Structure Learning in Graphical Modeling

Mathias Drton[1] and Marloes H. Maathuis[2]

[1]Department of Statistics, University of Washington, Seattle 98195; email: md5@uw.edu

[2]Seminar für Statistik, ETH Zürich, 8092 Zürich, Switzerland;
email: maathuis@stat.math.ethz.ch

Annu. Rev. Stat. Appl. 2017. 4:365–93

First published online as a Review in Advance on October 28, 2016

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi:
10.1146/annurev-statistics-060116-053803

Keywords

Bayesian network, graphical model, Markov random field, model selection, multivariate statistics, network reconstruction

Abstract

A graphical model is a statistical model that is associated with a graph whose nodes correspond to variables of interest. The edges of the graph reflect allowed conditional dependencies among the variables. Graphical models have computationally convenient factorization properties and have long been a valuable tool for tractable modeling of multivariate distributions. More recently, applications such as reconstructing gene regulatory networks from gene expression data have driven major advances in structure learning, that is, estimating the graph underlying a model. We review some of these advances and discuss methods such as the graphical lasso and neighborhood selection for undirected graphical models (or Markov random fields) and the PC algorithm and score-based search methods for directed graphical models (or Bayesian networks). We further review extensions that account for effects of latent variables and heterogeneous data sources.

1. INTRODUCTION

This article gives an overview of commonly used techniques for structure learning in graphical modeling. Structure learning is a model selection problem in which one estimates a graph that summarizes the dependence structure in a given data set.

Graphical models: conditional independence models for random vectors

Undirected graphical model: a multivariate statistical model in which each variable may be optimally predicted from neighboring variables

Directed graphical model: a multivariate statistical model that can capture dependencies induced by cause-effect relationships

1.1. What Is a Graphical Model?

A graphical model captures stochastic dependencies among a collection of random variables X_v, $v \in V$ (Lauritzen 1996). More precisely, a graphical model is a set of multivariate joint distributions that exhibit certain conditional independencies. Each model is associated with a graph $G = (V, E)$, where the vertex set V indexes the variables and the edge set E encodes the conditional independence constraints. These constraints require variables X_v and X_w to be conditionally independent given $X_C := (X_u : u \in C)$, denoted by $X_v \perp\!\!\!\perp X_w \mid X_C$, if every path between nodes v and w in G is suitably blocked by the nodes in C.

Markov chains constitute a familiar example of graphical models. We first demonstrate this in the context of undirected graphs, for which the edges are unordered pairs $\{v, w\}$ for distinct $v, w \in V$. We also write $v - w \in E$ to indicate that E contains the edge $\{v, w\}$. In an undirected graph, a path is blocked by C if it contains a node in C.

Example 1. Suppose that $(X_1, \ldots, X_5)$ belongs to the graphical model for the undirected graph in **Figure 1*a***. Then $X_v \perp\!\!\!\perp X_w \mid X_C$ whenever C contains a node c on the unique path between v and w, so $v < c < w$ or $w < c < v$. For example, $X_2 \perp\!\!\!\perp X_4 \mid (X_1, X_3)$. If we think of the indices as time, the past and the future are conditionally independent given the present. We recognize that $X_1, \ldots, X_5$ form a Markov chain. The generalization to a Markov chain of arbitrary length is obvious.

Let $\mathrm{nb}_G(v) = \{w \in V : \{w, v\} \in E\}$ be the neighbors of node v in an undirected graph $G = (V, E)$. As detailed in Section 2, a typically equivalent interpretation of the graph is that each X_v is conditionally independent of its nonneighbors $X_{V \setminus (\mathrm{nb}_G(v) \cup \{v\})}$ given its neighbors $X_{\mathrm{nb}_G(v)}$. A mean squared error optimal prediction of X_v can thus be made from its neighbors $X_{\mathrm{nb}_G(v)}$ alone.

Graphical models can also be built from directed graphs $G = (V, E)$. The edge set E then comprises ordered pairs (v, w) that represent an edge pointing from v to w. To indicate the presence of an edge from v to w, we also write $v \to w \in E$. Adopting the language of family trees, let $\mathrm{pa}_G(v) = \{w \in V : (w, v) \in E\}$ be the parents of node v in G, let $\mathrm{de}_G(v) = \{w \in V : w = v \text{ or } v \to \cdots \to w \text{ in } G\}$ be the descendants of v in G, and let $\mathrm{nd}_G(v) = V \setminus \mathrm{de}_G(v)$ be the nondescendants of v in G. In a directed acyclic graph, we then require that each variable X_v is conditionally independent of its nondescendants $X_{\mathrm{nd}_G(v) \setminus \mathrm{pa}_G(v)}$ given its parents $X_{\mathrm{pa}_G(v)}$. Such independencies arise when each variable X_v is a stochastic function of $X_{\mathrm{pa}_G(v)}$. This is the starting point for a connection to causal modeling (Spirtes et al. 2000, Pearl 2009). The notion of blocking a path in a directed graph is different and more subtle than blocking in an undirected graph. We give the details in Section 2.

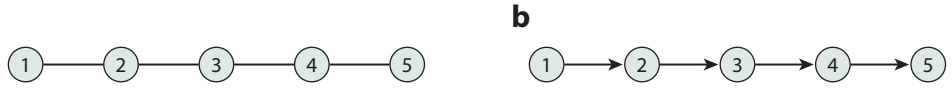

Figure 1

Two graphs that both induce a Markov chain, (*a*) an undirected graph and (*b*) a directed acyclic graph.

Example 2. Suppose that $(X_1, \ldots, X_5)$ belongs to the graphical model for the directed acyclic graph in **Figure 1*b***. Denoting this graph by G, we have $\mathrm{pa}_G(v) = \{v-1\}$ for $v = 2, \ldots, 5$. Therefore, $X_v \perp\!\!\!\perp (X_1, \ldots, X_{v-2}) \,|\, X_{v-1}$ for $v = 3, 4, 5$. This is precisely the standard Markov property of a Markov chain. Again, this example is easily generalized to a chain of arbitrary length.

Density factorization: the joint density is a product of terms that can be efficiently stored and computed with

Factorization: for undirected graphs, the joint density factors into a product of terms that are associated with the cliques of the graph

Graphical models can also be defined in terms of density factorizations. Indeed, much of the popularity of graphical models is due to the fact that factorizations allow efficient storage and computation with high-dimensional joint distributions (Wainwright & Jordan 2008). To explain, suppose for simplicity that $X_1, X_2, \ldots, X_m$ are binary and form a Markov chain, in which case the joint probability factorizes as

$$\Pr(X_1 = x_1, \ldots, X_m = x_m) \;=\; \Pr(X_1 = x_1) \prod_{v=2}^{m} \Pr(X_v = x_v \,|\, X_{v-1} = x_{v-1}). \qquad (1)$$

The $2^m - 1$ dimensional joint distribution is thus determined by merely $2m - 1$ parameters. In reference to the directed graph G in **Figure 1*b***, the product in Equation 1 is over the conditional probabilities $\Pr(X_v = x_v \,|\, X_{\mathrm{pa}_G(v)} = x_{\mathrm{pa}_G(v)})$, or simply $\Pr(X_v = x_v)$ when $\mathrm{pa}_G(v) = \emptyset$. For the undirected graph in **Figure 1*a***, the right-hand side of Equation 1 may be viewed as a product of $m - 1$ functions whose arguments are the pairs (x_v, x_{v+1}) that correspond to the edges of the graph. Section 2 gives generalizations of these observations.

1.2. Application: Reconstruction of Gene Regulatory Networks

The past decade has seen great advances in structure learning, with new methods being developed and older methods being viewed in a new light. These developments have largely been driven by problems in biology, such as inferring a network of regulatory relationships among genes from data on their expression levels (Friedman 2004).

Example 3. Reporting on a prediction challenge, Marbach et al. (2012) provide data on gene expression in *Escherichia coli*. We restrict attention to the $|V| = 87$ genes that form the only large connected component in the network of known interactions among transcription factors (known prior to the challenge). Ignoring heterogeneity across the $n = 804$ samples, we apply neighborhood selection in a Gaussian copula model (see Sections 3.4 and 3.6), under the defaults of the software of Zhao et al. (2012). The estimated undirected conditional independence graph (see Section 2.1) has 352 edges. It is plotted in **Figure 2**, where each one of the 24 red edges corresponds to one of the 124 pairs of transcription factors that are known to interact. Although the majority of known interactions fail to be part of the estimate, some signal is being detected. The probability that a random selection of 352 edges would comprise at least 24 of the 124 known interactions is approximately 1.5×10^{-4}.

Example 3 involves a modest number of preselected genes and is thus of moderate dimensionality. Modern experiments often yield data of far higher dimensionality, presenting a statistical challenge that is behind much of the recent fascination with graphical models.

1.3. Outline of the Review

This review begins with more background on graphical models in Section 2. Sections 3 and 4 discuss structure learning for undirected and directed graphical models. A brief treatment of issues arising from latent variables and heterogeneous data sources is given in Sections 5 and 6. We end with a discussion in Section 7. Although our review conveys some of the main ideas in

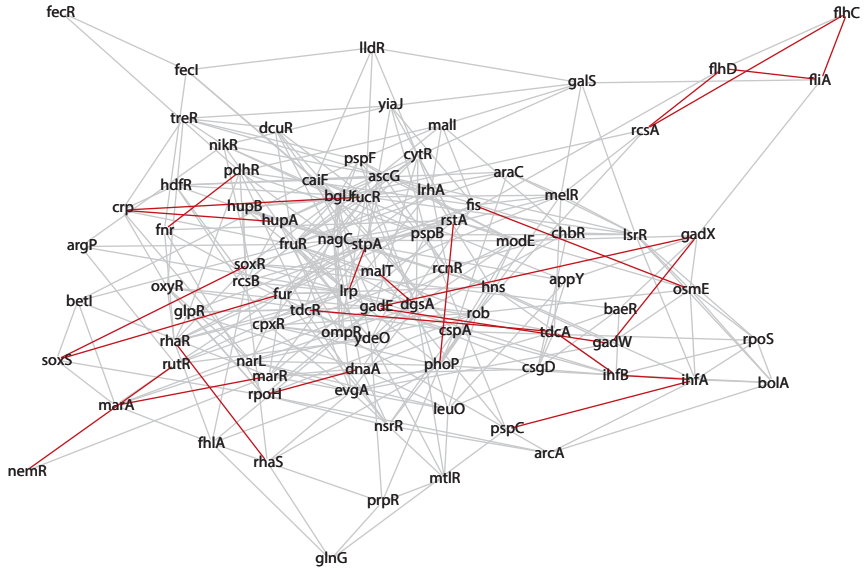

Figure 2

Estimated conditional independence graph in a Gaussian copula model for data on the expression of 87 transcription factors in *Escherichia coli*. The 24 red edges correspond to pairs of transcription factors that are known to interact.

structure learning, several interesting topics are beyond the scope of our review. For instance, we do not cover Bayesian inference, even though there is an active Bayesian community whose recent work tackles problems such as heterogeneous data (Peterson et al. 2015, Mitra et al. 2016), latent variables (Silva & Ghahramani 2009), Markov chain Monte Carlo for sampling graphs (He et al. 2013, Goudie & Mukherjee 2016, Kuipers & Moffa 2016), posterior convergence rates (Banerjee & Ghosal 2015), robustness (Finegold & Drton 2014), and context-specific independence (Nyman et al. 2014). We also do not treat dynamic graphical models for multivariate time series, but an example of work discussing such models is Eichler (2012). Another area omitted in this review is active learning; the reader is directed to, for example, Vats et al. (2014), Statnikov et al. (2015), and Dasarathy et al. (2016).

Finally, the literature on structure learning is vast and our references are necessarily selective. To limit the number of citations, we sometimes only cite a recent article, trusting that readers will follow the trail of literature to identify earlier work on the topic.

2. BASIC CONCEPTS IN GRAPHICAL MODELING

In this section, we review some essential concepts for undirected and directed graphical models (e.g., Lauritzen 1996, Studený 2005, Pearl 2009).

2.1. Undirected Graphical Models

Let $X = (X_v : v \in V)$ be a random vector indexed by the vertices of an undirected graph $G = (V, E)$. Then X satisfies the pairwise Markov property with respect to G if

$$X_v \perp\!\!\!\perp X_w \mid X_{V \setminus \{v,w\}} \tag{2}$$

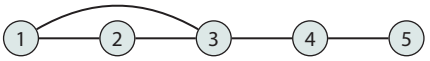

Figure 3

An undirected graph with five nodes.

whenever $\{v, w\} \notin E$. Moreover, X satisfies the local Markov property with respect to G if

$$X_v \perp\!\!\!\perp X_{V \setminus (\mathrm{nb}_G(v) \cup \{v\})} \mid X_{\mathrm{nb}_G(v)} \tag{3}$$

for every $v \in V$, where we recall that $\mathrm{nb}_G(v) = \{w \in V : \{w, v\} \in E\}$ is the set of neighbors of v. Finally, X satisfies the global Markov property with respect to G if $X_A \perp\!\!\!\perp X_B \mid X_C$ for all triples of pairwise disjoint subsets $A, B, C \subset V$ such that C separates A and B in G, i.e., such that every path between a node in A and a node in B contains a node in C.

It is easy to see that the global Markov property implies the pairwise and local properties. It can also be shown that the local Markov property implies the pairwise Markov property. Although not true in general, the three Markov properties are equivalent when X satisfies the intersection axiom for conditional independence (Lauritzen 1996). This equivalence is true in particular when the X_v are discrete with positive joint probabilities, or when X has a positive and continuous density with respect to Lebesgue measure.

Example 4. Let $X = (X_1, \ldots, X_5)$ belong to the graphical model with the undirected graph $G = (V, E)$ in **Figure 3**. If X satisfies the pairwise Markov property, the missing edges $1 - 4 \notin E$ and $2 - 4 \notin E$ imply $X_4 \perp\!\!\!\perp X_1 \mid (X_2, X_3, X_5)$ and $X_4 \perp\!\!\!\perp X_2 \mid (X_1, X_3, X_5)$. The local Markov property for node $v = 4$ implies $X_4 \perp\!\!\!\perp (X_1, X_2) \mid (X_3, X_5)$. The global Markov property explicitly requires many other conditional independencies, for example $X_4 \perp\!\!\!\perp X_2 \mid (X_1, X_3)$ or $(X_4, X_5) \perp\!\!\!\perp (X_1, X_2) \mid X_3$.

The pairwise Markov property for undirected graphical models translates each absent edge into a full conditional independence. For this reason, the smallest undirected graph G with respect to which X is pairwise Markov is also known as the conditional independence graph of X. Testing all $\binom{|V|}{2}$ pairwise full conditional independencies that might arise in Equation 2 yields a method to estimate this graph. Addressing the multiple testing issues in this approach allows for control of false edge discoveries. We will not discuss this in detail but refer the reader to Drton & Perlman (2007), Liu (2013), and Wasserman et al. (2014).

Under the local Markov property, each variable X_v can be optimally predicted from its neighbors $X_{\mathrm{nb}(v)}$, which is used in a method known as neighborhood selection (see Section 3). The more comprehensive global Markov property is very useful for reasoning about conditional independence.

The famous theorem of Hammersley and Clifford clarifies the construction of distributions that possess the Markov properties. Suppose $X = (X_v : v \in V)$ has a density with respect to a product measure $\mu = \otimes_{v \in V} \mu_v$ on $\mathbb{R}^V$. In applications, each μ_v is usually either Lebesgue or a counting measure, so each X_v is (absolutely) continuous or discrete. Let $\mathcal{C}(G)$ be the set of all complete subsets (or cliques) of G, i.e., $C \in \mathcal{C}(G)$ if $\{v, w\} \in E$ for all $v, w \in C$. Then the distribution of X is said to factorize with respect to G if it has a density of the form

$$f(x) = \prod_{C \in \mathcal{C}(G)} \phi_C(x_C), \quad x \in \mathbb{R}^V, \tag{4}$$

where each potential function ϕ_C has domain $\mathbb{R}^C$, and x_C is the subvector $(x_v : v \in C)$. (For a finite set A, the space $\mathbb{R}^A$ comprises real vectors of length $|A|$ with entries indexed by A.) The potential functions need not have a probabilistic interpretation as conditional densities.

Markov properties: for undirected graphs, different versions of the Markov properties require different sets of conditional independences; global ⇒ local ⇒ pairwise, with equivalence under the intersection axiom

Theorem 1 (Hammersley-Clifford). Suppose $X = (X_v : v \in V)$ has a positive density with respect to a product measure. Then the distribution of X factorizes with respect to $G = (V, E)$ if and only if X satisfies the pairwise Markov property with respect to G.

A graphical model may thus be defined by specifying families of potential functions. If the functions are positive, then the pairwise and global Markov properties are equivalent. Moreover, the global Markov property lists every conditional independence that holds in all distributions with factorizing densities. This is known as completeness of the global Markov property. Factorization of nonpositive distributions for categorical variables is treated by Geiger et al. (2006). Matúš (2012) gives a new perspective on factorization as a consequence of log-convexity of a set of distributions.

Example 5. Suppose $X = (X_1, \ldots, X_5)$ is centered and multivariate normal with positive definite covariance matrix Σ. Let $K = (\kappa_{vw}) = \Sigma^{-1}$. Then X has density

$$f(x) = \frac{1}{\sqrt{(2\pi)^5 \det(\Sigma)}} \exp\left(-\frac{1}{2} \sum_{v,w=1}^{5} \kappa_{vw} x_v x_w\right), \quad x \in \mathbb{R}^5.$$

This distribution factorizes with respect to the graph in **Figure 3** if and only if $\kappa_{14} = \kappa_{15} = \kappa_{24} = \kappa_{25} = \kappa_{35} = 0$. More generally, the Gaussian model associated with an undirected graph $G = (V, E)$ comprises all normal distributions with a positive definite inverse covariance matrix $K = (\kappa_{vw}) \in \mathbb{R}^{V \times V}$ such that $\kappa_{vw} = 0$ when $v - w \notin E$.

Let S be the sample covariance matrix for an independent and identically distributed (i.i.d.) Gaussian sample of size n. The maximum likelihood estimator (MLE) in the Gaussian graphical model maximizes the log-likelihood function

$$L(K) = \log\det(K) - \operatorname{tr}(SK), \tag{5}$$

subject to K being positive definite with zeros over non-edges of G. Strictly speaking, Equation 5 is obtained by maximizing over an unknown mean vector, dividing out $n/2$, and omitting an additive constant. If $n > |V|$, then L admits a unique maximizer with probability one, because S is almost surely positive definite. If $|V| \geq n$, then S is singular and L can be unbounded. However, if the graph G is sparse, then the MLE of K may exist uniquely with probability one even if n is much smaller than $|V|$ (see, e.g., Sullivant & Gross 2014).

2.2. Directed Graphical Models

Directed acyclic graphs (DAGs) are directed graphs without directed cycles.[1] A random vector $X = (X_v : v \in V)$ satisfies the local Markov property with respect to a DAG G if

$$X_v \perp\!\!\!\perp X_{\mathrm{nd}_G(v) \setminus \mathrm{pa}_G(v)} \mid X_{\mathrm{pa}_G(v)}$$

for every $v \in V$. Similarly, X satisfies the global Markov property with respect to G if $X_A \perp\!\!\!\perp X_B \mid X_C$ for all triples of pairwise disjoint subsets $A, B, C \subset V$ such that C d-separates A and B in G, which we denote by $A \perp_G B \mid C$. The notion of d-separation in directed graphs is more subtle than separation in undirected graphs and is given in Definition 1 below.

If X satisfies the global Markov property with respect to G, then G is called an independence map of X. A DAG G is a perfect map of X if $A \perp_G B \mid C$ if and only if $X_A \perp\!\!\!\perp X_B \mid X_C$ for all

[1] Although graph theory speaks of "acyclic directed graphs" (or "acyclic digraphs"), the phrase "directed acyclic graph," which leads to the catchy abbreviation "DAG," has established itself in the literature on graphical models.

pairwise disjoint sets $A, B, C \subset V$. A perfect map thus requires the global Markov property and its reverse implication, known as faithfulness. The assumption that G is a perfect map is important for the structure learning methods that we discuss in Section 4.

Perfect map: a DAG is a perfect map of a distribution if the d-separations in the DAG exactly match the conditional independencies in the distribution

Suppose X has a density with respect to a product measure. Then the distribution of X factorizes according to a DAG $G = (V, E)$ if it has a density of the form

$$f(x) = \prod_{v \in V} f(x_v \mid x_{\mathrm{pa}_G(v)}), \tag{6}$$

where the $f(x_v \mid x_{\mathrm{pa}_G(v)})$ are conditional densities with $f(x_v \mid x_\emptyset) = f(x_v)$. The global and local Markov properties are equivalent, and under the assumed existence of a density they are equivalent to the factorization property (Verma & Pearl 1988). The global Markov property is also complete, that is, it states all conditional independencies that are implied by the factorization.

A directed graphical model for DAG G can be specified by conditional densities $f(x_v \mid x_{\mathrm{pa}_G(v)})$, $v \in V$. If these do not share common parameters, the likelihood function of the model factorizes into $|V|$ local likelihood functions, and the MLEs of $f(x_v \mid x_{\mathrm{pa}_G(v)})$, $v \in V$, can be computed separately. For categorical data, this amounts to computing empirical frequencies for conditional probability tables. For multivariate Gaussian data, the conditional means and variances are found via linear regression of each X_v on $X_{\mathrm{pa}_G(v)}$.

We now define d-separation. In a DAG $G = (V, E)$, nodes v and w are adjacent if $v \to w \in E$ or $w \to v \in E$, and a path is a sequence of distinct nodes in which successive nodes are adjacent. If $\pi = (v_0, v_1, \ldots, v_k)$ is a path, then v_0 and v_k are the endpoints of the path. A non-endpoint v_i is a collider on π if $v_{i-1} \to v_i \leftarrow v_{i+1}$ is a subpath of π. Otherwise, v_i is a noncollider on π. If every edge on π is of the form $v_{i-1} \to v_i$, then v_0 is an ancestor of v_k and v_k is a descendent of v_0. We write $\mathrm{adj}_G(v)$, $\mathrm{an}_G(v)$, and $\mathrm{de}_G(v)$ for the sets of adjacent nodes, ancestors and descendants of v in G, respectively. We use the convention that v is an ancestor and descendant of itself, and apply the notions disjunctively to sets, for example, $\mathrm{an}_G(C) = \cup_{v \in C} \mathrm{an}_G(v)$. Finally, we define $\mathrm{nd}_G(C) = V \setminus \mathrm{de}_G(C)$.

Definition 1. Two nodes v and w in a DAG $G = (V, E)$ are d-connected given $C \subseteq V \setminus \{v, w\}$ if G contains a path π with endpoints v and w such that (*a*) all colliders on π are in $\mathrm{an}_G(C)$, and (*b*) no noncollider on π is in C. Generalizing to sets, two disjoint subsets $A, B \subset V$ are d-connected given $C \subseteq V \setminus (A \cup B)$ if there are two nodes $v \in A$ and $w \in B$ that are d-connected given C. If this is not the case, then C d-separates A and B.

Example 6. Let $X = (X_1, \ldots, X_5)$ belong to the graphical model with the DAG G in **Figure 4*a***. We see that node 2 is a collider on the path $1 \to 2 \leftarrow 3 \to 4$, whereas it is a noncollider on $1 \to 2 \to 5$. For node 4, the local Markov property requires $X_4 \perp\!\!\!\perp (X_1, X_2, X_5) \mid X_3$. Because 1 and 4 are d-connected given $C = \{2\}$, $C = \{5\}$ and $C = \{2, 5\}$, but d-separated given any other subset of $\{2, 3, 5\}$, the global

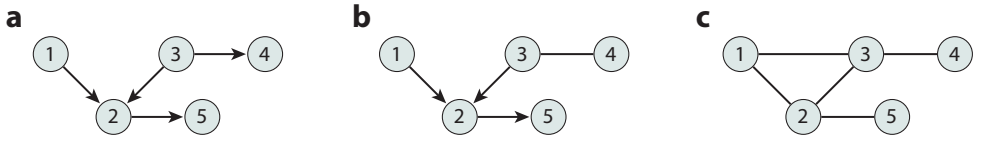

Figure 4

Example to illustrate Markov properties of directed acyclic graphs (DAGs). (*a*) A DAG, (*b*) the corresponding completed partially directed acyclic graph (CPDAG), and (*c*) the corresponding moral graph.

Markov property requires $X_1 \perp\!\!\!\perp X_4 \mid X_{C'}$ for any such other subset C' of $\{2, 3, 5\}$. We observe that, in contrast to separation in undirected graphs, d-separation in a DAG is not monotonic in the sense that $A \perp_G B \mid C$ does not imply that $A \perp_G B \mid C'$ for sets $C' \supsetneq C$. Note also that there does not exist an undirected graph that encodes the same conditional independencies as the DAG in **Figure 4*a***. Finally, the factorization for G takes the form $f(x) = f(x_1)f(x_2|x_1, x_3)f(x_3)f(x_4|x_3)f(x_5|x_2)$.

Markov equivalence: DAGs with the same skeleton and v-structures form a Markov equivalence class that can be represented by a completed partially directed acyclic graph (CPDAG)

Two DAGs G and G' are Markov equivalent if $A \perp_G B \mid C$ is equivalent to $A \perp_{G'} B \mid C$. Markov-equivalent DAGs are characterized by having the same skeleton and the same v-structures (Frydenberg 1990, Verma & Pearl 1990). A v-structure is a triple of nodes $u \to v \leftarrow w$ with u and w not adjacent. Each Markov equivalence class can be represented by a completed partially directed acyclic graph (CPDAG) that may have directed and undirected edges (e.g., Andersson et al. 1997, Roverato 2005). A CPDAG has edge $v \to w$ if and only if the edge $v \to w$ is common to all DAGs in its equivalence class. If the class contains a DAG with $v \to w$ and a DAG with $v \leftarrow w$, then the CPDAG has the undirected edge $v - w$. The DAG G in **Figure 4*a*** is Markov equivalent to exactly one other DAG, obtained by replacing the edge $3 \to 4$ by $3 \leftarrow 4$. The CPDAG of G is shown in **Figure 4*b***.

The skeleton of a (partially) directed graph is the undirected graph obtained by replacing all edges by undirected edges. The moral graph G^m of a DAG G is constructed by first shielding all v-structures and then taking the skeleton of the resulting graph. Shielding a v-structure $u \to v \leftarrow w$ means adding an edge between nodes u and w. **Figure 4*c*** shows the moral graph of the DAG in **Figure 4*a***. It is easy to see that if X satisfies the factorization property for directed graphs with respect to G, then X satisfies the factorization property for undirected graphs with respect to the moral graph G^m. Hence, if G is a perfect map of X, then G^m is the conditional independence graph of X. This implies that the skeleton of a DAG (or CPDAG) is a subgraph of its corresponding conditional independence graph.

The graphical model associated with a DAG $G = (V, E)$ can also be thought of as a structural equation model (Bollen 1989). Indeed, if $\epsilon = (\epsilon_v : v \in V)$ is a vector of independent random noise variables and g_v are measurable functions, then the random vector $X = (X_v : v \in V)$ given by

$$X_v = g_v(X_{\mathrm{pa}_G(v)}, \epsilon_v), \quad v \in V, \tag{7}$$

is Markov with respect to G. Conversely, if X is Markov with respect to G, then there are independent variables ϵ_v and functions g_v such that Equation 7 holds.

Example 7. If all functions g_v are linear and the ϵ_v are normal random variables, then we may assume that Equation 7 takes the form $X = BX + \epsilon$, where the matrix $B = (\beta_{vw}) \in \mathbb{R}^{V \times V}$ has $\beta_{vw} = 0$ if $w \notin \mathrm{pa}_G(v)$. The solution $X = (I - B)^{-1}\epsilon$ follows a multivariate normal distribution with covariance matrix $\mathrm{Cov}(X) = (I - B)^{-1}\mathrm{Cov}(\epsilon)(I - B)^{-T}$. Here, $I - B$ is invertible because $\det(I - B) = 1$ by acyclicity of G.

Structural equation models, and thus also directed graphical models, admit a natural causal interpretation. To this end, one views Equation 7 as specifying an assignment mechanism, which is clarified by writing

$$X_v \leftarrow g_v(X_{\mathrm{pa}_G(v)}, \epsilon_v), \quad v \in V. \tag{8}$$

The variables in $X_{\mathrm{pa}_G(v)}$ are then treated as direct causes of X_v, meaning that changes in $X_{\mathrm{pa}_G(v)}$ may lead to changes in X_v, but not the other way around. This interpretation allows statements about the distribution of X under experimental interventions. In particular, interventions to the

system can be modeled by changing the structural equations for precisely those variables that are affected by the intervention (Pearl 2009).

Example 8. The structural equation model for the DAG in **Figure 4*a*** postulates that

$$X_1 \leftarrow g_1(\epsilon_1), \quad X_3 \leftarrow g_3(\epsilon_3), \quad X_2 \leftarrow g_2(X_1, X_3, \epsilon_2), \quad X_4 \leftarrow g_4(X_3, \epsilon_4), \quad X_5 \leftarrow g_5(X_2, \epsilon_5).$$

We now consider an intervention on X_2, where X_2 is generated as an independent draw from the distribution of ϵ_2. Denoting the postintervention variables by $\tilde{X} = (\tilde{X}_v : v \in V)$, the distribution of $\tilde{X}$ is induced by the equation system

$$\tilde{X}_1 = g_1(\epsilon_1), \qquad \tilde{X}_3 = g_3(\epsilon_3), \qquad \tilde{X}_2 = \epsilon_2, \qquad \tilde{X}_4 = g_4(\tilde{X}_3, \epsilon_4), \qquad \tilde{X}_5 = g_5(\tilde{X}_2, \epsilon_5).$$

The postintervention DAG $\tilde{G}$ is obtained from G by removing the edges $1 \to 2$ and $3 \to 2$.

Structural equation models: causal interpretation of DAGs obtained by treating equations as assignment mechanisms

Chow-Liu trees: ML trees that are maximum spanning trees for a complete graph weighted by mutual information

Thus, the causal interpretation of a DAG allows predictions in changed environments, and hence the estimation of causal effects. These ideas can be combined with the structure learning methods that we discuss in Section 4 to estimate (bounds on) causal effects from observational data (Maathuis et al. 2009, 2010; Perković et al. 2015; Nandy et al. 2016b).

3. LEARNING UNDIRECTED GRAPHICAL MODELS

Our treatment of learning undirected graphical models begins with the special case of trees. We then move on to generally applicable greedy search and ℓ_1 penalization methods as well as techniques that avoid traditional assumptions of Gaussianity for continuous observations.

3.1. Chow-Liu Trees and Forests

A tree $G = (V, E)$ is an undirected graph with a unique path between any two nodes, and thus $|E| = |V| - 1$. Chow & Liu (1968) showed that one can efficiently find a tree-structured distribution that optimally approximates a given distribution. In the present context, we may view their algorithm as outputting a maximum likelihood (ML) estimate (MLE) of a conditional independence tree.

Owing to convenient factorizations, computation in tree-based graphical models is particularly tractable (Wainwright & Jordan 2008). Indeed, if the distribution of a random vector $X = (X_v : v \in V)$ factorizes with respect to a tree, then the joint density factorizes as

$$f(x) = \prod_{v-w \in E} \frac{f_{vw}(x_v, x_w)}{f_v(x_v) f_w(x_w)} \prod_{v \in V} f_v(x_v). \tag{9}$$

Here, f_v and f_{vw} are the marginal densities of X_v and (X_v, X_w), respectively. Equation 9 is a special case of a more general result exemplified in Equation 10. It coincides with Equation 6 when we create a directed tree by letting edges point away from one arbitrarily selected node.

Suppose for a moment that all variables are categorical, with X_v taking values in a finite set $\mathcal{X}_v$. For a joint state $x \in \mathcal{X} := \prod_{v \in V} \mathcal{X}_v$, let $N(x)$ be the number of times x appears in an i.i.d. sample of size n from the distribution of X. Writing $\hat{f}_G(x)$ for the MLE of the joint probability $f(x)$ in the graphical model given by tree G, we aim to find the tree G with largest maximum log-likelihood $\hat{L}(G) = \sum_{x \in \mathcal{X}} N(x) \log \hat{f}_G(x)$.

Let $\hat{f}_{vw}(x_v, x_w)$ and $\hat{f}_v(x_v)$ be the relative frequencies of seeing the pair (X_v, X_w) in state (x_v, x_w) and variable X_v in state x_v, respectively. The MLE $\hat{f}_G(x)$ is obtained by plugging the $\hat{f}_{vw}$ and $\hat{f}_v$

into Equation 9. It follows that

$$\frac{1}{n}\hat{L}(G) = \sum_{v-w\in E} I\left(\hat{f}_{vw}\right) + \text{const.},$$

where $I(\hat{f}_{vw})$ is the empirical mutual information of X_v and X_w, so

$$I\left(\hat{f}_{vw}\right) = \sum_{x_v\in\mathcal{X}_v}\sum_{x_w\in\mathcal{X}_w} \hat{f}_{vw}(x_v, x_w)\log\frac{\hat{f}_{vw}(x_v, x_w)}{\hat{f}_v(x_v)\hat{f}_w(x_w)}.$$

Because mutual information is nonnegative, the ML tree is a maximum spanning tree for the complete graph with edge weights $I(\hat{f}_{vw})$. The maximum spanning tree can be computed efficiently using, for example, Kruskal's algorithm, which adds edges $\{v, w\}$ in the order of decreasing mutual information $I(\hat{f}_{vw})$ but skips edges that create a cycle. The ML tree is found after addition of $|V|-1$ edges. If Kruskal's algorithm is stopped early, adding only k edges, then the output is a forest with maximum likelihood among all forests with k edges. A forest is an undirected graph that is a union of disconnected trees.

The Chow-Liu method is not limited to categorical data. Indeed, we may formulate statistical models for the bivariate marginal distributions f_{vw}, compute their ML estimates $\hat{f}_{vw}$, and find a maximum weight spanning tree from their mutual informations $I(\hat{f}_{vw})$. In particular, when the marginals are taken to be bivariate normal, then the joint density f is multivariate normal and $I(\hat{f}_{vw}) = -\frac{1}{2}\log(1-r_{vw}^2)$, where r_{vw} is the empirical correlation between X_v and X_w. In this case, the absolute correlations $|r_{vw}|$ can also be used as weights.

Although it is an older idea, there has been renewed interest in Chow and Liu's approach. Tan et al. (2010, 2011) study which trees/forests are most difficult to recover. Liu et al. (2011) use kernel density estimates in a nonparametric approach. Edwards et al. (2010) discuss mixed categorical and continuous data and incorporate information criteria into the algorithm. The output is then a forest because the penalties for model complexity may yield negative edge weights. Treating models with latent variables, Friedman et al. (2002) suggest a structural EM algorithm whose M-step optimizes over both parameters and tree structure. The Chow-Liu algorithm is also used in methods for learning latent locally tree-like graphs by Anandkumar & Valluvan (2013).

3.2. Greedy Search

Beyond the realm of trees, finding a graph that maximizes a (penalized) likelihood or information criterion is hard in a complexity-theoretic sense (Karger & Srebro 2001). Nevertheless, good estimates can be obtained from heuristic techniques, such as greedy or stepwise forward/backward search. Good implementations, for example as discussed by Højsgaard et al. (2012), evaluate the benefit of an edge addition or removal via local computation based on clique-sum decompositions (Lauritzen 1996, chapter 3).

Local computation: efficient computation of the likelihood after edge additions and removals by exploiting graph decompositions

Example 9. The graph G in **Figure 5*a*** is a clique-sum of three smaller graphs, the so-called prime components shown in **Figure 5*b***. If a distribution factorizes with respect to G, then its density f satisfies

$$f(x) = \frac{f_{1234}(x_1, x_2, x_3, x_4) f_{345}(x_3, x_4, x_5) f_{567}(x_5, x_6, x_7)}{f_{34}(x_3, x_4) f_5(x_5)}. \quad (10)$$

The marginal densities in the numerator correspond to the prime components, and those in the denominator correspond to the separating sets, i.e., the cliques along which the prime components are summed. Suppose now that we wish to compute the likelihood ratio statistic comparing the given graph to the graph with edge $3-5$ removed, in a setting of categorical data under multinomial sampling.

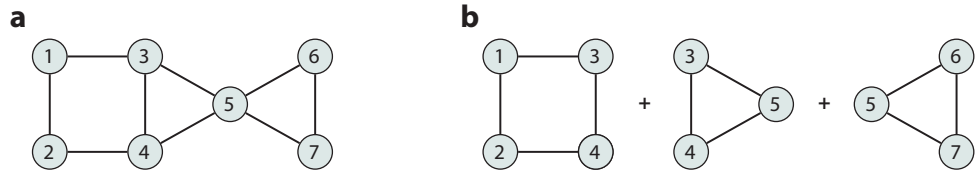

Figure 5

A clique-sum decomposition. (*a*) An undirected graph. (*b*) Clique-sum of the prime components of panel *a*.

The edge belongs to the clique $\{3, 4, 5\}$, and proposition 4.32 in Lauritzen (1996) implies that the likelihood ratio statistic can be obtained by computing with the data on (X_3, X_4, X_5) alone. Specifically, one computes the likelihood ratio of the graph $3 - 4 - 5$ with respect to the full triangle graph on $\{3, 4, 5\}$. Analogous results exist for the Gaussian case and, with additional subtleties, for mixed discrete and conditionally Gaussian observations (Lauritzen 1996, chapters 5 and 6).

Computation of likelihood ratios is particularly convenient for decomposable graphs, that is, graphs in which all prime components are complete. For example, trees are decomposable because their prime components are the individual edges, and the graph in **Figure 5*a*** is nondecomposable. For a decomposable graph, Gaussian models as well as models for categorical variables admit closed-form MLEs (Lauritzen 1996, sections 4.4 and 5.3). These are obtained by estimating marginal distributions (via sample means and covariances for Gaussian data, or via empirical frequencies for categorical data) and substituting these estimates in a factorization such as Equation 10. In contrast, computing MLEs for nondecomposable graphs involves solving higher degree polynomial equation systems (Drton et al. 2009, chapter 2.1).

More recently, greedy search has been applied in a framework of neighborhood selection, in which a graph G is selected by determining the neighborhood $\mathrm{nb}_G(v)$ of each node v. As discussed in Section 3.4, finding $\mathrm{nb}_G(v)$ often corresponds to variable selection in a regression problem. This avoids the need for iterative computation of MLEs when dealing with nondecomposable graphs, and a connection can be made to results on forward selection methods for variable selection in high-dimensional regression. Jalali et al. (2011) and Ray et al. (2015) leverage this connection and provide theoretical guarantees for greedy search in high-dimensional problems. The related method of Bresler (2015) greedily selects supersets of the neighborhoods that are subsequently pruned. These methods are competitive to the ℓ_1-regularization techniques that we discuss next.

3.3. Gaussian Models and ℓ_1-Penalization

Gaussian models provide the starting point for most graphical modeling of continuous observations. As noted in Example 5, a Gaussian conditional independence graph can be estimated by determining the zero entries of the inverse covariance matrix $K \in \mathbb{R}^{V \times V}$.

Let $S = (s_{vw})$ be the sample covariance matrix for a sample of n observations, with $s_{vv} > 0$ for all $v \in V$ to avoid trivialities. Yuan & Lin (2007) and Banerjee et al. (2008) proposed the graphical lasso (glasso) estimator

$$\hat{K}^{\mathrm{gl}} \;=\; \arg\min_{K} \; \{ -\log\det(K) + \mathrm{tr}\,(SK) + \lambda \|K\|_1 \}, \qquad (11)$$

Graphical lasso (glasso): estimator that optimizes ℓ_1-penalized joint Gaussian log-likelihood

where the minimization is over positive definite matrices $K = (\kappa_{vw}) \in \mathbb{R}^{V \times V}$ and $\lambda \geq 0$ is a tuning parameter. The objective adds to the log-likelihood function from Equation 5 a multiple of the (vector) ℓ_1-norm, i.e., $\|K\|_1 = \sum_{v,w \in V} |\kappa_{vw}|$. Some authors omit the positive diagonal entries κ_{vv}

when forming the norm. In either case, such a regularization term induces sparsity in $\hat{K}^{\mathrm{gl}}$, just as it does in lasso regression. The conditional independence graph is then estimated by the graph $\hat{G}^{\mathrm{gl}}$ that has edge $v - w$ if and only if $\hat{\kappa}^{\mathrm{gl}}_{vw} \neq 0$. For $\lambda > 0$, the minimum in Equation 11 is achieved uniquely because the objective is strictly concave and coercive irrespective of whether S has full rank. This is important in high-dimensional settings.

The coordinate-descent algorithm of Friedman et al. (2008) is a popular method for computation of $\hat{K}^{\mathrm{gl}}$; the reader is directed to Mazumder & Hastie (2012b) for a discussion of its properties. Recent implementations exploit the fact that simply thresholding the sample covariance matrix S yields the connected components of $\hat{G}^{\mathrm{gl}}$ (Witten et al. 2011, Mazumder & Hastie 2012a). Alternative approaches for computation of $\hat{K}^{\mathrm{gl}}$ are discussed by Hsieh et al. (2013). The estimation error of $\hat{K}^{\mathrm{gl}}$ and the consistency of $\hat{G}^{\mathrm{gl}}$ in high-dimensional problems are studied by Ravikumar et al. (2011).

There are several related methods to estimate a sparse covariance matrix. For example, Cai et al. (2011) minimize $\|K\|_1$ subject to a constraint on $\|SK - I\|_\infty$. Here, I is the identity and $\|A\|_\infty$ is the maximum absolute entry of A. Minimax optimality properties can be proven for an adaptive version of the estimator (Cai et al. 2016).

A conditional independence graph is sometimes expected to have particular structure, such as hub nodes with many neighbors. This motivated Defazio & Caetano (2012) to consider regularization with a sorted ℓ_1 norm. For tuning parameters $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$, the sorted ℓ_1 norm of $\beta \in \mathbb{R}^p$ is $\sum_{j=1}^{p} \lambda_j |\beta_{(j)}|$, where $\beta_{(1)}, \ldots, \beta_{(p)}$ are the entries of β listed in descending order of absolute values, so $|\beta_{(1)}| \geq |\beta_{(2)}| \geq \cdots \geq |\beta_{(p)}|$. Defazio & Caetano (2012) estimate the inverse covariance $K = (\kappa_{vw})$ by the optimal solution of

$$\min_K \left\{ -\log\det(K) + \operatorname{tr}(SK) + \sum_{v \in V} \sum_{j=1}^{|V|} \lambda_j |\kappa_{v,(j)}| \right\}. \tag{12}$$

Intuitively, the sorted ℓ_1 norm allows one to more easily detect a signal κ_{vw} if it concerns a variable X_v with other stronger signals κ_{vu}. Of course, the work just described is not the only one addressing hubs (see, e.g., Tan et al. 2014).

3.4. Neighborhood Selection

We may estimate a conditional independence graph $G = (V, E)$ by estimating all of its neighborhoods $\mathrm{nb}_G(v)$. According to the local Markov property, X_v depends on the other variables only through its neighbors X_w, $w \in \mathrm{nb}_G(v)$. Hence, we may proceed by estimating the conditional distribution of X_v given $X_{V \setminus \{v\}}$ and determining $\mathrm{nb}_G(v)$ as the index set of the variables X_w on which the estimated conditional distribution depends. (Alternatively, we could motivate the approach by referring to the pairwise Markov property, from which we conclude that $\{v, w\} \notin E$ if the conditional distribution does not depend on X_w.) The ideas behind this neighborhood selection have a longer tradition (Besag 1975). However, its widespread use in graphical modeling emerged more recently when Meinshausen & Bühlmann (2006) tackled high-dimensional problems through a connection to lasso regression.

Neighborhood selection: neighbors are found using techniques for variable selection such as ℓ_1 penalization

Example 10. Let $X = (X_v : v \in V)$ be multivariate normal with inverse covariance matrix $K = (\kappa_{vw})$. Then the conditional distribution of X_v given all remaining variables is normal with variance $1/\kappa_{vv}$ and expectation

$$\mathbb{E}[X_v \mid X_w,\ w \neq v] = \sum_{w \in V \setminus \{v\}} \left(-\frac{\kappa_{vw}}{\kappa_{vv}}\right) X_w = \sum_{w \in \mathrm{nb}_G(v)} \left(-\frac{\kappa_{vw}}{\kappa_{vv}}\right) X_w.$$

We may estimate $\mathrm{nb}_G(v)$ as the set of active covariates in a linear regression of X_v on all other variables X_w, $w \neq v$. Any technique for variable selection could be applied.

Example 11. In a symmetric Ising model with (real-valued) interaction parameters θ_{vw}, all random variables X_v take values in $\{-1, 1\}$ and joint probabilities have the form

$$\Pr(X_v = x_v,\ v \in V) \ \propto\ \exp\left\{ \sum_{\{v,w\} \in E} \theta_{vw} x_v x_w \right\}. \tag{13}$$

By the Hammersley-Clifford theorem, the conditional independence graph $G = (V, E)$ of such a distribution has edge $v - w \in E$ if and only if $\theta_{vw} \neq 0$. Because

$$\log\left(\frac{\Pr(X_v = 1 \,|\, X_w = x_w,\ w \neq v)}{1 - \Pr(X_v = 1 \,|\, X_w = x_w,\ w \neq v)} \right) \ =\ \sum_{w \in \mathrm{nb}_G(v)} (2\theta_{vw})\, x_w,$$

the neighborhood $\mathrm{nb}_G(v)$ can be estimated as the set of active covariates in a logistic regression of X_v on all other variables X_w, $w \neq v$. We note that the normalizing constant in Equation 13 is a sum over $2^{|V|}$ joint states and thus intractable unless $|V|$ is small.

Estimating each neighborhood in isolation may lead to inconsistencies that are commonly resolved post-hoc. Let $\widehat{\mathrm{nb}}(v)$, $v \in V$, be the estimated neighborhoods. If $w \in \widehat{\mathrm{nb}}(v)$ but $v \notin \widehat{\mathrm{nb}}(w)$, then the so-called and-rule excludes the edge $v - w$ from the estimated conditional independence graph, whereas the or-rule includes such edges.

The use of ℓ_1-regularization for neighborhood selection in Gaussian and Ising models (Examples 10 and 11) is studied by Meinshausen & Bühlmann (2006) and Ravikumar et al. (2010), respectively. Yang et al. (2015a) treat other exponential family models, and Chen et al. (2015) propose refinements of the and-/or-rules that take into account the distributional type of the nodes. Voorman et al. (2014) apply techniques for sparse additive models, in which a conditional expectation of the form $\mathbb{E}[X_v \,|\, X_w,\ w \neq v] \ =\ \sum_{w \neq v} f_{vw}(X_w)$ is estimated using basis expansions and a group lasso penalty that allows for zero functions as estimates of some of the univariate functions f_{vw}.

Neighborhood selection is related to the concept of pseudo-likelihood, which is based on the full conditional densities of a joint density f. The log-pseudo-likelihood function is

$$L_{\mathrm{pseudo}}(f) \ =\ \sum_{v \in V} \log f_{v|V\setminus\{v\}}\,(x_v \,|\, x_w,\ w \neq v)\,. \tag{14}$$

Non-Gaussian distributions specified via the Hammersley-Clifford theorem typically have intractable normalizing constants; recall Example 11. In contrast, the full conditionals in the pseudo-likelihood can often be normalized. Indeed, the conditional probabilities for discrete X_v can be normalized by summing over the state space of X_v alone as opposed to over all joint states. Similarly, it may be feasible to find the normalizing constant of the full conditional density for a continuous random variable by univariate integration.

Although neighborhood selection treats the different conditional densities $f_{v|V\setminus\{v\}}$ as unrelated, the conditionals share parameters. For instance, the interaction parameter θ_{vw} in the Ising model from Equation 13 appears in the conditionals for X_v and for X_w. As an alternative that avoids inconsistencies among estimated neighborhoods, we may maximize an ℓ_1 penalized version of L_{pseudo} from Equation 14 with respect to a symmetric interaction matrix. Höfling & Tibshirani (2009) explore this pseudo-likelihood method for Ising models but find rather little difference with neighborhood selection. Khare et al. (2015) give an overview of Gaussian pseudo-likelihood

methods that retain the symmetry of the inverse covariance matrix and address issues in the specification of a convex optimization objective.

Score matching: convex quadratic loss for continuous non-Gaussian graphical models

A full likelihood may be expected to yield more efficient estimators than neighborhood selection or pseudo-likelihood. However, under ℓ_1 regularization, the situation is subtle, as different irrepresentability conditions are needed to ensure consistency. Indeed, there are Gaussian examples in which neighborhood selection is consistent whenever the glasso is consistent, but in which the converse is false (Meinshausen 2008, Ravikumar et al. 2011).

3.5. Score Matching

The Hammersley-Clifford theorem allows the specification of graphical models as interaction models in the form of an exponential family. However, the normalizing constants in such models are tractable only in special cases. The score matching approach of Hyvärinen (2005, 2007) is well suited to address this challenge. We describe the basic version that applies to continuous observations supported on all of $\mathbb{R}^V$.

Let $X = (X_v : v \in V)$ be absolutely continuous with differentiable density f_0 and support $\mathbb{R}^V$. Let f be another density that is twice differentiable and has support $\mathbb{R}^V$. Writing ∇_x for the gradient with respect to x, define the Fisher information distance

$$J(f) = \int_{\mathbb{R}^V} f_0(x) \left\| \nabla_x \log f(x) - \nabla_x \log f_0(x) \right\|_2^2 \, \mathrm{d}x. \tag{15}$$

Although it is natural to minimize an estimate of $J(\cdot)$, this approach is complicated by the way the unknown true density f_0 appears in Equation 15. Hyvärinen (2005) circumvents this problem using integration by parts (Stein's identity), which yields under mild conditions that

$$J(f) = \int_{\mathbb{R}^V} f_0(x) \left[\Delta_x \log f(x) + \frac{1}{2} \|\nabla_x \log f(x)\|_2^2 \right] \mathrm{d}x + \text{const.}, \tag{16}$$

where $\Delta_x f(x) = \sum_v \partial^2 f(x)/\partial x_v^2$ is the Laplace operator. Writing $S(x, f) = \Delta_x \log f(x) + \frac{1}{2} \|\nabla_x \log f(x)\|_2^2$, a score matching estimator minimizes the empirical loss

$$\hat{J}(f) = \frac{1}{n} \sum_{i=1}^{n} S(x^{(i)}, f)$$

for f ranging over a model of interest. Importantly, if f is only known up to a normalizing constant, then this constant cancels in the logarithmic derivatives in $S(x, f)$. Moreover, in an exponential family with log-densities $\log f(x|\theta) = \theta^T t(x) - \psi(\theta) + b(x)$ for sufficient statistics $t(x)$, the loss $\hat{J}$ is a convex quadratic function of the natural parameter θ.

Example 12. Consider the family of centered multivariate normal distributions, parameterized by their inverse covariance matrices K. Then,

$$\hat{J}(K) = \frac{1}{2} \text{trace}(K^2 S) - \text{trace}(K) = \sum_{v \in V} \frac{1}{2} \kappa_v^T S \kappa_v - \kappa_v^T e_v, \tag{17}$$

where S is the sample covariance matrix of X, κ_v is the vth column of K, and e_v is the vth canonical basis vector of $\mathbb{R}^V$. If S is invertible then the score matching estimator equals the MLE $\hat{K} = S^{-1}$. However, for submodels that constrain K to lie in a linear subspace, the two estimators generally differ. The score matching estimator need not be asymptotically efficient, as Forbes & Lauritzen (2015) show in the context of symmetry constraints in graphical models.

Closed-form score matching estimators are available for any pairwise interaction model

$$\log f(x|\theta) = \sum_{a=1}^{A} \sum_{v \neq w} \theta_{vw}^{(a)} t_{vw}^{(a)}(x_v, x_w) + \sum_{l=1}^{L} \sum_{v} \theta_v^{(l)} t_v^{(l)}(x_v) - \psi(\theta) + b(x), \quad x \in \mathbb{R}^V, \tag{18}$$

for which $X_v \perp\!\!\!\perp X_w \mid X_{V \setminus \{v,w\}}$ if and only if $\theta_{vw}^{(a)} = 0$ for all $a = 1, \ldots, A$. Sparse estimates of the interaction matrices $(\theta_{vw}^{(a)})$ can be obtained by adding an ℓ_1 or group lasso penalty to the loss $\hat{J}$. The resulting estimators of conditional independence graphs are studied by Lin et al. (2016), who also treat nonnegative observations; by Janofsky (2015), who proposes a nonparametric exponential series approach; and by Sun et al. (2015), who consider infinite-dimensional exponential families. For Gaussian models, ℓ_1-regularized score matching is a simple but state-of-the-art method. It coincides with the method of Liu & Luo (2015).

3.6. Semiparametric and Robust Methods

Traditionally, methods for continuous observations have relied heavily on Gaussian models. A problem of obvious interest is to provide methods for non-Gaussian observations. We already mentioned several such methods and comment here on two other lines of work.

From a perspective of robustness, several authors explored the use of elliptical distributions (Bilodeau 2014, Vogel & Fried 2011, Vogel & Tyler 2014). Finegold & Drton (2011) consider the special case of t-distribution models in which a Gaussian random vector is observed under scaling with a single random divisor. They also propose nonelliptical alternative t-distributions, which result from dividing the different components of a latent Gaussian vector by independent scalars. Different divisors are useful for high-dimensional data with outliers in many observations, but where each observation only has a small number of corrupted entries.

In a different vein, Liu et al. (2009) propose semiparametric methods based on Gaussian copula models. Here, the observation $X = (X_v : v \in V)$ satisfies $X_v = h_v(Z_v)$ for a Gaussian random vector $Z = (Z_v : v \in V)$ and strictly increasing functions $h_v : \mathbb{R} \to \mathbb{R}$. Because the h_v are deterministic and one-to-one, X has the same Markov properties as Z. Liu et al. (2012a) and Xue & Zou (2012) observe that efficient estimation in the copula models can be based on Kendall's τ and Spearman's ρ. Indeed, for strictly increasing h_v, the observation X and the latent vector Z have the same rank correlations. One may thus apply Gaussian methods after estimating the latent Gaussian correlation matrix by suitably transformed pairwise estimates of τ or ρ. In the data analysis in Example 3, we applied this idea in conjunction with Gaussian neighborhood selection from Section 3.4.

It is noteworthy that Kendall's τ also provides a simple way to fit copula models based on elliptical distributions (Liu et al. 2012b). Extensions to mixed discrete and continuous data are treated by Fan et al. (2016). Avoiding the assumption of a Gaussian copula, Yang et al. (2014) use coarser data summaries than ranks to handle settings in which the full conditionals are generalized linear models with unknown base measure.

3.7. Tuning Parameter Selection

Many of the aforementioned methods depend on a tuning parameter. Varying this parameter typically yields a useful ranking of edges. However, it may also be desirable to select a single tuning parameter, for example by optimizing information criteria such as the Bayesian information criterion (BIC). However, in problems with a large number of variables $|V|$, the BIC tends to yield overly dense graphs. Foygel & Drton (2010), Gao et al. (2012), and Barber & Drton (2015) address

this issue by adapting ideas from sparse high-dimensional regression. Via a multiplicity-correcting prior, the BIC penalty is made dependent on $\log |V|$.

Another useful approach is stability selection (Meinshausen & Bühlmann 2010, Shah & Samworth 2013). This method records how often each edge is selected across random subsamples and different tuning parameters, and selects those edges for which there exists a tuning parameter so that the subsample selection frequency exceeds a specified threshold. The method aims to avoid false positives. A resampling technique that seeks to avoid false negatives was proposed by Liu et al. (2010). This method chooses the least amount of penalization for which graph estimates are suitably stable across subsamples.

Finally, we note that there are methods that aim to reduce or eliminate dependence on tuning parameters (see Lederer & Müller 2014 and references therein).

4. LEARNING DIRECTED GRAPHICAL MODELS

We now discuss methods for learning the structure of a directed graphical model. Textbooks on this problem include Spirtes et al. (2000), Neapolitan (2004), and Koller & Friedman (2009). Throughout, we assume that the DAG $G = (V, E)$ is a perfect map of $X = (X_v : v \in V)$, and that we observe n i.i.d. copies of X. The observed data are denoted by $\mathbf{x}$.

In general, G is not identifiable from the distribution of X, but we can identify its Markov equivalence class, or equivalently, its CPDAG. Thus, many structure learning methods aim to learn the CPDAG. We treat exact score-based search in problems of moderate dimensionality and review more broadly applicable methods based on greedy search or conditional independence tests, as well as hybrids of these two approaches. Finally, we discuss methods that impose additional assumptions that allow identification of the DAG.

4.1. Exact Score-Based Search

Score-based approaches learn a DAG by determining the graph G that optimizes a specified score $Q(G, \mathbf{x})$. Typically Q is a penalized likelihood score, for example the BIC. Such scores are often decomposable, meaning that $Q(G, \mathbf{x}) = \sum_{v \in V} q(v \mid \mathrm{pa}_G(v), \mathbf{x})$, where the summands are local scores of each node v given its parents. For many models, scores such as the BIC are also score-equivalent, meaning that $Q(G, \mathbf{x}) = Q(G', \mathbf{x})$ if G and G' are Markov equivalent. This is the case, in particular, for Gaussian models and for models for categorical data.

Finding an optimal DAG, or possibly CPDAG, is hard owing to the large search space and the acyclicity constraint. For instance, there are over 10^{36} (vertex-labeled) DAGs on 14 nodes. Nevertheless, for decomposable scores, an exact search is feasible more broadly than one might expect. Different approaches to exact search include branch and bound methods (e.g., de Campos et al. 2009), partial order covers (e.g., Parviainen & Koivisto 2009), and as we discuss in more detail below, dynamic programming and integer linear programming.

Silander & Myllymäki (2006) propose an elegant dynamic programming approach. It leverages the fact that a best DAG for a variable set $W \subseteq V$ can be thought of as a best sink $s \in W$, with best parents among subsets of $W \setminus \{s\}$, and a best DAG for $W \setminus \{s\}$. Using dynamic programming and starting from the singleton sets, a best sink can be found for all subsets $W \subseteq V$. Backtracking then yields an ordering of the nodes that is compatible with a best DAG on V. Given this ordering, one can use regression to select parents for each node from its predecessors in the ordering, yielding a best DAG on V.

Although the computational and memory requirements are exponential in $|V|$, this approach is feasible for problems with up to roughly 30 nodes.

More recently, authors such as Jaakkola et al. (2010), Cussens & Bartlett (2013) and Studený & Haws (2014) suggested integer linear programming approaches, in which the search over graph structures is formulated as a linear program over a polytope $\mathcal{P}$ representing DAGs. For instance, the vertices of $\mathcal{P}$ may be taken to be sparse binary vectors $\eta = (\eta_1, \ldots, \eta_{|V|})$, where each η_v is of length $2^{|V|-1}$, which is the number of possible parent sets. If $\mathrm{pa}(v) = s_v$, then we set $\eta_v(s_v) = 1$ and all other entries of η_v are zero. The polytope $\mathcal{P}$ is then the convex hull of all binary vectors η that correspond to DAGs. A key property of $\mathcal{P}$ is that cyclic graphs lie outside of $\mathcal{P}$. Using the notation η also for interior points of $\mathcal{P}$, the structure learning problem can be cast as

$$\max \sum_{v \in V} \sum_{s_v \subseteq V \setminus \{v\}} \eta_v(s_v) q(v \mid s_v, \mathbf{x}) \quad \text{s.t.} \quad \eta \in \mathcal{P}.$$

The complexity of this linear program is in the facets that define the polytope $\mathcal{P}$, and practical algorithms are based on relaxations of $\mathcal{P}$.

Greedy equivalence search (GES): greedy score-based search on the space of CPDAGs that provides consistency despite greedy search

4.2. Greedy Score-Based Search

For large graphs, exact search is infeasible, and one can turn to greedy search. A well-known algorithm of this type is the greedy equivalence search (GES) algorithm of Chickering (2002). Given a starting graph (often the empty graph) and a score, GES performs a greedy search on the space of CPDAGs. The algorithm performs a forward phase in which edges are added, and a backward phase in which edges can be removed. Efficient implementations use local computations to evaluate the benefit of an added or deleted edge, using the decomposability of the score. The forward phase tends to take much longer in practice than the backward phase.

Owing to the greedy search, GES will typically not find the global optimum of the score given data $\mathbf{x}$ with a sample of size n. Remarkably, however, Chickering (2002) showed that GES does find the global optimum with probability converging to 1 as $n \to \infty$, if the score is decomposable, score-equivalent, and consistent. The consistency property ensures that the true CPDAG gets the highest score with probability converging to 1 as $n \to \infty$. An important ingredient of Chickering's proof is the fact that the forward phase outputs an independence map of X, with probability converging to 1 as $n \to \infty$.

Although the number of added and deleted edges in GES is polynomial in the number of nodes $|V|$, the number of performed score evaluations can be exponential in $|V|$. Chickering & Meek (2015) show that the backward phase of GES can be made polynomial for sparse graphs. With a naive forward phase that simply gives the complete graph (which is trivially an independence map), this yields a polynomial time algorithm for sparse graphs.

The consistency result of Chickering (2002) is for a classical setting where $|V|$ is fixed and $n \to \infty$. Van de Geer & Bühlmann (2013) consider a high-dimensional setting where $|V|$ is allowed to grow with n. They show that the global optimum of the ℓ_0-penalized likelihood score is consistent, but they do not propose an algorithm to find the global optimum. Nandy et al. (2016a) show a first high-dimensional consistency result for GES.

4.3. Constraint-Based Methods

Constraint-based methods seek to find a DAG that is compatible with the conditional independencies seen in the given data set. We begin the discussion of such methods by focusing on the estimation of the skeleton of the DAG (or, equivalently, its CPDAG), which is typically the most computationally intensive step.

If G is a perfect map of X, then v and w are adjacent in G if and only if X_v and X_w are conditionally dependent given X_C for all $C \subseteq V \setminus \{v, w\}$. A naive approach thus determines if v

and w are adjacent by testing all $2^{|V|-2}$ conditional independencies, as done by the SGS (Spirtes, Glymour, and Scheines) or IC (inductive causation) algorithms (Verma & Pearl 1990, Spirtes et al. 2000). In contrast, the PC algorithm (named after its authors Peter Spirtes and Clark Glymour) of Spirtes et al. (2000) limits the number of conditional independence tests, by using the fact that v and w are not adjacent in a DAG G if and only if $X_v \perp\!\!\!\perp X_w \mid X_{\mathrm{pa}_G(v)}$ or $X_v \perp\!\!\!\perp X_w \mid X_{\mathrm{pa}_G(w)}$. Because we do not know the parent sets (this would mean knowing the DAG), the PC algorithm ensures that the following conditional independencies are tested: $X_v \perp\!\!\!\perp X_w \mid X_C$ for all $C \subseteq \mathrm{adj}_G(v)$ and all $C \subseteq \mathrm{adj}_G(w)$, and removes the edge if a conditional independence is found. This still appears infeasible, however, because we do not know $\mathrm{adj}_G(v)$ and $\mathrm{adj}_G(w)$. The PC algorithm circumvents this problem by always working with supergraphs of the true skeleton and testing conditional independencies given subsets of adjacency sets in these supergraphs.

Constraint-based methods: reverse-engineer the CPDAG based on conditional independencies inferred from data

PC algorithm: constraint-based method that uses a limited number of conditional independence tests with small conditioning sets, if the underlying CPDAG is sparse

Concretely, the PC algorithm starts with a complete graph on V. It then tests marginal independence for all pairs of nodes, and removes an edge if an independence is found. Next, for every pair of nodes that are still adjacent, it tests conditional independence given all subsets of cardinality 1 of the adjacency sets of the two nodes. The algorithm removes an edge if a conditional independence is found. The algorithm continues in this manner, each time increasing the cardinality of the conditioning set by 1, until the cardinality of the conditioning sets exceeds $\max_{v \in V} |\mathrm{adj}_{G'}(v)| - 1$, where G' is the current state of the skeleton. At this point, all required conditional independencies have been considered, and the skeleton is found.

In a second phase, the PC algorithm postprocesses the results of the conditional independence tests and learns the v-structures, as illustrated in Example 13 below. Finally, the algorithm applies a few simple orientation rules to orient some of the remaining edges, based on the fact that one may not create directed cycles or new v-structures. If the true conditional independencies are used as input for the PC algorithm, then the final output is the CPDAG of the underlying DAG G. We note that for graphs with bounded degree, that is, a bound on $\max_{v \in V} |\mathrm{adj}_G(v)|$, the PC algorithm has a running time that is polynomial in the number of variables. The running time depends exponentially on the degree.

Example 13. Suppose we have three random variables X_u, X_v, X_w with $X_u \perp\!\!\!\perp X_w$ as the sole conditional independence. The PC algorithm starts with a complete undirected graph on $\{u, v, w\}$. When testing marginal independencies, it removes the edge $u - w$ after finding $X_u \perp\!\!\!\perp X_w$. No other conditional independencies are found, so that $u - v - w$ is the final skeleton. Next, the algorithm detects that v is a collider on the path (u, v, w), because otherwise $u \not\perp\!\!\!\perp_G w$ is violated. Hence, the DAG (and corresponding CPDAG) is $u \to v \leftarrow w$.

The same skeleton is found if the only conditional independence is $X_u \perp\!\!\!\perp X_w \mid X_v$. Then, however, v must be a noncollider on the path (u, v, w) to ensure that $u \not\perp_G w \mid v$ holds. This leaves three possible DAGs, namely, $u \to v \to w$, $u \leftarrow v \leftarrow w$ or $u \leftarrow v \to w$, which form a Markov equivalence class represented by the CPDAG $u - v - w$.

In practice, conditional independencies need to be tested based on data. Standard tests are available for multivariate Gaussian and multinomial data. Usually, all tests are performed at the same significance level α, and an edge is removed if the null hypothesis of (conditional) independence is not rejected. Here, α does not control an overall type I error. Rather, it is a tuning parameter that typically gives sparser graphs for smaller values.

High-dimensional consistency of the PC algorithm for certain sparse Gaussian DAGs is shown in Kalisch & Bühlmann (2007). Harris & Drton (2013) generalize this result to Gaussian copulas. Colombo & Maathuis (2014) observe that the output of the PC algorithm can depend strongly on the variable ordering. They provide order-independent versions of the algorithm that are again

consistent in high-dimensional settings. Consistency of the PC algorithm rests on the assumption that the underlying DAG is a perfect map. This assumption is studied in depth by Uhler et al. (2013), who show that it can be restrictive.

4.4. Hybrid Algorithms

Hybrid algorithms combine ideas from constraint-based and score-based methods, by employing a greedy search over a restricted space, often determined using conditional independence tests. An example is the max-min hill climbing (MMHC) algorithm (Tsamardinos et al. 2006). Hybrid algorithms scale well with respect to the number of variables and exhibit good estimation performance (Tsamardinos et al. 2006, Nandy et al. 2016a).

The theoretical properties of hybrid algorithms are less well studied than those of purely score- or constraint-based algorithms. Nandy et al. (2016a) try to fill this gap by studying a simple hybrid algorithm: GES restricted to the search space determined by the conditional independence graph. Nandy et al. (2016a) show that this algorithm (and also MMHC) is not consistent. Indeed, even though the global optimum lies within the restricted search space (because the skeleton of the CPDAG is a subgraph of the conditional independence graph), the greedy search path may not find this global optimum without leaving the restricted search space. Nandy et al. (2016a) introduce a new hybrid algorithm, called adaptively restricted GES, which was shown to be consistent in classical and high-dimensional settings.

4.5. Structural Equation Models With Additional Restrictions

So far, we have discussed learning the Markov equivalence class of DAGs, as described by a CPDAG. Under some additional assumptions, it is possible to identify the unique DAG. To obtain intuition, we consider a simple example.

Example 14. Figure 6*a* shows a sample from the linear structural equation model (SEM): $X = \epsilon_X$, $Y = X + \epsilon_Y$, with ϵ_X and ϵ_Y i.i.d. Uniform(−0.5,0.5). The corresponding DAG is $Y \leftarrow X$, and we note that $\epsilon_Y \perp\!\!\!\perp X$. **Figure 6*b*** shows a sample from the analogous model with $X \leftarrow Y$, where $\epsilon_X \perp\!\!\!\perp Y$. The joint distributions are clearly different for the two different DAGs. (If the errors were Gaussian, however, the point clouds would be football shaped, and we would not be able to distinguish the two DAGs.)

Now suppose we are told that **Figure 6*a*** is generated from a linear SEM, and we are asked to decide whether the corresponding DAG is $X \rightarrow Y$ or $Y \rightarrow X$. If the DAG were $Y \rightarrow X$, then regressing X on Y would yield residuals that are roughly independent of Y. This is clearly not the case in **Figure 6*a***. If the DAG were $X \rightarrow Y$, then regressing Y on X would yield residuals that are roughly independent of X. This is indeed the case in **Figure 6*a***. Hence, we can learn that the DAG is $X \rightarrow Y$.

Such ideas were first used for linear SEMs with non-Gaussian noise, or equivalently, for linear non-Gaussian acyclic models (LiNGAM) (Shimizu et al. 2006). Recall from Example 8 that a linear SEM can be written as $X = (I - B)^{-1}\epsilon$, meaning that X is a linear, invertible mixture of independent errors with mixing matrix $A = (I - B)^{-1}$. In the case of non-Gaussian errors, independent component analysis can identify A up to scaling and permutation of the columns. This idea forms the basis of the original LiNGAM algorithm. Shimizu et al. (2011) give a more advanced implementation that is suitable for larger numbers of variables. The LiNGAM approach has also been extended to allow for latent variables, time series data and feedback loops (see Shimizu 2014 for an overview).

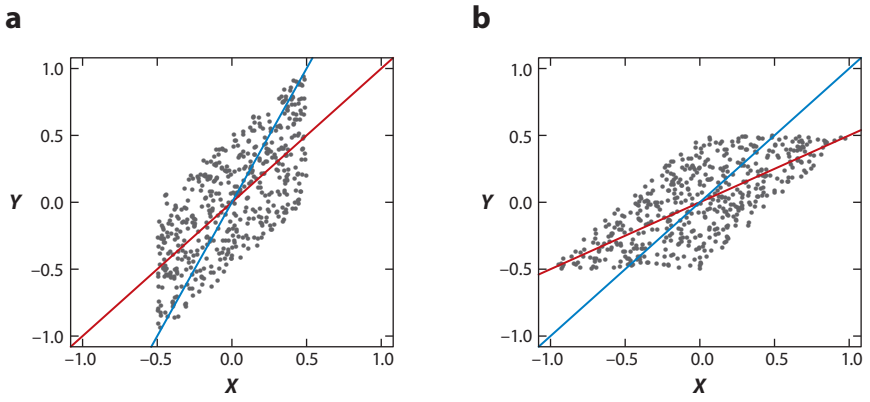

Figure 6

Simulated data from linear structural equation models with uniform errors, showing (*a*) the directed acyclic graph (DAG) $X \to Y$, and (*b*) the DAG $Y \to X$. Red lines indicate the regression of Y on X, and blue lines indicate the regression of X on Y.

Identifiability of the DAG can also be achieved by various other restrictions on SEMs with additive noise, such as nonlinear structural equations (Peters et al. 2014), additive models (Bühlmann et al. 2014), or equal error variances (Peters & Bühlmann 2014).

5. LATENT VARIABLES

The methods discussed so far rely on data being available for all relevant variables. However, many applications of graphical models involve latent, that is, unobserved variables. Sometimes, these are specific variables of interest, such as features of extinct species in phylogenetics. In other settings, there may simply be a concern that observed correlations are induced by latent variables. We will review some of the ideas proposed to address this latter issue.

5.1. Low-Rank Structure in Undirected Graphical Models

Let $X = (X_v : v \in V)$ be multivariate normal with covariance matrix Σ, and let $K = \Sigma^{-1}$. Suppose only the variables indexed by $O \subsetneq V$ are observed. Letting $H = V \setminus O$, we have that $X_O = (X_v : v \in O)$ is multivariate normal with inverse covariance matrix

$$\left(\Sigma_{O,O}\right)^{-1} = K_{O,O} - K_{O,H}\left(K_{H,H}\right)^{-1} K_{H,O}. \tag{19}$$

If X satisfies the Markov property with respect to an undirected graph $G = (V, E)$, then the matrix $K_{O,O}$ is supported over the induced subgraph G_O, i.e., the (v, w) entry of $K_{O,O}$ is nonzero only if $v = w$ or $v - w \in E$. In contrast, the matrix $K_{O,H}(K_{H,H})^{-1}K_{H,O}$ will typically be dense, as its (v, w) entry is generically nonzero whenever the graph G contains a path $v - h_1 - \cdots - h_k - w$ with $h_1, \ldots, h_k \in H$. However, $K_{O,H}(K_{H,H})^{-1}K_{H,O}$ has rank at most $|H|$. Therefore, if G_O is sparse and the number of latent variables is small, then the inverse covariance matrix of X_O is the sum of a sparse and a low-rank matrix.

Example 15. Taking up Example 5, let $(X_1, \ldots, X_5)$ be multivariate normal with the graph from **Figure 3** as a conditional independence graph. Let $K = (\kappa_{vw})$ be the inverse covariance matrix. If

$O = \{1, 2, 4, 5\}$ and $H = \{3\}$, then X_O has inverse covariance matrix

$$\begin{pmatrix} \kappa_{11} & \kappa_{12} & 0 & 0 \\ \kappa_{12} & \kappa_{22} & 0 & 0 \\ 0 & 0 & \kappa_{44} & \kappa_{45} \\ 0 & 0 & \kappa_{45} & \kappa_{55} \end{pmatrix} - \frac{1}{\kappa_{33}} \begin{pmatrix} \kappa_{13} \\ \kappa_{23} \\ \kappa_{34} \\ 0 \end{pmatrix} \begin{pmatrix} \kappa_{13} \\ \kappa_{23} \\ \kappa_{34} \\ 0 \end{pmatrix}^T.$$

Here, the first matrix is sparse and supported over the subgraph with node 3 removed, and the second matrix has rank $|H| = 1$. In this example, the low-rank matrix has two zero entries corresponding to the pairs (X_1, X_5) and (X_2, X_5). This can be seen from the global Markov property because X_5 can be separated from (X_1, X_2) by the observed variable X_4.

Suppose we have an i.i.d. sample of size n from the distribution of X_O, with sample covariance matrix S, and wish to learn (*a*) the edges between nodes $v, w \in O$ in the conditional independence graph of $X = (X_O, X_H)$ and (*b*) the number of latent variables $|H|$. By Equation 19, both tasks can be solved simultaneously by estimating a sparse-plus-low-rank decomposition of the inverse covariance matrix of X_O. Chandrasekaran et al. (2012) propose a penalized maximum likelihood approach in which one solves

$$\min_{K^{\mathrm{sp}}, K^{\mathrm{lr}}} \Big\{ -\log\det(K^{\mathrm{sp}} - K^{\mathrm{lr}}) + \mathrm{tr}\big[S(K^{\mathrm{sp}} - K^{\mathrm{lr}})\big] + \lambda\big[\gamma \|K^{\mathrm{sp}}\|_1 + \mathrm{tr}(K^{\mathrm{lr}})\big] \Big\}, \tag{20}$$

subject to $K^{\mathrm{sp}} - K^{\mathrm{lr}}$ being positive definite and K^{lr} being positive semidefinite. Here, K^{sp} and K^{lr} stand for the sparse and low-rank components of K and have corresponding ℓ_1- and trace/nuclear norm penalties. For tuning parameters $\lambda, \gamma \geq 0$, this optimization problem is convex. Let $(\hat{K}^{\mathrm{sp}}, \hat{K}^{\mathrm{lr}})$ be a minimizer. Chandrasekaran et al. (2012) show that under identifiability conditions the sparsity pattern of $\hat{K}^{\mathrm{sp}}$ and the rank of $\hat{K}^{\mathrm{lr}}$ consistently estimate the subgraph G_O and the number of hidden variables. The theory covers settings in which $|O|$ may roughly be as large as n. Possible modifications of the procedure are proposed in discussion pieces published along with Chandrasekaran et al. (2012). Larger instances of Equation 20 can be solved using an ADMM (alternating direction method of multipliers) algorithm (Ma et al. 2013).

The Gaussian example we treated is only one very special case of graphical modeling with latent variables. Indeed, many mixture and latent factor models can be thought of as graphical models with latent variables. (The states of discrete latent variables index different mixture components.) Nevertheless, the example illustrates a general phenomenon: latent variables induce low-rank structure in tensors of moments. This fact is exploited also in methods of Anandkumar et al. (2014) and Chaganty & Liang (2014).

5.2. Latent Variables in Directed Graphical Models

If a DAG model has latent variables, then the marginal distribution of the observed variables can generally not be represented by a DAG. Moreover, if the marginal distribution can be represented by a DAG, this DAG may have no causal interpretation.

Example 16. Let the DAG G: $1 \to 2 \leftarrow 3 \to 4 \leftarrow 5$ be a perfect map of the distribution of $(X_1, \ldots, X_5)$, and suppose that X_3 is latent. There is no DAG on $\{1, 2, 4, 5\}$ that encodes exactly the same d-separation relations among $\{1, 2, 4, 5\}$ as G. Hence, there does not exist a perfect map of the marginal distribution of (X_1, X_2, X_4, X_5).

Example 17. Let the DAG G: $1 \leftarrow 2 \rightarrow 3 \leftarrow 4 \rightarrow 5$ be a perfect map of the distribution of $(X_1, \ldots, X_5)$, and suppose that X_2 and X_4 are latent. The only conditional independence among the observed variables is $X_1 \perp\!\!\!\perp X_5$, which is encoded by the DAG G': $1 \rightarrow 3 \leftarrow 5$. Indeed, G' is a perfect map of the distribution of $X = (X_1, X_3, X_5)$ and would be found when applying consistent methods such as the PC algorithm to a large sample of (X_1, X_3, X_5). However, G' does not reflect the causal interpretation of G. For example, G' suggests X_1 as a cause of X_3, but there is no directed path from X_1 to X_3 in G.

Mixed graphs: represent dependencies from DAGs with latent variables and can be learned using the FCI algorithm

Mixed graphs provide a useful approach to address these problems without explicit modeling of latent variables (e.g., Spirtes et al. 2000, Pearl 2009, Wermuth 2011). The nodes of these graphs index the observed variables only. The edges, however, may be of two types, directed and bidirected. This added flexibility allows one to represent the more complicated dependence structures arising from a DAG with latent variables. A straightforward generalization of d-separation determines conditional independencies in mixed graph models. For instance, the mixed graph $1 \rightarrow 2 \longleftrightarrow 4 \leftarrow 5$ is a perfect map for the distribution in Example 16. To facilitate constraint-based structure learning in settings with latent variables, Richardson & Spirtes (2002) introduce a class of mixed graphs known as maximal ancestral graphs (MAGs). (The work also considers selection bias, which we ignore here.) Every DAG $G = (V, E)$ with $V = O \cup H$, where O and H index the observed and latent variables, respectively, can be transformed into a unique MAG on O such that conditional independencies among the observed variables are preserved. The MAG also encodes ancestral relationships in the underlying DAG G, as follows. If the MAG has edge $v \rightarrow w$, then $v \in \mathrm{an}_G(w)$ but $w \notin \mathrm{an}_G(v)$. Similarly, $v \leftrightarrow w$ encodes $v \notin \mathrm{an}_G(w)$ and $w \notin \mathrm{an}_G(v)$. In general, several MAGs may describe the same set of conditional independencies. The resulting Markov equivalence class of MAGs can be described by a partial ancestral graph (PAG) (Ali et al. 2009).

PAGs can be learned by a generalization of the PC algorithm called the fast causal inference (FCI) algorithm (Spirtes et al. 1999, Richardson & Spirtes 2002). As noted in Section 4.3, the PC algorithm is of polynomial time for graphs of bounded degree, by exploiting the fact that the edge $v - w$ is absent in the skeleton of the DAG G if and only if $X_v \perp\!\!\!\perp X_w \mid X_{\mathrm{pa}_G(v)}$ or $X_v \perp\!\!\!\perp X_w \mid X_{\mathrm{pa}_G(w)}$. In the presence of latent variables, this fact no longer holds. The FCI algorithm therefore performs additional tests, and the number of such tests can be exponential in the number of nodes, even for sparse graphs. The FCI algorithm also uses more complicated orientation rules, which were extended and proved to be complete by Zhang (2008). Colombo et al. (2012) and Claassen et al. (2013) introduce fast modifications of the FCI algorithm that are of polynomial time for graphs of bounded degree.

Although MAGs can represent all conditional independencies in the marginal distribution of the observed variables, there can be equality and inequality constraints that MAGs cannot represent. An example is the so-called Verma constraint (Verma & Pearl 1990) (see also example 3.3.14 in Drton et al. 2009). To represent constraints beyond conditional independencies, there is current work on new classes of graphs (Shpitser et al. 2012, 2014; Evans 2016).

6. HETEROGENEOUS DATA

Heterogeneous data that do not form an i.i.d. sample from a single population are encountered, for instance, in gene expression studies involving different organisms or experimental conditions, or in the comparative analysis of brain networks for patients with different neurological disorders. In such settings, it is of interest to learn the structure of graphical models for subpopulations. More generally, a graph may depend on covariates.

Guo et al. (2011) propose an extension of the graphical lasso from Equation 11 to estimate undirected conditional independence graphs of several related Gaussian populations. The authors sum up the log-likelihood functions for m populations with inverse covariance matrices $K_1, \ldots, K_m \in \mathbb{R}^{V \times V}$ and then add a sparsity-inducing penalty. To share common structure, the inverse covariances are reparametrized as $K_{i,vw} = \theta_{vw}\gamma_{i,vw}$. The penalty then adds the (vector) ℓ_1 norms of the matrix (θ_{vw}) and the m matrices $(\gamma_{i,vw})$. Sparsity in an estimate of (θ_{vw}) results in edges being simultaneously absent from all m graph estimates, each of which may have further edges absent through zero estimates of $\gamma_{i,vw}$. A downside of this method is that its optimization problem is not convex. Danaher et al. (2014) propose instead the use of group or fused lasso penalties specified in terms of the inverse covariances. The group lasso penalty takes the form $\sum_{v \neq w} \sqrt{K_{1,vw}^2 + \cdots + K_{m,vw}^2}$ and leads to edges being simultaneously absent from all m graph estimates; a version of this approach that separates positive from negative signals was also proposed by Chiquet et al. (2011). The fused lasso penalty $\sum_{1 \leq i < j \leq m} \sum_{v,w} |K_{i,vw} - K_{j,vw}|$ yields pairwise similar edge patterns in the different populations. Yang et al. (2015b) consider generalizations of this fused lasso penalty and characterize when the resulting optimization problem can be decomposed into smaller problems arising from block-diagonal inverse covariance matrices. Saegusa & Shojaie (2016) treat settings in which some populations may be more closely related than others. Finally, Zhao et al. (2014) show how the difference between two conditional independence graphs can be estimated without estimating the two graphs.

Heterogeneous data also arise from time-course observations, for which we may wish to learn time-varying structure. Zhou et al. (2010) compute glasso estimates at different time points, taking as input a weighted sample covariance matrix that is a kernel estimate of the covariance matrix at time t. Alternatively, Kolar et al. (2010) and Kolar & Xing (2012) use fused lasso penalties. Matrix/tensor-normal models, in which one of the dimensions could be time, constitute another approach; the reader is directed to He et al. (2014) and references therein.

Similar issues arise for directed graphical models. In particular, so-called dynamic Bayesian networks have long been used to model temporal dependencies, and there is a natural connection to vector-autoregressive processes in the time series literature (e.g., Shojaie & Michailidis 2010). A detailed discussion is beyond the scope of this review.

Finally, we note that heterogeneity from different experimental conditions can also be exploited in structure learning for directed graphical models (Danks et al. 2009, Hauser & Bühlmann 2012, Hyttinen et al. 2013, Triantafillou & Tsamardinos 2015). These papers generally assume i.i.d. observations from various known experimental conditions, and some of them allow cycles and/or latent variables. Peters et al. (2016) study a scenario where data may come from different experimental conditions, but the type of these conditions is unknown. They provide a method built on the idea that a variable can be predicted from its causes equally well across different experimental conditions.

7. DISCUSSION

Stimulated to a large extent by applications in gene expression analysis, the field of structure learning in graphical modeling has undergone rapid development, with much of the new work focusing on high-dimensional problems. In this review, we have treated some of the main ideas behind these developments, including ℓ_1-regularization techniques, greedy search approaches, and methods based on conditional independence tests. Extensions to cope with latent variables have long been of interest and continue to be addressed in new ways. More recently, challenges arising in connection with heterogeneous and dependent data have inspired methods that generalize

those for i.i.d. data. Similarly, methods for continuous but non-Gaussian data are an active area of research.

In a different vein, it is of interest to provide an uncertainty assessment for estimates of graph structure. Bayesian approaches naturally include an uncertainty assessment but frequentist techniques that are able to cope with high-dimensional data are also being developed (Janková & van de Geer 2015, Ren et al. 2015, Xia et al. 2015).

Finally, our treatment of directed graphical models considered acyclic graphs, in which no feedback loops exist in the cause-effect relationships that the model captures. Effective methods for structure learning exist for the acyclic case, but coping with feedback loops is a far more difficult problem. Although certain forms of feedback can be represented in the paradigm of linear structural equation modeling (Spirtes et al. 2000, Mooij & Heskes 2013), and conditional independence can then be exploited in structure learning (Richardson 1996), these models can generally not be described using solely conditional independence (see example 3.6 and appendix A in Drton 2009). New ideas are still needed to effectively learn cyclic cause-effect relationships from possibly high-dimensional observational data.

FUTURE ISSUES

1. Methods that account for the effects of latent variables need to be developed further.
2. Emerging techniques for formal statistical inference for high-dimensional graphical models should see continued development.
3. New methods for cyclic directed graphical models are needed to effectively learn cause-effect networks with feedback loops.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Ali RA, Richardson TS, Spirtes P. 2009. Markov equivalence for ancestral graphs. *Ann. Stat.* 37:2808–37

Anandkumar A, Ge R, Hsu D, Kakade SM, Telgarsky M. 2014. Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.* 15:2773–832

Anandkumar A, Valluvan R. 2013. Learning loopy graphical models with latent variables: efficient methods and guarantees. *Ann. Stat.* 41:401–35

Andersson SA, Madigan D, Perlman MD. 1997. A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Stat.* 25:505–41

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

Banerjee S, Ghosal S. 2015. Bayesian structure learning in graphical models. *J. Multivar. Anal.* 136:147–62

Barber RF, Drton M. 2015. High-dimensional Ising model selection with Bayesian information criteria. *Electron. J. Stat.* 9:567–607

Besag J. 1975. Statistical analysis of non-lattice data. *J. R. Stat. Soc. D.* 24:179–95

Bilodeau M. 2014. Graphical lassos for meta-elliptical distributions. *Can. J. Stat.* 42:185–203

Bollen KA. 1989. *Structural Equations with Latent Variables*. New York: Wiley

Bresler G. 2015. Efficiently learning Ising models on arbitrary graphs. *Proc. 47th Annu. ACM Symp. Theory Comp. (STOC 2015)*, pp. 771–82. New York, NY: ACM

Bühlmann P, Peters J, Ernest J. 2014. CAM: causal additive models, high-dimensional order search and penalized regression. *Ann. Stat.* 42:2526–56

Cai TT, Liu W, Luo X. 2011. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Stat. Assoc.* 106:594–607

Cai TT, Liu W, Zhou HH. 2016. Estimating sparse precision matrix: optimal rates of convergence and adaptive estimation. *Ann. Stat.* 44:455–88

Chaganty AT, Liang P. 2014. Estimating latent-variable graphical models using moments and likelihoods. *Proc. 31st Int. Conf. Mach. Learn. (ICML 2014)*, pp. 1872–80. New York: ACM

Chandrasekaran V, Parrilo PA, Willsky AS. 2012. Latent variable graphical model selection via convex optimization. *Ann. Stat.* 40:1935–67

Chen S, Witten DM, Shojaie A. 2015. Selection and estimation for mixed graphical models. *Biometrika* 102:47–64

Chickering DM. 2002. Optimal structure identification with greedy search. *J. Mach. Learn. Res.* 3:507–54

Chickering DM, Meek C. 2015. Selective greedy equivalence search: finding optimal Bayesian networks using a polynomial number of score evaluations. *Proc. 31st Conf. Uncertain. Artif. Intell. (UAI 2015)*, pp. 211–19. Corvallis, OR: AUAI Press

Chiquet J, Grandvalet Y, Ambroise C. 2011. Inferring multiple graphical structures. *Stat. Comput.* 21:537–53

Chow CK, Liu C. 1968. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inform. Theory* IT-14:462–67

Claassen T, Mooij JM, Heskes T. 2013. Learning sparse causal models is not NP-hard. *Proc. 29th Conf. Uncertain. Artif. Intell. (UAI 2013)*, pp. 172–81. Corvallis, OR: AUAI Press

Colombo D, Maathuis MH. 2014. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* 15:3741–82

Colombo D, Maathuis MH, Kalisch M, Richardson TS. 2012. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Stat.* 40:294–321

Cussens J, Bartlett M. 2013. Advances in Bayesian network learning using integer programming. *Proc. 29th Conf. Uncertain. Artif. Intell. (UAI 2013)*, pp. 182–91. Corvallis, OR: AUAI Press

Danaher P, Wang P, Witten DM. 2014. The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Stat. Soc. B* 76:373–97

Danks D, Glymour C, Tillman RE. 2009. Integrating locally learned causal structures with overlapping variables. *Proc. Adv. Neural Inf. Proc. Syst. 21 (NIPS 2008)*, pp. 1665–72. Red Hook, NY: Curran

Dasarathy G, Singh A, Balcan MF, Park JH. 2016. Active learning algorithms for graphical model selection. arXiv:1602.00354

de Campos C, Zeng Z, Ji Q. 2009. Structure learning of Bayesian networks using constraints. *Proc. 26th Int. Conf. Mach. Learn. (ICML 2009)*, pp. 113–20. New York: ACM

Defazio A, Caetano TS. 2012. A convex formulation for learning scale-free networks via submodular relaxation. *Proc. Adv. Neural Inf. Proc. Syst. 25 (NIPS 2012)*, pp. 1250–58. Red Hook, NY: Curran

Drton M. 2009. Likelihood ratio tests and singularities. *Ann. Stat.* 37:979–1012

Drton M, Perlman MD. 2007. Multiple testing and error control in Gaussian graphical model selection. *Stat. Sci.* 22:430–49

Drton M, Sturmfels B, Sullivant S. 2009. *Lectures on Algebraic Statistics*. Oberwolfach Semin. 39. Basel, Switz: Birkhäuser Verlag

Edwards D, de Abreu GC, Labouriau R. 2010. Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. *BMC Bioinform.* 11:1–13

Eichler M. 2012. Graphical modelling of multivariate time series. *Probab. Theory Related Fields*. 153:233–68

Evans RJ. 2016. Graphs for margins of Bayesian networks. *Scand. J. Stat.* doi: 10.1111/sjos.12194

Fan J, Liu H, Ning Y, Zou H. 2016. High dimensional semiparametric latent graphical model for mixed data. *J. R. Stat. Soc. B.* doi: 10.1111/rssb.12168

Finegold M, Drton M. 2011. Robust graphical modeling of gene networks using classical and alternative *t*-distributions. *Ann. Appl. Stat.* 5:1057–80

Finegold M, Drton M. 2014. Robust Bayesian graphical modeling using Dirichlet *t*-distributions. *Bayesian Anal.* 9:521–50

Forbes PGM, Lauritzen S. 2015. Linear estimating equations for exponential families with application to Gaussian linear concentration models. *Linear Algebra Appl.* 473:261–83

Foygel R, Drton M. 2010. Extended Bayesian information criteria for Gaussian graphical models. *Proc. Adv. Neural Inf. Process. Syst. 23 (NIPS 2010)*, pp. 2020–28. Red Hook, NY: Curran

Friedman J, Hastie T, Tibshirani R. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9:432–41

Friedman N. 2004. Inferring cellular networks using probabilistic graphical models. *Science* 303:799–805

Friedman N, Ninio M, Pe'er I, Pupko T. 2002. A structural EM algorithm for phylogenetic inference. *J. Comput. Biol.* 9:331–53

Frydenberg M. 1990. The chain graph Markov property. *Scand. J. Stat.* 17:333–53

Gao X, Pu DQ, Wu Y, Xu H. 2012. Tuning parameter selection for penalized likelihood estimation of Gaussian graphical model. *Stat. Sinica* 22:1123–46

Geiger D, Meek C, Sturmfels B. 2006. On the toric algebra of graphical models. *Ann. Stat.* 34:1463–92

Goudie RJB, Mukherjee S. 2016. A Gibbs sampler for learning DAGs. *J. Mach. Learn. Res.* 17:1–39

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

Harris N, Drton M. 2013. PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.* 14:3365–83

Hauser A, Bühlmann P. 2012. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.* 13:2409–64

He S, Yin J, Li H, Wang X. 2014. Graphical model selection and estimation for high dimensional tensor data. *J. Multivar. Anal.* 128:165–85

He Y, Jia J, Yu B. 2013. Reversible MCMC on Markov equivalence classes of sparse directed acyclic graphs. *Ann. Stat.* 41:1742–79

Höfling H, Tibshirani R. 2009. Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *J. Mach. Learn. Res.* 10:883–906

Højsgaard S, Edwards D, Lauritzen S. 2012. *Graphical Models with R*. New York: Springer

Hsieh CJ, Sustik MA, Dhillon IS, Ravikumar PK, Poldrack R. 2013. BIG & QUIC: sparse inverse covariance estimation for a million variables. *Proc. Adv. Neural Inf. Proc. Syst. 26 (NIPS 2013)*, pp. 3165–73. Red Hook, NY: Curran

Hyttinen A, Hoyer P, Eberhardt F, Jarvisalo M. 2013. Discovering cyclic causal models with latent variables: a general SAT-based procedure. *Proc. 29th Conf. Uncertain. Artif. Intell. (UAI 2013)*, pp. 301–10. Corvallis, OR: AUAI Press

Hyvärinen A. 2005. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.* 6:695–709

Hyvärinen A. 2007. Some extensions of score matching. *Comput. Stat. Data Anal.* 51:2499–512

Jaakkola T, Sontag D, Globerson A, Meila M. 2010. Learning Bayesian network structure using LP relaxations. *J. Mach. Learn. Res.* 9:358–65

Jalali A, Johnson CC, Ravikumar PK. 2011. On learning discrete graphical models using greedy methods. *Proc. Adv. Neural Inf. Proc. Syst. 24 (NIPS 2011)*, pp. 1935–43. Red Hook, NY: Curran

Janková J, van de Geer S. 2015. Confidence intervals for high-dimensional inverse covariance estimation. *Electron. J. Stat.* 9:1205–29

Janofsky E. 2015. *Exponential series approaches for nonparametric graphical models*. PhD thesis, Univ. Chicago. arXiv:1506.03537 [math.ST]

Kalisch M, Bühlmann P. 2007. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* 8:613–36

Karger D, Srebro N. 2001. Learning Markov networks: maximum bounded tree-width graphs. *Proc. 12th ACM-SIAM Symp. Discret. Algorithms (SODA 2001)*, pp. 392–401. Philadelphia: SIAM

Khare K, Oh SY, Rajaratnam B. 2015. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *J. R. Stat. Soc. B* 77:803–25

Kolar M, Song L, Ahmed A, Xing EP. 2010. Estimating time-varying networks. *Ann. Appl. Stat.* 4:94–123

Kolar M, Xing EP. 2012. Estimating networks with jumps. *Electron. J. Stat.* 6:2069–106

Koller D, Friedman N. 2009. *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press

Kuipers J, Moffa G. 2016. Partition MCMC for inference on acyclic digraphs. *J. Am. Stat. Assoc.* doi: 10.1080/01621459.2015.1133426
Lauritzen SL. 1996. *Graphical Models*. Oxford Stat. Sci. Ser. Vol. 17. Oxford, UK: Clarendon
Lederer J, Müller C. 2014. Topology adaptive graph estimation in high dimensions. arXiv:1410.7279 [stat.ML]
Lin L, Drton M, Shojaie A. 2016. Estimation of high-dimensional graphical models using regularized score matching. *Electron. J. Stat.* 10:806–54
Liu H, Han F, Yuan M, Lafferty J, Wasserman L. 2012a. High-dimensional semiparametric Gaussian copula graphical models. *Ann. Stat.* 40:2293–326
Liu H, Han F, Zhang CH. 2012b. Transelliptical graphical models. *Proc. Adv. Neural Inf. Proc. Syst. 25 (NIPS 2012)*, pp. 809–17. Red Hook, NY: Curran
Liu H, Lafferty J, Wasserman L. 2009. The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* 10:2295–328
Liu H, Roeder K, Wasserman L. 2010. Stability approach to regularization selection (StARS) for high-dimensional graphical models. *Proc. Adv. Neural Inf. Process. Syst. 23 (NIPS 2010)*, pp. 1432–40. Red Hook, NY: Curran
Liu H, Xu M, Gu H, Gupta A, Lafferty J, Wasserman L. 2011. Forest density estimation. *J. Mach. Learn. Res.* 12:907–51
Liu W. 2013. Gaussian graphical model estimation with false discovery rate control. *Ann. Statist.* 41:2948–78
Liu W, Luo X. 2015. Fast and adaptive sparse precision matrix estimation in high dimensions. *J. Multivar. Anal.* 135:153–62
Ma S, Xue L, Zou H. 2013. Alternating direction methods for latent variable Gaussian graphical model selection. *Neural Comput.* 25:2172–98
Maathuis MH, Colombo D, Kalisch M, Bühlmann P. 2010. Predicting causal effects in large-scale systems from observational data. *Nat. Methods* 7:247–48
Maathuis MH, Kalisch M, Bühlmann P. 2009. Estimating high-dimensional intervention effects from observational data. *Ann. Stat.* 37:3133–64
Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, et al. 2012. Wisdom of crowds for robust gene network inference. *Nat. Methods* 9:796–804
Matúš F. 2012. On conditional independence and log-convexity. *Ann. Inst. Henri Poincaré Probab. Stat.* 48:1137–47
Mazumder R, Hastie T. 2012a. Exact covariance thresholding into connected components for large-scale graphical lasso. *J. Mach. Learn. Res.* 13:781–94
Mazumder R, Hastie T. 2012b. The graphical lasso: new insights and alternatives. *Electron. J. Stat.* 6:2125–49
Meinshausen N. 2008. A note on the Lasso for Gaussian graphical model selection. *Stat. Probab. Lett.* 78:880–84
Meinshausen N, Bühlmann P. 2006. High-dimensional graphs and variable selection with the lasso. *Ann. Stat.* 34:1436–62
Meinshausen N, Bühlmann P. 2010. Stability selection. *J. R. Stat. Soc. B* 72:417–73
Mitra R, Müller P, Ji Y. 2016. Bayesian graphical models for differential pathways. *Bayesian Anal.* 11:99–124
Mooij JM, Heskes T. 2013. Cyclic causal discovery from continuous equilibrium data. In *Proc. 29th Conf. Uncertain. Artif. Intell. (UAI 2013)*, pp. 431–39. Corvallis, OR: AUAI Press
Nandy P, Hauser A, Maathuis MH. 2016a. High-dimensional consistency in score-based and hybrid structure learning. arXiv:1507.02608 [math.ST]
Nandy P, Maathuis MH, Richardson TS. 2016b. Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *Ann. Stat.* In press
Neapolitan RE. 2004. *Learning Bayesian Networks*. Upper Saddle River, NJ: Pearson Prentice Hall
Nyman H, Pensar J, Koski T, Corander J. 2014. Stratified graphical models—context-specific independence in graphical models. *Bayesian Anal.* 9:883–908
Parviainen P, Koivisto M. 2009. Exact structure discovery in Bayesian networks with less space. *Proc. 25th Conf. Uncertain. Artif. Intell. (UAI 2009)*, pp. 436–43. Arlington, VA: AUAI Press
Pearl J. 2009. *Causality: Models, Reasoning and Inference*. Cambridge, UK: Cambridge Univ. Press. 2nd ed.
Perković E, Textor J, Kalisch M, Maathuis MH. 2015. A complete adjustment criterion. *Proc. 31st Conf. Uncertain. Artif. Intell. (UAI 2015)*, pp. 682–91. Corvallis, OR: AUAI Press

Peters J, Bühlmann P. 2014. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika* 101:219–28
Peters J, Bühlmann P, Meinshausen N. 2016. Causal inference using invariant prediction: identification and confidence intervals. *J. R. Stat. Soc. B* 78:1–42
Peters J, Mooij JM, Janzing D, Schölkopf B. 2014. Causal discovery with continuous additive noise models. *J. Mach. Learn. Res.* 15:2009–53
Peterson C, Stingo FC, Vannucci M. 2015. Bayesian inference of multiple Gaussian graphical models. *J. Am. Stat. Assoc.* 110:159–74
Ravikumar P, Wainwright MJ, Lafferty JD. 2010. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Stat.* 38:1287–319
Ravikumar P, Wainwright MJ, Raskutti G, Yu B. 2011. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* 5:935–80
Ray A, Sanghavi S, Shakkottai S. 2015. Improved greedy algorithms for learning graphical models. *IEEE Trans. Inf. Theory* 61:3457–68
Ren Z, Sun T, Zhang CH, Zhou HH. 2015. Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Stat.* 43:991–1026
Richardson TS. 1996. A discovery algorithm for directed cyclic graphs. *Proc. 12th Conf. Uncertain. Artif. Intell. (UAI 1996)*, pp. 454–61. San Francisco: Morgan Kaufmann
Richardson TS, Spirtes P. 2002. Ancestral graph Markov models. *Ann. Stat.* 30:962–1030
Roverato A. 2005. A unified approach to the characterization of equivalence classes of DAGs, chain graphs with no flags and chain graphs. *Scand. J. Stat.* 32:295–312
Saegusa T, Shojaie A. 2016. Joint estimation of precision matrices in heterogeneous populations. *Electron. J. Stat.* 10:1341–92
Shah RD, Samworth RJ. 2013. Variable selection with error control: another look at stability selection. *J. R. Stat. Soc. B* 75:55–80
Shimizu S. 2014. LiNGAM: Non-Gaussian methods for estimating causal structures. *Behaviormetrika* 41:65–98
Shimizu S, Hoyer PO, Hyvärinen A, Kerminen A. 2006. A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* 7:2003–30
Shimizu S, Inazumi T, Sogawa Y, Hyvärinen A, Kawahara Y, et al. 2011. DirectLiNGAM: a direct method for learning a linear non-Gaussian structural equation model. *J. Mach. Learn. Res.* 12:1225–48
Shojaie A, Michailidis G. 2010. Discovering graphical Granger causality using the truncating lasso penalty. *Bioinformatics* 26:i517–23
Shpitser I, Evans R, Richardson T, Robins J. 2014. Introduction to nested Markov models. *Behaviormetrika* 41:3–39
Shpitser I, Richardson TS, Robins JM, Evans R. 2012. Parameter and structure learning in nested Markov models. arXiv:1207.5058 [stat.ML]
Silander T, Myllymäki P. 2006. A simple approach for finding the globally optimal Bayesian network structure. *Proc. 22nd Conf. Uncertain. Artif. Intell. (UAI 2006)*, pp. 445–52. Arlington, VA: AUAI Press
Silva R, Ghahramani Z. 2009. The hidden life of latent variables: Bayesian learning with mixed graph models. *J. Mach. Learn. Res.* 10:1187–238
Spirtes P, Glymour C, Scheines R. 2000. *Causation, Prediction, and Search*. Cambridge, MA: MIT Press. 2nd ed.
Spirtes P, Meek C, Richardson TS. 1999. An algorithm for causal inference in the presence of latent variables and selection bias. In *Computation, Causation, and Discovery*, pp. 211–52. Menlo Park, CA: AAAI Press
Statnikov A, Ma S, Henaff M, Lytkin N, Efstathiadis E, et al. 2015. Ultra-scalable and efficient methods for hybrid observational and experimental local causal pathway discovery. *J. Mach. Learn. Res.* 16:3219–67
Studený M. 2005. *Probabilistic conditional independence structures*. London: Springer
Studený M, Haws D. 2014. Learning Bayesian network structure: towards the essential graph by integer linear programming tools. *Int. J. Approx. Reason.* 55:1043–71
Sullivant S, Gross E. 2014. The maximum likelihood threshold of a graph. arXiv:1404.6989 [math.CO]
Sun S, Kolar M, Xu J. 2015. Learning structured densities via infinite dimensional exponential families. *Proc. Adv. Neural Inf. Proc. Syst. 28 (NIPS 2015)*, pp. 2287–95. Red Hook, NY: Curran

Tan KM, London P, Mohan K, Lee SI, Fazel M, Witten D. 2014. Learning graphical models with hubs. *J. Mach. Learn. Res.* 15:3297–331
Tan VYF, Anandkumar A, Willsky AS. 2010. Learning Gaussian tree models: analysis of error exponents and extremal structures. *IEEE Trans. Signal Process.* 58:2701–14
Tan VYF, Anandkumar A, Willsky AS. 2011. Learning high-dimensional Markov forest distributions: analysis of error rates. *J. Mach. Learn. Res.* 12:1617–53
Triantafillou S, Tsamardinos I. 2015. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *J. Mach. Learn. Res.* 16:2147–205
Tsamardinos I, Brown LE, Aliferis CF. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.* 65:31–78
Uhler C, Raskutti G, Bühlmann P, Yu B. 2013. Geometry of the faithfulness assumption in causal inference. *Ann. Stat.* 41:436–63
Van de Geer S, Bühlmann P. 2013. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Ann. Stat.* 41:536–67
Vats D, Nowak R, Baraniuk R. 2014. Active learning for undirected graphical model selection. *Proc. 17th Int. Conf. Artif. Intell. Stat. (AISTATS) 2014, Reykjavik, Iceland*, pp. 958–67. Cambridge, MA: MIT Press
Verma T, Pearl J. 1988. Causal networks: Semantics and expressiveness. *Proc. 4th Conf. Uncertain. Artif. Intell. (UAI 1988)*, pp. 352–59. Amsterdam: Elsevier
Verma TS, Pearl J. 1990. Equivalence and synthesis of causal models. *Proc. 6th Conf. Uncertain. Artif. Intell. (UAI 1990)*, pp. 255–68. Amsterdam: Elsevier
Vogel D, Fried R. 2011. Elliptical graphical modelling. *Biometrika* 98:935–51
Vogel D, Tyler DE. 2014. Robust estimators for nondecomposable elliptical graphical models. *Biometrika* 101:865–82
Voorman A, Shojaie A, Witten D. 2014. Graph estimation with joint additive models. *Biometrika* 101:85–101
Wainwright MJ, Jordan MI. 2008. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* 1:1–305
Wasserman L, Kolar M, Rinaldo A. 2014. Berry-Esseen bounds for estimating undirected graphs. *Electron. J. Stat.* 8:1188–224
Wermuth N. 2011. Probability distributions with summary graph structure. *Bernoulli* 17:845–79
Witten DM, Friedman JH, Simon N. 2011. New insights and faster computations for the graphical lasso. *J. Comput. Graph. Stat.* 20:892–900
Xia Y, Cai T, Cai TT. 2015. Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika* 102:247–66
Xue L, Zou H. 2012. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Stat.* 40:2541–71
Yang E, Ravikumar P, Allen GI, Liu Z. 2015a. Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.* 16:3813–47
Yang S, Lu Z, Shen X, Wonka P, Ye J. 2015b. Fused multiple graphical lasso. *SIAM J. Optim.* 25:916–43
Yang Z, Ning Y, Liu H. 2014. On semiparametric exponential family graphical models. arXiv:1412.8697 [stat.ML]
Yuan M, Lin Y. 2007. Model selection and estimation in the Gaussian graphical model. *Biometrika* 94:19–35
Zhang J. 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.* 172:1873–96
Zhao SD, Cai TT, Li H. 2014. Direct estimation of differential networks. *Biometrika* 101:253–68
Zhao T, Liu H, Roeder K, Lafferty J, Wasserman L. 2012. The huge package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* 13:1059–62
Zhou S, Lafferty J, Wasserman L. 2010. Time varying undirected graphs. *Mach. Learn.* 80:295–319

Annual Review of Statistics and Its Application

Volume 4, 2017

Contents

p-Values: The Insight to Modern Statistical Inference
D.A.S. Fraser 1

Curriculum Guidelines for Undergraduate Programs in Data Science
Richard D. De Veaux, Mahesh Agarwal, Maia Averett, Benjamin S. Baumer, Andrew Bray, Thomas C. Bressoud, Lance Bryant, Lei Z. Cheng, Amanda Francis, Robert Gould, Albert Y. Kim, Matt Kretchmar, Qin Lu, Ann Moskol, Deborah Nolan, Roberto Pelayo, Sean Raleigh, Ricky J. Sethi, Mutiara Sondjaja, Neelesh Tiruviluamala, Paul X. Uhlig, Talitha M. Washington, Curtis L. Wesley, David White, and Ping Ye 15

Risk and Uncertainty Communication
David Spiegelhalter 31

Exposed! A Survey of Attacks on Private Data
Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman 61

The Evolution of Data Quality: Understanding the Transdisciplinary Origins of Data Quality Concepts and Approaches
Sallie Keller, Gizem Korkmaz, Mark Orr, Aaron Schroeder, and Stephanie Shipp 85

Is Most Published Research Really False?
Jeffrey T. Leek and Leah R. Jager 109

Understanding and Assessing Nutrition
Alicia L. Carriquiry 123

Hazard Rate Modeling of Step-Stress Experiments
Maria Kateri and Udo Kamps 147

Online Analysis of Medical Time Series
Roland Fried, Sermad Abbas, Matthias Borowski, and Michael Imhoff 169

Statistical Methods for Large Ensembles of Super-Resolution Stochastic Single Particle Trajectories in Cell Biology
Nathanäel Hozé and David Holcman 189

Statistical Issues in Forensic Science
Hal S. Stern 225

Bayesian Modeling and Analysis of Geostatistical Data
Alan E. Gelfand and Sudipto Banerjee 245

Modeling Through Latent Variables
Geert Verbeke and Geert Molenberghs 267

Two-Part and Related Regression Models for Longitudinal Data
V.T. Farewell, D.L. Long, B.D.M. Tom, S. Yiu, and L. Su 283

Some Recent Developments in Statistics for Spatial Point Patterns
Jesper Møller and Rasmus Waagepetersen 317

Stochastic Actor-Oriented Models for Network Dynamics
Tom A.B. Snijders 343

Structure Learning in Graphical Modeling
Mathias Drton and Marloes H. Maathuis 365

Bayesian Computing with INLA: A Review
Håvard Rue, Andrea Riebler, Sigrunn H. Sørbye, Janine B. Illian, Daniel P. Simpson, and Finn K. Lindgren 395

Global Testing and Large-Scale Multiple Testing for High-Dimensional Covariance Structures
T. Tony Cai 423

The Energy of Data
Gabór J. Székely and Maria L. Rizzo 447

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics