MODEL SELECTION AND LOCAL GEOMETRY

BY ROBIN J. EVANS

Department of Statistics, University of Oxford, evans@stats.ox.ac.uk

We consider problems in model selection caused by the geometry of models close to their points of intersection. In some cases—including common classes of causal or graphical models, as well as time series models—distinct models may nevertheless have identical tangent spaces. This has two immediate consequences: first, in order to obtain constant power to reject one model in favour of another we need local alternative hypotheses that decrease to the null at a slower rate than the usual parametric $n^{-1/2}$ (typically we will require $n^{-1/4}$ or slower); in other words, to distinguish between the models we need large effect sizes or very large sample sizes. Second, we show that under even weaker conditions on their tangent cones, models in these classes cannot be made simultaneously convex by a reparameterization.

This shows that Bayesian network models, amongst others, cannot be learned directly with a convex method similar to the graphical lasso. However, we are able to use our results to suggest methods for model selection that learn the tangent space directly, rather than the model itself. In particular, we give a generic algorithm for learning Bayesian network models.

1. Introduction. Consider a class of probabilistic models $\mathcal{M}_i$ indexed by elements of some set $i \in I$, and suppose that we have data from some distribution P; model selection is the task of deducing, from the data, which $\mathcal{M}_i$ contains P. Typically, there will be multiple such models, in which case one may appeal to parsimony or—if the model class is closed under intersection—select the smallest such model by inclusion.

There have been dramatic advancements in certain kinds of statistical model selection, including methods for working with large datasets and very high-dimensional problems (see, e.g., Bühlmann and van de Geer (2011)). However, model selection in some settings is more difficult; for example, selecting an optimal Bayesian network for discrete data is known to be a NP-complete problem (Chickering (1996)). In this paper, we consider why some model classes are so much harder to learn with than others. Taking a geometric approach, we find that some classes contain models which are distinct but—in a sense that will be made precise—are locally very similar to one another. The task of distinguishing between them using data is therefore fundamentally more difficult, both statistically and computationally.

EXAMPLE 1.1. To illustrate the main idea in simple terms, consider a model space smoothly described by a two-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \theta_2)^T \in \mathbb{R}^2$, and with four submodels of interest:

$$\mathcal{M}_\varnothing : \theta_1 = \theta_2 = 0, \qquad \mathcal{M}_1 : \theta_2 = 0,$$
$$\mathcal{M}_2 : \theta_1 = 0, \qquad \mathcal{M}_{12} : \text{unrestricted}.$$

In a setting with independent data, we would expect to have statistical power sufficient to distinguish between $\mathcal{M}_{12}$ and $\mathcal{M}_2$ (i.e., to determine whether or not $\theta_1 = 0$) provided that the
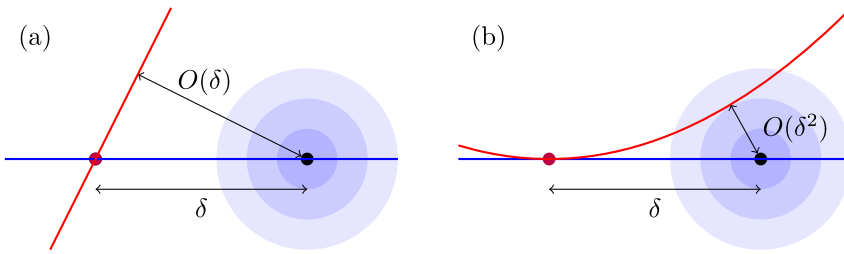
FIG. 1.   *Illustration of model selection close to points of intersection. On the left, the models $\mathcal{M}_1$ and $\mathcal{M}_2$ (in blue and red, resp.) have different tangent spaces, so the distance between them increases linearly as one moves away from the intersection. On the right, the models $\mathcal{M}_1$ (in blue) and $\mathcal{M}_2'$ (in red) have the* same *tangent space at the intersection, so they diverge only quadratically with distance from the intersection.*

magnitude of θ_1 is large compared to $n^{-1/2}$, where n is the number of independent samples available. We might also expect to be able to distinguish between $\mathcal{M}_1$ and $\mathcal{M}_2$ at the same asymptotic rate; this is the picture in Figure 1(a), in which the distance between the two models is proportional to distance from their intersection $\mathcal{M}_\varnothing = \mathcal{M}_1 \cap \mathcal{M}_2$ (the constant of proportionality being determined by the angle between the two models).

Suppose now that we define a model $\mathcal{M}_2' : \psi_1 = 0$, where $\psi_1 \equiv \theta_1^2 - \theta_2$, and have to select between $\mathcal{M}_\varnothing, \mathcal{M}_1, \mathcal{M}_2', \mathcal{M}_{12}$ (note that we still have $\mathcal{M}_\varnothing = \mathcal{M}_1 \cap \mathcal{M}_2'$), as illustrated in Figure 1(b). Superficially, the task of choosing between these four models seems no different to our first scenario, but in fact the models $\mathcal{M}_1$ and $\mathcal{M}_2'$ are locally linearly identical at the point of intersection $\psi_1 = \theta_2 = 0$; that is, the tangent spaces of the two models at this point are the same, so up to a linear approximation they are indistinguishable.

Models that overlap linearly in the manner above lead to two major consequences relating to statistical power and computational efficiency. First, as illustrated in Figure 1(b), if $\mathcal{M}_1$ is correct the distance[1] between the true parameter value $(\theta_1, 0)$ and the closest point on $\mathcal{M}_2'$ grows quadratically rather than linearly in θ_1. Hence, while $|\theta_1| = \Omega(n^{-1/2})$ is sufficient to gain power against $\mathcal{M}_\varnothing$, one needs $|\theta_1| = \Omega(n^{-1/4})$ to ensure power[2] against $\mathcal{M}_2'$. This is potentially a very stringent condition indeed; if the effect size is halved, then we will need 16 times the sample size to maintain power against the alternative model.

Second, if two models have the same tangent space then we cannot choose a parameterization under which both models are convex sets. Note that in Figure 1(a) all four models are convex, but in Figure 1(b) the model $\mathcal{M}_2'$ is not. If we reparameterize to make $\mathcal{M}_2'$ convex, then $\mathcal{M}_1$ will not be.[3] This prevents penalized methods such as the lasso being used in a computationally efficient way.

EXAMPLE 1.2 (Directed Gaussian graphical models).   A common class of models in which the phenomenon described above occurs is Gaussian Bayesian networks. Consider the two graphs shown in Figure 2, each representing certain multivariate Gaussian distributions over variables X, Y, Z with joint correlation matrix Π. The graph in Figure 2(a) corresponds to the marginal independence model $X \perp\!\!\!\perp Y$, so that there is a zero in the corresponding entry

---

[1]Typically, this would be approximated by the Mahalanobis distance using the Fisher information. For regular statistical models, this is locally equivalent to the Hellinger distance or the square-root of the KL-divergence.

[2]Recall that for positive functions f, g we have $f(x) = \Omega(g(x))$ if and only if $g(x) = O(f(x))$, and $f(x) = \omega(g(x))$ if and only if $g(x) = o(f(x))$.

[3]By 'reparameterize', we mean under a twice differentiable bijection with an invertible Jacobian. So we do not allow the map $(\theta_1, \theta_2) \mapsto (\psi_1, \theta_2)$ used in the earlier example.

FIG. 2. *Two Bayesian networks in which, for Gaussian random variables, the tangent spaces of the models are identical at some points of intersection.*

in Π: $\rho_{xy} = 0$. Figure 2(b), on the other hand, corresponds to the conditional independence model $X \perp\!\!\!\perp Y \mid Z$, that is, to a zero in the X, Y entry of Π^{-1}, or equivalently to $\rho_{xy} - \rho_{xz}\rho_{zy} = 0$.

The two models intersect along two further submodels: for example, if $\rho_{xz} = 0$ (so that $X \perp\!\!\!\perp Z$), then $\rho_{xy} = 0$ if and only if $\rho_{xy} - \rho_{xz}\rho_{zy} = 0$. The same thing happens if $\rho_{yz} = 0$ (i.e., $Y \perp\!\!\!\perp Z$). When we are at the intersection between *all* these submodels—so $\rho_{xy} = \rho_{xz} = \rho_{yz} = 0$ and all variables are jointly independent—we find that the tangent spaces of the two original models are the same, giving rise to the phenomenon described above. Indeed, we will see that this arises whenever two models intersect along two or more such—suitably distinct—further submodels (Theorem 3.1).

1.1. *Background and prior work.* Crudely speaking, there are two flavours of statistical model, and consequently two main reasons for wishing to select one. The first, called *substantive* or *explanatory*, emphasizes the use of models to explain underlying phenomena, and such models are sometimes viewed as approximations to an unknown scientific 'ground truth' (Cox (1990), Skrondal and Rabe-Hesketh (2004)). The second kind, referred to as *empirical* or *predictive*, is mainly concerned with predicting outcomes from future observations, generally assuming that such observations will arise from the same population as previous data (Breiman (2001)). A discussion of these two camps, together with some finer distinctions can be found in Cox (1990).

Our focus will primarily be on substantive models, in which case different models may lead to rather different practical conclusions, even if the probability distributions associated with them are 'close' in the sense observed above. The case of causal models such as the graphs in Figure 2 is particularly stark: the reversal of an arrow will significantly affect our understanding of how a system will behave under an intervention. For interpolative prediction—that is, with new data from the same population as the data used to learn the model—such concerns are generally lessened; if two probability distributions are similar, then they should give similar conclusions. However, the computational concerns we raise will affect model selection performed for any reason whatsoever.

For Bayesian network (BN) models specifically, there has been a great deal of work dealing with the problem of accurate learning from data. Chickering (1996) showed that the problem of finding the BN which maximizes a penalized likelihood criterion is NP-complete in the case of discrete data with several common penalties. Uhler et al. (2013) give geometric proofs that directed Gaussian graphical models are, in a global sense, very hard to learn using sequential independence tests; this is because the volume of 'unfaithful' points that will mislead at least one hypothesis test for a given sample size is very large, and in settings where the number of parameters is larger than the number of observations will overwhelm the model. Our approach is considerably simpler and is applicable to arbitrary model classes and model selection procedures, but cannot make global statements about the model.

Shojaie and Michailidis (2010) provide a penalized method for learning sparse high-dimensional graphs, but they assume a known topological ordering for the variables in the

graph, that is, the direction of each possible edge is known. Ni, Stingo and Baladandayutha-pani (2015) develop a Bayesian approach that is similar in spirit, and apply it to gene regulatory networks. Fu and Zhou (2013), Gu, Fu and Zhou (2019) and Aragam and Zhou (2015) all use penalization to learn BNs without a pre-specified topological order, in the former paper even allowing for interventional data; however in each case the resulting optimization problem is nonconvex. Other approaches based on assumptions such as non-Gaussianity or nonlinearity are also available (Shimizu et al. (2006), Bühlmann, Peters and Ernest (2014)).

1.2. *Contribution.* In this paper, we develop the notion of using local geometry as a heuristic for how closely related two models are, and how rich a class of models is. In several classes of models for which model selection is known to be difficult, we find that they contain distinct models that are linearly equivalent at certain points in the parameter space. This makes it *statistically* difficult to tell which model is correct.

Under even weaker conditions, we find that distinct models may have directions that can be approximately obtained in both models, but not in their intersection. This means the models cannot be simultaneously convex, and prevents efficient algorithms from being used to learn which model is correct; this makes it *computationally* hard to pick the best model. To our knowledge, these are both completely new contributions to the literature.

The remainder of the paper is organized as follows. In Section 2, we formalize the intuition given above by carefully defining local similarity between models, and then proving results relating to local asymptotic power and convex parameterizations. In Section 3, we give sufficient conditions for this situation to occur, including the intersection of two models along multiple distinct submodels. In Section 4, we apply these results to show that the lasso cannot be used directly to learn Bayesian networks. Section 5 provides further examples of how the result can be applied, while Section 6 considers related phenomena such as double robustness. Section 7 suggests methods to exploit model classes in which the tangent spaces are distinct but in which we cannot make the models simultaneously convex, and Section 8 contains a discussion.

2. Models. Consider a class of finite-dimensional probability distributions $\{P_\theta : \theta \in \Theta\}$, where Θ is an open subset of $\mathbb{R}^k$; each P_θ has density p_θ with respect to a measure μ. We assume throughout that the parameter θ describes a smooth (twice differentiable with full rank Jacobian) bijective map between the set of distributions and Θ. Consequently, we will refer interchangeably to a subset of parameters and the corresponding set of probability distributions as a *model*.

Suppose we have models corresponding to subsets $\mathcal{M}_i \subseteq \Theta$, $i = 1, 2, \ldots$. We take our models to be either differentiable manifolds or semialgebraic sets,[4] that is, a finite union of sets defined by a finite set of polynomial equalities and inequalities. Semialgebraic sets include a wide range of models of interest; see Drton and Sullivant (2007) for further examples. Our formal discussion of the similarity of these models is based on their tangent cones and tangent spaces at points of intersection.

DEFINITION 2.1. The *tangent cone* $C_\theta(\mathcal{M})$ of a model $\mathcal{M} \subseteq \Theta$ at $\theta \in \mathcal{M}$ is defined as the set of limits of sequences $\alpha_n(\theta_n - \theta)$, such that $\alpha_n > 0$, $\theta_n \in \mathcal{M}$ and $\theta_n \to \theta$.

The *tangent space* of $\mathcal{M}$ at θ is the vector space $T_\theta(\mathcal{M})$ spanned by elements of the tangent cone.

---

[4]Note that these guarantee Chernoff regularity (Drton (2009a), Lemma 3.3 and Remark 3.4).

Clearly, the tangent space contains the tangent cone; the model is said to be *regular at* θ if the two are equal; in particular this means that $\mathcal{M}$ looks like a differentiable manifold (embedded in Θ) at θ, and regular parametric asymptotics apply: in other words, the maximum likelihood estimator is asymptotically normal with covariance given by the inverse Fisher information (van der Vaart (1998)). Most of the models we will initially consider are regular everywhere, though their intersections may not be.

2.1. *c-Equivalence.* Our next definition considers the classification of models based on their local similarity. We work with a local version of the *Hausdorff distance* between sets; this is the furthest distance from any point on one set to the nearest point on the other. Denote this by

$$D(A, B) \equiv \max\Big\{\sup_{a \in A} \inf_{b \in B} \|a - b\|, \ \sup_{b \in B} \inf_{a \in A} \|a - b\|\Big\}.$$

DEFINITION 2.2. We say that $\mathcal{M}_1, \mathcal{M}_2$ are *c-equivalent* at $\theta \in \mathcal{M}_1 \cap \mathcal{M}_2$ if the Hausdorff distance between the sets in a ball of radius ε is $o(\varepsilon^c)$. Formally,

$$\lim_{\varepsilon \downarrow 0} \varepsilon^{-c} D\big(\mathcal{M}_1 \cap N_\varepsilon(\theta), \ \mathcal{M}_2 \cap N_\varepsilon(\theta)\big) = 0,$$

where $N_\varepsilon(\theta)$ is an ε-ball around θ. In other words, within an ε-ball of θ, the maximum distance between the two models is $o(\varepsilon^c)$.

If the limit above is bounded but not necessarily zero—that is, the distance is $O(\varepsilon^c)$—we will say that $\mathcal{M}_1$ and $\mathcal{M}_2$ are *c-near-equivalent* at θ.

The definition of *c*-equivalence is given by Ferrarotti, Fortuna and Wilson (2002), who also derive some of its elementary properties. If $\theta \in \mathcal{M}_1 \cap \mathcal{M}_2$, then 1-near-equivalence is trivial, while 1-equivalence means that to a *linear approximation* around θ the models are the same (formally they have the same tangent cone at θ); this is illustrated by the two surfaces in Figure 3(a). Similar considerations can be applied to higher orders: 2-equivalence means that the quadratic surface best approximating one model is the same as that approximating the other. For D^k-models[5] with $l \leq k - 1$, we have that $(l - 1)$-equivalence for $l \in \mathbb{N}$ implies l-near-equivalence.

PROPOSITION 2.3 (Ferrarotti, Fortuna and Wilson (2002), Proposition 1.3). *Two semi-algebraic models* $\mathcal{M}_1, \mathcal{M}_2$ *are* 1-*equivalent at* $\theta \in \mathcal{M}_1 \cap \mathcal{M}_2$ *if and only if* $C_\theta(\mathcal{M}_1) = C_\theta(\mathcal{M}_2)$.

EXAMPLE 2.4. Consider the two graphical models in Figure 2, defined respectively by the independence constraints $X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y \mid Z$. If we take the set of trivariate Gaussian distributions satisfying these two restrictions, then as discussed in Section 1, the two models are 2-near-equivalent at all diagonal covariance matrices.

For finite discrete variables, however, these two models are not 1-equivalent. Suppose X, Y, Z take n_X, n_Y, n_Z levels, respectively. The models are defined by the equations

$$X \perp\!\!\!\perp Y: \quad p(x, y) - p(x) \cdot p(y) = 0 \quad \forall x, y,$$

$$X \perp\!\!\!\perp Y \mid Z: \quad p(x, y, z) \cdot p(z) - p(x, z) \cdot p(y, z) = 0 \quad \forall x, y, z,$$

where, for example, $p(x, z) = P(X = x, Z = z)$. In particular, the marginal independence model is subject to $(n_X - 1)(n_Y - 1)$ restrictions, but the conditional independence model is subject to that many restrictions for each level of Z. One can use this to show that the dimension of the conditional independence model is smaller than the marginal independence model, and so they cannot both be approximated by the same linear space.

---

[5]that is, k-times differentiable

2.2. *Overlap.*   In spite of the differing dimensions in the example above, we will show that the issue—noted in the Introduction—of the models not being simultaneously convex still arises. This motivates a slightly weaker definition for models meeting in an intuitively 'irregular' manner, which we term *overlap*.

DEFINITION 2.5.   Suppose we have two models $\mathcal{M}_1, \mathcal{M}_2$ with a common point $\theta \in \mathcal{M}_1 \cap \mathcal{M}_2$. We say that $\mathcal{M}_1$ and $\mathcal{M}_2$ *overlap* if there exist points $h \in C_\theta(\mathcal{M}_1) \cap C_\theta(\mathcal{M}_2)$ but $h \notin C_\theta(\mathcal{M}_1 \cap \mathcal{M}_2)$.

In other words, there are tangent vectors that can be obtained in either model, but not in their intersection. Note that necessarily we have $C_\theta(\mathcal{M}_1 \cap \mathcal{M}_2) \subseteq C_\theta(\mathcal{M}_1) \cap C_\theta(\mathcal{M}_2)$, so the definition asks that this is a strict inclusion. For distinct, regular algebraic models, overlap is implied by 1-equivalence.

LEMMA 2.6.   *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be algebraic models that are 1-equivalent and regular at θ, but not equal in a neighbourhood of θ. Then $\mathcal{M}_1$ and $\mathcal{M}_2$ overlap at θ.*

PROOF.   Suppose that, at θ, the models do not overlap and are 1-equivalent. We will show that they are equal in a neighbourhood of θ.

Since $\mathcal{M}_1$ is regular at θ, we can assume that it is an irreducible model, else replace it with its unique irreducible component containing θ (similarly for $\mathcal{M}_2$). The condition that the models are 1-equivalent implies $C_\theta(\mathcal{M}_1) = C_\theta(\mathcal{M}_2)$, and no overlap means $C_\theta(\mathcal{M}_1 \cap \mathcal{M}_2) = C_\theta(\mathcal{M}_1) \cap C_\theta(\mathcal{M}_2) = C_\theta(\mathcal{M}_1)$.

On the other hand, $\mathcal{M}_1 \cap \mathcal{M}_2$ is an algebraic submodel of $\mathcal{M}_1$, so it is either equal to $\mathcal{M}_1$ or has strictly lower dimension. If the latter, then $C_\theta(\mathcal{M}_1 \cap \mathcal{M}_2)$ would be of this same lower dimension (Cox, Little and O'Shea (2007), Theorem 9.7.8), and hence is a strict subset. Since we have already seen that $\mathcal{M}_1 \cap \mathcal{M}_2$ has the same tangent cone as $\mathcal{M}_1$, it must be that the former possibility holds, that is, $\mathcal{M}_1 \cap \mathcal{M}_2 = \mathcal{M}_1$, in a neighbourhood of θ. Clearly, the same is true for $\mathcal{M}_2$, which proves the result.   □

The same result could be applied to *analytic* models, defined by the zeroes of analytic functions, since these functions are (by definition) arbitrarily well approximated by their Taylor series.

REMARK 2.7.   Without assuming that models are analytic, one can construct examples that are 'regular' in most of the usual statistical senses, but for which the previous result fails. As an example of how things can go wrong, consider the function $f(x) = e^{-1/x^2} \sin(1/x^2)$ (taking $f(0) = 0$); this is a C^∞ function, but is not analytic at $x = 0$ (all its derivatives being zero at this point).

Now let $\mathcal{M}_1 = \{(x, 0) : x \in \mathbb{R}\}$ and $\mathcal{M}_2 = \{(x, f(x)) : x \in \mathbb{R}\}$. These models are c-equivalent for every $c \in \mathbb{N}$, but are not equal. However, in contrast to the result of Lemma 2.6 they do *not* overlap. Both sets have tangent cone equal to $\mathcal{M}_1$, and since f has infinitely many roots in any neighbourhood of 0, the tangent cone of $\mathcal{M}_1 \cap \mathcal{M}_2$ is also $\mathcal{M}_1$.

A canonical example of sets that overlap but are not 1-equivalent is given by the subsets of $\mathbb{R}^3$ shown in Figure 3(c); $\mathcal{M}_1 = \{(x, y, z) : y = z = 0\}$ (in blue) and $\mathcal{M}_2 = \{(x, y, z) : z = -x^2\}$ (in red) have $\mathcal{M}_1 \cap \mathcal{M}_2 = \{0\}$. The tangent cone of $\mathcal{M}_2$ is the plane $z = 0$, while $\mathcal{M}_1$ is its own tangent cone and, therefore, $C_0(\mathcal{M}_1) \subseteq C_0(\mathcal{M}_2)$ and $C_0(\mathcal{M}_1) \cap C_0(\mathcal{M}_2) = \mathcal{M}_1$. However, $C_0(\mathcal{M}_1 \cap \mathcal{M}_2) = \{0\}$, so the inclusion is strict. In words, we can approach the origin along a line that becomes tangent to the x-axis in either model, but not in the intersection. The blue model has smaller dimension than the red so they are clearly not 1-equivalent.
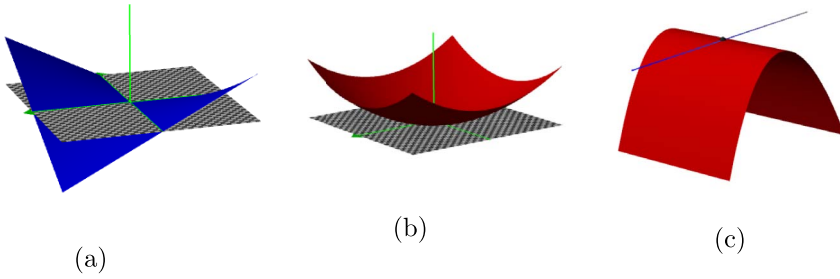
FIG. 3. *Illustration of two models that* (a), (b) *are 1-equivalent;* (c) *overlap. In* (a) *and* (b), *the two surfaces have the same tangent space, but meet in this way for different reasons.*

Typically, it is hard to show that two models are *not* 1-equivalent (or do not overlap) anywhere in the parameter space: indeed the difficulty of verifying such global conditions is one of our motivations for considering these local criteria instead. However, for algebraic models we can often show that such points are at most a set of zero measure within a large class of 'interesting' submodels. To be more precise, if two models are not 1-equivalent at some model of total independence (say θ_0), then they are also 1-equivalent almost nowhere within any algebraic model that contains θ_0.

2.3. *Statistical power.* We will assume that all the models we consider satisfy standard parametric regularity conditions, in particular *differentiability in quadratic mean* (DQM), which yields the familiar asymptotic expansion of the log-likelihood $\ell(\theta)$ (see, e.g., van der Vaart (1998)):

$$(1) \qquad \ell(\theta + h_n) - \ell(\theta) = \frac{h^T}{\sqrt{n}} \dot{\ell}(\theta) - \frac{1}{2} h^T I(\theta) h + o_p(1),$$

where $\dot{\ell}(\theta)$ is the data dependent score, $I(\theta)$ the Fisher information for one observation and $n^{1/2} h_n \to h$. For the purpose of distinguishing between models from data, we need the difference between the log-likelihoods at points close to the MLE not to vanish as sample size $n \to \infty$. The expansion above shows that this requires $h \neq 0$ for the right-hand side to contain a stable term. Hence the distance between the two parameter values needs to shrink no faster than $n^{-1/2}$, the standard parametric rate of statistical convergence.

We consider settings in which $h_n, \tilde{h}_n \to 0$ with $n^{1/2}(h_n - \tilde{h}_n) \to k$, so that we may compare alternatives in two different models. For this reason, we impose the stronger condition that the model is *doubly* differentiable in quadratic mean (DDQM) in some neighbourhood of θ. This is closely related to existence and continuity of the Fisher information and will hold, for example, on the interior of the parameter space of any regular exponential family model.

DEFINITION 2.8. Say that p_θ is *doubly differentiable in quadratic mean* (DDQM) at $\theta \in \Theta$ if for any sequences $h, \tilde{h} \to 0$, we have

$$\int \left(\sqrt{p_{\theta+h}} - \sqrt{p_{\theta+\tilde{h}}} - \frac{1}{2}(h - \tilde{h})^T \dot{\ell}(\theta + \tilde{h}) \sqrt{p_{\theta+\tilde{h}}} \right)^2 d\mu = o(\|h - \tilde{h}\|^2).$$

Note that DDQM reduces to DQM in the special case $\tilde{h} = 0$, and that (by symmetry) we could replace $\dot{\ell}(\theta + \tilde{h}) \sqrt{p_{\theta+\tilde{h}}}$ by $\dot{\ell}(\theta + h) \sqrt{p_{\theta+h}}$. On the other hand, it is strictly stronger than DQM at θ (see Example A.2 in the Appendix). See Appendix A.1 for more details.

THEOREM 2.9. *Let Θ be a model that is DDQM at some θ, and further let $h_n, \tilde{h}_n \to 0$ be sequences such that $k = \lim_n n^{1/2}(h_n - \tilde{h}_n)$. Then*

$$\ell(\theta + h_n) - \ell(\theta + \tilde{h}_n) = \frac{k^T}{\sqrt{n}}\dot{\ell}(\theta + \tilde{h}_n) - \frac{1}{2}k^T I(\theta)k + o_p(1)$$

$$\longrightarrow^d N\left(-\frac{1}{2}k^T I(\theta)k, k^T I(\theta)k\right).$$

The proof is given in Appendix A.1. The theorem shows that it is not possible to distinguish between the two models as n grows if $k = 0$, that is, if the difference in the alternatives shrinks at a rate faster than $n^{-1/2}$. The geometry of the model determines the relationship between the rate at which $h_n, \tilde{h}_n$ shrink and the size of $\tilde{h}_n - h_n$; in general, $\|\tilde{h}_n - h_n\|$ may be of smaller order than both $\|h_n\|$ and $\|\tilde{h}_n\|$.

This leads to our first main result about c-equivalence, which is a direct application of its definition.

THEOREM 2.10. *Suppose S_1, S_2 are c-equivalent sets at x, and consider $h_n = O(n^{-\frac{1}{2c}})$ with $x + h_n \in S_1$. There exists $\tilde{h}_n$ with $x + \tilde{h}_n \in S_2$ such that $\sqrt{n}(h_n - \tilde{h}_n) \to 0$.*

This result is proved in Appendix A.2. Combining the previous two theorems gives the following useful corollary.

COROLLARY 2.11. *Let Θ be a model that is DDQM at some θ, and $\mathcal{M}_1, \mathcal{M}_2 \subset \Theta$ be submodels. If $\mathcal{M}_1, \mathcal{M}_2$ are c-equivalent (resp., c-near-equivalent) at θ then they cannot be asymptotically distinguished under local alternatives to θ of order $O(n^{-\frac{1}{2c}})$ (resp., $o(n^{-\frac{1}{2c}})$).*

All models that intersect are 1-near-equivalent and, therefore, we recover the usual parametric rate; we have power only if h_n shrinks at $n^{-1/2}$ or slower. For 1-equivalent models, this is not enough; a rate of $n^{-1/2}$ will be too fast for us to tell whether our parameters are in $\mathcal{M}_1$ or $\mathcal{M}_2$. In practice, regular models that are 1-equivalent are also 2-near-equivalent, so any rate quicker than $n^{-1/4}$ is also too fast.

2.4. *Convexity.* A second consequence of having 1-equivalent regular models is that it is not possible to have a parameterization with respect to which both models are convex; in fact, this is true even under the weaker assumption of overlap. Some automatic model selection methods such as the lasso (Tibshirani (1996)) rely on a convex parameterization, usually constructed by ensuring that interesting submodels correspond to coordinate zeroes (e.g., $\theta_1 = 0$). Such selection methods cannot be directly applied to model classes that contain overlapping models.

THEOREM 2.12. *Suppose $\mathcal{M}_1$ and $\mathcal{M}_2$ are distinct models that overlap at a point θ on their relative interiors. Then there is no parameterization with respect to which both these models are convex in a neighbourhood of θ.*

PROOF. This is just a statement about sets under differentiable maps with differentiable inverses; note that tangent cones and spaces are isomorphically preserved under such transformations. We will prove the contrapositive: if two sets C, D with $x \in C \cap D$ are convex (in a neighbourhood of x) then they do not overlap at x.

The affine hull of any convex set C is the unique minimal affine space A containing C. Furthermore, by definition of the relative interior, for every $x \in C^{\text{int}}$ there is a neighbourhood of x in which C and A are identical, that is, C coincides with the affine space A at x.
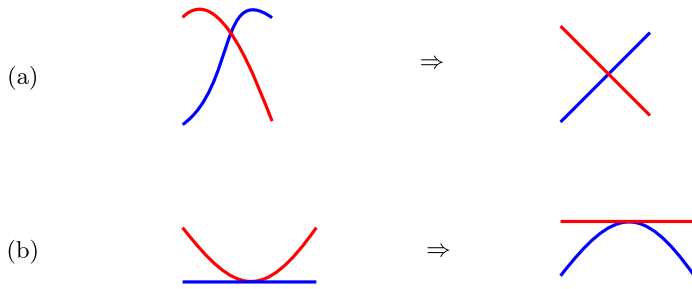
FIG. 4. *Cartoon illustrating why overlapping models cannot be made locally convex. In (a), the red and blue models intersect at a nonzero angle and so the parameterization can be smoothly transformed to make the models locally linear. In (b), there is no angle between the models, and this fact is invariant in any smooth reparameterization.*

Now, suppose two sets C and D are both convex and $x \in C^{\text{int}} \cap D^{\text{int}}$. Then they coincide with affine spaces say A, B, and hence are their own tangent spaces at x. This implies that $C \cap D$ coincides with the affine space $A \cap B$, and hence the tangent cone of $C \cap D$ is just $A \cap B$. Hence C and D do not overlap. Since isomorphic maps cannot alter the tangent cones of these sets, this is true regardless of the parameterization chosen. □

Intuitively, one can reparameterize a regular model such that (at least locally to some point θ) the model is convex. However, this cannot be done simultaneously for two overlapping models. This is illustrated in Figure 4. The result fails if θ is not required to be a point on the relative interior, a counterexample being models $\mathcal{M}_1 = \{\theta_2 \geq \theta_1^2\}$ and $\mathcal{M}_2 = \{\theta_2 \leq -\theta_1^2\}$ with $\theta = (0, 0)$.

REMARK 2.13. Many convex methods, including the lasso, proceed by optimizing a convex function over a parameter space containing all submodels of interest. In the case of the lasso, this function has 'cusps' on submodels of interest that lead to a nonzero probability that the optimum lies exactly on the model. Theorem 2.12 shows that such a procedure cannot be convex if the class contains overlapping models, since the submodels themselves cannot be made convex. Any submodel that is not convex cannot represent a cusp in a convex function and, therefore, we cannot obtain a nonzero probability of selecting such a model.

The nature of this impossibility result should perhaps not be surprising, since all versions of the lasso in the context of ordinary linear models place a restriction on the collinearity of the different parameters in the form of restricted isometry properties or similar (for an overview see, e.g., Bühlmann and van de Geer (2011)). In the case of 1-equivalent models, as we move closer to the point of intersection the angle between the two models shrinks to zero, so no such property could possibly hold. For models that overlap the problem is similar, but may only apply as we approach from certain directions. The result implies, in particular, that an algorithm such as the lasso cannot be used directly to learn Bayesian networks, whether Gaussian or discrete. We expand upon this in Section 4.

If a class of models is nonconvex when parameterized in a canonical way, it may be possible to reparameterize so that they are all convex, but *only if* no two models overlap. For example, suppose we are interested in Gaussian models of marginal independence, that is, models defined by the pattern of zeroes in the off-diagonal elements of the covariance matrix. The log-likelihood of a multivariate Gaussian with zero mean and covariance matrix Σ is

$$\ell(\Sigma; S) = \frac{n}{2}\{\log \det \Sigma^{-1} - \text{tr}(S\Sigma^{-1})\},$$

where S is the sample covariance matrix and n is the sample size. This is a simple function of the inverse covariance matrix Σ^{-1}—the canonical parameter for this exponential family—but a complicated function of the mean parameter Σ. Indeed, while the log-likelihood is a convex function of Σ^{-1}, it is typically not as a function of Σ (see Zwiernik, Uhler and Richards (2017) for an overview of related problems).

However, if we are prepared to accept some loss of efficiency, there is nothing to stop us estimating Σ via a moment matching approach, such as by solving the convex program:

$$\hat{\Sigma} = \arg\min_{\Sigma \succ 0} \left\{ \|\Sigma - S\|^2 + \nu \sum_{i<j} |\sigma_{ij}| \right\};$$

here, $\Sigma \succ 0$ denotes that Σ belongs to the convex set of positive definite symmetric matrices. If the penalty ν is chosen to grow at an appropriate rate (n^δ for some $\frac{1}{2} < \delta < 1$), then under some conditions on the Fisher information for Σ, this will still be consistent for model selection (see, e.g., Rocha, Wang and Yu (2009), Theorem 5).

3. Submodels, equivalence and overlap.

In this section, we consider sufficient conditions for models to be c-equivalent or to overlap. We will see that it is often a consequence of having models whose intersections are themselves expressible as a union of two or more distinct models. Let $\mathcal{M}$ be an algebraic model that can be written as $\mathcal{M} = V_1 \cup V_2$ for incomparable algebraic submodels V_1 and V_2; in this case, we say $\mathcal{M}$ is *reducible*, and otherwise *irreducible*.

3.1. *Identifying c-equivalence.* For the example in Figure 2, the 1-equivalence of the two models $\mathcal{M}_1 = \{\Pi : X \perp\!\!\!\perp Y\}$ and $\mathcal{M}_2 = \{\Pi : X \perp\!\!\!\perp Y \mid Z\}$ was closely related to the fact that if *either* $X \perp\!\!\!\perp Z$ *or* $Y \perp\!\!\!\perp Z$ holds, then $\mathcal{M}_1$ and $\mathcal{M}_2$ intersect. This means that their intersection is a reducible model, as it can be nontrivially expressed as the union of two or more models. It follows that at any points where both the submodels $\mathcal{M}_{X \perp\!\!\!\perp Z}$ and $\mathcal{M}_{Y \perp\!\!\!\perp Z}$ hold, the two original models are 1-equivalent. This is because any direction in $\mathcal{M}_1$ (or $\mathcal{M}_2$) away from the point of intersection $\mathcal{M}_1 \cap \mathcal{M}_2$ can be written as a linear combination of (limits of) vectors that lie in one of $\mathcal{M}_{X \perp\!\!\!\perp Z}$ or $\mathcal{M}_{Y \perp\!\!\!\perp Z}$; the partial derivatives of the distance between $\mathcal{M}_1$ and $\mathcal{M}_2$ in these directions are zero, and so any directional derivative of this quantity is also zero. This is illustrated in Figure 3(a), which shows two surfaces intersecting along two lines: in directions that are diagonal to these lines, the two surfaces separate at a rate that is at most quadratic.
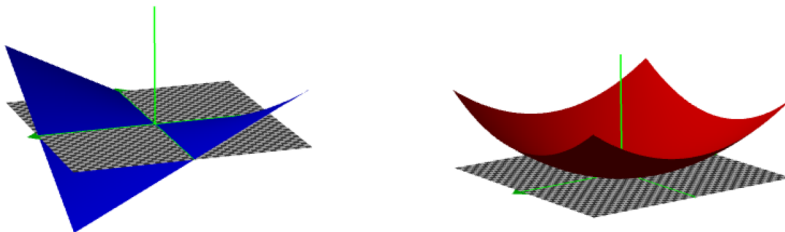
It is not *necessary* for models to intersect in this manner in order to be 1-equivalent: for example, the surfaces $z = 0$ and $z = x^2 + y^2$ intersect only at the point $(0, 0, 0)$ (see Figure 3(b)). However, many models that are 1-equivalent do intersect along at least two submodels, including most of the substantive examples that we are aware of; the time series models in Section 5.2 are an exception. If $c > 2$ submodels are involved in the intersection, then we will see that the original models are c-near-equivalent; consequently asymptotic rates for local alternatives will be even slower than $n^{-1/4}$.

To formalize this, we use the next result. Define the *normal space* of a D^1 surface $\mathcal{M}$ to be the orthogonal complement of its tangent space; i.e. $T_\theta(\mathcal{M})^\perp$.
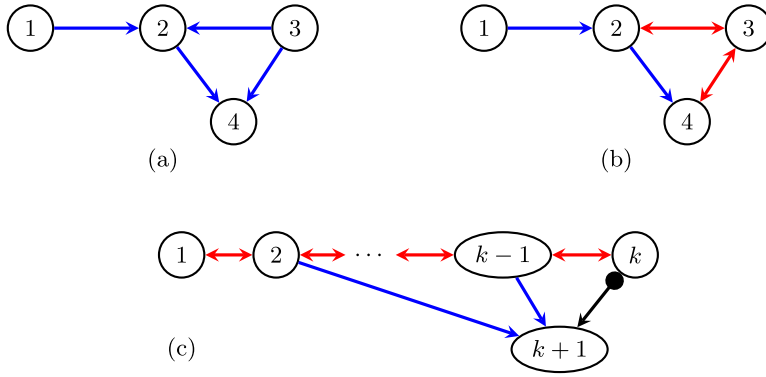
FIG. 5. (a) *and* (b) *are two graphs that differ only by the arrow heads present at* 3. (c) *A discriminating path of length* k.

THEOREM 3.1. *Suppose* $\mathcal{M}_1, \mathcal{M}_2$ *and* $\mathcal{N}_1, \ldots, \mathcal{N}_m$ *are* D^m *manifolds all containing a point* θ, *and such that* $\mathcal{N}_i \cap \mathcal{M}_1 = \mathcal{N}_i \cap \mathcal{M}_2$ *for* $i = 1, \ldots, m$. *Suppose also that* $T_\theta(\mathcal{N}_i \cap \mathcal{M}_j) = T_\theta(\mathcal{N}_i) \cap T_\theta(\mathcal{M}_j)$ *for each* $i = 1, \ldots, m$ *and* $j = 1, 2$, *and further that the normal vector spaces* $T_\theta(\mathcal{N}_1)^\perp, \ldots, T_\theta(\mathcal{N}_m)^\perp$ *all have linearly independent bases. Then* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *are* m-*near-equivalent at* θ.

Note that an algebraic set is always a D^m manifold within a ball around a regular point. In words, there are m distinct submodels on which $\mathcal{M}_1$ and $\mathcal{M}_2$ intersect; in order to distinguish between $\mathcal{M}_1$ and $\mathcal{M}_2$, we need to 'move away' from all the submodels $\mathcal{N}_i$. Linear independence of the normal spaces ensures that we cannot move directly away from several submodels at once.

PROOF. See the Appendix, Section A.3. $\square$

Perhaps an easier way to think about the normal vector spaces is in terms of the submodels being defined by 'independent constraints'; in particular, by constraints defined on different parts of the model. If $\mathcal{N}_1$ is defined by the set of points that are zeros of the functions $f_1, \ldots, f_k$, then $T_\theta(\mathcal{N}_1)^\perp$ contains the space spanned by the Jacobian $J(f_1, \ldots, f_k)$. If $\mathcal{N}_2$ is similarly defined by $g_1, \ldots, g_l$, then the condition is equivalent to saying that the Jacobian of all $k + l$ functions has full rank $k + l$ at θ.

EXAMPLE 3.2 (Discriminating paths). The graphs in Figures 5(a) and (b) are examples of ancestral graphs, which will be introduced more fully in Section 4. Both of these graphs are associated with probabilistic models in which $X_1 \perp\!\!\!\perp X_3$, but in the case of (a) we also have $X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3$, whereas (b) implies $X_1 \perp\!\!\!\perp X_4 \mid X_2$.

In other words, for multivariate Gaussian distributions, both models imply $\rho_{13} = 0$, and (b) gives $f_b(\Pi) = \rho_{14} - \rho_{12}\rho_{24} = 0$, whereas (a) has

$$f_a(\Pi) = \rho_{14} - \rho_{12}\rho_{24} - \rho_{13}\rho_{34} + \rho_{12}\rho_{34}\rho_{23} + \rho_{13}\rho_{24}\rho_{23} - \rho_{14}\rho_{23}^2$$

$$= \rho_{14} - \rho_{12}\rho_{24} + \rho_{12}\rho_{34}\rho_{23} - \rho_{14}\rho_{23}^2 = 0.$$

(Recall that Π is the correlation matrix for the model.) Note that $f_a(\Pi) = f_b(\Pi) + O(\|\Pi - I\|^3)$ when $\rho_{13} = 0$, which strongly suggests these models should be 3-near-equivalent at $\Pi = I$.

Indeed, if any of the three edges between the pairs (1,2), (2,3) and (3,4) are removed, then the models do become Markov equivalent, that is, they represent the same conditional

independences and, therefore, are identical. In the case of multivariate Gaussian distributions, this corresponds to any of the (partial) correlations ρ_{12}, ρ_{23} or $\rho_{34.12}$ being zero. Applying Theorem 3.1 shows that these models are 3-near-equivalent at points where all variables are independent.

This example can be expanded to arbitrarily long *discriminating paths* of the kind shown in Figure 5(c): these differ only by the edges incident to the vertex k, and this makes them distinguishable. However, if any of the edges $(i, i+1)$ for $i = 1, \ldots, k$ are missing, the submodels are Markov equivalent, and hence the Gaussian graphical models are k-near-equivalent. This means that the 'discrimination' between different models that is theoretically possible may be quite limited in practice, absent extraordinarily large sample sizes. We provide a simulation study to illustrate this in Section 4.3.

This example has serious ramifications for the FCI (fast causal inference) algorithm, which uses discriminating paths to orient edges (Zhang (2008)). It suggests only very strong dependence will allow an unambiguous conclusion to be reached for moderate to long paths.

3.2. *Identifying overlap.* Overlap between two regular models occurs when their intersection is not itself a regular model, but rather a union of such models. Unlike with c-equivalence, there is no requirement that normal spaces be linearly independent—incomparability is enough.

THEOREM 3.3. *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be algebraic models, regular at some $\theta \in \mathcal{M}_1 \cap \mathcal{M}_2$, and suppose that $\mathcal{M}_1 \cap \mathcal{M}_2$ is reducible into two (or more) further models that have incomparable tangent cones at θ. Then the models $\mathcal{M}_1$ and $\mathcal{M}_2$ overlap at θ.*

This condition is similar in spirit to Theorem 3.1, but note that here we do not require the normal vector spaces of the two submodels to be linearly independent at the point of intersection, only that the tangent spaces of those pieces are incomparable.

PROOF. Let $\mathcal{M}_1 \cap \mathcal{M}_2 = V_1 \cup V_2$ be the reduction into submodels. Taking $\theta \in V_1 \cap V_2$, note that $C_\theta(\mathcal{M}_1 \cap \mathcal{M}_2) = C_\theta(V_1 \cup V_2) = C_\theta(V_1) \cup C_\theta(V_2)$, the second equality following from the definition of a tangent cone. Now, since $V_1, V_2 \subseteq \mathcal{M}_1$ and $\mathcal{M}_1$ is regular at θ, this implies that any vector in $C_\theta(V_1) + C_\theta(V_2)$ is contained in $C_\theta(\mathcal{M}_1)$; similarly for $\mathcal{M}_2$. Therefore, the condition for not overlapping,

$$C_\theta(\mathcal{M}_1 \cap \mathcal{M}_2) = C_\theta(\mathcal{M}_1) \cap C_\theta(\mathcal{M}_2),$$

holds only if $C_\theta(V_1) + C_\theta(V_2) \subseteq C_\theta(V_1) \cup C_\theta(V_2)$. This occurs only if one of $C_\theta(V_1)$ or $C_\theta(V_2)$ is a subspace of the other, but this was ruled out by hypothesis. $\square$

EXAMPLE 3.4. As already noted in Examples 1.2 and 2.4, the Gaussian graphical models defined respectively by the independences $\mathcal{M}_1 : X \perp\!\!\!\perp Y$ and $\mathcal{M}_2 : X \perp\!\!\!\perp Y \mid Z$ are 1-equivalent at diagonal covariance matrices, but the corresponding discrete models are not. This is because—taking X, Y, Z to be binary—the three-way interaction parameter

$$\lambda_{XYZ} \equiv \frac{1}{8} \sum_{x,y,z \in \{0,1\}} (-1)^{|x+y+z|} \log P(X = x, Y = y, Z = z)$$

is zero in the conditional independence model,[6] but essentially unrestricted in the marginal independence model. However, the intersection of $\mathcal{M}_1$ and $\mathcal{M}_2$ for binary X, Y, Z is the set

---

[6]This is equivalent to $\prod_{x+y+z \text{ even}} p(x, y, z) = \prod_{x+y+z \text{ odd}} p(x, y, z)$, and so certainly still a polynomial condition.

of distributions such that *either* $X \perp\!\!\!\perp Y, Z$ or $Y \perp\!\!\!\perp X, Z$, and these correspond respectively to the submodels

$$\lambda_{XY} = \lambda_{XZ} = \lambda_{XYZ} = 0 \quad \text{or} \quad \lambda_{XY} = \lambda_{YZ} = \lambda_{XYZ} = 0,$$

also defined by zeros of polynomials in P (see Appendix B for full definitions). These models satisfy the conditions of Theorem 3.3 at points of total independence $\perp\!\!\!\perp \{X, Y, Z\}$ and, therefore, $\mathcal{M}_1$ and $\mathcal{M}_2$ *do* overlap.

4. Directed and ancestral graph models. In this section, we focus on two classes of graphical models: Bayesian network models, and the more general ancestral graph models. A more detailed explanation of the relevant theory can be found in Spirtes, Glymour and Scheines (2000) and Richardson and Spirtes (2002).

4.1. *Ancestral graphs.* A *maximal ancestral graph* (MAG) is a simple, mixed graph with three kinds of edge, undirected $(-)$, directed $(\rightarrow)$ and bidirected $(\leftrightarrow)$. Special cases of ancestral graphs include directed acyclic graphs, undirected graphs and bidirected graphs, but not chain graphs. There are some technical restrictions on the structure of the graph which we omit here for brevity: the key detail is that—under the usual Markov property—the model implies a conditional independence constraint between each pair of vertices if (and only if) they are not joined by any sort of edge in the graph (Richardson and Spirtes (2002)). The set that needs to be conditioned upon to obtain the independence depends on the presence of *colliders* in the graph. A collider is a pair of edges that meet with two arrowheads at a vertex k: for example, $i \rightarrow k \leftarrow j$ or $i \rightarrow k \leftrightarrow j$. Any other configuration is called a *noncollider*. We say the collider or noncollider is *unshielded* if i and j are not joined by an edge.

The special case of an ancestral graph model in which all edges are directed yields a *Bayesian network (BN) model*, widely used in causal inference and in machine learning (Bishop (2006), Pearl (2009)). The additional undirected and bidirected edges allow MAGs to represent the set of conditional independence models generated by marginalizing and conditioning a BN model. Ancestral graphs are therefore useful in causal modelling, since they represent the conditional independence model implied by a causal structure with hidden and selection variables.

EXAMPLE 4.1. Consider the maximal ancestral graphs in Figure 5(a) and (b). The graph in (a) is fully directed and represents the model defined by the conditional independences:

$$X_1 \perp\!\!\!\perp X_3, \qquad X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3.$$

The graph in (b), on the other hand, represents

$$X_1 \perp\!\!\!\perp X_3, \qquad X_1 \perp\!\!\!\perp X_4 \mid X_2.$$

The difference in the conditioning sets above is due to the fact that $2 \leftarrow 3 \rightarrow 4$ is a noncollider in the first graph, but a collider in the second: $2 \leftrightarrow 3 \leftrightarrow 4$.

An *independence model*, $\mathcal{I}$, is a collection of (conditional) independence statements of the form $X_i \perp\!\!\!\perp X_j \mid X_C$, for $i \neq j$ and possibly empty C. We will say $\mathcal{I}$ is *simple* if it can be written so that it contains at most one independence statement for each unordered pair $\{i, j\}$. If there is no independence statement between X_i, X_j in a simple independence model, we say i and j are *adjacent*.

THEOREM 4.2. *Let $\mathcal{I}_1, \mathcal{I}_2$ be simple independence models on the space of $p \times p$ Gaussian covariances matrices. Then the two models are 2-near-equivalent if they have the same adjacencies.*

Further, if the two models have different adjacencies then they are 1-equivalent on at most a null set within any parametric independence model.

PROOF.   Let E denote the set of adjacencies in a simple independence model $\mathcal{I}$. Parameterizing using the set of correlation matrices, we will show that the tangent space of $\mathcal{I}$ at the identity matrix I is

$$(2) \qquad\qquad T_I(\mathcal{I}) = \bigoplus_{\substack{i<j \\ \{i,j\}\in E}} D^{ij},$$

where the matrices D^{ij} have zeroes everywhere except in the (i, j) and (j, i)th entries, which are 1.

If $\{i, j\} \in E$, it is easy to see that $I + \lambda D^{ij}$ is in the model for all $\lambda \in (-1, 1)$, since this means that all conditional independences except those between X_i and X_j hold; hence $D^{ij} \in T_I(\mathcal{I})$. Conversely, if i and j are not adjacent, then some independence restriction $X_i \perp\!\!\!\perp X_j \mid X_C$ holds, so

$$f(\Pi) = \Pi_{ij} - \Pi_{iC}(\Pi_{CC})^{-1}\Pi_{Cj} = 0.$$

The derivative of f at $\Pi = I$ is just D^{ij}, so it follows that $D^{ij} \notin T_I(\mathcal{I})$. Hence the tangent space at I is in the form (2). By Proposition 2.3, the models are 1-equivalent, and since these constraints are linearly independent at $\Pi = I$, they are regular and, therefore, also 2-near-equivalent.

Conversely, suppose that there is some pair i, j subject to the restriction $X_i \perp\!\!\!\perp X_j \mid X_C$ in $\mathcal{I}_1$ but not to any such restriction in $\mathcal{I}_2$. By the above analysis, these models have distinct tangent spaces at $\Pi = I$. Since these are models defined by polynomials in Π, the set of points on which the tangent spaces are identical (say W) is an algebraic model; its intersection with any irreducible model V is therefore either equal to V or of strictly smaller dimension than V (indeed this follows from the usual definition of dimension in such sets; see Bochnak, Coste and Roy (2013), Section 2.8). However, if the identity matrix is contained in V then clearly $W \cap V \subset V$, since we have established that the tangent spaces do not intersect at the identity. Hence $W \cap V$ has smaller dimension, and is a null subset of V. $\square$

COROLLARY 4.3.   *Two Gaussian maximal ancestral graph models are 2-near-equivalent if they have the same adjacencies* (*when viewed either as an independence model or a graph*), *and are otherwise 1-equivalent almost nowhere on any submodel of independence.*

PROOF.   This follows from the pairwise Markov property of Richardson and Spirtes (2002). $\square$

We conjecture that, in fact, two ancestral graph models of the form given in Theorem 4.2 will overlap nowhere in the set of positive definite correlation matrices if they do not share the same adjacencies (rather than almost nowhere). It is not hard to see that this holds for models of different dimension, since these models are regular and therefore will share this dimension everywhere. To prove it in general seems challenging; the result above is sufficient for most practical purposes.

COROLLARY 4.4.   *Let $\mathcal{G}, \mathcal{H}$ be chain graphs with the same adjacencies under any of the interpretations given in* Drton (2009b) (*not necessarily the same interpretation*). *Then the corresponding models are 2-near-equivalent.*

Note that Remark 5 of Drton (2009b) makes clear that pairwise independences are sufficient to define Gaussian chain graph models.
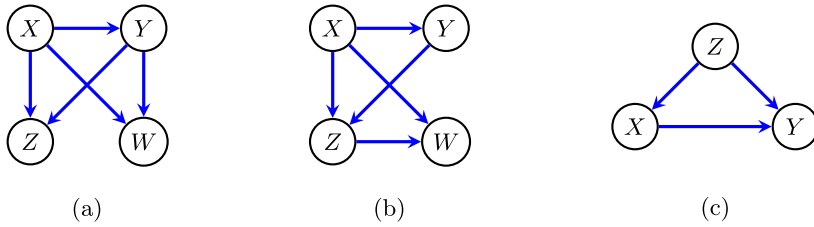
FIG. 6. (a) and (b) two Bayesian networks which do not overlap but are not consistent with a single topological order. (c) A simple causal model.

4.2. *Discrete data.* The picture is slightly rosier if we consider discrete data instead. A well-known result of Chickering (1996) shows that finding an optimal Bayesian network for discrete data is a NP-hard problem; in other words, it is computationally difficult. However, we find that from a *statistical* point of view, it is somewhat easier than in the Gaussian case.

THEOREM 4.5. *No two distinct, binary, maximal ancestral graph models are 1-equivalent at the model of total independence.*

PROOF.   See the Appendix, Section C.   □

Although distinct, discrete MAG models (and, therefore, BN models) are never 1-equivalent, they do still overlap, as our next result demonstrates.

PROPOSITION 4.6.   *Let $\mathcal{M}(\mathcal{G}_1), \mathcal{M}(\mathcal{G}_2)$ be two discrete Bayesian network models such that $i \to k \leftarrow j$ is an unshielded collider in $\mathcal{G}_1$ but an unshielded noncollider in $\mathcal{G}_2$. Then, if X_k is binary, the two models overlap.*

PROOF.   See the Appendix, Section C.   □

REMARK 4.7.   The condition that some variables are binary is, in fact, unnecessary; see Remark C.2 for more details on the general finite discrete case.

Bayesian network models that are consistent with a single topological ordering of the vertices do not overlap, because their intersection is always another BN model. We can therefore work with a class defined by the subgraphs of a single complete BN in order to avoid the problems associated with overlap.

This leads to the question of whether any other, perhaps larger, subclasses share this property. The previous result shows that any such subclass would be restricted fairly severely, since any two graphs must never disagree about a specific unshielded collider. Note that the result does not imply that it is *necessary* for graphs to be consistent with a single topological order in order for the corresponding models not to overlap; the graphs in Figure 6(a) and (b) provide a counterexample to this.

The easiest way to ensure that a class of models does not overlap at the independence model is to associate each potentially missing edge with a single constraint using a pairwise Markov property, as with the set of BN models that are consistent with a given topological order, or the set of undirected graph models.

Note that Proposition 4.6 and Theorem 2.12 combine to show that the impossibility result discussed in Remark 2.13 applies to binary Bayesian networks, and we will never be able to use a lasso-like method to consistently select from this class under standard conditions.
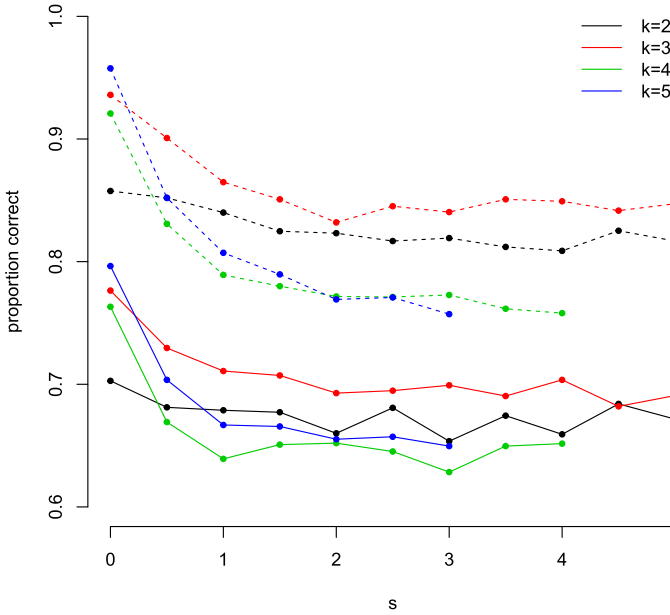
FIG. 7. *Ability to discriminate between $\mathcal{G}_k$ and $\mathcal{G}'_k$ for various k and s. A sample of 2500 data sets were drawn from the Gaussian graphical model according to the scheme described in the text, with effect sizes $\rho_s = 0.4 \times 2^{-s}$; the sample size was fixed at $n_{init} \times 2^{2ks}$ for some n_{init}. The y-axis gives the proportion of times $\mathcal{G}_k$ correctly gave a lower deviance than $\mathcal{G}'_k$. Each solid line corresponds to a value of $k \in \{2, 3, 4, 5\}$ and a corresponding sample size $n_{init} \in \{32, 250, 800, 5000\}$ (dashed lines correspond to an initial sample size of $4n_{init}$). The flattening out each line shows agreement with the prediction of Corollary 2.11. Note that the vertical ordering of the different lines is merely a function of the choice of n_{init}. Larger values of k and s were excluded because the relevant sample sizes grew too quickly ($k = s = 5$ corresponds to $n = 5.6 \times 10^{18}$ for the solid line).*

4.3. *Discriminating paths.* Example 3.2 introduced the notion of a discriminating path, which allows the identification of colliders in an ancestral graph. Formally, define the ancestral graphs $\mathcal{G}_k$ and $\mathcal{G}'_k$ as having vertices $1, \ldots, k+1$, with a path $1 \leftrightarrow 2 \leftrightarrow \cdots \leftrightarrow k$ and directed edges from each of $2, \ldots, k-1$ to $k+1$. In addition, $\mathcal{G}_k$ has the edges $k \leftrightarrow k+1$, while $\mathcal{G}'_k$ has $k \rightarrow k+1$; the graphs are shown in Figure 5(c), with only the final edge left ambiguous. The path from 1 to $k+1$ is known as a *discriminating path*, and its structure allows us to determine whether there is a collider at k (as in $\mathcal{G}_k$) or not (as in $\mathcal{G}'_k$).

As noted already, if any of the edges $i \leftrightarrow i+1$ for $i = 1, \ldots, k-1$ are missing (i.e., if $X_i \perp\!\!\!\perp X_{i+1}$), then the two models coincide. In addition, if the final edge between k and $k+1$ is missing (so $X_{k+1} \perp\!\!\!\perp X_k \mid X_1, \ldots, X_{k-1}$), the two models coincide. We consider the Gaussian graphical models associated with these graphs and, by application of Theorem 3.1, the two ancestral graph models for $\mathcal{G}_k$ and $\mathcal{G}'_k$ are k-near-equivalent at points where these additional independences hold.

We now perform a small simulation to show that, in order to maintain power as the effect sizes shrink, the sample sizes need to grow at the rates claimed in Corollary 2.11. To do this, we take the structural equation model parameterization from Richardson and Spirtes (2002), Section 8; for $\mathcal{G}_k$ we set the weight of each edge on the discriminating path to be $\rho_s = 0.4 \times 2^{-s}$, and of the other edges to be 0.5. We take a sample size of $n_{\text{init}} \times 2^{2ks}$ for some n_{init}, so that this grows at the rate suggested by Corollary 2.11 to keep the power constant in s. The plot in Figure 7 shows that the simulations agree with the predictions of that result, since the power stays roughly constant as s grows.

Table 1 lists some of the colossal sample sizes that are needed to keep power constant in longer discriminating paths as effect sizes shrink. In the $k = 3$ case, we have approximately

TABLE 1
Table giving effect sizes ρ_s and sample sizes used in each of the graphs $\mathcal{G}_k$ for $k = 2, 3, 4, 5$, and corresponding to the solid lines in Figure 7. For $k = 4, 5$, the sample sizes grow extremely quickly, and some larger entries are excluded because they led to numerical problems. 'acc.' gives the proportion of runs for which the correct graph was selected

		$k = 2$		$k = 3$		$k = 4$		$k = 5$	
s	ρ_s	n	acc.	n	acc.	n	acc.	n	acc.
0	0.4	32	0.703	250	0.776	800	0.763	5000	0.796
0.5	0.283	128	0.681	2000	0.730	12,800	0.669	160,000	0.704
1	0.2	512	0.679	16,000	0.711	204,800	0.639	5,120,000	0.667
1.5	0.141	2048	0.677	128,000	0.707	3,276,800	0.651	163,840,000	0.666
2	0.1	8192	0.660	1,024,000	0.693	52,428,800	0.652	5.24×10^9	0.655
2.5	0.071	32,768	0.681	8,192,000	0.695	838,860,800	0.645	1.68×10^{11}	0.657
3	0.05	131,072	0.654	65,536,000	0.699	1.34×10^{10}	0.628	5.37×10^{12}	0.650
3.5	0.035	524,288	0.674	524,288,000	0.690	2.15×10^{11}	0.650		
4	0.025	2,097,152	0.659	4.19×10^9	0.704	3.44×10^{12}	0.652		
4.5	0.018	8,388,608	0.684	3.36×10^{10}	0.682				
5	0.012	33,554,432	0.672	2.68×10^{11}	0.691				

70% accuracy when $\rho = 0.2$ and $n = 16,000$, but to maintain this when $\rho = 0.1$ we need $n = 1,024,000$. For $k = 4$, the respective sample sizes are an increase from $n = 2 \times 10^5$ to $n = 5 \times 10^7$ (a 256-fold increase), and for $k = 5$ from $n = 5 \times 10^6$ to $n = 5 \times 10^9$ (a 1024-fold increase).

5. Other classes of model.

5.1. *Undirected graphs.* An undirected graphical model associates a simple undirected graph with a collection of probability distributions. Under the pairwise Markov property, the model consists of distributions such that $X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}$ whenever i and j are not joined by an edge in the graph. For Gaussian graphical models, this is equivalent to enforcing a zero in the (i, j) entry in the inverse covariance matrix. It is a simple matter to see that no two distinct undirected Gaussian graphical models ever overlap.

This helps to explain why undirected models are fundamentally easier to learn than other classes, something which has been much exploited in high-dimensional statistics. For example, the graphical lasso (Friedman, Hastie and Tibshirani (2008), Witten, Friedman and Simon (2011)) and neighbourhood selection (Meinshausen and Bühlmann (2006)) methods allow very fast consistent model selection amongst undirected Gaussian graphical models, including in high-dimensional settings.

5.2. *Time series.* The *autoregressive moving average* process of order (p, q), or ARMA(p, q) model, is a time series model that assumes

$$X_t = \varepsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}, \quad t \in \mathbb{Z},$$

where $\varepsilon_t \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ and the parameters $\boldsymbol{\phi} = (\phi_i)_{i=1}^{p}$ and $\boldsymbol{\theta} = (\theta_i)_{i=1}^{q}$ are unknown. The model is stationary and Gaussian and, therefore, parameterized by σ^2 and the autocorrelations, that is, $\gamma_i \equiv \text{Cor}(X_t, X_{t+i})$ for $i = 1, 2, \ldots$. Let the space spanned by γ_i be L^i.

The special cases where $p = 0$ and $q = 0$ are respectively the MA(q) and AR(p) models. These are identifiable, and the unique parameter values that lead to independence of the X_ts
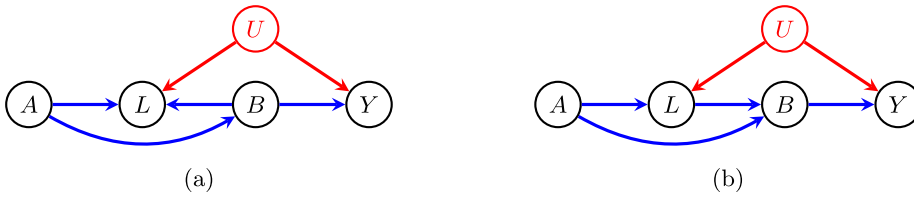
FIG. 8. *Two causal models on five variables, differing in the direction of the L − B edge. The variable U is assumed to be unobserved.*

are $(\phi, \theta) = (\mathbf{0}, \mathbf{0})$. It is not hard to see that the derivative of the joint correlation matrix with respect to either ϕ_i or θ_i is just L^i. Hence the tangent space of an AR(p) or MA(p) model at this point is of the form

$$C_{(\mathbf{0},\mathbf{0})}(\mathcal{M}_p) = \bigoplus_{i=1}^{p} L^i,$$

and so the two AR(p) and MA(p) models are 1-equivalent at the point of joint independence. This suggests it will be hard to distinguish between these two types of process when correlations are weak—though this may not matter if the aim of an analysis is predictive.

5.3. *Nested Markov models.* Richardson et al. (2017) introduce *nested Markov models*, which are defined by a generalized form of conditional independence that may hold under a Bayesian network model with hidden variables. For example, the two causal Bayesian networks in Figure 8 differ only in the direction of a single edge (that between L and B). Assuming that the variable U is unobserved, the model (a) implies a single observable conditional independence constraint: $Y \perp\!\!\!\perp A \mid B$.

The model in (b) does not imply any conditional independence constraint over the observed variables, but does impose the restriction that

$$p(y \mid \mathrm{do}(a, b)) \equiv \sum_{l} p(l \mid a) \cdot p(y \mid a, l, b)$$

does not depend on a. This can be interpreted as the statement that A does not causally affect Y, except through B. Note that, if $L \perp\!\!\!\perp B \mid A$, then

$$p(y \mid \mathrm{do}(a, b)) \equiv \sum_{l} p(l \mid a) \cdot p(y \mid a, l, b)$$

$$= \sum_{l} p(l \mid a, b) \cdot p(y \mid a, l, b)$$

$$= p(y \mid a, b),$$

so the statement that this quantity does not depend upon a is the same as the ordinary independence statement $Y \perp\!\!\!\perp A \mid B$; a similar conclusion can be reached if $Y \perp\!\!\!\perp L \mid A, B$. Thus there are two distinct submodels along which the two models in Figure 8 intersect. These two submodels also have linearly independent normal vector spaces because they correspond to restrictions on $p(a \mid y, b)$ and $p(l \mid a, y, b)$. It follows that the two models are 2-near-equivalent at points where $Y \perp\!\!\!\perp A, L \mid B$ (where both submodels hold) by Theorem 3.1; note that unlike for the case of ancestral graph models, this applies even in the case of discrete variables.

So, even though in principle one can distinguish the two models in Figure 8 and determine the orientation of the $L − B$ edge, in practice the distinction may be hard to show with

data. This suggests it will be very difficult to learn the correct model without further information, something borne out by the simulations in Shpitser et al. (2013). Interestingly, the approach in that paper of setting higher-order parameters to zero in order to equalize parameter counts between models may actually have made learning significantly harder, since in some cases this means removing the only directions in the tangent space that differ between models.

6. Related phenomena.

6.1. *Double robustness.* The phenomenon of *double robustness* of estimators has been exploited to improve estimation in causal models (Scharfstein, Rotnitzky and Robins (1999)); see Kang and Schafer (2007) for an overview. The essential element of double robustness is to choose an estimating equation with two components, such that the solution is a $\sqrt{n}$-consistent estimator when one of the two parts is correctly specified, even if the other is misspecified. For simplicity, we will just consider examples where we assume independence rather than fitting a model.

Consider the simple causal model depicted in Figure 6(c), and suppose we are interested in the causal effect of a binary variable X on the expectation of Y, but there is a (potentially continuous) measured confounder Z. The causal distribution for X on Y is given by $p(y|do(x)) = \sum_z p(z) \cdot p(y|x, z)$; this is generally different from the ordinary conditional $p(y|x) = \sum_z p(z|x) \cdot p(y|x, z)$.

What happens if we use the ordinary conditional anyway? One can easily check that $p(y|x) = p(y|do(x))$ if either $X \perp\!\!\!\perp Z$ or $Y \perp\!\!\!\perp Z \mid X$, and hence the set of distributions where our estimate is correct contains the union of points $\mathcal{M}_{X \perp\!\!\!\perp Z} \cup \mathcal{M}_{Y \perp\!\!\!\perp Z|X}$. If we apply Theorem 3.1, we find that any error in estimating the causal effect will be quadratic in the distance from the point where $Z \perp\!\!\!\perp X, Y$.

6.2. *Triple robustness.* Another phenomenon known as *triple robustness*[7] is observed in some causal models related to mediation (Tchetgen Tchetgen and Shpitser (2012)). In this case, an estimator will be consistent provided at least two out of three other quantities are correctly specified. Here, we introduce another result related to Theorem 3.1.

PROPOSITION 6.1. *Let $\mathcal{N}_1, \ldots, \mathcal{N}_m$ be algebraic submodels containing 0, and let f be a polynomial such that $f(x) = 0$ for any $x \in \mathcal{N}_i \subset \Theta$ for $i = 1, \ldots, m$. Suppose also that the tangent spaces $\mathrm{T}_0(\mathcal{N}_i)$ jointly span Θ. Then $f(x) = O(\|x\|^2)$.*

PROOF. For any twice differentiable function with $f(0) = 0$, whichever direction we move away from 0 in can be written as a linear combination of directions in the tangent spaces of submodels $\mathcal{N}_i$. It follows that the directional derivative of any such function is zero, in any direction. Hence $f(x) = O(\|x\|^2)$. □

In light of this, suppose we have three submodels $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ each defined by constraints on linearly independent parts of the full model Θ, and such that an estimator is consistent on the intersection of *any two* of them. The condition on the definition of the $\mathcal{M}_i$ means that their normal vector spaces are linearly independent, so any vector is in the tangent space of at least two such submodels. Their pairwise intersections thus satisfy the conditions of the theorem, and the error in estimating the relevant parameter is quadratic in the distance to the joint intersection $\mathcal{M}_1 \cap \mathcal{M}_2 \cap \mathcal{M}_3$.

---

[7]Perhaps misleadingly, since it is strictly weaker than double robustness.

6.3. *Post-double-selection.* Belloni, Chernozhukov and Hansen (2014) consider the problem of estimating a causal effect $p(y|do(x))$ in the presence of a high-dimensional measured confounder Z_I, $I = \{1, \ldots, p\}$ where $p \gg n$. We can try to find a subset $S \subseteq I$ such that S is much smaller than I, and Z_S is sufficient to control for the confounding, that is, $p(y|do(x)) = \sum_{z_S} p(z_S) p(y|x, z_S)$. Formally, this will be satisfied if we ignore any Z_i such that $Z_i \perp\!\!\!\perp X \mid Z_{-i}$ or $Z_i \perp\!\!\!\perp Y \mid X, Z_{-i}$. However, in finite samples selecting variables creates an *omitted-variable bias*, in which the decision boundary of whether to drop a particular variable leads to a bias of order $O(n^{-1/2})$ at some points in the parameter space.

If we only exclude components of Z_i for which *both* $Z_i \perp\!\!\!\perp X \mid Z_{-i}$ *and* $Z_i \perp\!\!\!\perp Y \mid X, Z_{-i}$, then the order of the bias on our causal estimate is effectively squared and becomes $O(n^{-1})$; this is because—for the same reason as in the discussion of double robustness—the bias induced by a component Z_i is at most quadratic in the distance of the true distribution from the intersection of these two independences. Since this bias is small compared to the sampling variance, it can effectively be ignored; this idea is referred by Belloni, Chernozhukov and Hansen (2014) to as *post-double-selection*, and can be viewed as another consequence of the local geometry of the model.

7. Algorithms for learning models with overlap.

Suppose we have a class of models $\mathcal{M}_i$ that overlap but are not 1-equivalent, that is, the models all have different tangent cones. We have seen already that overlapping models place restrictions on one class of computationally efficient methods, because they cannot be made convex. This suggests that a method which attempts to learn the (linear) tangent cone rather than selecting the model directly may be computationally advantageous. This can be achieved by learning from a set of 'surrogate' models that have the same tangent spaces as the original models, but that do not overlap.

To take a simple example, consider again the graphical models in Figure 2 for binary variables. Letting

$$\lambda_{XY} = \frac{1}{8} \sum_{x,y,z \in \{0,1\}} (-1)^{|x+y|} \log P(X = x, Y = y, Z = z),$$

then the model $\mathcal{M}_2$ in (b) consists of the zero parameters: $\lambda_{XY} = \lambda_{XYZ} = 0$, while (a) involves log-linear parameters over the X, Y-margin $\mathcal{M}_1 : \lambda'_{XY} = 0$. This apparently makes model selection tricky because some models involve zeroes of ordinary log-linear parameters, and some of marginal log-linear parameters.

However, one could try replacing $\mathcal{M}_1$ with a model that corresponds to the zero of the ordinary log-linear parameter, for example, $\mathcal{M}'_1 : \lambda_{XY} = 0$. This has the same tangent space as $\mathcal{M}_1$ at the uniform distribution, and so it is 'close' to $\mathcal{M}_1$ in a precise sense. This suggests that if we pursued a model selection strategy for ordinary log-linear parameters and learned $\lambda_{XY} = 0$ but $\lambda_{XYZ} \neq 0$ (i.e. we chose $\mathcal{M}'_1$), then we could conclude that this is sufficiently similar to $\mathcal{M}_1$ to select this model from the graphical class.

7.1. *Model selection.* Define $\Lambda_i = \langle e_i \rangle$ to be the vector space spanned by the ith coordinate axis. Suppose we have a class of regular algebraic models $\mathcal{M}_i \subseteq \Theta \subseteq \mathbb{R}^k$ such that each model has a tangent space at $\theta = 0$ defined by a subset of coordinate axes, that is, for each model there is some set $c(\mathcal{M}_i) \subseteq \{1, \ldots, k\}$ such that

$$C_0(\mathcal{M}_i) = \bigoplus_{j \in c(\mathcal{M}_i)} \Lambda_j.$$

Suppose further that our class of models is such that $c(\mathcal{M}_i) \neq c(\mathcal{M}_j)$ for any $i \neq j$. We have already seen that, if two models overlap, then it may be computationally difficult to learn the correct one due to a lack of convexity. We will show that, under certain assumptions about the

true parameter being sufficiently close to $\theta = 0$, we can learn the tangent space itself, thereby circumventing the models' lack of convexity.

Denote the sparsity pattern of a parameter by $c(\theta) = \{i : \theta_i \neq 0\}$; then $\theta \in T_0(\mathcal{M})$ implies that $c(\theta) \subseteq c(T_0(\mathcal{M})) = c(\mathcal{M})$. However, note that $\theta \in \mathcal{M}$ *does not* imply this, and in general the sparsity patterns of parameters in $\mathcal{M}$ is arbitrary.

Suppose we have a model selection procedure which returns a set $\hat{S}$ estimating $S \subseteq \{1, \ldots, k\}$ such that $\theta_S \neq 0$ and $\theta_{S^c} = 0$. We will assume that the procedure is consistent, in the sense that if $\theta_{S^c}^n = o(n^{-1/2})$ and $|\theta_s^n| = \omega(n^{-1/2})$ for each $s \in S$, we have $P(\hat{S} = S) \to 1$. These conditions are satisfied by many common model selection methods such as BIC, or an L_1-penalized selection method with appropriate penalty (and with Fisher information matrix satisfying certain irrepresentability conditions, Nardi and Rinaldo (2012)).

The following result shows that we can adapt a model selection procedure of this kind to learn models that overlap, by replacing each (possibly nonconvex) model of interest by a convex surrogate model with the same tangent space.

THEOREM 7.1. *Consider a sequence of parameters of the form $\theta_n = \theta_0 + n^{-\gamma} h + O(n^{-2\gamma})$ such that each $\theta_n \in \mathcal{M}$; here, $\frac{1}{4} < \gamma < \frac{1}{2}$ and $h_i \neq 0$ for any $i \in c(\mathcal{M})$. Suppose our model selection procedure provides a sequence of parameter estimates $\tilde{\theta}_n$.*
Then $P(c(\tilde{\theta}_n) = c(\mathcal{M})) \to 1$ as $n \to \infty$.

PROOF. If $i \notin c(\mathcal{M})$, then $\theta_{ni} = O(n^{-2\gamma}) = o(n^{-1/2})$ since $\gamma > \frac{1}{4}$, while if $i \in c(\mathcal{M})$ then $\theta_{ni} = \Omega(n^{-\gamma}) = \omega(n^{-1/2})$ since $\gamma < \frac{1}{2}$. By the conditions on our model selection procedure, then we have the required consistency. $\square$

The condition $\gamma > \frac{1}{4}$ is to ensure that the bias induced by using the surrogate model is too small to detect at the specific sample size, and that our procedure will set the relevant parameters to zero. Slower rates of convergence (i.e., $0 < \gamma \leq \frac{1}{4}$) could also lead to a satisfactory model selection procedure if we were simply to subsample our data or otherwise 'pretend' that n is smaller than it actually is. If $\gamma > \frac{1}{2}$, on the other hand, we will not have asymptotic power to identify the truly nonzero parameters.

Note that $P(c(\tilde{\theta}_n) = c(\theta_n))$ does *not* tend to 1, since the sparsity pattern of the true parameter is not the same as that of the tangent space of the model: it is merely 'close' to having the correct sparsity.

The assumption that θ_n tends to 0 at the required rate may seem rather artificial: some assumption of this form is unavoidable, simply because our results only hold in a neighbourhood of points of intersection. The precise rate at which $\theta_n \to 0$ just needs to be such that 'real' effects do not disappear faster than we can statistically detect them (i.e., slower than $n^{-1/2} \to 0$), and that any other effects shrink fast enough that they are taken to be zero. Any more realistic framework would require conditions on the global geometry of the models, and would be extremely challenging to verify.

7.2. *Application to Bayesian networks.* Suppose that we have a sequence of binary distributions p_n, with $\|p_n - p_0\| = O(n^{-\gamma})$ for $\frac{1}{4} < \gamma < \frac{1}{2}$, where p_0 is the uniform distribution, and such that each p_n is Markov with respect to a Bayesian network $\mathcal{G}$; assume also that $\lambda_{ij} = \omega(n^{-1/2})$ if i, j are adjacent, and $\lambda_{ijk} = \omega(n^{-1/2})$ if $i \to k \leftarrow j$ is an unshielded collider. Then a consistent method to determine $\mathcal{G}$ would be as follows:

- select the model for p_n using the log-linear lasso with penalty $\nu = n^{\delta}$ for some $\frac{1}{2} < \delta < 1$;
- then find the graph. Asymptotically, the sparsity pattern of the log-linear model is the same as that of the original graph, so this can be done by simply finding the graph with skeleton given by $i - j$ when $\lambda_{ij} \neq 0$ and orienting unshielded triples as colliders $i \to k \leftarrow j$ if and only if $\lambda_{ijk} \neq 0$.

This is far from an optimal approach, but does give an idea of how one might be able to overcome the nonconvexity inherent to overlapping models. The algorithm could also be extended to ancestral graphs, via Theorem 4.5.

8. Discussion. We have proposed that the geometry of two models at points of intersection is a useful measure of how statistically difficult it will be to distinguish between them, and shown that when models' tangent spaces are not closed under intersection this restricts the possibility of using convex methods to perform model selection in the class. We have also given examples of model classes in which this occurs and noted that in several cases, model selection is indeed known to be difficult.

We suggest that special consideration should be given in model selection problems to whether or not the class contains models that overlap or are 1-equivalent and—if it does— to whether a smaller and simpler model class can be used instead. Alternatively, additional experiments may need to be performed to help distinguish between models. The results in this paper provide a point of focus for new model selection methods and also for experimental design. If we are able to work with a class of models that is less rich and, therefore, easier to select from, then perhaps we ought to. If we cannot, it is useful to know in advance at what points in the parameter space it is likely to be difficult to draw clear distinctions between models, so that we can either power our experiments appropriately, or report that we do not know which of several models is correct.

APPENDIX A: TECHNICAL RESULTS

A.1. Asymptotics. We start with the definition of differentiability in quadratic mean.

DEFINITION A.1.   Let $(p_\theta : \theta \in \Theta)$ be a class of densities with respect to a measure μ indexed by some open $\Theta \subseteq \mathbb{R}^k$. We say that this class is *differentiable in quadratic mean* (DQM) at θ if there exists a vector $\dot{\ell}(\theta) \in \mathbb{R}^k$ such that

$$\int \left[\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2} h^T \dot{\ell}(\theta) \sqrt{p_\theta} \right]^2 d\mu = o(\|h\|^2).$$

Recall also our definition of a model that is doubly DQM.

Definition 2.8 Say that p_θ is *doubly differentiable in quadratic mean* (DDQM) at $\theta \in \mathcal{M}$ if for any sequences $h, \tilde{h} \to 0$, we have

$$\int \left(\sqrt{p_{\theta+h}} - \sqrt{p_{\theta+\tilde{h}}} - \frac{1}{2}(h - \tilde{h})^T \dot{\ell}(\theta + \tilde{h}) \sqrt{p_{\theta+\tilde{h}}} \right)^2 d\mu = o(\|h - \tilde{h}\|^2).$$

Recall also that DDQM reduces to DQM in the special case $\tilde{h} = 0$, and that (by symmetry) we could replace $\dot{\ell}_{\theta+\tilde{h}} \sqrt{p_{\theta+\tilde{h}}}$ by $\dot{\ell}_{\theta+h} \sqrt{p_{\theta+h}}$. On the other hand, it is strictly stronger than DQM at θ.

EXAMPLE A.2.   Suppose that $(X, Y)^T \sim N(\eta, I)$ where $\eta(\theta) = (\theta_1 \theta_2^{1/3}, \theta_2)$. We claim that p_θ is DQM at $(0, 0)$ but not DDQM.

Obviously, $\eta(0, 0) = (0, 0)$ and p_η is DQM at $\eta = (0, 0)$, so

$$\int \left(\sqrt{p_\eta} - \sqrt{p_0} - \eta^T \dot{\ell}(0) \sqrt{p_0} \right)^2 d\eta = o(\|\eta\|^2).$$

But now for any θ, we have

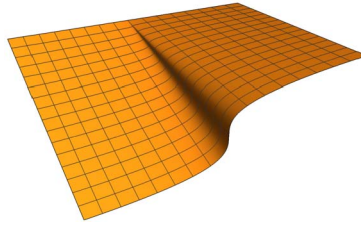$$\frac{p_\theta}{p_0} = \exp\left\{ -\frac{1}{2} [(x - \eta_1)^2 + (y - \eta_2)^2 - x^2 - y^2] \right\}$$

FIG. 9. *Surface plot of the function $f(x, y) = xy^{1/3}$.*

$$= 1 + x\theta_1\theta_2^{1/3} + y\theta_2 + o(\|\theta\|)$$

$$= 1 + y\theta_2 + o(\|\theta\|),$$

$$\sqrt{\frac{p_\theta}{p_0}} = 1 + \frac{1}{2}y\theta_2 + o(\|\theta\|).$$

Hence

$$\mathbb{E}_0\left(\sqrt{\frac{p_\theta}{p_0}} - 1 - \theta_2 Y\right)^2 = o(\|\theta\|^2),$$

and p_θ is also DQM at $(0, 0)$.

Now let $\theta^n = (n^{-1/6}, n^{-1/2})$ and $\tilde{\theta}^n = (n^{-1/6}, -n^{-1/2})$. Then $\eta^n = (n^{-1/3}, n^{-1/2})$ and $\tilde{\eta}^n = (-n^{-1/3}, -n^{-1/2})$. In particular, the sequence $\|\eta^n - \tilde{\eta}^n\|$ is of order $n^{-1/3}$, and hence we will have power to choose between the two sequences.

This (somewhat pathological) construction has a 'wrinkle' in the surface that maps θ to η; see Figure 9 for an illustration. The wrinkle is too small for DQM to fail at $\theta = 0$, but pairs of points close to the wrinkle and to each other in θ space may be far apart in η space. This model therefore fails to satisfy the condition of $\sqrt{p_\theta}$ being continuously differentiable at $\theta = (0, 0)$.

LEMMA A.3. *Let p_θ be DDQM at $\theta \in \Theta$. Then the Fisher information $I(\theta) \equiv \mathbb{E}\dot{\ell}(\theta)\dot{\ell}(\theta)^T$ exists and is continuous in a neighbourhood of $\theta \in \Theta$.*

PROOF. Since p_θ is DDQM at θ it is also DQM, and hence $I(\theta)$ exists by van der Vaart (1998), Theorem 7.2. In addition, the symmetry of DDQM shows that $\lim_{t\to 0} h^T I(\theta + th)h = h^T I(\theta)h$ for any h, so this matrix must indeed exist in a neighbourhood of θ.

By the symmetry property noted above, $\int[(h - t\tilde{h})^T \dot{\ell}(\theta + t^2\tilde{h})\sqrt{p_{\theta+t^2\tilde{h}}}]^2 d\mu = (h - t\tilde{h})^T I(\theta + t^2\tilde{h})(h - t\tilde{h})$ and $(h - t\tilde{h})^T I(\theta + th)(h - t\tilde{h})$ have the same limit, and these are in turn the same as the respective limits of $h^T I(\theta + t^2\tilde{h})h$ and $h^T I(\theta + th)h$. Since h and $\tilde{h}$ are arbitrary, this shows that $I(\theta)$ is continuous at θ. □

We now prove Theorem 2.9, closely following the proof of Theorem 7.2 of van der Vaart (1998).

PROOF OF THEOREM 2.9. Let p_n and $\tilde{p}_{,n}$ respectively, denote $p_{\theta+h_n}$ and $p_{\theta+\tilde{h}_n}$. By DDQM we have that $\sqrt{n}(\sqrt{p_n} - \sqrt{\tilde{p}_n}) - \frac{1}{2}k^T \dot{\ell}(\theta + \tilde{h}_n)\sqrt{\tilde{p}_n}$ converges in quadratic mean to 0; since the second term is bounded in squared expectation (given by $k^T I(\theta + \tilde{h}_n)k/4$) so is the first, and hence $n^\gamma(\sqrt{p_n} - \sqrt{\tilde{p}_n}) \to 0$ in quadratic mean for any $\gamma < \frac{1}{2}$.

Let $g_n = k^T \dot{\ell}(\theta + \tilde{h}_n)$. Note that DDQM implies that $\frac{1}{2} g_n \sqrt{\tilde{p}_n}$ has the same limit as $\sqrt{n}(\sqrt{p}_n - \sqrt{\tilde{p}_n})$. By continuity of the inner product, we have

$$\lim_n \mathbb{E}_{\theta + \tilde{h}_n} g_n = \lim_n \int g_n \tilde{p}_n \, d\mu$$

$$= \lim_n \int \frac{1}{2} g_n \sqrt{\tilde{p}_n} 2 \sqrt{\tilde{p}_n} \, d\mu$$

$$= \lim_n \sqrt{n} \int (\sqrt{p_n} - \sqrt{\tilde{p}_n})(\sqrt{p_n} + \sqrt{\tilde{p}_n}) \, d\mu$$

$$= \lim_n \sqrt{n} \int (p_n - \tilde{p}_n) \, d\mu$$

$$= 0,$$

since $\int (p_n - \tilde{p}_n) \, d\mu = 1 - 1 = 0$ for all densities p_n, $\tilde{p}_n$.

Let $W_{ni} = 2(\sqrt{p_n/\tilde{p}_n}(X_i) - 1)$ where $X_i \sim \tilde{p}_n$. Then

$$n \mathbb{E} W_{ni} = 2n \int \sqrt{p_n \tilde{p}_n} \, d\mu - 2n = -n \int (\sqrt{p_n} - \sqrt{\tilde{p}_n})^2 \, d\mu,$$

which, by the DDQM condition, has the same limit as

$$-\frac{1}{4} k^T \left(\mathbb{E}_{\theta + \tilde{h}_n} \dot{\ell}(\theta + \tilde{h}_n)^T \dot{\ell}(\theta + \tilde{h}_n) \right) k = -\frac{1}{4} k^T I(\theta + \tilde{h}_n) k.$$

Then

$$\text{Var}_{\theta + \tilde{h}_n} \left(\sum_i W_{ni} - \frac{1}{\sqrt{n}} \sum_i g_n(X_i) \right)$$

$$\leq \mathbb{E}_{\theta + \tilde{h}_n} (\sqrt{n} W_{ni} - g_n(X_i))^2$$

$$= \int (\sqrt{n} W_{ni} - g_n(X_i))^2 \tilde{p}_n \, d\mu$$

(3)

$$= \int (\sqrt{n} 2(\sqrt{p_n/\tilde{p}_n} - 1) - k^T \dot{\ell}(\theta + \tilde{h}_n))^2 \tilde{p}_n \, d\mu$$

$$= \int 4 \left(\sqrt{n}(\sqrt{p_n} - \sqrt{\tilde{p}_n}) - \frac{1}{2} k^T \dot{\ell}(\theta + \tilde{h}_n) \sqrt{\tilde{p}_n} \right)^2 d\mu$$

$$= o(1)$$

by DDQM. It follows from all this that the sequence of random variables

$$\sum_i W_{ni} - \frac{1}{\sqrt{n}} \sum_i g_n(X_i) + \frac{1}{4} k^T I(\theta + \tilde{h}_n) k$$

has mean and variance tending to zero, and hence they converge to zero in probability.

Using a Taylor expansion, we obtain

$$\log \prod_{i=1}^n \frac{p_{\theta + h_n}}{p_{\theta + \tilde{h}_n}}(X_i) = 2 \sum_{i=1}^n \log(1 + W_{ni}/2)$$

(4)

$$= \sum_{i=1}^n W_{ni} - \frac{1}{4} \sum_{i=1}^n W_{ni}^2 + \frac{1}{2} \sum_{i=1}^n W_{ni}^2 R(W_{ni})$$

for some function R such that $\lim_{x \to 0} R(x) = 0$. By the right-hand side of (3), we have $nW_{ni}^2 = g_n(X_i)^2 + A_{ni}$ for some A_{ni} such that $\mathbb{E}|A_{ni}| \to 0$, and hence $\bar{A}_n = n^{-1}\sum_i A_{ni}$ converges in probability to 0. Then

$$\sum_i W_{ni}^2 - n^{-1}\sum_i g_n(X_i)^2 = \bar{A}_n = o_p(1).$$

We also have

$$P\Big(\max_i |W_{ni}| > \varepsilon\sqrt{2}\Big) \leq n P\big(|W_{ni}| > \varepsilon\sqrt{2}\big)$$

$$\leq n P\big(g_n(X_i)^2 > n\varepsilon^2\big) + n P\big(|A_{ni}| > n\varepsilon^2\big)$$

$$\leq \varepsilon^{-2}\mathbb{E}g_n(X_i)^2 \mathbb{I}_{\{g_n(X_i)^2 > n\varepsilon^2\}} + \varepsilon^{-2}\mathbb{E}|A_{ni}|.$$

We already have $\mathbb{E}|A_{ni}| \to 0$, and since $\mathbb{E}g_n(X_i)^2 = k^T I(\theta + \tilde{h}_n)k$ is continuous (and hence bounded) by Lemma A.3, the first term also tends to 0. It follows that $\max_i |W_{ni}| = o_p(1)$, and thus $\max_i |R(W_{ni})| = o_p(1)$. Note therefore that the final term is bounded by $\max_{1 \leq i \leq n} |R(W_{ni})| \cdot \sum_{i=1}^n W_{ni}^2 = o_p(1)O_p(1)$ which converges to zero in probability.

Putting this back into (4) gives

$$\log \prod_{i=1}^n \frac{p_{\theta+h_n}}{p_{\theta+\tilde{h}_n}}(X_i) = \sum_{i=1}^n W_{ni} - \frac{1}{4}k^T I(\theta + \tilde{h}_n)k + o_p(1).$$

We then directly obtain

$$\ell(\theta + h_n) - \ell(\theta + \tilde{h}_n) = \frac{1}{\sqrt{n}}\sum_i g_n(X_i) - \frac{1}{2}k^T I(\theta + \tilde{h}_n)k + o_p(1)$$

$$= \frac{1}{\sqrt{n}}k^T \dot{\ell}(\theta + \tilde{h}_n) - \frac{1}{2}k^T I(\theta + \tilde{h}_n)k + o_p(1).$$

Using the fact that $I(\cdot)$ is continuous at θ, then gives the required result. $\quad\square$

Sufficient conditions for DQM are given in Lemma 7.6 of van der Vaart (1998); in fact, these conditions are also sufficient for DDQM.

LEMMA A.4. *Assume that $\theta \mapsto s_\theta(x) := \sqrt{p_\theta(x)}$ is μ-almost everywhere continuously differentiable, and that the matrix $I(\theta) := \int (\dot{p}_\theta/p_\theta)(\dot{p}_\theta^T/p_\theta)p_\theta \, d\mu$ has well-defined continuous entries. Then p_θ is DDQM.*

PROOF. We follow the same proof method as Lemma 7.6 in van der Vaart (1998). By the chain rule, p_θ is also differentiable with $\dot{p}_\theta = 2s_\theta \dot{s}_\theta$, and hence $\dot{s}_\theta = \frac{1}{2}(\dot{p}_\theta/p_\theta)\sqrt{p_\theta}$.

Since $\theta \mapsto s_\theta(x)$ is continuously differentiable (assuming for now that x excludes the set of measure zero on which this fails), we can write

$$\frac{s_{\theta+th} - s_{\theta+tg}}{t} = (h - g)^T \dot{s}_{\theta+t(g+u(h-g))}$$

for some $u \in [0, 1]$ by the mean value theorem.

By Cauchy–Schwarz and Fubini, we have

$$\int \left(\frac{s_{\theta+th} - s_{\theta+tg}}{t}\right)^2 d\mu \leq \int \int_0^1 \big((h - g)^T \dot{s}_{\theta+t(g+u(h-g))}\big)^2 \, du \, d\mu$$

$$= \frac{1}{4}\int_0^1 (h - g)^T I\big(\theta + t(g + u(h - g))\big)(h - g) \, du$$

$$\longrightarrow \frac{1}{4}(h - g)^T I(\theta)(h - g);$$

here we have used the continuity of $I(\cdot)$. Continuous differentiability of $s(\cdot)$ shows that $t^{-1}(s_{\theta+th} - s_{\theta+tg}) - (h-g)^T \dot{s}_{\theta+tg} \to 0$ pointwise, and hence its integral converges to zero by Proposition 2.29 of van der Vaart (1998). $\square$

We remark that Lemmas A.3 and A.4 show a close correspondence between continuity of the Fisher information and DDQM.

It is a standard result that the conditions of Lemma A.4 are satisfied by an exponential family provided that θ is in the interior of the natural parameter space.

A.2. Proof of Theorem 2.10.

PROOF.   Assume $x = 0$ without loss of generality. Given $h_n \in S_1$ with $h_n \to 0$, and the fact that S_1 and S_2 are c-equivalent at 0, there exists a sequence $\tilde{h}_n = h_n + o(\|h_n\|^c)$. Now, if $h_n = O(n^{-\frac{1}{2c}})$, then $\tilde{h}_n - h_n = o(n^{-1/2})$, which gives the required result. A similar argument holds for c-near-equivalence, giving $\tilde{h}_n - h_n = O(n^{-1/2})$. $\square$

A.3. Proof of Theorem 3.1.

LEMMA A.5.   *Let the conditions of Theorem 3.1 be satisfied with $m \geq 2$. Then $\mathcal{M}_1$ and $\mathcal{M}_2$ have the same dimension and tangent space at θ.*

PROOF.   We have that $T_\theta(\mathcal{M}_1 \cap \mathcal{N}_i) = T_\theta(\mathcal{M}_1) \cap T_\theta(\mathcal{N}_i)$ for $i = 1, 2$. Since $\mathcal{N}_1$ and $\mathcal{N}_2$ have disjoint normal spaces, it follows that

$$
\begin{aligned}
T_\theta(\mathcal{M}_1) &= T_\theta(\mathcal{M}_1) \cap \big(T_\theta(\mathcal{N}_1) + T_\theta(\mathcal{N}_2)\big) \\
&= T_\theta(\mathcal{M}_1 \cap \mathcal{N}_1) + T_\theta(\mathcal{M}_1 \cap \mathcal{N}_2) \\
&= T_\theta(\mathcal{M}_2 \cap \mathcal{N}_1) + T_\theta(\mathcal{M}_2 \cap \mathcal{N}_2) \\
&= T_\theta(\mathcal{M}_2) \cap \big(T_\theta(\mathcal{N}_1) + T_\theta(\mathcal{N}_2)\big) \\
&= T_\theta(\mathcal{M}_2).
\end{aligned}
$$

Since $\mathcal{M}_1, \mathcal{M}_2$ are regular at θ, this completes the proof. $\square$

The proof below makes modest use of differential geometry; the basics may be found in Conlon (2008).

PROOF OF THEOREM 3.1.   We choose $\theta = 0$ for convenience. The result is clear for $m = 1$, so assume $m \geq 2$. By Lemma A.5, $\mathcal{M}_1$ and $\mathcal{M}_2$ share a common dimension and tangent space at 0. Since $\mathcal{M}_1, \mathcal{M}_2$ are D^m surfaces at 0, they can each be locally represented by a D^m parametric function, say $\phi_1, \phi_2 : U \to \mathbb{R}^k$. Assume that $\phi_1(0) = \phi_2(0) = 0$, and that these functions share a derivative at 0 with respect to $u \in U$.

Choose ϕ_1 so that $\phi_1(u) = (u, 0) \in \mathbb{R}^d \times \mathbb{R}^{k-d}$, by the constant rank theorem (Conlon (2008), Theorem 2.4.6). Note this means that $\phi_2(u) = (u, O(\|u\|^2))$ by the definition of the tangent space and the fact that ϕ_2 is at least D^2. Also set $\phi_1^{-1}(A) = \phi_2^{-1}(A)$ for all $A \subseteq \mathcal{M}_j \cap \mathcal{N}_i$; then for each $u : \phi_1(u) \in \mathcal{M}_1 \cap \mathcal{N}_i$ we have $\phi_2(u) = (u, 0)$.

By the implicit function theorem, each of the remaining $k - d$ coordinates of ϕ_2 can be written as a D^m function of the first d. By a further invertible D^m transformation, we can ensure that $\phi_1(u), \phi_2(u) \in \mathcal{N}_i$ whenever $u_{c_{i-1}+1} = \cdots = u_{c_i} = 0$ (where $c_i - c_{i-1}$ is the codimension of $\mathcal{M}_j \cap \mathcal{N}_i$ in $\mathcal{M}_j$).

Note that this means that not only is $\frac{\partial \phi_1(0)}{\partial u_a} = \frac{\partial \phi_2(0)}{\partial u_a}$ for all a, but indeed $\frac{\partial^{m-1}\phi_1(0)}{\partial u_{a_1}\cdots\partial u_{a_{m-1}}} = \frac{\partial^{m-1}\phi_2(0)}{\partial u_{a_1}\cdots\partial u_{a_{m-1}}}$ for all $a_1, \ldots, a_{m-1}$, because there could still be some $i \in \{1, \ldots, m\}$ such that $u_{c_{i-1}+1} = \cdots = u_{c_i} = 0$; therefore, we are still (potentially) in at least one of the m submodels, and $\phi_1^{-1}(y) = \phi_2^{-1}(y)$ holds at this point.

It follows that the Taylor expansions of ϕ_1 and ϕ_2 at 0 to order $m-1$ are identical, and the first term in which there is any difference will be of the form

$$\frac{1}{m!} u_{a_1} \cdots u_{a_m} \frac{\partial^m \phi_j(0)}{\partial u_{a_1}\cdots\partial u_{a_m}},$$

where each $a_i \in \{c_{i-1}+1, \ldots, c_i\}$. As a consequence of the product $u_{a_1}\cdots u_{a_m}$, it follows that $\|\phi_1(u) - \phi_2(u)\| = O(\|u\|^m)$, and hence $\mathcal{M}_1$ and $\mathcal{M}_2$ are m-near-equivalent.   □

APPENDIX B: LOG-LINEAR PARAMETERS

Let V be a finite set, and define the log-linear design matrix as a $2^{|V|} \times 2^{|V|}$ matrix with rows and columns indexed by subsets of V, such that

$$M_{A,B} = (-1)^{|A \cap B|}.$$

We denote the Bth column (or equivalently row) of M by M_B. Note that

$$M_B = \bigodot_{v \in B} M_{\{v\}},$$

where $\odot$ denotes the Hadamard (or pointwise) product. As an example, here is a log-linear design matrix for three items.

$$M = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix} \begin{matrix} \varnothing \\ \{1\} \\ \{2\} \\ \{1,2\} \\ \{3\} \\ \{1,3\} \\ \{2,3\} \\ \{1,2,3\} \end{matrix}$$

For example, the fourth column of M is $M_{\{1,2\}}$ and is given by the pointwise product of the second and third columns $M_{\{1\}}$ and $M_{\{2\}}$. Note also that M is *involutory*—that is, its own inverse—up to a constant: $M^{-1} = 2^{-|V|}M$.

Let $X_V = (X_v)_{v \in V}$ be a vector of binary random variables. We abbreviate the event $\{X_v = 0$ for all $v \in C\}$ to 0_C, and similarly $\{X_v = 1$ for all $v \in C\}$ to 1_C. Let $\eta_A = \log p(1_A, 0_{V \setminus A})$. Then we define the log-linear parameters via the identities

$$\eta = M\lambda, \qquad \lambda = M^{-1}\eta.$$

Letting $p = (p(x_V) : x_V \in \mathfrak{X}_V)$, assumed to be ordered in the same way as η so that $\eta = \log p$, we have

$$\frac{\partial \lambda}{\partial p} = \frac{\partial \lambda}{\partial \eta}\frac{\partial \eta}{\partial p} = M^{-1}(\operatorname{diag} p(x_V))^{-1}.$$

Of interest to us is the connection between log-linear parameters within different marginal distributions, known as *marginal log-linear parameters* (Bergsma and Rudas (2002)). Denote the log-linear parameters within a marginal distribution X_K by $\lambda^K \equiv (M^K)^{-1} \log p(x_K)$,

where M^K is the appropriate restriction of M to rows and columns indexed by subsets of K. We continue to denote the ordinary log-linear parameter associated with a particular interaction set A by $\lambda_A = \lambda_A^V$.

LEMMA B.1. *The derivative of the parameter λ_A^K (with respect to $\boldsymbol{p}$) lies in the span of the derivatives of $\lambda_A^V, \ldots, \lambda_V^V$ if $X_K \perp\!\!\!\perp X_{V \setminus K}$. If $K = A$, then the converse also holds.*

Additionally, if $p(x_V)$ is uniform then λ_A^K and λ_A^V have the same derivative M_A.

PROOF. Let $x_V = (1_B, 0_{V \setminus B})$. We have

$$\frac{\partial \lambda_A^K}{\partial p(x_V)} = \frac{M_{A,B}}{2^{|K|} p(x_K)} \quad \text{and}$$

(5)

$$\sum_{C \subseteq V \setminus A} \alpha_C \frac{\partial \lambda_{AC}^V}{\partial p(x_V)} = \frac{M_{A,B}}{2^{|V|} p(x_V)} \sum_{C \subseteq V \setminus A} \alpha_C M_{C,B}$$

since $M_{A \cup C, B} = M_{A,B} \cdot M_{C,B}$. For the derivatives $\frac{\partial \lambda_A^K}{\partial \boldsymbol{p}}$ to lie in the span of the derivatives $\frac{\partial \lambda_{AC}^V}{\partial \boldsymbol{p}}$ we need to find α_C to solve

$$\frac{M_{A,B}}{2^{|V|} p(x_V)} \sum_{C \subseteq V \setminus A} \alpha_C M_{C,B} = \frac{M_{A,B}}{2^{|K|} p(x_K)}$$

or equivalently

(6)
$$\sum_{C \subseteq V \setminus A} \alpha_C M_{C,B} = 2^{|V \setminus K|} p(x_{V \setminus K} | x_K)$$

for each x_V. If $X_{V \setminus K} \perp\!\!\!\perp X_K$, this becomes $\sum_{C \subseteq V \setminus K} \alpha_C M_{C,B} = 2^{|V \setminus K|} p(x_{V \setminus K})$, which has a solution because it amounts to $2^{|V \setminus K|}$ linearly independent equations in $2^{|V \setminus A|} \geq 2^{|V \setminus K|}$ variables.

In the case that $K = A$, note that there are precisely as many variables as equations, and since the coefficients $M_{C,B}$ in the expression on the left of (6) do not vary with x_K (since $C \cap K = \varnothing$), it is necessary for $X_{V \setminus K} \perp\!\!\!\perp X_K$ in order for a solution to exist. For $A \subset K$, a similar argument shows that $X_{V \setminus K} \perp\!\!\!\perp X_D \mid X_{K \setminus D}$ for some $D \subset K$ with $|D| \leq |K \setminus A|$ is sufficient.

If $p(x_V)$ is uniform, then note that the derivative of λ_A^K does not depend upon K. $\square$

Since the map from $\boldsymbol{\lambda}$ to $\boldsymbol{p}$ is a smooth one, this yields us the following corollary.

COROLLARY B.2. *If $\varepsilon \equiv \| p - p_0 \|$ for p_0 under which all variables are uniform, then*

$$\lambda_A^K = \lambda_A^L + O(\varepsilon^2).$$

B.1. Log-linear models are algebraic. Note that a log-linear model is algebraic, since $\lambda_A = c$ if and only if

$$\prod_{\|x_A\|_1 \text{ even}} p(x_A, x_{V \setminus A}) - e^{c 2^{|V|}} \prod_{\|x_A\|_1 \text{ odd}} p(x_A, x_{V \setminus A}) = 0.$$

Hence they are defined by the zeroes of polynomials in p.

APPENDIX C: ANCESTRAL GRAPHS

PROOF OF THEOREM 4.5.   Let $\mathcal{G}$ be a graph with vertices V. We can parameterize the binary probability simplex using log-linear parameters λ_K for $\varnothing \neq K \subseteq V$ (see Appendix B for details). We consider the tangent space of the model at the uniform distribution, that is, at $\lambda_K = 0$ for every K.

The conditional independence $X_a \perp\!\!\!\perp X_b \mid X_C$ is equivalent to $\lambda'_{abD} = 0$ for each $D \subseteq C$, where λ'_K are the log-linear parameters for the marginal distribution over X_a, X_b, X_C (Rudas, Bergsma and Németh (2010)). However, within δ of the uniform distribution we have $\lambda'_K = \lambda_K + O(\delta^2)$ (see Corollary B.2). Hence the constraint to the tangent cone imposed by $\lambda'_{abD} = 0$ is the same as that imposed by $\lambda_{abD} = 0$ for each $D \subseteq C$.

Now, two MAGs are Markov equivalent if and only if they have the same adjacencies, unshielded colliders and discriminating paths (e.g., Zhang (2008), Proposition 2). If they differ in adjacencies (say i, j), then a log-linear parameter λ_{ij} will appear in the tangent cone of one model but not the other. If they differ in an unshielded collider $i *\!\!\rightarrow k \leftarrow\!\!* j$, then in the model without a collider $\lambda_{ijk} = 0$ but in the other this direction is not restricted. The same holds for a discriminating path between i and j for a potential collider k. $\square$

REMARK C.1.   Note that this proof also demonstrates that the skeleton of the graph is determined solely by the two-way interaction parameters, and that the remainder of the model can be deduced entirely from the three-way interaction parameters. This suggests that it might be possible to develop a model selection procedure using only this information, something we have done in Section 7. It also illustrates that in cases with strong three-way interactions it should be easier to learn the correct model rather than just the correct skeleton. A 'noisy-OR' model would, for example, have the desired property. The well-known ALARM dataset (Beinlich et al. (1989)) has strong interaction effects and is—at least in part for this reason—considered to be relatively easy to learn.

PROOF OF PROPOSITION 4.6.   We claim that the spaces spanned by λ_{ik} and λ_{jk} are contained in the tangent cones of both models, but not of their intersection; the first claim follows from Theorem 4.5. For the second, if X_k is binary then $\lambda'_{ij} = \lambda_{ij} = 0$ if and only if either $\lambda_{ik} = 0$ or $\lambda_{jk} = 0$ (see Drton, Sturmfels and Sullivant (2009), Example 3.1.7). Clearly then directions in which they are both nonzero will not appear in the intersection model. $\square$

REMARK C.2.   Note that these results can easily be extended to a general finite discrete case, though the notation becomes rather cumbersome. In particular, suppose that the states-pace is $\mathfrak{X}_V = \prod_{v \in V} \mathfrak{X}_v$ for some finite sets $\mathfrak{X}_v$. In this case, λ_A represents a collection of parameters of dimension $\prod_{a \in A}(|\mathfrak{X}_a| - 1)$; these are redundant whenever some x_a is equal to a suitable reference value (say $0_a \in \mathfrak{X}_a$), since they can be inferred from the remaining values.

Then we define

$$\lambda_A(x_A) = |\mathfrak{X}_V|^{-1} \sum_{y_V \in \mathfrak{X}_V} \log p(y_V) \prod_{v \in A} \left(|\mathfrak{X}_v| \mathbb{1}_{\{x_v = y_v\}} - 1\right).$$

Hence $\lambda_\varnothing = |\mathfrak{X}_V|^{-1} \sum_{y_V} \log p(y_V)$ and

$$\lambda_1(x_1) = |\mathfrak{X}_V|^{-1} \sum_{y_V \in \mathfrak{X}_V} \left(|\mathfrak{X}_1| \mathbb{1}_{\{x_1 = y_1\}} - 1\right) \log p(y_V),$$

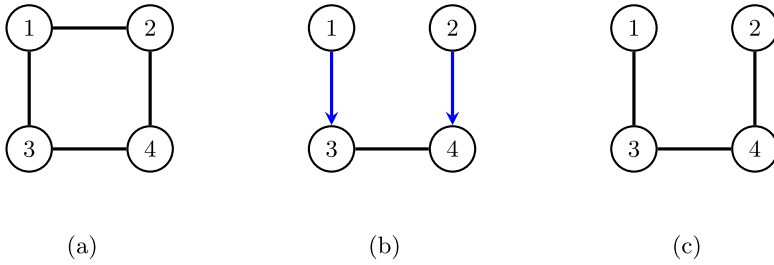for example. The results in Section 4 still hold with these parameters at analogous locations.

FIG. 10. *Three chain graphs. When interpreted under the LWF Markov property, all associated models satisfy* $X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3$ *and* $X_2 \perp\!\!\!\perp X_3 \mid X_1, X_4$; *these constraints define the model in* (a). *The submodel* (b) *additionally implies that* $X_1 \perp\!\!\!\perp X_2$, *whereas* (c) *implies* $X_1 \perp\!\!\!\perp X_2 \mid X_3, X_4$.

Now, in the case of Proposition 4.6, the same result will hold even if X_k is not binary. Evans (2015), Theorem 3.1, shows that $\lambda_{ij}^{Ak} = \lambda_{ij}^A + g(\lambda_{k|A})$, where $g = 0$ whenever $X_k \perp\!\!\!\perp X_l \mid X_{A \setminus \{l\}}$ for any $l \in V$. This means that $X_k \perp\!\!\!\perp X_i \mid X_{A \setminus \{i\}}$ or $X_k \perp\!\!\!\perp X_j \mid X_{A \setminus \{j\}}$ are included in the intersections of the two models. Further, one can check that the models are not identical, since adding $\varepsilon > 0$ to $P(X_A = x_A, X_k = x_k)$ and subtracting it from $P(X_A = x_A, X_k = x'_k)$ will not change the fact that $X_i \perp\!\!\!\perp X_j \mid X_{A \setminus \{i,j\}}$, but we will no longer have $X_i \perp\!\!\!\perp X_j \mid X_{A \setminus \{i,j\}}, X_k$. It follows that the two models overlap by Theorem 3.3.

APPENDIX D: EXAMPLE: DISCRETE LWF CHAIN GRAPHS

EXAMPLE D.1. Consider the graphs shown in Figure 10. Interpreted using the Lauritzen–Frydenberg–Wermuth (LWF) Markov property (Lauritzen and Wermuth (1989)), these three graphs all represent distinct models. The graph in (a) satisfies the usual Markov property for undirected graphs:

$$X_1 \perp\!\!\!\perp X_4 \mid X_2, X_3, \qquad X_2 \perp\!\!\!\perp X_3 \mid X_1, X_4.$$

The graphs in (b) and (c) satisfy these independences, as well as the respective independences $X_1 \perp\!\!\!\perp X_2$ (in the case of (b)), and $X_1 \perp\!\!\!\perp X_2 \mid X_3, X_4$ for (c). These two additional constraints are generally distinct, but they coincide if any of the three remaining edges are not present: that is, if any of the conditional independences

$$X_1 \perp\!\!\!\perp X_3 \mid X_2, X_4, \qquad X_3 \perp\!\!\!\perp X_4 \mid X_1, X_2, \qquad X_2 \perp\!\!\!\perp X_4 \mid X_1, X_3,$$

also hold. Under either model, each of these constraints corresponds to a single zero log-linear parameter:

$$\lambda_{13} = 0, \qquad \lambda_{34} = 0, \qquad \lambda_{24} = 0$$

[recall that $\lambda_A := 2^{-|V|} \sum_{x_V} (-1)^{|x_A|} \log P(X_V = x_V)$]. Taking the models defined by these three constraints, we can apply Theorem 3.1 and find that the two models in (b) and (c) are 3-near-equivalent at all points in the model of complete independence.

One can extend this example arbitrarily by drawing a graph of the form $1 \to 3 - 4 - \cdots - k \leftarrow 2$ and comparing it to its undirected counterpart. In this case, if the effect corresponding to any of the $k - 1$ edges is missing, then the two models intersect. Hence, by Theorem 3.1 these two models are $(k - 1)$-near-equivalent.

REFERENCES

ARAGAM, B. and ZHOU, Q. (2015). Concave penalized estimation of sparse Gaussian Bayesian networks. *J. Mach. Learn. Res.* **16** 2273–2328. MR3450507

BEINLICH, I. A., SUERMONDT, H. J., CHAVEZ, R. M. and COOPER, G. F. (1989). The ALARM monitoring system. In *Proceedings of AIME 89: Second European Conference on Artificial Intelligence in Medicine* (J. Hunter, J. Cookson and J. Wyatt, eds.) 247–256. Springer, Berlin Heidelberg.

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983 https://doi.org/10.1093/restud/rdt044

BERGSMA, W. P. and RUDAS, T. (2002). Marginal models for categorical data. *Ann. Statist.* **30** 140–159. MR1892659 https://doi.org/10.1214/aos/1015362188

BISHOP, C. M. (2006). *Pattern Recognition and Machine Learning. Information Science and Statistics*. Springer, New York. MR2247587 https://doi.org/10.1007/978-0-387-45528-0

BOCHNAK, J., COSTE, M. and ROY, M.-F. (2013). *Real algebraic geometry*. Springer, Berlin.

BREIMAN, L. (2001). Statistical modeling: The two cultures. *Statist. Sci.* **16** 199–231. MR1874152 https://doi.org/10.1214/ss/1009213726

BÜHLMANN, P., PETERS, J. and ERNEST, J. (2014). CAM: Causal additive models, high-dimensional order search and penalized regression. *Ann. Statist.* **42** 2526–2556. MR3277670 https://doi.org/10.1214/14-AOS1260

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: Methods, Theory and Applications. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

CHICKERING, D. M. (1996). Learning Bayesian networks is NP-complete. In *Learning from Data* (*Fort Lauderdale*, *FL*, 1995). *Lect. Notes Stat.* **112** 121–130. Springer, New York. MR1473013 https://doi.org/10.1007/978-1-4612-2404-4_12

CONLON, L. (2008). *Differentiable Manifolds*, 2nd ed. *Modern Birkhäuser Classics*. Birkhäuser, Boston, MA. MR2413709

COX, D. R. (1990). Role of models in statistical analysis. *Statist. Sci.* **5** 169–174. MR1062575

COX, D., LITTLE, J. and O'SHEA, D. (2007). *Ideals, Varieties, and Algorithms*: An Introduction to Computational Algebraic Geometry and Commutative Algebra, 3rd ed. *Undergraduate Texts in Mathematics*. Springer, New York. MR2290010 https://doi.org/10.1007/978-0-387-35651-8

DRTON, M. (2009a). Likelihood ratio tests and singularities. *Ann. Statist.* **37** 979–1012. MR2502658 https://doi.org/10.1214/07-AOS571

DRTON, M. (2009b). Discrete chain graph models. *Bernoulli* **15** 736–753. MR2555197 https://doi.org/10.3150/08-BEJ172

DRTON, M., STURMFELS, B. and SULLIVANT, S. (2009). *Lectures on Algebraic Statistics*. *Oberwolfach Seminars* **39**. Birkhäuser, Basel. MR2723140 https://doi.org/10.1007/978-3-7643-8905-5

DRTON, M. and SULLIVANT, S. (2007). Algebraic statistical models. *Statist. Sinica* **17** 1273–1297. MR2398596

EVANS, R. J. (2015). Smoothness of marginal log-linear parameterizations. *Electron. J. Stat.* **9** 475–491. MR3326132 https://doi.org/10.1214/15-EJS1009

FERRAROTTI, M., FORTUNA, E. and WILSON, L. (2002). Local approximation of semialgebraic sets. *Ann. Sc. Norm. Super. Pisa Cl. Sci.* (5) **1** 1–11. MR1994799

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

FU, F. and ZHOU, Q. (2013). Learning sparse causal Gaussian networks with experimental intervention: Regularization and coordinate descent. *J. Amer. Statist. Assoc.* **108** 288–300. MR3174620 https://doi.org/10.1080/01621459.2012.754359

GU, J., FU, F. and ZHOU, Q. (2019). Penalized estimation of directed acyclic graphs from discrete data. *Stat. Comput.* **29** 161–176. MR3905546 https://doi.org/10.1007/s11222-018-9801-y

KANG, J. D. Y. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458 https://doi.org/10.1214/07-STS227

LAURITZEN, S. L. and WERMUTH, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17** 31–57. MR0981437 https://doi.org/10.1214/aos/1176347003

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

NARDI, Y. and RINALDO, A. (2012). The log-linear group-lasso estimator and its asymptotic properties. *Bernoulli* **18** 945–974. MR2948908 https://doi.org/10.3150/11-BEJ364

NI, Y., STINGO, F. C. and BALADANDAYUTHAPANI, V. (2015). Bayesian nonlinear model selection for gene regulatory networks. *Biometrics* **71** 585–595. MR3402594 https://doi.org/10.1111/biom.12309

PEARL, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166 https://doi.org/10.1017/CBO9780511803161

RICHARDSON, T. EVANS, J. ROBINS, M. and SHPITSER, I. (2017). Nested Markov properties for acyclic directed mixed graphs. Available at arXiv:1701.06686.

RICHARDSON, T. and SPIRTES, P. (2002). Ancestral graph Markov models. *Ann. Statist.* **30** 962–1030. MR1926166 https://doi.org/10.1214/aos/1031689015

ROCHA, G. V., WANG, X. and YU, B. (2009). Asymptotic distribution and sparsistency for ℓ_1-penalized parametric M-estimators with applications to linear SVM and logistic regression. Preprint. Available at arXiv:0908.1940.

RUDAS, T., BERGSMA, W. P. and NÉMETH, R. (2010). Marginal log-linear parameterization of conditional independence models. *Biometrika* **97** 1006–1012. MR2746171 https://doi.org/10.1093/biomet/asq037

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Rejoinder to "Adjusting for nonignorable drop-out using semiparametric nonresponse models". *J. Amer. Statist. Assoc.* **94** 1135–1146.

SHIMIZU, S., HOYER, P. O., HYVÄRINEN, A. and KERMINEN, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.* **7** 2003–2030. MR2274431

SHOJAIE, A. and MICHAILIDIS, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika* **97** 519–538. MR2672481 https://doi.org/10.1093/biomet/asq038

SHPITSER, I., EVANS, R., RICHARDSON, T. and ROBINS, J. (2013). Sparse nested Markov models with log-linear parameters. In *UAI*-2013 576–585.

SKRONDAL, A. and RABE-HESKETH, S. (2004). *Generalized Latent Variable Modeling*: *Multilevel*, *Longitudinal*, *and Structural Equation Models*. *Interdisciplinary Statistics*. CRC Press/CRC, Boca Raton, FL. MR2059021 https://doi.org/10.1201/9780203489437

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (2000). *Causation*, *Prediction*, *and Search*, 2nd ed. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR1815675

TCHETGEN TCHETGEN, E. J. and SHPITSER, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. *Ann. Statist.* **40** 1816–1845. MR3015045 https://doi.org/10.1214/12-AOS990

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

UHLER, C., RASKUTTI, G., BÜHLMANN, P. and YU, B. (2013). Geometry of the faithfulness assumption in causal inference. *Ann. Statist.* **41** 436–463. MR3099109 https://doi.org/10.1214/12-AOS1080

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

WITTEN, D. M., FRIEDMAN, J. H. and SIMON, N. (2011). New insights and faster computations for the graphical lasso. *J. Comput. Graph. Statist.* **20** 892–900. MR2878953 https://doi.org/10.1198/jcgs.2011.11051a

ZHANG, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence* **172** 1873–1896. MR2459793 https://doi.org/10.1016/j.artint.2008.08.001

ZWIERNIK, P., UHLER, C. and RICHARDS, D. (2017). Maximum likelihood estimation for linear Gaussian covariance models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 1269–1292. MR3689318 https://doi.org/10.1111/rssb.12217