any $x \in \mathcal{X}$. Vertices b and c are joined by a $\{1, 2\}$-labelled dotted edge in $\mathcal{G}_{\mathcal{B}}$, then $(b, c) \notin E_B(x)$ for $x \in \{1, 2\}$ and $(b, c) \in E_B(0)$. There is a full edge between b and d in $\mathcal{G}_{\mathcal{B}}$ and in any $B(x) \in B_{V|\mathcal{X}}$.*

The proposition below shows that, from a profile bi-directed graph model for the joint distributions of $Y_{V|\mathcal{X}}$, it can be derived the induced bi-directed graph model for any $Y_V(x) \in Y_{V|\mathcal{X}}$.

Proposition 3.5. *Consider a profile bi-directed graph $\mathcal{G}_{\mathcal{B}} = (V, \mathcal{E}_{\mathcal{B}})$ associated to the profile outcome vectors $Y_{V|\mathcal{X}}$ and the induced class of multiple bi-directed graph $\mathcal{B}_{V|\mathcal{X}}$. If the probability distributions $P[Y_{V|\mathcal{X}}]$ satisfy the $\mathcal{B}$-CSMP wrt $\mathcal{G}_{\mathcal{B}}$, the probability distribution $P[Y_V(x)]$ of each profile vector $Y_V(x) \in Y_V(x)$ satisfies the connected set Markov property wrt the induced bi-directed graph $B(x) \in \mathcal{B}_{V|\mathcal{X}}$.*

4 Chain graph compatibility

4.1 Profile undirected graphs and LWF chain graphs

For any profile undirected graph $\mathcal{G}_{\mathcal{U}}$, we derive an induced class of two-block LWF chain graphs $\mathcal{C}_U = \{C_U\}$, with generic element $C_U = [\{V, X\}, E_{C_U}]$. We show that any profile undirected graph model for the set of profile distributions $P[Y_{V|\mathcal{X}}]$ is compatible, in terms of independence models, with a class of LWF chain graph models for the joint distribution $P(Y_V, X)$. Compatibility is formally defined within Theorem 4.1; in summary, we say that an LWF chain graph C_U is compatible with a profile undirected graph $\mathcal{G}_{\mathcal{U}}$ if all independencies between variables in its response component can be read from the profile undirected graph.

A joint probability distribution $P(Y_V, X)$ satisfies the LWF Global Markov property (LWF-GMP) wrt the LWF chain graph $C_U = [\{V, X\}, E_{C_U}]$ if (Frydenberg, 1990; Drton,

2009):

for any disconnected set $D \subseteq V$ with connected components $K_1, \ldots, K_r$,

$$Y_{K_1} \perp\!\!\!\perp \ldots \perp\!\!\!\perp Y_{K_r} | \{Y_{V \setminus D}, X\}; \tag{4.1}$$

for any subset $A \subseteq V$ such that there is a missing arrow between any vertex $a \in A$ and X,

$$Y_A \perp\!\!\!\perp X | Y_{V \setminus A} \tag{4.2}$$

We remark that Equation (4.1) directly derives from Theorem 3.1.

Definition 4.1. *Given a profile undirected graph $\mathcal{G}_{\mathcal{U}} = (V, \mathcal{E}_{\mathcal{U}})$ associated to the profile outcome vectors $Y_{V|\mathcal{X}}$, let $\mathcal{C}_U$ be the induced class of two-block LWF chain graphs where a graph $C_U = [\{V, X\}, E_{C_U}]$ belongs to $\mathcal{C}_U$ if*

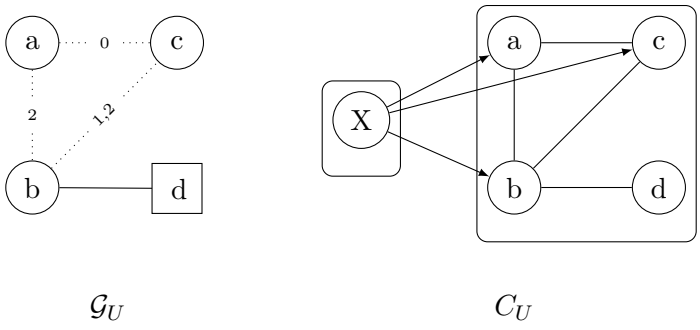
(i) any couple $a, b \in V$ is joined by an undirected edge in C_U if $\mathcal{Z} \subset \mathcal{X}$ for the pair $(a, b)^{\mathcal{Z}} \in \mathcal{E}_{\mathcal{U}}$;

(ii) for any couple $a, b \in V$, a and b are both reached by an arrow in C_U starting from X if $\mathcal{Z} \subset \mathcal{X}$ and $\mathcal{Z} \neq \emptyset$ for the pair $(a, b)^{\mathcal{Z}} \in \mathcal{E}_{\mathcal{U}}$.

Necessary conditions (i) and (ii) in Definition 4.1 ensure that it will always exist at least one compatible LWF chain graph for any given profile undirected graph. Condition (i) is related to the missing/non-missing undirected edges for any induced chain graph; it states that dotted and full edges in profile undirected graphs correspond to full edges in chain graphs. Condition (ii) is related to missing/non-missing directed edges for any induced chain graph; it states that vertices joined by a dotted edge in a profile undirected graph cannot be disjoined from X in the induced chain graph. Since condition (ii) may not be intuitive, the following counterexample shows that this is a necessary condition.

Example 4.1. *Let $V = \{a, b, c\}$ be a set of response variables and X a factor with state-space $\mathcal{X} = \{0, 1\}$. Consider a chain graph $\mathcal{C}_U = \{(V, X), E_{C_U}\}$ with $E_{C_U} = \{(a, b), (b, c), (X, c)\}$, where vertices a and b are both disjoined from X. For the condition (4.2), we have $\{Y_a, Y_b\} \perp\!\!\!\perp X | Y_c$, i.e., $P(Y_a(0), Y_b(0) | Y_c(0)) = P(Y_a(1), Y_b(1) | Y_c(1))$. For any $x \in \{0, 1\}$, we have that*

$$Y_a(x) \perp\!\!\!\perp Y_b(x) | Y_c(x), \qquad \text{or} \qquad Y_a(x) \not\perp\!\!\!\perp Y_b(x) | Y_c(x). \tag{4.3}$$



$\mathcal{G}_U$ C_U

Figure 3: A profile undirected graph with a compatible LWF chain graph.

Consider the profile undirected graph $\mathcal{G}_{\mathcal{U}} = (V, \mathcal{E}_{\mathcal{U}})$ *with* $\mathcal{E}_{\mathcal{U}} = \{(a,b)^{\{0\}}, (a,c)^{\emptyset}, (b,c)^{\mathcal{X}}\}$ *where the pair* a, b *is joined by a* $\{0\}$*-dotted edge that implies*

$$Y_a(0) \perp\!\!\!\perp Y_b(0)|Y_c(0) \qquad \text{and} \qquad Y_a(1) \not\perp\!\!\!\perp Y_b(1)|Y_c(1). \tag{4.4}$$

Statements (4.3) and (4.4) are not compatible, then C_U *does not belong to the induced class* $\mathcal{C}_U$*. Now, consider the chain graph* $C_U^{'}$ *with* $E_{C_U^{'}} = \{(a,b), (b,c), (X,b), (X,c)\}$*, where only* b *is joined to* X*. Equation (4.2) implies that* $Y_a \perp\!\!\!\perp X|\{Y_b, Y_c\}$*, that is,* $P[Y_a(0)|Y_b(0), Y_c(0)] = P[Y_a(1)|Y_b(1), Y_c(1)]$*. Condition in Equation (4.3) still holds since either* $P[Y_a(1)|Y_c(1)] = [Y_a(0)|Y_c(0)]$ *or* $P[Y_a(1)|Y_b(1), Y_c(1)] = [Y_a(0)|Y_b(1)Y_c(0)]$*. Then,* $C^{'}$ *does not belongs to the induced class* $\mathcal{C}_U$*. Finally, consider the chain graph* $C_U^{''}$ *with the set* $E_{C_U^{''}} = \{(a,b), (b,c), (X,a), (X,b), (X,c)\}$ *of edges , where* a *and* b *are both joined to* X*. This chain graph is compatible with the profile graph* $\mathcal{G}_{\mathcal{U}}$ *since statement (4.3) is no more required. Then* $C_U^{''}$ *satisfies both conditions (i) and (ii) in Definition 4.1 and belongs to the induced class* $\mathcal{C}_U$*.*

In essence, given a profile undirected graph $\mathcal{G}_{\mathcal{U}}$, the induced class $\mathcal{C}_U$ includes LWF chain graphs $C_U = \{(V, X), E_{C_U}\}$ where the chain component V has the same skeleton of $\mathcal{G}_{\mathcal{U}}$, and differ only according to which arrows are missing. Within this class, we can identify the *maximum element*, i.e., the chain graph with no missing arrows and the *minimum element*, i.e., the chain graph with a set of arrows that point to all vertices $a \in V$ such that $nb_{\mathcal{Z}}(a) \neq \emptyset$ with $Z \subset \mathcal{X}$ and $Z \neq \emptyset$.

Theorem 4.1. *Consider a profile undirected graph* $\mathcal{G}_U = (V, \mathcal{E}_{\mathcal{U}})$ *associated to the profile outcome vectors* $Y_{V|\mathcal{X}}$*. If the probability distributions* $P[Y_{V|\mathcal{X}}]$ *satisfy the* $\mathcal{U}$*-GMP for* $\mathcal{G}_U$*, then also the independence statements in Equation (4.1) are satisfied for the induced class* $\mathcal{C}_U$ *of two-block LWF chain graphs.*

In order to account also for the independence statements (4.2) and to establish a one-to-one relationship between profile undirected graphs and LWF chain graphs, we generalize the class of profile undirected graphs. Given a profile undirected graph $\mathcal{G}_{\mathcal{U}} = (V, \mathcal{E}_{\mathcal{U}})$, consider the partition $V = V_{\bigcirc} \cup V_{\square}$ of the vertex set so that we distinguish between two types of vertices, a *circle vertex* $a_{\bigcirc} \in V_{\bigcirc}$ and a *square vertex* $a_{\square} \in V_{\square}$, drawn as $\bigcirc$ and $\square$, respectively. For every $a_{\square} \in V_{\square}$, we assume $Y_a \perp\!\!\!\perp X | Y_{V \setminus a}$; that is the univariate profile distribution of $Y_a(x)$ is invariant for any $x \in \mathcal{X}$, given the remaining variables $Y_{V \setminus a}$; otherwise if $a_{\bigcirc} \in V_{\bigcirc}$, we assume $Y_a \not\perp\!\!\!\perp X | Y_{V \setminus a}$.

The profile graph in this generalized representation includes information also about the independence structure between subsets of response variables Y_A, with $A \subseteq V$, and the external factor X. In particular, for any $A \subseteq V_{\square}$, we assume that $Y_A \perp\!\!\!\perp X | Y_{V \setminus A}$. Then, given a profile undirected graph $\mathcal{G}_{\mathcal{U}} = (V_{\bigcirc}, V_{\square}, \mathcal{E}_{\mathcal{U}})$, the compatible two-block LWF chain graph $C_U = [\{V, X\}, E_{C_U}]$ in the class $\mathcal{C}_U$ is unique and is defined by a chain graph where the undirected graph of the response component V has the same skeleton of $\mathcal{G}_{\mathcal{U}}$ and there are missing arrows between X and any square vertex $a_{\square} \in V_{\square}$.

Example 4.2. *Consider the profile undirected graph and the induced chain graph in Figure 3. Vertices a, b, c are circled vertices while d is a square vertex wrt $\mathcal{G}_{\mathcal{U}}$, i.e., $\{a, b, c\} \in V_{\bigcirc}$ and $d \in V_{\square}$. Then, both the profile undirected graph and the chain graph imply the independence statement $Y_d \perp\!\!\!\perp X | \{Y_a, Y_b, Y_c\}$. Also, both graphs imply that $\{Y_a, Y_c\} \perp\!\!\!\perp Y_d | \{Y_b, X\}$. Unlike the profile graph, the chain graph does not provide information about the effect of X on the Y_V association structure, e.g., $Y_a(2) \perp\!\!\!\perp Y_b(2) | \{Y_c(2), Y_d(2)\}$.*

4.2 Profile bi-directed graphs and regression graphs

For any profile bi-directed graph $\mathcal{G}_{\mathcal{B}}$ we may define an induced class of two-block regression graphs $\mathcal{C}_B = \{C_B\}$. Any $C_B = [\{V, X\}, E_{C_B}]$ in the class $\mathcal{C}_B$ represents a regression graph such that vertices $a, b \in V$ in the response variables component can be joined by full bi-directed edges. Our aim is to show that any profile bi-directed graph model for the set of profile distributions $P[Y_{V|\mathcal{X}}]$ is compatible in terms of independence models with a class of regression graph models for the joint distribution $P(Y_V, X)$. Compatibility between profile bi-directed graphs and regression graphs is formally defined within Theorem 4.2. We briefly recall the Global Markov property for a two-block regression graph C_B; see also Drton (2009) and Wermuth and Sadeghi (2012).

A joint probability distribution $P(Y_V, X)$ satisfies the regression Global Markov property (R-GMP) wrt the regression graph C_B if,

for any disconnected set $D \subseteq V$ with connected components $K_1, \ldots, K_r$,

$$Y_{K_1} \perp\!\!\!\perp \ldots \perp\!\!\!\perp Y_{K_r} | X; \tag{4.5}$$

for any subset $A \subseteq V$ such that there is a missing arrow between any vertex $a \in A$ and X,

$$Y_A \perp\!\!\!\perp X \tag{4.6}$$

Definition 4.2. *Given a profile bi-directed graph $\mathcal{G}_\mathcal{B} = (V, \mathcal{E}_\mathcal{B})$, consider the induced class $\mathcal{C}_B = \{C_B\}$ of regression graphs where a graph $C_B = [\{V, X\}, E_{C_B}]$ is in $\mathcal{C}_B$ if,*

(i) any couple $a, b \in V$ is joined by a bi-directed edge in C_B if $\mathcal{Z} \subset \mathcal{X}$ for the pair $(a, b)^\mathcal{Z} \in \mathcal{E}_\mathcal{B}$;

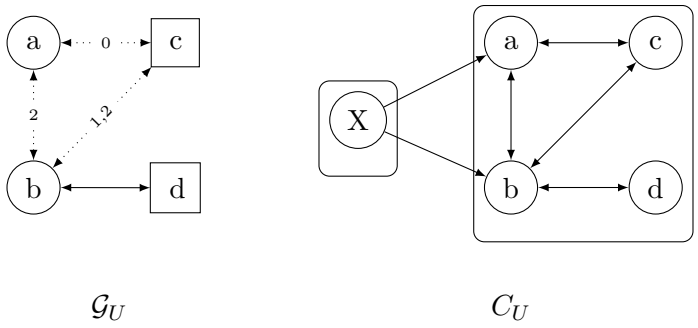
(ii) for any couple $a, b \in V$, a or b are reached by an arrow in C_B starting from X if $\mathcal{Z} \subset \mathcal{X}$ and $\mathcal{Z} \neq \emptyset$ for the pair $(a, b)^\mathcal{Z} \in \mathcal{E}_\mathcal{B}$.

Given a profile bi-directed graph $\mathcal{G}_\mathcal{B}$, the induced class of regression graphs includes a set of regression graphs C_B where the bi-directed graph of the response chain component has the same skeleton of $\mathcal{G}_\mathcal{B}$. Therefore, regression graphs $C_B \in \mathcal{C}_B$ differ only according to which arrows are present. Within this class, we can identify the *maximum element*, and the *minimum element*, that in this case may not be unique; we provide in the Supplementary Material an iterative procedure to find the size, in terms of number of arrows, of the minimum element.

Necessary conditions (i) and (ii) in Definition 4.2 ensure that it will always exist at least one compatible regression graph for any given profile bi-directed graph. Since condition (ii) may not be intuitive, the following counterexample shows that this is a necessary condition.

Example 4.3. *Following Example 4.1, let $V = \{a, b, c\}$ be a set of response variables and X a factor with state-space $\mathcal{X} = \{0, 1\}$. Consider a regression graph $C_B = \{(V, X), E_{C_B}\}$ with $E_{C_B} = \{(a, b), (b, c), (X, c)\}$, where vertices a and b are both disjointed from X. For the R-GMP in (4.6), we have $\{Y_a, Y_b\} \perp\!\!\!\perp X$. Then, $P[Y_a(0), Y_b(0)] = P[Y_a(1), Y_b(1)]$. Therefore, for any $x \in \{0, 1\}$, we have that*

$$Y_a(x) \perp\!\!\!\perp Y_b(x), \qquad \text{or} \qquad Y_a(x) \not\perp\!\!\!\perp Y_b(x). \tag{4.7}$$



$\mathcal{G}_U$ C_U

Figure 4: A profile bi-directed graph with a compatible regression graph.

Consider the profile bi-directed graph $\mathcal{G}_{\mathcal{B}} = (V, \mathcal{E}_{\mathcal{B}})$ *with* $\mathcal{E}_{\mathcal{B}} = \{(a,b)^{\{0\}}, (a,c)^{\emptyset}, (b,c)^{\mathcal{X}}\}$ *where the pair* a, b *is joined by a* $\{0\}$*-dotted edge that implies*

$$Y_a(0) \perp\!\!\!\perp Y_b(0) \qquad \text{and} \qquad Y_a(1) \not\perp\!\!\!\perp Y_b(1). \tag{4.8}$$

Statements (4.7) *and* (4.8) *are not compatible, then* C_B *does not belong to the induced class* $\mathcal{C}_B$*. Now, consider the regression graph* $C_B^{'}$ *with* $E_{B_U^{'}} = \{(a,b), (b,c), (X,b), (X,c)\}$*, where only* b *is joined to* X*. Equation* (4.6) *implies* $Y_a \perp\!\!\!\perp X$*, .i.e.,* $P[Y_a(0)] = P[Y_a(1)]$ *which is compatibile with the profile independence in* (4.8)*. Then,* $C_B^{'}$ *satisfies both conditions (i) and (ii) in Definition 4.2 and belongs to the induced class* $\mathcal{C}_B$*. We draw similar conclusion for the regression graph* $C_B^{''}$ *with* $E_{C_B^{''}} = \{(a,b), (b,c), (X,a), (X,b), (X,c)\}$*, where* a *and* b *are both joined to* X*.*

Theorem 4.2. *Consider a profile bi-directed graph* $\mathcal{G}_B = (V, \mathcal{E}_{\mathcal{B}})$ *associated to the profile outcome vectors* $Y_{V|\mathcal{X}}$*. If the probability distributions* $P[Y_{V|\mathcal{X}}]$ *satisfy the* $\mathcal{B}$*-GMP for* $\mathcal{G}_B$*, then the independence statements in Equation* (4.5) *are satisfied for the induced class* $\mathcal{C}_B$ *of two-block regression graphs.*

Assuming the partition of the vertex set $V = V_{\bigcirc} \cup V_{\square}$ into circled and square vertices, we can also generalize the class of profile bi-directed graphs in order to account for the independence statements in Equation (4.6) and to establish a one-to-one relationship between profile bi-directed graphs and regression graphs. Given a square vertex $a_{\square} \in V_{\square}$, we assume that $Y_a \perp\!\!\!\perp X$, that is the univariate profile variable distribution of $Y_a(x)$ is invariant for any $x \in \mathcal{X}$; otherwise we have a circled vertex $a_{\bigcirc} \in V_{\bigcirc}$. More generally, for any subset $A \subseteq V_{\square}$, we assume $Y_A \perp\!\!\!\perp X$. Then, given a profile bi-directed graph $\mathcal{G}_{\mathcal{B}} = (V_{\bigcirc}, V_{\square}, \mathcal{E}_{\mathcal{B}})$, the

corresponding compatible regression graph $C_B = [\{V, X\}, E_{C_B}]$ is defined by a two-block regression graph where the bi-directed graph of the response component V has the same skeleton of $\mathcal{G}_\mathcal{B}$ with missing arrows between any vertex $a \in V_\square$ and X.

Example 4.4. *Consider the profile bi-directed graph in Figure 4 where vertices a and b are circled vertices while c and d are square vertices; the graph $\mathcal{G}_\mathcal{B}$ implies $\{Y_c, Y_d\} \perp\!\!\!\perp X$ which also holds in the compatible regression graph C_B because there is no arrow pointing to $c, d \in V$. Both the profile and the regression graph imply that $\{Y_a, Y_c\} \perp\!\!\!\perp Y_d | X$ because d and $\{a, c\}$ are the connected components of the disconnected set $\{a, c, d\}$. The induced regression graph does not provide information about the profile marginal independencies, e.g., $Y_a(2) \perp\!\!\!\perp Y_b(2)$.*

5 Parametrization and inference for Gaussian profile graph models

We define the class of *Gaussian profile graphical models* for both the undirected and bi-directed types. For all $x \in \mathcal{X}$, let $Y_V(x) \sim N(\alpha + \beta_x, \Sigma(x))$ where $[\alpha_a + \beta_{ax}]_{a \in V} = \mathbb{E}[Y_a(x)]_{a \in V}$ is the *profile marginal mean vector* and $\Sigma(x)$ is the *profile covariance matrix* with entries $\omega_{ab}(x)$, $x \in \mathcal{X}$. Let γ_{ax}, $a \in V$, be the linear effect of the external factor on the *profile conditional mean vector* $\mathbb{E}[Y_a(x)|Y_{V \setminus a}(x)]_{a \in V}$ and let $\Lambda(x) = \Sigma^{-1}(x)$ be the *profile precision matrix* with entries $\lambda_{ab}(x)$, $x \in \mathcal{X}$; note that $\gamma_x = \Lambda(x)\beta_x$ where $\gamma_x = [\gamma_{ax}]_{a \in V}$ and $\beta_x = [\beta_{ax}]_{a \in V}$ (Andersson et al., 2001).

The two classes of Gaussian profile undirected and bi-directed graphs are defined by different zero-constraints over the model parameters; these constraints naturally follow from the Markov equivalence between profile graphs and multiple graphs, and the compatibility between profile graphs and chain graphs established in the previous Sections 3 and 4.

The *Gaussian profile undirected graphical model* for $Y_{V|\mathcal{X}}$ wrt $\mathcal{G}_\mathcal{U} = (V, \mathcal{E}_\mathcal{U})$ is such that,

(i) for any $a \in V_\square$, $\gamma_{ax} = 0$ for all $x \in \mathcal{X}$,

(ii) for any $(a, b)^\mathcal{Z} \in \mathcal{E}_\mathcal{U}$, with $\mathcal{Z} \subseteq \mathcal{X}$, $\lambda_{ab}(x) = 0$, for each $x \in \mathcal{Z}$.

The *Gaussian profile bi-directed graphical model* for $Y_{V|\mathcal{X}}$ wrt $\mathcal{G}_\mathcal{B} = (V, \mathcal{E}_\mathcal{B})$ is such that,

(i) for any $a \in V_\square$, $\beta_{ax} = 0$ for all $x \in \mathcal{X}$,

(ii) for any $(a,b)^{\mathcal{Z}} \in \mathcal{E}_{\mathcal{B}}$, with $\mathcal{Z} \subseteq \mathcal{X}$, $\omega_{ab}(x) = 0$, for each $x \in \mathcal{Z}$.

Estimation of Gaussian profile undirected graphical models can be virtually based on any method previously developed for inference of Gaussian chain graphs or multiple graphs. The neighborhood selection approach proposed by Meinshausen and Bühlmann (2006) results in a asymptotically consistent estimator of high-dimensional graph structures. Given $n(x)$ i.i.d. observations, neighborhood selection aims at estimating the neighbours of each vertex $a \in V$. In a nutshell, neighborhood selection can be framed as a standard regression problem and can be solved efficiently with the Lasso (Meinshausen and Bühlmann, 2006). For each $a \in V$ and $x \in \mathcal{X}$, a regression model with $Y_a(x)$ as response variable and all remaining variables $Y_{V\setminus a}(x)$ as covariates is estimated with a Lasso penalty. The coefficients estimated to be zero define the zero patterns that correspond to a given profile undirected graph $\mathcal{G}_{\mathcal{U}} = (V, \mathcal{E}_{\mathcal{U}})$. The finite sample properties of the proposed Lasso procedure have been explored through a simulation study. The adopted simulation scheme and the results are described in the Supplementary Material. As expected, the performances of some indices of interest, e.g., true and false positive rate, improve as the sample size increases. The extension of the aforementioned procedure to Gaussian profile bi-directed graphical model is trivial; technical aspects related to the implementation of these algorithms are discussed in the Supplementary Material. Note that alternative modeling frameworks may be used. For example, penalty terms that encourage shared network structures across profiles, in the same spirit of the group graphical lasso of Danaher et al. (2014), may be appropriate in case of x-profile outcome vectors that follows distributions with a very similar graph structure. The development of more tailored modeling frameworks is beyond the scope of this manuscript.

6 Application

We illustrate the utility of our method with an application in cancer genomics. We analyze protein expression data from patients affected by acute myeloid leukemia (AML) with the goal of reconstructing and comparing protein networks across disease subtypes; comparing the networks for these groups provides insight into the differences in protein signaling that may affect whether treatments for one subtype will be effective in another. The R-code is available and located at `https://github.com/kinglaz90/phd`.

A set of protein levels, collected using the reverse phase protein array (RPPA) technology, is observed in a sample of 213 newly diagnosed AML patients (Kornblau et al., 2009)[1]. Patients are classified by subtype according to the French-American-British (FAB) classification system. We consider 4 different profiles given by 4 AML subtypes, for which a reasonable sample size is available: M0 (17 subjects), M1 (34 subjects), M2 (68 subjects), and M4 (59 subjects). These profiles, based on criteria including cytogenetics and cellular morphology, show varying prognosis. We expect to observe different protein interactions in the subtypes. We focus on 18 proteins relevant to the apoptosis and cell cycle regulation KEGG pathways (Kanehisa et al., 2011).

Our interest is modelling the effect of the AML subtype on the joint independence structure of the protein levels. Profile undirected graphical models are an encompassing tool that coherently and jointly performs all inferential tasks of interest of learning how the protein dependency structure changes across subtypes as well as the mean protein levels. Therefore, considering the $p = 18$ protein levels following a multivariate Gaussian distribution and $q = 4$ different profiles of AML, where the levels $x \in \mathcal{X} = \{0, 1, 2, 3\}$ denote the subtypes M0, M1, M2, M4 respectively, we estimate and select the profile undirected graphical model represented in Figure 5. For the sake of comparison, we represent the corresponding compatible chain graph in Figure 6; this chain graph is more dense and then harder to read. Most importantly, the many profile specific independencies are obviously missed by the chain graph.

For instance, from the selected profile graph we learn that for the profiles $x \in \{0, 1\}$,

$$Y_{\text{AKTp.308}}(x) \perp\!\!\!\perp Y_{\text{BCI.2}}(x) | Y_{V \setminus \{\text{AKTp.308,BCI.2}\}}(x);$$

for any profile $x \in \mathcal{X}$,

$$Y_{\text{AKTp.308}}(x) \perp\!\!\!\perp Y_{\text{BAD}}(x) | Y_{V \setminus \{\text{AKTp.308,BAD}\}}(x).$$

The level of proteins AKTp.473, BAX, BCI.XL, PTEN, PTEN.p. TP53 and XIAP and all the pairwise associations among them are independent to the AML subtypes.

[1] http://bioinformatics.mdanderson.org/Supplements/Kornblau-AML-RPPA/aml-rppa.xls

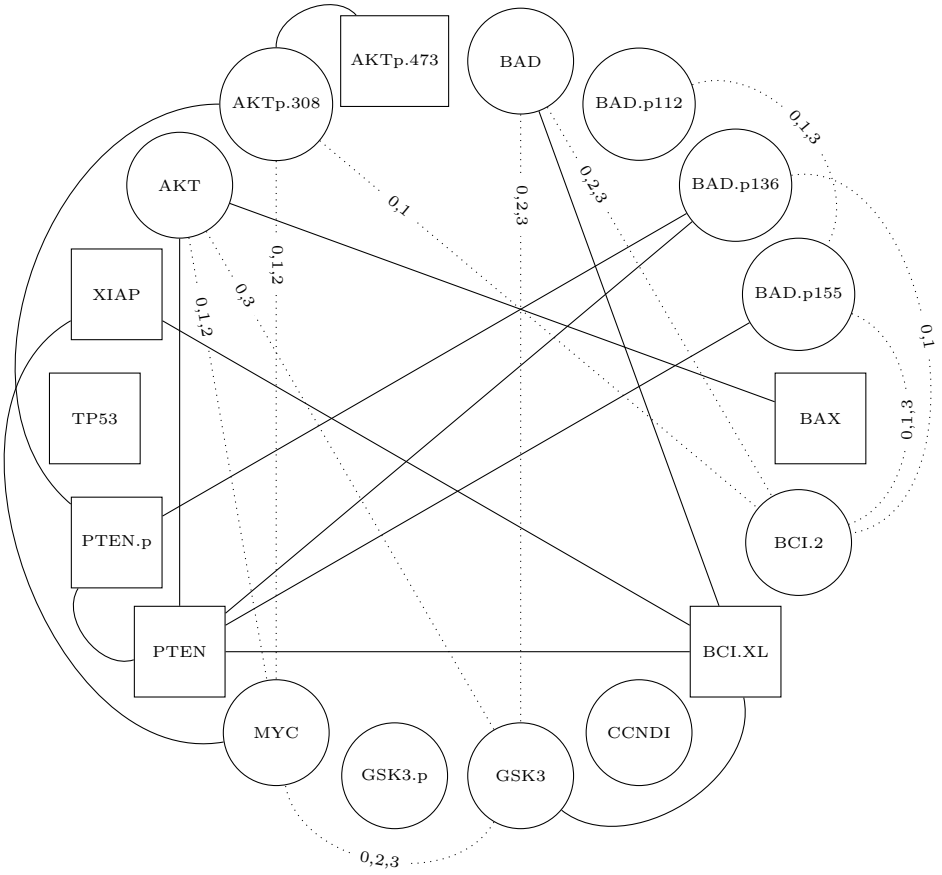


Figure 5: The selected profile undirected graph model for protein data

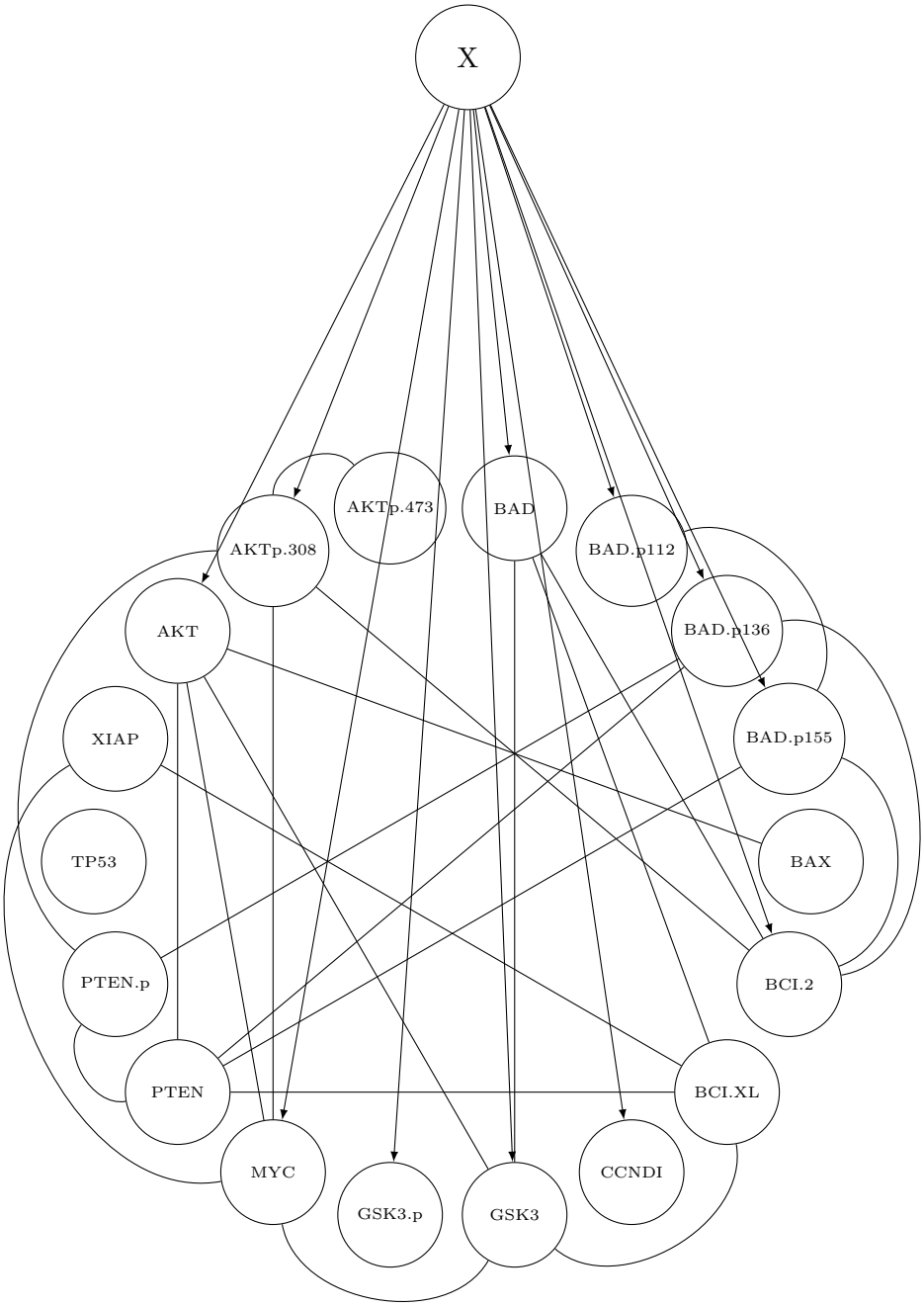


Figure 6: The compatible chain graph model for protein data

7 Discussion

We propose a class of graphical models that generalizes both chain graphs and multiple graphs and, for the first time, we establish compatibility between these two types of graph. The proposed approach is compatible with different types of chain graph models, which provide different regression frameworks for data analysis. In line with LWF chain graphs, profile undirected graphs can be used for modelling the profile conditional independencies resulting from a sequence of non-independent regression models involving all response variables. On the other hand, in line with regression graphs, profile bi-directed graph models can be used for modelling profile marginal independencies resulting from a sequence of non-independent marginal regression models which ignore other response variables. See Wermuth and Sadeghi (2012) for a discussion about chain graphs in terms of sequences of non-independent regression models. From this perspective, the specification of a class of profile chain graphs represents an interesting generalization to explore profile independencies in a multivariate regression setting. We conjecture that, under certain assumptions, this generalization is more feasible for the class of LWF chain graphs rather than for the class of regression graphs. Nevertheless, there are some crucial aspects, mainly related to the specification of the Markov properties, which need to be better investigated. The class of profile graphs we propose can be used for modelling the profile distributions of both continuous and discrete outcomes. The parameterization discussed in Section 5 for the Gaussian case is quite standard and it is based on the idea that these models correspond to sequences of non-independent regressions. Under the assumption of a Multinomial sampling scheme for the multivariate outcome vector, a parameterization based on the log-linear transformation (Lauritzen, 1996) and on the log-mean linear transformation (Roverato et al., 2013) could be used, respectively, for profile undirected and bi-directed graph models. In alternative to the selection strategy based on a Lasso penalty, model comparison within the class of profile undirected or profile bi-directed graphs can be based on the likelihood ratio test in case of nested models; these graphical models are smooth and belong to the curve exponential family, so the likelihood ratio test has an asymptotic chi-square distributions. However, if undirected and bi-directed graphs are both viable options for a given data analysis, the likelihood ratio test for model comparison between the two types of graphs cannot be used and we need to rely on some type of information criteria, such as AIC or BIC.

Appendix

Profile local Markov property

The probability distributions $P[Y_{V|\mathcal{X}}]$ of the profile outcome vectors $Y_{V|\mathcal{X}}$ satisfy the *profile undirected Local Markov Property* ($\mathcal{U}$-LMP) wrt the graph $\mathcal{G}_{\mathcal{U}} = (V, \mathcal{E}_{\mathcal{U}})$ if, for any vertex $a \in V$

$$Y_a(x) \perp\!\!\!\perp Y_{V \setminus \{a \cup nb_x(a)\}}(x) | Y_{nb_x(a)}(x), \qquad x \in \mathcal{X}. \tag{7.1}$$

The probability distributions $P[Y_{V|\mathcal{X}}]$ of the profile outcome vectors $Y_{V|\mathcal{X}}$ satisfy the *profile bi-directed Local Markov Property* ($\mathcal{B}$-LMP) wrt the graph $\mathcal{G}_{\mathcal{B}} = (V, \mathcal{E}_{\mathcal{B}})$ if, for any vertex $a \in V$

$$Y_a(x) \perp\!\!\!\perp Y_{V \setminus \{a \cup nb_x(a)\}}(x), \qquad x \in \mathcal{X}. \tag{7.2}$$

Proofs

Proof. ***of Theorem 3.1.*** Let $\mathcal{G}_{\mathcal{U}} = (V, \mathcal{E}_{\mathcal{U}})$ be a profile undirected graph and $D \subseteq V$ be any x-disconnected set with x-connected components $K_1, \ldots, K_r$ such that for every pair K_i, K_j with $i, j = 1, \ldots, r$, $i \neq j$, for the $\mathcal{U}$-CSMP wrt $\mathcal{G}_{\mathcal{U}}$ we have

$$Y_{K_i}(x) \perp\!\!\!\perp Y_{K_j}(x) | Y_{V \setminus \{K_i, K_j\}}(x), \qquad x \in \mathcal{X}. \tag{7.3}$$

For any pair $K_i, K_j \subset D$ with $i, j = 1, \ldots, r$, $i \neq j$, the set $S_{ij} = V \setminus \{K_i, K_j\}$ is an x-separator. Then, the $\mathcal{U}$-CSMP implies the $\mathcal{U}$-GMP wrt $\mathcal{G}_{\mathcal{U}}$. Conversely, consider any x-connected set C wrt $\mathcal{G}_{\mathcal{U}}$ and let $nb_x(C) = \bigcup_{a \in C} nb_x(a)$ be the neighbour set including C and let $S_C = nb_x(C) \setminus C$ be an x-separator for the sets C and $V \setminus nb_x(C)$, for any $x \in \mathcal{X}$. The $\mathcal{U}$-GMP implies that

$$Y_C(x) \perp\!\!\!\perp Y_{V \setminus nb_x(C)}(x) | Y_{S_C}(x), \qquad x \in \mathcal{X}. \tag{7.4}$$

Note that $C \cup \{V \setminus nb_x(C)\}$ is an x-disconnected set, for $x \in \mathcal{X}$. We distinguish two cases, whether $V \setminus nb_x(C)$ is x-connected or x-disconnected. In the first case, the x-connected components of $C \cup \{V \setminus nb_x(C)\}$ are C and $V \setminus nb_x(C)$, then the $\mathcal{U}$-CSMP is satisfied. If $V \setminus nb_x(C)$ is x-disconnected with $K_1 \cup \cdots \cup K_r$ connected components, the $\mathcal{U}$-GMP also implies that

$$Y_C(x) \perp\!\!\!\perp Y_{K_1}(x) \perp\!\!\!\perp \ldots \perp\!\!\!\perp Y_{K_r}(x) | Y_{S_C}(x), \qquad x \in \mathcal{X}. \tag{7.5}$$

Then the $\mathcal{U}$-GMP implies the $\mathcal{U}$-CSMP wrt $\mathcal{G}_{\mathcal{U}}$. □

Proof. ***of Proposition 3.1.*** Consider a profile undirected graph $\mathcal{G}_{\mathcal{U}} = (V, \mathcal{E}_{\mathcal{U}})$ associated to the profile outcome vectors $Y_{V|\mathcal{X}}$ and the induced class $U_{V|\mathcal{X}}$ of multiple undirected graphs. If the probability distributions $P[Y_{V|\mathcal{X}}]$ satisfy the $\mathcal{U}$-CSMP wrt $\mathcal{G}_{\mathcal{U}}$, the $\mathcal{U}$-GMP is also satisfied from Theorem 3.1. So, given three disjoint subsets A, B, C of V,

$$Y_A(x) \perp\!\!\!\perp Y_B(x) | Y_C(x), \tag{7.6}$$

where A and B are x-separated by C, with $x \in \mathcal{X}$. The result follows by Definition 3.1, since A and B are x-separated by C in $\mathcal{G}_{\mathcal{B}}$ if and only if they are x-separated by C in $U(x) \in U_{V|\mathcal{X}}$, with $x \in \mathcal{X}$. □

Proof. ***of Proposition 3.2.*** From a well-established result we have that, given an undirected graph $U = (V, E_U)$ associated to a random vector Y_V, the global and the pairwise Markov properties are equivalent if the joint probability distribution $P(Y_V)$ is strictly positive; see Lauritzen (1996). The proposition follows by applying this result to the strictly positive probability distribution $P[Y_V(x)]$ of any profile outcome vector $Y_V(x) \in Y_{V|\mathcal{X}}$. □

Proof. ***of Proposition 3.3.*** From a result of Drton and Richardson (2008), given a bi-directed graph $B = (V, E_B)$ associated to a random vector Y_V, the joint probability distribution $P(Y_V)$ satisfies the global Markov property if and only if the connected set Markov property is satisfied. The proposition follows by applying this result to the profile distribution $P[Y_V(x)]$ of any profile outcome vector $Y_V(x) \in Y_{V|\mathcal{X}}$ associated to a profile bi-directed graph $\mathcal{G}_{\mathcal{B}} = (V, \mathcal{E}_{\mathcal{B}})$. □

Proof. ***of Proposition 3.4.*** Consider a profile bi-directed graph $\mathcal{G}_{\mathcal{B}}$ associated to the profile outcome vectors $Y_{V|\mathcal{X}}$. If $P[Y_{V|\mathcal{X}}]$ satisfy the $\mathcal{B}$-GMP, the $\mathcal{B}$-CSMP is satisfied by Proposition 3.3, and, for every pair $(a, b)^{\mathcal{Z}} \in \mathcal{E}_{\mathcal{U}}$ with $Z \neq \emptyset$, the $\mathcal{B}$-PMP is also satisfied since

$$Y_a(x) \perp\!\!\!\perp Y_b(x), \qquad x \in \mathcal{Z}.$$

□

Proof. ***of Proposition 3.5.*** Consider a profile bi-directed graph $\mathcal{G}_{\mathcal{B}} = (V, \mathcal{E}_{\mathcal{B}})$ associated to the profile outcome vectors $Y_{V|\mathcal{X}}$. If the probability distributions $P[Y_{V|\mathcal{X}}]$ satisfy the $\mathcal{B}$-CSMP wrt $\mathcal{G}_{\mathcal{U}}$, for every x-disconnected set $D \subseteq V$ with x-connected components

$K_1, \ldots, K_r$,

$$Y_{K_1}(x) \perp\!\!\!\perp \ldots \perp\!\!\!\perp Y_{K_r}(x), \qquad x \in \mathcal{X}. \tag{7.7}$$

The result follows by Definition 3.2, since any $D \subseteq V$ is an x-disconnected set in $\mathcal{G}_{\mathcal{B}}$ if and only if it is a disconnected set with the same connected components in $B(x) \in B_{V|\mathcal{X}}$, for any $x \in X$. □

Proof. ***of Theorem 4.1.*** Given a profile undirected graph $\mathcal{G}_{\mathcal{U}}$ associated to the profile outcome vectors $Y_{V|\mathcal{X}}$, if the probability distributions $P[Y_{V|\mathcal{X}}]$ satisfies the $\mathcal{U}$-GMP, the $\mathcal{U}$-CSMP is also satisfied from Theorem (3.1). For every x-disconnected set D with x-connected components $K_1, \ldots, K_r$,

$$Y_{K_1}(x) \perp\!\!\!\perp \ldots \perp\!\!\!\perp Y_{K_r}(x) | Y_{V \setminus D}(x), \quad x \in \mathcal{X}.$$

Then, the independence statement in Equation (4.1) is satisfied since, by Definition 4.1, for any chain graph $C_U = [\{V, X\}, E_{C_U}]$ in $\mathcal{C}_U$, each disconnected set $D \subseteq V$ is also a disconnected set in $\mathcal{G}_{\mathcal{U}}$ with same connected components. □

Proof. ***of Theorem 4.2.*** A proof is given along the same line of the proof for Theorem 4.1. Then, the result follows since, for any regression graph $\mathcal{C}_B = [\{V, X\}, E_{C_B}]$ in $\mathcal{C}_B$, each disconnected set $D \subseteq V$ is also a disconnected set in $\mathcal{G}_{\mathcal{B}}$ with same connected components, by Definition 4.2. □

Supplementary materials

Simulation study

We perform two simulation studies to investigate the finite sample properties of the neighborhood selection approach proposed in Section 7. This method estimates the set of non-zero elements of precision (covariance) matrices corresponding to the presence of edges for profile undirected (bi-directed) graphs. Moreover, this method estimates the non-zero coefficient vector γ_x (β_x) corresponding to circled nodes of a profile undirected (bidirected) graph. The R-code is available and located at https://github.com/kinglaz90/phd.

Simulation study I

Data generating process: we generate observations from a set $Y_{V|\mathcal{X}}$ of random vectors associated to a profile undirected graph $\mathcal{G}_{\mathcal{U}}$ with $p = 20$ nodes and $q = 5$ levels of X, such that $x \in \mathcal{X} = \{0, 1, 2, 3, 4\}$. Following Peterson et al. (2015), we first construct $\Lambda(0)$, the precision matrix of the baseline level $x = 0$. We set $\Lambda(0)$ to be a $p \times p$ symmetric matrix with main diagonal entries $\lambda_{aa}(0) = a$, with $a = 1, \ldots, p$, and off-diagonal entries $\lambda_{(a+1)a}(0) = \lambda_{a(a+1)}(0) = 0.5$ with $a = 1, \ldots, 19$ and $\lambda_{(a+2)a}(0) = \lambda_{a(a+2)}(0) = 0.4$ with $a = 1, \ldots, 18$. For all $a \in V$, we set both α_a and γ_{ax} to zero. For $x = \{1, 2, 3, 4\}$, we also set $\gamma_{ax} = 0$ for $a = 5, \ldots 20$ and $\gamma_{ax} = 1$ for $a = 1, \ldots 4$, i.e., the external factor X affects only the first four response variables. The remaining precision matrices $\Lambda(x)$ for $x = \{1, 2, 3, 4\}$ are obtained as follow, first we set $\Lambda(x) = \Lambda(0)$, then with probability 0.5 we set to zero its non-zero entries, given that constraint (ii) in subsection 4.1 is respected. The resulting precision matrices have about 17.5% of non-zero elements, on average. Data are generated by drawing a random sample of size $n(x) = 10, 10^2, 10^3$ from the distribution $\mathcal{N}(\beta_x, \Sigma(x))$ where $\Sigma(x) = \Lambda^{-1}(x)$ and $\beta_x = \Sigma(x)\gamma_x$, for all $x \in \mathcal{X}$. We use a similar procedure to generate data from a bi-directed profile graph: in the procedure we set $\Sigma(x) = \Lambda(x)$ and $\beta_x = \gamma_x$, for all $x \in \mathcal{X}$ and we draw samples from a multivariate normal distribution with such parameters.

Performance metrics and results: to assess the accuracy of graph structure estimation, we compute the true positive rate (TPR) and the true negative rate (TNR) of the non-zero elements of γ_x and $\Lambda(x)$ for the undirected graph and the non-zero elements of β_x and $\Sigma(x)$ for the bi-directed graph, for all $x \in \mathcal{X}$, along with associated standard errors (SE). The rates of precision and covariance matrices are obtained averaging over all the five matrices. We report the results in Table 1 and Table 2. The results show that all the TPR-s converge to 1 and all the SE go to 0, as $n(x)$ increases; in particular, with $n(x) = 10^3$ all the TPR-s are equal to 1, for all $x \in \mathcal{X}$.

Simulation study II

We perform a second simulation study to evaluate the proposed neighborhood selection method about the ability to correctly identify a full edge (F), a dotted edge (D) or a missing edge (M) for each couple $a, b \in V$, $a \neq b$. We construct a profile undirected graph $\mathcal{G}^*_{\mathcal{U}}$ and a

profile bi-directed graph $\mathcal{G}^*_{\mathcal{B}}$ by following the same procedure as in the first simulation study with the only difference of allowing X to affect the first ten variables; the resulting graphs have 19 full edges and 18 dotted edges (and therefore 153 missing edges). We generate 100 simulated datasets for each $n(x) = 10, 10^2, 10^3$, for all $x \in \mathcal{X}$. To assess the accuracy of edge-selection, we compute the true positive edge-rate (TPR.e) with associated SE. For true positive edge-rate we mean the proportion of F, D or M edges correctly classified. We report the results in Table 3 and Table 4. The results show that all TPR.e-s converge to 1 and all the SE go to 0, as $n(x)$ increases; in particular, with $n(x) = 10^3$ both TPR.e(F) and TPR.e(D) are equal to 1, for all $x \in \mathcal{X}$.

Table 1: True positive rate (TPR) and true negative rate (TNR) for non-zero elements of the precision matrices and regression coefficients with associated standard error (SE) across 100 simulated datasets corresponding to the profile undirected graph $\mathcal{G}_{\mathcal{U}}$ of the simulation study.

$n(x)$	TPR[$\Lambda(x)$](SE)	TNR[$\Lambda(x)$](SE)	TPR[γ_x](SE)	TNR[γ_x](SE)
10	0.75(0.08)	0.89(0.03)	0.36(0.24)	0.75(0.10)
10^2	1.00(0.00)	0.91(0.02)	0.99(0.04)	0.86(0.07)
10^3	1.00(0.00)	0.91(0.01)	1.00(0.00)	0.96(0.05)

Table 2: True positive rate (TPR) and true negative rate (TNR) for non-zero elements of the covariance matrices and regression coefficients with associated standard error (SE) across 100 simulated datasets corresponding to the profile bi-directed graph $\mathcal{G}_{\mathcal{B}}$ of the simulation study.

$n(x)$	TPR[$\Sigma(x)$](SE)	TNR[$\Sigma(x)$](SE)	TPR[β_x](SE)	TNR[β_x](SE)
10	0.72(0.09)	0.95(0.02)	0.61(0.31)	0.93(0.07)
10^2	0.95(0.02)	0.97(0.02)	1.00(0.00)	0.95(0.06)
10^3	1.00(0.00)	0.97(0.02)	1.00(0.00)	0.95(0.06)

Table 3: True positive edge-rate (TPR.e) of the three types of edge F, D and M with associated standard error (SE) across 100 simulated datasets of the profile undirected graph $\mathcal{G}^*_{\mathcal{U}}$ of simulation study

$n(x)$	TPR(F)(SE)	TPR(D)(SE)	TPR(M)(SE)
10	0.77(0.09)	0.09(0.17)	0.89(0.03)
10^2	1.00(0.02)	0.96(0.10)	0.91(0.02)
10^3	1.00(0.00)	1.00(0.00)	0.91(0.01)

Table 4: True positive edge-rate (TPR.e) of the three types of edge F, D and M with associated standard error (SE) across 100 simulated datasets of the profile bi-directed graph $\mathcal{G}^*_{\mathcal{B}}$ of simulation study

$n(x)$	TPR.e(F)(SE)	TPR.e(D)(SE)	TPR.e(M)(SE)
10	0.77(0.14)	0.08(0.07)	0.97(0.02)
10^2	0.98(0.04)	0.69(0.07)	0.99(0.01)
10^3	1.00(0.00)	1.00(0.00)	0.99(0.01)

Number of arrows in minimal compatible regression graph

Consider a profile bi-directed graph $\mathcal{G}_{\mathcal{B}} = (V, \mathcal{E}_{\mathcal{U}})$ and let $nb_{\mathbf{x}}(a) = \bigcup_{x\in\mathcal{Z}} nb_x(a)$ for all $a \in V$, with $\mathcal{Z} \subset \mathcal{X}$ and $\mathcal{Z} \neq \emptyset$. The number of arrows in the minimal compatible regression graph wrt $\mathcal{G}_{\mathcal{B}}$, denoted by $\min_{\mathcal{C}_B}$, can be obtained through the following iterative procedure.

For all $a, b \in V$, $\mathcal{Z} \subset \mathcal{X}$ and $\mathcal{Z} \neq \emptyset$

1. Initialize $min_{\mathcal{C}_B} = 0$,
2. Identify the vertex $a \in V$ such that $|nb_{\mathbf{x}}(a)| = \max_{a\in V} |nb_{\mathbf{x}}(a)|$,
3. Set $\mathcal{E} = \mathcal{E} \setminus \{\bigcup_{b\in nb_{\mathbf{x}}(a)}(a, b)^{\mathcal{Z}}\}$ and $\min_{\mathcal{C}_B} = \min_{\mathcal{C}_B} +1$,
4. Stop if does not exists any $(a, b)^{\mathcal{Z}} \in \mathcal{E}$ else goes to 2.

Model selection: the algorithm

In this section we describe the algorithm used for the selection of profile graphs. This algorithm builds upon the neighboring selection approach of Meinshausen and Bühlmann (2006), and obviously differ according to the type of the graph. This algorithm is implemented in R and uses functions form the R package *glmnet* (Friedman et al., 2010). Let y be a data matrix with n independent observations of the $(p+1)$ variables $Y_1, Y_2, \ldots, Y_p, X$. For all $a \in V$, we denote with y_a the a-th column vector of y, whereas $y_{(a)}$ is the matrix y after removing the a-th column vector; the same also applies to the residuals matrix r and design matrix D. We consider $x = 0$ the baseline level of X.

Selection of a profile undirected graph $\mathcal{G}_{\mathcal{U}} = (V, \mathcal{E}_{\mathcal{U}})$ is performed as follows:

1. For all $a \in V$, we perform a Lasso regression using y_a as response vector and $y_{(a)}$ as design matrix. The external factor y_X is included into the model as $(q-1)$ dummy variables. For each response variable $a \in V$, we estimate a vector of $(p+q-1)$ coefficients, including $\hat{\phi}_{ax}$ for any $x \in \{\mathcal{X} \setminus 0\}$. The penalty parameter is selected with cross validation. For all $a \in V$, if $\hat{\phi}_{ax} = 0$ for all $x \in \{\mathcal{X} \setminus 0\}$ then $a \in V_{\square}$; otherwise $a \in V_{\bigcirc}$.

2. For all $a \in V$, we compute the residuals vector r_a, through a standard regression with response y_a and the only predictor y_X. We obtain a matrix r with columns $r_1, r_2, \ldots, r_p$. Then, we order the rows of r accordingly to the level $x \in \mathcal{X} = 0, 1, \ldots, q$ of the external factor.

3. For all $a \in V$, we perform again a Lasso regression with response r_a and predictors all $b \in V \setminus a$. We construct the design matrix D, where for any predictor $b \in V \setminus a$, if at least one of $a, b \in V_{\square}$, then

$$D_b = r_b,$$

otherwise

$$D_b = \begin{pmatrix} r_b(0) & 0 & 0 & 0 & 0 & 0 \\ 0 & r_b(1) & 0 & \vdots & 0 & 0 \\ \vdots & 0 & r_b(2) & 0 & \vdots & 0 \\ 0 & \vdots & 0 & \ddots & 0 & \vdots \\ 0 & 0 & \vdots & 0 & r_b(q-1) & 0 \\ 0 & 0 & 0 & 0 & 0 & r_b(q) \end{pmatrix},$$

where $r_b(x)$, $x \in \mathcal{X} = 0, 1, \ldots, q$, is the subvector of r_b that corresponds to the observations with $X = x$. Therefore, for all $a, b \in V$, if at least one of $a, b \in V_{\square}$ then we get a unique Lasso estimate $\hat{\phi}_{ab}$; if $\hat{\phi}_{ab} = 0$, then $(a, b)^{\mathcal{X}} \in \mathcal{E}_{\mathcal{U}}$, otherwise $(a, b)^{\emptyset} \in \mathcal{E}_{\mathcal{U}}$. Conversely, if both $a, b \in V_{\bigcirc}$, then we get q Lasso estimates, one for each level of X, i.e. $\hat{\phi}_{ab}(0), \hat{\phi}_{ab}(1), \ldots, \hat{\phi}_{ab}(q)$. If $\hat{\phi}_{ab}(x) = 0$ for all $x \in \mathcal{X}$, then $(a, b)^{\mathcal{X}} \in \mathcal{E}_{\mathcal{U}}$; if $\hat{\phi}_{ab}(x) \neq 0$ for all $x \in \mathcal{X}$, then $(a, b)^{\emptyset} \in \mathcal{E}_{\mathcal{U}}$; if $\hat{\phi}_{ab}(x) = 0$ for all $x \in \mathcal{Z}$ while $\hat{\phi}_{ab}(x) \neq 0$ for all $x \in \mathcal{X} \setminus \mathcal{Z}$, with $\mathcal{Z} \subset \mathcal{X}$, $\mathcal{Z} \neq \emptyset$, then $(a, b)^{\mathcal{Z}} \in \mathcal{E}_{\mathcal{U}}$.

4. From the previous step of this procedure, we obtain two Lasso estimates of the same coefficient $\phi_{ab}(x)$ or $\phi_{ba}(x)$, for any couple $a, b \in V$, for all $x \in \mathcal{X}$. We use the AND or OR strategy (Meinshausen and Bühlmann, 2006) to identify the non-zero estimates.

A profile bi-directed graph $\mathcal{G}_{\mathcal{B}} = (V, \mathcal{E}_{\mathcal{B}})$ is selected as follows:

1. For all $a \in V$, we perform a standard regression using y_a as response and y_X as the only predictor. For each response variable $a \in V$, we obtain an estimate $\hat{\psi}_{ax}$ for all $x \in \{\mathcal{X} \setminus 0\}$ and we evaluate the significance of X through a LRT test. If we fail to reject the null hypothesis, then we set to zero $\hat{\psi}_{ax}$ for all $x \in \{\mathcal{X} \setminus 0\}$. For all $a \in V$, if $\hat{\psi}_{ax} = 0$ for all $x \in \{\mathcal{X} \setminus 0\}$ then $a \in V_{\square}$; otherwise $a \in V_{\bigcirc}$.

2. From the previous step, we obtain the residuals vector r_a, for all $a \in V$. We construct

a matrix r with columns $r_1, r_2, \ldots, r_p$. Then, we order the rows of r according to each level $x \in \mathcal{X} = 0, 1, \ldots, q$.

3. For all $a, b \in V$, if both $a, b \in V_{\square}$, then we perform a standard regression using r_a as response and r_b as the only predictor. We obtain the point estimate $\hat{\psi}_{ab}$ and we evaluate the significance through a Wald test. The non-significant $\hat{\psi}_{ab}$ are set to zero. If $\hat{\psi}_{ab} = 0$, then $(a, b)^{\mathcal{X}} \in \mathcal{G}_{\mathcal{B}}$ otherwise $(a, b)^{\emptyset} \in \mathcal{G}_{\mathcal{B}}$. Conversely, for all $a, b \in V$, if at least one of $a, b \in V_{\bigcirc}$, then we perform a Lasso regression with response y_a and design matrix D_b. We obtain the design matrix D_b as described above. In this case we get q Lasso estimates, one for each level of X, i.e. $\hat{\psi}_{ab}(0), \hat{\psi}_{ab}(1), \ldots, \hat{\psi}_{ab}(q)$. If $\hat{\psi}_{ab}(x) = 0$ for all $x \in \mathcal{X}$, then $(a, b)^{\mathcal{X}} \in \mathcal{E}_{\mathcal{B}}$; if $\hat{\psi}_{ab}(x) \neq 0$ for all $x \in \mathcal{X}$, then $(a, b)^{\emptyset} \in \mathcal{E}_{\mathcal{B}}$; if $\hat{\psi}_{ab}(x) = 0$ for all $x \in \mathcal{Z}$ while $\hat{\psi}_{ab}(x) \neq 0$ for all $x \in \mathcal{X} \setminus \mathcal{Z}$, with $\mathcal{Z} \subset \mathcal{X}$, $\mathcal{Z} \neq \emptyset$, then $(a, b)^{\mathcal{Z}} \in \mathcal{E}_{\mathcal{B}}$.

4. We use the AND or OR strategy to identify the non-zero estimates.

Following the same rational of other neighbor selection procedures (Meinshausen and Bühlmann, 2006), in step 3 of our algorithms we select edges based on the equivalence of the following zero constrains $\phi_{ab}(x) = 0 \iff \lambda_{ab}(x) = 0$ and $\psi_{ab}(x) = 0 \iff \omega_{ab}(x) = 0$ for profile undirected and bi-directed graphs, respectively. Similarly, in step 1 we select arrows based on the equivalence of the following zero constrains $\phi_{ax} = 0 \iff \gamma_{ax} = 0$ (undirected graphs) and on the equivalence, by definition, of the following parameters $\psi_{ax} \equiv \beta_{ax}$ (bi-directed graphs).

References

Andersson, S. A., D. Madigan, and M. D. Perlman (2001). Alternative markov properties for chain graphs. *Scandinavian journal of statistics 28*(1), 33–85.

Bhadra, A. and B. K. Mallick (2013). Joint high-dimensional bayesian variable and covariance selection with an application to eqtl analysis. *Biometrics 69*(2), 447–457.

Cai, T. T., H. Li, W. Liu, and J. Xie (2012). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, ass058.

Chen, M., Z. Ren, H. Zhao, and H. Zhou (2016). Asymptotically normal and efficient estimation of covariate-adjusted gaussian graphical model. *Journal of the American Statistical Association 111*(513), 394–406.

Consonni, G., L. La Rocca, and S. Peluso (2017). Objective bayes covariate-adjusted sparse graphical model selection. *Scandinavian Journal of Statistics 44*(3), 741–764.

Corander, J. (2003). Labelled graphical models. *Scandinavian Journal of Statistics 30*(3), 493–508.

Cox, D. and N. Wermuth (2003). A general condition for avoiding effect reversal after marginalization. *Journal of the Royal Statistical Society, Series B 65*, 937–941.

Danaher, P., P. Wang, and D. Witten (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B 76*, 373–397.

Drton, M. (2009). Discrete chain graph models. *Bernoulli 15*(3), 736–753.

Drton, M. and T. Richardson (2008). Binary models for marginal independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(2), 287–309.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Frydenberg, M. (1990). The chain graph markov property. *Scandinavian Journal of Statistics*, 333–353.

Guo, J., E. Levina, G. Michailidis, and J. Zhu (2011). Joint estimation of multiple graphical models. *Biometrika 98*(1), 1–15.

Hojsgaard, S. (2003). Split models for contingency tables. *Computational Statistics & Data Analysis 42*(4), 621–645.

Kanehisa, M., S. Goto, Y. Sato, M. Furumichi, and M. Tanabe (2011). Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research 40*(D1), D109–D114.

Kornblau, S. M., R. Tibes, Y. H. Qiu, W. Chen, H. M. Kantarjian, and M. Andreeff (2009). Functional proteomic profiling of aml predicts response and survival. *Blood 113*(1), 154–164.

Lauritzen, S. L. (1996). *Graphical Models*. New York: Oxford Univ. Press.

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

Lee, W. and Y. Liu (2012). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood. *Journal of Multivariate Analysis 111*, 241–255.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*(3), 1436–1462.

Nyman, H., J. Pensar, T. Koski, and J. Corander (2014). Stratified graphical models - context-specific independence in graphical models. *Bayesian Analysis 9*(4), 883–908.

Nyman, H., J. Pensar, T. Koski, and J. Corander (2016). Context-specific independence in graphical log-linear models. *Computational Statistics 31*(4), 1493–1512.

Peterson, C. B., F. Stingo, and M. Vannucci (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association 110*(509), 159–174.

Rothman, A. J., E. Levina, and J. Zhu (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics 19*(4), 947–962.

Roverato, A., M. Lupparelli, and L. La Rocca (2013). Log-mean linear models for binary data. *Biometrika 100*(2), 485–494.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B 13*, 238–241.

Wermuth, N. and D. R. Cox (2004). Joint response graphs and separation induced by triangular systems. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 66*(3), 687–717.

Wermuth, N. and K. Sadeghi (2012). Sequences of regressions and their independences. *TEST 21*(2), 215–252.

Yin, J. and H. Li (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics 5*(4), 2630.

Chapter 3

Bayesian model selection of multiple Ising undirected graphs

Contents

List of Tables

List of Figures

Abstract

In this manuscript, we propose two Bayesian approaches for the selection of log-linear models associated to multiple Ising graphs. We devise a Bayesian exact-likelihood inference for low-dimensional binary response data, based on conjugate priors for log-linear parameters, where we implement a computational strategy that uses Laplace approximations and a Metropolis-Hastings algorithm that allows us to perform a stochastic model search. We also propose a quasi-likelihood Bayesian approach for fitting high-dimensional Ising multiple graphs, where the normalization constant results computationally intractable, with spike-and-slab priors to encode sparsity and MCMC algorithms for sampling from the quasi-posterior distribution which enables variable selection and estimation simultaneously. In both methods, we define a Markov Random Field prior on the graph structures, which encourages the selection of the same edges in related graphs. We finally perform simulation studies to compare the proposed approaches with competing methods.

1 Introduction

The Ising model (Ising, 1925) is a type of graphical model for binary vectors. Graphical models are effective tools to model the joint distribution of a set of variables and graphically represent their conditional independence structure (Lauritzen, 1996). Applications of Ising models include the work of Banerjee et al. (2008) where associations between US senators are founded from their binary voting records. Ballout and Viallon (2019) present an application of these models to study associations among the injuries suffered by victims of road accidents according to road user type. In many applications it is more realistic to consider a collection of graphical models, due to the heterogeneity of the data involved, where the dependence structure of the variables may differ with respect to one or more factors. An example can be found in gene networks describing different subtypes of the same cancer: there are some shared pathways across different subtypes, and there are also links that are unique to a particular subtype. The purpose of this work is to develope Bayesian methodologies to select multiple related Ising undirected graphical models. Some approaches for inferring both Gaussian and discrete graphical models for two or more sample groups have been proposed in recent years (Guo et al., 2011; Peterson et al., 2015; Ballout and Viallon, 2019). Inference for Ising models is particularly challenging. For the case of a single discrete graph with a low number of variables, Massam et al. (2009) propose to use conjugate priors for log-linear parameters with a computational strategy that uses Laplace approximations. In this framework, Dobra and Massam (2010) devise MCMC methods for a stochastic search of the best model. Unfortunatelly, these methods do not scale well with the number of variables; specifically, these methods require to perform numerical approximations of the normalizing constant, and these calculations become unfeasable as soon as the number of variables included in the model is greater than about 10. The lack of a closed form of the normalizing constant implies that maximum likelihood estimation can generally not be performed. Various solutions have arisen in both the frequentist and Bayesian literature. In the frequentist literature there is a long history of fitting discrete graphical models using (quasi/pseudo)-likelihood methods instead of the exact-likelihood (Besag, 1974) . We find several examples of quasi-likelihood approaches in the frequentist literature when dealing with large graphical models (Meinshausen and Bühlmann, 2006; Ravikumar et al., 2010; Guo, 2015). In the Bayesian framework, the idea of using a non-likelihood function to carry out inference

is currently in growing popularity (Jiang and Tanner, 2008; Kato, 2013; Bhattacharyya and Atchade, 2019). In this manuscript, we propose two Bayesian approaches for the selection of log-linear models associated to multiple Ising graphs. Following Massam et al. (2009), we devise a Bayesian exact-likelihood inference for low-dimensional binary response data, based on conjugate priors for log-linear parameters, where we implement a computational strategy that uses Laplace approximations and a Metropolis-Hastings algorithm that allows us to perform a stochastic model search. We also propose a quasi-likelihood Bayesian approach, extending the work of Bhattacharyya and Atchade (2019), for fitting high-dimensional Ising multiple graphs, where the normalization constant results computationally intractable, with spike-and-slab priors to encode sparsity and MCMC algorithms for sampling from the quasi-posterior distribution which enables variable selection and estimation simultaneously. In both methods, we define a Markov Random Field prior on the graph structures, which encourages the selection of the same edges in related graphs (Peterson et al., 2015). In ongoing simulation studies we compared the proposed approaches with the competing methods Indep-SepLogit (Meinshausen and Bühlmann, 2006) and DataShared-SepLogit (Ollier and Viallon, 2017). We also compared our methods with the same ones using identical and independent Bernoulli distributions for the prior distribution of the model, as in Bhattacharyya and Atchade (2019). Overall the proposed approaches perform comparatively well; moreover, as unique features, these learn which groups are related and provide measure of uncertainty for model selection and parameter inference.

2 Background

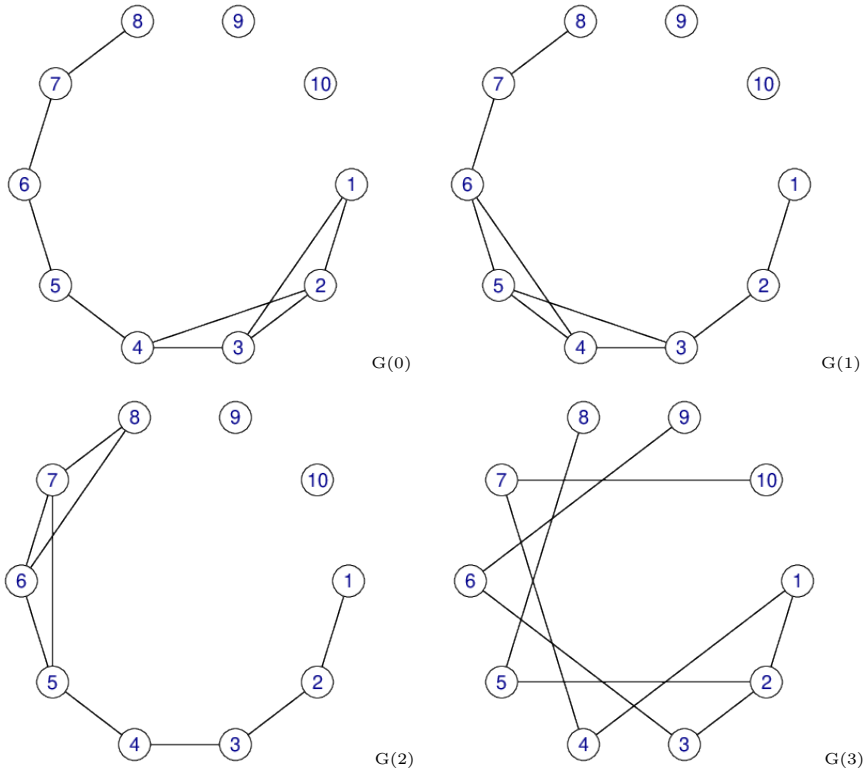
In the following subsection 2.1 we introduce some background materials on the multiple undirected graphs. In subsection 2.2 we show the likelihood and quasi-likelihood function of an Ising model.

2.1 Multiple undirected graphs

A graphical Markov model is a statistical model defined over a graph whose vertices correspond to random variables (Lauritzen, 1996; Edwards, 2000). The missing edges of the graph are translated into conditional independence restrictions that the model imposes on the joint distribution of the variables. We consider a collection of graphical models, where

the dependence structure of the variables may differ with respect to one or more factors (Guo et al., 2011; Peterson et al., 2015). More formally, let Y_V be the random vector corresponding to V, a set of p binary attributes, and X be the random variable corresponding to a categorical factor external to V, taking value $x \in \mathcal{X}$, with $|\mathcal{X}| = q$. We consider a collection of *multiple undirected graphs* $G_{V|\mathcal{X}} = [G(x)]_{x\in\mathcal{X}}$, where each graph $G(x) = (V, E(x))$ is associated to the random vector $Y_V|\{X = x\}$, where V is the *node set* and $E(x)$ is the *edge set* which depends on x, $x \in \mathcal{X}$. For any couple $r, j \in V$ and $x \in \mathcal{X}$, if $(r, j) \in E(x)$ then we have an edge between r and j in the corresponding graph $G(x)$ while if $(r, j) \notin E(x)$ then the two nodes are disjoined in $G(x)$. Missing edges in the graph correspond to conditional independencies for the associated joint probability distribution (Lauritzen, 1996); for instance, for the *pairwise Markov property*, if $(r, j) \notin E(x)$ then $Y_r \perp\!\!\!\perp Y_j|\{Y_{V\setminus\{r,j\}}, X = x\}$. We report an example of multiple undirected graphs in Figure 2.1 with $p = 10$ nodes and $q = 4$ levels of X, $\mathcal{X} = \{0, 1, 2, 3\}$, where, for instance, $Y_1 \perp\!\!\!\perp Y_{10}|\{Y_{V\setminus\{1,10\}}, X = x\}$ for all $x \in \mathcal{X}$ while $Y_1 \perp\!\!\!\perp Y_3|\{Y_{V\setminus\{1,3\}}, X = x\}$ for any $x \in \{1, 2, 3\}$.

Figure 1: A collection of multiple undirected graphs for $p = 10$ variables and $q = 4$ levels of X, with $\mathcal{X} = \{0, 1, 2, 3\}$.



2.2 The Ising model

We consider the case of binary random variables response data. An issue about modelling such data is that, as the number of the variables increases, the number of parameters can become so large to be intractable. A possible solution to this problem is to assume the Ising model (Besag, 1974), which implies a simplification in terms of number of possible non-zero parameters in the model. Indeed, in this model all higher than two-way interaction parameters vanish. So, let's assume we have observed $n^{(x)}$ realizations of $Y_V|\{X = x\} \sim \text{Ising}(\lambda^{(x)})$ where $\lambda^{(x)} = [\lambda^{(x)}_{rj}]_{r,j\in V} \in \mathbb{R}^{p+(p\times(p-1))/2}$, for all $x \in \mathcal{X}$. Let $\lambda^{(x)}_r = [\lambda^{(x)}_{rj}]_{j\in V}$ denotes the r-th vector of log-linear parameters, for any $r \in V$. Let $Z^{(x)}$ be the corresponding $n^{(x)} \times p$ observed binary matrix, with i-th row $z^{i(x)} \in (0, 1)^p$ and entries $z^{i(x)}_r \in (0, 1)$.

The likelihood function of $\lambda^{(x)}$ can be expressed as

$$p(Z^{(x)}|\lambda^{(x)}) = \prod_{i=1}^{n^{(x)}} \frac{1}{\Psi(\lambda^{(x)})} \exp\Big\{ \sum_{r=1}^{p} \lambda^{(x)}_{rr} z^{i(x)}_r + \sum_{r=1}^{p} \sum_{j<r} \lambda^{(x)}_{rj} z^{i(x)}_r z^{i(x)}_j \Big\}, \quad x \in \mathcal{X}, \tag{2.1}$$

where

$$\Psi(\lambda^{(x)}) = \sum_{z^{i(x)}\in\{0,1\}^p} \exp\Big\{ \sum_{r=1}^{p} \lambda^{(x)}_{rr} z^{i(x)}_r + \sum_{r=1}^{p} \sum_{j<r} \lambda^{(x)}_{rj} z^{i(x)}_r z^{i(x)}_j \Big\} \tag{2.2}$$

is the normalization constant. In a high-dimensional setting, likelihood based inference on $\lambda^{(x)}$ is computationally intractable because $\Psi(\lambda^{(x)})$ is hard to calculate, since it requires to compute a sum that is exponential in p and quickly blows up for even moderate values of p. In this situation a possible strategy relies in using a *quasi-likelihood* approach (Bhattacharyya and Atchade, 2019).

We express the r-th node conditional likelihood for $\lambda^{(x)}_r \in \mathbb{R}^p$, $r \in V$, as

$$p_r(Z^{(x)}|\lambda^{(x)}_r) = \prod_{i=1}^{n^{(x)}} \frac{1}{\Psi^i_r(\lambda^{(x)}_r)} \exp\Big\{ \lambda^{(x)}_{rr} z^{i(x)}_r + \sum_{j<r} \lambda^{(x)}_{rj} z^{i(x)}_r z^{i(x)}_j \Big\}, \quad x \in \mathcal{X}, \tag{2.3}$$

where in this case the normalization constant is

$$\Psi^i_r(\lambda^{(x)}_r) = 1 + \exp\Big\{ \lambda^{(x)}_{rr} + \sum_{j<r} \lambda^{(x)}_{rj} z^{i(x)}_j \Big\}, \tag{2.4}$$

We approximate the likelihood in (2.1) with a quasi-likelihood obtained as the product of p conditional likelihoods, i.e.

$$p_q(Z^{(x)}|\lambda^{(x)}) = \prod_{r=1}^{p} p_r(Z^{(x)}|\lambda^{(x)}_r), \quad x \in \mathcal{X}, \tag{2.5}$$

and so the inference problem on $\lambda^{(x)} \in \mathbb{R}^{p+(p\times(p-1))/2}$ simplifies into p separable subproblems on $\mathbb{R}^p$.

3 The prior

3.1 Linking-graphs prior

Our aim is to select the best set of graphs $G_{V|\mathcal{X}}$ taking into account the possible similarities among them. Note that, for all $r, j \in V$ and for all $x \in \mathcal{X}$, setting to zero $\lambda_{rj}^{(x)}$ correspond to the missing edge (r, j) wrt $G(x)$. Therefore we want to estimate $\lambda^{(x)}$ for all levels $x \in \mathcal{X}$, considering the similarities among these graphs. We encourage the selection of the same edges in related graphs with a Markov Random Field (MRF) prior on the graph structures (Peterson et al., 2015). The MRF replaces indicators of variable inclusion with indicators of edge inclusion. We introduce for each parameter $\lambda_{rj}^{(x)} \in \mathbb{R}$, for all $r, j \in V$, a selection parameter $\delta_{rj}^{(x)} \in (0, 1)$, $x \in \mathcal{X}$. The conditional probability of the inclusion of edge (r, j) in $G(x)$, given the inclusion of edge (r, j) in the remaining graphs $[G(h)]_{h\in\{\mathcal{X}\setminus x\}}$, for any $r, j \in V$, is

$$\pi(\delta_{rj}^{(x)}|\delta_{rj}^{(-x)}, \nu_{rj}, \theta^{(x)}) = \frac{\exp[\delta_{rj}^{(x)}(\nu_{rj} + \mathbf{1}^T\theta^{(x)}\delta_{rj}^{(-x)})]}{1 + \exp[\delta_{rj}^{(x)}(\nu_{rj} + \mathbf{1}^T\theta^{(x)}\delta_{rj}^{(-x)})]}, \quad x \in \mathcal{X}, \tag{3.1}$$

where $\delta_{rj}^{(-x)} = [\delta_{rj}^{(h)}]_{h\in\{\mathcal{X}\setminus x\}}$, $\nu_{rj} \in \mathbb{R}$ is a sparsity parameter specific for the edge (r, j) and $\theta^{(x)} = [\theta^{(x,h)}]_{h\in\{\mathcal{X}\setminus x\}}$, where $\theta^{(x,h)} \in \mathbb{R}$ is a linking-graphs parameter, representing the relatedness between the graphs $G(x)$ and $G(h)$, for all $x, h \in \mathcal{X}, x \neq h$.

We write the conditional probability of the r-vector $\delta_r^{(x)} = [\delta_{rj}^{(x)}]_{j\in V}$, for any $r \in V$, as

$$\pi(\delta_r^{(x)}|\delta_r^{(-x)}, \nu_r, \theta^{(x)}) = \prod_{j=1}^{p} \pi(\delta_{rj}^{(x)}|\delta_{rj}^{(-x)}, \nu_{rj}, \theta^{(x)}), \quad x \in \mathcal{X}, \tag{3.2}$$

where $\delta_r^{(-x)} = [\delta_r^{(h)}]_{h\in\{\mathcal{X}\setminus x\}}$ and $\nu_r = [\nu_{rj}]_{j\in V}$.

We also write the conditional probability of the entire vector $\delta^{(x)} = [\delta_{rj}^{(x)}]_{r,j\in V, j<r}$ as

$$\pi(\delta^{(x)}|\delta^{(-x)}, \nu, \theta^{(x)}) = \prod_{r=1}^{p}\prod_{j<r} \pi(\delta_{rj}^{(x)}|\delta_{rj}^{(-x)}, \nu_{rj}, \theta^{(x)}), \quad x \in \mathcal{X}, \tag{3.3}$$

where $\delta^{(-x)} = [\delta^{(h)}]_{h\in\{\mathcal{X}\setminus x\}}$ and $\nu = [\nu_{rj}]_{r,j\in V, j<r}$.

3.1.1 Prior of $\theta^{(x,h)}$

Following Peterson et al. (2015), we have previously introduced the graphs-linking parameter $\theta^{(x,h)}$ where its magnitude measures the pairwise similarity between graphs $G(x)$ and

$G(h)$, for any $x, h \in \mathcal{X}$, such that if $\theta^{(x,h)} = 0$ then the two graphs $G(x)$ and $G(h)$ are independent. The linking-graphs parameter $\theta^{(x,h)}$, for any $x, h \in \mathcal{X}$, is learned from the data. We place a spike-and-slab prior on $\theta^{(x,h)}$, for any $x, h \in \mathcal{X}$ (George and McCulloch, 1997). Since the probability density function Gamma$(x|\alpha, \beta)$ is equal to zero at the point $x = 0$ and is nonzero on the domain $x > 0$, an appropriate choice for the "slab" portion of the mixture prior is the Gamma$(x|\alpha, \beta)$ density. We formalize our prior by using a latent indicator variable $\epsilon^{(x,h)}$ to represent the event that graphs x and h are related, for all $x, h \in \mathcal{X}$. The mixture prior on $\theta^{(x,h)}$ can then be written in terms of the latent indicator as

$$\pi(\theta^{(x,h)}|\epsilon^{(x,h)}) = (1 - \epsilon^{(x,h)})d_0 + \epsilon^{(x,h)} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{(x,h)^{\alpha-1}} e^{-\beta\theta^{(x,h)}}, \quad x, h \in \mathcal{X}, \tag{3.4}$$

where $\Gamma(\cdot)$ represents the Gamma function and α and β are fixed hyperparameters. The joint prior for $\theta = [\theta^{(x,h)}]_{x<h}$ given $\epsilon = [\epsilon^{(x,h)}]_{x<h}$ can be written as the product of the marginal densities of any $\theta^{(x,h)}$, because the parameters are variation independent such that

$$\pi(\theta|\epsilon) = \prod_{x<h} \pi(\theta^{(x,h)}|\epsilon^{(x,h)}). \tag{3.5}$$

We place independent Bernoulli priors on the latent indicators

$$\pi(\epsilon^{(x,h)}) = w^{\epsilon^{(x,h)}}(1 - w)^{(1-\epsilon^{(x,h)})} \tag{3.6}$$

3.1.2 Prior of ν_{rj}

We use the edge-specific parameter $\nu_{rj} \in \mathbb{R}$, for all $r, j \in V$, to give sparsity, indeed a negative value of ν_{rj} reduces the prior probability of the inclusion of edge (r, j) in any graph of $G_{V|\mathcal{X}}$. The probability of inclusion of edge (r, j), for all $r, j \in V$, in $G(x)$, for all $x \in X$, can be written as

$$\pi(\delta_{rj}^{(x)}|\nu_{rj}) = \frac{e^{\nu_{rj}}}{1 + e^{\nu_{rj}}} = q_{rj}. \tag{3.7}$$

In cases where no prior knowledge on the graph structure is available, a prior that favors lower values, such as $q_{rj} \sim$ Beta(a, b) with $a < b$ for all edges (r, j), can be chosen to encourage overall sparsity. This determines a prior on ν_{rj} since $\nu_{rj} =$ logit(q_{rj}). After applying a univariate transformation of variables to the Beta(a, b) prior on q_{rj}, the prior on

ν_{rj}, for all $r, j \in V$, can be written as

$$\pi(\nu_{rj}) = \frac{1}{B(a,b)} \frac{e^{a\nu_{rj}}}{(1+e^{\nu_{rj}})^{a+b}}, \tag{3.8}$$

where $B(\cdot)$ represents the beta function.

3.2 Prior of $\lambda^{(x)}$

We follow different inference approaches depending on the dimension of p, with different priors for $\lambda^{(x)}$, $x \in \mathcal{X}$. We denote with ${}_1\lambda^{(x)} = [{}_1\lambda^{(x)}_{rj}]_{r,j\in V, j\leq r}$ the sparse vector with elements ${}_1\lambda^{(x)}_{rj} = \lambda^{(x)}_{rj}$ if $\delta^{(x)}_{rj} = 1$, else ${}_1\lambda^{(x)}_{rj} = 0$ if $\delta^{(x)}_{rj} = 0$. In the same way let ${}_0\lambda^{(x)} = [{}_0\lambda^{(x)}_{rj}]_{r,j\in V, j\leq r}$ be the vector with elements ${}_0\lambda^{(x)}_{rj} = \lambda^{(x)}_{rj}$ if $\delta^{(x)}_{rj} = 0$, else ${}_0\lambda^{(x)}_{rj} = 0$ if $\delta^{(x)}_{rj} = 1$, such that ${}_0\lambda^{(x)}_r +_1 \lambda^{(x)}_r = \lambda^{(x)}_r$, $x \in \mathcal{X}$.

3.2.1 Low-dimensional case ($p \leq 10$)

We firstly present our prior choice for a marginal-likelihood based approach in a low-dimensional framework. Let $\iota^{(x)}$ be the ι-th cell of the contingency table $\mathcal{I}^{(x)} = \times(0,1)$ for $Y_V|\{X = x\}$, with $\iota \in \mathcal{I}^{(x)}$ and $x \in \mathcal{X}$. We denote with $\iota^{(x)}_\emptyset$ the baseline cell. We denote with $y^{(x)} = [y^{(x)}_{rj}]_{r,j\in V}$, where $y^{(x)}_{rj} = \sum_{i=1}^{n^{(x)}} \mathbb{1}_{\{z^{i(x)}_r = x, z^{i(x)}_j = x\}}$, the vector of marginal counts for (Y_r, Y_j), $r, j \in V$. Consider the case of p small ($p \leq 10$); we choose the Diaconis and Ylvisaker prior distribution (Diaconis and Ylvisaker, 1979), that is conjiugate for $\lambda^{(x)}$ (Massam et al., 2009). We rewrite the likelihood (2.1) in the sparsified version and as function of the marginal counts

$$\begin{aligned} p({}_1\lambda^{(x)}|y^{(x)}) = \exp\Big\{ &\sum_{r=1}^{p} {}_1\lambda^{(x)}_{rr} y^{(x)}_{rr} + \sum_{r=1}^{p}\sum_{j<r} {}_1\lambda^{(x)}_{rj} y^{(x)}_{rj} \\ &- n^{(x)} \log\Big[\sum_{\{\mathcal{I}^{(x)} \setminus \iota^{(x)}_\emptyset\}} \exp\Big(\sum_{r=1}^{p} {}_1\lambda^{(x)}_{rr} + \sum_{r=1}^{p}\sum_{j<r} {}_1\lambda^{(x)}_{rj} \Big)\Big]\Big\}, \quad x \in \mathcal{X}. \end{aligned} \tag{3.9}$$

The Diaconis and Ylvisaker prior distribution for (3.9) is given by

$$\begin{aligned} \pi({}_1\lambda^{(x)}|s^{(x)}, g^{(x)}) = &\, C(s^{(x)}, \alpha^{(x)})^{-1} \\ &\times \exp\Big\{ \sum_{r=1}^{p} {}_1\lambda^{(x)}_{rr} s^{(x)}_{rr} + \sum_{r=1}^{p}\sum_{j<r} {}_1\lambda^{(x)}_{rj} s^{(x)}_{rj} \\ &- g^{(x)} \log\Big[\sum_{\{\mathcal{I}^{(x)} \setminus \iota^{(x)}_\emptyset\}} \exp\Big(\sum_{r=1}^{p} {}_1\lambda^{(x)}_{rr} + \sum_{r=1}^{p}\sum_{j<r} {}_1\lambda^{(x)}_{rj} \Big)\Big]\Big\}, \quad x \in \mathcal{X}, \end{aligned} \tag{3.10}$$

where $C(s^{(x)}, g^{(x)})$ is an uknown normalization constant that depends on $g^{(x)} \in \mathbb{R}$ and $s^{(x)} = [s_{rj}^{(x)}]_{r,j \in V}$, $s^{(x)} \in \mathbb{R}^{p+(p\times(p-1))/2}$, which are the hyperparameters.

3.2.2 High-dimensional case ($p > 10$)

When the number of variables p is large ($p > 10$), we follow a Bayesian quasi-likelihood approach to make inference on $\lambda_r^{(x)}$, for all $r \in V$. In this case, we focus on the r-th node conditional likelihood (2.3), for all $r \in V$. We use a relaxed form of the spike and slab prior for $\lambda_r^{(x)}$, such that $\lambda_{rj}^{(x)}|(\delta_{rj}^{(x)} = 1, \rho) \sim \mathrm{N}(0, \rho), \rho > 0$ and $\lambda_{rj}^{(x)}|(\delta_{rj}^{(x)} = 0, \gamma) \sim \mathrm{N}(0, \gamma), \gamma > 0$, $x \in \mathcal{X}$.

The conditional distribution of $\lambda_r^{(x)}$ given $\delta_r^{(x)}$ is given by

$$\pi(\lambda_r^{(x)}|\delta_r^{(x)}, \rho, \gamma) \propto (2\pi\rho)^{\frac{-||\delta_r^{(x)}||}{2}} \exp\Big(-\frac{||_1\lambda_r^{(x)}||_2^2}{2\rho}\Big) \times (2\pi\gamma)^{\frac{-||1-\delta_r^{(x)}||}{2}} \exp\Big(-\frac{||_0\lambda_r^{(x)}||_2^2}{2\gamma}\Big) \tag{3.11}$$

where $||\cdot||$ and $||\cdot||_2$ denote the L1 and L2 norm respectively.

4 Posterior inference

The goal is to select the graph with higher posterior probability taking into account the possible relatedness of the graphs in $G_{V|\mathcal{X}}$. In subsection 4.1 we follow a Bayesian exact-likelihood approach for low-dimensional cases, where we compute the marginal likelihood through the Laplace approximation. In subsection 4.1 we deal with high-dimensional cases, following a Bayesian approximate-likelihood approach that uses MCMC methods to sample from the quasi-posterior distribution.

4.1 Low-dimensional case ($p \leq 10$)

In the low-dimensional setting the posterior inference is based on the computation of the marginal likelihood. For any $x \in \mathcal{X}$, let $G(x)$ be the graph selected from a set of competing graphs $\Omega(x)$; we denote with $\pi(G(x))$ and $\Pi(G(x)|Z^{(x)})$ the prior and the posterior probability of $G(x)$, $x \in \mathcal{X}$. The posterior probability of $G(x)$ is proportional to the product of the prior distribution $\pi(G(x))$ and the marginal likelihood $m(Z|G(x))$, i.e.

$$\Pi(G(x)|Z) \propto \pi(G(x))\ m(Z|G(x)), \quad x \in \mathcal{X}, \tag{4.1}$$

where

$$m(Z|G(x)) = \int_{\Theta_{G(x)}} \pi({}_1\lambda^{(x)}|s^{(x)}, g^{(x)})\ p({}_1\lambda^{(x)}|y^{(x)})\ d_1\lambda^{(x)} \tag{4.2}$$

and

$$\pi(G(x)) = \pi(\delta^{(x)}|\delta^{(-x)}, \nu, \theta^{(x)}). \tag{4.3}$$

We combine the prior in (3.10) with the likelihood in (3.9) to obtain the posterior of $\lambda^{(x)}$, i.e.

$$\begin{aligned}\Pi({}_1\lambda^{(x)}|y^{(x)}) = {} & C(y^{(x)} + s^{(x)}, n^{(x)} + g^{(x)})^{-1} \\ & \times \exp\Big\{ \sum_{r=1}^{p} {}_1\lambda_{rr}^{(x)}(s_{rr}^{(x)} + y_{rr}^{(x)}) + \sum_{r=1}^{p}\sum_{j<r} {}_1\lambda_{rj}^{(x)}(s_{rj}^{(x)} + y_{rj}^{(x)}) \\ & - (g^{(x)} + n^{(x)}) \log\Big[\sum_{\{\mathcal{I}^{(x)} \setminus \iota_{\emptyset}^{(x)}\}} \exp\Big(\sum_{r=1}^{p} {}_1\lambda_{rr}^{(x)} + \sum_{r=1}^{p}\sum_{j<r} {}_1\lambda_{rj}^{(x)} \Big) \Big] \Big\}, \qquad x \in \mathcal{X},\end{aligned} \tag{4.4}$$

and in this way the integral in (4.2) is analitically derived as

$$\frac{C(y^{(x)} + s^{(x)}, n^{(x)} + g^{(x)})}{C(s^{(x)}, g^{(x)})}, \qquad x \in \mathcal{X}, \tag{4.5}$$

the ratio between the normalising constants of the posterior and prior distributions of $\lambda^{(x)}$ (Massam et al., 2009). We calculate both the normalizing constants through the Laplace approximation (Tierney and Kadane, 1986) such that, for instance

$$C(s^{(x)}, g^{(x)}) = Ke({}_1\lambda^{(x)*}) \frac{(2\pi)^{||\delta^{(x)}||/2}}{|A^{(x)}|^{1/2}}, \qquad x \in \mathcal{X}, \tag{4.6}$$

where $Ke({}_1\lambda^{(x)*})$ is the kernel of the prior (3.10) and $A^{(x)}$ is the Hessian matrix ($||\delta^{(x)}|| \times ||\delta^{(x)}||$), both evaluated in a stationary point ${}_1\lambda^{(x)^*}$.

4.2 High-dimensional case ($p > 10$)

In the high-dimensional setting, we follow Bhattacharyya and Atchade (2019). We combine the prior distribution in (3.3) together with the r-th conditional likelihood in (2.3) and we obtain the r-th posterior distribution of $(\delta_r^{(x)}, \lambda_r^{(x)})$ given by

$$\Pi(\delta_r^{(x)}, \lambda_r^{(x)}|Z^{(x)}) \propto \pi(\delta_r^{(x)}|\delta_r^{(-x)}, \nu_r, \theta^{(x)}) \times \Big(\frac{\gamma}{\rho}\Big)^{\frac{||\delta_r^{(x)}||}{2}} \exp\Big(p_r(Z^{(x)}|{}_1\lambda^{(x)}) - \frac{||{}_1\lambda_r^{(x)}||_2^2}{2\rho} - \frac{||{}_0\lambda_r^{(x)}||_2^2}{2\gamma} \Big) \tag{4.7}$$

such that the quasi-posterior of $(\delta^{(x)}, \lambda^{(x)})$ for the graph $G(x)$, for all $x \in \mathcal{X}$, is given by

$$\Pi_q(\delta^{(x)}, \lambda^{(x)}|Z^{(x)}) = \prod_{r=1}^{p} \Pi(\delta_r^{(x)}, \lambda_r^{(x)}|Z^{(x)}) \tag{4.8}$$

Note that in this case for any $\lambda_{rj}^{(x)}$, $r, j \in V, r < j$, we obtain two estimates $\hat{\lambda}_{rj}^{(x)}$ and $\hat{\lambda}_{jr}^{(x)}$. We use a post-estimation symmetrization resulting in a singular estimate by taking the simple mean of the two estimates.

5 Posterior computation

In this section, we present the MCMC methods used for the posterior inference. In the exact-likelihood method, we perform a stochastic search for the model with high posterior probability while in the approximate-likelihood case we sample from the quasi-posterior distribution to update the model parameter.

5.1 Updating of $G(x)$

5.1.1 Low-dimensional case $(p < 10)$

We firstly show the MCMC algorithm for the low-dimensional case. Since the size of the set of possible graphs $\Omega^{(x)}$ is too large to be explored entirely, we perform a stochastic model search. For $x \in \mathcal{X}$, we start from $G(x)^{(t-1)}$ the graph accepted at time $(t-1)$ and we propose a new graph $G^{(x)(t)}$ by randomly sampling one element of $\delta^{(x)(t-1)}$ and switching its value. We finally accept the new model $G^{(x)(t)}$ with probability

$$r = \min\left(1, \frac{\Pi(G(x)^{(t)}|Z)}{\Pi(G(x)^{(t-1)}|Z)}\right). \tag{5.1}$$

5.1.2 High-dimensional case $(p > 10)$

We now discuss the construction of the Markov Chain Monte Carlo (MCMC) algorithm to draw samples from the quasi-posterior distribution 4.8. In particular, we use a general Metropolis Adjusted Langevin Algorithm (MALA) which updates ${}_1\lambda_r^{(x)}, {}_0\lambda_r^{(x)}, \delta_r^{(x)}, \theta^{(x)}$ and ν_r respectively, for any $x \in \mathcal{X}$.

We define

$$h(\delta_r^{(x)}, \lambda_r^{(x)}|Z^{(x)}) = \left(p_r(Z^{(x)}|{}_1\lambda_r^{(x)}) - \frac{||{}_1\lambda_r^{(x)}||_2^2}{2\rho} - \frac{||{}_0\lambda_r^{(x)}||_2^2}{2\gamma}\right) \tag{5.2}$$

which has a gradient given by

$$\mathbb{G} = \nabla_{\lambda_r^{(x)}} \; h(\delta_r^{(x)}, \lambda_r^{(x)} | Z^{(x)}) = \nabla_{{}_1\lambda_r^{(x)}} \; p_r(Z^{(x)} | {}_1\lambda_r^{(x)}) - \frac{{}_1\lambda_r^{(x)}}{\rho} - \frac{{}_0\lambda_r^{(x)}}{\gamma} \tag{5.3}$$

For any j-th component of $\lambda_r^{(x)}$ such that $\delta_{rj}^{(x)} = 1$, we propose a new value

$$\lambda_{rj}^{(x)*} | \lambda_r^{(x)} \sim \mathrm{N}(\lambda_{rj}^{(x)} + \frac{\sigma}{2}\mathbb{G}_j, \sigma^2), \tag{5.4}$$

where σ is some constant step size and $\mathbb{G}_j$ represents the j-th component of the gradient. Let $f(\lambda_{rj}^{(x)*} | \lambda_r^{(x)})$ denote the density of the proposal distribution in 5.4. We also define $\lambda_r^{(x)*} = (\lambda_{r1}^{(x)}, \ldots, \lambda_{rj}^{(x)*}, \ldots, \lambda_{rp}^{(x)})$ and the acceptance probability as

$$\zeta_{rj} = \min\left(1, \frac{f(\lambda_{rj}^{(x)} | \lambda_r^{(x)*})}{f(\lambda_{rj}^{(x)*} | \lambda_r^{(x)})} \times \frac{\Pi(\delta_r^{(x)}, \lambda_r^{(x)*} | Z^{(x)})}{\Pi(\delta_r^{(x)}, \lambda_r^{(x)} | Z^{(x)})}\right), \tag{5.5}$$

such that we set $\lambda_{rj}^{(x)} = \lambda_{rj}^{(x)*}$ with probability ζ_{rj}.

Conversely, for any j-th component of $\lambda_r^{(x)}$ such that $\delta_{rj}^{(x)} = 0$, we update its value

$$\lambda_{rj}^{(x)} \sim \mathrm{N}(0, \gamma). \tag{5.6}$$

Finally, for each $j \in V$, we define $\bar{\delta}_r^{(x)} = (\delta_{r1}^{(x)}, \ldots, (1 - \delta_{rj}^{(x)}), \ldots, \delta_{rp}^{(x)})$ and set

$$\tau_{rj} = \min\left(1, \frac{\Pi(\bar{\delta}_r^{(x)}, \lambda_r^{(x)} | Z^x)}{\Pi(\delta_r^{(x)}, \lambda_r^{(x)} | Z^x)}\right); \tag{5.7}$$

we set $\delta_{rj} = 1 - \delta_{rj}$ with probability τ_{rj}.

5.2 Updating of $\theta^{(x,h)}$

Given the prior on $\theta^{(x,h)}$ from Equation (3.4) and the prior on $\epsilon^{(x,h)}$ from Equation (3.6), the posterior full conditional of $\theta^{(x,h)}$ and $\epsilon^{(x,h)}$ can be written as

$$\begin{aligned} \Pi(\theta^{(x,h)}, \epsilon^{(x,h)} | \cdot) \propto & \left(\prod_{r<j} C(\nu_{rj}, \theta)^{-1} \exp\{2\theta^{(x,h)} \delta_{rj}^{(x)} \delta_{rj}^{(h)}\}\right) \\ & \times \left((1 - \epsilon^{(x,h)}) d_0 + \epsilon^{(x,h)} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{(x,h)^{\alpha-1}} e^{-\beta\theta^{(x,h)}}\right) \\ & \times \left(w^{\epsilon^{(x,h)}} (1 - w)^{(1-\epsilon^{(x,h)})}\right) \end{aligned} \tag{5.8}$$

where the normalizing constant $C(\nu_{rj}, \theta) = \sum_{\delta_{rj} \in \{0,1\}^q} \exp\{\nu_{rj} \mathbf{1}^T \boldsymbol{\delta}_{rj} + \boldsymbol{\delta}_{rj}^T \boldsymbol{\theta} \boldsymbol{\delta}_{rj}\}$ with $\boldsymbol{\delta}_{rj} = [\delta_{rj}^{(x)}]_{x \in \mathcal{X}}$ and $\boldsymbol{\theta}$ the $(q \times q)$ symmetric matrix with entries $\theta^{(x,h)}$, for all $x, h \in \mathcal{X}$. Since

the normalizing constant is analytically intractable, we use Metropolis–Hastings steps to sample $\theta^{(x,h)}$ and $\epsilon^{(x,h)}$ from their joint posterior full conditional distribution for each pair $x, h \in \mathcal{X}$. At each iteration we perform two steps: a between-model and a within-model move; see Gottardo and Raftery (2008) for more details. For the between-model move, if in the current state $\epsilon^{(x,h)} = 1$, we propose $\epsilon^{(x,h)*} = 0$ and $\theta^{(x,h)*} = 0$. If in the current state $\epsilon^{(x,h)} = 0$, we propose $\epsilon^{(x,h)*} = 1$ and sample $\theta^{(x,h)*}$ from the proposal density $f(\theta^{(x,h)*}) = \text{Gamma}(\theta^{(x,h)*}|\alpha^*, \beta^*)$.

When moving from $\epsilon^{(x,h)} = 1$ to $\epsilon^{(x,h)*} = 0$, the Metropolis–Hastings ratio is

$$r = \frac{\Pi(\theta^{(x,h)*}, \epsilon^{(x,h)*}|\cdot) f(\theta^{(x,h)})}{\Pi(\theta^{(x,h)}, \epsilon^{(x,h)}|\cdot)}, \tag{5.9}$$

while if we move from $\epsilon^{(x,h)} = 0$ to $\epsilon^{(x,h)*} = 1$, the Metropolis–Hastings ratio is

$$r = \frac{\Pi(\theta^{(x,h)*}, \epsilon^{(x,h)*}|\cdot)}{\Pi(\theta^{(x,h)}, \epsilon^{(x,h)}|\cdot) f(\theta^{(x,h)*})}. \tag{5.10}$$

We then perform a within-model move whenever the value of $\epsilon^{(x,h)}$ sampled from the between-model move is 1. For this step, we propose a new value of $\theta^{(x,h)}$ using the same proposal density as before. The Metropolis–Hastings ratio for this step is

$$r = \frac{\Pi(\theta^{(x,h)*}, \epsilon^{(x,h)*}|\cdot) f(\theta^{(x,h)})}{\Pi(\theta^{(x,h)}, \epsilon^{(x,h)}|\cdot) f(\theta^{(x,h)*})}. \tag{5.11}$$

5.3 Updating of ν_{rj}

Given the prior from Equation (3.14), the posterior full conditional of ν_{rj} given the data and all remaining parameters is proportional to

$$\Pi(\nu_{rj}|\cdot) \propto \frac{\exp(a\nu_{rj})}{(1+e^{\nu_{rj}})^{a+b}} C(\nu_{rj}, \theta)^{-1} \exp(\nu_{rj} \mathbf{1}^T \boldsymbol{\delta}_{rj}). \tag{5.12}$$

For each pair $r, j \in V, r \neq j$, we propose a value q^* from the density Beta$(1, 2)$, then set $\nu^* = \text{logit}(q^*)$. The proposal density can be written in terms of ν^* as

$$f(\nu^*) = \frac{1}{B(a^*, b^*)} \frac{e^{a^*\nu^*}}{(1+e^{\nu_*})^{a^*+b^*}} \tag{5.13}$$

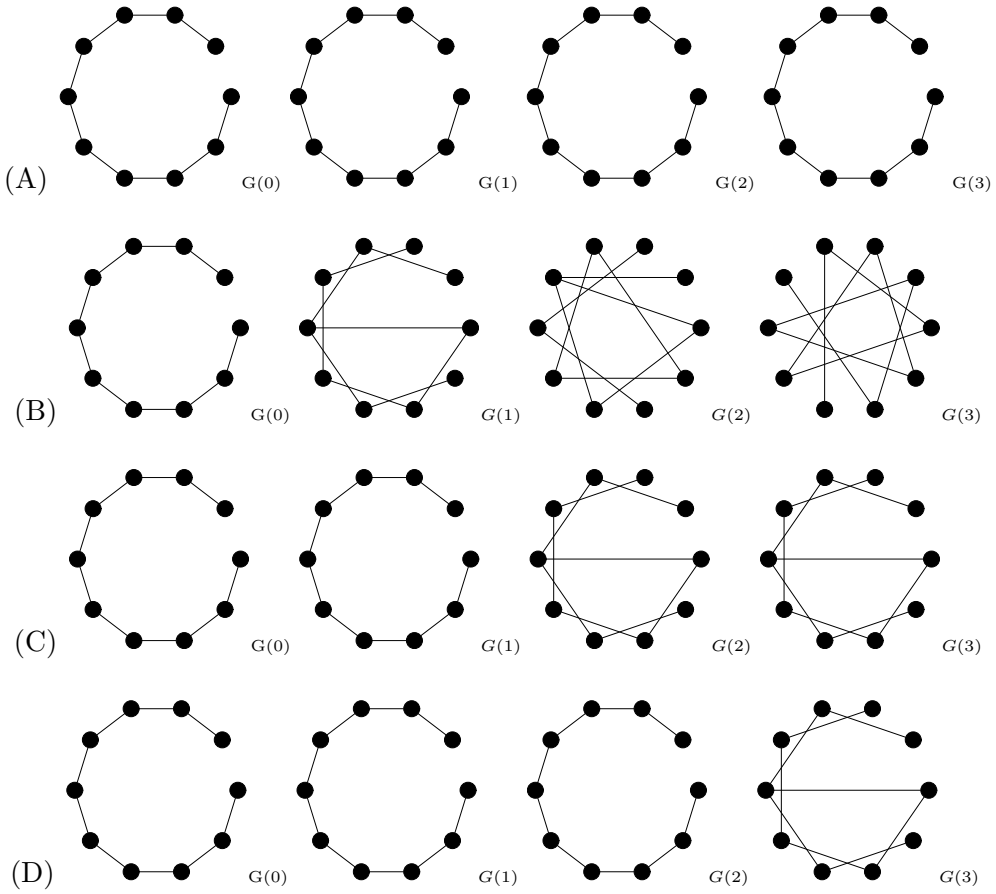
and the Metropolis–Hastings ratio is

$$r = \frac{\Pi(\nu^*|\cdot) f(\nu_{rj})}{\Pi(\nu_{rj}|\cdot) f(\nu^*)}. \tag{5.14}$$

6 Simulation studies

We empirically assess the two Bayesian approaches proposed above, that we call Bayesian Exact Linking (BEL) and Bayesian Approximate Linking (BAL), in the selection of different profile undirected graphs on simulated data. We compare our methods with the same ones using identical and independent Bernoulli distributions for the prior distribution of the model, as in Bhattacharyya and Atchade (2019), and we call them Bayesian Exact (BE) and Bayesian Approximate (BA). We also compare our Bayesian approaches with the frequentists Indep-Seplogit (SL) (Meinshausen and Bühlmann, 2006) and DataShared-SepLogit (DSSL) (Ollier and Viallon, 2017), that we implemented using the glmnet package (Friedman et al., 2010). Selection of tuning parameters was performed using the BIC. We generate the data from the Ising model using a Gibbs sampler. The R-code is available and located at `https://github.com/kinglaz90/phd`.

Figure 2: Multiple undirected graphs in the four different scenarios of simulation study I.



6.1 Simulation study I

6.1.1 Data generation

In simulation study I we assess our methods for the low-dimensional case. For any $x \in \mathcal{X}$, we sample $n^{(x)} = 100$ observations from $Y_V|\{X = x\} \sim \text{Ising}(\lambda^{(x)})$ with associated undirected graph $\mathcal{G}$ with $p = 10$ nodes and $q = 4$ levels of X, such that $x \in \mathcal{X} = \{0, 1, 2, 3\}$. For all $x \in \mathcal{X}$ and for all $r \in V$, we set $\lambda_{rr}^{(x)} = -1$. For all $x \in \mathcal{X}$ and for all $r, j \in V, j < r$, the non-zero interactions $\lambda_{rj}^{(x)}$ are set to 1.5. Note that $\mathcal{G}$ includes at most $p(p-1)/2 = 45$ edges that are identified by the selection parameter $\delta^{(x)}$, $x \in \mathcal{X}$. We consider 4 different profile undirected graphs $\mathcal{G}(A)$, $\mathcal{G}(B)$, $\mathcal{G}(C)$ and $\mathcal{G}(D)$. In Scenario (A) the four graphs are identical. In Scenario (B) the four graphs are completely different. In Scenario (C) $G(0)$ and $G(1)$ are identical but completely different to $G(2)$ and $G(3)$, which again are identical to each other. In Scenario (D) $G(0)$, $G(1)$ and $G(2)$ are identical and completely different to $G(3)$. We report the graphs of all the four scenarios in Figure 2.

6.1.2 Parameters setting

For the frequentists approaches SL and DSSL we select the penalization parameter using the BIC. Also, In DSSL method we set the parameter that controls the degree of sharing between the levels of X as $r = \frac{1}{\sqrt{q}}$, after having standardized the columns of the design matrix; see Ollier and Viallon (2017). In the linking-graphs prior we set the hyper-parameters $a = 1$, $b = 3$, $\alpha = 1$, $\beta = 2$ and $\omega = 0.6$. In the Bayesian exact-likelihood approaches we set $g^{(x)} = 0.02$ and the vector $s^{(x)}$ in such a way that the prior probability of each cell of the contingency table $\mathcal{I}^{(x)} = \times(0, 1)^p$ is equal to $g^{(x)}/|\mathcal{I}^{(x)}|$, for all $x \in \mathcal{X}$. In the Bayesian approximate-likelihood approaches we set $\rho = 2$ and $\gamma = 0.5$. Finally, for the approaches BE and BA, we set the prior of the model as the product of p Bernoulli(0.2).

6.2 Simulation study II

6.2.1 Data generation

In simulation study II we assess our methods for the high-dimensional case. For any $x \in \mathcal{X}$, we sample $n^{(x)} = 200$ observations from $Y_V|\{X = x\} \sim \text{Ising}(\lambda^{(x)})$ with associated undirected graph $\mathcal{G}$ with $p = 50$ nodes and $q = 4$ levels of X, such that $x \in \mathcal{X} = \{0, 1, 2, 3\}$. For all $x \in \mathcal{X}$ and for all $r \in V$, we set $\lambda_{rr}^{(x)} = -1$. For all $x \in \mathcal{X}$ and for all $r, j \in V, j < r$, the

non-zero interactions $\lambda_{rj}^{(x)}$ are set to 1.5. Note that $\mathcal{G}$ includes at most $p(p-1)/2 = 1225$ edges that are identified by the selection parameter $\delta^{(x)}$, $x \in \mathcal{X}$. Also in this case, we consider 4 different profile undirected graphs $\mathcal{G}(A)$, $\mathcal{G}(B)$, $\mathcal{G}(C)$ and $\mathcal{G}(D)$ where in Scenario (A) the four graphs are identical, in Scenario (B) the four graphs are completely different, in Scenario (C) the graphs $G(0)$ and $G(1)$ are identical but completely different to $G(2)$ and $G(3)$, which again are identical to each other and finally in Scenario (D) the grapjs $G(0)$, $G(1)$ and $G(2)$ are identical and completely different to $G(3)$.

6.2.2 Parameters setting

The parameter setting of the frequentists approaches is the same of Simulation study I. We set the spike and slab prior parameters $\rho = 10$ and $\gamma = 10^{-1}$ in the BAL approach and we consider a product of p Bernoulli(0.1) for the prior of the model with respect the BA approach.

6.3 Performance results

To assess the accuracy of graph structure estimation, we compute for any scenario, the Matthews correlation coefficient (MCC) and the F1 score (F1) of the true non-zero elements of $\lambda^{(x)}$, for all $x \in \mathcal{X}$, along with associated standard errors (SE). MCC is a balanced measure of binary classification that takes values between -1 (total disagreement) and +1 (perfect classification). F1 score is the harmonic mean of the precision and recall with the highest possible value equal to 1, indicating perfect precision and recall, and the lowest possible value is 0, if either the precision or the recall is zero. We report the results for both Simulation study I and II, averaged over the four graphs, in Table 1 and 2.

We comment the results obtained. As expected, the SL and BA methods tend to perform similarly, both being based on approximate inference. Compared to the latter, the BE approach, being based on exact inference, has a better performance. The methods SL, BA and BE do not take into account the homogeneity among the graphs. The DSSL method, conversely to SL, takes into account the homogeneity among the graphs and shows the overall best performance in scenario (A), where all the graphs are equal, but the worst in the remaining 3 scenarios, resulting in very little flexibility. Conversely, the Bayesian linking-graphs approaches BAL and BEL, in addition to show better results in scenarios where there is strong homogeneity among the graphs (Scenarios (A)-(D)), they do not

worsen noticeably in scenarios where there is very little or no homogeneity (Scenario (B)-(C)), resulting in better adaptation to the context. Also, the BAL and BEL methods standard errors tend not to grow compared to the corresponding non-linking methods BA and BE.

Table 1: MCC and F1 score with associated standard error (SE) across 10 simulated datasets for all scenarios of simulation study I

		(A)	(B)	(C)	(D)
MCC	SL	0.771(0.070)	0.759(0.063)	0.781(0.049)	0.806(0.037)
	DSSL	0.987(0.025)	0.581(0.144)	0.578(0.079)	0.726(0.038)
	BA	0.775(0.072)	0.753(0.049)	0.803(0.061)	0.827(0.038)
	BAL	0.831(0.055)	0.747(0.032)	0.807(0.061)	0.863(0.041)
	BE	0.828(0.050)	0.811(0.053)	0.819(0.050)	0.847(0.055)
	BEL	0.909(0.047)	0.737(0.056)	0.831(0.065)	0.875(0.047)
		(A)	(B)	(C)	(D)
F1	SL	0.812(0.057)	0.803(0.051)	0.816(0.044)	0.763(0.047)
	DSSL	0.989(0.020)	0.640(0.140)	0.663(0.062)	0.657(0.050)
	BA	0.803(0.063)	0.780(0.047)	0.826(0.059)	0.799(0.040)
	BAL	0.855(0.050)	0.785(0.030)	0.834(0.056)	0.836(0.046)
	BE	0.863(0.040)	0.849(0.042)	0.854(0.041)	0.810(0.068)
	BEL	0.927(0.038)	0.790(0.044)	0.863(0.053)	0.844(0.058)

Table 2: MCC and F1 score with associated standard error (SE) across 10 simulated datasets for all scenarios of simulation study II

		(A)	(B)	(C)	(D)
MCC	SL	0.911(0.006)	0.923(0.012)	0.916(0.011)	0.922(0.011)
	DSSL	0.989(0.011)	0.830(0.015)	0.723(0.015)	0.788(0.013)
	BA	0.934(0.008)	0.928(0.012)	0.941(0.013)	0.929(0.010)
	BAL	0.958(0.009)	0.925(0.012)	0.945(0.011)	0.936(0.011)
		(A)	(B)	(C)	(D)
F1	SL	0.914(0.006)	0.925(0.011)	0.918(0.011)	0.924(0.011)
	DSSL	0.990(0.010)	0.829(0.016)	0.710(0.016)	0.781(0.013)
	BA	0.935(0.008)	0.929(0.012)	0.942(0.013)	0.930(0.010)
	BAL	0.960(0.008)	0.928(0.012)	0.948(0.010)	0.940(0.11)

7 Conclusion

In this work, we propose two Bayesian approaches to infer multiple related Ising undirected graphical models. In particular, we extend the approach of Massam et al. (2009) and Bhattacharyya and Atchade (2019) to select multiple graphs. To take into account the similarities among the graphs we follow Peterson et al. (2015), using a Markov random field prior on the graph structure, which encourages the selection of the same edges in related graphs. In particular, we use a Bayesian exact-likelihood inference for low-dimensional binary data, based on conjugate priors for log-linear parameters. We randomly propose a new model and we compute the marginal likelihood through the Laplace approximation. We finally perform a stochastic search of the model with higher posterior probability using MCMC methods. We also propose a Bayesian approximate-likelihood inference for high-dimensional binary data, using a quasi-likelihood approach that enable to computationally manage the normalization constant. We use spike-and-slab priors for log-linear parameters to encode sparsity and MCMC algorithms for sampling from the quasi-posterior distribution and update the model parameters. We performed a simulation study to compare our methods with the competing Lasso approaches proposed in Meinshausen and Bühlmann (2006) and Ollier and Viallon (2017). We also checked the performance of our methods

respect to the same ones where we replace the linking-prior of the graph with identical and independent Bernoulli distributions. Our approaches show good performances compared to the competing approaches. In addition, as an unique feature, the proposed approaches can learn which groups are related and which are not in terms of graph structure. Future developments may be the extension of our approaches to different types of random variables, for instance, categorical response variables, relaxing the assumption of the Ising model. In addition, we may assume two or more external factors.

References

Ballout, N. and V. Viallon (2019). Structure estimation of binary graphical models on stratified data: application to the description of injury tables for victims of road accidents. *Statistics in Medicine 38*(14), 2680–2703.

Banerjee, O., L. El Ghaoui, and A. d'Aspremont (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. Mach. Learn. Res. 9*.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B 36*(2), 192–225.

Bhattacharyya, A. and Y. Atchade (2019). A bayesian analysis of large discrete graphical models. *ArXiv*.

Diaconis, P. and D. Ylvisaker (1979). Conjugate priors for exponential families. *The Annals of Statistics 7*(2), 269–281.

Dobra, A. and H. Massam (2010). The mode oriented stochastic search (moss) algorithm for log-linear models with conjugate priors. *Statistical Methodology 7*(3), 240 – 253.

Edwards, D. (2000). *Introduction to Graphical Modelling* (2nd ed.). Springer.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

George, E. I. and R. E. McCulloch (1997). Approaches for bayesian variable selection. *Statistica Sinica 7*(2), 339–373.

Gottardo, R. and A. E. Raftery (2008). Markov chain monte carlo with mixtures of mutually singular distributions. *Journal of Computational and Graphical Statistics 17*(4), 949–975.

Guo, J., C. J. L. E. M. G. Z. J. (2015). Estimating heterogeneous graphical models for discete data with an application to roll call voting. *The annals of applied statistics 30*(9), 821–848.

Guo, J., E. Levina, G. Michailidis, and J. Zhu (2011). Joint estimation of multiple graphical models. *Biometrika 98*(1), 1–15.

Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift Für Physik A Hadrons and Nuclei 31*, 253–258.

Jiang, W. and M. A. Tanner (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics 36*(5), 2207–2231.

Kato, K. (2013, 10). Quasi-bayesian analysis of nonparametric instrumental variables models. *Ann. Statist. 41*(5), 2359–2390.

Lauritzen, S. L. (1996). *Graphical Models*. New York: Oxford Univ. Press.

Massam, H., J. Liu, and A. Dobra (2009, 12). A conjugate prior for discrete hierarchical log-linear models. *Ann. Statist. 37*(6A), 3431–3467.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*(3), 1436–1462.

Ollier, E. and V. Viallon (2017, 01). Regression modelling on stratified data with the lasso. *Biometrika 104*(1), 83–96.

Peterson, C. B., F. Stingo, and M. Vannucci (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association 110*(509), 159–174.

Ravikumar, P., M. J. Wainwright, and J. D. Lafferty (2010). High-dimensional ising model selection using l1-regularized logistic regression. *Ann. Statist. 38*(3), 1287–1319.

Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association 81*(393), 82–86.

Chapter 4

Final remarks

Chapter 2: Profile graphical models

We aim to model the effect of a categorical factor on the dependence structure of a set of random response variables. In particular our interest focuses on the effect of the categorical factor on the interactions among the response variables. We propose two novel classes of graphical models, termed *profile undirected and bi-directed graphical models*, which preserve the convenient aspects of a graphical approach and enhance, at the same time, the modelling prospects given by chain graphs and multiple graphs. Under the assumption of a Multinomial sampling scheme for the multivariate outcome vector, a parameterization based on the log-linear transformation (Lauritzen, 1996) and on the log-mean linear transformation (Roverato et al., 2013) could be used, respectively, for profile undirected and bi-directed graph models. In this manuscript some aspects still require further investigations. Firstly, at this stage, inference and model selection are performed by means of an independent Lasso Sep-Logit approach (Meinshausen and Bühlmann, 2006) which does not account for the dependence between sub-group models. In this context, more efforts are needed for the implementation of a joint selection procedure based, for instance, of a Data-shared Lasso strategy (Ollier and Viallon, 2017). From a modelling perspective, an interesting development could be given by the generalization of the profile approach to chain graph models to explore profile dependence structures among variables grouped in chain components. This generalization is not trivial in terms of Markov property specification, since we need to consider the effect of an external factor on variables collected both within and between chain components. However, we conjecture that profile chain graph models would provide useful insights to investigate data generating processes for data which, in principle, might be different in each sub-group.

Chapter 3: Bayesian model selection of multiple Ising undirected graphs

The purpose of this work is to develope Bayesian methodologies to select multiple related Ising undirected graphical models. Following Massam et al. (2009), we devise a Bayesian exact-likelihood inference for low-dimensional binary response data, based on conjugate priors for log-linear parameters, where we implement a computational strategy that uses

Laplace approximations and a Metropolis-Hastings algorithm that allows us to perform a stochastic model search. We also propose a quasi-likelihood Bayesian approach, extending the work of Bhattacharyya and Atchade (2019), for fitting high-dimensional Ising multiple graphs, where the normalization constant results computationally intractable, with spike-and-slab priors to encode sparsity and MCMC algorithms for sampling from the quasi-posterior distribution which enables variable selection and estimation simultaneously. In both methods, we define a Markov Random Field prior on the graph structures, which encourages the selection of the same edges in related graphs (Peterson et al., 2015). We performed two simulation studies to compare our methods with the competing Lasso approaches proposed in Meinshausen and Bühlmann (2006) and Ollier and Viallon (2017). We also checked the performance of our methods with respect to the same ones where we replace the linking-prior of the graph with identical and independent Bernoulli distributions. Our approaches show good performances compared to the competing approaches. In addition, as an unique feature, the proposed approaches can learn which groups are related and which are not in terms of graph structure. Future developments may be the extension of our approaches to different types of random variables, for instance, categorical response variables, relaxing the assumption of the Ising model. In addition, we may assume two or more external factors.

References

Bhattacharyya, A. and Y. Atchade (2019). A bayesian analysis of large discrete graphical models. *ArXiv*.

Lauritzen, S. L. (1996). *Graphical Models*. New York: Oxford Univ. Press.

Massam, H., J. Liu, and A. Dobra (2009, 12). A conjugate prior for discrete hierarchical log-linear models. *Ann. Statist. 37*(6A), 3431–3467.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*(3), 1436–1462.

Ollier, E. and V. Viallon (2017, 01). Regression modelling on stratified data with the lasso. *Biometrika 104*(1), 83–96.

Peterson, C. B., F. Stingo, and M. Vannucci (2015). Bayesian inference of multiple Gaussian graphical models. *Journal of the American Statistical Association 110*(509), 159–174.

Roverato, A., M. Lupparelli, and L. La Rocca (2013). Log-mean linear models for binary data. *Biometrika 100*(2), 485–494.