# A review of Gaussian Markov models for conditional independence

Irene Córdoba *, Concha Bielza, Pedro Larrañaga

*Universidad Politécnica de Madrid, Madrid, Spain*

## ARTICLE INFO

## ABSTRACT

Markov models lie at the interface between statistical independence in a probability distribution and graph separation properties. We review model selection and estimation in directed and undirected Markov models with Gaussian parametrization, emphasizing the main similarities and differences. These two model classes are similar but not equivalent, although they share a common intersection. We present the existing results from a historical perspective, taking into account the amount of literature existing from both the artificial intelligence and statistics research communities, where these models were originated. We cover classical topics such as maximum likelihood estimation and model selection via hypothesis testing, but also more modern approaches like regularization and Bayesian methods. We also discuss how the Markov models reviewed fit in the rich hierarchy of other, higher level Markov model classes. Finally, we close the paper overviewing relaxations of the Gaussian assumption and pointing out the main areas of application where these Markov models are nowadays used.

## 1. Introduction

Markov models, or probabilistic graphical models, explicitly establish a correspondence between statistical independence in a probability distribution and certain separation criteria holding in a graph. They were originated at the interface between statistics, where Markov random fields were predominant (Darroch et al., 1980), and artificial intelligence, with a focus on Bayesian networks (Pearl, 1985, 1986). These two model classes are now considered the traditional ones, but still are widely applied and nowadays there is a significant amount of research devoted to them (Daly et al., 2011; Uhler, 2012). They both share the modelling of conditional independences: Bayesian networks relate them with acyclic directed graphs, whereas in Markov fields they are associated with undirected graphs. However, the models they represent are only equivalent under additional assumptions on the respective graphs.

In this paper, we review the existing methods for model selection and estimation in undirected and acyclic directed Markov models with a Gaussian parametrization. The multivariate Gaussian distribution is among the most widely developed and applied statistical family in this context (Werhli et al., 2006; Ibáñez et al., 2016), and allows for an explicit parametric comparison of their similarities and differences. The highly interdisciplinary nature of these Markov model classes has led to a wide range of terminology in methodological developments and theoretical results. They have usually been studied separately, with some exceptions (Wermuth, 1980; Pearl, 1988), and most unifying works (Sadeghi and Lauritzen, 2014; Wermuth, 2015) are characterized by a high-level view, where the models are embedded in other, more expressive classes, and the focus is on the properties of these container classes. By contrast, in this paper we review

---

* Corresponding author.
*E-mail addresses:* irene.cordoba@upm.es (I. Córdoba), mcbielza@fi.upm.es (C. Bielza), pedro.larranaga@fi.upm.es (P. Larrañaga).
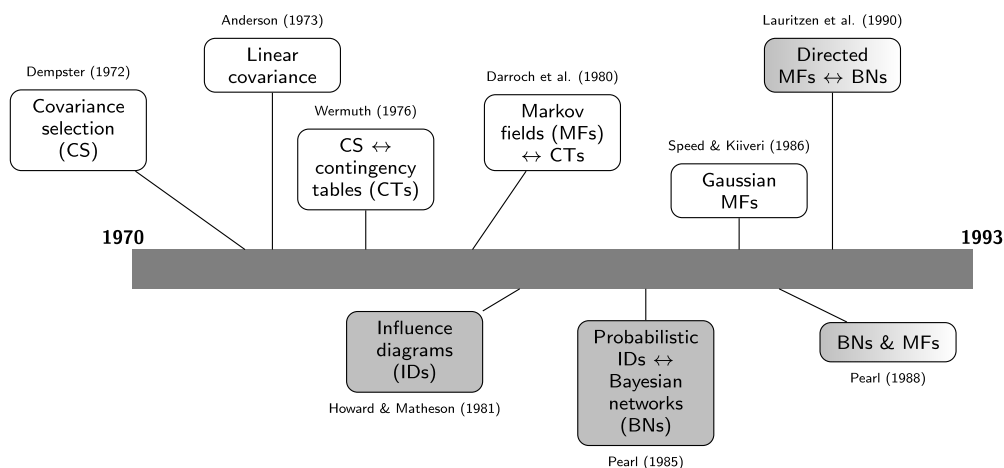
**Fig. 1.** Timeline on the origins of Gaussian Markov models. Papers from the statistical community appear at the top, while papers from other research areas appear below. Thematically, grey filled squares are papers about acyclic directed Markov models, the white ones are about the undirected case, and those gradient filled treat both classes.

them from a low-level perspective. In doing so, we use a unified notation that allows for a direct comparison between the two types of classes. Furthermore, throughout each section we explicitly compare them, in terms of both methodological and theoretical developments.

The paper is structured as follows. A historical introduction to Markov models is presented in Section 2, emphasizing the different research areas that contributed to their birth. Afterwards, preliminary concepts from graph theory are presented in Section 3. In Section 4, undirected and acyclic directed Markov model classes are introduced, under no distributional assumptions. This is because many foundational relationships between them can already be established from this general perspective. Next, we restrict their parametrization to multivariate Gaussian distributions, and explore the main derived properties from this in Section 5. We review maximum likelihood estimation in Section 6. These estimates are used for model selection via hypothesis testing, as we present in Section 7. When maximum likelihood estimators are not guaranteed to exist, a popular technique is to employ regularization, which we overview in Section 8. Finally, the alternative Bayesian approach for model selection and estimation is treated in Section 9. We explore the relationship of Gaussian acyclic directed and undirected Markov models with other, higher level model classes in Section 10. Alternatives to the Gaussian distribution are discussed in Section 11. We close the paper discussing the main real applications of the Gaussian Markov model classes reviewed in Section 12.

## 2. A historical perspective

We will now introduce the main terminology for Gaussian Markov models that can be found nowadays, from a historical perspective. In Fig. 1 we have depicted a timeline on the origins of Markov models, containing most of the key works we will refer to in this section.

Undirected Markov models for conditional independence are the oldest type of Markov models, preceded only by special cases such as the Ising model for ferromagnetic materials (Kindermann and Snell, 1980; Isham, 1981). In fact, they are a generalization of the Ising model, which is at the same time a generalization of Markov chains. Originally, undirected Markov models were called *Markov random fields* (Grimmett, 1973), since they generalized the correspondence between Gibbs measures (Besag, 1974) and Markov properties. The terminology *graphical model* was not introduced until Darroch et al. (1980) linked the graphical ideas for contingency tables with Markov properties of discrete Markov fields. Furthermore, we also find them called *Markov networks* (Pearl, 1988), from researchers in artificial intelligence, as a parallel to the terminology *Bayesian networks*, used for acyclic directed Markov fields.

Regarding the Gaussian parametrization, we can find that one of the first works to impose some structure on the covariance matrix of a multivariate Gaussian distribution, in order to reduce the number of parameters to be estimated, was Anderson (1973). He considered the mean vector and covariance matrix to be linear combinations of known linearly independent vectors and matrices, respectively. Closely following this work was Dempster (1972), who suggested to estimate the inverse of the covariance matrix (concentration matrix) by assuming certain entries equal to zero, motivated by the representation of the multivariate Gaussian distribution as an exponential family. His work was later referred to as *covariance selection* models. Interestingly, although Dempster did not have any graphical interpretation in mind, such zero entries in the concentration matrix are directly associated with missing edges in an undirected Gaussian Markov models, and these correspondence was analysed some years later in Wermuth (1976a). This is why, even nowadays, these Markov models with a Gaussian parametrization are sometimes called covariance selection models.
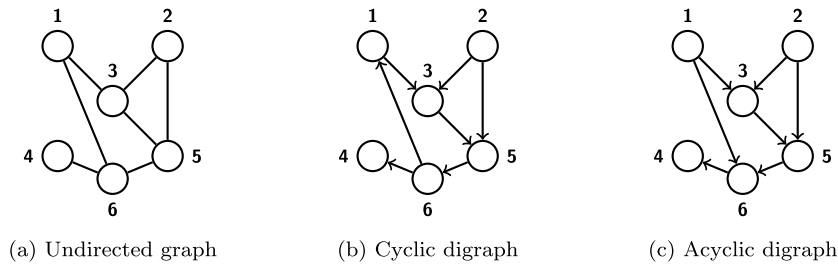
**Fig. 2.** Examples of an undirected graph and two digraphs.

Acyclic digraphs, in contrast, were intensely used as models for multivariate probability distributions after the definition of *influence diagrams*. These are used to model decision-making processes, and were introduced by Howard and Matheson in 1981 (article reprinted in Howard and Matheson (2005)). Their probabilistic reduction coincides with acyclic directed Markov models, and was subsequently extensively studied by Pearl (1988), who renamed probabilistic influence diagrams as *Bayesian networks* or *influence networks* (Pearl, 1985). Some researchers working on Markov fields also developed theory regarding these directed counterparts, calling them *directed Markov fields* (Lauritzen et al., 1990).

Earlier works than the previously outlined, employing or referencing acyclic directed Markov models, are available. Wermuth (1980) implicitly studied them in the Gaussian case as *linear recursive regression systems*, although the main focus was rather on covariance selection models. In fact, we can trace the use of directed graphs as graphical models for dependencies among random variables at least to the work of geneticist Sewall Wright in 1918, who developed the method of *path coefficients* (Wright, 1934), nowadays known as *path analysis*. Linearly related variables were represented using a directed acyclic graph, whereas their correlation was represented by bi-directed edges joining them.

## 3. Graph preliminaries

A graph is defined as a pair $\mathcal{G} = (V, E)$ where $V$ is the vertex set and $E$ is the edge set. Throughout all the paper, and unless otherwise stated, the graphs will be labelled and simple, which means that the elements in $V$ are labelled, for example, as $1, \ldots, p$; and $E$ is formed by pairs of distinct elements in $V$. A graph is called *undirected* if these latter pairs are unordered ($E \subseteq \{\{u, v\} \ : \ u, v \in V\}$), and *directed* or *digraph* otherwise ($E \subseteq \{(u, v) \ : \ u, v \in V\}$). Edges $\{u, v\}$ in an undirected graph are usually denoted as $uv$ and graphically represented as a line (see Fig. 2(a)); while in a digraph they are called *arcs* or *directed edges* and represented as arrows (Figs. 2(b) and 2(c)).

### 3.1. Undirected graphs

In an undirected graph $\mathcal{G} = (V, E)$, if $uv \in E$, $u$ and $v$ are called *neighbours*. For $v \in V$, the set of its neighbours is denoted as $\mathrm{ne}(v)$, and the *closure* of $v$ is $\mathrm{cl}(v) := \{v\} \cup \mathrm{ne}(v)$. $\mathcal{G}$ is called *complete* if for every $u, v \in V$, $uv \in E$. A maximal $C \subseteq V$ such that $\mathcal{G}_C$ is complete is called a *clique*. Let $\mathcal{H} = (V_{\mathcal{H}}, E_{\mathcal{H}})$ be another undirected graph. $\mathcal{H}$ is a *sub-graph* of $\mathcal{G}$ (written as $\mathcal{H} \subseteq \mathcal{G}$) if $V_{\mathcal{H}} \subseteq V$ and $E_{\mathcal{H}} \subseteq E$. If $E_{\mathcal{H}} = \{uv \in E : u, v \in V_{\mathcal{H}}\}$, then $\mathcal{H}$ is called the *induced sub-graph* and denoted $\mathcal{G}_{V_{\mathcal{H}}}$.

A *walk* between $u$ and $v$ is an ordered sequence of vertices $(u =)u_0, u_1, \ldots, u_{k-1}, u_k(= v)$ where $u_{i-1}u_i \in E$ for $i \in \{1, \ldots, k\}$. The number $k$ is called the *length* of the walk. If $u = v$ the walk is *closed*, and when $u_0, \ldots, u_{k-1}$ are distinct, the walk is called a *path*. A closed path of length $k \geq 3$ is called a *cycle* or $k$-*cycle*. $\mathcal{G}$ is called *chordal* or *triangulated* if all minimal $k$-cycles are of length $k = 3$. A *chordal cover* of a graph $\mathcal{G}$ is a graph $\mathcal{G}^*$ such that $\mathcal{G} \subseteq \mathcal{G}^*$ and $\mathcal{G}^*$ is chordal.

$S \subseteq V$ *separates* $u$ and $v$ in $\mathcal{G} = (V, E)$ if there is no path between $u$ and $v$ in the sub-graph $\mathcal{G}_{V \setminus S}$. If we consider $A, B, S \subseteq V$, $A$ and $B$ are said to be *separated* by $S$ if $u$ and $v$ are separated by $S$ for all $u \in A, v \in B$. Let $V$ be partitioned into disjoint sets $A, B, S \subseteq V$. $(A, B, S)$ is called a *decomposition* of $\mathcal{G}$ if $S$ separates $A$ and $B$ in $\mathcal{G}$ and $\mathcal{G}_S$ is complete. If $A \neq \emptyset$ and $B \neq \emptyset$ the decomposition is said to be *proper*. An undirected graph is *decomposable* if (i) it is complete or (ii) it admits a proper decomposition into *decomposable* sub-graphs. An undirected graph is decomposable if and only if it is chordal.

### 3.2. Acyclic digraphs

In a digraph $\mathcal{D} = (V, A)$ the definitions of (induced) sub-graph, walk, path, and cycle are analogous to the undirected case. The undirected graph $\mathcal{D}^U := (V, A^U)$ with $A^U := \{uv : (u, v) \in A\}$ is called the *skeleton* of $\mathcal{D}$, and $\mathcal{D}$ is one of its *orientations*. A digraph $\mathcal{D}$ is said to be *complete* when $\mathcal{D}^U$ is complete.

In the following, assume that $\mathcal{D}$ is acyclic (see Fig. 2(b) for a cyclic digraph, and Fig. 2(c) for an acyclic one). The *parent set* of $v \in V$ is $\mathrm{pa}(v) := \{u \in V : (u, v) \in A\}$; conversely, the *child set* is $\mathrm{ch}(v) := \{u \in V : (v, u) \in A\}$. The *ancestors* of $v$, $\mathrm{an}(v)$, are those $u \in V$ such that there exists a directed path from $u$ to $v$; the *descendants* of $v$, $\mathrm{de}(v)$, are those $u \in V$ such that there exists a directed path from $v$ to $u$. We will let $\mathrm{nd}(v) := V \setminus (\{v\} \cup \mathrm{de}(v))$ be the set of *non-descendants* of

$v \in V$, and $\text{An}(A) := A \cup (\cup_{a \in A} \text{an}(a))$ the *ancestral set* of $A \subseteq V$. Note that a total order $\prec$ can be defined over the set of vertices $V$ in an acyclic digraph $\mathcal{D} = (V, A)$, such that if $(u, v) \in A$, then $u \prec v$. This ordering is usually called *ancestral*, and it is a linear extension of the partial order naturally defined as $u \preceq v$ if $u \in \text{an}(v)$. For $v \in V$, the set of *successors* of $v$ with respect to $\prec$ is $\text{su}(v) = \{u \in V : u \succ v\}$; the set of *predecessors* of $v$ is $\text{pr}(v) = \{u \in V : u \prec v\}$.

Finally, let $u, w_1, w_2 \in V$ with $(w_1, u), (w_2, u) \in A$ and $(w_1, w_2), (w_2, w_1) \notin A$ (see vertices 1, 2 and 3 in Fig. 2(c)). Such configurations are usually called *v-structures* and denoted as $w_1 \rightarrow u \leftarrow w_2$. The *moral graph* of $\mathcal{D}$ is defined as the undirected graph $\mathcal{D}^m = (V, A^m)$ with $A^m := A^U \cup \{w_1 w_2 : w_1 \rightarrow u \leftarrow w_2 \text{ for some } u \in V\}$.

## 4. Undirected and acyclic directed Markov model classes

The Markov model classes we will review associate conditional independences in random vectors $\boldsymbol{X} = (X_1, \ldots, X_p)^t$ with undirected graph and acyclic digraph separation properties. This is made explicit via the *Markov properties* of the distribution of $\boldsymbol{X}$, which are in turn based on what are known as *independence relations*.

In the following, for arbitrary $I \subseteq \{1, \ldots, p\}$, we will denote the $|I|$-dimensional sub-vector of $\boldsymbol{X}$ as $\boldsymbol{X}_I := (X_i)_{i \in I}$. Conditional independence will be expressed as in Dawid (1979): $\boldsymbol{X}_I \perp\!\!\!\perp \boldsymbol{X}_J \mid \boldsymbol{X}_K$ represents the statement '$\boldsymbol{X}_I$ is conditionally independent from $\boldsymbol{X}_J$ given $\boldsymbol{X}_K$' (see, e.g. Studenỳ, 2018, §1.3).

### 4.1. Independence relations

An *independence relation* over a set $V = \{1, \ldots, p\}$ is a collection $\mathcal{I}$ of triples $(A, B, C)$ where $A$, $B$ and $C$ are pairwise disjoint subsets of $V$. It is called a *semi-graphoid* when the following conditions are met,

if $(A, B, C) \in \mathcal{I}$ then $(B, A, C) \in \mathcal{I}$,

if $(A, B \cup C, D) \in \mathcal{I}$ then $(A, C, D) \in \mathcal{I}$ and $(A, B, C \cup D) \in \mathcal{I}$,

if $(A, B, C \cup D) \in \mathcal{I}$ and $(A, C, D) \in \mathcal{I}$ then $(A, B \cup C, D) \in \mathcal{I}$;

and a *graphoid* when it additionally satisfies that if $(A, B, C \cup D) \in \mathcal{I}$ and $(A, C, B \cup D) \in \mathcal{I}$ then $(A, B \cup C, D) \in \mathcal{I}$ (Pearl and Paz, 1987).

Independence relations arise in different contexts relevant for Markov models. Specifically, an independence relation $\mathcal{I}$ over $V = \{1, \ldots, p\}$ is said to be *induced* by

- an undirected graph $\mathcal{G} = (V, E)$ if $(A, B, S) \in \mathcal{I} \iff A$ and $B$ are separated by $S$ in $\mathcal{G}$,
- an acyclic digraph $\mathcal{D} = (V, A)$ if $(A, B, S) \in \mathcal{I} \iff A$ and $B$ are separated by $S$ in $(\mathcal{D}_{\text{An}(A \cup B \cup S)})^m$,
- a $p$-dimensional random vector $\boldsymbol{X}$ if $(A, B, S) \in \mathcal{I} \iff \boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_B \mid \boldsymbol{X}_S$.

Graph-induced independence relations are always graphoids, while probabilistic ones are always semi-graphoids and require additional assumptions on the probability spaces involved to be graphoids (Dawid, 1980). See Studenỳ (2018) §1.5 and §1.11 for a detailed exposition of graphoid theory and how to compute and represent their closures, that is, all the triplets that can be derived from a given independence relation by using the graphoid axioms.

The core of Markov model classes is the relationship between induced independence relations, which we will denote as $\mathcal{I}(\cdot)$ with the argument being the inducing element. Specifically, if $\mathcal{G}$ is an undirected (acyclic directed) graph, an undirected (directed) *Markov model* is defined as

$$\mathcal{M}(\mathcal{G}) := \{P_{\boldsymbol{X}} \ : \ \mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(\boldsymbol{X})\}.$$

where the random vectors $\boldsymbol{X}$ are defined over the same probability space and $P_{\boldsymbol{X}}$ denotes their distribution. These classes are non-empty (Geiger and Pearl, 1990, 1993); that is, for any undirected or acyclic directed graph, we can always find a probability distribution whose independence model contains the one generated by the graph.

Graphoids can be generalized to what are known as *separoids* (Dawid, 2001), which are algebraic structures usually appearing whenever a notion of 'irrelevance' is being mathematically treated (see, e.g. Studenỳ, 2018, §1.1.3). Further research on these axiom systems from an abstract point of view could shed more light on how the apparently different mathematical contexts in which such structures arise are related, and also provide an explicit bridge between them and the recently defined *independence logic* (Grädel and Väänänen, 2013), closely related.

### 4.2. Markov properties

When a distribution $P_{\boldsymbol{X}}$ belongs to $\mathcal{M}(\mathcal{G})$ for an undirected or acyclic directed graph $\mathcal{G}$, it is said that $P_{\boldsymbol{X}}$ is globally $\mathcal{G}$-*Markov* or satisfies the *global Markov property* with respect to $\mathcal{G}$. There are other weaker Markov properties that usually allow to simplify the model. Specifically, if $\mathcal{G} = (V, E)$ is an undirected graph, then the probability distribution $P_{\boldsymbol{X}}$ of $\boldsymbol{X}$ is said to be

- *pairwise $\mathcal{G}$-Markov* if $X_u \perp\!\!\!\perp X_v \mid \boldsymbol{X}_{V \setminus \{u, v\}}$ for all $uv \notin E$,
- *locally $\mathcal{G}$-Markov* if $X_v \perp\!\!\!\perp \boldsymbol{X}_{V \setminus \text{cl}(v)} \mid \boldsymbol{X}_{\text{ne}(v)}$ for all $v \in V$;

whereas if $\mathcal{G}$ is an acyclic digraph, then $P_{\mathbf{X}}$ is called

- *pairwise $\mathcal{G}$-Markov* if $X_u \perp\!\!\!\perp X_v \mid \mathbf{X}_{\mathrm{nd}(u)\setminus\{v\}}$ for all $u \in V$, $v \in \mathrm{nd}(u) \setminus \mathrm{pa}(u)$;
- *locally $\mathcal{G}$-Markov* if $X_v \perp\!\!\!\perp \mathbf{X}_{\mathrm{nd}(v)\setminus\mathrm{pa}(v)} \mid \mathbf{X}_{\mathrm{pa}(v)}$ for all $v \in V$.

The three Markov properties are equivalent when $\mathcal{G}$ is acyclic directed (Lauritzen et al., 1990), while if $\mathcal{G}$ is undirected this equivalence is only guaranteed when $\mathcal{I}(\mathbf{X})$ is a graphoid (Pearl, 1988). A sufficient condition for this to happen is that $P_{\mathbf{X}}$ admits a continuous and strictly positive density. This result was proved in different forms by several authors, but it is usually attributed to Hammersley and Clifford (1971), who were the first to outline the proof for the discrete case (Speed, 1979). It relies on an additional characterization of a probability distribution with respect to $\mathcal{G}$: denoting as $\mathfrak{C}(\mathcal{G})$ the class of cliques of $\mathcal{G}$, the density function $f$ of $P_{\mathbf{X}}$ is said to *factorize according to $\mathcal{G}$* when there exists a set $\{\psi_C(\mathbf{x}_C) : C \in \mathfrak{C}(\mathcal{G}), \psi_C \geq 0\}$ such that

$$f(\mathbf{x}) = \prod_{C \in \mathfrak{C}(\mathcal{G})} \psi_C(\mathbf{x}_C). \tag{1}$$

When (1) holds, then $P_{\mathbf{X}}$ is globally $\mathcal{G}$-Markov, while if $f$ is continuous and strictly positive, the pairwise Markov property implies (1), which gives the equivalence of Markov properties. Positivity is a straightforward sufficient condition for checking whether an independence model originated from a distribution is a graphoid. Necessary and sufficient conditions are given in measure theoretic terms by Dawid (1980), and recently by Peters (2014) in terms of special functions over the sample space.

Finally, recall that the nodes of an acyclic digraph $\mathcal{D} = (V, A)$ can be totally ordered such that if $(u, v) \in A$, then $u \in \mathrm{pr}(v)$. This gives rise to another Markov property, exclusive for acyclic digraphs: $P_{\mathbf{X}}$ is said to be *ordered $\mathcal{D}$-Markov* if $X_v \perp\!\!\!\perp \mathbf{X}_{\mathrm{pr}(v)\setminus\mathrm{pa}(v)} \mid \mathbf{X}_{\mathrm{pa}(v)}$ for all $v \in V$. This property is also equivalent to the global, local and pairwise Markov properties (Lauritzen et al., 1990). The classical theory of undirected and acyclic directed Markov properties can be found in Lauritzen (1996), whereas Studenỳ (2018) §1.7 and §1.8 provides a recent overview.

### 4.3. Independence and Markov equivalence

When the Markov models defined by two graphs $\mathcal{G}$ and $\tilde{\mathcal{G}}$, with the same vertex set $V$, coincide, such graphs are said to be *Markov equivalent*. A simpler notion, which implies Markov equivalence, is *independence equivalence*, holding when $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\tilde{\mathcal{G}})$. Independence equivalence is implied by Markov equivalence under fairly general circumstances (Studenỳ, 2005, §6.1), which is why most authors treat them as the same notion. These equivalences allow to choose the most suited graph for the Markov model.

We will first characterize equivalence within undirected graphs. For each graphoid $\mathcal{I}$ over $V$ there exists a unique edge-minimal undirected graph $\mathcal{G}$ such that $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}$ (Pearl and Paz, 1987). It follows that $\mathcal{I}(\mathcal{G}) = \mathcal{I}(\tilde{\mathcal{G}})$ (independence equivalence) if and only if $\mathcal{G}$ and $\tilde{\mathcal{G}}$ are identical. Furthermore, if we assume that $\mathcal{I}(\mathbf{X})$ is a graphoid for all $P_{\mathbf{X}} \in \mathcal{M}(\mathcal{G})$, then a unique edge-minimal $\tilde{\mathcal{G}}$ exists, with $\tilde{\mathcal{G}} \subseteq \mathcal{G}$, such that $\mathcal{M}(\mathcal{G}) = \mathcal{M}(\tilde{\mathcal{G}})$ (Markov equivalence); that is, a unique undirected graph can be chosen as representative of each undirected Markov model.

In contrast, acyclic digraphs are not, in general, unique representations of a Markov model, since $\mathcal{I}(\mathcal{D}) = \mathcal{I}(\tilde{\mathcal{D}})$ if and only if $\mathcal{D}$ and $\tilde{\mathcal{D}}$ have the same skeleton and the same v-structures (Verma and Pearl, 1991). However, unique representatives can be constructed: let $\mathfrak{D}_p$ be the set of acyclic digraphs over $V = \{1, \ldots, p\}$ and define an equivalence relation $\sim$ in $\mathfrak{D}_p$ as $\mathcal{D} \sim \tilde{\mathcal{D}} \iff \mathcal{I}(\mathcal{D}) = \mathcal{I}(\tilde{\mathcal{D}})$. The quotient space of $\sim$ is $\mathfrak{D}_p/\!\!\sim = \{[\mathcal{D}] : \mathcal{D} \in \mathfrak{D}_p\}$, where $[\mathcal{D}] := \{\tilde{\mathcal{D}} \in \mathfrak{D}_p : \tilde{\mathcal{D}} \sim \mathcal{D}\}$ is the *Markov equivalence class*; indeed, $\mathcal{M}(\tilde{\mathcal{D}}) = \mathcal{M}(\mathcal{D})$ for all $\tilde{\mathcal{D}} \in [\mathcal{D}]$, that is, $[\mathcal{D}]$ is the unique representative of the directed Markov model.

The asymptotic ratio $l = \lim_{p\to\infty}|\mathfrak{D}_p|/|\mathfrak{D}_p/\!\!\sim|$ influences the computational gain obtained by using $\mathfrak{D}_p/\!\!\sim$ instead of $\mathfrak{D}_p$ as a search space for model selection. Steinsky (2004) analytically calculates an upper bound of $l$ as 13.65. Exact computations by Gillispie and Perlman (2002), for $p \leq 10$, and approximations by Sonntag et al. (2015), up to $p = 31$, seem to indicate that $l \sim 3.7$. However, its analytical deduction remains an open problem. Note that the computational gain is not only influenced by $l$, but also by other factors, such as how the element size in $\mathfrak{D}_p/\!\!\sim$ is distributed. An algorithm to compute such sizes can be found in He et al. (2015). Recently, Radhakrishnan et al. (2018) have provided tight lower and upper bounds on the number and size of Markov equivalence classes when $\mathcal{D}$ is a tree.

Finally, we will characterize equivalence between directed and undirected graphs, firstly obtained by Wermuth (1980) for multivariate Gaussian distributions, Wermuth and Lauritzen (1983) for contingency tables, and generalized in Frydenberg (1990) for graphoid-inducing distributions. When $\mathcal{G}$ is an undirected graph, $\mathcal{M}(\mathcal{G}) = \mathcal{M}(\mathcal{D})$ for some acyclic digraph $\mathcal{D}$ if and only if $\mathcal{G}$ is chordal. Conversely, an acyclic digraph $\mathcal{D}$ is Markov equivalent to its skeleton $\mathcal{D}^U$ if and only if $\mathcal{D}$ contains no v-structures. Furthermore, a relation with the moral graph can be established, which requires an analogous to (1): a density function $f$ is said to *recursively factorize* according to $\mathcal{D}$ when

$$f(\mathbf{x}) = \prod_{v \in V} f(x_v \mid \mathbf{x}_{\mathrm{pa}(v)}).$$

This characterization is equivalent to the Markov properties, and also implies that $f$ factorizes as in (1) with respect to the moral graph $\mathcal{D}^m$ (Lauritzen et al., 1990). This means that $P_{\mathbf{X}}$ is always globally $\mathcal{D}^m$-Markov for continuous $\mathbf{X}$, and thus $\mathcal{M}(\mathcal{D}) \subseteq \mathcal{M}(\mathcal{D}^m)$, with the equality only holding when $\mathcal{D}^m = \mathcal{D}^U$.
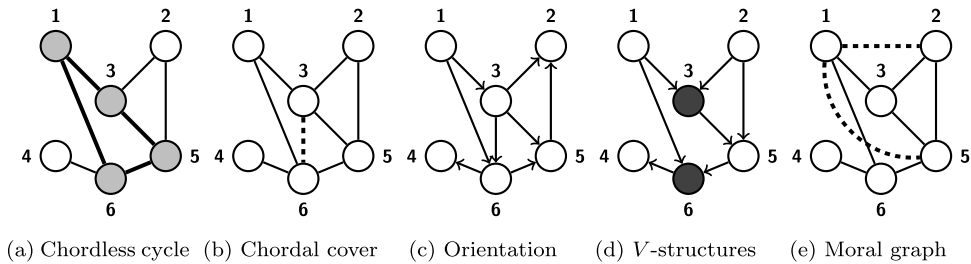
(a) Chordless cycle (b) Chordal cover (c) Orientation (d) *V*-structures (e) Moral graph

**Fig. 3.** Markov equivalence.

**Example 1.** An illustration of the previous concepts can be found in Fig. 3. The graph in 3(a) is not chordal, and thus there is no Markov equivalent acyclic digraph. 3(b) is a chordal cover of 3(a), and a Markov equivalent orientation is depicted in 3(c). The acyclic digraph in 3(d) has v-structures, emphasized in dark grey, and thus cannot be Markov equivalent to its skeleton 3(a). The moral graph of 3(d) is 3(e), which in fact is another chordal cover of 3(a), and thus none of its orientations will be Markov equivalent to 3(c).

## 5. Gaussian parametrization

When restricting to multivariate Gaussian distributions, we find connections between conditional and vanishing parameters. This correspondence can be used for providing a direct interpretation of Markov properties, both in the undirected and directed case, allowing an enhanced manipulation of these Markov models.

In the following, the elements of a real $q \times r$ matrix $\mathbf{M} \in \mathbb{M}_{q \times r}(\mathbb{R})$ will be denoted as $m_{ij}$, where $i \in \{1, \ldots, q\}$ and $j \in \{1, \ldots, r\}$. $\mathbf{M}_{IJ}$ will be the $|I| \times |J|$ sub-matrix of $\mathbf{M}$, where $I \subseteq \{1, \ldots, q\}$ and $J \subseteq \{1, \ldots, r\}$; and we will use $\mathbf{M}_{IJ}^{-1}$ as $(\mathbf{M}_{IJ})^{-1}$. $\mathbb{S}^{>0}$ and $\mathbb{S}^{\geq 0}$ will represent the sets of positive and semi-positive definite symmetric matrices, respectively. The $p$-variate Gaussian distributions is denoted as $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} \in \mathbb{R}^p$ is the mean vector and $\boldsymbol{\Sigma} \in \mathbb{S}^{>0}$ is the covariance matrix. $\mathbf{I}_p$ will denote the identity $p$-dimensional square matrix, whereas $\mathbf{1}_p$ will denote the $p$-vector with all entries equal to 1; many times, dimensionality sub-scripts will be dropped if the dimension of the respective object is clear from the context.

### 5.1. Conditional independence and the multivariate Gaussian distribution

Let $V = \{1, \ldots, p\}$. When a random vector $\boldsymbol{X}$ is distributed as $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then for $i, j \in V$, $X_i \perp\!\!\!\perp X_j$ is equivalent to $\sigma_{ij} = 0$. If we consider a partition $(I, J)$ of $V$, then $\boldsymbol{X}_I \mid \boldsymbol{x}_J$ is distributed as $\mathcal{N}_{|I|}(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_{I \cdot J})$, where $\boldsymbol{\Sigma}_{I \cdot J} = \boldsymbol{\Sigma}_{II} - \boldsymbol{\Sigma}_{IJ} \boldsymbol{\Sigma}_{JJ}^{-1} \boldsymbol{\Sigma}_{JI}$ (Anderson, 2003). Thus, for $i, k \in I$, we have that $X_i \perp\!\!\!\perp X_k \mid \boldsymbol{x}_J$ is equivalent to $\sigma_{ik \cdot J} = 0$, the $(i, k)$ element in the conditional covariance matrix $\boldsymbol{\Sigma}_{I \cdot J}$.

A correspondence can be established between the zeros in $\boldsymbol{\Sigma}_{I \cdot J}$ and zero patterns in other representative matrices (Wermuth, 1976a, 1980; Uhler, 2018, §9.1), as follows. Let the concentration matrix of $\boldsymbol{X}$ be $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, with elements $\omega_{uv}$ for $u, v \in V$. The matrix $\boldsymbol{\Sigma}_{IJ} \boldsymbol{\Sigma}_{JJ}^{-1}$ is usually denoted as $\mathbf{B}_{I \cdot J}$ and called the matrix of *regression coefficients* of $\boldsymbol{X}_I$ on $\boldsymbol{X}_J$. Letting $\boldsymbol{\Omega}_{I \cdot J} := \boldsymbol{\Sigma}_{JJ}^{-1}$, we have the following matrix identity (Horn and Johnson, 2012)

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Sigma}_{II} & \boldsymbol{\Sigma}_{IJ} \\ \boldsymbol{\Sigma}_{JI} & \boldsymbol{\Sigma}_{JJ} \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{I \cdot J}^{-1} & -\boldsymbol{\Sigma}_{I \cdot J}^{-1} \mathbf{B}_{I \cdot J} \\ -\mathbf{B}_{I \cdot J}^t \boldsymbol{\Sigma}_{I \cdot J}^{-1} & \boldsymbol{\Omega}_{I \cdot J} + \mathbf{B}_{I \cdot J}^t \boldsymbol{\Sigma}_{I \cdot J}^{-1} \mathbf{B}_{I \cdot J} \end{pmatrix}.$$

This allows us to relate $\boldsymbol{\Sigma}_{I \cdot J}$ with $\boldsymbol{\Omega}$ and $\mathbf{B}_{I \cdot J}$ as

$$\boldsymbol{\Sigma}_{I \cdot J} = \boldsymbol{\Omega}_{II}^{-1}, \tag{2}$$

$$\mathbf{B}_{I \cdot J} = -\boldsymbol{\Omega}_{II}^{-1} \boldsymbol{\Omega}_{IJ}, \tag{3}$$

which implies that, dually, $\boldsymbol{\Omega}_{II}$ is identically equal to the concentration matrix of $\boldsymbol{X}_I \mid \boldsymbol{x}_J$, while $\boldsymbol{\Omega}_{I \cdot J}$ is the concentration matrix of $\boldsymbol{X}_J$. Finally, from (2) we get, for $i, k \in V$,

$$X_i \perp\!\!\!\perp X_j \mid \boldsymbol{X}_{V \setminus \{i, k\}} \iff \omega_{ik} = 0, \tag{4}$$

whereas from (3) it follows that, for $J \subseteq V$, $i, k \in V \setminus J$,

$$X_i \perp\!\!\!\perp X_k \mid \boldsymbol{X}_J \iff \beta_{ik \cdot J \cup \{k\}} = 0, \tag{5}$$

where $\beta_{ik \cdot J \cup \{k\}}$ is the $v$ entry in the vector $\boldsymbol{\beta}_{i \cdot J \cup \{k\}}^t$, that is, the coefficient of $X_k$ on the regression of $X_i$ on $\boldsymbol{x}_{J \cup \{k\}}$. The original notation for this, introduced in Yule (1907), was $\beta_{ik \cdot J}$; that is, $k$ is implicitly considered as included in the conditioning indexes. We have however chosen the alternative, explicit notation $\beta_{ik \cdot J \cup \{k\}}$, since it provides more notational simplicity in later sections.

## 5.2. Gaussian Markov models

In the Gaussian case, undirected Markov models are in correspondence with the concentration matrix, while for acyclic digraphs this correspondence is with the regression coefficients. Both rely on the auxiliary Markov properties that we presented in Section 4.2.

Let $\mathcal{G} = (V, E)$ be an undirected graph and consider $\boldsymbol{X}$ distributed as $P_{\boldsymbol{X}} \equiv \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $P_{\boldsymbol{X}} \in \mathcal{M}(\mathcal{G})$. Since $P_{\boldsymbol{X}}$ is globally $\mathcal{G}$-Markov, it is also pairwise $\mathcal{G}$-Markov, and thus (4) directly gives that $\omega_{uv} = 0$ for all $u, v \in V$ such that $uv \in E$. This means that, if we define the set $\mathbb{S}^{>0}(\mathcal{G}) := \{ \mathbf{M} \in \mathbb{S}^{>0} \; : \; m_{uv} = 0 \text{ for all } uv \notin E \}$, we have $\boldsymbol{\Omega} \in \mathbb{S}^{>0}(\mathcal{G})$ if and only if $P_{\boldsymbol{X}}$ is pairwise $\mathcal{G}$-Markov. Furthermore, since the multivariate Gaussian distribution has positive density, $\mathcal{I}(\boldsymbol{X})$ is a graphoid and thus the three Markov properties are equivalent. This allows us to redefine the *Gaussian undirected Markov model* as

$$\mathcal{N}(\mathcal{G}) = \left\{ \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \; : \; \boldsymbol{\Sigma}^{-1} \in \mathbb{S}^{>0}(\mathcal{G}), \boldsymbol{\mu} \in \mathbb{R}^p \right\}. \tag{6}$$

In the directed case, the redefinition is not so direct. Let $\mathcal{D} = (V, A)$ be an acyclic digraph, and assume, for notational simplicity, that the nodes are already ancestrally ordered as $1 \preceq \cdots \preceq p$. If $\boldsymbol{X}$ is distributed as $P_{\boldsymbol{X}} \equiv \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $P_{\boldsymbol{X}} \in \mathcal{M}(\mathcal{D})$, it satisfies the ordered Markov property. Thus, whenever $v \in \mathrm{pr}(u) \setminus \mathrm{pa}(u)$, we have $X_u \perp\!\!\!\perp X_v \mid \boldsymbol{X}_{\mathrm{pa}(u)}$, which is equivalent to $\beta_{uv \cdot \mathrm{pa}(u) \cup \{v\}} = 0$ as in (5). Since we have assumed an ancestral order, $\beta_{uv \cdot \mathrm{pa}(u) \cup \{v\}} = \beta_{uv \cdot \mathrm{pr}(u)}$ for all $u \in V$, $v \in \mathrm{pr}(u) \setminus \mathrm{pa}(u)$, which leads to $P_{\boldsymbol{X}}$ being ordered $\mathcal{D}$-Markov if and only if $\beta_{uv \cdot \mathrm{pr}(u)} = 0$ for all $u \in V, v \in \mathrm{pr}(u) \setminus \mathrm{pa}(u)$. Such triangular requirement on the regression coefficients can be expressed with the matrix $\mathbf{B}$ defined, for $v < u$ as $b_{uv} = 0$ if $v \notin \mathrm{pa}(u)$, and $b_{uv} = \beta_{uv \cdot \mathrm{pa}(u)} \equiv \beta_{uv \cdot \mathrm{pr}(u)}$ otherwise.

If we let $v_u := \sigma_{uu \cdot \mathrm{pr}(u)}$, the previous characterization leads to a matrix form of the linear regressions involved as

$$\boldsymbol{X} = \boldsymbol{\mu} + \mathbf{B}(\boldsymbol{X} - \boldsymbol{\mu}) + \boldsymbol{E}, \tag{7}$$

where $E_u \sim \mathcal{N}(0, v_u)$. We can rearrange it as $\boldsymbol{X} = \mathbf{U}^{-1}\boldsymbol{\xi} + \mathbf{U}^{-1}\boldsymbol{E}$, where $\boldsymbol{\xi} := \mathbf{U}\boldsymbol{\mu}$ and $\mathbf{U} := \mathbf{I}_p - \mathbf{B}$, since $\mathbf{U}$ is invertible. Let $\mathbf{V}$ be the diagonal matrix of conditional variances $v$. Sometimes $\boldsymbol{\xi}, \mathbf{B}$ and $\mathbf{V}$ are called the $\mathcal{D}$-parameters of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (Andersson and Perlman, 1998). In fact, $\mathbf{U}$ and $\mathbf{V}$ allow a decomposition of $\boldsymbol{\Sigma}$ (and $\boldsymbol{\Sigma}^{-1}$) as $\boldsymbol{\Sigma} = \mathbf{U}^{-1}\mathbf{V}\mathbf{U}^{-t}$. Furthermore, this decomposition uniquely determines $\boldsymbol{\Sigma}$ via $\mathbf{U}/\mathbf{B}$ and $\mathbf{V}$ (Horn and Johnson, 2012). Thus, in analogy with (6), if we define the set

$$\mathbb{M}(\mathcal{D}) := \left\{ \mathbf{M} \in \mathbb{M}_{p \times p}(\mathbb{R}) \; : \; m_{uv} = 0 \text{ for all } (u, v) \notin A \right\}$$

and the set $\boldsymbol{\Delta}_p$ of $p \times p$ diagonal matrices, we can redefine the *Gaussian directed Markov model* as

$$\mathcal{N}(\mathcal{D}) = \left\{ \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \; : \; \boldsymbol{\Sigma}^{-1} = (\mathbf{I}_p - \mathbf{B})^t \mathbf{V}^{-1}(\mathbf{I}_p - \mathbf{B}), \mathbf{B} \in \mathbb{M}(\mathcal{D}), \mathbf{V} \in \boldsymbol{\Delta}_p \right\}. \tag{8}$$

## 6. Maximum likelihood estimation

Maximum likelihood estimation is greatly simplified in exponential family theory (Barndorff-Nielsen, 1978). The multivariate Gaussian distribution is a regular exponential family, and thus both undirected and directed Gaussian Markov models can be expressed as special subfamilies of it.

### 6.1. The Gaussian family and maximum likelihood

In the multivariate Gaussian family the canonical parameter is $\boldsymbol{\eta} = (\boldsymbol{\Omega}\boldsymbol{\mu}, -\boldsymbol{\Omega}/2)$, over the space $\mathcal{H} = \{(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \; : \; \boldsymbol{\eta}_1 \in \mathbb{R}^p, -\boldsymbol{\eta}_2 \in \mathbb{S}^{>0}\}$ and the sufficient statistics are $\boldsymbol{t}(\boldsymbol{X}) = (\boldsymbol{X}, \boldsymbol{X}\boldsymbol{X}^t)$. Let $\{\boldsymbol{x}^{(n)} \; : \; 1 \leq n \leq N\}$ be $N$ independent observations, where $\boldsymbol{X}^{(n)} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for each $n \in \{1, \ldots, N\}$, arranged in $\mathbf{x} \in \mathbb{M}_{p \times N}(\mathbb{R})$, the respective random matrix being $\mathbf{X}$. The random sample is also a regular exponential family with canonical parameter $\boldsymbol{\eta} = (\boldsymbol{\Omega}\boldsymbol{\mu}, -\boldsymbol{\Omega}/2)$ over the space $\mathcal{H}$. The sufficient statistics in this case are $\boldsymbol{t}(\mathbf{X}) = (N\bar{\mathbf{X}}, \mathbf{X}\mathbf{X}^t)$ with $N\bar{\mathbf{X}} = \sum_{n=1}^N \boldsymbol{X}^{(n)}$.

In a regular exponential family $\mathcal{F}_{\mathcal{H}}$, a maximum of the likelihood function, $\mathcal{L}(\boldsymbol{\eta})$, given a random sample $\mathbf{X} = \mathbf{x}$, is reached in $\mathcal{H}$ if and only if $\boldsymbol{t}(\mathbf{x})$ belongs to the interior of $\mathcal{C}(\boldsymbol{t})$, the closed convex hull of the support of the distribution of $\boldsymbol{t}$, denoted as $\mathrm{int}(\mathcal{C}(\boldsymbol{t}))$. In such case, it is unique and given by the $\boldsymbol{\eta} \in \mathcal{H}$ satisfying $\mathbb{E}[\boldsymbol{t}(\mathbf{X})] = \boldsymbol{t}(\mathbf{x})$.

For the multivariate Gaussian random sample, we have that $\mathbb{E}[N\bar{\mathbf{X}}] = N\boldsymbol{\mu}$ and $\mathbb{E}[\mathbf{X}\mathbf{X}^t] = N\boldsymbol{\Sigma} + N\boldsymbol{\mu}\boldsymbol{\mu}^t$, thus the convex support of $\boldsymbol{t}(\mathbf{X}) = (N\bar{\mathbf{X}}, \mathbf{X}\mathbf{X}^t)$ is $\mathcal{C}(\boldsymbol{t}) = \{(\boldsymbol{v}, \mathbf{M}) \in \mathbb{R}^p \times \mathbb{S}_p \; : \; \mathbf{M} - \boldsymbol{v}\boldsymbol{v}^t/N \in \mathbb{S}^{\geq 0}\}$. This gives that the maximum likelihood estimator for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ exists if and only if $\mathbf{x}\mathbf{x}^t - N\bar{\mathbf{x}}\bar{\mathbf{x}}^t \in \mathbb{S}^{>0}$, which happens with probability one whenever $N > p$ and never otherwise. The solution in such case is $(\bar{\mathbf{x}}, \mathbf{Q}/N)$, where

$$\mathbf{Q} = \sum_{n=1}^N \left( \boldsymbol{X}^{(n)} - \bar{\mathbf{X}} \right) \left( \boldsymbol{X}^{(n)} - \bar{\mathbf{X}} \right)^t = \mathbf{X}\mathbf{X}^t - N\bar{\mathbf{X}}\bar{\mathbf{X}}^t.$$

A particular situation, usually assumed, is when $\boldsymbol{\mu} = \mathbf{0}$. The canonical parameter now is $\boldsymbol{\eta} = -\boldsymbol{\Omega}/2$ in the space $\{\boldsymbol{\eta} \; : \; -\boldsymbol{\eta} \in \mathbb{S}^{>0}\}$, and the sufficient statistic is $\boldsymbol{t}(\mathbf{X}) = \mathbf{X}\mathbf{X}^t$. The maximum likelihood estimator exists if and only if $\mathbf{x}\mathbf{x}^t \in \mathbb{S}^{>0}$, and in such case it is $\mathbf{X}\mathbf{X}^t/N = \mathbf{Q}/N$.

## 6.2. Gaussian Markov models as exponential families

When $\mathcal{G}$ is an undirected graph, the set $\mathbb{S}^{\succ 0}(\mathcal{G})$ is a convex (linear) cone inside the positive definite cone $\mathbb{S}^{\succ 0}$ (e.g. Uhler, 2018, §9.2), which means that $\mathbb{R}^p \times \mathbb{S}^{\succ 0}(\mathcal{G})$ is an affine subspace of $\mathbb{R}^p \times \mathbb{S}^{\succ 0}$, and thus $\mathcal{N}(\mathcal{G})$ is also a regular exponential family (Barndorff-Nielsen, 1978). Assume that $\boldsymbol{\mu} = \mathbf{0}$ and let $\mathbf{Q}^{\mathcal{G}}$ be the projection of $\mathbf{Q}$ on $E \cup \{uu \; : \; u \in V\}$, that is, such that $q_{uv}^{\mathcal{G}} = 0$ for all $uv \notin E$ with $u \neq v$. Since $\mathcal{L}(\boldsymbol{\Omega}) \propto \det(\boldsymbol{\Omega})^{\frac{1}{2}} \exp\left(-\operatorname{tr}(\boldsymbol{\Omega}\mathbf{Q})\right)$ and $\boldsymbol{\Omega} \in \mathbb{S}^{\succ 0}(\mathcal{G})$, we have $\operatorname{tr}(\boldsymbol{\Omega}\mathbf{Q}) = \operatorname{tr}(\boldsymbol{\Omega}\mathbf{Q}^{\mathcal{G}})$ and the sufficient statistic for $\mathcal{N}(\mathcal{G})$ is $\boldsymbol{t}(\mathbf{x}) = \mathbf{Q}^{\mathcal{G}}$ (Lauritzen, 1996). Its convex support is $\mathcal{C}(\boldsymbol{t}) = \{\mathbf{P}^{\mathcal{G}} \; : \; \mathbf{P} \in \mathbb{S}^{\geq 0}\}$, equivalently called the set of projections extendible to full positive definite matrices. Thus, the maximum likelihood estimator for $\boldsymbol{\Sigma}$ exists if and only if $\mathbf{Q}^{\mathcal{G}} \in \operatorname{int}(\mathcal{C}(\boldsymbol{t}))$, which is equivalent to say that $\mathbf{Q}^{\mathcal{G}}$ is extendible to a full positive definite matrix. Whenever it exists, it is the only extendible matrix $\widehat{\boldsymbol{\Sigma}}$ that also satisfies the model restriction $\widehat{\boldsymbol{\Sigma}}^{-1} \in \mathbb{S}^{\succ 0}(\mathcal{G})$. A sufficient condition thus is that $\mathbf{Q} \in \mathbb{S}^{\succ 0}$, which happens almost surely for $N \geq p$. Recovering $\widehat{\boldsymbol{\Sigma}}$ is a convex optimization problem, Uhler (2018) §9.6 overviews some of the algorithms available for its computation. Note however that if $\mathcal{G}$ is chordal, then there is a closed form expression for $\widehat{\boldsymbol{\Omega}}$ (Lauritzen, 1996)

The existence of $\widehat{\boldsymbol{\Sigma}}$ has been completely characterized when $\mathcal{G}$ is chordal by Grone et al. (1984) and Frydenberg and Lauritzen (1989), independently. Since finding $\widehat{\boldsymbol{\Sigma}}$ is equivalent to a positive definite matrix completion problem (Uhler, 2018, §9.3), the problem lies at the interface between statistics and linear algebra. Therefore, this problem has been solved from an algebraic (Sturmfels and Uhler, 2010; Uhler, 2012, 2018, §9.4) viewpoint for a general, non-chordal $\mathcal{G}$. However, the conditions on the sample size $N$ are still unknown except for certain non-chordal graph types, see Uhler (2018) §9.5 for an up-to-date overview of the advances made so far.

Now we turn on the case where the random sample $\mathbf{X}$ is assumed to follow multivariate Gaussian distribution constrained by the separation properties in an acyclic digraph. The restriction in (8), however, is not linear in the canonical parameter; in fact, Spirtes et al. (1997) show that they are curved exponential families. To obtain the maximum likelihood estimates, theory from multivariate linear regression can be applied (Andersson and Perlman, 1998). Recall that if $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathcal{N}(\mathcal{D})$, then $\boldsymbol{X}$ can be expressed as (7). Thus, we can estimate the $\mathcal{D}$-parameters for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as the usual least squares estimators,

$$\widehat{\boldsymbol{\beta}}_{u \cdot \operatorname{pa}(u)}^t = \mathbf{Q}_{u \operatorname{pa}(u)} \mathbf{Q}_{\operatorname{pa}(u) \operatorname{pa}(u)}^{-1},$$

$$\widehat{\xi}_u = \bar{x}_u - \widehat{\boldsymbol{\beta}}_{u \cdot \operatorname{pa}(u)}^t \bar{\mathbf{x}}_{\operatorname{pa}(u)},$$

$$N\widehat{v}_{uu} = q_{uu} - \widehat{\boldsymbol{\beta}}_{u \cdot \operatorname{pa}(u)}^t \mathbf{Q}_{u \operatorname{pa}(u)}^t,$$

respectively for each $u \in V$, where $q_{uu}$ is the $u$th diagonal entry in $\mathbf{Q}$. We can then obtain directly the maximum likelihood estimator for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from their respective $\mathcal{D}$-parameter estimators (see Andersson and Perlman, 1998, for an algorithm). As opposed to the undirected case, $(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$ exist with probability one if and only if $N \geq p + \max\{|\operatorname{pa}(u)| \; : \; u \in V\}$ (Anderson, 2003). Recently, Ben-David and Rajaratnam (2012) analyse in detail the relationship between the $\mathcal{D}$-parameters and $\boldsymbol{\Sigma}$ as a positive definite matrix completion problem, in analogy with the undirected case.

## 7. Model selection via hypothesis testing

Maximum likelihood estimators, presented in the previous section, can be used to address the problem of model estimation, and require either prior knowledge or a statistical procedure that allows model selection; that is, selecting the graph that will define the Markov model. In this section we will review the main hypothesis testing methods for such task.

Throughout the section, for $u, v \in V = \{1, \ldots, p\}$ and $U \subseteq V \setminus \{u, v\}$, we will denote as $\rho_{uv \cdot U}$ the partial correlation coefficient between $u$ and $v$ given the variables in $U$, and as $r_{uv \cdot U}$ its maximum likelihood estimator, the sample partial correlation (see e.g. Anderson, 2003, §4.3 for an introduction to partial correlation theory).

## 7.1. Stepwise selection

In the undirected case, we are interested in testing the hypothesis $H_0 : \boldsymbol{\Omega} \in \mathbb{S}^{\succ 0}(\mathcal{G}_0)$ against $H_1 : \boldsymbol{\Omega} \in \mathbb{S}^{\succ 0}(\mathcal{G})$, where $\mathcal{G}_0 = (V, E_{\mathcal{G}_0}) \subseteq \mathcal{G} = (V, E_{\mathcal{G}})$. The result of such test determines whether the edges in $E_{\mathcal{G}} \setminus E_{\mathcal{G}_0}$ should be excluded from the selected model; that is why these tests are usually known as *edge exclusion tests*. Note also that this is *backward* model selection, since our null hypothesis consists on a subgraph. Let $\widehat{\boldsymbol{\Sigma}}_0$ and $\widehat{\boldsymbol{\Sigma}}$ be the maximum likelihood estimators for a covariance matrix in the Markov model determined by $\mathcal{G}_0$ and $\mathcal{G}$, respectively. The likelihood ratio statistic is

$$T_L = \frac{\det(\widehat{\boldsymbol{\Sigma}})^{N/2}}{\det(\widehat{\boldsymbol{\Sigma}}_0)^{N/2}} = \left(\frac{\det(\widehat{\boldsymbol{\Omega}}_0)}{\det(\widehat{\boldsymbol{\Omega}})}\right)^{N/2}.$$

Under $H_0$, $-2\log(T_L) = N(\log\det(\widehat{\boldsymbol{\Omega}}) - \log\det\widehat{\boldsymbol{\Omega}}_0)$ is asymptotically distributed as a $\chi^2$ distribution with $|E_{\mathcal{G}}| - |E_{\mathcal{G}_0}|$ degrees of freedom; however, this is a poor approximation in many cases (Porteous, 1989). More accurate distributional results have been derived by Eriksen (1996), as follows. Let $\mathcal{G}_0 \subset \ldots \subset \mathcal{G}_k(= \mathcal{G})$ be a sequence of graphs where, for

$1 \leq i \leq k$, $E_{\mathcal{G}_{i-1}} = E_{\mathcal{G}_i} \setminus \{e_i\}$ for some $e_i = u_i v_i \in E_{\mathcal{G}_i}$ (sequence of edge deletions). Then, under $H_0$, $T_L^{2/N}$ is distributed as the product $\prod_{i=1}^{k} B_i$ of univariate Beta variables, where, for $1 \leq i \leq k$,

$$B_i \sim \mathcal{B}\left(\frac{1}{2}(N - |\mathrm{ne}_{\mathcal{G}_i}(u_i) \cap \mathrm{ne}_{\mathcal{G}_i}(v_i)| - 2), \frac{1}{2}\right).$$

The above result is exact whenever $\mathcal{G}$ and $\mathcal{G}_0$ are chordal or share the same non-chordal maximal subgraphs (Eriksen, 1996). Specifically, denote as $C_i^* = \mathrm{ne}(u_i) \cap \mathrm{ne}(v_i)$ the unique clique in $\mathcal{G}_i$ of which edge $e_i$ is a member. Then, under $H_0$ (see e.g. Lauritzen, 1996, Proposition 5.14)

$$T_L^{2/N} = \prod_{i=1}^{k} (1 - r_{u_i v_i \cdot C_i^* \setminus \{u_i v_i\}}^2),$$

giving that $T_L^{2/N}$ is distributed as $\prod_{i=1}^{k} \mathcal{B}((N - |C_i^*|)/2, 1/2)$. Note that in this decomposable case one avoids to actually compute $\widehat{\Omega}$ and $\widehat{\Omega}_0$. The statistic $T_L$ has been used for model selection in undirected Gaussian Markov models by Wermuth (1976b).

In the case of a directed Gaussian Markov model over an acyclic digraph $\mathcal{D} = (V, A)$, most of the results are adaptations from analogues in multivariate linear Gaussian models. The likelihood ratio, whose moments are also characterized in Andersson and Perlman (1998), is

$$T_L = \frac{\det(\widehat{\Sigma})^{\frac{N}{2}}}{\det(\tilde{\Sigma})^{\frac{N}{2}}} = \frac{\prod_{v \in V} \left| \widehat{\sigma}_{vv} - \widehat{\boldsymbol{\sigma}}_{v \cdot \mathrm{pa}(v)}^t \widehat{\Sigma}_{\mathrm{pa}(v)}^{-1} \widehat{\boldsymbol{\sigma}}_{v \cdot \mathrm{pa}(v)} \right|}{\prod_{v \in V} \left| \tilde{\sigma}_{vv} - \tilde{\boldsymbol{\sigma}}_{v \cdot \mathrm{pa}(v)}^t \tilde{\Sigma}_{\mathrm{pa}(v)}^{-1} \tilde{\boldsymbol{\sigma}}_{v \cdot \mathrm{pa}(v)} \right|},$$

where $\tilde{\Sigma}$ and $\widehat{\Sigma}$ are the respective maximum likelihood estimators for $\tilde{\mathcal{D}}$ and $\mathcal{D}$, $\tilde{\mathcal{D}} \subseteq \mathcal{D}$.

A backward stepwise method has become popular for selecting $\mathcal{D}$, commonly called the *PC algorithm* (Spirtes et al., 2000). This method proceeds by first finding an estimator of the skeleton, $\widehat{\mathcal{D}^U}$, from a complete undirected graph, and then orienting it. At iteration $i$ of the first step (finding $\widehat{\mathcal{D}^U}$), $H_0 : X_u \perp\!\!\!\perp X_v \mid X_C$ is tested, with $C = \widehat{\mathrm{ne}}(u) \setminus \{v\}$ and $|C| = i$. The edge $uv$ is removed from $\widehat{\mathcal{D}^U}$ if $H_0$ is not rejected. Note that $\widehat{\mathcal{D}^U}$ depends on the order in which $H_0$ is tested at each iteration, problem circumvented in the modification by Colombo and Maathuis (2014). Assuming that $\mathcal{I}(\boldsymbol{X}) = \mathcal{I}(\mathcal{D})$ (see Section 4), commonly called the *faithfulness* assumption, Robins et al. (2003) showed that the PC algorithm is pointwise consistent but may not be uniformly consistent, regardless of the method used for testing $H_0$. Zhang and Spirtes (2003) approached this problem by introducing a stronger condition, called *strong faithfulness*, which, by requiring nonzero partial correlations to have a common lower bound, gives uniform consistency, even in a high-dimensional setting (Kalisch and Bühlmann, 2007). However, despite the set of 'unfaithful' distributions has Lebesgue measure zero (Meek, 1995), those 'strongly unfaithful' constitute a non-zero Lebesgue measure set, which can in some cases be very large (Uhler et al., 2013; Lin et al., 2014).

### 7.2. Multiple testing

When performing model selection with these tests, multiple testing error rates need to be controlled. For overcoming this, Drton and Perlman (2004) propose an alternative to the previous stepwise methods. First, note that both acyclic directed and undirected Gaussian Markov models over $V = \{1, \ldots, p\}$ are characterized by certain partial correlation coefficients, since for $u, v \in V$ and $U \subseteq V \setminus \{u, v\}$, we have $X_u \perp\!\!\!\perp X_v \mid \boldsymbol{X}_U \iff \rho_{uv \cdot U} = 0$. Assuming conditional independence, that is, $\rho_{uv \cdot U} = 0$, then $\sqrt{N - |U| - 2}\, r_{uv \cdot U} / \sqrt{1 - r_{uv \cdot U}^2}$ has a $t$ distribution with $N - |U| - 2$ degrees of freedom. However, a faster Gaussian approximation can be obtained using Fisher's $Z$-transform,

$$Z(x) = \frac{1}{2} \log\left(\frac{1+x}{1-x}\right) = \tanh^{-1}(x).$$

In such case the distribution of $\sqrt{N - |U| - 3}\, Z(r_{uv \cdot U})$ tends to a standard Gaussian.

Based on the above discussion, Drton and Perlman (2004) propose a method where a set of simultaneous $p$-values and confidence intervals is obtained such that the edge set is estimated, for a significance level $\alpha$ and using Sidak (1967) inequality, as

$$\widehat{E}^\alpha := \left\{ uv : \sqrt{N - p - 1} \left| Z(r_{uv \cdot V \setminus \{u,v\}}) \right| > \Phi^{-1}\left( \frac{1}{2}(1 - \alpha)^{\frac{2}{p(p-1)}} + \frac{1}{2} \right) \right\}, \tag{9}$$

where $\Phi$ is the cumulative distribution function of a standard Gaussian. Denoting as $\widehat{\mathcal{G}}^\alpha = (V, \widehat{E}^\alpha)$, it holds $\liminf_{N \to \infty} P(\widehat{\mathcal{G}}^\alpha = \mathcal{G}) \geq 1 - \alpha$ if the distribution under consideration $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \mathcal{N}(\mathcal{G})$ is faithful to $\mathcal{G}$, that is, if $\omega_{uv} = 0 \iff uv \notin E$. If faithfulness is not satisfied, then the result holds with respect to the smallest graph $\mathcal{H}$ such that $\mathcal{G} \subseteq \mathcal{H}$ and $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is faithful to $\mathcal{H}$.

The multiple testing procedure in (9) has also been extended in Drton and Perlman (2008), obtaining an estimate of the arc set as

$$\widehat{A}^{\alpha} := \left\{ (v, u) : v < u \text{ and } \sqrt{N - u - 1} \left| Z(r_{uv \cdot \mathrm{pr}(u) \setminus \{v\}}) \right| > \Phi^{-1} \left( \frac{1}{2}(1 - \alpha)^{\frac{2}{p(p-1)}} + \frac{1}{2} \right) \right\}, \quad (10)$$

where an ancestral ordering $\prec$ is being assumed in $V$ such that the resulting permutation is the identity; that is, such that $v \prec u \iff v < u$. Consistency is established as in the undirected case; note the symmetry with (9). See Drton and Perlman (2007) for a general discussion on some variations of (9) and (10) and their impact on overall error control. Recently, Liu (2013) has extended the methodology of Drton and Perlman (2004) to the high dimensional scenario, with $p > N$.

A related testing procedure has emerged motivated by the field of gene network learning from microarray data. Instead of testing full partial correlations $\rho_{uv \cdot V \setminus \{u,v\}}$ in an undirected model, only limited $q$-order partial correlations $\rho_{uv \cdot U} = 0$, where $U \subseteq V \setminus \{u, v\}$ and $|U| = q$, are tested (Castelo and Roverato, 2006). An edge is added to the resulting graph, called *q-partial graph*, only when all of the $q$-partial correlations are rejected to be zero. This procedure is specially suited for situations where the number of variables is substantially larger than the number of instances, as happens in the case of microarray data, where low order conditional independence relationships (up to $q \le 3$) have been popular (Wille and Bühlmann, 2006; de la Fuente et al., 2004; Magwene and Kim, 2004). Castelo and Roverato (2006) generalize and formalize these approaches, and provide a robust model selection procedure for $q$-partial graphs. This is intended to serve as an intermediate step for model selection of a classical undirected Markov model $\mathcal{N}(\mathcal{G})$, and yields to a great simplification when $\mathcal{G}$ is sparse (Castelo and Roverato, 2006).

## 8. Regularization

Regularization approaches, which perform model selection and estimation in a simultaneous way, have become popular in the context of Markov models. They are usually applied when $N < p$, and thus the existence of the maximum likelihood estimator is not guaranteed. The main consistency results available for both the directed and undirected cases share sparseness and high-dimensionality assumptions, as we will see below. There are two different approaches, those that penalize the likelihood and those that instead focus on the regression coefficients.

Throughout this section, we will employ the asymptotic notation, specifically symbols $O(\cdot)$ and $\Theta(\cdot)$, asymptotic inferiority and equivalence, respectively. For $\mathbf{M} \in \mathbb{M}_{q \times r}(\mathbb{R})$, vec$(\mathbf{M})$ will denote the vectorized function of $\mathbf{M}$, $(m_{11}, \ldots, m_{q1}, \ldots, m_{1r}, \ldots, m_{qr})^t$. This way, the operator norm of $\mathbf{M}$ will be denoted as $\|\mathbf{M}\|$; whereas $\|\mathbf{M}\|_{q+r}$ will be used to denote $\|\text{vec}(\mathbf{M})\|_{q+r}$, being $\|\cdot\|_p$ the $p$-norm function. If $v$ is a $p$-vector, diag$(v)$ will denote the matrix $\mathbf{M}$ in $\Delta_p$ with main diagonal $v$; analogously, diag$(\mathbf{M}) \in \Delta_p$ will have the same diagonal as $\mathbf{M}$, and $\mathbf{M}^-$ will be used for $\mathbf{M} - \text{diag}(\mathbf{M})$.

### 8.1. Node-wise regression

Let $\mathcal{G} = (V, E)$ be an undirected graph, with $V = \{1, \ldots, p\}$. Let $X$ be a random vector whose distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ belongs to the undirected Gaussian Markov model $\mathcal{N}(\mathcal{G})$. Assume, for notational simplicity, that $\boldsymbol{\mu} = \mathbf{0}$, and, following the notation of Section 6, let $\mathbf{X} = \mathbf{x}$ be a $p \times N$ random sample from $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. Since for each $u, v \in V$, $\beta_{uv \cdot V \setminus \{u\}} = -\omega_{uv}/\omega_{uu}$ (Eq. (3)), then

$$X_u \perp\!\!\!\perp X_v \mid \boldsymbol{X}_{V \setminus \{u,v\}} \iff \omega_{uv} = 0 \iff \beta_{uv \cdot V \setminus \{u\}} = \beta_{vu \cdot V \setminus \{v\}} = 0.$$

This means that an analogue of the matrix $\mathbf{B}$ in directed Gaussian Markov models (Eq. (7)) can be used for determining the missing edges in the undirected case. In Meinshausen and Bühlmann (2006), this is done in the regression function, as

$$\widehat{\boldsymbol{b}}_u^{\lambda} := \underset{\boldsymbol{b}_u \in \mathbb{R}^p, \, b_{uu}=0}{\arg\min} \left( \frac{1}{N} \left\| \mathbf{x}_u^t - \mathbf{x}^t \boldsymbol{b}_u \right\|_2^2 + \lambda f(\boldsymbol{b}_u) \right), \quad (11)$$

where $\lambda \ge 0$, $\mathbf{x}_u$ is the $u$-the row vector of $\mathbf{x}$ and $f(\cdot)$ is the penalty function. For each $v \in V \setminus \{u\}$, $\widehat{b}_{uv}^{\lambda}$ gives an estimate of $\beta_{uv \cdot V \setminus \{v\}}$. Let $\widehat{\text{ne}}(v) := \{u \in V : \widehat{b}_{vu}^{\lambda} \ne 0\}$. while $u \in \text{ne}(v) \iff v \in \text{ne}(u)$ for all $u, v \in V$, this may not be true for $\widehat{\text{ne}}(u)$ and $\widehat{\text{ne}}(v)$. Hence, two different estimators for the edge set $E$ may be defined

$$\widehat{E}_{\wedge} := \{uv : u \in \widehat{\text{ne}}(v) \text{ and } v \in \widehat{\text{ne}}(u)\},$$
$$\widehat{E}_{\vee} := \{uv : u \in \widehat{\text{ne}}(v) \text{ or } v \in \widehat{\text{ne}}(u)\}.$$

Let $f(\cdot) = \|\cdot\|_1$, commonly known as the *lasso* penalty (Tibshirani, 1996) or $l_1$ regularization. Then both estimators $\widehat{E}_{\wedge}$ and $\widehat{E}_{\vee}$ are consistent for certain choice of $\lambda$. This result was independently discovered by Meinshausen and Bühlmann (2006), Zhao and Yu (2006), Zou (2006) and Yuan and Lin (2007b). It relies on the following almost necessary and sufficient condition

$$\left| \sum_{z \in \text{ne}(v)} \text{sign}(\beta_{vz \cdot \text{ne}(v)}) \beta_{uz \cdot \text{ne}(v)} \right| < 1, \quad (12)$$

**Table 1**

Comparison of assumptions for consistency results on regression based penalized estimation in acyclic directed and undirected Gaussian Markov models.

| $\mathcal{N}(\mathcal{G})$ (Meinshausen and Bühlmann, 2006) | $\mathcal{N}(\mathcal{D})$ (van de Geer and Bühlmann, 2013) |
|---|---|
| $l_1$ regularization | $l_0$ regularization |
| Lower bound on $\left\vert \rho_{uv \cdot V \setminus \{u,v\}} \right\vert$ | Lower bound on $\left\vert \beta_{uv \cdot \mathrm{pr}(u)} \right\vert$ |
| Upper bound on $\vert \mathrm{ne}(v) \vert$ | Upper bound on $\vert \mathrm{pa}(v) \vert$ |
| Bounded neighbourhood perturbations | Bounded permutation perturbations |

This node-wise regression approach may also be used to perform model selection for acyclic directed Gaussian Markov models if there is a known order among the variables, see for example Shojaie and Michailidis (2010) or Yu and Bien (2017) and references therein. From Eq. (7), the regression function to penalize in this case would be, for each $u \in V$,

$$X_u = \mu_u + \sum_{v \in \mathrm{pr}(u)} \beta_{uv \cdot \mathrm{pr}(u)}(X_v - \mu_v) + E_u \tag{13}$$

The condition of Eq. (12), commonly called the 'irrepresentable condition' (Zhao and Yu, 2006) or 'neighbourhood stability' (Meinshausen and Bühlmann, 2006), is inherent to model selection in linear regression with $l_1$ regularization, and thus it also holds when penalizing (13) with the $l_1$ penalty. However, some variants have been proposed because it is rather restrictive. These alternatives usually rely on thresholding the regression coefficients or adding weights in the $l_1$ penalty, that under milder assumptions still achieve model selection consistency (Meinshausen and Yu, 2009) or other attractive, 'oracle' properties (van de Geer and Bühlmann, 2009); see Bühlmann and van de Geer (2011), §7 for a review. van de Geer and Bühlmann (2009) show that although model selection consistency for neighbourhood selection may be restrictive, sufficient conditions for such oracle properties hold fairly generally.

### 8.2. Penalized likelihood

In van de Geer and Bühlmann (2013), $l_0$ regularization is alternatively used in the context of directed Gaussian Markov models, without assuming a known order. As in Meinshausen and Bühlmann (2006), the regression coefficients are penalized in their approach, more generally, the $\mathcal{D}$-parameters in the likelihood function (assuming $\boldsymbol{\mu} = \mathbf{0}$). As such, the assumptions required for the consistency of both methods share some symmetry, as we have outlined in Table 1. The estimators in this case are obtained as

$$(\widehat{\mathbf{V}}^\lambda, \widehat{\mathbf{B}}^\lambda) = \underset{\mathbf{\Omega} = (\mathbf{I}_p - \mathbf{B})^t \mathbf{V}^{-1}(\mathbf{I}_p - \mathbf{B}),\, \mathbf{B} \in \mathbb{M}(\mathcal{D}),\, \mathbf{V} \in \mathbf{\Delta}_p}{\arg\min} \; (\mathrm{tr}(\mathbf{\Omega S}) - N \log \det(\mathbf{\Omega}) + \lambda f(\mathbf{\Omega})),$$

where $\lambda \geq 0$, $\mathbb{M}(\mathcal{D})$, $\mathbf{\Delta}_p$ are as in Eq. (8) and $\mathbf{S} = \mathbf{xx}^t/N$. When $f(\mathbf{\Omega}) = |\{(u,v) \; : \; b_{uv} \neq 0\}|$ ($l_0$ regularization), $\widehat{\mathbf{V}}^\lambda$ and $\widehat{\mathbf{B}}^\lambda$ are equal among Markov equivalent models and the resulting estimator of the concentration matrix $\widehat{\mathbf{\Omega}}^\lambda$ is consistent for certain choice of $\lambda$ (van de Geer and Bühlmann, 2013). The strong faithfulness condition for the PC algorithm, bounding nonzero partial correlations, resembles the assumptions for regularization methods (Table 1). In fact, $l_0$ regularization has been suggested as an alternative for the PC, in order to avoid the restrictive strong faithfulness assumption (Uhler et al., 2013); however, it is unclear how the assumptions of both methods are related. For recent extensions of the work by van de Geer and Bühlmann (2013), see Aragam and Zhou (2015), Aragam et al. (2017).

In undirected Gaussian Markov models conditional independences can be read from $\mathbf{\Omega}$. Therefore, the penalized likelihood approach can be formulated more directly, for $\lambda \geq 0$, as

$$\widehat{\mathbf{\Omega}}^\lambda = \underset{\mathbf{\Omega} \in \mathbb{S}^{\succ 0}(\mathcal{G})}{\arg\min} \; (\mathrm{tr}(\mathbf{\Omega S}) - N \log \det(\mathbf{\Omega}) + \lambda f(\mathbf{\Omega})). \tag{14}$$

Yuan and Lin (2007a) were the first to pursue this approach, and they chose $f(\mathbf{\Omega}) = \left\Vert \mathbf{\Omega}^- \right\Vert_1$, that is, the off-diagonal elements in $\mathbf{\Omega}$, which determine the edges of the resulting undirected graph, are penalized. Later, in Banerjee et al. (2008) the diagonal elements are included in the regularization function, that is, $f(\mathbf{\Omega}) = \Vert \mathbf{\Omega} \Vert_1$; however, since $1/\omega_{uu} = \sigma_{uu \cdot V \setminus \{u\}}$, this choice for the penalty favours larger values for the error variances in the regression of $X_u$ on the rest of variables (Bühlmann and van de Geer, 2011). Nonetheless, this estimator is the one chosen in the extensively used algorithm *Graphical Lasso* of Friedman et al. (2008), although model selection consistency has only been proved for $f(\mathbf{\Omega}) = \left\Vert \mathbf{\Omega}^- \right\Vert_1$ (Lam and Fan, 2009; Ravikumar et al., 2011). It is not known whether the sufficient conditions required for this consistency are strictly stronger than the irrepresentable condition, as some examples (Meinshausen, 2008) seem to indicate.

For the penalization of Yuan and Lin (2007a) ($f(\mathbf{\Omega}) = \left\Vert \mathbf{\Omega}^- \right\Vert_1$), the convergence rate is (Rothman et al., 2008)

$$\left\Vert \widehat{\mathbf{\Omega}}^\lambda - \mathbf{\Omega} \right\Vert_2 \in O\left( \sqrt{\frac{(|E| + p) \log(p)}{N}} \right) \text{ as } N \to \infty.$$

A relaxation of this rate can be obtained based on the correlation matrix, as follows. Since $\boldsymbol{\Sigma} = \mathbf{DPD}$ with $\mathbf{P}$ the correlation matrix and $\mathbf{D}$ the diagonal matrix of standard deviations, if we let the corresponding sample estimates be $\widehat{\mathbf{D}}^2 = \operatorname{diag}(\widehat{\boldsymbol{\Sigma}})$ and $\widehat{\mathbf{P}} = \widehat{\mathbf{D}}^{-1}\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{D}}^{-1}$, we can then estimate $\mathbf{K} = \mathbf{P}^{-1}$ as

$$\widehat{\mathbf{K}}^\lambda = \underset{\mathbf{K} \in \mathbb{S}^{\succ 0}(\mathcal{G})}{\arg\min} \left( \operatorname{tr}(\mathbf{K}\widehat{\mathbf{P}}) - N \log \det(\mathbf{K}) + \lambda f(\mathbf{K}) \right),$$

for $\lambda \geq 0$. The concentration matrix can then be alternatively estimated as $\tilde{\Omega}^\lambda = \widehat{\mathbf{D}}^{-1}\widehat{\mathbf{K}}^\lambda\widehat{\mathbf{D}}^{-1}$, yielding a convergence rate of Rothman et al. (2008)

$$\left\| \tilde{\Omega}^\lambda - \Omega \right\| \in O\left( \sqrt{\frac{(|E| + 1)\log(p)}{N}} \right) \text{ as } N \to \infty.$$

Convergence rates in other norms are provided in Ravikumar et al. (2011), and they have been generalized by Lam and Fan (2009) for other penalty functions.

## 9. Bayesian model selection and estimation

Consider a continuous multivariate family $\mathcal{F}_\theta$ parametrized by $\theta$, and denote as $f(\mathbf{x} \mid \theta)$ the density function of a random sample $\mathbf{X}$ from $P \in \mathcal{F}_\theta$ for a given value of $\theta$. In Bayesian statistics, $\theta$ is treated as a random variable with known distribution, $f(\theta)$, usually called the *prior* distribution of $\theta$. Inference is then performed based on the value of $f(\theta \mid \mathbf{x}) \propto f(\theta)f(\mathbf{x} \mid \theta)$, the *posterior* distribution of $\theta$ given the information in $\mathbf{X} = \mathbf{x}$.

In Gaussian Markov models, $\theta = (\mu, \Omega, \mathcal{G})$, where in our case $\mathcal{G}$ is either undirected or acyclic directed. Therefore the target probability is $f(\mathcal{G}, \mu, \Omega \mid \mathbf{x}) \propto f(\mathbf{x} \mid \mu, \Omega, \mathcal{G})f(\mu, \Omega \mid \mathcal{G})f(\mathcal{G})$. Integrating out $\mu$ and $\Omega$, we obtain the posterior density of model $\mathcal{G}$, $f(\mathcal{G} \mid \mathbf{x}) \propto f(\mathbf{x} \mid \mathcal{G})f(\mathcal{G})$. The prior for the graph space, $f(\mathcal{G})$, is usually set as uniform. However, this choice is biased towards middle size graphs, and thus other prior distributions (Scutari, 2013, e.g.) have been proposed; see Massam (2018) §10.4.1 and references therein for a recent detailed overview. Bayesian inference for Gaussian graphical models, is usually meant for moderate sample sizes, since it relies on sampling from the resulting posterior distribution, which becomes infeasible in high dimensions (see e.g. Jones et al., 2005 or Massam, 2018).

In the following, the *p-variate Wishart distribution* will be denoted as $\mathcal{W}_p(n, \Lambda)$ with $n \in \mathbb{R}$, $n > p - 1$ and $\Lambda \in \mathbb{M}_{p \times p}(\mathbb{R})$, $\Lambda \succ 0$; analogously, the *p-variate inverse Wishart distribution* will be $\mathcal{W}_p^{-1}(\nu, \Psi)$ with $\nu \in \mathbb{R}$, $\nu > p - 1$ and $\Psi \in \mathbb{M}_{p \times p}(\mathbb{R})$, $\Psi \succ 0$.

### 9.1. Hyper Markov laws

When $\mathcal{G}$ is undirected and decomposable, and assuming $\mu = \mathbf{0}$, Dawid and Lauritzen (1993) proposed for $f(\Omega \mid \mathcal{G})$ what are known as the *hyper Markov laws*. These are defined in terms of properties of the graph associated with the Markov model, mimicking Markov properties. Specifically, let $\theta$ be a random variable taking values over $\mathcal{N}(\mathcal{G})$ and for subsets $A, B \subseteq V$, denote as $\theta_A$ and $\theta_{B|A}$ the parameters of the marginal distribution of $\mathbf{X}_A$ and the conditional distribution of $\mathbf{X}_A$ given values of $\mathbf{X}_B$, respectively. The probability distribution of $\theta$ is said to be *(weakly) hyper $\mathcal{G}$-Markov* if, for any decomposition $(A, B, S)$ of $\mathcal{G}$, it holds that $\theta_{A \cup S} \perp\!\!\!\perp \theta_{B \cup S} \mid \theta_S$; if it further holds $\theta_{B \cup S|A \cup S} \perp\!\!\!\perp \theta_{A \cup S}$, it is called *strongly hyper $\mathcal{G}$-Markov*. For chordal graphs, if the probability distribution of $\theta$ is strongly hyper $\mathcal{G}$-Markov with respect to $\mathcal{G}$, then the probability distribution of $\theta \mid \mathbf{x}$ is the unique (strong) hyper $\mathcal{G}$-Markov distribution specified by the clique-marginal distributions $\{\mathbb{P}(\theta_C \mid \mathbf{x}_C) : C \in \mathfrak{C}(\mathcal{G})\}$; and, when densities exist, $f(\theta_C \mid \mathbf{x}) \propto f(\mathbf{x}_C \mid \theta_C)f(\theta_C)$ (Dawid and Lauritzen, 1993), where $\mathbf{x}_C$ stands for all the observations in $\mathbf{x}$ corresponding to the variables in $C$. That is, under these assumptions, it is possible to localize computations over the graph cliques when performing Bayesian inference.

In a multivariate Gaussian distribution $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, the inverse Wishart is a conjugate prior for $\boldsymbol{\Sigma}$; that is, if $\boldsymbol{\Sigma} \sim \mathcal{W}_p^{-1}(\nu, \Psi)$, then $\boldsymbol{\Sigma} \mid \mathbf{Q}/N \sim \mathcal{W}_p^{-1}(N + \nu, \mathbf{Q} + \Psi)$ (recall $\mathbf{Q} = \mathbf{x}\mathbf{x}^t$). We can thus construct the *hyper inverse Wishart distribution*, as the unique hyper Markov distribution associated with inverse Wishart clique marginals: $\boldsymbol{\Sigma}_{CC} \sim \mathcal{W}_{|C|}^{-1}(\nu, \Psi^C)$, for each clique $C \in \mathfrak{C}(\mathcal{G})$. This hyper Markov distribution is denoted as $\mathcal{HW}_p^{-1}(\nu, \Psi)$, where $\Psi \in \mathbb{S}^{\succ 0}$ such that $\Psi_{CC} = \Psi^C$ for each clique $C \in \mathfrak{C}(\mathcal{G})$. From the discussion above, we know that this distribution is strongly hyper $\mathcal{G}$-Markov. The main advantage of this prior is that it has many properties that mirror those for Markov models, since hyper Markov distributions are also defined in terms of an underlying graph.

Since its introduction, the hyper inverse Wishart prior for decomposable graphs has been extensively studied. The explicit expression for its density is devised in, e.g., Giudici (1996) or Roverato (2000). In order to set its parameters, a hierarchical approach such as in Giudici and Green (1999) can be followed, where $\nu$ and $\Psi$ are assumed to have a Gamma and Wishart distribution, respectively. Since the absent edges of $\mathcal{G}$ correspond to zeros in $\Omega$ (Eq. (4)), Roverato (2000) derives the distribution induced on $\Omega$ by assuming $f(\boldsymbol{\Sigma} \mid \mathcal{G}) = \mathcal{HW}_p^{-1}(\nu, \Psi)$. He shows that in that case, the density is proportional to that of a Wishart matrix conditioned on the event $\Omega \in \mathbb{S}^{\succ 0}(\mathcal{G})$, and calls such prior distribution on $\Omega$ the *$\mathcal{G}$-conditional Wishart*. Recently, Massam (2018) §10.3.2 has provided a detailed overview of the properties of the hyper inverse Wishart, and technical considerations as how to sample from it or perform Bayesian model selection using Bayes factors. Letac and Massam (2007) generalize both the $\mathcal{G}$-conditional and hyper inverse Wishart to a broader

conjugate family, allowing for more than one shape parameter, which is used for model selection by Rajaratnam et al. (2008) (see also Massam, 2018, §10.3.3).

The hyper inverse Wishart has been extended to non-chordal graphs by Roverato (2002), based on properties of the Isserlis matrix of $\mathbf{\Sigma}$ (Roverato and Whittaker, 1998). However, Bayesian model selection in this scenario requires the evaluation of the $\mathcal{G}$-conditional Wishart normalizing constant, which becomes a problem since it did not have a known closed-form expression for a general non-chordal $\mathcal{G}$ until very recently (Uhler et al., 2018). Much of the literature therefore has been devoted to this issue: Atay-Kayis and Massam (2005) analysed the Cholesky decomposition of $\mathbf{\Omega}$ and its relation with the cone $\mathbb{S}^{>0}(\mathcal{G})$ and positive definite matrix completions (Section 6), Carvalho et al. (2007) and Wang and Carvalho (2010) used such theoretical analysis to provide a direct sampler from the hyper inverse Wishart prior, etc. A recent detailed presentation of this computational body of research can be found in Massam (2018), §10.4. Note that, although Uhler et al. (2018) provide exact formulas and examples for special types of graphs, it still remains to find efficient methods for their computation.

*9.2. Priors for acyclic directed models*

The methodology by Geiger and Heckerman (2002) for acyclic directed Gaussian Markov models can be seen as an extension of hyper Markov distributions to such context, since they both coincide for chordal skeletons.

If $\mathcal{D}_c$ is an arbitrary complete digraph, then, under some assumptions on $f(\boldsymbol{\mu}, \mathbf{\Omega} \mid \mathcal{D})$ and $f(\mathbf{x} \mid \mathcal{D}, \boldsymbol{\mu}, \mathbf{\Omega})$, computations can be localized as

$$f(\mathbf{x} \mid \mathcal{D}) = \prod_{v \in V} \frac{f(\mathbf{x}_{\{v\} \cup \text{pa}_{\mathcal{D}}(v)} \mid \mathcal{D}_c)}{f(\mathbf{x}_{\text{pa}_{\mathcal{D}}(v)} \mid \mathcal{D}_c)}, \tag{15}$$

The posterior in Eq. (15) is equal among Markov equivalent acyclic digraphs (Geiger and Heckerman, 2002).

The conjugate prior for $(\boldsymbol{\mu}, \mathbf{\Omega})$ is the normal-Wishart distribution, where $\mathbf{\Omega} \sim \mathcal{W}_p(\alpha_{\mathbf{\Omega}}, \mathbf{\Lambda})$ and $\boldsymbol{\mu} \mid \mathbf{\Omega} \sim \mathcal{N}_p\left(\boldsymbol{\mu}_0, (\alpha_{\boldsymbol{\mu}} \mathbf{\Omega})^{-1}\right)$. This yields a normal-Wishart posterior distribution for $(\boldsymbol{\mu}, \mathbf{\Omega}) \mid \mathcal{D}_c$. Using this, Geiger and Heckerman (1994), obtain an explicit expression for each factor in Eq. (15): for $U \subseteq V$,

$$f(\mathbf{x}_U \mid \mathcal{D}_c) = \left(\frac{\alpha_{\boldsymbol{\mu}}}{\alpha_{\boldsymbol{\mu}} + N}\right)^{\frac{|U|}{2}} 2\pi^{-\frac{lN}{2}} \frac{\Gamma_{|U|}\left(\frac{N + \alpha_{\mathbf{\Omega}} - p + |U|}{2}\right)}{\Gamma_{|U|}\left(\frac{\alpha_{\mathbf{\Omega}} - p + |U|}{2}\right)} \frac{|\mathbf{\Lambda}_{UU}|^{\frac{\alpha_{\mathbf{\Omega}} - p + |U|}{2}}}{|\mathbf{R}_{UU}|^{\frac{\alpha_{\mathbf{\Omega}} - p + |U| + N}{2}}},$$

where $\Gamma_p(\cdot)$ is the $p$-dimensional Gamma function, and

$$\mathbf{R} = \mathbf{\Lambda} + \mathbf{Q} + \frac{N \alpha_{\mathbf{\Omega}}}{N + \alpha_{\mathbf{\Omega}}} (\boldsymbol{\mu}_0 - \bar{\mathbf{x}})(\boldsymbol{\mu}_0 - \bar{\mathbf{x}})^t.$$

Furthermore, Geiger and Heckerman (2002) characterize the normal-Wishart prior for $(\boldsymbol{\mu}, \mathbf{\Omega})$ as the only distribution satisfying the global parameter independence assumption,

$$f(\boldsymbol{\theta} \mid \mathcal{D}) = \prod_{v \in V} f(\theta_v \mid \mathcal{D})$$

for every $P_X \in \mathcal{N}(\mathcal{D})$. This condition is required for Eq. (15) to hold.

The above priors have been used by Consonni and Rocca (2012) and Altomare et al. (2013) for objective Bayesian model selection, where $f(\boldsymbol{\theta} \mid \mathcal{D})$ might be improper. For overcoming this, they use fractional Bayes factors (O'Hagan, 1995), which had been also previously used by Carvalho and Scott (2009) for chordal undirected models (see Massam, 2018, §10.6 and references therein for more details). Recently, Ben-David et al. (2016) have proposed a family of priors extending those by Geiger and Heckerman (2002) but including more shape parameters, that is, mimicking those in Letac and Massam (2007) for undirected models. See Rajaratnam (2012) and Cao et al. (2019) for further discussion on these priors.

# 10. Higher level Markov model classes with mixed graphs

As we have seen throughout this review, the classes of acyclic directed and undirected Markov models are intimately related. Therefore, one approach for their unified treatment could be to step to a higher level, and define new Markov model classes containing them as subclasses. In this section we will overview this approach, which has been particularly active in the past few years. The graphs used for these new Markov models are usually called mixed graphs, because unlike purely undirected or acyclic directed graphs, they allow for more than one edge type. We do not aim in this section for a thorough account of the achievements and drawbacks of the different developments since that would take another full paper.

Chain graphs are the first higher level attempt at this unification: they allow two edge types, directed and undirected, and forbid semi-directed cycles. Drton (2009) provides a unifying view of these model classes, focusing on their discrete parametrization: both the undirected and directed edges can have two different interpretations, thus giving rise to four different chain graph model classes. Among them, *AMP chain graphs* (Andersson et al., 2001), and *LWF chain graphs*

(Lauritzen and Wermuth, 1989; Frydenberg, 1990), named in such way because of the respective paper authors, contain both acyclic directed and undirected Markov model classes.

*Multivariate regression* (MVR) *chain graphs* (Cox and Wermuth, 1993, 1996), or Type IV in Drton (2009), are possibly the most traditional ones and can be viewed as a special case of the path diagrams by Wright (1934). Although they do not contain undirected models, their extension, *regression graphs* (Wermuth and Sadeghi, 2012; Wermuth, 2011), do contain both classes treated in this review, by allowing up to three edge types. The class of regression graphs allows to represent additional relationships in classical sequences of multivariate regressions by means of a bi-directed edge, which could not be otherwise modelled using only the acyclic directed or undirected graphs. These bi-directed edges represent interactions with *latent variables*. The recovery of latent variables is specially relevant in social studies, where the presence of these *confounding* variables may affect the prediction. It has been recently shown that a Gaussian MVR chain graph is Markov equivalent to an acyclic directed Gaussian Markov model with latent variables when its bidirected part is chordal (Fox et al., 2015). Pure Gaussian bidirected graphs represent marginal independences among variables, therefore they impose zero constraints directly in the covariance $\Sigma$.

Finally, the *Type III chain graph* so far has not been devoted much attention (Lauritzen and Sadeghi, 2018). All of the mentioned chain graph model classes are smooth, whereas for their discrete counterparts only the classes of LWF and multivariate regression chain graphs consist of smooth models (Drton, 2009).

The semi-directed cycle constraint on chain graphs can be relaxed, and new graphical model classes are obtained by forbidding only directed cycles. By doing so, we arrive at three different classes of what are called acyclic directed mixed graphs: the so-called *original acyclic directed mixed graph* (oADMG, Richardson, 2003), the *alternative* (aADMG, Peña, 2016a), and *UDAGs* (Peña, 2018), which relax MVR, AMP and LWF chain graphs, respectively. Each oADMG model contains a model obtained from a Bayesian network after marginalizing some of its nodes, (the latent variables). However, other constraints may arise after marginalizing that cannot be represented in terms of conditional independence with this class, for example, the *Verma constraints* (Richardson and Spirtes, 2002, §7.3.1) and *inequality constraints* (Drton et al., 2012). In order to deal with these, Richardson et al. (2012) introduced nested Markov models which also allow for hyper-edges between more than two nodes; however, we are not aware of any Gaussian parametrization.

The classes of both oADMGs and aADMGs are subsumed by the class of *ADM graphs* (Peña, 2018), consisting, naturally, of three edge types. When parametrized with the Gaussian distribution, ADM graph models can be represented as recursive linear equations with two blocks of variables and possibly correlated errors (Koster, 1999; Spirtes, 1995; Peña, 2016a,b). Bidirected edges in these models represent latent confounding effects, whereas undirected edges account for dependence between the errors. Note that, although the classes of ADM graphs and regression graphs allow the same edge types, they are not equivalent since the former contains AMP chain graph models, while the latter does not.

There are other models allowing for up to three edge types, besides the already mentioned ADM and regression graphs: *anterial* and *chain mixed graphs* (Sadeghi, 2016), *ribbonless graphs* (Sadeghi, 2013), *MC graphs* (Koster, 2002), *summary graphs* (Wermuth, 2011; Cox and Wermuth, 1996), *ancestral graphs* (Richardson and Spirtes, 2002), etc. These model classes share rich relationships, which have been recently discussed by Lauritzen and Sadeghi (2018). Ancestral graphs extend regression graphs by relaxing the cycle constraint, but they are not a maximal class; that is, if an edge is removed from the graph, we may remain on the same Markov model. Maximality is convenient because it is what allows to define pairwise Markov properties, so that each edge absent implies a conditional independence. Fortunately, for an arbitrary ancestral graph we may always find a maximal one which is Markov equivalent to it, therefore many times authors speak of the class of *maximal ancestral graphs* (MAGs). This class is closed under marginalization and conditioning, and every MAG can be obtained from an acyclic digraph after performing such operations on its nodes. Just as marginalization leads to latent confounders, conditioning is sometimes called *selection bias* in the literature on social sciences.

The class of summary graphs, although in correspondence with ancestral graphs, is not easily parametrized. One of the main drawbacks is that they allow more than one edge type between the nodes, which means that in principle more than one parameter can be associated between a pair of variables. Furthermore, they are not maximal and thus cannot have a pairwise Markov property, which implies that fewer independences can be deduced from the model. Most of the other three-edge-type models mentioned share these drawbacks for defining a parametrization (Richardson and Spirtes, 2002; Sadeghi and Marchetti, 2012).

The proliferation of higher level Markov model classes has led Lauritzen and Sadeghi (2018) to recently propose a class of mixed graphs consisting of up to four edge types, in an attempt to unify most of them under a unique Markov property (see also Sadeghi and Lauritzen, 2014; Evans, 2018). Nowadays, a great amount of research is focused on characterizing basic foundational properties for these higher level models: for example, Markov equivalence, definition and equivalence of Markov properties, factorization properties, etc.

## 11. Relaxing the Gaussian assumption

In some real problems, the Gaussian assumption is too restrictive, and thus some alternative models to overcome this have been proposed. Although these are outside the scope of this review, we will survey here the main proposals to relax the Gaussian assumption.

As we have seen, Gaussian Bayesian networks are equivalent to a set of recursive regressions where the errors are Gaussian. In Shimizu et al. (2006), an analogous model is proposed, called LiNGAM, where the errors are assumed to be

non Gaussian. The work by Loh and Bühlmann (2014) further generalizes this by not making distributional assumptions on the errors. As an alternative, Peters et al. (2014) and Bühlmann et al. (2014) maintain Gaussian errors but the additive regression is now assumed to be non linear. Other families of continuous distributions that have been used for parametrizing Markov and Bayesian networks are nonparametric Gaussian copulas (Liu et al., 2009) and elliptical distributions (Vogel and Fried, 2011), both of which generalize the Gaussian distribution. Copula graphical models are usually referred to as 'nonparanormal' models, and model selection and estimation have been researched by Harris and Drton (2013), Xue and Zou (2012) and Liu et al. (2012), including high-dimensional scenarios. Estimation results for elliptical graphical models have been obtained by Vogel and Tyler (2014).

Another approach is to extend the model and allow for both discrete and continuous variables. In such case, a challenge is posed specially on inference in Bayesian networks, where the usual operations may not allow for a direct and efficient implementation as in the pure cases. The main source of this problem is the integration that appears in marginalization of continuous variables. To overcome this, the classical approach is to use the conditional Gaussian distribution (Olkin and Tate, 1961). It is characterized by a multinomial distribution on the discrete variables and a Gaussian distribution for the continuous variables when conditioned on the discrete ones. Therefore, it contains the pure multinomial and Gaussian models as particular cases. Markov properties of this distribution with respect to an undirected graph were defined by Lauritzen and Wermuth (1989). With respect to an acyclic digraph, a further assumption is that no discrete variable may have continuous parents, which leads to conditional linear Gaussian Bayesian networks (Lauritzen, 1992). Exact inference in these networks is applicable thanks to these constraints imposed on the network topology.

In order to avoid the structural constraints of conditional linear Gaussian Bayesian networks, nonparametric density estimation techniques have been proposed. Moral et al. (2001) approximated the joint density by mixtures of truncated exponentials. In this model, discrete nodes with continuous parents are allowed, while exact inference remains possible. A similar approach is that of Shenoy and West (2011), where mixtures of polynomials are used instead for approximating the joint density. These two models have been generalized by Langseth et al. (2012) as mixtures of truncated basis functions. However, there are limited results about maximum likelihood estimation and model selection for these models (Langseth et al., 2010, 2014; Varando et al., 2015).

## 12. Main application areas

Graphical or Markov models have been widely applied since their conception and continue to be nowadays an essential tool in many fields, since they are intuitive for visualizing the associations between the components in a system. We will first outline applications of Gaussian Markov and Gaussian Bayesian networks and then illustrate other areas where graphical models have played an important role.

Markov and Bayesian networks with Gaussian parametrization have been specially useful in biomedical sciences. For example, Gaussian Bayesian networks have been used for extracting knowledge from fMRI studies (Mumford and Ramsey, 2014; Zhou et al., 2016), where nodes are identified with brain regions, and arrows are interpreted as direct influences between the respective regions. Another example where both models have been applied is the modelling of gene regulatory networks, which are high-dimensional and complex by nature. In fact, the challenge posed by this problem has served as an impulse for methodological developments in both models. A vast amount of literature can be found regarding the main computational aspects involved on this subject, as well as interpretability issues, see Lauritzen and Sheehan (2003), Friedman (2004), Markowetz and Spang (2007) and Ness et al. (2016) for reviews.

In social sciences, Bayesian networks have been used since their conception, in fact, we could say that research in this application area helped to settle the foundations of graphical models (Kiiveri and Speed, 1982). In terms of interpretability, the directed arcs in Bayesian networks are usually given a causal interpretation (Pearl, 2000; Cox and Wermuth, 1996), since ultimately the main goal of social studies is to identify the causes of a resulting event of interest. In recent years, graphical models have found a natural area of application which is social network analysis (Farasat et al., 2015), which include problems such as influence analysis, privacy protection, web browsing, etc.

Some other traditional models from bioinformatics can also be seen as graphical models. These include phylogenetic trees, which model evolutionary relationships between different species or organisms, and pedigrees, which are diagrams showing the occurrence and variants of a gene from one generation of organisms to the next (Jordan, 2004). Apart from fMRI studies, Bayesian networks have also been applied in different subareas of neuroscience (Bielza and Larrañaga, 2014), such as morphological and electrophysiological studies. Finally, other, more technical application areas include information retrieval (de Campos et al., 2004), where relevant documents about some matter are collected from an available set of sources, and linguistics, with subfields such as speech recognition (Deng and Li, 2013), and natural language processing (Cambria and White, 2014).

## Acknowledgments

# References

Altomare, D., Consonni, G., La Rocca, L., 2013. Objective Bayesian search of Gaussian directed acyclic graphical models for ordered variables with non-local priors. Biometrics 69 (2), 478–487.

Anderson, T.W., 1973. Asymptotically efficient estimation of covariance matrices with linear structure. Ann. Statist. 1 (1), 135–141.

Anderson, T.W., 2003. An Introduction to Multivariate Statistical Analysis, third ed. John Wiley & Sons.

Andersson, S.A., Madigan, D., Perlman, M.D., 2001. Alternative Markov properties for chain graphs. Scand. J. Stat. 28 (1), 33–85.

Andersson, S.A., Perlman, M.D., 1998. Normal linear regression models with recursive graphical Markov structure. J. Multivariate Anal. 66 (2), 133–187.

Aragam, B., Amini, A.A., Zhou, Q., 2017. Learning directed acyclic graphs with penalized neighbourhood regression. arXiv:1511.08963.

Aragam, B., Zhou, Q., 2015. Concave penalized estimation of sparse Gaussian Bayesian networks. J. Mach. Learn. Res. 16, 2273–2328.

Atay-Kayis, A., Massam, H., 2005. A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. Biometrika 92 (2), 317–335.

Banerjee, O., El Ghaoui, L., d'Aspremont, A., 2008. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. J. Mach. Learn. Res. 9, 485–516.

Barndorff-Nielsen, O., 1978. Information and Exponential Families in Statistical Theory. John Wiley & Sons.

Ben-David, E., Li, T., Massam, H., Rajaratnam, B., 2016. High dimensional Bayesian inference for Gaussian directed acyclic graph models. arXiv: 1109.4371.

Ben-David, E., Rajaratnam, B., 2012. Positive definite completion problems for Bayesian networks. SIAM J. Matrix Anal. Appl. 33 (2), 617–638.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. R. Stat. Soc. Ser. B Stat. Methodol. 36 (2), 192–236.

Bielza, C., Larrañaga, P., 2014. Bayesian networks in neuroscience: A survey. Front. Comput. Neurosci. 8, 131.

Bühlmann, P., van de Geer, S., 2011. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer.

Bühlmann, P., Peters, J., Ernest, J., 2014. CAM: Causal additive models, high-dimensional order search and penalized regression. Ann. Statist. 42 (6), 2526–2556.

Cambria, E., White, B., 2014. Jumping NLP curves: A review of natural language processing research. IEEE Comput. Intell. Mag. 9 (2), 48–57.

de Campos, L.M., M.F.ernández-Luna, J., F.H.uete, J., 2004. Bayesian networks and information retrieval: an introduction to the special issue. Inf. Process. Manag. 40 (5), 727–733.

Cao, X., Khare, K., Ghosh, M., 2019. Posterior graph selection and estimation consistency for high-dimensional Bayesian DAG models. Ann. Statist. 47 (1), 319–348.

Carvalho, C.M., Massam, H., West, M., 2007. Simulation of hyper-inverse Wishart distributions in graphical models. Biometrika 94 (3), 647–659.

Carvalho, C.M., Scott, J.G., 2009. Objective Bayesian model selection in Gaussian graphical models. Biometrika 96 (3), 497–512.

Castelo, R., Roverato, A., 2006. A robust procedure for Gaussian graphical model search from microarray data with $p$ larger than $n$. J. Mach. Learn. Res. 6, 2621–2650.

Colombo, D., Maathuis, M.H., 2014. Order-independent constraint-based causal structure learning. J. Mach. Learn. Res. 15, 3921–3962.

Consonni, G., Rocca, L., 2012. Objective Bayes factors for Gaussian directed acyclic graphical models. Scand. J. Stat. 39 (4), 743–756.

Cox, D.R., Wermuth, N., 1993. Linear dependencies represented by chain graphs. Stat. Sci. 8 (3), 204–218.

Cox, D.R., Wermuth, N., 1996. Multivariate Dependencies: Models, Analysis and Interpretation. Chapman & Hall.

Daly, R., Shen, Q., Aitken, S., 2011. Learning Bayesian networks: Approaches and issues. Knowl. Eng. Rev. 26, 99–157.

Darroch, J.N., Lauritzen, S.L., Speed, T.P., 1980. Markov fields and log-linear interaction models for contingency tables. Ann. Statist. 8 (3), 522–539.

Dawid, A.P., 1979. Conditional independence in statistical theory. J. R. Stat. Soc. Ser. B Stat. Methodol. 41 (1), 1–31.

Dawid, A.P., 1980. Conditional independence for statistical operations. Ann. Statist. 8 (3), 598–617.

Dawid, A.P., 2001. Separoids: A mathematical framework for conditional independence and irrelevance. Ann. Math. Artif. Intell. 32 (1), 335–372.

Dawid, A.P., Lauritzen, S.L., 1993. Hyper Markov laws in the statistical analysis of decomposable graphical models. Ann. Statist. 21 (3), 1272–1317.

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

Deng, L., Li, X., 2013. Machine learning paradigms for speech recognition: An overview. IEEE Trans. Audio, Speech, Lang. Process. 21 (5), 1060–1089.

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

Drton, M., Fox, C., Käufl, A., 2012. Comments on: Sequences of regressions and their independencies. TEST 21 (2), 255–261.

Drton, M., Perlman, M.D., 2004. Model selection for Gaussian concentration graphs. Biometrika 91 (3), 591–602.

Drton, M., Perlman, M.D., 2007. Multiple testing and error control in Gaussian graphical model selection. Stat. Sci. 22 (3), 430–449.

Drton, M., Perlman, M.D., 2008. A SINful approach to Gaussian graphical model selection. J. Statist. Plann. Inference 138 (4), 1179–1200.

Eriksen, P.S., 1996. Tests in covariance selection models. Scand. J. Stat. 23 (3), 275–284.

Evans, R., 2018. Markov properties for mixed graphical models. In: Handbook of Graphical Models. CRC Press, pp. 57–78.

Farasat, A., Nikolaev, A., Srihari, S.N., Blair, R.H., 2015. Probabilistic graphical models in modern social network analysis. Soc. Netw. Anal. Min. 5 (1), 62.

Fox, C.J., Käufl, A., Drton, M., 2015. On the causal interpretation of acyclic mixed graphs under multivariate normality. Linear Algebra Appl. 473 (Suppl. C), 93–113.

Friedman, N., 2004. Inferring cellular networks using probabilistic graphical models. Science 303 (5659), 799–805.

Friedman, J., Hastie, T., Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9 (3), 432–441.

Frydenberg, M., 1990. The chain graph Markov property. Scand. J. Stat. 17 (4), 333–353.

Frydenberg, M., Lauritzen, S., 1989. Decomposition of maximum likelihood in mixed graphical interaction models. Biometrika 76 (3), 539–555.

de la Fuente, A., Bing, N., Hoeschele, I., Mendes, P., 2004. Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics 20 (18), 3565–3574.

Geiger, D., Heckerman, D., 1994. Learning Gaussian networks. In: Proc. of the Tenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco, pp. 235–243.

Geiger, D., Heckerman, D., 2002. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. Ann. Statist. 30 (5), 1412–1440.

Geiger, D., Pearl, J., 1990. On the logic of causal models. In: Proc. of the Fourth Annual Conference on Uncertainty in Artificial Intelligence. AUAI Press, Corvallis, pp. 3–14.

Geiger, D., Pearl, J., 1993. Logical and algorithmic properties of conditional independence and graphical models. Ann. Statist. 21 (4), 2001–2021.

Gillispie, S.B., Perlman, M.D., 2002. The size distribution for Markov equivalence classes of acyclic digraph models. Artificial Intelligence 141 (1–2), 137–155.

Giudici, P., 1996. Learning in graphical Gaussian models. Bayesian Stat. 5, 621–628.

Giudici, P., Green, P.J., 1999. Decomposable graphical Gaussian model determination. Biometrika 86 (4), 785–801.

Grädel, E., Väänänen, J., 2013. Dependence and independence. Stud. Log. 101 (2), 399–410.

Grimmett, G.R., 1973. A theorem about random fields. Bull. Lond. Math. Soc. 5 (1), 81–84.

Grone, R., Johnson, C.R., Sá, E.M., Wolkowicz, H., 1984. Positive definite completions of partial Hermitian matrices. Linear Algebra Appl. 58, 109–124.

Hammersley, J.M., Clifford, P., 1971. Markov fields on finite graphs and lattices. Unpublished manuscript.

Harris, N., Drton, M., 2013. PC algorithm for nonparanormal graphical models. J. Mach. Learn. Res. 14, 3365–3383.

He, Y., Jia, J., Yu, B., 2015. Counting and exploring sizes of Markov equivalence classes of directed acyclic graphs. J. Mach. Learn. Res. 16, 2589–2609.

Horn, R.A., Johnson, C.R., 2012. Matrix Analysis, second ed. Cambridge University Press.

Howard, R.A., Matheson, J.E., 2005. Influence diagrams. Decis. Anal. 2 (3), 127–143.

Ibáñez, A., Armañanzas, R., Bielza, C., Larrañaga, P., 2016. Genetic algorithms and Gaussian Bayesian networks to uncover the predictive core set of bibliometric indices. J. Assoc. Inf. Sci. Technol. 67 (7), 1703–1721.

Isham, V., 1981. An introduction to spatial point processes and Markov random fields. Int. Stat. Rev. 49 (1), 21–43.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., West, M., 2005. Experiments in stochastic computation for high-dimensional graphical models. Stat. Sci. 20 (4), 388–400.

Jordan, M.I., 2004. Graphical models. Stat. Sci. 19 (1), 140–155.

Kalisch, M., Bühlmann, P., 2007. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. J. Mach. Learn. Res. 8, 613–636.

Kiiveri, H., Speed, T., 1982. Structural analysis of multivariate data: A review. Sociol. Methodol. 13, 209–289.

Kindermann, R., Snell, J.L., 1980. Markov Random Fields and their Applications. American Mathematical Society.

Koster, J.T.A., 1999. On the validity of the Markov interpretation of path diagrams of Gaussian structural equations systems with correlated errors. Scand. J. Stat. 26 (3), 413–431.

Koster, J.T., 2002. Marginalizing and conditioning in graphical models. Bernoulli 8 (6), 817–840.

Lam, C., Fan, J., 2009. Sparsistency and rates of convergence in large covariance matrix estimators. Ann. Statist. 37 (6B), 4254–4278.

Langseth, H., Nielsen, T., Pérez-Bernabé, I., Salmerón, A., 2014. Learning mixtures of truncated basis functions from data. Internat. J. Approx. Reason. 55, 940–956.

Langseth, H., Nielsen, T., Rumí, R., Salmerón, A., 2010. Parameter estimation and model selection for mixtures of truncated exponentials. Internat. J. Approx. Reason. 51, 485–498.

Langseth, H., Nielsen, T., Rumí, R., Salmerón, A., 2012. Mixtures of truncated basis functions. Internat. J. Approx. Reason. 53, 212–227.

Lauritzen, S.L., 1992. Propagation of probabilities, means, and variances in mixed graphical association models. J. Amer. Stat. Assoc. 87 (420), 1098–1108.

Lauritzen, S.L., 1996. Graphical Models. Oxford University Press.

Lauritzen, S.L., Dawid, A.P., Larsen, B.N., Leimer, H.-G., 1990. Independence properties of directed Markov fields. Networks 20 (5), 491–505.

Lauritzen, S., Sadeghi, K., 2018. Unifying Markov properties for graphical models. Ann. Statist. 46 (5), 2251–2278.

Lauritzen, S.L., Sheehan, N.A., 2003. Graphical models for genetic analyses. Stat. Sci. 18 (4), 489–514.

Lauritzen, S.L., Wermuth, N., 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. Ann. Statist. 17 (1), 31–57.

Letac, G., Massam, H., 2007. Wishart distributions for decomposable graphs. Ann. Statist. 35 (3), 1278–1323.

Lin, S., Uhler, C., Sturmfels, B., Bühlmann, P., 2014. Hypersurfaces and their singularities in partial correlation testing. Found. Comput. Math. 14 (5), 1079–1116.

Liu, W., 2013. Gaussian graphical model estimation with false discovery rate control. Ann. Statist. 41 (6), 2948–2978.

Liu, H., Han, F., Yuan, M., Lafferty, J., Wasserman, L., 2012. High-dimensional semiparametric Gaussian copula graphical models. Ann. Statist. 40 (4), 2293–2326.

Liu, H., Lafferty, J., Wasserman, L., 2009. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. J. Mach. Learn. Res. 10, 2295–2328.

Loh, P.-L., Bühlmann, P., 2014. High-dimensional learning of linear causal networks via inverse covariance estimation. J. Mach. Learn. Res. 15, 3065–3105.

Magwene, P.M., Kim, J., 2004. Estimating genomic coexpression networks using first-order conditional independence. Genome Biol. 5 (12), R100.

Markowetz, F., Spang, R., 2007. Inferring cellular networks – a review. BMC Bioinform. 8 (6), S5.

Massam, H., 2018. Bayesian inference in graphical Gaussian models. In: Handbook of Graphical Models. CRC Press, pp. 257–282.

Meek, C., 1995. Strong completeness and faithfulness in Bayesian networks. In: Proc. of the Eleventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco, pp. 411–418.

Meinshausen, N., 2008. A note on the lasso for Gaussian graphical model selection. Statist. Probab. Lett. 78 (7), 880–884.

Meinshausen, N., Bühlmann, P., 2006. High-dimensional graphs and variable selection with the lasso. Ann. Statist. 34 (3), 1436–1462.

Meinshausen, N., Yu, B., 2009. Lasso-type recovery of sparse representations for high-dimensional data. Ann. Statist. 37 (1), 246–270.

Moral, S., Rumí, R., Salmerón, A., 2001. Mixtures of truncated exponentials in hybrid Bayesian networks. In: Symbolic and Quantitative Approaches To Reasoning with Uncertainty. In: Lecture Notes in Artificial Intelligence, vol. 2143, Springer, pp. 156–167.

Mumford, J.A., Ramsey, J.D., 2014. Bayesian networks for fMRI: A primer. NeuroImage 86, 573–582.

Ness, R.O., Sachs, K., Vitek, O., 2016. From correlation to causality: Statistical approaches to learning regulatory relationships in large-scale biomolecular investigations. J. Proteom. Res. 15 (3), 683–690.

O'Hagan, A., 1995. Fractional Bayes factors for model comparison. J. R. Stat. Soc. Ser. B Stat. Methodol. 57 (1), 99–138.

Olkin, I., Tate, R., 1961. Multivariate correlation models with mixed discrete and continuous variables. Ann. Math. Stat. 32 (2), 448–465.

Peña, J.M., 2016a. Alternative Markov and causal properties for acyclic directed mixed graphs. In: Proc. of the Thirty-Second Conference on Uncertainty in Artificial Intelligence. AUAI Press, Arlington, pp. 577–586.

Peña, J.M., 2016b. Learning acyclic directed mixed graphs from observations and interventions. In: Proc. of the Eighth International Conference on Probabilistic Graphical Models. In: Proceedings of Machine Learning Research, vol. 52, PMLR, Lugano, pp. 392–402.

Peña, J.M., 2018. Unifying DAGs and UGs. In: Proc. of the Ninth International Conference on Probabilistic Graphical Models. In: Proceedings of Machine Learning Research, vol. 72, PMLR, Prague, pp. 308–319.

Pearl, J., 1985. Bayesian networks: A model of self-activated memory for evidential reasoning. Tech. Rep. R-43, University of California, Los Angeles.

Pearl, J., 1986. Fusion, propagation, and structuring in belief networks. Artificial Intelligence 29 (3), 241–288.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann.

Pearl, J., 2000. Causaliy: Models, Reasoning and Inference. Cambridge University Press.

Pearl, J., Paz, A., 1987. Graphoids: A graph-based logic for reasoning about relevance relations. In: Advances in Artificial Intelligence. Vol. 2. Elsevier, pp. 357–363.

Peters, J., 2014. On the intersection property of conditional independence and its application to causal discovery. J. Causal Inference 3 (1), 97–108.

Peters, J., Mooij, J.M., Janzing, D., Schölkopf, B., 2014. Causal discovery with continuous additive noise models. J. Mach. Learn. Res. 15, 2009–2053.

Porteous, B.T., 1989. Stochastic inequalities relating a class of log-likelihood ratio statistics to their asymptotic $\chi^2$ distribution. Ann. Statist. 17 (4), 1723–1734.

Radhakrishnan, A., Solus, L., Uhler, C., 2018. Counting Markov equivalence classes for DAG models on trees. Discrete Appl. Math. 244, 170–185.

Rajaratnam, B., 2012. Comment on: Sequences of regressions and their independences. TEST 21 (2), 268–273.

Rajaratnam, B., Massam, H., Carvalho, C.M., 2008. Flexible covariance estimation in graphical Gaussian models. Ann. Statist. 36 (6), 2818–2849.

Ravikumar, P., Wainwright, M.J., Raskutti, G., Yu, B., 2011. High-dimensional covariance estimation by minimizing $l_1$-penalized log-determinant divergence. Electron. J. Stat. 5, 935–980.

Richardson, T., 2003. Markov properties for acyclic directed mixed graphs. Scand. J. Stat. 30 (1), 145–157.

Richardson, T.S., Robins, J.M., Shpitser, I., 2012. Nested Markov properties for acyclic directed mixed graphs. In: Proc. of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence. AUAI Press, Arlington, p. 13.

Richardson, T., Spirtes, P., 2002. Ancestral graph Markov models. Ann. Statist. 30 (4), 962–1030.

Robins, J.M., Scheines, R., Spirtes, P., Wasserman, L., 2003. Uniform consistency in causal inference. Biometrika 90 (3), 491–515.

Rothman, A.J., Bickel, P.J., Levina, E., Zhu, J., 2008. Sparse permutation invariant covariance estimation. Electron. J. Stat. 2, 494–515.

Roverato, A., 2000. Cholesky decomposition of a hyper inverse Wishart matrix. Biometrika 87 (1), 99–112.

Roverato, A., 2002. Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. Scand. J. Stat. 29 (3), 391–411.

Roverato, A., Whittaker, J., 1998. The Isserlis matrix and its application to non-decomposable graphical Gaussian models. Biometrika 85 (3), 711–725.

Sadeghi, K., 2013. Stable mixed graphs. Bernoulli 19 (5B), 2330–2358.

Sadeghi, K., 2016. Marginalization and conditioning for LWF chain graphs. Ann. Statist. 44 (4), 1792–1816.

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

Sadeghi, K., Marchetti, G.M., 2012. Graphical Markov models with mixed graphs in R. The R J. 4 (2), 65–73.

Scutari, M., 2013. On the prior and posterior distributions used in graphical modelling. Bayesian Anal. 8 (3), 505–532.

Shenoy, P., West, J.C., 2011. Inference in hybrid Bayesian networks using mixtures of polynomials. Internat. J. Approx. Reason. 52 (5), 641–657.

Shimizu, S., Hoyer, P.O., Hyvarinen, A., Kerminen, A., 2006. A linear non-Gaussian acyclic model for causal discovery. J. Mach. Learn. Res. 7, 2003–2030.

Shojaie, A., Michailidis, G., 2010. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. Biometrika 97 (3), 519–538.

Sidak, Z., 1967. Rectangular confidence regions for the means of multivariate normal distributions. J. Amer. Stat. Assoc. 62 (318), 626–633.

Sonntag, D., Peña, J.M., Gómez-Olmedo, M., 2015. Approximate counting of graphical models via MCMC revisited. Intern. J. Intell. Sys. 30 (3), 384–420.

Speed, T.P., 1979. A note on nearest-neighbour Gibbs and Markov probabilities. Sankhya mphA 41 (3/4), 184–197.

Spirtes, P., 1995. Directed cyclic graphical representations of feedback models. In: Proc. of the Eleventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufman, San Francisco, pp. 491–498.

Spirtes, P., Glymour, C., Scheines, R., 2000. Causation, Prediction, and Search. MIT Press.

Spirtes, P., Richardson, T., Meek, C., 1997. The dimensionality of mixed ancestral graphs. Tech. Rep. CMU-PHIL-83, Carnegie Mellon University.

Steinsky, B., 2004. Asymptotic behaviour of the number of labelled essential acyclic digraphs and labelled chain graphs. Graphs Combin. 20 (3), 399–411.

Studenỳ, M., 2005. On Probabilistic Conditional Independence Structures. Springer London.

Studenỳ, M., 2018. Conditional independence and basic Markov properties. In: Handbook of Graphical Models. CRC Press, pp. 21–56.

Sturmfels, B., Uhler, C., 2010. Multivariate Gaussians, semidefinite matrix completion, and convex algebraic geometry. Ann. Inst. Stat. Math. 62 (4), 603–638.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.

Uhler, C., 2012. Geometry of maximum likelihood estimation in Gaussian graphical models. Ann. Statist. 40 (1), 238–261.

Uhler, C., 2018. Gaussian graphical models. In: Handbook of Graphical Models. CRC Press, pp. 235–256.

Uhler, C., Lenkoski, A., Richards, D., 2018. Exact formulas for the normalizing constants of Wishart distributions for graphical models. Ann. Statist. 46 (1), 90–118.

Uhler, C., Raskutti, G., Bühlmann, P., Yu, B., 2013. Geometry of the faithfulness assumption in causal inference. Ann. Statist. 41 (2), 436–463.

van de Geer, S.A., Bühlmann, P., 2009. On the conditions used to prove oracle results for the lasso. Electron. J. Stat. 3, 1360–1392.

van de Geer, S., Bühlmann, P., 2013. $\ell_0$-penalized maximum likelihood for sparse directed acyclic graphs. Ann. Statist. 41 (2), 536–567.

Varando, G., López-Cruz, P., Nielsen, T., Larrañaga, P., Bielza, C., 2015. Conditional density approximations with mixtures of polynomials. Intern. J. Intell. Sys. 30, 236–264.

Verma, T., Pearl, J., 1991. Equivalence and synthesis of causal models. In: Proc. of the Sixth Annual Conference on Uncertainty in Artificial Intelligence. AUAI Press, Corvallis, pp. 255–270.

Vogel, D., Fried, R., 2011. Elliptical graphical modelling. Biometrika 98 (4), 935–951.

Vogel, D., Tyler, D.E., 2014. Robust estimators for nondecomposable elliptical graphical models. Biometrika 101 (4), 865.

Wang, H., Carvalho, C.M., 2010. Simulation of hyper-inverse Wishart distributions for non-decomposable graphs. Electron. J. Stat. 4, 1470–1475.

Werhli, A.V., Grzegorczyk, M., Husmeier, D., 2006. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. Bioinformatics 22 (20), 2523–2531.

Wermuth, N., 1976a. Analogies between multiplicative models in contingency tables and covariance selection. Biometrics 32 (1), 95–108.

Wermuth, N., 1976b. Model search among multiplicative models. Biometrics 32 (2), 253–263.

Wermuth, N., 1980. Linear recursive equations, covariance selection, and path analysis. J. Amer. Statist. Assoc. 75 (372), 963–972.

Wermuth, N., 2011. Probability distributions with summary graph structure. Bernoulli 17 (3), 845–879.

Wermuth, N., 2015. Graphical Markov models, unifying results and their interpretation. Wiley StatsRef: Stat. Ref. Online 1–29.

Wermuth, N., Lauritzen, S.L., 1983. Graphical and recursive models for contingency tables. Biometrika 70 (3), 537–552.

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

Wille, A., Bühlmann, P., 2006. Low-order conditional independence graphs for inferring genetic networks. Stat. Appl. Genet. Mol. Biol. 5 (1), Article 1.

Wright, S., 1934. The method of path coefficients. Ann. Math. Stat. 5 (3), 161–215.

Xue, L., Zou, H., 2012. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. Ann. Statist. 40 (5), 2541–2571.

Yu, G., Bien, J., 2017. Learning local dependence in ordered data. J. Mach. Learn. Res. 18, 1–60.

Yuan, M., Lin, Y., 2007a. Model selection and estimation in the Gaussian graphical model. Biometrika 94 (1), 19–35.

Yuan, M., Lin, Y., 2007b. On the non-negative garrotte estimator. J. R. Stat. Soc. Ser. B Stat. Methodol. 69 (2), 143–161.

Yule, G.U., 1907. On the theory of correlation for any number of variables, treated by a new system of notation. Proc. R. Soc. A 79 (529), 182–193.

Zhang, J., Spirtes, P., 2003. Strong faithfulness and uniform consistency in causal inference. In: Proc. of the Nineteenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Francisco, pp. 632–639.

Zhao, P., Yu, B., 2006. On model selection consistency of lasso. J. Mach. Learn. Res. 7, 2541–2563.

Zhou, L., Wang, L., Liu, L., Ogunbona, P., Dinggang, S., 2016. Learning discriminative Bayesian networks from high-dimensional continuous neuroimaging data. IEEE Trans. Pattern Anal. Mach. Intell. 38 (11), 2269–2283.

Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Stat. Assoc. 101 (476), 1418–1429.