Semidefinite Tests for Latent Causal Structures

Aditya Kela, Kai von Prillwitz, Johan Åberg, Rafael Chaves, and David Gross

***Abstract*— Testing whether a probability distribution is compatible with a given Bayesian network is a fundamental task in the field of causal inference, where Bayesian networks model causal relations. Here we consider the class of causal structures where all correlations between observed quantities are solely due to the influence from latent variables. We show that each model of this type imposes a certain signature on the observable covariance matrix in terms of a particular decomposition into positive semidefinite components. This signature, and thus the underlying hypothetical latent structure, can be tested in a computationally efficient manner via semidefinite programming. This stands in stark contrast with the algebraic geometric tools required if the full observable probability distribution is taken into account. The semidefinite test is compared with tests based on entropic inequalities.**

***Index Terms*— Cause effect analysis.**

I. Introduction

In SPITE of the primal importance of discovering causal relations in science, statistical analysis has historically shied away from causality. Only relatively recently has a rigorous theory of causality emerged (see, for instance, [1], [2]). Since then, causal inference has quickly become influential. Examples range from applications to the inference of genetic [3] and social networks [4], to a better understanding of the role of causality within quantum physics [5]–[13].

To formalize causal mechanisms, it has become popular to use directed acyclic graphs (DAGs) where nodes denote random variables and directed edges (arrows) account for their causal relations. The most common method to determine the set of possible DAGs compatible with a given distribution is based on the Markov condition and the faithfulness assumption [1], [2]. Under these conditions, and in the case where all variables composing a given DAG can be observed, the conditional independences implied by the graph contain all the information required to test for the compatibility with the causal structure. However, for a variety of practical and fundamental reasons, we generally need to deal with latent (hidden) variables, that is, variables that may play an important role in the causal model, but nonetheless cannot be observed. In this case we have to characterize the set of marginal probability distributions that a given DAG can give rise to. Unfortunately, as is widely recognized, generic causal models with latent variables impose highly non-trivial constraints on the possible correlations compatible with it [14]–[27]. It is worth noting that these challenges arise already at the idealized (oracular) level, where we assume that we have access to the probability distribution of the observables.

For distributions over finite sets, as well as for normal distributions, the marginal compatibility can in principle be completely characterized in terms of semi-algebraic sets [16]. However, it appears that the resulting tests in practice are computationally intractable beyond a few variables [18], [22]. One approach to deal with the apparent intractability is to consider relaxations of the original problem, that is, to design tests that define incomplete lists of constraints (outer approximations) to the set of compatible distributions [17]–[20], [28]–[30]. This approach has previously been considered in [29]–[33], with tests based on entropic information theoretic inequalities; an idea originally conceived to tackle foundational questions in quantum mechanics [34]–[40]. Here we consider a relaxation in a similar spirit, but based on covariances rather than entropies. We primarily focus on tests for determining whether a given distribution is compatible with a given DAG, although we briefly discuss how one could turn this into a test also in the statistical sense, when we only have access to finite samples of data.

The effect of latent variables on the observable covariance matrix has previously been characterized in the context of factor analysis, yielding, e.g., tetrad constraints [2], [41]–[44]. There the aim is to decide if observations are compatible with a given number of univariate latent variables, or factors, where each such factor is allowed to influence every observable. These approaches are typically based on linear Gaussian models (although see [45], [46] for generalizations) where the observables are linear functions of the latent variables with added noise, or on the assumption of a global Gaussian distribution (see, e.g., [42]). In contrast, we wish to determine if the observations are compatible with a hypothetical causal structure, irrespective of the nature of the constituent variables, which can be categorical, real-valued, or multivariate, and with no restriction on the type of distribution (as long as the relevant conditional covariances are well defined). Furthermore, we do not assume any particular functional form for how the observables depend on the latent variables. The constraints on the observable covariance matrix do in our case instead

Manuscript received July 17, 2018; revised May 27, 2019; accepted July 29, 2019. Date of publication August 16, 2019; date of current version December 23, 2019. This work was supported in part by the Excellence Initiative of the German Federal and State Governments under Grants ZUK 43 and 81, in part by the ARO under Contract W911NF-14-1-0098 (Quantum Characterization, Verification, and Validation), and in part by the DFG under Grant SPP1798 CoSIP.

A. Kela, J. Åberg, and D. Gross are with the Institute for Theoretical Physics, University of Cologne, 50937 Cologne, Germany.

K. von Prillwitz is with the Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, 26111 Oldenburg, Germany.

R. Chaves is with the International Institute of Physics, Federal University of Rio Grande do Norte, Natal 59070-405, Brazil.

Communicated by E. Abbe, Associate Editor for Machine Learning.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2019.2935755

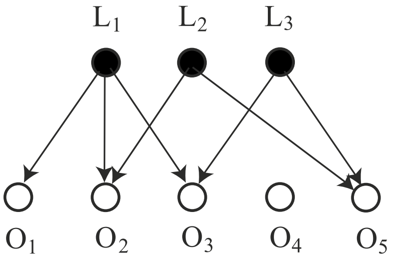

Fig. 1. **Bipartite DAGs.** In this investigation we focus on the class of causal models where all correlations among the observables are due to a collection of independent latent variables. This setting can be described in terms of DAGs that are bipartite, where the latter means that all edges are directed from latent variables (L_1, L_2, L_3) to the observables $(O_1, O_2, O_3, O_4, O_5)$, and where there are no edges within each of these subsets.

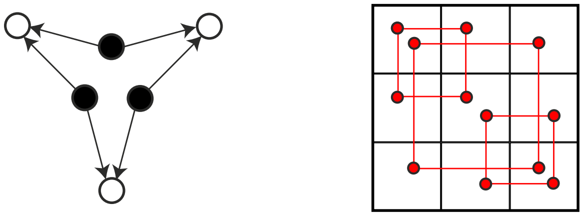

Fig. 2. **Example: Triangular bipartite DAG.** The covariance matrix resulting from the observables in a bipartite DAG is subject to a decomposition where each latent variable gives rise to a positive semidefinite component, and where the support of that component is determined by the children of the corresponding latent variable. In the case of the 'triangular' scenario of the bipartite DAG to the left, each of the three latent variables has two children. The covariance matrix, schematically depicted to the right, can consequently be decomposed into three positive semidefinite components, each with bipartite supports. This observation yields a method (which we refer to as the 'semidefinite test') to falsify a given bipartite DAG as an explanation of an observed covariance matrix.

stem solely from the assumption that the global distribution is compatible with the assumed causal structure, as represented by the given DAG.

Another characterization of the observable covariance matrix has been obtained in [47], where the constraints are derived from the same class of causal structures that we consider (compare Section 6 in [47] with the description of our setup in Section I-A below). In particular, Corollary 5.4 in [47] provides a characterization for the triangular scenario in Figure 2. However, [47] assumes Gaussian linear models, while we, as mentioned above, make no restrictions on the nature of the random variables, beyond the given causal structure.

A. Main Assumptions and Results

We focus on a particular class of latent causal structures, where we assume that there are no direct causal influences between the observables, but only from latent variables to observables (see Figure 1). Hence, all correlations among the observables are due to the latent variables. This setting can be described by the class of DAGs where all edges are directed from latent vertices to observable vertices, but no edges within these two groups (see Figure 1). In other words, we consider the case of DAGs that are bipartite, with the coloring 'observable' and 'latent'. Alternatively, this can be described in terms of hypergraphs, where each independent latent cause is associated with a hyperedge consisting of the affected observable vertices (see e.g. [48]).

This class of graphs has previously been considered in the context of marginalization of Bayesian networks [28], [31], [48], and Gaussian models with latent variables [47]. They moreover provide examples of the difficulties that arise when characterizing latent structures [6], [29], [30], [49]–[51], where standard techniques based on the use of conditional independencies even can yield erroneous results (for a discussion, see e.g. [52]). This type of latent structures furthermore emerges in the context of Bell's theorem [53], as well as in recent generalizations [6], [23], [24], [49]–[51], [54], [55], where they can be used to show that quantum correlations between distant observers–thus without direct causal influences between them–are incompatible with our most basic notions of cause and effect.

Irrespective of the nature of the observables (categorical or continuous) we are free to assign vectors to each possible outcome of the observables. Our main result is to show that each bipartite DAG implies a particular decomposition of the resulting covariance matrix into positive semidefinite components. Hence, we can test whether the observed covariance matrix is compatible with a hypothetical bipartite DAG by checking whether it satisfies the corresponding positive semidefinite decomposition, and we somewhat colloquially refer to this as the 'semidefinite test'. The semidefinite test can thus be phrased as a semidefinite membership problem, which in turn can be solved via semidefinite programming. The latter is known to be computationally efficient from a theoretical point of view, and has a good track record concerning algorithms that are efficient also in practice (see discussions in [56]).

B. Structure of the Paper

In Section II we derive the main result, namely that every bipartite DAG implies a particular semidefinite decomposition of the observable covariance matrix. Section III relates the semidefinite decomposition to previous types of operator inequalities introduced in [57]. To obtain a covariance matrix we may be required to assign vectors to the outcomes of the random variables, and Section IV discusses the dependence of the semidefinite test on this assignment. In Section V we briefly discuss the fact that the compatibility with a given bipartite DAG is not affected if the observables are processed locally, and that the semidefinite test respects this basic property under suitable conditions. Section VI considers a specific class of distributions where it is possible to analytically determine the conditions for a semidefinite decomposition. This class of distributions does in Section VII serve as a testbed for comparisons with the above mentioned entropic tests. We conclude with a summary and outlook in Section VIII.

II. Decomposition of the Covariance Matrix for Bipartite DAGs

We consider a collection of observable variables $O_1, \ldots, O_M$. To each of these variables O_m we associate a mapping $Y^{(m)}$, in some contexts referred to as a 'feature map' [59], into a finite-dimensional vector space $\mathcal{V}_m$. We denote the resulting vector-valued random variables by $Y_m := Y^{(m)}(O_m)$, and for the sake of simplicity we often abuse the terminology and refer to the vectors Y_m themselves as feature maps. We also define the joint random vector $Y := \sum_{m=1}^{M} Y_m$ on $\mathcal{V} := \bigoplus_{m=1}^{M} \mathcal{V}_m$. One should note that while we regard the observable variables O_m as being part of the setup that is 'given', the feature maps $Y^{(m)}$ are part of the analysis, and we are free to assign these as we see fit. (Concerning the question of how the test depends on this choice, see Section IV.) Let P_m denote the projector onto the subspace $\mathcal{V}_m$ in $\mathcal{V}$.

For a vector-valued random variable Y, in a real or complex inner product space $\mathcal{V}$, we define the covariance matrix of Y as $\mathrm{Cov}(Y) := E(YY^{\dagger}) - E(Y)E(Y)^{\dagger}$, where $E(Y)$ denotes the expectation of Y and $\dagger$ denotes the transposition if the underlying vector space is real, and the Hermitian conjugation if the space is complex. We also make use of the cross-covariances $\mathrm{Cov}(Y_m, Y_{m'}) := E(Y_m Y_{m'}^{\dagger}) - E(Y_m)E(Y_{m'})^{\dagger}$, as well as the conditional cross-covariance $\mathrm{Cov}(Y_m, Y_{m'}|X) := E(Y_m Y_{m'}^{\dagger}|X) - E(Y_m|X)E(Y_{m'}|X)^{\dagger}$, with respect to some random variable X. We can thus divide the total covariance matrix $\mathrm{Cov}(Y)$ into the cross-covariances between the separate observable quantities $\mathrm{Cov}(Y) = [\mathrm{Cov}(Y_m, Y_{m'})]_{m,m'=1}^{M}$. Note that $\mathrm{Cov}(Y_m, Y_{m'}) = P_m \mathrm{Cov}(Y) P_{m'}$.

We define a bipartite DAG as a finite DAG $G = (V, E)$ with vertices V and edges E, with a bipartition $V = O \cup L$, $O \cap L = \emptyset$ such that all edges in E are directed from the elements in L (the latent variables) to the elements in O (the observables). Since G is finite, we enumerate the elements of O as $O_1, \ldots, O_M$ and the elements of L as $L_1, \ldots, L_N$. We often overload the notation and let O_m and L_n denote the vertices in the underlying bipartite DAG, as well as denoting the random variables corresponding to these vertices.

For a vertex v in a directed graph G we let $\mathrm{ch}(v)$ denote the children of v, i.e., the set of vertices v' for which there is an edge directed from v to v'. We let $\mathrm{pa}(v)$ denote the parents of v, i.e., the set of vertices v' for which there is an edge directed from v' to v. For bipartite DAGs an element in L can only have children in O (and have no parents), and an element in O can only have parents in L (and no children). As an example, for the bipartite DAG in Figure 1 we have $\mathrm{ch}(L_1) = \{O_1, O_2, O_3\}$, $\mathrm{ch}(L_2) = \{O_2, O_5\}$, and $\mathrm{ch}(L_3) = \{O_3, O_5\}$, and $\mathrm{pa}(O_1) = \{L_1\}$, $\mathrm{pa}(O_2) = \{L_1, L_2\}$, $\mathrm{pa}(O_3) = \{L_1, L_3\}$, $\mathrm{pa}(O_4) = \emptyset$, and $\mathrm{pa}(O_5) = \{L_2, L_3\}$.

For a causal model defined by a general DAG $G = (V, E)$ the underlying probability distribution can be described via the Markov condition where each edge represents a direct causal influence, and thus each vertex v can only be directly influenced by its parents $\mathrm{pa}(v)$, resulting in distributions of the form $P = \Pi_{v \in V} P\big(v \big| \mathrm{pa}(v)\big)$. Hence, for a bipartite DAG we get $P = \Pi_m P\big(O_m \big| \mathrm{pa}(O_m)\big) \Pi_n P(L_n)$, and thus all the latent variables are independent, and the observables are independent when conditioned on the latent variables.

We map the observables $O_1, \ldots, O_M$ to vectors $Y_1, \ldots, Y_M$ in vector spaces $\mathcal{V}_1, \ldots, \mathcal{V}_M$. For each n we define the projector $P^{(n)}$ in $\mathcal{V}$ by

$$P^{(n)} := \sum_{m \in \mathrm{ch}(L_n)} P_m. \tag{1}$$

Hence, $P^{(n)}$ is the projector onto all subspaces of $\mathcal{V}$ that correspond to the children $\mathrm{ch}(L_n)$ of the latent variable L_n. (We often write $m \in \mathrm{ch}(L_n)$ rather than $O_m \in \mathrm{ch}(L_n)$, and $n \in \mathrm{pa}(O_m)$ rather than $L_n \in \mathrm{pa}(O_m)$.)

Proposition 1. *For a bipartite DAG with latent variables $L_1, \ldots, L_N$ and observables $O_1, \ldots, O_M$ with assigned feature maps $Y_1, \ldots, Y_M$ into finite-dimensional real or complex inner-product spaces $\mathcal{V}_1, \ldots, \mathcal{V}_M$, the covariance matrix of $Y = \sum_{m=1}^{M} Y_m$ satisfies*

$$\mathrm{Cov}(Y) = R + \sum_{n=1}^{N} C_n, \quad R \geq 0, \quad C_n \geq 0, \tag{2}$$

where

$$P^{(n)} C_n P^{(n)} = C_n, \quad R = \sum_{m=1}^{M} P_m R P_m. \tag{3}$$

and where the projectors $P^{(n)}$ are as defined in (1) with respect to the given bipartite DAG, and where P_m is the projector onto $\mathcal{V}_m$ in $\bigoplus_{m=1}^{M} \mathcal{V}_m$.

If the span of the supports of $\{P^{(n)}\}_{n=1}^{N}$ covers $\mathcal{V}$, then we can distribute the blocks $P_m R P_m$ of R and add them to the different C_n in such a way that the new operators still are positive semidefinite and satisfy the support structure of the original C_ns. The exception is if there is some observable that has no parent (as O_4 in Figure 1).

Although the operators C_n may change if we generate them via a permutation of the sequence $L_1, \ldots, L_N$, the resulting projectors $P^{(n)}$ would not change. Hence, the support-structure described by (2) and (3) is stable under rearrangements.

Proof: For each $n = 1, \ldots, N$, let $\mathcal{F}_n$ be the σ-algebra spanned by $L_1, \ldots, L_n$. Let $\mathcal{F}_0 = \{\emptyset, \Omega\}$, and let $\mathcal{F}_{N+1}$ be spanned by $L_1, \ldots, L_N, O_1, \ldots, O_M$. For each $m = 1, \ldots, M$ define $X_0^{(m)} := E(Y_m|\mathcal{F}_0) = E(Y_m)$, $X_n^{(m)} := E(Y_m|\mathcal{F}_n) = E(Y_m|L_n, \ldots, L_1)$, for all $n = 1, \ldots, N$, and $X_{N+1}^{(m)} := E(Y_m|\mathcal{F}_{N+1}) = E(Y_m|L_N, \ldots, L_1, O_M, \ldots, O_1) = Y_m$, where the last equality follows from Y_m being a (deterministic) function of O_m. One can confirm that $X_1^{(m)}, \ldots, X_{N+1}^{(m)}$ is a martingale sequence with respect to $\mathcal{F}_0, \mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_N, \mathcal{F}_{N+1}$, i.e., that $E(X_n^{(m)}|\mathcal{F}_{n-1}) = X_{n-1}^{(m)}$, for $n = 1, \ldots, N+1$. Define the martingale difference sequence $\Delta_n^{(m)} := X_n^{(m)} - X_{n-1}^{(m)}$, for $n = 1, \ldots, N+1$. From $\Delta_n^{(m)}$ being a martingale difference sequence, one can confirm that $E(\Delta_n^{(m)} \Delta_{n'}^{(m')^{\dagger}}) = \delta_{n,n'} E(\Delta_n^{(m)} \Delta_n^{(m')^{\dagger}})$. Together with the observation that $Y_m - E(Y_m) = \sum_{n=1}^{N+1} \Delta_n^{(m)}$, this yields

$$\mathrm{Cov}(Y_m, Y_{m'}) = \sum_{n=1}^{N+1} E(\Delta_n^{(m)} \Delta_n^{(m')^{\dagger}}). \tag{4}$$

This gives the decomposition in (2), with the definitions $C_n := [C_n^{m,m'}]_{m,m'=1}^M$, for $n = 1,\dots,N$, and $R := [R^{m,m'}]_{m,m'=1}^M$, where $C_n^{m,m'} := E(\Delta_n^{(m)} \Delta_n^{(m')\dagger})$ and $R^{m,m'} := E(\Delta_{N+1}^{(m)} \Delta_{N+1}^{(m')\dagger})$. If $L_n \notin \mathrm{pa}(O_m)$, then it means that Y_m is conditionally independent of L_n, given $L_{n-1},\dots,L_1$, and thus $\Delta_n^{(m)} = E(Y_m|L_n,\dots,L_1) - E(Y_m|L_{n-1},\dots,L_1) = 0$. The analogous statement is true if $L_n \notin \mathrm{pa}(O_{m'})$. Consequently,

$$C_n^{m,m'} = E(\Delta_n^{(m)} \Delta_n^{(m')\dagger}) = 0,$$
$$\text{if} \quad L_n \notin \mathrm{pa}(O_m) \cap \mathrm{pa}(O_{m'}), \quad n = 1,\dots,N. \quad (5)$$

Note that $L_n \in \mathrm{pa}(O_m) \cap \mathrm{pa}(O_{m'}) \Leftrightarrow O_m, O_{m'} \in \mathrm{ch}(L_n)$. By comparing (5) with the definition of the projector $P^{(n)}$ in (1), we see that $P^{(n)} C_n P^{(n)} = C_n$. By the definition, $C_n^{m,m'} = E(\Delta_n^{(m)} \Delta_n^{(m')\dagger})$, one can also see that $C_n \geq 0$. Next, note that $R^{m,m'} = E(\Delta_{N+1}^{(m)} \Delta_{N+1}^{(m')\dagger}) = E\big(\mathrm{Cov}(Y_m, Y_{m'}|L_1,\dots,L_N)\big)$. By construction, all the observables $O_1,\dots,O_M$ and thus also $Y_1,\dots,Y_M$ are independent when conditioned on all the latent variables. Hence, $R^{m,m'} = \delta_{m,m'} R^{m,m}$, and consequently $R = \sum_m P_m R P_m$. Since $R^{m,m} = E(\Delta_{N+1}^{(m)} \Delta_{N+1}^{(m)\dagger})$ one can see that each $R^{m,m}$, and thus also R, is positive semidefinite. □

Deciding whether a given matrix is of the form (2) can be done via semi-definite programming (SDP). We end this section by describing an explicit SDP formulation. The optimization is over matrices Z, which can be interpreted as the direct sum of candidates for R and the C_n's. More precisely, let

$$\mathcal{Z} \;:=\; \mathcal{V}_1 \oplus \cdots \oplus \mathcal{V}_M \oplus \mathcal{W}_1 \oplus \cdots \oplus \mathcal{W}_N, \quad (6)$$
$$\mathcal{W}_i \;:=\; \bigoplus_{m\in \mathrm{ch}(L_i)} \mathcal{V}_m. \quad (7)$$

Let Z be a matrix on $\mathcal{Z}$. According to the direct sum decomposition (6), the matrix Z is a block matrix with $(M+N)\times(M+N)$ blocks. We think of the fist M diagonal blocks as carrying candidates for $R_m = P_m R P_m$ (which completely defines R, according to (3)); while the rear N diagonal blocks correspond to candidate C_n's. Note that the N rear summands in (6) are direct sums themselves. It therefore makes sense to use double indices to refer to spaces inside the $\mathcal{W}_i$'s. Concretely, the SDP includes affine constraints on the blocks $Z^{(M+n,m),(M+n,m')}$. The first part of the indices selects the space $\mathcal{W}_n$ in (6). The second part refers to the space $\mathcal{V}_m$ within $\mathcal{W}_n$ according to (7). We use the convention that $Z^{(M+n,m),(M+n,m')}$ denotes 0 if either $\mathcal{V}_m$ or $\mathcal{V}_{m'}$ does not occur in $\mathcal{W}_n$.

With these definitions, the semi-definite program that verifies whether a covariance matrix $\mathrm{Cov}(Y)$ is of the form (2) reads

$$\text{maximize} \quad 0 \quad (8)$$
$$\text{subject to} \quad \delta_{m,m'} \sum_{m=1}^{M} Z^{(m),(m)} + \sum_{n=1}^{N} Z^{(M+n,m),(M+n,m')} = \mathrm{Cov}(Y)^{m,m'}, \quad (m,m'=1,\dots M) \quad (9)$$
$$Z \geq 0, \quad (10)$$

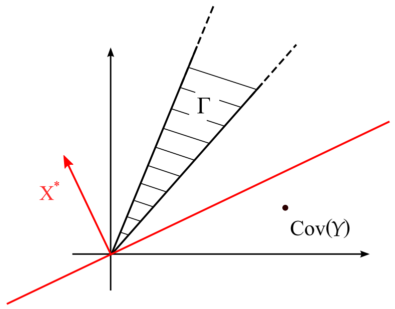

Fig. 3. **Dual Certificates.** The set of covariance matrices compatible with a certain causal structure in the sense of Proposition 1 forms a convex cone Γ. The cone is the feasible set of the SDP (8). If a given covariance matrix $\mathrm{Cov}(Y)$ is *not* an element of that cone, then there exists a hyperplane (depicted in red) separating the two convex sets. A normal vector $X^\star$ for the separating hyperplane can be found using the dual SDP (12).

where the optimization is over symmetric (hermitian) matrices Z on $\mathcal{Z}$. Up to a trivial re-expression of the linear functions of Z in terms of trace inner products with suitable matrices F_i, the optimization problem above is in the (dual) standard form of an SDP [56, Section 3]. The left-hand side of (9) implicitly defines a linear map $\mathcal{A}$ from matrices on $\mathcal{Z}$ to matrices on $\mathcal{V}$. Explicitly, $\mathcal{A}$ maps off-diagonal blocks to 0 and acts on block-diagonal matrices as

$$\mathcal{A} : R_1 \oplus \cdots \oplus R_M \oplus C_1 \oplus \cdots \oplus C_N \mapsto \sum_m R_m + \sum_n C_n.$$

The constraints of the SDP can thus be written slightly more transparently as

$$\mathcal{A}(Z) = \mathrm{Cov}(Y), \quad Z \geq 0. \quad (11)$$

In this language, the dual of the above SDP is

$$\text{minimize} \quad \mathrm{tr}\,\big(X\,\mathrm{Cov}(Y)\big)$$
$$\text{subject to} \quad \mathcal{A}^\dagger(X) \geq 0. \quad (12)$$

Let $X^\star$ be the optimizer of (12). If $\mathrm{tr}\big(X^\star\,\mathrm{Cov}(Y)\big) < 0$, then the original SDP is infeasible and therefore, $\mathrm{Cov}(Y)$ is not of the form (2). Indeed, by construction, such an $X^\star$ has a negative trace inner product with the covariance matrix, but a positive trace inner product $\mathrm{tr}\,\big(\mathcal{A}(Z)X\big) = \mathrm{tr}\,\big(Z\mathcal{A}^\dagger(X)\big) \geq 0$, $\forall Z \geq 0$, with all matrices $\mathcal{A}(Z), Z \geq 0$ that could potentially be feasible for the primal SDP (11). Thus, the dual SDP (12) can be used to find a *witness* or a *dual certificate* $X^\star$ for the incompatibility of a covariance matrix with a presumed causal structure. The geometry of the involved objects is shown in Figure 3. We refer to this dual construction in Section VIII, where we sketch possibilities to base statistical hypothesis tests on such witnesses.

III. Implied Operator Inequalities

Here we show that the existence of positive semidefinite decompositions as in Proposition 1 implies operator inequalities of a type studied in [57]. These operator inequalities yield simplified (but cruder) tests compared to the semidefinite test; by applying the mapping (13) to the covariance matrix, a hypothetical DAG is falsified if the resulting operator fails to be positive semidefinite.

Consider a bipartite DAG with latent variables $L_1, \dots, L_N$ and observables $O_1, \dots, O_M$ with assigned feature maps $Y_1, \dots, Y_M$ into vector spaces $\mathcal{V}_1, \dots, \mathcal{V}_M$. For a number d we define the following map on the space of operators on $\mathcal{V} = \oplus_{m=1}^{M} \mathcal{V}_m$

$$\Phi(Q) := (d-1)P_1 Q P_1 + \sum_{m=2}^{M} (P_m Q P_m + P_1 Q P_m + P_m Q P_1), \tag{13}$$

where P_m are the projectors onto the spaces $\mathcal{V}_m$ defined in Section II. Theorem 4.1 in [57] does in essence say that if all the latent variables L_n in the given bipartite DAG have degree at most d, then the resulting covariance matrix $\mathrm{Cov}(Y)$ satisfies

$$\Phi\big(\mathrm{Cov}(Y)\big) \geq 0, \tag{14}$$

Note that Q being positive semidefinite is not enough to guarantee that $\Phi(Q)$ is positive semidefinite. Hence, (14) can be used as a test of the latent structure. Equations (14) singles out observable 1, but by relabeling we can obtain analogous inequalities for all observables.

The following proposition shows that the semidefinite decomposition implies the operator inequality (14) under the assumption that all the latent variables (regarded as vertices in a bipartite graph) have the degree at most d.

Proposition 2. *For a bipartite DAG with latent variables $L_1, \dots, L_N$, each with degree at most d, and observables $O_1, \dots, O_M$ with assigned feature maps $Y_1, \dots, Y_M$ into finite-dimensional real or complex inner-product spaces $\mathcal{V}_1, \dots, \mathcal{V}_M$, the covariance matrix of $Y = \sum_{m=1}^{M} Y_m$ satisfies $\Phi\big(\mathrm{Cov}(Y)\big) \geq 0$, where Φ is as defined in (13).*

Proof: We know from Proposition 1 that $\mathrm{Cov}(Y) = R + \sum_{n=1}^{N} C_n$ with $P^{(n)} C_n P^{(n)} = C_n$, $C_n \geq 0$, and where R is such that $\sum_m P_m R P_m = R$ and $R \geq 0$. Due to this,

$$\Phi(R) = (1-d)P_1 R P_1 + \sum_{m=2}^{M} P_m R P_m \geq 0. \tag{15}$$

For each C_n we can distinguish two cases. In the first case, C_n has no support on $\mathcal{V}_1$, i.e., $P_1 C_n P_1 = 0$. Due to the positive semidefiniteness of C_n it also follows that $P_1 C_n P_j = 0$ for $j = 2, \dots, M$, and thus $\Phi(C_n) = \sum_{m=2}^{M} P_m C_n P_m \geq 0$. In the second case, C_n does have a support on $\mathcal{V}_1$, meaning that $P_1 C_n P_1 \neq 0$. By assumption, the latent variable L_n has degree at most d, which means that C_n has support on at most d of the subspaces $\mathcal{V}_1, \dots, \mathcal{V}_M$. Hence, apart from $\mathcal{V}_1$, there are at most $d-1$ further spaces involved. We enumerate these spaces as $\mathcal{V}_{m(2)}, \dots, \mathcal{V}_{m(d)}$, and let $\mathcal{V}_{m(1)} = \mathcal{V}_1$. Hence, it may be the case that $P_{m(j)} C_n P_{m(j)} \neq 0$ for $j = 1, \dots, d$, while $P_m C_n P_m = 0$ for the remaining values of m. Due to the positive semidefiniteness of C_n, we can analogously have $P_1 C_n P_{m(j)} \neq 0$, and $P_{m(j)} C_n P_1 \neq 0$, but $P_1 C_n P_m = 0$, and $P_m C_n P_1 = 0$ for the other values of m. We can conclude that

$$\begin{aligned}\Phi(C_n) &= (d-1)P_1 C_n P_1 + \sum_{j=2}^{d} (P_{m(j)} C_n P_{m(j)} \\ &\quad + P_1 C_n P_{m(j)} + P_{m(j)} C_n P_1) \\ &= \sum_{j=2}^{d} \Big(P_1 + P_{m(j)}\Big) C_n \Big(P_1 + P_{m(j)}\Big) \geq 0. \end{aligned} \tag{16}$$

The combination of (15) with the above cases yields $\Phi\big(\mathrm{Cov}(Y)\big) = \Phi(R) + \sum_{n=1}^{N} \Phi(C_n) \geq 0$, which proves (14). □

IV. Universal Feature Maps for Finite Categorical Variables

The semi-definite test depends on the choice of feature maps $Y^{(m)}$. Hence, if a particular choice results in compatibility, there may still be another choice that yields a violation. However, in the case of observables with only finite number of outcomes, we shall here see that one can make a single test. For an observable O with a finite number of possible outcomes $o_1, \dots, o_d$, we refer to a feature map Y as 'universal' if its components $y_1, \dots, y_d$ are linearly independent.

Lemma 1. *Let $\tilde{Y} = (\tilde{y}_1, \dots, \tilde{y}_d)$ be an arbitrary feature map and $Y = (y_1, \dots, y_d)$ be a universal feature map on a vector space $\mathcal{V}$. Then there exists a linear map $\phi_{\mathcal{V}} \to \mathcal{V}$; such that $\tilde{y}_k = \phi y_k$.*

Proof: Let $G := [G_{j,j'}]_{j,j'=1}^{d_m}$ with $G_{j,j'} := (y_j^m, y_{j'}^m)$. G is invertible since $(y_1, \dots, y_d)$ is linearly independent. One can confirm that ϕ defined by $\phi(v) := \sum_{jj'} \tilde{y}_j [G^{-1}]_{jj'} (y_{j'}, v)$ satisfies $\phi y_j = \tilde{y}_j$. □

A direct consequence of this lemma is the following.

Lemma 2. *Let $O_1, \dots, O_M$ be observables with finite number of outcomes. Let $Y_1, \dots, Y_M$ be universal feature maps on the spaces $\mathcal{V}_1, \dots, \mathcal{V}_M$, and let $\tilde{Y}_1, \dots, \tilde{Y}_M$ be feature maps on the same spaces. Then there exists a linear map ϕ on $\bigoplus_{m=1}^{M} \mathcal{V}_m$, such that $\mathrm{Cov}(\tilde{Y}) = \phi \mathrm{Cov}(Y) \phi^{\dagger}$, where $Y := \sum_{m=1}^{M} Y_m$ and $\tilde{Y} := \sum_{m=1}^{M} \tilde{Y}_m$.*

We can conclude that if $\mathrm{Cov}(Y)$ satisfies the decomposition in Proposition 1 for a given bipartite DAG, then $\mathrm{Cov}(\tilde{Y})$ also satisfies the decomposition. Hence, if $\mathrm{Cov}(\tilde{Y})$ fails to satisfy the decomposition, then $\mathrm{Cov}(Y)$ also fails to satisfy it.

V. Monotonicity Under Local Operations

Suppose that we would process each observable variable in a collection $O_1, \dots, O_M$ 'locally', i.e., the output $\tilde{O}_m$ is a (possibly random) function only of O_m. If we restrict to discrete random variables, then this type of mapping from an input distribution P^M of the $O_1, \dots, O_M$, to the output distribution $\tilde{P}^M$ of $\tilde{O}_1, \dots, \tilde{O}_M$ can be written

$$\begin{aligned}&\tilde{P}^M(\tilde{x}_1, \dots, \tilde{x}_M) := \\ &\sum_{x_1 \dots, x_M} P^1(\tilde{x}_1|x_1) \cdots P^M(\tilde{x}_M|x_M) P^M(x_1, \dots, x_M), \end{aligned} \tag{17}$$

where all $P^m(\tilde{x}_m|x_m)$ are conditional distributions. Since these local operations do not change the structure of the underlying bipartite DAG, we can immediately conclude the following.

Proposition 3. *If a distribution P^M over $O_1, \dots, O_M$ is compatible with the given bipartite DAG, then the distribution $\tilde{P}^M$ on $\tilde{O}_1, \dots, \tilde{O}_M$ as defined by (17) is also compatible with the same DAG.*

Another way to phrase this is that compatibility with a DAG is monotone with respect to local operations. There is no reason to expect that relaxations of the compatibility problem would respect this monotonicity. However, the following proposition shows that the semidefinite test also is monotonous, if the test is based on universal feature maps.

Proposition 4. *Let $O_1, \dots, O_M$ and $\tilde{O}_1, \dots, \tilde{O}_M$ be random variables on finite alphabets, that are related by local operations as in (17). Let Y_m be universal feature maps assigned to O_m on space $\mathcal{V}_m$, and $\tilde{Y}_m$ be feature maps assigned to $\tilde{O}_m$ on $\tilde{\mathcal{V}}_m$, for $m = 1, \dots, M$. Let $Y := \sum_{m=1}^M Y_m$ and $\tilde{Y} := \sum_{m=1}^M \tilde{Y}_m$. If $\mathrm{Cov}(Y)$ satisfies the decomposition (2) in Proposition 1 for a given bipartite DAG, then $\mathrm{Cov}(\tilde{Y})$ also satisfies the decomposition.*

Proof: Let $y_1^m, \dots, y_K^m$ be the components of Y_m, and $\tilde{y}_1^m, \dots, \tilde{y}_L^m$ be the components of $\tilde{Y}_m$. $G = [G_{x,x'}]_{x,x'=1}^K$, with $G_{x,x'} = (y_x^m, y_{x'}^m)$, is invertible since $y_1^m, \dots, y_K^m$ are linearly independent. Define $\psi_m(v) := \sum_{\tilde{x},x',x''} \tilde{y}_{\tilde{x}} P^m(\tilde{x}|x')[G^{-1}]_{x',x''}(y_{x''}, v)$. (We omit the superscript 'm' on the vectors y.) With $\psi := \sum_m \psi_m$, we get $E(\tilde{Y}) = \psi\big(E(Y)\big)$. Moreover,

$$\begin{aligned}\mathrm{Cov}(\tilde{Y}_m) &= \psi_m \mathrm{Cov}(Y_m)\psi_m^\dagger + W_m, \\ W_m &:= \sum_{\tilde{x},x} \tilde{y}_{\tilde{x}} \tilde{y}_{\tilde{x}}^\dagger P^m(\tilde{x}|x) P(O_m = x) \\ &\quad - \sum_{\tilde{x},\tilde{x}',x} \tilde{y}_{\tilde{x}} \tilde{y}_{\tilde{x}'}^\dagger P^m(\tilde{x}|x) P^m(\tilde{x}'|x) P(O_m = x).\end{aligned} \tag{18}$$

Let $c \in \tilde{\mathcal{V}}_m$, and define $z_{\tilde{x}} = (c, \tilde{y}_{\tilde{x}})$. Then

$$\begin{aligned}(c, W_m c) =& \sum_{x,\tilde{x}} P(O_m = x) P^m(\tilde{x}|x) \\ &\times \Big| z_{\tilde{x}} - \sum_{\tilde{x}'} P^m(\tilde{x}'|x) z_{\tilde{x}'} \Big|^2 \geq 0.\end{aligned}$$

Hence, $W_m \geq 0$. Moreover, each W_m is supported only on the subspace $\tilde{\mathcal{V}}_m$. If $\mathrm{Cov}(Y)$ satisfies the decomposition (2) in Proposition 1 for some bipartite DAG, then it follows that $\psi \mathrm{Cov}(Y) \psi^\dagger$ also satisfies the corresponding decomposition with respect to the subspaces $\{\tilde{\mathcal{V}}_m\}_m$. Moreover, since the correction terms W_m are positive semidefinite and block-diagonal with respect to these subspaces, it follows that $\mathrm{Cov}(\tilde{Y}) = \psi \mathrm{Cov}(Y) \psi^\dagger + \sum_m W_m$ also satisfies the decomposition. □

VI. A Monotone Family of Distributions

Here we consider a specific family of multi-partite distributions that is monotone in the sense of the previous section. In particular, we consider the case of the triangular scenario in Figure 2, which turns out to be convenient for the comparison with the entropic tests, which we consider in Section VII.

Suppose that we have a collection of variables, each of which has $D \geq 2$ possible outcomes. In equation (17) we described local operations transforming an initial distribution P^M. Here we consider a specific class of such local operations, parametrized by $p \in [0, 1]$

$$P_p(\tilde{x}|x) := (1-p)\delta_{\tilde{x},x} + p\frac{1}{D}. \tag{19}$$

Hence, on each variable we (independently) apply the same type of process, where with probability p we replace the input with a uniformly distributed output, and with probability $1-p$ leave the input intact. Here we choose the input distribution $P^M(x_1, \dots, x_M) = \delta_{x_1,\dots,x_M}/D$, where $\delta_{x_1,\dots,x_M} = 1$ if $x_1 = \dots = x_M$, while zero otherwise. By applying (17) with the local operations (19) we thus obtain the global distribution

$$\begin{aligned}\tilde{P}_p^{M:D}(\tilde{x}_1, \dots, \tilde{x}_M) :=& \frac{1}{D} \sum_{x_1,\dots,x_M} P_p(\tilde{x}_1|x_1) \cdots \\ & \cdots P_p(\tilde{x}_M|x_M) \delta_{x_1,\dots,x_M},\end{aligned} \tag{20}$$

where the extra superscript D indicates the alphabet size of the local random variables. Since $\tilde{P}_1^{M:D}$ is a product distribution over all the observable variables, it is compatible with every bipartite DAG, while $\tilde{P}_0^{M:D}$ is perfectly correlated, and is thus only compatible with bipartite DAGs where some latent variable has edges to all observable variables.

Corollary 1. *For a given bipartite DAG G, and numbers M and $D \geq 2$, there exists a $0 \leq p^* \leq 1$ such that $P_p^{M:D}$ is compatible with G for all $p > p^*$ and incompatible with G for all $p < p^*$. Moreover, there exists a $0 \leq \overline{p} \leq 1$, such that $P_p^{M:D}$ is compatible with the semidefinite test, with respect to universal feature maps, for all $p > \overline{p}$ and incompatible for all $p < \overline{p}$. Moreover, $\overline{p} \leq p^*$.*

Proof: The local operations in (19) are such that if $1 \geq p' \geq p \geq 0$, then there exists a $1 \geq q \geq 0$ such that $P_{p'}(\tilde{x}|x) = \sum_{x'} P_q(\tilde{x}|x') P_p(x'|x)$. Here, any $1 \geq q \geq 0$ is a valid choice if $p = 1$, while $q = (p'-p)/(1-p)$ if $1 > p \geq 0$. Consequently, if $p' \geq p$, then $\tilde{P}_{p'}^{M:D}$ can be generated from $\tilde{P}_p^{M:D}$ by local operations. Let p^* be the infimum of all p such that $P_p^{M:D}$ is compatible with the given DAG. By Proposition 3 it follows that $P_p^{M:D}$ is compatible with the given DAG for all $p > p^*$. Similarly, let $\overline{p}$ be the infimum of all p for which $P_p^{M:D}$ is compatible with the semidefinite test. By Proposition 4, it follows that $P_p^{M:D}$ is compatible with the semidefinite test for all $p > \overline{p}$. Since a distribution that is compatible with a given DAG also satisfies the semidefinite test, it follows that $\overline{p} \leq p^*$. □

Note that Corollary 1 does not specify the compatibility with respect to the DAG and the semidefinite test at $p = p^*$ and $p = \overline{p}$, respectively. One might speculate that generally $p^* = 1$. However, for $\tilde{P}_p^{3:2}$ and the triangular scenario in Figure 2, it is the case that $p^* \leq 1/2$. To see this, let the three parties possess the variables (x_{1a}, x_{1b}), (x_{2a}, x_{2b}) and (x_{3a}, x_{3b}), with joint distribution $P(x_{1a}, x_{1b}, x_{2a}, x_{2b}, x_{3a}, x_{3b}) = \delta_{x_{1a},x_{2b}} \delta_{x_{2a},x_{3b}} \delta_{x_{3a},x_{1b}}/8$, for $x_{ka}, x_{kb} = 1, 2$, which manifestly satisfies the triangular scenario. If each party applies the local operation $P(\tilde{x}_k|x_{ka}, x_{kb}) = \delta_{\tilde{x}_k, x_{ka}}\delta_{x_{ka}, x_{kb}} + (1 - \delta_{x_{ka}, x_{kb}})/2$, for $k = 1, 2, 3$, then one can confirm that the resulting distribution is $\tilde{P}^{3:2}_{1/2}$. Hence, by Proposition 3, it follows that $\tilde{P}^{3:2}_{1/2}$ is compatible with the triangular DAG, and thus $p^* \leq 1/2$.

In the following we determine $\overline{p}$ for the triangular scenario in Figure 2. The family of distributions $\tilde{P}^{M:D}_p$, defined in (20), in the tripartite case becomes

$$\begin{aligned}\tilde{P}^{3:D}_p(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3) =&(1 - p)^3\frac{1}{D}\delta_{\tilde{x}_1, \tilde{x}_2, \tilde{x}_3}\\ &+p(1 - p)^2\frac{1}{D^2}[\delta_{\tilde{x}_1, \tilde{x}_2} + \delta_{\tilde{x}_1, \tilde{x}_3} + \delta_{\tilde{x}_2, \tilde{x}_3}]\\ &+p^2(3 - 2p)\frac{1}{D^3}. \end{aligned} \tag{21}$$

For the construction of the covariance matrix, we assume feature maps Y_1, Y_2, Y_3 that have orthonormal components. Hence, the total space $\mathcal{V} = \mathcal{V}_1 \oplus \mathcal{V}_2 \oplus \mathcal{V}_3$ is $3D$-dimensional, and we can write it as a tensor product $\mathcal{V} = \mathcal{V}^D \otimes \mathcal{L}$ of a D-dimensional space $\mathcal{V}^D$ and a 3-dimensional space $\mathcal{L}$. By choosing an orthonormal basis $\{e_m\}_{m=1}^3$ of $\mathcal{L}$, we can identify $\mathcal{V}_m = \mathcal{V}^D \otimes \mathrm{Sp}\{e_m\}$. In Section II we defined the projectors P_m onto the subspaces $\mathcal{V}_m$, and we can write these projectors as

$$P_m = \hat{1}_D \otimes \tilde{P}_m, \tag{22}$$

where $\hat{1}_D$ is the identity operator on $\mathcal{V}^D$, and $\tilde{P}_m$ is the projector onto e_m. The projector $P^{(n)}$ in Proposition 1 can thus be written

$$P^{(n)} = I_D \otimes \tilde{P}^{(n)}, \quad \tilde{P}^{(n)} = \sum_{m \in \mathrm{ch}(L_n)} \tilde{P}_m. \tag{23}$$

The covariance matrix $\mathrm{Cov}(Y)$ for the random variable $Y = Y_1 + Y_2 + Y_3$ is a $3D \times 3D$ matrix and takes the form

$$\mathrm{Cov}(Y) = \frac{1}{D}Q \otimes C(p), \tag{24}$$

where we define the 3×3 matrix $C(p)$ with components

$$C(p)_{mm'} := [1 - (1 - p)^2]\delta_{mm'} + (1 - p)^2, \tag{25}$$

and the $D \times D$ matrix Q with elements

$$Q_{\tilde{x}, \tilde{x}'} := \delta_{\tilde{x}, \tilde{x}'} - \frac{1}{D}, \quad \tilde{x}, \tilde{x}' = 1, \ldots, D. \tag{26}$$

Note that $Q = \hat{1}_D - cc^\dagger$, where $c = (1, \ldots, 1)^\dagger/\sqrt{D} \in \mathcal{V}^D$ is normalized. Hence, Q is projector onto a $(D-1)$-dimensional subspace. From Q being a projector, it follows that $Q \geq 0$.

Lemma 3. *For* $p \in \mathbb{R}$ *it is the case that* $\left[\begin{smallmatrix} \frac{1}{2} & (1-p)^2 \\ (1-p)^2 & \frac{1}{2} \end{smallmatrix}\right] \geq 0$ $\Leftrightarrow 1 - \frac{1}{\sqrt{2}} \leq p \leq 1 + \frac{1}{\sqrt{2}}$.

Lemma 4. *Let* $a, b, r \in \mathbb{C}$, *then* $\left[\begin{smallmatrix} a & r \\ r & b \end{smallmatrix}\right] \geq 0 \Leftrightarrow \left[\begin{smallmatrix} b & r \\ r & a \end{smallmatrix}\right] \geq 0$.

Proposition 5. *For the family* $\tilde{P}^{3:D}_p$ *in equation (21), with* $D \geq 2$, *and for feature maps with orthonormal components, the covariance matrix* $\mathrm{Cov}(Y)$ *has a semidefinite decomposition with respect to the bipartite DAG in Figure 2, if and only if* $1 - 1/\sqrt{2} \leq p \leq 1$. *Hence,* $\overline{p} = 1 - 1/\sqrt{2}$.

Proof: We begin by showing that $1 - 1/\sqrt{2} \leq p \leq 1$ is a sufficient condition for $\mathrm{Cov}(Y)$ satisfying the semidefinite decomposition. Define the matrices

$$\begin{aligned}\tilde{C}_1 &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{2} & (1-p)^2 \\ 0 & (1-p)^2 & \frac{1}{2} \end{bmatrix}, \quad \tilde{C}_2 = \begin{bmatrix} \frac{1}{2} & 0 & (1-p)^2 \\ 0 & 0 & 0 \\ (1-p)^2 & 0 & \frac{1}{2} \end{bmatrix}, \\ \tilde{C}_3 &= \begin{bmatrix} \frac{1}{2} & (1-p)^2 & 0 \\ (1-p)^2 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \tilde{R} = 0. \end{aligned} \tag{27}$$

By Lemma 3 it follows that these matrices are positive semidefinite for $1 - 1/\sqrt{2} \leq p \leq 1$. Define $R := Q \otimes \tilde{R}/D$ and $C_n := Q \otimes \tilde{C}_n/D$, for Q as in (26). One can confirm that

$$\begin{aligned} R + \sum_n C_n &= \frac{1}{D}Q \otimes \tilde{R} + \sum_n \frac{1}{D}Q \otimes \tilde{C}_n \\ &= \frac{1}{D}Q \otimes C(p) = \mathrm{Cov}(Y). \end{aligned}$$

Moreover, $R \geq 0$ and $C_n \geq 0$. Since $R = 0$, we have $\sum_m P_m R P_m = R$. Moreover, $P^{(n)} C_n P^{(n)} = Q \otimes \tilde{P}^{(n)} \tilde{C}_n \tilde{P}^{(n)}/D$. By inspection of (27), one can confirm that $\tilde{P}^{(n)} \tilde{C}_n \tilde{P}^{(n)} = \tilde{C}_n$.

Next we prove that $1 - 1/\sqrt{2} \leq p \leq 1$ is a necessary condition for $\mathrm{Cov}(Y)$ to satisfy the semidefinite decomposition. Thus assume that there exists a decomposition of $\mathrm{Cov}(Y)$ as in (2) and (3), which provide R and C_n. Let $v \in \mathcal{V}^D$ be normalized, and such that $Qv = v$. Such v always exists, since Q is a projector onto a $(D-1)$-dimensional subspace of $\mathcal{V}^D$ and $D \geq 2$. Define $\overline{R} := Dv^\dagger R v \geq 0$ and $\overline{C}_n := Dv^\dagger C_n v \geq 0$. Hence, by (2), (3) and (24) it follows that $\overline{R} + \sum_n \overline{C}_n = v^\dagger Q v C(p) = C(p)$, with $C(p)$ as in (25). By the conditions in (2), (3) and (23), it follows that $\tilde{P}^{(n)} \overline{C}_n \tilde{P}^{(n)} = \overline{C}_n$. Moreover, (2), (3) and (22) yields $\sum_m \tilde{P}_m \overline{R} \tilde{P}_m = \overline{R}$. This means that $\overline{R}$ is a diagonal matrix, and without loss of generality, we can find new positive semidefinite matrices $\widetilde{C}_n$, such that $\sum_n \widetilde{C}_n = \overline{R} + \sum_n \overline{C}_n = C(p)$ and $\tilde{P}^{(n)} \widetilde{C}_n \tilde{P}^{(n)} = \widetilde{C}_n$. The last condition, together with $\widetilde{C}_n \geq 0$, yields that the most general decomposition $\widetilde{C}_1 + \widetilde{C}_2 + \widetilde{C}_3 = C(p)$ possible, is of the form

$$\begin{aligned}\widetilde{C}_1 &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & b_2 & (1-p')^2 \\ 0 & (1-p')^2 & c_1 \end{bmatrix}, \quad \widetilde{C}_2 = \begin{bmatrix} a_1 & 0 & (1-p')^2 \\ 0 & 0 & 0 \\ (1-p')^2 & 0 & c_2 \end{bmatrix}, \\ \widetilde{C}_3 &= \begin{bmatrix} a_2 & (1-p')^2 & 0 \\ (1-p')^2 & b_1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \end{aligned}$$

where $a_1, a_2, b_1, b_2, c_1, c_2 \geq 0$ and $a_1 + a_2 = 1$, $b_1 + b_2 = 1$, $c_1 + c_2 = 1$. By the positive semidefiniteness of these matrices, follow the positive semidefiniteness of the following matrices,

$$\begin{aligned} M_1 &:= \begin{bmatrix} a_1 & (1-p')^2 \\ (1-p')^2 & c_2 \end{bmatrix} \geq 0, \quad M_2 := \begin{bmatrix} a_2 & (1-p')^2 \\ (1-p')^2 & b_1 \end{bmatrix} \geq 0, \\ M_3 &:= \begin{bmatrix} b_2 & (1-p')^2 \\ (1-p')^2 & c_1 \end{bmatrix} \geq 0. \end{aligned} \tag{28}$$

By Lemma 4 it follows that (28) implies

$$\begin{aligned} M_4 &:= \begin{bmatrix} c_2 & (1-p')^2 \\ (1-p')^2 & a_1 \end{bmatrix} \geq 0, \quad M_5 := \begin{bmatrix} b_1 & (1-p')^2 \\ (1-p')^2 & a_2 \end{bmatrix} \geq 0, \\ M_6 &:= \begin{bmatrix} c_1 & (1-p')^2 \\ (1-p')^2 & b_2 \end{bmatrix} \geq 0. \end{aligned}$$

Since these matrices all are positive semidefinite, it follows that every convex combination is also positive semidefinite.

Thus

$$\frac{1}{6}M_1 + \frac{1}{6}M_2 + \frac{1}{6}M_3 + \frac{1}{6}M_4 + \frac{1}{6}M_5 + \frac{1}{6}M_6 \\ = \begin{bmatrix} \frac{1}{2} & (1-p')^2 \\ (1-p')^2 & \frac{1}{2} \end{bmatrix} \tag{29}$$

is positive semidefinite. By Lemma 3 it follows that we must have $1 - 1/\sqrt{2} \leq p \leq 1$. □

VII. Comparison With Entropic Tests

Outer relaxations of the compatibility set of latent structures, based on information theoretic inequalities, have been considered previously [29]–[33]. Here, we compare numerically the performance of these entropic tests with the semidefinite test. A challenge is that we in practice do not know the true set of compatible distributions. However, since we are dealing with outer approximations, a reasonable approach is to compare how 'strict' the tests are, i.e., if one test generally tends to reject more distributions than the other. Given the rather radical difference in appearance of the semidefinite test and the entropic tests, it is not clear if there is a clear-cut relation between them, in the sense that one would be systematically stronger than the other. In [57] it was found that tests based on operator inequalities, of the type described in Section III, appear to be stronger than the entropic ones for small alphabet sizes, but that there seems to be a switchover for larger alphabets (see Section 4.5 of [57]). Here we confirm similar trends.

A. *Entropy Inequalities for the Triangular DAG*

We focus on the triangular DAG in Figure 2, since this is a rather well investigated scenario with several entropic inequalities. For the three observables O_1, O_2, O_3, let $H(1) := H(O_1) := -\sum_j P(O_1 = j)\log_2 P(O_1 = j)$ denote the Shannon entropy, and in a similar manner $H(12) := H(O_1, O_2)$, etc., where '$\log_2$' denotes the base 2 logarithm. The first inequality (30) for the triangular scenario was obtained in [6] (see also [29] and [32])

$$E_1 := -H(1) - H(2) - H(3) + H(13) + H(12) \geq 0. \tag{30}$$

The following two inequalities were derived in [29]

$$E_2 := -3H(1) - 3H(2) - 3H(3) \\ +2H(12) + 2H(13) + 3H(23) - H(123) \geq 0, \tag{31}$$

$$E_3 := -5H(1) - 5H(2) - 5H(3) \\ +4H(12) + 4H(13) + 4H(23) - 2H(123) \geq 0. \tag{32}$$

Finally, inequalities (33) to (35) were obtained in [32] (explicitly as equation (23) in arXiv version 2 of [32]).

$$E_4 := -4H(1) - 4H(2) - 4H(3) \\ +3H(12) + 3H(13) + 4H(23) - 2H(123) \geq 0, \tag{33}$$

$$E_5 := -2H(1) - 2H(2) - 2H(3) \\ +3H(12) + 3H(13) + 3H(23) - 4H(123) \geq 0, \tag{34}$$

$$E_6 := -8H(1) - 8H(2) - 8H(3) \\ +7H(12) + 7H(13) + 7H(23) - 5H(123) \geq 0. \tag{35}$$

The expressions in (30), (31), and (33) are not symmetric under permutations of the O_1, O_2, O_3, and thus each of these generate two more inequalities. Whenever one of these inequalities is violated we can conclude that the observable distribution cannot originate from the bipartite DAG in Figure 2. All of these entropic inequalities, apart from (30), depend on the full tripartite distribution, while the semidefinite test only uses the mono- and bipartite marginals. One may thus suspect that the semidefinite test would be at a disadvantage compared to these tripartite entropic tests.

B. *Rejection Rates in Random Ising Models: The Binary Case*

For a numerical comparison between the entropic and the semidefinite test for the triangular scenario in Figure 2, we assume binary variables $O_1, O_2, O_3 \in \{-1, 1\}$, and distributions $P(\overline{x}) := P(O_1 = x_1, O_2 = x_2, O_3 = x_3)$, $\overline{x} := (x_1, x_2, x_3)$ given by an Ising interaction model [60], [61], $P(\overline{x}) = e^{-\overline{x}^{\dagger}J\overline{x}}/Z$, with Z being the normalization constant, and where J is a real 3×3 matrix. For each single instance of this model we draw the elements of J independently from a Gaussian distribution with zero mean and variance 1. For the semidefinite test we choose (universal) feature maps that associate the outcomes of the random variables to elements of orthonormal bases, thus resulting in a 6×6 covariance matrix. The semidefinite test was implemented via a semidefinite program that minimizes a constant function, thus effectively testing whether there exist any feasible elements. For each instance over 10^6 independent repetitions of the Ising model, we performed the semidefinite test, as well as tested the entropic inequalities (30) to (35) together with all their permutations. The following table gives the approximate fraction of rejections. In the table, $E_1^{\cup}$ (and analogously for $E_2^{\cup}$ and $E_4^{\cup}$) means that we test the inequality in (30) as well as its two permutations, and we count the fraction of the sample that violates any of these three inequalities, i.e., we take the union of the corresponding rejection regions. The entry 'Combined' signifies the fraction of rejections due to violations of at least one of the inequalities (30) to (35) or any of their permutations. Finally 'Semidefinite' denotes the fraction of rejections for the semidefinite test.

$$E_1^{\cup} : 0.57, \quad E_2^{\cup} : 0.60, \quad E_3 : 0.54,$$
$$E_4^{\cup} : 0.63, \quad E_5 : 0.40, \quad E_6 : 0.60,$$

Combined : 0.64,

Semidefinite : 0.77

Since the fraction of rejections is higher for the semidefinite test than for all the entropic inequalities combined, this suggests that the semidefinite test in some sense has a 'larger' region of rejection, and thus would be the stronger test. To get some information on the relation between the two regions of rejections, we checked whether we could find any case where the semidefinite test accepted an instance that had been rejected by some of the entropic inequalities. However, we could find no such case, which suggests that the region of rejection for the collection of entropic inequalities is contained in the region of rejection for the semidefinite test.

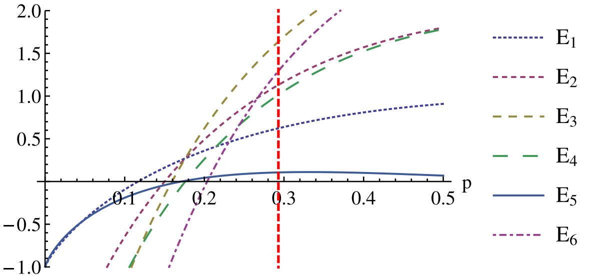

Fig. 4. **Entropic versus semidefinite for binary variables.** For three binary variables with the distribution $\tilde{P}_p^{3:2}$ in (21), we calculate $E_1,\ldots,E_6$ defined in (30) to (35) as functions of the parameter p. When one of these functions turns negative, it implies that the distribution $\tilde{P}_p^{3:2}$ is not compatible with the triangular bipartite DAG in Figure 2. Moreover, we determine the 6×6 covariance matrix with respect to feature maps that assign orthogonal vectors to the outcomes. The red vertical line indicates the value $p = 1-1/\sqrt{2}\approx 0.29$, determined in Proposition 5, below which the semidefinite test rejects the resulting covariance matrix. As one can see, the semidefinite test has the larger region of rejection, and is in this sense the stronger test for this particular binary setup.

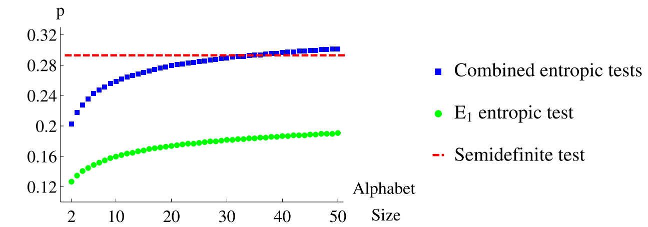

Fig. 5. **Entropic versus semidefinite tests for increasing alphabet sizes.** For the distribution $\tilde{P}_p^{3:D}$ in (21) we compare the entropic and semidefinite test as functions of D. We determine the smallest values of p for which each test accepts $\tilde{P}_p^{3:D}$, as a function of the local alphabet size D. By Proposition 5 we know that the transition point for the semidefinite test is $p = 1-1/\sqrt{2}\approx 0.29$, independently of D (the red dashed line). We also plot (blue squares) the minimal value of p for which all of the entropic inequalities (30) to (35) are satisfied, as a function of D. The transition point for this entropic test crosses the red line at $D = 32$. Hence, for the class of functions $\tilde{P}_p^{3:D}$, the entropic tests becomes stronger than the semidefinite test for alphabet sizes beyond 32. Finally, we plot (green circles) the minimal value of p for which $E_1(p)\geq 0$, as a function of D. By Section VII-C.2 we know that this transition point asymptotically reaches $1-1/\sqrt{2}$.

C. Comparison on a Monotone Family of Distributions

Here we compare the performance of the entropic tests with the semidefinite test on the distributions $\tilde{P}_p^{3:D}$ defined in (21).

1) Binary Variables: In the case of three binary variables, $\tilde{P}_p^{3:2}(\tilde{x}_1,\tilde{x}_2,\tilde{x}_3) = (4-6p+3p^2)/2$ if $\tilde{x}_1=\tilde{x}_2=\tilde{x}_2$, while otherwise $\tilde{P}_p^{3:2}(\tilde{x}_1,\tilde{x}_2,\tilde{x}_3) = p(2-p)/8$. In Figure 4 we plot $E_1,\ldots,E_6$ as functions of the parameter p. The entropic test rejects the model for a given p whenever one of these functions become negative. For the calculation of the covariance matrix we choose feature maps that assign orthonormal vectors to the outcomes of the three random variables, thus being universal. As one can see from Figure 4, the semidefinite test starts to reject at higher values of p than all the entropic tests, and is thus closer to the true value of the transition point p^* (in Corollary 1) than any of the entropic tests.

2) Asymptotics of the E_1 Test: E_1 defined in (30) is the only of the entropic quantities (30) to (35) that solely includes mono- and bipartite marginals. Since the test based on E_1 and the semidefinite test thus are on 'equal footing' in this regard, we compare these two tests further. By Proposition 5 we know that $\tilde{P}_p^{3:D}$ satisfies the semidefinite test if and only if $p\geq 1-1/\sqrt{2}$, irrespective of the alphabet size D. Hence, the 'transition point' for the semidefinite test is independent of D for this family of distributions. Here we show that the transition point for the test based on E_1 lies below $1-1/\sqrt{2}$, but asymptotically approaches this value as D increases. For $\tilde{P}_p^{3:D}$ in (21) we get

$$\begin{aligned}E_1 =&-3\log D\\&-2(1-\frac{1}{D})p(2-p)\log\Big[p(2-p)\frac{1}{D^2}\Big]\\&-2\Big[(1-p)^2+\frac{p(2-p)}{D}\Big]\log\Big[\frac{(1-p)^2}{D}+\frac{p(2-p)}{D^2}\Big].\end{aligned}\tag{36}$$

One can confirm that $E_1(0) = -\log D$, $E_1(1) = \log D$, and

$$\frac{dE_1}{dp} = 4(1-\frac{1}{D})(1-p)\log\Big[1+D\frac{(1-p)^2}{p(2-p)}\Big],\tag{37}$$

which is non-negative for $0\leq p\leq 1$. Hence, for each fixed D, the function E_1 is monotonically increasing for $0\leq p\leq 1$, and thus the equation $E_1(p) = 0$ has exactly one root, which is situated somewhere in the open interval $(0,1)$. Thus, analogous to the semidefinite test, the test based on E_1 rejects all elements in the family $\tilde{P}_p^{3:D}$ below a certain transition point, and accept all distributions above that value. Next one can confirm that

$$E_1(1-\frac{1}{\sqrt{2}}) = \log\frac{2D}{D+1}+\frac{1}{D}\log\frac{2^D}{D+1} > 0,\quad D = 2,3,\ldots.$$

Since E_1 is monotonously increasing with respect to p, we can conclude that the root $\tilde{p}$ of $E_1(\tilde{p}) = 0$ is such that $\tilde{p} < 1-1/\sqrt{2}$ for all $D\geq 2$. Finally we wish to the determine the asymptotic value of the root $\tilde{p}$ as $D\to\infty$. Let $1/4 > \delta > 0$. For each $1-1/\sqrt{2} > \epsilon > 0$, it follows from (36) that $E_1(p) > 0$ for all $1-1/\sqrt{2}+\epsilon\leq p\leq 1-\delta$, and $E_1(p)\leq 0$ for all $0\leq p\leq 1-1/\sqrt{2}-\epsilon$, for all sufficiently large D. Hence, the root $\tilde{p}$ of $E_1(\tilde{p}) = 0$ in the interval $0\leq\tilde{p}\leq 1$ approaches $1-1/\sqrt{2}$ as $D\to\infty$.

3) Comparison on Increasing Alphabets: Here we compare the semidefinite and the entropic tests for increasing D in the family $\tilde{P}_p^{3:D}$. By Proposition 5, the semidefinite test is independent of D for this particular family of distributions. The entropic test could thus become stronger than the semidefinite test for sufficiently large alphabet sizes. This is indeed what we find in the numerical evaluation of the entropic test, which we display in Figure 5. As pointed out in Section VII-C.2, all the entropic inequalities, apart from E_1, depend on the full tripartite distribution, while E_1 and the semidefinite test only utilize the bi- and mono-partite margins. We already know from the previous section that the test based on E_1 always is weaker than the semidefinite test for the family $\tilde{P}_p^{3:D}$, but that it approaches the semidefinite test in the limit of large alphabet sizes D. The plot in Figure 5 suggests that the convergence is very slow. Moreover, for $D = 10^7$, the root of the equation $E_1(p) = 0$ is $p\approx 0.26$, while the limit is $p\approx 0.29$.

VIII. Conclusion

In this work we have considered the constraints imposed by a large class of causal structures on the covariance matrix of the observed variables. More specifically, we have shown that each bipartite DAG induces a decomposition that every covariance matrix resulting from the causal model has to satisfy. Such decompositions can be formulated in terms of semidefinite programs that allow for a straightforward and efficient computational treatment of the problem (as opposed to techniques such as quantifier elimination). A violation of the condition imposed by the bipartite DAG under test (or in other terms, the non-feasibility of the semidefinite program) thus implies that the observed covariance matrix is not compatible with it. We have moreover compared the performance of the semidefinite test and tests based on information theoretic inequalities formulated in terms of entropies, where the results indicate that the semidefinite test outperforms the entropic test for moderate alphabet sizes of the random variables, while the latter become more powerful for large alphabet sizes.

These results open several directions for future research. Here, we have restricted attention to characterizing the set of covariance matrices compatible with a given causal structure. In real-world situations however, the covariance matrix is unknown and has to be estimated from a limited number of samples drawn from the underlying distribution. This raises the question of how to turn the theory developed here into statistical hypothesis tests for a presumed causal structure. An obvious idea would be to construct a confidence region for the estimated covariance matrix and reject the hypothesis if the confidence region does not intersect the set compatible with the causal assumption. We speculate, though, that it might be simpler to obtain statistically sound results by employing convex duality, as explained in the context of Figure 3. Indeed, assume that X is such that all compatible covariance matrices have non-negative inner product with X. The inner product between X and the true covariance matrix is a scalar linear function of the distribution of the observable variables. A one-sided statistical hypothesis test for $\operatorname{tr}\big(X\operatorname{Cov}(Y)\big) \leq 0$ with any desired significance level is therefore easy to construct. It automatically also tests the causal hypothesis at the same significance level. While any X gives rise to such a test, their power to identify a given true incompatible distribution may vary wildly. One way of making an informed choice for X would be as follows: Split the samples into two parts. If the empirical covariance matrix of the first part is compatible with the hypothesis, accept. If not, the dual SDP (12) identifies a witness $X^\star$ that separates the empirical matrix from the compatible set. Now use the test based on $X^\star$ with the second part of the samples. We leave the details to future work.

Another immediate question is to better understand the relation between the semidefinite and the entropic tests. Similarly, it would be highly desirable to combine our results with other tools that have very recently been proposed in order to characterize complex DAGs [23]–[25]. In view of the common focus on bipartite DAGs, it would also be interesting to explore the links to the characterization of observables covariances that has been developed in [47].

Here we have focused on comparing the semidefinite and entropic tests in terms of their capacity for rejection. However, another question is how the computational complexity of these methods compare. Is there a trade-off between how ‘close’ an outer approximation is to the true compatibility region, and its computational complexity?

On a more general level it is noteworthy that by restricting to covariance we turn a highly non-linear problem into what essentially is a convex optimization. Understanding how far this can be pushed (considering higher order moments, for instance) would certainly give us new geometric insights on the nature of this problem. Since we here have focused on a setting where all correlations of observed variables are due to latent variables, it is very reasonable to ask if tests based on covariances can be extended to more general types of DAGs that do not have this bipartite structure.

Recently, covariance constraints have been obtained in the context of Bell scenarios [62], corresponding to a simple causal structure where two observable nodes have their correlations mediated by a single latent factor. The difference, however, is the fact that each observable node can be influenced by another local observable variable (an input), that in the quantum realization corresponds to the choice of a measurement basis. Even though the problem in this case can be solved by standard tools from convex optimization, it points towards an interesting direction for future research, where our approach could be extended to include inputs.

From a more fundamental perspective, our work may have implications for the current research program on the foundations of quantum physics. Bayesian networks have attracted growing attention as means to understand the role of causality in quantum mechanical systems [5]–[13]. One may thus ask whether the methods we have employed here can be generalized to the case of quantum causal structures, where for example some nodes in the graph represent quantum states without a classical analogue. Any positive results along this line would certainly be highly relevant in the context of quantum causal modeling and once more highlight the very fruitful interplay between the fields of causal inference and foundational aspects of quantum mechanics.

Acknowledgment

The authors would like to thank Thomas Kahle and Johannes Textor for productive discussions during the early stages of this project. They also thank Thomas Richardson for drawing their attention to the literature on tetrads, and Mathias Drton for pointing out reference [47]. Furthermore, they thank an anonymous referee for pointing out a simplification of the proof of Proposition 1 in terms of martingale difference sequences.

References

[1] J. Pearl, *Causality* Cambridge, U.K.: Cambridge Univ. Press, 2009.

[2] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, 2nd ed. Cambridge, MA, USA: MIT Press, 2000.

[3] N. Friedman, “Inferring cellular networks using probabilistic graphical models,” *Science*, vol. 303, pp. 85–799, Feb. 2004.

[4] G. V. Steeg and A. Galstyan, “A sequence of relaxations constraining hidden variable models,” in *Proc. 27th Conf. Uncertainty Artif. Intell.*, Jul. 2011, pp. 717–726.

[5] M. S. Leifer and R. W. Spekkens, "Towards a formulation of quantum theory as a causally neutral theory of Bayesian inference," *Phys. Rev. A, Gen. Phys.*, vol. 88, Nov. 2013, Art. no. 052130.

[6] T. Fritz, "Beyond Bell's theorem: Correlation scenarios," *New J. Phys.*, vol. 14, Oct. 2012, Art. no. 103001.

[7] T. Fritz, "Beyond Bell's theorem II: Scenarios with arbitrary causal structure," *Commun. Math. Phys.*, vol. 341, pp. 391–434, Jan. 2016.

[8] J. Henson, R. Lal, and M. F. Pusey, "Theory-independent limits on correlations from generalized Bayesian networks," *New J. Phys.*, vol. 16, Nov. 2014, Art. no. 113043.

[9] R. Chaves, C. Majenz, and D. Gross, "Information-theoretic implications of quantum causal structures," *Nat. Commun.*, vol. 6, p. 5766, Jan. 2015.

[10] J. Pienaar and Č. Brukner, "A graph-separation theorem for quantum causal models," *New J. Phys.*, vol. 17, Jul. 2015, Art. no. 073020.

[11] K. Ried, M. Agnew, L. Vermeyden, D. Janzing, R. W. Spekkens, and K. J. Resch, "A quantum advantage for inferring causal structure," *Nature Phys.*, vol. 11, pp. 414–420 (2015).

[12] F. Costa and S. Shrapnel, "Quantum causal modelling," *New J. Phys.*, vol. 18, May 2016, Art. no. 063032.

[13] D. Horsman, C. Heunen, M. F. Pusey, J. Barrett, and R. W. Spekkens, "Can a quantum state over time resemble a quantum state at a single time?" *Proc. Royal Soc. A, Math., Phys. Eng. Sci.*, vol. 473, Sep. 2017, Art. no. 20170395.

[14] I. Pitowsky, "Correlation polytopes: Their geometry and complexity," *Math. Program.*, vol. 50, pp. 395–414, Mar. 1991.

[15] J. Pearl, "On the testability of causal models with latent and instrumental variables," in *Proc. 11th Conf. Uncertainty Artif. Intell.*, Apr. 1995, pp. 435–443.

[16] D. Geige and C. Meek, "Quantifier elimination for statistical problems," in *Proc. 15th Conf. Uncertainty Artif. Intell.*, 1999, pp. 226–235.

[17] B. Bonet, "Instrumentality tests revisited," in *Proc. 17th Conf. Uncertainty Artif. Intell.*, 2001, pp. 48–55.

[18] L. D. Garcia, M. Stillman, and B. Sturmfels, "Algebraic geometry of Bayesian networks," *J. Symbolic Comput.*, vol. 39, pp. 331–355, Mar. 2005.

[19] C. Kang and J. Tian, "Inequality constraints in causal models with hidden variables," in *Proc. 22nd Conf. Uncertainty Artif. Intell.*, 2006, pp. 233–240.

[20] C. Kang and J. Tian, "Polynomial constraints in causal Bayesian networks," in *Proc. 23rd Conf. Uncertainty Artif. Intell.*, 2007, pp. 200–208.

[21] R. J. Evans, "Graphical methods for inequality constraints in marginalized DAGs," in *Proc. 2012 IEEE Int. Workshop Mach. Learn. Signal Process.*, Sep. 2012, pp. 1–6.

[22] C. M. Lee and R. W. Spekkens, "Causal inference via algebraic geometry: Feasibility tests for functional causal structures with two binary observed variables," *J. Causal Inference*, vol. 5, Sep. 2017, Art. no. 20160013.

[23] R. Chaves, "Polynomial bell inequalities," *Phys. Rev. Lett.*, vol. 116, Jul. 2016, Art. no. 010402.

[24] D. Rosset, C. Branciard, T. J. Barnea, G. Pütz, N. Brunner, and N. Gisin, "Nonlinear bell inequalities tailored for quantum networks," *Phys. Rev. Lett.*, vol. 116, Jan. 2016, Art. no. 010403.

[25] E. Wolfe, R. W. Spekkens, and T. Fritz, "The inflation technique for causal inference with latent variables," 2016, *arXiv:1609.00672*. [Online]. Available: https://arxiv.org/abs/1609.00672

[26] T. S. Verma and J. Pearl, "Equivalence and synthesis of causal models," in *Proc. 6th Conf. Uncertainty Artif. Intell.*, 1990, pp. 220–227.

[27] J. Tian and J. Pearl, "On the testable implications of causal models with hidden variables," in *Proc. 18th Conf. Uncertainty Artif. Intell.*, Aug. 2002, pp. 519–527.

[28] P. Moritz, J. Reichardt, and N. Ay, "Discriminating between causal structures in Bayesian Networks given partial observations," *Kybernetika*, vol. 50, pp. 284–295, Feb. 2014.

[29] R. Chaves, L. Luft, and D. Gross, "Causal structures from entropic information: Geometry and novel scenarios," *New J. Phys.*, vol. 16, Aug. 2014, Art. no. 043001.

[30] R. Chaves, L. Luft, T. O. Maciel, D. Gross, D. Janzing, and B. Schölkopf, "Inferring latent structures via information inequalities," in *Proc. 30th Conf. Uncertainty Artif. Intell.*, 2014, pp. 112–121.

[31] B. Steudel and N. Ay, "Information-theoretic inference of common ancestors," *Entropy*, vol. 17, pp. 2304–2327, Apr. 2015.

[32] M. Weilenmann and R. Colbeck, "Non-Shannon inequalities in the entropy vector approach to causal structures," 2018, *arXiv:1605.02078*. [Online]. Available: https://arxiv.org/abs/1605.02078

[33] M. Weilenmann and R. Colbeck, "Analysing causal structures with entropy," *Proc. Royal Soc. A, Math., Phys. Eng. Sci.*, vol. 473, Nov. 2017, Art. no. 20170483.

[34] S. L. Braunstein and C. M. Caves, "Information-theoretic bell inequalities," *Phys. Rev. Lett.*, vol. 61, pp. 662–665, Aug. 1988.

[35] N. J. Cerf and C. Adami, "Entropic Bell inequalities," *Phys. Rev. A, Gen. Phys.*, vol. 55, pp. 3371–3374, May 1997.

[36] R. Chaves and T. Fritz, "Entropic approach to local realism and noncontextuality," *Phys. Rev. A, Gen. Phys.*, vol. 85, Mar. 2012, Art. no. 032113.

[37] T. Fritz and R. Chaves, "Entropic inequalities and marginal problems," *IEEE Trans. Inf. Theory*, vol. 59, no. 2, pp. 803–817, Feb. 2013.

[38] R. Chaves, "Entropic inequalities as a necessary and sufficient condition to noncontextuality and locality," *Phys. Rev. A, Gen. Phys.*, vol. 87, Feb. 2013, Art. no. 022102.

[39] R. Chaves, J. Bohr Brask, and N. Brunner, "Device-independent tests of entropy," *Phys. Rev. Lett.*, vol. 115, Sep. 2015, Art. no. 110501.

[40] R. Chaves and C. Budroni, "Entropic nonsignaling correlations," *Phys. Rev. Lett.*, vol. 116, Jun. 2016, Art. no. 240501.

[41] R. Silva, R. Scheines, C. Glymor, and P. Spirtes, "Learning the structure of linear latent variable models," *J. Mach. Learn. Res.*, vol. 7, pp. 191–246, Feb. 2006.

[42] M. Drton, B. Sturmfels, and S. Sullivant, "Algebraic factor analysis: Tetrads, pentads and beyond," *Probab. Thory Relat. Fields*, vol. 138, pp. 463–493, Jul. 2007.

[43] S. Sullivant, K. Talaska, and J. Draisma, "Trek separation for Gaussian graphical models," *Ann. Statist.*, vol. 38, pp. 1665–1685, Mar. 2010.

[44] J. Draisma, S. Sullivant, and K. Talaska, "Positivity for Gaussian graphical models," 2012, *arXiv:1210.0390*. [Online]. Available: https://arxiv.org/abs/1210.0390

[45] R. Silva and R. Scheines, "New d-separation identification results for learning continuous latent variable models," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 808–815.

[46] P. Spirtes, "Calculation of entailed rank constraints in partially nonlinear and cyclic models," in *Proc. 29th Conf. Uncertainty Artif. Intell.*, May 2013, pp. 606–615.

[47] M. Drton and J. Yu, "On a parametrization of positive semidefinite matrices with zeros," *SIAM J. Matrix. Anal. Appl.*, vol. 31, pp. 2665–2680, Apr. 2010.

[48] R. J. Evans, "Graphs for margins of Bayesian networks," *Scandin. J. Statist.*, vol. 3, pp. 625–648, May 2016.

[49] C. Branciard, N. Gisin, and S. Pironio, "Characterizing the nonlocal correlations created via entanglement swapping," *Phys. Rev. Lett.*, vol. 104, Jul. 2010, Art. no. 170401.

[50] C. Branciard, D. Rosset, N. Gisin, and S. Pironio, "Bilocal versus nonbilocal correlations in entanglement-swapping experiments," *Phys. Rev. A, Gen. Phys.*, vol. 85, Apr. 2012, Art. no. 032119.

[51] A. Tavakoli, P. Skrzypczyk, D. Cavalcanti, and A. Acín, "Nonlocal correlations in the star-network configuration," *Phys. Rev. A, Gen. Phys.*, vol. 90, Apr. 2014, Art. no. 062109.

[52] C. J. Wood and R. W. Spekkens, "The lesson of causal discovery algorithms for quantum correlations: Causal explanations of Bell-inequality violations require fine-tuning," *New J. Phys.*, vol. 17, Mar. 2015, Art. no. 033002.

[53] J. S. Bell, "On the einstein-podolsky-rosen paradox," *Physics*, vol. 1, pp. 195–200, Mar. 1964.

[54] D. J. Saunders, A. J. Bennet, C. Branciard, and G. J. Pryde, "Experimental demonstration of non-bilocal quantum correlations," *Sci. Adv.*, vol. 3, Apr. 2017, Art. no. e1602743.

[55] G. Carvacho, F. Andreoli, L. Santodonato, M. Bentivegna, R. Chaves, and F. Sciarrino, "Experimental violation of local causality in a quantum network," *Nature Commun.*, vol. 8, p. 14775, Apr. 2017.

[56] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Rev.*, vol. 38, pp. 49–95, Apr. 1996.

[57] K. von Prillwitz, "Statistical aspects of inferring Bayesian networks from marginal observations," M. S. thesis, Dept. Math. Phys., Univ. Freiburg, Freiburg, Germany, 2015.

[58] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd Ed. Hoboken, NJ, USA: Wiley, 2012.

[59] B. Schölkopf and A. J. Smola, *Learning With Kernels*. Cambridge, MA, USA: MIT Press, 2002.

[60] G. Gallavotti, *Statistical Mechanics*. New York, NY, USA: Springer, 1999.

[61] D. Koller and N. Friedman, *Probabilistic Graphical Models*. Cambridge, MA, USA: MIT Press, 2009.

[62] V. Pozsgay, F. Hirsch, C. Branciard, and N. Brunner, "Covariance bell inequalities," *Phys. Rev. A, Gen. Phys.*, vol. 96, Jul. 2017, Art. no. 062128.