HIGH-DIMENSIONAL CONSISTENCY IN SCORE-BASED AND HYBRID STRUCTURE LEARNING

By Preetam Nandy[*,1], Alain Hauser[†] and Marloes H. Maathuis[*,1]

ETH Zürich[*] and Bern University of Applied Sciences[†]

Main approaches for learning Bayesian networks can be classified as constraint-based, score-based or hybrid methods. Although high-dimensional consistency results are available for constraint-based methods like the PC algorithm, such results have not been proved for score-based or hybrid methods, and most of the hybrid methods have not even shown to be consistent in the classical setting where the number of variables remains fixed and the sample size tends to infinity. In this paper, we show that consistency of hybrid methods based on greedy equivalence search (GES) can be achieved in the classical setting with adaptive restrictions on the search space that depend on the current state of the algorithm. Moreover, we prove consistency of GES and adaptively restricted GES (ARGES) in several sparse high-dimensional settings. ARGES scales well to sparse graphs with thousands of variables and our simulation study indicates that both GES and ARGES generally outperform the PC algorithm.

1. Introduction. A Bayesian network consists of a directed acyclic graph (DAG) on a set of variables and conditional distributions for each node given its parents in the DAG. Bayesian networks can be used for various purposes, such as efficiently modeling the joint distribution of the variables, constructing decision support systems, probabilistic reasoning in expert systems, and causal inference.

In a Bayesian network, the DAG encodes conditional independence relationships that must hold among the corresponding random variables. Several DAGs can encode exactly the same set of conditional independence relationships. Such DAGs are called Markov equivalent and form a Markov equivalence class (see Section 2.2). A Markov equivalence class can be uniquely represented by a completed partially directed acyclic graph (CPDAG). We consider estimating the CPDAG of a Bayesian network from observational data, and we refer to this as structure learning. Main approaches for structure learning can be classified as constraint-based, score-based or hybrid methods.

Constraint-based methods, such as the PC algorithm [Spirtes, Glymour and Scheines (2000)], are based on conditional independence tests. The PC algorithm

and its variants [e.g., Harris and Drton (2013), Colombo and Maathuis (2014)] have been widely applied to high-dimensional datasets [e.g., Maathuis et al. (2010), Schmidberger, Lennert and Mansmann (2011), Stekhoven et al. (2012), Verdugo et al. (2013), Le et al. (2013), Gao and Cui (2015)], partly because they were shown to be consistent in sparse high-dimensional settings where the number of variables is allowed to grow with the sample size [Kalisch and Bühlmann (2007), Harris and Drton (2013), Colombo and Maathuis (2014)], and partly because they scale well to sparse graphs with thousands of variables.

Score-based methods aim to optimize a scoring criterion over the space of possible CPDAGs or DAGs, typically through a greedy search procedure (see Section 2.4). Greedy equivalence search (GES) [Chickering (2003)] is a popular score-based method, which was shown to be consistent in the classical setting where the number of variables remains fixed and the sample size goes to infinity. This consistency result of Chickering (2003) is remarkable, since it involves a *greedy* search. However, GES has not been shown to be consistent in high-dimensional settings. van de Geer and Bühlmann (2013) proved high-dimensional consistency of the global optimum of an ℓ_0-penalized likelihood score function under the multivariate Gaussian assumption, but it has not been proved that a greedy search method like GES can find the global optimum in a high-dimensional setting. Another obstacle for applying score-based methods like GES to high-dimensional data is that they do not scale well to large graphs.

A hybrid method combines a score-based method either with a constraint-based method or with a variable selection method. Such methods often use a greedy search on a restricted search space in order to achieve computational efficiency, where the restricted space is estimated using conditional independence tests or variable selection methods [Tsamardinos, Brown and Aliferis (2006), Schmidt, Niculescu-Mizil and Murphy (2007), Schulte et al. (2010), Alonso-Barba et al. (2013)]. Common choices for the restricted search space are an estimated skeleton of the CPDAG (CPDAG-skeleton) or an estimated conditional independence graph (CIG). The CIG (or moral graph or Markov network) of a joint distribution of p variables $X_1, \ldots, X_p$ is an undirected graph where two nodes X_i and X_j are adjacent if and only if they are conditionally dependent given $\{X_1, \ldots, X_p\} \setminus \{X_i, X_j\}$. The CIG is a supergraph of the CPDAG-skeleton. Hybrid algorithms generally scale well with respect to the number of variables, but their consistency results are generally lacking even in the classical setting, except for Alonso-Barba et al. (2013).

In a preliminary simulation study, we compared the performances of PC, GES, and GES restricted to an estimated CIG (RGES-CIG) in high-dimensional settings (see Section 1 of the Supplementary Material). Table 1 summarizes our findings from these preliminary simulations and the existence of consistency results in the literature.

Although GES and RGES-CIG outperform PC in terms of estimation performance in our high-dimensional simulations, we find that PC is the most commonly

TABLE 1
Summary of performance and existing consistency results, where tick marks represent good performance or existence of consistency results, cross marks represent bad performance, and question marks represent nonexistence of consistency results

	Speed	Estimation performance	Consistency	High-dimensional consistency
PC	✓	✗	✓	✓
GES	✗	✓	✓	?
RGES-CIG	✓	✓	?	?

applied method in high-dimensional applications. We suspect that the main reasons for the lack of popularity of score-based and hybrid methods in high-dimensional applications are that they lack consistency results in high-dimensional settings and/or that they do not scale well to large graphs. In this paper, we prove high-dimensional consistency of GES, and we propose new hybrid algorithms based on GES that are consistent in several sparse high-dimensional settings *and* scale well to large sparse graphs. To the best of our knowledge, these are the first results on high-dimensional consistency of score-based and hybrid methods.

The paper is organized as follows. Section 2 discusses some necessary background knowledge. In Section 3, we show with an explicit example that naive hybrid versions of GES that restrict the search space to the CIG or the CPDAG-skeleton are inconsistent. This shows that the search path of GES may have to leave the search space determined by the CIG or the CPDAG-skeleton, even though the true CPDAG lies within these search spaces.

In Section 4 we provide a novel insight into how consistency can be achieved with hybrid algorithms based on GES, by imposing a restriction on the search space that changes adaptively depending on the current state of an algorithm. In particular, we propose a new method called adaptively restricted greedy equivalence search (ARGES), where in addition to the edges of the CIG (or the CPDAG-skeleton), we allow the shields of v-structures (or unshielded triples) in the current CPDAG, at every step of the algorithm. Our consistency proofs are based on a new characterization of independence maps (Theorem 4.1), which is an interesting result in itself.

In Section 5 we prove consistency of GES and ARGES in certain sparse high-dimensional settings with multivariate Gaussian distributions. As a key ingredient of our proof, we show a connection between constraint-based and score-based methods. This connection enables us to extend our high-dimensional consistency results to linear structural equation models with sub-Gaussian errors (Section 6). Furthermore, it motivated us to define a scoring criterion based on rank correlations, and hence to derive high-dimensional consistency results for GES and ARGES for nonparanormal distributions (Section 7). This result is analogous to the

high-dimensional consistency result of the Rank-PC algorithm [Harris and Drton (2013)]. Section 8 contains simulation results, where we compare the finite sample performances and runtimes of PC, GES, ARGES and max-min hill-climbing [Tsamardinos, Brown and Aliferis (2006)] in certain sparse high-dimensional settings. We end with a discussion and problems for future research in Section 9.

All proofs are given in the Supplementary Material [Nandy, Hauser and Maathuis (2018)]. An implementation of ARGES has been added to the R-package *pcalg* [Kalisch et al. (2012)].

2. Preliminaries.

2.1. *Graph terminology.* We consider graphs $\mathcal{A} = (\mathbf{X}, E)$, where the nodes (or vertices) $\mathbf{X} = \{X_1, \ldots, X_p\}$ represent random variables and the edges represent relationships between pairs of variables. The edges can be either *directed* ($X_i \rightarrow X_k$) or *undirected* ($X_i - X_k$). An *(un)directed graph* can only contain (un)directed edges, whereas a *partially directed graph* may contain both directed and undirected edges. For partially directed graphs $\mathcal{A} = (\mathbf{X}, E)$ and $\mathcal{A}' = (\mathbf{X}, E')$, we write $\mathcal{A} \subseteq \mathcal{A}'$ and $\mathcal{A} = \mathcal{A}'$ to denote $E \subseteq E'$ and $E = E'$ respectively. The *skeleton* of a partially directed graph $\mathcal{A}$, denoted as skeleton($\mathcal{A}$), is the undirected graph that results from replacing all directed edges of $\mathcal{A}$ by undirected edges.

Two nodes X_i and X_k are *adjacent* if there is an edge between them. Otherwise they are *nonadjacent*. The set of all adjacent node of X_i in $\mathcal{A}$ is denoted by $\mathbf{Adj}_{\mathcal{A}}(X_i)$. The degree of a node X_i in $\mathcal{A}$ equals $|\mathbf{Adj}_{\mathcal{A}}(X_i)|$. If $X_i \rightarrow X_k$, then X_i is a *parent* of X_k. The set of all parents of X_k in $\mathcal{A}$ is denoted by $\mathbf{Pa}_{\mathcal{A}}(X_k)$. A triple of nodes (X_i, X_j, X_k) is an unshielded triple in $\mathcal{A}$ if X_i and X_k are nonadjacent in $\mathcal{A}$ and $\{X_i, X_k\} \subseteq \mathbf{Adj}_{\mathcal{A}}(X_j)$. An unshielded triple (X_i, X_j, X_k) is a *v-structure* if $X_i \rightarrow X_j \leftarrow X_k$. If (X_i, X_j, X_k) is an unshielded triple in $\mathcal{A}$, then the edge $X_i \rightarrow X_k$ (or $X_k \rightarrow X_i$), which is not present in $\mathcal{A}$, is called a shield of the unshielded triple (X_i, X_j, X_k).

A *path* between X_i and X_k in a graph $\mathcal{A}$ is sequence of distinct nodes $(X_i, \ldots, X_k)$ such that all pairs of successive nodes in the sequence are adjacent in $\mathcal{A}$. We use the shorthand $\pi_{\mathcal{A}}(X_i, \ldots, X_k)$ to denote a path in $\mathcal{A}$ with endpoint nodes X_i and X_k. A nonendpoint node X_r on a path $\pi_{\mathcal{A}}(X_i, \ldots, X_{r-1}, X_r, X_{r+1}, \ldots, X_k)$ is a *collider* on the path if $X_{r-1} \rightarrow X_r \leftarrow X_{r+1}$. Otherwise it is a *noncollider* on the path. An endpoint node on a path is neither a collider nor a noncollider on the path. A path without colliders is a *noncollider path*. A path of two nodes is a trivial noncollider path.

A *directed path* from X_i to X_k is a path between X_i and X_k, where all edges are directed towards X_k. If there is a directed path from X_i to X_k, then X_k is a *descendant* of X_i. Otherwise it is a *nondescendant*. We use the convention that each node is a descendant of itself. The set of all descendants (nondescendants) of X_i in $\mathcal{A}$ is denoted by $\mathbf{De}_{\mathcal{A}}(X_i)$ [$\mathbf{Nd}_{\mathcal{A}}(X_i)$].

A path between X_i and X_k of at least three nodes, together with edge between X_i and X_k forms a *cycle*. A directed path from X_i to X_k together with $X_k \rightarrow X_i$ forms a *directed cycle*. A directed graph or partially directed graph without directed cycles is called *acyclic*. A graph that is both (partially) directed and acyclic, is a *(partially) directed acyclic graph* or (P)DAG.

We will typically denote an arbitrary DAG by $\mathcal{H}$ (or $\mathcal{H}_n$), and an arbitrary partially directed graph by $\mathcal{A}$ (or $\mathcal{A}_n$). Graphs are always assumed to have vertex set $\mathbf{X} = \{X_1, \ldots, X_p\}$ (or $\mathbf{X}_n = \{X_{n1}, \ldots, X_{np_n}\}$).

2.2. *Bayesian network terminology.* We consider a random vector $\mathbf{X} = (X_1, \ldots, X_p)^T$ with a parametric density $f(\cdot)$. The density *factorizes* according to a DAG $\mathcal{H}$ if there exists a set of parameter values $\mathbf{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p\}$ such that

$$(1) \qquad f(x_1, \ldots, x_p) = \prod_{i=1}^{p} f_i\big(x_i \mid \mathbf{Pa}_{\mathcal{H}}(X_i) = \mathbf{Pa}_{\mathcal{H}}(x_i), \boldsymbol{\theta}_i\big),$$

where $\boldsymbol{\theta}_i$ specifies the conditional density of X_i given its parents in $\mathcal{H}$ and $\mathbf{Pa}_{\mathcal{H}}(x_i)$ denotes the subvector of $(x_1, \ldots, x_p)$ that corresponds to $\mathbf{Pa}_{\mathcal{H}}(X_i)$. Such a pair $(\mathcal{H}, \mathbf{\Theta})$ is a Bayesian network that defines the joint distribution. The DAG $\mathcal{H}$ of a Bayesian network $(\mathcal{H}, \mathbf{\Theta})$ encodes conditional independence constraints that must hold in any distribution that factorizes according to $\mathcal{H}$. Conditional independence constraints encoded by a DAG can be read off from the DAG using the notion of *d-separation*.

DEFINITION 2.1 [d-separation; see Definition 1.2.3 of Pearl (2009)]. Let $\mathbf{S}$ be a subset of nodes in a DAG $\mathcal{H}$, where $\mathbf{S}$ does not contain X_i and X_k. Then $\mathbf{S}$ *blocks* a path $\pi_{\mathcal{H}}(X_i, \ldots, X_k)$ if at least one of the following holds: (i) $\pi_{\mathcal{H}}$ contains a noncollider that is in $\mathbf{S}$, or (ii) $\pi_{\mathcal{H}}$ contains a collider that has no descendant in $\mathbf{S}$. Otherwise $\pi_{\mathcal{H}}$ is *open* given $\mathbf{S}$. For pairwise disjoint sets of nodes $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{S}$, we say that $\mathbf{W}_1$ and $\mathbf{W}_2$ are *d-separated* by $\mathbf{S}$ in $\mathcal{H}$ if every path between a node in $\mathbf{W}_1$ and a node in $\mathbf{W}_2$ is blocked by $\mathbf{S}$. This is denoted by $\mathbf{W}_1 \perp_{\mathcal{H}} \mathbf{W}_2 \mid \mathbf{S}$. Otherwise, $\mathbf{W}_1$ and $\mathbf{W}_2$ are *d-connected* given $\mathbf{S}$ in $\mathcal{H}$, denoted by $\mathbf{W}_1 \not\perp_{\mathcal{H}} \mathbf{W}_2 \mid \mathbf{S}$.

The distribution of $\mathbf{X}$ is *DAG-perfect* [Chickering (2003)] if there exists a DAG $\mathcal{G}_0$ such that (i) every independence constraint encoded by $\mathcal{G}_0$ holds in the distribution of $\mathbf{X}$, and (ii) every independence constraint that holds in the distribution is encoded by $\mathcal{G}_0$. Such a DAG $\mathcal{G}_0$ is called a *perfect map* of the distribution. Condition (i) is known as the *global Markov property*, and condition (ii) is the so-called *faithfulness* condition [see, for example, Definition 3.8 of Koller and Friedman (2009)]. In this paper, we only consider DAG-perfect distributions [as in Chickering (2003)].

DAGs that encode exactly the same set of conditional independence constraints form a *Markov equivalence class* [Verma and Pearl (1990)]. Two DAGs belong

to the same Markov equivalence class if and only if they have the same skeleton and the same v-structures [Verma and Pearl (1990)]. A Markov equivalence class of DAGs can be uniquely represented by a *completed partially directed acyclic graph* (CPDAG), which is a PDAG that satisfies the following: $X_i \rightarrow X_k$ in the CPDAG if $X_i \rightarrow X_k$ in every DAG in the Markov equivalence class, and $X_i - X_k$ in the CPDAG if the Markov equivalence class contains a DAG in which $X_i \rightarrow X_k$ as well as a DAG in which $X_i \leftarrow X_k$ [Verma and Pearl (1990), Andersson, Madigan and Perlman (1997), Chickering (2003)]. We use the notation CPDAG($\mathcal{H}$) to denote the CPDAG that represents the Markov equivalence class of a DAG $\mathcal{H}$.

NOTATION 2.1.   We reserve the notation $\mathcal{G}_0, \mathcal{C}_0, \mathcal{I}_0$ and $\mathcal{D}_n$ for the following: $\mathcal{G}_0$ denotes a perfect map of the distribution of $\mathbf{X}$ with $\mathcal{C}_0 = \text{CPDAG}(\mathcal{G}_0)$, $\mathcal{I}_0$ is the conditional independence graph of $\mathbf{X}$, and $\mathcal{D}_n$ denotes the data, consisting of n i.i.d. observations of $\mathbf{X}$.

For a DAG $\mathcal{H}$, let $\mathbb{CI}(\mathcal{H})$ denote the set of all conditional independence constraints encoded by $\mathcal{H}$. By definition, $\mathcal{H}_1$ and $\mathcal{H}_2$ are Markov equivalent DAGs if and only if $\mathbb{CI}(\mathcal{H}_1) = \mathbb{CI}(\mathcal{H}_2)$. Thus, for a CPDAG $\mathcal{C}$, we use the notation $\mathbb{CI}(\mathcal{C})$ unambiguously to denote the set of all conditional independence constraints encoded by any DAG $\mathcal{H}$ in the Markov equivalence class of $\mathcal{C}$.

DEFINITION 2.2.   A (CP)DAG $\mathcal{A}$ is an *independence map* of a (CP)DAG $\mathcal{A}'$ if $\mathbb{CI}(\mathcal{A}) \subseteq \mathbb{CI}(\mathcal{A}')$.

2.3. *Properties of a scoring criterion.*   Score-based and hybrid structure learning methods require a scoring criterion $\mathcal{S}(\mathcal{H}, \mathcal{D}_n)$ that measures the quality of a candidate DAG $\mathcal{H}$ with respect to given data $\mathcal{D}_n$. Throughout this paper, we assume without loss of generality that optimizing a scoring criterion corresponds to minimizing it. Hence, we say that the score improves by moving from $\mathcal{H}$ to $\mathcal{H}'$ if $\mathcal{S}(\mathcal{H}', \mathcal{D}_n) < \mathcal{S}(\mathcal{H}, \mathcal{D}_n)$.

We consider scoring criterions that are *score equivalent*, *decomposable* and *consistent* (see Section 2 of the Supplementary Material). These properties are also assumed by Chickering (2003) as basic requirements of a scoring criterion to be used in GES. Score equivalence ensures that all DAGs in a Markov equivalence class get the same score, and the common score is defined as the score of the Markov equivalence class or its representative CPDAG. Decomposability of a scoring criterion facilitates fast computation of the score difference between two DAGs that differ by a few edges. Consistency of $\mathcal{S}$ assures that $\mathcal{G}_0$ has a lower score than any DAG that is not in the Markov equivalence class of $\mathcal{G}_0$, with probability approaching one as $n \rightarrow \infty$ [Proposition 8 of Chickering (2003)]. For multivariate Gaussian distributions, the Bayesian information criterion (BIC) is an example of a scoring criterion that is score equivalent, decomposable and consistent. BIC was chosen as the scoring criterion of GES [Chickering (2003)].

2.4. *Greedy equivalence search* (*GES*) [*Chickering* (2002*b*)].   GES is a greedy search algorithm that aims to optimize a score function on the space of CPDAGs. GES uses a forward phase and a backward phase. The forward phase starts with an initial CPDAG (often an empty graph) and sequentially obtains larger CPDAGs by adding exactly one edge at each step. Among all the possible single edge additions at every step, it selects the one that minimizes the score function. The forward phase ends when the score of the CPDAG can no longer be improved by a single edge addition. The backward phase starts with the output of the forward phase and sequentially obtains smaller CPDAGs by deleting exactly one edge at each step. It selects the optimal single edge deletion at each step and stops when the score can no longer be improved by a single edge deletion.

Conceptually, the single edge additions and deletions are defined as follows. At every step in the forward (or backward) phase of GES, one can first list all DAGs in the Markov equivalence class of the current CPDAG, then consider all DAGs that can be obtained from the aforementioned DAGs by a single edge addition (or deletion), and finally move to the CPDAG that corresponds to a resulting DAG that has the minimum score (if the minimum score is smaller than the score of the current CPDAG). Thus, at each step, the skeleton of the CPDAG changes by exactly one edge, but the orientations may change for several edges. These moves, however, can be determined much more efficiently, and we refer to Chickering (2003) for details.

Pseudocodes of the forward and the backward phases are given in Section 3 of the Supplementary Material and we refer to Figure 3 of the Supplementary Material for an illustration of the search path of GES for Example 1 (Section 3).

3. Inconsistency of GES restricted to the CIG/CPDAG-skeleton.

Naive hybrid versions of GES restrict the search space to an estimated CIG or CPDAG-skeleton. We refer to these hybrid methods as RGES-CIG or RGES-skeleton. More precisely, we restrict the search space of GES by allowing an edge $X_i \to X_j$ for addition only if X_i and X_j are adjacent in the CIG or in the CPDAG-skeleton. We will show inconsistency of these methods using an explicit example, where we assume that the CIG or the CPDAG-skeleton can be estimated consistently meaning that we restrict the search space to the true CIG or the true CPDAG-skeleton.

EXAMPLE 1.   We consider $\mathbf{X} = (X_1, X_2, X_3, X_4)^T$ with a joint distribution defined via the following linear structural equation model (SEM): $X_1 = \varepsilon_1$, $X_2 = \varepsilon_2$, $X_3 = 1.4X_1 + 1.3X_2 + \varepsilon_3$ and $X_4 = 1.2X_2 + 0.9X_3 + \varepsilon_4$, where $\varepsilon_1, \ldots, \varepsilon_4$ are independently distributed standard Gaussian random variables. We write the linear SEM in matrix notation as $\mathbf{X} = B\mathbf{X} + \boldsymbol{\varepsilon}$, where B is a lower triangular matrix of coefficients and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_4)^T$. Thus $\mathbf{X} = (I - B)^{-1}\boldsymbol{\varepsilon}$ and $\mathbf{X}$ has a zero-mean multivariate Gaussian distribution with covariance matrix $\Sigma_0 = (I - B)^{-1}(I - B)^{-T}$.
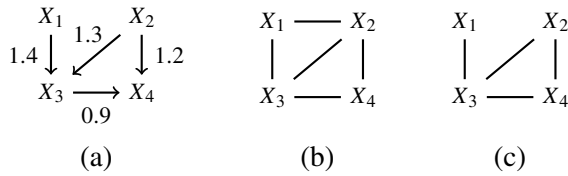
FIG. 1.   *The weighted DAG in* (a) *represents the data generating process. The corresponding CIG and the CPDAG-skeleton are given in* (b) *and* (c) *respectively.*

This linear SEM can be represented by the DAG $\mathcal{G}_0$ in Figure 1(a), where an edge $X_i \to X_j$ is present if and only if $B_{ji} \neq 0$ and then the weight of the edge $X_i \to X_j$ is B_{ji}. Note that $\mathcal{G}_0$ is a perfect map of the distribution of $\mathbf{X}$. The Markov equivalence class of $\mathcal{G}_0$ contains only one DAG and hence the corresponding CPDAG is identical to $\mathcal{G}_0$. The CIG contains all possible undirected edges except for the edge $X_1 - X_4$ [Figure 1(b)], whereas the CPDAG-skeleton additionally excludes the undirected edge $X_1 - X_2$ [Figure 1(c)].

Figure 2 shows that the large sample limit output (with the BIC criterion) of GES equals $\mathcal{G}_0 = \mathrm{CPDAG}(\mathcal{G}_0)$, but the large sample limit outputs (with the BIC criterion) of RGES-CIG and RGES-skeleton are different from $\mathcal{G}_0$. The corresponding search paths of all three algorithms are given in Section 4.1 of the Supplementary Material.

We initialized all algorithms by the empty graph. We determined the large sample limit outputs of all algorithms by choosing the scoring criterion to be the expected negative log-likelihood scores. A detailed description is given in Section 4.2 of the Supplementary Material.

REMARK 3.1.   We note that the linear Gaussian SEM given in Example 1 can also be used to show inconsistency of hill-climbing DAG search, hill-climbing DAG search restricted to the CIG [e.g., Schmidt, Niculescu-Mizil and Murphy (2007)] and hill-climbing DAG search restricted to the CPDAG-skeleton [e.g., the max-min hill-climbing algorithm of Tsamardinos, Brown and Aliferis (2006)]. Details are given in Section 4.3 of the Supplementary Material.

REMARK 3.2.   The consistency of score-based and hybrid algorithms corresponds to the soundness of these algorithms with an appropriate oracle scoring
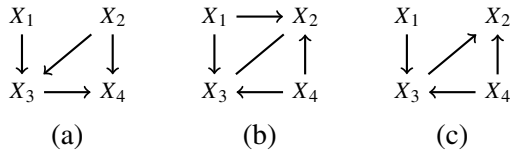

FIG. 2.   *The CPDAGs in* (a), (b) *and* (c) *are the large sample limit outputs of GES, RGES-CIG and RGES-skeleton respectively.*

criterion. An oracle scoring criterion depends on the joint distribution instead of the data and can be viewed as the large sample limit of its finite sample counterpart. For example, the (penalized) expected log-likelihood score is an oracle score which is the large sample limit of the (penalized) average log-likelihood. In fact, Example 1 shows that RGES-CIG and RGES-skeleton with the expected negative log-likelihood scoring criterion are not sound (cf. Section 4.2 of the Supplementary Material).

4. Adaptively restricted greedy equivalence search (ARGES). In the previous section, we have seen that naive hybrid versions of GES can be inconsistent, although GES is consistent [Chickering (2003)]. We now propose adaptively restricted hybrid versions of GES that will be shown to be consistent in Section 4.2.

We recall that the consistency proof of Chickering (2003) consists of two main steps. The first step shows that the output of the forward phase is an independence map of $\mathcal{C}_0$, with probability approaching one. The second step shows that $\mathbb{P}(\tilde{\mathcal{C}}_n = \mathcal{C}_0) \to 1$ given the result of the first step [Lemma 10 of Chickering (2003)], where $\tilde{\mathcal{C}}_n$ denotes the output of GES based on n i.i.d. samples.

We consider hybrid versions of GES that modify only the forward phase of GES, by restricting edge additions. To retain consistency in such hybrid versions, it therefore suffices to ensure that the output of the forward phase is an independence map of $\mathcal{C}_0$ with probability approaching one. This motivated us to provide the following novel characterization of independence maps in Theorem 4.1. Our adaptively restricted hybrid versions and their consistency will follow as a natural consequence of Theorem 4.1 and Chickering's consistency proof of GES.

THEOREM 4.1. *A DAG $\mathcal{H}$ is not an independence map of a DAG $\mathcal{G}$ if and only if*

1. skeleton($\mathcal{G}$) $\nsubseteq$ skeleton($\mathcal{H}$), *or*
2. *there exists a triple of nodes $\{X_i, X_j, X_k\}$ such that X_i and X_k are nonadjacent in $\mathcal{H}$, $\pi_{\mathcal{H}}(X_i, X_j, X_k)$ is a noncollider path, and $\pi_{\mathcal{G}}(X_i, X_j, X_k)$ is a v-structure, or*
3. *there exists a triple of nodes $\{X_i, X_j, X_k\}$ such that $\pi_{\mathcal{H}}(X_i, X_j, X_k)$ is a v-structure and $X_i \not\perp_{\mathcal{G}} X_k \mid \mathbf{Pa}_{\mathcal{H}}(X_k)$, where without loss of generality we assume $X_i \in \mathbf{Nd}_{\mathcal{H}}(X_k)$.*

We note that Proposition 27 and Lemma 28 of Chickering (2003) imply that if one of the first two conditions of Theorem 4.1 hold, then $\mathcal{H}$ is not an independence map. If the third condition of Theorem 4.1 holds, then $X_i \not\perp_{\mathcal{G}} X_k \mid \mathbf{Pa}_{\mathcal{H}}(X_k)$ and $X_i \perp_{\mathcal{H}} X_k \mid \mathbf{Pa}_{\mathcal{H}}(X_k)$ [since $X_i \in \mathbf{Nd}_{\mathcal{H}}(X_k) \setminus \mathbf{Pa}_{\mathcal{H}}(X_k)$], and hence $\mathcal{H}$ is not an independence map of $\mathcal{G}$.

The proof of the "only if" part of theorem is rather involved and we provide some intuition by considering two special cases in Section 9.1 of the Supplementary Material.

We will use Theorem 4.1 with $\mathcal{G} = \mathcal{G}_0$ to derive that consistency can be achieved with a hybrid algorithm that allows an edge addition between X_i and X_k at any step in the forward phase if (i) X_i and X_k are adjacent in skeleton($\mathcal{G}_0$), (ii) $\pi_{\mathcal{G}_0}(X_i, X_j, X_k)$ is a v-structure for some X_j, or (iii) (X_i, X_j, X_k) is a v-structure in the current CPDAG for some X_j.

Recall that X_i and X_k are adjacent in the CIG of the distribution of $\mathbf{X}$ if and only if X_i and X_k are adjacent in $\mathcal{G}_0$ or $\pi_{\mathcal{G}_0}(X_i, X_j, X_k)$ is a v-structure for some X_j. Thus, we modify RGES-CIG by additionally allowing edges that are shields of v-structures in the current CPDAG at every step of the forward phase.

RGES-skeleton allows an edge addition between X_i and X_k only if X_i and X_k are adjacent in skeleton($\mathcal{G}_0$). Therefore, we modify RGES-skeleton by additionally allowing edges that are shields of unshielded triples in the current CPDAG at every step of the forward phase.

We call these modified versions ARGES-CIG or ARGES-skeleton, and we describe them in detail below. We often refer to both ARGES-CIG and ARGES-skeleton as ARGES in statements that hold for both of them.

4.1. *The ARGES algorithm.*   Given an estimated CIG (or CPDAG-skeleton), ARGES greedily optimizes a scoring criterion in two phases: a forward phase that depends on the estimated CIG (or CPDAG-skeleton) and on the current state of the algorithm, and a backward phase that is identical to the backward phase of GES.

The pseudocode of the forward phase of ARGES is given in Algorithm 4.1. It starts with an initial CPDAG $\mathcal{C}_{\text{start}}$ (often an empty graph) and sequentially obtains a larger CPDAG by adding exactly one edge at each step. At every step, it selects

---

Algorithm 4.1 The forward phase of ARGES (based on $\mathcal{A}$)

Input: A scoring criterion $\mathcal{S}$, the data $\mathcal{D}_n$, an initial CPDAG $\mathcal{C}_{\text{start}}$, and an undirected graph $\mathcal{A}$ (that equals either an (estimated) CIG or an (estimated) CPDAG-skeleton).

Output: A CPDAG

1: $\mathcal{C}_{\text{new}} \leftarrow \mathcal{C}_{\text{start}}$;
2: **repeat**
3:     $\hat{\mathcal{C}}_n^f \leftarrow \mathcal{C}_{\text{new}}$;
4:     $\mathfrak{C} \leftarrow$ the set of all CPDAGs $\mathcal{C}$ such that $\mathcal{S}(\mathcal{C}, \mathcal{D}_n) < \mathcal{S}(\hat{\mathcal{C}}_n^f, \mathcal{D}_n)$, and $\mathcal{C}$ can be obtained by an admissible move from $\hat{\mathcal{C}}_n^f$ with respect to $\mathcal{A}$;
5:     **if** $\mathfrak{C} \neq \varnothing$ **then**
6:         choose $\mathcal{C}_{\text{new}}$ to be the CPDAG that minimizes the scoring criterion among the CPDAGs in $\mathfrak{C}$;
7:     **end if**
8: **until** $\mathfrak{C} = \varnothing$;
9: **return** $\hat{\mathcal{C}}_n^f$.

an optimal move (that minimizes the score) from a given set of *admissible moves*, which depend on an estimated CIG (or CPDAG-skeleton) and the current state of the algorithm. To define the set of admissible moves, we introduce the notion of *admissible edge* for ARGES-CIG and ARGES-skeleton.

DEFINITION 4.1.   (Admissible edge for ARGES-CIG) Let $\{X_i, X_k\}$ be a pair of nonadjacent nodes in a CPDAG $\mathcal{C}$. Then an edge between X_i and X_k is *admissible* for $\mathcal{C}$ with respect to an undirected graph $\mathcal{I}$ if at least one of the following hold:

1. X_i and X_k are adjacent in $\mathcal{I}$; or
2. There exists a node X_j such that (X_i, X_j, X_k) is a v-structure in $\mathcal{C}$.

If $\mathcal{I}$ equals the true CIG $\mathcal{I}_0$ or an estimate thereof, then the first condition of Definition 4.1 corresponds to the restriction of the search space to the (estimated) CIG. The second condition is our adaptive relaxation that allows shields of v-structures.

DEFINITION 4.2 (Admissible edge for ARGES-skeleton).   Let $\{X_i, X_k\}$ be a pair of nonadjacent nodes in a CPDAG $\mathcal{C}$. Then an edge between X_i and X_k is *admissible* for $\mathcal{C}$ with respect to an undirected graph $\mathcal{U}$ if at least one of the following hold:

1. X_i and X_k are adjacent in $\mathcal{U}$; or
2. There exists a node X_j such that (X_i, X_j, X_k) is an unshielded triple in $\mathcal{C}$.

For ARGES-skeleton, we will choose $\mathcal{U}$ to be (an estimate of) skeleton($\mathcal{C}_0$), which is typically a smaller graph than (an estimate of) $\mathcal{I}_0$. Therefore, the first condition in Definition 4.2 is more restrictive than the first condition of Definition 4.1 when $\mathcal{U} = $ skeleton($\mathcal{C}_0$) and $\mathcal{I} = \mathcal{I}_0$. This is somewhat compensated by the second condition, which is less restrictive in Definition 4.2 than in Definition 4.1.

DEFINITION 4.3.   (Admissible move) Let $\mathcal{A}$ be an undirected graph and $\mathcal{C}$ a CPDAG, such that the edge $X_i \rightarrow X_k$ is admissible for a CPDAG $\mathcal{C}$ with respect to $\mathcal{A}$, where we apply Definition 4.1 if $\mathcal{A}$ is (an estimate of) $\mathcal{I}_0$ and Definition 4.2 if $\mathcal{A}$ is (an estimate of) skeleton($\mathcal{C}_0$). Then the move from $\mathcal{C}$ to another CPDAG $\mathcal{C}'$ is admissible with respect to $\mathcal{A}$ if there exist DAGs $\mathcal{H}$ and $\mathcal{H}'$ in the Markov equivalence classes described by $\mathcal{C}$ and $\mathcal{C}'$ respectively, such that $\mathcal{H}'$ can be obtained from $\mathcal{H}$ by adding the edge $X_i \rightarrow X_k$ [i.e., $X_i \in \mathbf{Nd}_{\mathcal{H}}(X_k) \setminus \mathbf{Pa}_{\mathcal{H}}(X_k)$].

The forward phase of ARGES resembles the forward phase of GES (Algorithm 3.1 of the Supplementary Material), with the difference that at each step an edge between two nonadjacent nodes can only be added if (i) the nodes are adjacent in an estimated CIG (or CPDAG-skeleton), or (ii) the edge shields a v-structure (or an unshielded triple) in the current CPDAG. Therefore, the forward phase of GES is the same as Algorithm 4.1 with $\mathcal{A}$ being the complete undirected graph.

4.2. *Consistency in the classical setting.* In this subsection we prove consistency of ARGES in the classical setting, where the sample size n tends to infinity and the number of variables p remains fixed. We fix an initial CPDAG $\mathcal{C}_{start}$ and a score equivalent and consistent scoring criterion $\mathcal{S}$ (see Section 2.3).

As we discussed before, it suffices to show that the output of the forward phase of ARGES is an independence map of $\mathcal{C}_0$. In the proof of Lemma 9 of Chickering (2003), Chickering argued that if the output of the forward phase of GES is not an independence map, its score can be improved (asymptotically) by an edge addition, which is a contradiction. The following corollary of Theorem 4.1 shows that if $\mathcal{H}$ is not an independence map of $\mathcal{G}_0$, then the score can be improved (asymptotically) by adding an admissible edge $X_i \rightarrow X_k$ (as in Definition 4.1 or Definition 4.2). This additional result allows us to follow Chickering's argument for showing that the output of the forward phase of ARGES is an independence map of $\mathcal{C}_0$, leading to the consistency results given in Theorems 4.2 and 4.3.

COROLLARY 4.1. *If $\mathcal{H}$ is not an independence map of $\mathcal{G}_0$, then there exists a pair of nonadjacent nodes $\{X_i, X_k\}$ in $\mathcal{H}$ such that $X_i \in \mathbf{Nd}_{\mathcal{H}}(X_k)$, $X_i \not\perp_{\mathcal{G}_0} X_k \mid \mathbf{Pa}_{\mathcal{H}}(X_k)$, and the edge $X_i \rightarrow X_k$ is admissible for CPDAG($\mathcal{H}$) with respect to $\mathcal{I}_0$ (for ARGES-CIG) and with respect to* skeleton($\mathcal{C}_0$) *(for ARGES-skeleton).*

THEOREM 4.2. *Let the CPDAG $\hat{\mathcal{C}}_n$ be the output of ARGES-CIG based on the estimated CIG $\hat{\mathcal{I}}_n$, where $\hat{\mathcal{I}}_n$ satisfies $\lim_{n \rightarrow \infty} \mathbb{P}(\hat{\mathcal{I}}_n \supseteq \mathcal{I}_0) = 1$. Then $\lim_{n \rightarrow \infty} \mathbb{P}(\hat{\mathcal{C}}_n = \mathcal{C}_0) = 1$.*

THEOREM 4.3. *Let the CPDAG $\check{\mathcal{C}}_n$ be the output of ARGES-skeleton based on the estimated CPDAG-skeleton $\hat{\mathcal{U}}_n$, where $\hat{\mathcal{U}}_n$ satisfies $\lim_{n \rightarrow \infty} \mathbb{P}(\hat{\mathcal{U}}_n \supseteq$ skeleton($\mathcal{C}_0$)) $= 1$. Then $\lim_{n \rightarrow \infty} \mathbb{P}(\check{\mathcal{C}}_n = \mathcal{C}_0) = 1$.*

5. High-dimensional consistency of GES and ARGES in the multivariate Gaussian setting. We prove high-dimensional consistency of (AR)GES with an ℓ_0-penalized log-likelihood scoring criterion $\mathcal{S}_\lambda$ (Definition 5.1), using the following steps. We define a collection of oracle versions of (AR)GES with an oracle scoring criterion $\mathcal{S}_\lambda^*$ (Definition 5.2), and prove soundness of them. We complete the proof by showing that the sample version of (AR)GES with scoring criterion $\mathcal{S}_{\lambda_n}$ is identical to one of the oracle versions with probability approaching one, for a suitably chosen sequence of penalty parameters λ_n.

5.1. *ℓ_0-Penalized log-likelihood score in the multivariate Gaussian setting.*

DEFINITION 5.1. Let $\mathcal{H} = (\mathbf{X}, E)$ be a DAG. The ℓ_0-penalized log-likelihood score with penalty parameter λ is given by

$$\mathcal{S}_\lambda(\mathcal{H}, \mathcal{D}_n) = -\sum_{i=1}^{p} \frac{1}{n} \log\left(L\left(\hat{\boldsymbol{\theta}}_i(\mathcal{H}), \mathcal{D}_n(X_i) \mid \mathcal{D}_n\left(\mathbf{Pa}_{\mathcal{H}}(X_i)\right)\right)\right) + \lambda |E|,$$

where $L(\boldsymbol{\theta}_i(\mathcal{H}), \mathcal{D}_n(X_i)|\mathcal{D}_n(\mathbf{Pa}_{\mathcal{H}}(X_i)))$ is the likelihood function that corresponds to the conditional density of X_i given $\mathbf{Pa}_{\mathcal{H}}(X_i)$ and

$$\hat{\boldsymbol{\theta}}_i(\mathcal{H}) = \underset{\boldsymbol{\theta}_i(\mathcal{H})}{\operatorname{argmax}} \frac{1}{n} \log\big(L\big(\boldsymbol{\theta}_i(\mathcal{H}), \mathcal{D}_n(X_i)|\mathcal{D}_n\big(\mathbf{Pa}_{\mathcal{H}}(X_i)\big)\big)\big)$$

is the maximum likelihood estimate (MLE) of the parameter vector $\boldsymbol{\theta}_i(\mathcal{H})$.

REMARK 5.1. The BIC criterion is a special case of the ℓ_0-penalized log-likelihood score. In particular, the BIC score of a DAG $\mathcal{H}$ equals $2n\mathcal{S}_{\lambda_n}(\mathcal{H}, \mathcal{D}_n)$ with $\lambda_n = \frac{\log(n)}{2n}$.

The following lemma shows that when the distribution of $\mathbf{X}$ is multivariate Gaussian and λ is suitably chosen, the ℓ_0-penalized log-likelihood score of a DAG $\mathcal{H}$ can be improved by adding an edge $X_i \to X_j$ if and only if the sample partial correlation between X_i and X_j given $\mathbf{Pa}_{\mathcal{H}}(X_j)$ is nonzero. This is one of the key results for our proof of high-dimensional consistency of (AR)GES.

LEMMA 5.1. *Let $\mathcal{H} = (\mathbf{X}, E)$ be a DAG such that $X_i \in \mathbf{Nd}_{\mathcal{H}}(X_k) \setminus \mathbf{Pa}_{\mathcal{H}}(X_k)$. Let $\mathcal{H}' = (\mathbf{X}, E \cup \{X_i \to X_k\})$. If the distribution of $\mathbf{X}$ is multivariate Gaussian, then the ℓ_0-penalized log-likelihood score difference between $\mathcal{H}'$ and $\mathcal{H}$ is*

$$(2) \qquad \mathcal{S}_{\lambda}\big(\mathcal{H}', \mathcal{D}_n\big) - \mathcal{S}_{\lambda}\big(\mathcal{H}, \mathcal{D}_n\big) = \frac{1}{2}\log\big(1 - \hat{\rho}^2_{ik|\mathbf{Pa}_{\mathcal{H}}(k)}\big) + \lambda,$$

where $\hat{\rho}_{ik|\mathbf{Pa}_{\mathcal{H}}(k)}$ denotes the sample partial correlation between X_i and X_k given $\mathbf{Pa}_{\mathcal{H}}(X_k)$.

The first term on the right-hand side of (2) equals the negative of the conditional mutual information between X_i and X_k given $\mathbf{Pa}_{\mathcal{H}}(X_k)$. Thus, Lemma 5.1 shows that score-based methods like GES essentially use conditional independence tests [cf. Anandkumar et al. (2012)] for sequentially adding and deleting edges starting from an initial graph. This shows that the score-based GES algorithm and the constraint-based PC algorithm use the same basic principle (a conditional independence test) in the multivariate Gaussian setting. Although this connection (see Section 9 for more details) between PC and GES is not very surprising, we were unable to find it in the literature.

Further, Lemma 5.1 also opens the possibility to define generalized scoring criterions, by replacing the Gaussian conditional mutual information in (2) with a more general measure of conditional independence. In fact, we exploit this in Section 6 to extend our high-dimensional consistency results to nonparanormal distributions.

We define an oracle version of the ℓ_0-penalized log-likelihood scoring criterion by replacing the log-likelihood in Definition 5.1 by its expectation with respect to the distribution F of $\mathbf{X}$. The oracle score given by Definition 5.2 will be used to define a collection of oracle versions of (AR)GES in Section 5.2.

DEFINITION 5.2.   (Oracle score) Let $\mathcal{H} = (\mathbf{X}, E)$ be a DAG. We define the oracle score of $\mathcal{H}$ with respect to the distribution F of $\mathbf{X}$ as

$$\mathcal{S}_\lambda^*(\mathcal{H}, F) = -\sum_{i=1}^p \mathrm{E}\big[\log\big(L\big(\boldsymbol{\theta}_i^*(\mathcal{H}), X_i | \mathbf{Pa}_\mathcal{H}(X_i)\big)\big)\big] + \lambda|E|,$$

where $L(\boldsymbol{\theta}_i, X_i | \mathbf{Pa}_\mathcal{H}(X_i))$ is the likelihood function that corresponds to the conditional density of X_i given $\mathbf{Pa}_\mathcal{H}(X_i)$ and

$$\boldsymbol{\theta}_i^*(\mathcal{H}) = \underset{\boldsymbol{\theta}_i}{\operatorname{argmax}}\, \mathrm{E}\big[\log\big(L\big(\boldsymbol{\theta}_i, X_i | \mathbf{Pa}_\mathcal{H}(X_i)\big)\big)\big].$$

We note that both $\mathcal{S}_\lambda$ and $\mathcal{S}_\lambda^*$ are decomposable (see Definition 2.2 of the Supplementary Material). Moreover, they are score equivalent (see Definition 2.1 of the Supplementary Material) when the distribution F is multivariate Gaussian. The scoring criterion used to compute the large sample outputs in Example 1 equals $\mathcal{S}_\lambda^*$ with $\lambda = 0$ (see Section 4.2 of the Supplementary Material).

The following lemma is analogous to Lemma 5.1. We do not provide a proof of Lemma 5.2, since it can be obtained from the proof of Lemma 5.1 by replacing the sample quantities by the corresponding population quantities (e.g., averages should be replaced by expectations and sample regression coefficients should be replaced by their population counterparts).

LEMMA 5.2.   *Let $\mathcal{H} = (\mathbf{X}, E)$ be a DAG such that $X_i \in \mathbf{Nd}_\mathcal{H}(X_k) \setminus \mathbf{Pa}_\mathcal{H}(X_k)$. Let $\mathcal{H}' = (\mathbf{X}, E \cup \{X_i \rightarrow X_k\})$. If the distribution $\mathbf{X}$ is multivariate Gaussian, then the oracle score difference between $\mathcal{H}'$ and $\mathcal{H}$ is*

$$\mathcal{S}_\lambda^*(\mathcal{H}', F) - \mathcal{S}_\lambda^*(\mathcal{H}, F) = \frac{1}{2}\log\big(1 - \rho_{ik|\mathbf{Pa}_\mathcal{H}(k)}^2\big) + \lambda,$$

where $\rho_{ik|\mathbf{Pa}_\mathcal{H}(k)}$ denotes the partial correlation between X_i and X_k given $\mathbf{Pa}_\mathcal{H}(X_k)$.

5.2. *High-dimensional consistency of (AR)GES.*   First, we define a collection of oracle versions of (AR)GES using the oracle scoring criterion $\mathcal{S}_{\lambda_n}^*$ (Definition 5.2). Every move in the forward or backward phase of (AR)GES from a CPDAG $\mathcal{C}_{\text{current}}$ to $\mathcal{C}_{\text{new}}$ corresponds to an edge addition or edge deletion in a DAG in the Markov equivalence class of $\mathcal{C}_{\text{current}}$, and thus corresponds to a partial correlation, by Lemma 5.2. We denote the partial correlation associated with a move from $\mathcal{C}_{\text{current}}$ to $\mathcal{C}_{\text{new}}$ by $\rho(\mathcal{C}_{\text{current}}, \mathcal{C}_{\text{new}})$.

At every step in the forward phase of (AR)GES, an optimal CPDAG is chosen among a set of possible choices for the next step (see line 6 of Algorithm 4.1). These optimal choices in the forward phase are not crucial for consistency of (AR)GES [Chickering and Meek (2002)]. Thus, we define below oracle versions of (AR)GES that allow suboptimal choices for edge additions. One of our assumptions [see (A5) below] will be based on this definition.

DEFINITION 5.3 (δ-optimal oracle version of (AR)GES).   Let $\delta \in [0, 1]$. A δ-optimal oracle version of (AR)GES with scoring criterion $\mathcal{S}_\lambda^*$ consists of two phases: a δ-optimal oracle forward phase and an oracle backward phase. A δ-optimal oracle forward phase of (AR)GES is Algorithm 4.1 based on $\mathcal{I}_0$ (for ARGES-CIG) or skeleton($\mathcal{C}_0$) (for ARGES-skeleton) or the complete undirected graph (for GES), using the oracle scoring criterion $\mathcal{S}_\lambda^*$, where at each step with $\mathfrak{C} \neq \varnothing$, $\mathcal{C}_{\text{new}}$ is chosen to be any CPDAG in $\mathfrak{C}$ (see line 6 of Algorithm 4.1) such that

$$\left| \rho(\mathcal{C}_{\text{current}}, \mathcal{C}_{\text{new}}) \right| \geq \max_{\mathcal{C} \in \mathfrak{C}} \left| \rho(\mathcal{C}_{\text{current}}, \mathcal{C}) \right| - \delta.$$

An oracle backward phase of (AR)GES equals the backward phase of GES (Algorithm 3.2 of the Supplementary Material), using the oracle scoring criterion $\mathcal{S}_\lambda^*$. At each step with $\mathfrak{C} \neq \varnothing$, if there are several CPDAGs with the same optimal score, then one of these is chosen arbitrarily as $\mathcal{C}_{\text{new}}$ (see line 6 of Algorithm 3.2 of the Supplementary Material).

THEOREM 5.1 (Soundness).   *Assume that the distribution of* $\mathbf{X}$ *is multivariate Gaussian and DAG-perfect. Let* $\delta \in [0, 1]$. *Let m be such that the maximum degree in the output of the forward phase of every* δ-optimal oracle version of (AR)GES *with scoring criterion* $\mathcal{S}_\lambda^*$ *is bounded by m for all* $\lambda \geq 0$. *If* $\lambda < -\frac{1}{2}\log(1 - \rho_{ij|S}^2)$ *for all* $i, j \in \{1, \ldots, p\}$ *and* $S \subseteq \{1, \ldots, p\} \setminus \{i, j\}$ *such that* $|S| \leq m$ *and* $\rho_{ij|S} \neq 0$, *then the outputs of all* δ-optimal oracle versions of (AR)GES *with scoring criterion* $\mathcal{S}_\lambda^*$ *are identical and equal to* $\mathcal{C}_0$.

Note that the edge additions in the forward phase of a δ-optimal oracle version of (AR)GES are only slightly suboptimal for small values of δ. In fact, we let δ_n tend to zero as $n \to \infty$ in assumption (A5) below. For $\delta = 0$, we refer to the forward phase of the δ-optimal oracle version of (AR)GES as the oracle forward phase of (AR)GES.

We now consider an asymptotic scenario where the number of variables in $\mathbf{X}$ and the distribution of $\mathbf{X}$ are allowed to change with the sample size n. Thus, let $\{\mathbf{X}_n\}$ be a sequence of random vectors such that the distribution of each $\mathbf{X}_n$ is multivariate Gaussian and DAG-perfect. Further, we slightly modify Notation 2.1 as follows.

NOTATION 5.1.   We reserve the notation F_n, $\mathcal{G}_{n0}$, $\mathcal{C}_{n0}$, $\mathcal{I}_{n0}$ and $\mathcal{D}_n$ for the following: F_n denotes the distribution of $\mathbf{X}_n = (X_{n1}, \ldots, X_{np_n})^T$, $\mathcal{G}_{n0}$ denotes a perfect map of F_n, $\mathcal{C}_{n0} = \text{CPDAG}(\mathcal{G}_{n0})$ is the corresponding CPDAG, $\mathcal{I}_{n0}$ is the CIG of F_n, and $\mathcal{D}_n$ denotes the data, consisting of n i.i.d. observations from F_n.

We make the following assumptions to prove high-dimensional consistency of (AR)GES.

(A1) (Gaussianity) The distribution of $\mathbf{X}_n$ is multivariate Gaussian and DAG-perfect for all n.

(A2) (high-dimensional setting) $p_n = \mathcal{O}(n^a)$ for some $0 \leq a < \infty$.

(A3) (sparsity condition) Let $q_n = \max_{1 \leq i \leq p_n} |\mathbf{Adj}_{\mathcal{C}_{n0}}(X_{ni})|$ be the maximum degree in $\mathcal{C}_{n0}$. Then $q_n = \mathcal{O}(n^{1-b_1})$ for some $0 < b_1 \leq 1$.

(A4) (consistent estimators of the CIG or the CPDAG-skeleton) There exists a sequence of estimators $\hat{\mathcal{I}}_n$ (for ARGES-CIG) or a sequence of estimators $\hat{\mathcal{U}}_n$ (for ARGES-skeleton) such that

$$\lim_{n \to \infty} \mathbb{P}(\hat{\mathcal{I}}_n = \mathcal{I}_{n0}) = 1 \quad \text{or} \quad \lim_{n \to \infty} \mathbb{P}(\hat{\mathcal{U}}_n = \text{skeleton}(\mathcal{C}_{n0})) = 1.$$

(A5) (bounds on the growth of oracle versions) The maximum degree in the output of the forward phase of every δ_n-optimal oracle version of (AR)GES with scoring criterion $\mathcal{S}^*_{\lambda_n}$ is bounded by $K_n q_n$, for all $\lambda_n \geq 0$ and some sequences $\delta_n \to 0$ and $\{K_n\}$ such that $\delta_n^{-1} = \mathcal{O}(n^{d_1})$ and $K_n = \mathcal{O}(n^{b_1 - b_2})$ for some constants b_2 and d_1 satisfying $0 \leq d_1 < b_2/2 \leq 1/2$, where q_n is given by (A3).

(A6) (bounds on partial correlations) The partial correlations $\rho_{nij|S}$ between X_{ni} and X_{nj} given $\{X_{nr} : r \in S\}$ satisfy the following upper and lower bounds for all n, uniformly over $i, j \in \{1, \ldots, p_n\}$ and $S \subseteq \{1, \ldots, p_n\} \setminus \{i, j\}$ such that $|S| \leq K_n q_n$ [where K_n and q_n are from (A3) and (A5)]:

$$\sup_{i \neq j, S} |\rho_{nij|S}| \leq M < 1 \quad \text{and} \quad \inf_{i,j,S} \{|\rho_{nij|S}| : \rho_{nij|S} \neq 0\} \geq c_n,$$

for some sequence $c_n \to 0$ such that $c_n^{-1} = \mathcal{O}(n^{d_2})$ for some constant d_2 satisfying $0 < d_2 < b_2/2$, where b_2 is given by (A5).

Kalisch and Bühlmann (2007) proved high dimensional consistency of the PC algorithm assuming (A1), (A2), (A3) and a slightly weaker version of (A6). More precisely, the authors assumed (A6) with $d_2 < b_1/2$. The most criticized assumption among these four assumptions is probably (A6) [Uhler et al. (2013)], which is also known as the strong faithfulness condition. Interestingly, van de Geer and Bühlmann (2013) proved high-dimensional consistency of the global optimum of the ℓ_0-penalized likelihood score function without assuming strong faithfulness, but assuming a permutation beta-min condition. Unfortunately, we cannot guarantee that a greedy search method like (AR)GES can always find the global optimum without (A6) and thus we cannot substitute the strong faithfulness assumption here.

We make two additional assumptions compared to Kalisch and Bühlmann (2007), namely (A4) and (A5). Assumption (A4) is natural and it is not a strong assumption, since there are various estimation methods for the CIG or the CPDAG-skeleton that are consistent in sparse high-dimensional settings [e.g., Meinshausen and Bühlmann (2006), Banerjee, El Ghaoui and d'Aspremont (2008), Friedman, Hastie and Tibshirani (2008), Ravikumar et al. (2008, 2011), Cai, Liu and Luo (2011), Kalisch and Bühlmann (2007), Ha, Sun and Xie (2016)]. We will discuss (A5) in Section 5.3 below.

THEOREM 5.2.   *Assume* (A1)–(A6). *Let* $\hat{\mathcal{C}}_n$, $\check{\mathcal{C}}_n$ *and* $\tilde{\mathcal{C}}_n$ *be the outputs of ARGES-CIG based on* $\hat{\mathcal{I}}_n$, *ARGES-skeleton based on* $\hat{\mathcal{U}}_n$ *and GES respectively, with the scoring criterion* $\mathcal{S}_{\lambda_n}$. *Then there exists a sequence* $\lambda_n \to 0$ *such that*

$$\lim_{n\to\infty} \mathbb{P}(\hat{\mathcal{C}}_n = \mathcal{C}_{n0}) = \lim_{n\to\infty} \mathbb{P}(\check{\mathcal{C}}_n = \mathcal{C}_{n0}) = \lim_{n\to\infty} \mathbb{P}(\tilde{\mathcal{C}}_n = \mathcal{C}_{n0}) = 1.$$

We choose $\lambda_n = \frac{1}{9} \log(1 - c_n^2)$ to prove the above theorems (see Section 9.7 of the Supplementary Material), where c_n is as in (A6). However, similar arguments hold for any sequence $\{\lambda_n\}$ satisfying $\lambda_n < \frac{1}{8} \log(1 - c_n^2)$ and $\frac{K_n q_n \log(p_n)}{n\lambda_n} \to 0$, where p_n, q_n and K_n are as in (A2), (A3) and (A5) respectively. The penalty parameter of the BIC criterion (see Remark 5.1), which is known to be weak for sparse high-dimensional settings [see, for example, Foygel and Drton (2010)], does not satisfy $\frac{K_n q_n \log(p_n)}{n\lambda_n} \to 0$, except for some very restricted choices of $\{p_n\}$, $\{q_n\}$ and $\{K_n\}$. Further, we note that van de Geer and Bühlmann (2013) proved their high-dimensional consistency result with $\lambda_n = \mathcal{O}(\sqrt{\log(p_n)/n})$. This choice of $\{\lambda_n\}$ satisfies $\frac{K_n q_n \log(p_n)}{n\lambda_n} \to 0$ under (A1)–(A6), when the constant b_2 in (A5) is greater than $1/2$.

5.3. *Discussion on assumption* (A5).   Note that for every δ_n-optimal forward phase of (AR)GES with scoring criterion $\mathcal{S}_{\lambda_n}^*$, the maximum degree in the output increases or remains unchanged as λ_n decreases. Thus, without loss of generality, we fix the scoring criterion for all δ_n-optimal versions of (AR)GES to be $\mathcal{S}_0^*$ in (A5) and in the remainder of this subsection.

We first consider (A5) for $\{\delta_n\} = 0$. Then (A3) and (A5) together imply that the output of the oracle forward phase of (AR)GES is bounded by $K_n q_n = \mathcal{O}(n^{1-b_2})$, where K_n is an upper bound on the ratio of the maximum degrees in the output of the oracle forward phase of (AR)GES and in $\mathcal{C}_{n0}$. It follows from the soundness of the oracle version of (AR)GES that $K_n \geq 1$ and hence $b_2 \leq b_1$, where b_1 is given by (A3). Therefore, (A5) roughly states that the oracle forward phase of (AR)GES does not add "too many" edges that do not belong to the CPDAG-skeleton. In Section 5 of the Supplementary Material, we empirically verify this in certain sparse high-dimensional settings and we find that in all but very few cases the maximum degree of the outputs of the oracle forward phase of (AR)GES are reasonably close to the maximum degree of the true CPDAG-skeleton. Further, note that (A5) is slightly different for GES, ARGES-CIG and ARGES-skeleton (see Figure 5 of the Supplementary Material).

The intuition for having a δ_n-optimal oracle version in (A5), with $\delta_n > 0$ is the following: If

$$\left|\rho(\mathcal{C}_{\text{current}}, \mathcal{C}_{\text{new}})\right| \geq \max_{\mathcal{C}\in\mathfrak{C}} \left|\rho(\mathcal{C}_{\text{current}}, \mathcal{C})\right| - \delta_n$$

(see Definition 5.3), then the move from $\mathcal{C}_{\text{current}}$ to $\mathcal{C}_{\text{new}}$ is so close to the optimal move that it cannot be identified as suboptimal in the sample version.

We note that the output of a δ-optimal oracle forward phase of (AR)GES depends not only on the structure $\mathcal{C}_0$ but also on the absolute values of the nonzero partial correlations (even for $\delta = 0$). The latter makes it very difficult to characterize a family of distributions for which (A5) holds. However, we provide two strong structural conditions under which (A5) holds. The conditions are based on the following two results.

LEMMA 5.3. *Assume that the distribution of* $\mathbf{X}$ *is multivariate Gaussian and DAG-perfect. Let* X_i *and* X_j *be two nodes in* $\mathcal{C}_0$. *If there is no path between* X_i *and* X_j *in* $\mathcal{C}_0$, *then the output of a* δ-*optimal oracle forward phase of* (AR)GES *does not contain an edge between* X_i *and* X_j.

THEOREM 5.3. *Assume that the distribution of* $\mathbf{X}$ *is multivariate Gaussian and DAG-perfect. If* skeleton($\mathcal{C}_0$) *is a forest* (*i.e., contains no cycle*), *then the output of the oracle forward phase of* (AR)GES *equals* $\mathcal{C}_0$.

The proof of Lemma 5.3 is trivial, since if X_i and X_j are not connected by a path in $\mathcal{C}_0$, then $\rho_{ij|S} = 0$ for all $S \subseteq \{1, \ldots, p\} \setminus \{i, j\}$. Further, note that Lemma 5.3 implies that it is sufficient to prove Theorem 5.3 when skeleton($\mathcal{C}_0$) is an undirected tree (i.e., a connected graph containing no cycle) or equivalently, when $\mathcal{G}_0$ is a polytree (i.e., a directed graph whose skeleton is an undirected tree).

We note that Theorem 5.3 shows a connection between the oracle forward phase of GES and the Chow-Liu algorithm [Chow and Liu (1968)] for multivariate Gaussian distributions. The Chow-Liu algorithm is a greedy forward search procedure for learning optimal undirected trees, based on mutual information between pairs of variables [which equals $-\frac{1}{2} \log(1 - \rho_{ij}^2)$ for a multivariate Gaussian distribution]. Theorem 5.3 shows that the oracle forward phase of GES is a greedy forward search procedure for learning optimal polytrees, based on conditional mutual information $-\frac{1}{2} \log(1 - \rho_{ij|S}^2)$. To our knowledge, this connection between GES and the Chow-Liu algorithm cannot be found in the literature on the Chow-Liu algorithm and extensions thereof for learning polytrees [Rebane and Pearl (1987), Huete and de Campos (1993), de Campos (1998), Ouerd, Oommen and Matwin (2004)].

We are now ready to state our sufficient conditions for assumption (A5) as the following immediate corollaries of Lemma 5.3 and Theorem 5.3.

COROLLARY 5.1. *Assume* (A1). *If the number of nodes in each connected component of* $\mathcal{C}_{n0}$ *is bounded by* q'_n *satisfying* $q'_n = \mathcal{O}(n^{a'})$ *for some* $0 \leq a' < 1$, *then the maximum degree in the output of a* δ_n-*optimal oracle forward phase of* (AR)GES *is bounded by* q'_n *for all* $\delta_n \geq 0$.

COROLLARY 5.2. *Assume* (A1) *and* (A3). *If* skeleton($\mathcal{C}_{n0}$) *is a forest, then the maximum degree in the output of the oracle forward phase of* (AR)GES *equals* q_n, *where* q_n *is given by* (A3).

In order to extend Theorem 5.3 or Corollary 5.2 to a δ_n-optimal version of (AR)GES, we need the following additional assumption.

(A7) (bounds on the gaps between marginal correlations and first order partial correlations)

$$\inf_{(i,j,k)\in T}\{||\rho_{nij|k}| - |\rho_{nij}|| : |\rho_{nij|k}| \neq |\rho_{nij}|\} \geq c_n',$$

where $T = \{(i, j, k) : (X_{ni}, X_{nj}, X_{nk})$ is an unshielded triple in $\mathcal{C}_{n0}\}$ and $c_n' \to 0$ such that $c_n'^{-1} = \mathcal{O}(n^{d_3})$ for some $0 \leq d_3 < b_1/2$ where b_1 is given by (A3).

To understand (A7), assume that skeleton$(\mathcal{C}_{n0})$ is a forest containing an unshielded triple (X_{ni}, X_{nj}, X_{nk}) such that $\pi_{\mathcal{C}_{n0}}(X_{ni}, X_{nj}, X_{nk})$ is either $X_{ni} \to X_{nj} \leftarrow X_{nk}$ or $X_{ni} - X_{nj} - X_{nk}$. Let $\mathcal{C}_n$ be the CPDAG representing the Markov equivalence class of the DAG $\mathcal{G}_n$ obtained by deleting the edge between X_{ni} and X_{nj} in $\mathcal{G}_{n0}$. Note that $\rho_{nij|k}$ corresponds to the improvement in the oracle score for adding the edge $X_{ni} \to X_{nj}$ in a DAG in the Markov equivalence class of $\mathcal{C}_n$ while creating a new v-structure $X_{ni} \to X_{nj} \leftarrow X_{nk}$, whereas ρ_{nij} corresponds to the improvement in the oracle score for adding the edge $X_{ni} \to X_{nj}$ in a DAG in the Markov equivalence class of $\mathcal{C}_n$ without creating the v-structure $X_{ni} \to X_{nj} \leftarrow X_{nk}$. Thus (A7) ensures that for each $\delta_n < c_n'$, a δ_n-optimal oracle forward phase of (AR)GES would move from $\mathcal{C}_n$ to $\mathcal{C}_{n0}$ by correctly deciding if a new v-structure should be created.

THEOREM 5.4. *Assume* (A1), (A3), (A6) *with* $K_n = 1$ *and* (A7). *If* skeleton$(\mathcal{C}_{n0})$ *is a forest, then the output of a δ_n-optimal oracle forward phase of (AR)GES equals $\mathcal{C}_{n0}$ for all $\delta_n < \min((1 - M)c_n, c_n')$, where M and c_n are given by* (A6) *and c_n' is given by* (A7). *Hence,* (A5) *holds with $b_2 = b_1$ and $d_1 = \max(d_2, d_3)$, where b_1 is given by* (A3).

REMARK 5.2. An obvious extension of Corollary 5.1 and Theorem 5.4 is the following: if the assumption of Corollary 5.1 holds for the connected components of skeleton$(\mathcal{C}_{n0})$ that are not trees and the assumptions of Theorem 5.4 hold for the connected components of skeleton$(\mathcal{C}_{n0})$ that are trees, then (A5) holds with $b_2 = a'$ and $d_1 = \max(d_2, d_3)$.

6. High-dimensional consistency of GES and ARGES for linear structural equation models. In this section, we present a slightly modified version of the result from the previous section: we prove high-dimensional consistency of (AR)GES for linear structural equation models with sub-Gaussian error variables.

DEFINITION 6.1. Let $\mathcal{G}_0 = (\mathbf{X}, E)$ be a DAG and let $B_{\mathcal{G}_0}$ be a $p \times p$ matrix such that $(B_{\mathcal{G}_0})_{ji} \neq 0$ if and only if $X_i \in \mathbf{Pa}_{\mathcal{G}_0}(X_j)$. Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)^T$ be a random vector of jointly independent error variables. Then $\mathbf{X} = (X_1, \ldots, X_p)^T$ is said to be generated from a linear structural equation model (linear SEM) characterized by $(B_{\mathcal{G}_0}, \boldsymbol{\varepsilon})$ if $\mathbf{X} = B_{\mathcal{G}_0}\mathbf{X} + \boldsymbol{\varepsilon}$.

We assume the distribution of $\mathbf{X}$ is faithful to $\mathcal{G}_0$, implying that $\mathcal{G}_0$ is a perfect map of the distribution of $\mathbf{X}$. We refer to this as $\mathbf{X}$ is generated from a DAG-perfect linear SEM.

Note that if $\varepsilon_1, \ldots, \varepsilon_p$ are Gaussian random variables, then the joint distribution of $\mathbf{X}$ is multivariate Gaussian with covariance matrix $\Sigma_0 = (I - B_{\mathcal{G}_0})^{-1} \text{Cov}(\boldsymbol{\varepsilon})(I - B_{\mathcal{G}_0})^{-T}$. In this case, for any $i \neq j$ and $S \subseteq \{1, \ldots, p\} \setminus \{i, j\}$, we have

$$(3) \qquad X_i \perp_{\mathcal{G}_0} X_j \mid \{X_r : r \in S\} \quad \Longleftrightarrow \quad \rho_{ij|S} = 0.$$

Since $\rho_{ij|S} = 0$ depends on the distribution of $\mathbf{X}$ via Σ_0, (3) holds regardless of the distribution of $\boldsymbol{\varepsilon}$ [see Spirtes et al. (1998)]. Therefore, if $\mathbf{X}$ is generated from a DAG-perfect linear SEM, then $X_i \perp\!\!\!\perp X_j \mid \{X_r : r \in S\}$ if and only if $\rho_{ij|S} = 0$. Consequently, Lemma 5.2 and Theorem 5.1 imply the soundness of the (AR)GES algorithm with the ℓ_0-penalized Gaussian log-likelihood oracle score $\mathcal{S}_\lambda^*(\cdot, \tilde{F})$ (Definition 5.2) for a DAG-perfect linear SEM characterized by $(B_{\mathcal{G}_0}, \boldsymbol{\varepsilon})$, where $\tilde{F}$ is the distribution of a Gaussian random vector with covariance matrix $\Sigma_0 = (I - B_{\mathcal{G}_0})^{-1} \text{Cov}(\boldsymbol{\varepsilon})(I - B_{\mathcal{G}_0})^{-T}$.

We use Notation 5.1 to present the following high-dimensional consistency result for linear SEMs with sub-Gaussian error variables. We replace (A1) given in Section 5.2 by (A1*) below, and we make an additional assumption.

(A1*) For each n, $\mathbf{X}_n$ is generated from a DAG-perfect linear SEM with sub-Gaussian error variables satisfying $\|\varepsilon_{ni}\|_{\psi_2} \leq C_1 \sqrt{\text{Var}(\varepsilon_{ni})}$ for some absolute constant $C_1 > 0$, where $\|\cdot\|_{\psi_2}$ denotes the sub-Gaussian norm.

(A8) For any $(K_n q_n + 2) \times (K_n q_n + 2)$ principal submatrix Σ_n of $\Sigma_{n0} = \text{Cov}(\mathbf{X}_n)$,

$$C_2 \leq 1/\|\Sigma_n^{-1}\|_2 \leq \|\Sigma_n\|_2 \leq C_3$$

for some absolute constants $C_2, C_3 > 0$, where $\|\cdot\|_2$ denotes the spectral norm and q_n and K_n are given by (A3) and (A5) respectively.

THEOREM 6.1. *Assume* (A1*), (A8) *and* (A2)–(A6) *given in Section* 5.2. *Then there exists a sequence* $\lambda_n \to 0$ *such that*

$$\lim_{n \to \infty} \mathbb{P}(\hat{\mathcal{C}}_n = \mathcal{C}_{n0}) = \lim_{n \to \infty} \mathbb{P}(\check{\mathcal{C}}_n = \mathcal{C}_{n0}) = \lim_{n \to \infty} \mathbb{P}(\tilde{\mathcal{C}}_n = \mathcal{C}_{n0}) = 1,$$

where $\hat{\mathcal{C}}_n, \check{\mathcal{C}}_n$ *and* $\tilde{\mathcal{C}}_n$ *are the outputs of ARGES-CIG and ARGES-skeleton and GES respectively, with the* ℓ_0-*penalized Gaussian log-likelihood scoring criterion* $\hat{\mathcal{S}}_{\lambda_n}$ (*see Definition* 5.1).

REMARK 6.1. As in Harris and Drton (2013), we can allow an ultra high-dimensional setting by replacing (A2) by $p_n = \mathcal{O}(\exp(n^a))$ (for some $0 \leq a < 1$) and allow an $\mathcal{O}(n^f)$ (for some $0 \leq f < 1/4$) growth rate of $\|\Sigma_n^{-1}\|_2$ for each $(K_n q_n + 2) \times (K_n q_n + 2)$ submatrix Σ_n of Σ_{n0} under the additional restriction that $a + 4f < b_2 - 2 \max(d_1, d_2)$.

REMARK 6.2.   We note that the results of Section 5.3 continue to hold if we replace the multivariate Gaussian assumption by the assumption that $\mathbf{X}_n$ is generated from a linear SEM with arbitrary error variables.

REMARK 6.3.   We note that a similar high-dimensional consistency result holds for the PC algorithm when the sample partial correlations are used for testing conditional independence.

7. High-dimensional consistency of GES and ARGES in the nonparanormal setting.   In this section, we prove high-dimensional consistency under the assumption that each $\mathbf{X}_n$ has a nonparanormal distribution.

DEFINITION 7.1 [Harris and Drton (2013)].   Let $\Sigma_0 = (\rho_{ij})$ be a positive definite correlation matrix and let $\mathbf{Z} = (Z_1, \ldots, Z_p)^T \sim N(0, \Sigma_0)$. For a collection of strictly increasing functions $\mathbf{g} = (g_1, \ldots, g_p)^T$, the random vector $\mathbf{X} = (g_1(Z_1), \ldots, g_p(Z_p))^T$ has a nonparanormal distribution NPN$(\mathbf{g}, \Sigma_0)$.

A variant of the PC algorithm, called Rank PC, was shown to be consistent in high-dimensional settings under the nonparanormal distribution assumption [Harris and Drton (2013)]. First, we briefly discuss the Rank PC algorithm. Then we define a new scoring criterion motivated by the rank-based measures of correlations used in the rank PC algorithm. Finally, we prove high-dimensional consistency of (AR)GES with this scoring criterion in the nonparanormal setting.

Let Σ_0, $\mathbf{Z}$, $\mathbf{g}$, and $\mathbf{X}$ be as in Definition 7.1. First, note that since the marginal transformations g_i are deterministic, for any $i \neq j$ and $S \subseteq \{1, \ldots, p\} \setminus \{i, j\}$

$$
\begin{aligned}
X_i \perp\!\!\!\perp X_j \mid \{X_r : r \in S\} &\iff Z_i \perp\!\!\!\perp Z_j \mid \{Z_r : r \in S\} \\
&\iff \rho_{ij|S} = 0,
\end{aligned}
$$

(4)

where $\rho_{ij|S}$ is the partial correlation coefficient between Z_i and Z_j given $\{Z_r : r \in S\}$. Next, note that since g_i are strictly increasing functions, a sample rank correlation coefficient (Spearman's ρ or Kendall's τ) between X_i and X_j (denoted as $\hat{\rho}_{ij}^S$ or $\hat{\rho}_{ij}^K$) is identical to the corresponding rank correlation between Z_i and Z_j. Further, Liu et al. (2012) showed that $2\sin(\frac{\pi}{6}\hat{\rho}_{ij}^S)$ and $\sin(\frac{\pi}{2}\hat{\rho}_{ij}^K)$ are consistent estimators of ρ_{ij}.

In the remainder of this section, we generically denote a rank based estimator $2\sin(\frac{\pi}{6}\hat{\rho}_{ij}^S)$ or $\sin(\frac{\pi}{2}\hat{\rho}_{ij}^K)$ by $\hat{\rho}_{ij}$. We denote the corresponding estimator of the correlation matrix Σ_0 by $\hat{\Sigma} = (\hat{\rho}_{ij})$. Following Harris and Drton (2013), we define rank based estimators of partial correlations $\rho_{ij|S}$ through the matrix inversion formula

(5)
$$
\hat{\rho}_{ij|S} := -\frac{\hat{\Psi}_{12}^{-1}}{\sqrt{\hat{\Psi}_{11}^{-1}\hat{\Psi}_{22}^{-1}}},
$$

where $\hat{\Psi}$ is the submatrix of $\hat{\Sigma}$ that corresponds to X_i, X_j and $\{X_r : r \in S\}$ in this order, and $\hat{\Psi}_{st}^{-1}$ denotes the (s, t)th entry of $\hat{\Psi}^{-1}$.

Given the estimators $\hat{\rho}_{ij|S}$ defined above, the rank PC algorithm is just the PC algorithm based on the following conditional independence tests: reject the null hypothesis $X_i \perp\!\!\!\perp X_j \mid \{X_r : r \in S\}$ if and only if $|\hat{\rho}_{ij|S}| > \nu$, where the critical value ν is chosen to be the same for each individual test, and it is a tuning parameter of the algorithm that controls sparsity of the output.

We now define the scoring criterion $\tilde{\mathcal{S}}_{\lambda_n}$ (Definition 7.2), which is motivated by Lemma 5.1. We show below that (AR)GES with $\tilde{\mathcal{S}}_{\lambda_n}$ is consistent in certain sparse high-dimensional settings with nonparanormal distributions.

DEFINITION 7.2. We define the scoring criterion $\tilde{\mathcal{S}}_{\lambda_n}$ by setting the score of an empty DAG to zero, and by defining the score difference between two DAGs that differ by exactly one edge as follows. Let $\mathcal{H}_n = (\mathbf{X}_n, E_n)$ be a DAG such that $X_{ni} \in \mathbf{Nd}_{\mathcal{H}_n}(X_{nk}) \setminus \mathbf{Pa}_{\mathcal{H}_n}(X_{nk})$. Let $\mathcal{H}_n' = (\mathbf{X}_n, E_n \cup \{X_{ni} \to X_{nk}\})$. Then

$$(6) \qquad \tilde{\mathcal{S}}_{\lambda_n}(\mathcal{H}_n', \mathcal{D}_n) - \tilde{\mathcal{S}}_{\lambda_n}(\mathcal{H}_n, \mathcal{D}_n) := \frac{1}{2} \log(1 - \hat{\rho}_{nik|\mathbf{Pa}_{\mathcal{H}_n}(k)}^2) + \lambda_n,$$

where $\mathbf{Pa}_{\mathcal{H}_n}(k) = \{j : X_{nj} \in \mathbf{Pa}_{\mathcal{H}_n}(X_{nk})\}$ and $\hat{\rho}_{nik|\mathbf{Pa}_{\mathcal{H}_n}(k)}$ is a rank correlation based estimate defined by (5).

For a DAG $\mathcal{H}_n = (\mathbf{X}_n, E)$, the score $\tilde{\mathcal{S}}_{\lambda_n}(\mathcal{H}_n, \mathcal{D}_n)$ can be obtained by summing up the score differences while sequentially adding directed edges from E starting from the empty graph.

LEMMA 7.1. $\tilde{\mathcal{S}}_{\lambda_n}(\mathcal{H}_n, \mathcal{D}_n)$ is well defined for all DAGs $\mathcal{H}_n$, that is, the score does not depend on the order in which the directed edges are added to the empty graph.

LEMMA 7.2. $\tilde{\mathcal{S}}_{\lambda_n}$ is score equivalent, that is, $\tilde{\mathcal{S}}_{\lambda_n}(\mathcal{H}_n', \mathcal{D}_n) = \tilde{\mathcal{S}}_{\lambda_n}(\mathcal{H}_n, \mathcal{D}_n)$ for any two Markov equivalent DAGs $\mathcal{H}_n$ and $\mathcal{H}_n'$, and for all $\mathcal{D}_n$.

For a nonparanormal distribution $\mathrm{NPN}(\mathbf{g}_n, \Sigma_{n0})$, we define the oracle score $\tilde{\mathcal{S}}_{\lambda_n}^*(\mathcal{H}_n, \Sigma_{n0})$ by replacing (6) in Definition 7.2 with the following:

$$\tilde{\mathcal{S}}_{\lambda_n}^*(\mathcal{H}_n', \Sigma_{n0}) - \tilde{\mathcal{S}}_{\lambda_n}^*(\mathcal{H}_n, \Sigma_{n0}) := \frac{1}{2} \log(1 - \rho_{nik|\mathbf{Pa}_{\mathcal{H}_n}(k)}^2) + \lambda_n,$$

where the partial correlations are given by Σ_{n0}. In fact, $\tilde{\mathcal{S}}_{\lambda_n}^*(\cdot, \Sigma_{n0})$ is identical to the scoring criterion $\mathcal{S}_{\lambda_n}^*(\cdot, \tilde{F}_n)$ (Definition 5.2), where $\tilde{F}_n$ is the distribution of $\mathbf{Z}_n \sim N(\mathbf{0}, \Sigma_{n0})$.

THEOREM 7.1. Assume that the distribution of $\mathbf{X}_n$ is $\mathrm{NPN}(\mathbf{g}_n, \Sigma_{n0})$ and DAG-perfect. Assume (A2)–(A6) given in Section 5.2, with δ_n-optimal oracle versions based on $\tilde{\mathcal{S}}_{\lambda_n}^*$ [for (A5)], and with partial correlations based on Σ_{n0} [for

(A6)]. *Further, assume that the constants b_1, b_2, d_1, d_2 in* (A3)–(A6) *satisfy the following stronger restrictions:* $1/2 < b_2 \le b_1 \le 1$ *and* $\max(d_1, d_2) < b_2 - 1/2$. *Finally, assume that* $1/\|\Sigma_n^{-1}\|_2$ *is bounded below by an absolute constant* $C_2 > 0$ *for all* $(K_n q_n + 2) \times (K_n q_n + 2)$ *principal submatrices* Σ_n *of* Σ_{n0}, *where* q_n *and* K_n *are given by* (A3) *and* (A5) *respectively. Then there exists a sequence* $\lambda_n \to 0$ *such that*

$$\lim_{n\to\infty} \mathbb{P}(\hat{\mathcal{C}}_n = \mathcal{C}_{n0}) = \lim_{n\to\infty} \mathbb{P}(\check{\mathcal{C}}_n = \mathcal{C}_{n0}) = \lim_{n\to\infty} \mathbb{P}(\tilde{\mathcal{C}}_n = \mathcal{C}_{n0}) = 1,$$

where $\hat{\mathcal{C}}_n$, $\check{\mathcal{C}}_n$ *and* $\tilde{\mathcal{C}}_n$ *are the outputs of ARGES-CIG, ARGES-skeleton and GES, respectively, with the scoring criterion* $\tilde{\mathcal{S}}_{\lambda_n}$ *given by Definition* 7.2.

Our assumptions in Theorem 7.1 are similar to the assumptions of Harris and Drton (2013),[2] except that we additionally assume (A4) and (A5), where we require (A4) only for ARGES. Note that (A4) is not a strong assumption since there are high-dimensionally consistent estimators of the CIG or CPDAG-skeleton [Liu et al. (2012), Harris and Drton (2013)]. In fact, Liu et al. (2012) proposed to use rank based estimators of the correlation matrix as described above, and showed that high-dimensional consistency can be retained in the nonparanormal setting by plugging in such an estimated correlation matrix in many CIG estimation methods that are developed for multivariate Gaussian distributions [e.g., Friedman, Hastie and Tibshirani (2008), Cai, Liu and Luo (2011)].

8. Simulations. Having shown that (AR)GES has similar theoretical guarantees as the PC algorithm in high-dimensional settings, we now compare the finite sample performance and computational efficiency of (AR)GES and PC. In fact, we compare (AR)GES with the order independent version of PC [Colombo and Maathuis (2014)], but in the remaining of this section, we simply refer to it as PC. Additionally, we include another popular hybrid structure learning method, called Max-Min Hill-Climbing (MMHC) [Tsamardinos, Brown and Aliferis (2006)]. MMHC first estimates the CPDAG-skeleton by applying the Max-Min Parents and Children (MMPC) algorithm [Tsamardinos, Brown and Aliferis (2006)], and then performs a hill-climbing DAG search on the space restricted to the estimated CPDAG-skeleton (see also Remark 3.1). We use the R-package *pcalg* [Kalisch et al. (2012)] for (AR)GES and PC and use the R-package *bnlearn* [Scutari (2010)] for MMHC. In fact, we use a slightly modified version of (AR)GES that additionally includes a *turning phase* [Chickering (2002a), Hauser and Bühlmann (2012)] and an iteration over all three phases (see Section 6 of the Supplementary Material for details).

---

[2]As in Harris and Drton (2013), we can allow an ultra high-dimensional setting by replacing (A2) by $p_n = \mathcal{O}(\exp(n^a))$ (for some $0 \le a < 1$) and allow an $\mathcal{O}(n^f)$ (for some $0 \le f < 1/4$) growth rate of $\|\Sigma_n^{-1}\|_2$ for each $(K_n q_n + 2) \times (K_n q_n + 2)$ submatrix Σ_n of Σ_{n0} if we replace the condition $\max(d_1, d_2) < b_2 - 1/2$ by $a + 2(1 - b_2) + 2\max(d_1, d_2) + 4f < 1$.

TABLE 2
Simulation settings

n	100	200	300	400
p_n	300	600	1200	2400
e_n	300	840	2100	4800

8.1. *Simulation settings.* For each of the four settings given in Table 2, we use the R-package *pcalg* [Kalisch et al. (2012)] to simulate $r = 100$ random weighted DAGs $\{\mathcal{G}_n^{(1)}, \ldots, \mathcal{G}_n^{(r)}\}$ with p_n vertices and expected number of edges e_n, where each pair of nodes in a randomly generated DAG has the probability $e_n / \binom{p_n}{2}$ of being adjacent. The edge weights are drawn independently from a uniform distribution on $(-1, -0.1) \cup (0.1, 1)$.

Let $B_n^{(t)}$ denote the weight matrix of the weighted DAG $\mathcal{G}_n^{(t)}$, that is, $(B_n^{(t)})_{ij} \neq 0$ if and only if the edge $X_j \to X_i$ is present in $\mathcal{G}_n^{(t)}$ and it then equals the corresponding edge weight. For $t = 1, \ldots, r$, the weight matrix $B_n^{(t)}$ and a random vector $\boldsymbol{\varepsilon}_n^{(t)} = (\varepsilon_{n1}^{(t)}, \ldots, \varepsilon_{np_n}^{(t)})^T$ define a distribution on $\mathbf{X}_n^{(t)} = (X_{1n}^{(t)}, \ldots, X_{np_n}^{(t)})^T$ via the linear structural equation model $\mathbf{X}_n^{(t)} = B_n^{(t)} \mathbf{X}_n^{(t)} + \boldsymbol{\varepsilon}_n^{(t)}$. We choose $\varepsilon_{n1}^{(t)}, \ldots, \varepsilon_{np_n}^{(t)}$ to be zero mean Gaussian random variables with variances independently drawn from a Uniform[1, 2] distribution. We aim to estimate CPDAG($\mathcal{G}_n^{(t)}$) from n i.i.d. samples from the multivariate Gaussian distribution of $\mathbf{X}_n^{(t)}$.

8.2. *Estimation of the CIG and the CPDAG-skeleton for ARGES.* We estimate the CIG for ARGES-CIG using neighborhood selection with the LASSO[3] of Meinshausen and Bühlmann (2006), where we use the implementation in the R-package *huge* [Zhao et al. (2012)]. Neighborhood selection involves a tuning parameter γ_n that corresponds to the LASSO penalization, where larger values of γ_n yield sparser estimated graphs. We choose $\gamma_n = 0.16, 0.14, 0.12$ and 0.10 for $p_n = 300, 600, 1200$ and 2400 respectively. In Section 7 of the Supplementary Material, we empirically investigate the influence of γ_n on the performance of ARGES-CIG. We find that the performance of ARGES-CIG is not very sensitive to the choice of γ_n in this simulation setting, as long as γ_n is reasonably small (i.e., the estimated CIG is reasonably dense). In particular, estimation quality can be slightly improved by choosing a smaller γ_n, but with a loss of computational efficiency.

We estimate the CPDAG-skeleton for ARGES-skeleton using the MMPC algorithm of Tsamardinos, Brown and Aliferis (2006), where we use the implementation in the R-package *bnlearn* [Scutari (2010)]. MMPC involves a tuning

---

[3]We will later use adaptive LASSO [Zou (2006)] for this step (see Section 8.4).

parameter κ_n that corresponds to the significance level of the conditional independence tests, where smaller values of κ_n yield sparser estimated graphs. We choose $\kappa_n = 0.2, 0.15, 0.10$ and 0.05 for $p_n = 300, 600, 1200$ and 2400 respectively. We note that the MMPC algorithm is computationally expensive for large values of κ_n.

8.3. *Results.* As the scoring criterion for (AR)GES and MMHC, we use the ℓ_0-penalized likelihood score (see Definition 5.1) with a number of choices for the penalty parameter λ_n. Similarly, we apply PC with a number of choices for its tuning parameter α_n (the significance level for conditional independence tests). Finally, we compare their estimation quality with averaged receiver operating characteristic (ROC) curves, where we average true positive rates and false positive rates for each value of the tuning parameters over r iterations (cf. threshold averaging of ROC curves [Fawcett (2006)]). We do not apply GES and MMHC for the case $p_n = 2400$, since they are too slow to handle such large graphs. We additionally apply ARGES-CIG with the true CIG, and we call it ARGES-CIG*.

In Figure 3, we see that the averaged ROC curves get better as n and p_n increase together. This supports the high-dimensional consistency theory of (AR)GES and PC. Based on Figure 3, the performances of the algorithms can be summarized as follows:

$$\text{PC} < \text{MMHC} < \text{(A)RGES-skeleton} < \text{(A)RGES-CIG} \approx \text{GES} < \text{ARGES-CIG}^*,$$
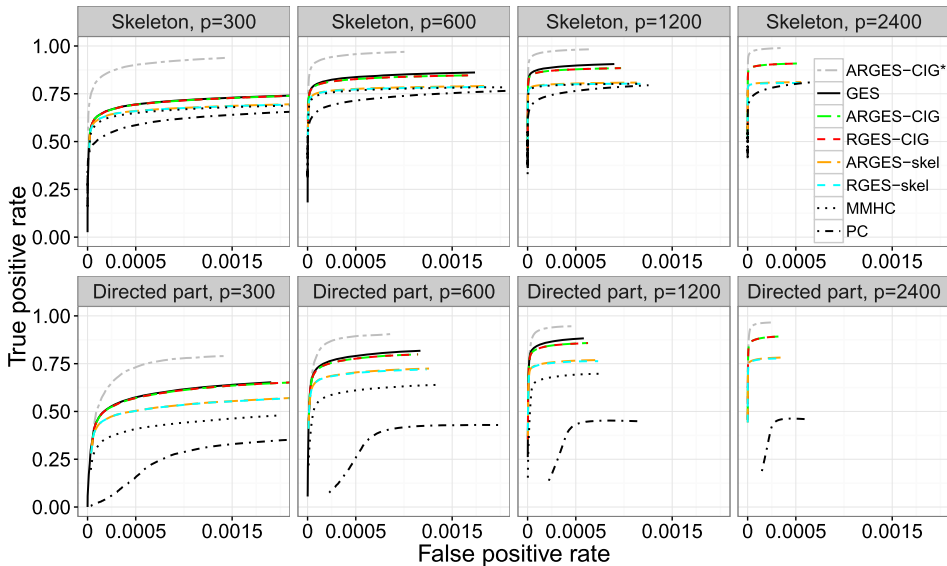

FIG. 3. *Averaged ROC curves for estimating the skeletons (upper panel) and the directed parts (lower panel) of the underlying CPDAGs with ARGES-CIG*, ARGES-CIG, ARGES-skeleton, GES, MMHC and PC, for simulation settings given in Table 2.*

where A < B represents that B outperformed A and A ≈ B represents that A and B performed similarly. Below, we list our main findings and possible explanations in detail.

1. ARGES-CIG* is the best performing method, but it is infeasible in practice as it requires knowledge of the true CIG.

2. GES is the next best performing method, closely followed by (A)RGES-CIG. However, the fact that ARGES-CIG* outperforms GES indicates the possibility that ARGES-CIG can outperform GES when combined with a better CIG estimation technique. We explore this in Section 8.4.

3. The fact that the performances of ARGES and RGES are almost identical, indicates that the adaptive part of ARGES does not have a significant influence on the performance.

4. (A)RGES-CIG outperformed (A)RGES-skeleton because (i) the true positive rate of (A)RGES is approximately bounded by the estimated CIG or CPDAG-skeleton, and (ii) the true positive rates of the estimated CIGs are larger than that of the estimated CPDAG-skeletons (see Table 4 of the Supplementary Material).

5. Although the performance of MMHC is similar to ARGES-skeleton for estimating the CPDAG-skeleton, it is significantly worse for estimating the directed part of the CPDAG. We suspect that this is due to some arbitrary choices of edge orientations, made in hill-climbing DAG search to resolve its nonuniqueness (see Section 4.3 of the Supplementary Material).

6. The constraint-based PC is the worst performing method in terms of estimation quality in these simulations.

Figure 4 shows the average runtimes of the algorithms.[4] We see that GES does not scale well to large graphs and that the runtimes of (AR)GES and PC are somewhat similar. The runtimes of (A)RGES do not include the CIG or the CPDAG-skeleton estimation (see Table 4 of the Supplementary Material). The average runtimes of MMHC are much worse than the runtimes of (A)RGES-skeleton, although they are based on the same estimated CPDAG-skeleton.

8.4. *Simulation with decreasing sparsity level.* We have seen that in very sparse high-dimensional settings where GES performed very well, ARGES-CIG based on neighborhood selection could not outperform GES in terms of estimation. One would expect, however, that using a good restricted search space can also be beneficial for the estimation performance. We investigate this in the following simulations (see Table 3), where we increased the adversity of the the problem by decreasing the sparsity level, and we paid more attention to the estimation of the CIG

---

[4]We use implementations of the algorithms from several R-packages. Hence, Figure 4 does not represent the computational efficiency of the algorithms, but rather represents a comparison of their currently available implementations in R.
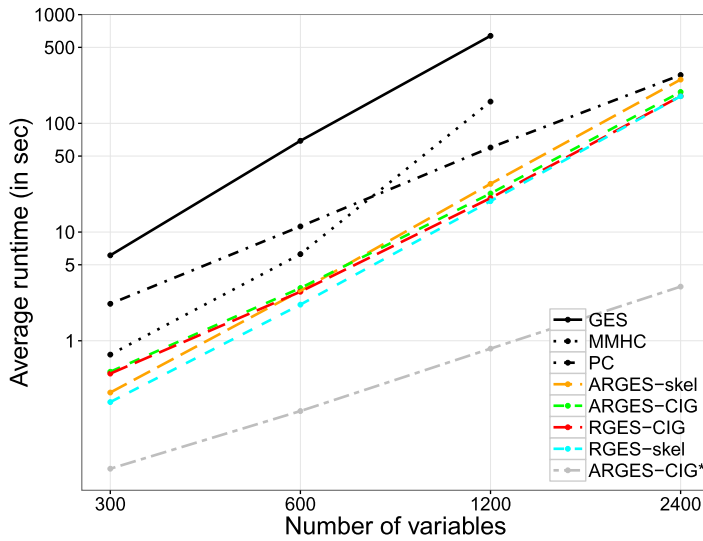
FIG. 4.   *Average runtimes* (*in seconds*) *for ARGES-CIG**, *ARGES-CIG*, *ARGES-skeleton*, *GES*, *MMHC and PC*, *where the tuning parameters* α_n *and* λ_n *are chosen to get roughly the right sparsity. The runtimes of* (A)*RGES do not include the CIG or the CPDAG-skeleton estimation part* (*see Table* 4 *of the Supplementary Material*).

by including adaptive LASSO [Zou (2006)]. Moreover, to make the method applicable in practice, we no longer choose some pre-specified γ_n's (cf. Section 8.2), but choose it via cross-validation.

We compare the following methods: ARGES-CIG based on LASSO, ARGES-CIG based on adaptive LASSO, GES and PC. We choose the tuning parameter for each LASSO[5] by 10-fold cross-validation, optimizing the prediction error (e.g., [Hastie, Tibshirani and Friedman (2009)]). For each adaptive LASSO,[6] we use the weights $w_i = 1/\hat{\beta}_i$ where $\hat{\beta}_i$ is the initial estimate of the ith linear regression coefficient given by the corresponding LASSO regression with a 10-fold cross

TABLE 3
Simulation settings with fixed number of variables and decreasing sparsity

n	50	100	150	200
p_n	100	100	100	100
e_n	100	200	300	400

---

[5]We use the implementation in the R-package *glmnet* [Friedman, Hastie and Tibshirani (2010)].

[6]We use the aforementioned implementation of LASSO after removing the variables with infinite weights and rescaling the other variables with the corresponding weights.
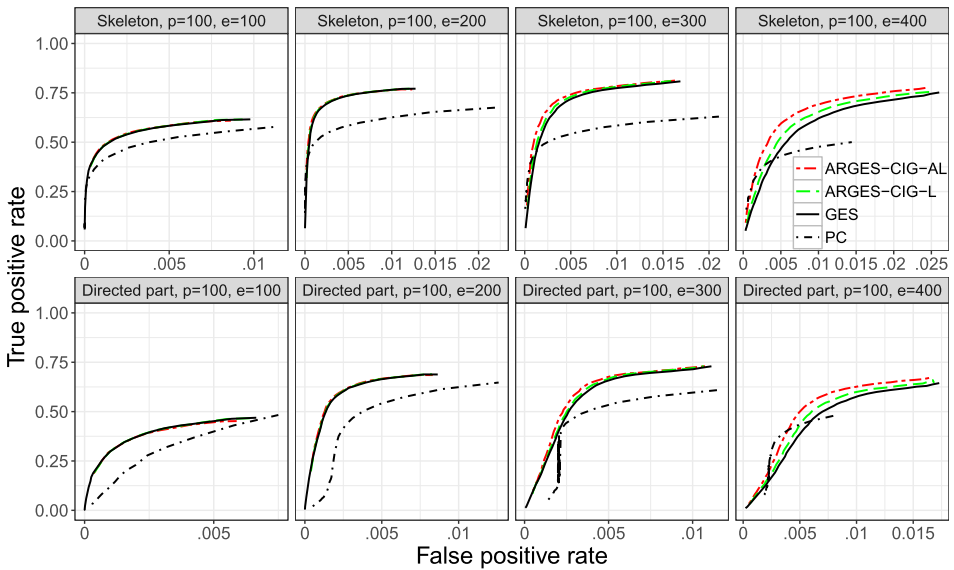
FIG. 5. *Averaged ROC curves for estimating the skeletons* (*upper panel*) *and the directed parts* (*lower panel*) *of the underlying CPDAGs with ARGES-CIG based on adaptive LASSO* (*ARGES–CIG-AL*), *ARGES-CIG based on LASSO* (*ARGES-CIG-L*), *GES and PC, for simulation settings given in Table* 3.

validation. After fixing the weights, we choose the tuning parameter in the second stage of the adaptive LASSO based on an additional 10-fold cross validation. ROC curves for estimating the skeleton of the CPDAG and the directed part of the CPDAG are obtained by varying λ_n for all GES based methods, and by varying α_n for the PC algorithm.

Figure 5 shows that both versions of ARGES-CIG and GES perform equally well for the first three settings, while ARGES-CIG outperforms GES in the most adverse setting with $e_n = 400$. Interestingly, although the estimated CIG based on the adaptive LASSO is always a subgraph of the estimated CIG based on the LASSO, the use of the adaptive LASSO for the CIG estimation enhances the performance gain as the sparsity decreases (cf. Section 7 of the Supplementary Material).

9. Discussion. We propose two new hybrid methods, ARGES-CIG and ARGES-skeleton, consisting of restricted versions of GES, where the restriction on the search space is not simply given by an estimated CIG or an estimated CPDAG-skeleton, but also depends adaptively on the current state of the algorithm. We include the adaptive part in our algorithms to ensure that the output is a consistent estimate of the underlying CPDAG. The fact that the adaptive modification is rather small, provides an explanation for the empirical success of inconsistent hybrid methods that restrict the search space to an estimated CIG or an estimated

CPDAG-skeleton [e.g., Tsamardinos, Brown and Aliferis (2006), Schulte et al. (2010)].

We prove consistency of GES and ARGES in sparse high-dimensional settings with multivariate Gaussian distributions, linear SEMs with sub-Gaussian errors, or nonparanormal distributions (see Sections 5, 6 and 7). To the best of our knowledge, these are the first results on high-dimensional consistency of score-based and hybrid methods. Our simulation results indicate that GES and ARGES generally outperform the PC algorithm (see Section 8), which has so far been the most popular structure learning algorithm in high-dimensional settings. Moreover, an advantage of (AR)GES compared to PC is that its output is always a CPDAG, which is important for some applications such as (joint-)IDA [Maathuis, Kalisch and Bühlmann (2009), Nandy, Maathuis and Richardson (2017)] and the generalized adjustment criterion [Perkovic et al. (2015a), Perkovic et al. (2015b)]. Note that the sample version of the PC algorithm provides no such guarantee and often produces a partially directed graph that is not a CPDAG.

A disadvantage of ARGES compared to GES is that it requires an additional tuning parameter to estimate the CIG or the CPDAG-skeleton. Our simulation results suggest that ARGES-CIG can achieve a very similar performance as GES in a much shorter time as long as we choose a sufficiently dense estimated CIG (by adjusting the corresponding tuning parameter) while respecting the computation limit. In some settings, however, the restricted search space is not only beneficial from a computational point of view, but also in terms of estimation performance (see Section 8.4).

Tuning the penalty parameter λ of a scoring criterion of (AR)GES is a well-known practical problem. We recommend to apply the stability selection approach of Meinshausen and Bühlmann (2010) or to use the extended BIC criterion [Chen and Chen (2008), Foygel and Drton (2010)], which has been shown to work better in sparse high-dimensional settings than the BIC criterion.

There have been some recent theoretical and practical attempts to speed up GES and we note that they can be applied to ARGES as well. Chickering and Meek (2015) proposed a modification of the backward phase of GES that has polynomial complexity. Further, the authors showed that the final output of this modified version of GES, called selective GES (SGES), is consistent in the classical setting if the output of the forward phase SGES is an independence map of the CPDAG in the limit of large samples. The forward phase of ARGES can be combined with the backward phase of SGES and consistency of such an algorithm follows from the fact that the output of the forward phase of ARGES is an independence map of the CPDAG in the limit of large samples (see the proof of Theorem 4.2). Ramsey (2015) showed that with an efficient implementation and parallel computing, GES can be scaled up to thousands of variables. Similar efficient implementations and parallel computations are possible for hybrid algorithms like ARGES, and this would push the computation limit even further.

We establish a novel connection between score-based and constraint-based methods (see Section 5.1). In particular, Lemma 5.1 shows that the score-based GES algorithm and the constraint-based PC algorithm are in fact closely related approaches in the multivariate Gaussian setting. The fundamental principle of the PC algorithm is to start with a complete graph and to delete edges sequentially by testing conditional independencies. In the multivariate Gaussian setting, conditional independence tests are equivalent to tests for zero partial correlations. Lemma 5.1 shows that GES checks if sample partial correlations are large enough in order to add edges in the forward phase or small enough to delete edges in the backward phase. This insight opens the door to study new score-based and hybrid algorithms that are applicable to broader classes of models. For example, in Section 7, we defined a new scoring criterion based on rank correlations (see Definition 7.2) for nonparanormal distributions. Analogously, one can define scoring criteria based on more general conditional independence tests, leading to score-based (or hybrid) competitors of the PC algorithms based on such conditional independence tests [e.g., Zhang et al. (2011), Doran et al. (2014)].

Although both GES and PC use partial correlation-based conditional independence tests in the multivariate Gaussian setting, we found that GES outperforms PC in terms of estimation quality. A possible explanation for the better performance of GES is that GES considers the skeleton and the orientations of the edges simultaneously, whereas the PC algorithm first determines the skeleton and then orients the edges by determining v-structures and subsequently applying a set of rules [Meek (1995)].

Recall that our high-dimensional consistency proofs require an assumption on the growth of oracle versions of (AR)GES [see assumption (A5) of Section 5.2]. Such an assumption is not required for high-dimensional consistency of the PC algorithm [Kalisch and Bühlmann (2007)]. We discussed this assumption in Section 5.3 and provided some strong structural conditions under which this assumption holds. In order to derive these sufficient conditions, we show a connection between GES and the Chow-Liu algorithm, which may be of independent interest.

We emphasize that our consistency result in the classical setting (where the number of variables remains fixed and the sample size goes to infinity) does not require any distributional assumption (it even holds for discrete distributions), except that the joint distribution of the variables is DAG-perfect, that is, there exists a DAG $\mathcal{G}$ such that all conditional independence relationships encoded by $\mathcal{G}$ hold in the joint distribution and vice versa. Chickering and Meek (2002) showed that without this assumption (but with a weaker assumption on the joint distribution), GES is consistent for learning a *minimal independence map* of the joint distribution. A DAG is a minimal independence map of a distribution if it is an independence map of the distribution and no proper subgraph is an independence map. An interesting direction for future work is to investigate (in)consistency of ARGES for learning a minimal independence map under similar assumptions as in Chickering and Meek (2002).

SUPPLEMENTARY MATERIAL

Supplement to "High-dimensional consistency in score-based and hybrid structure learning" (DOI: 10.1214/17-AOS1654SUPP; .pdf). All proofs, additional simulation results, additional details for Example 1 can be found in the supplementary material.

REFERENCES

ALONSO-BARBA, J. I., DELAOSSA, L., GÁMEZ, J. A. and PUERTA, J. M. (2013). Scaling up the greedy equivalence search algorithm by constraining the search space of equivalence classes. *Internat. J. Approx. Reason.* **54** 429–451. MR3041110

ANANDKUMAR, A., TAN, V. Y. F., HUANG, F. and WILLSKY, A. S. (2012). High-dimensional Gaussian graphical model selection: Walk summability and local separation criterion. *J. Mach. Learn. Res.* **13** 2293–2337. MR2973603

ANDERSSON, S. A., MADIGAN, D. and PERLMAN, M. D. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.* **25** 505–541. MR1439312

BANERJEE, O., EL GHAOUI, L. and D'ASPREMONT, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9** 485–516. MR2417243

CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189

CHICKERING, D. M. (2002). Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.* **2** 445–498. MR1929415

CHICKERING, D. M. (2003). Optimal structure identification with greedy search. *J. Mach. Learn. Res.* **3** 507–554. MR1991085

CHICKERING, D. M. and MEEK, C. (2002). Finding optimal Bayesian networks. In *UAI* 2002.

CHICKERING, D. M. and MEEK, C. (2015). Selective greedy equivalence search: Finding optimal Bayesian networks using a polynomial number of score evaluations. In *UAI* 2015.

CHOW, C. and LIU, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inform. Theory* **14** 462–467.

COLOMBO, D. and MAATHUIS, M. H. (2014). Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.* **15** 3741–3782. MR3291411

DE CAMPOS, L. M. (1998). Independency relationships and learning algorithms for singly connected networks. *J. Exp. Theor. Artif. Intell.* **10** 511–549.

DORAN, G., MUANDET, K., ZHANG, K. and SCHÖLKOPF, B. (2014). A permutation-based kernel conditional independence test. In *UAI* 2014.

FAWCETT, T. (2006). An introduction to ROC analysis. *Pattern Recogn. Lett.* **27** 861–874.

FOYGEL, R. and DRTON, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In *NIPS* 2010.

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

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

GAO, B. and CUI, Y. (2015). Learning directed acyclic graphical structures with genetical genomics data. *Bioinformatics* **31** 3953–3960.

HA, M. J., SUN, W. and XIE, J. (2016). PenPC: A two-step approach to estimate the skeletons of high-dimensional directed acyclic graphs. *Biometrics* **72** 146–155. MR3500583

HARRIS, N. and DRTON, M. (2013). PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.* **14** 3365–3383. MR3144465

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York. MR2722294

HAUSER, A. and BÜHLMANN, P. (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *J. Mach. Learn. Res.* **13** 2409–2464. MR2973606

HUETE, J. F. and DE CAMPOS, L. M. (1993). Learning causal polytrees. In *ECSQARU* 1993.

KALISCH, M. and BÜHLMANN, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* **8** 613–636.

KALISCH, M., MÄCHLER, M., COLOMBO, D., MAATHUIS, M. and BÜHLMANN, P. (2012). Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.* **47** 1–26.

KOLLER, D. and FRIEDMAN, N. (2009). *Probabilistic Graphical Models*: *Principles and Techniques*. MIT Press, Cambridge, MA. MR2778120

LE, T. D., LIU, L., TSYKIN, A., GOODALL, G. J., LIU, B., SUN, B.-Y. and LI, J. (2013). Inferring microRNA-mRNA causal regulatory relationships from expression data. *Bioinformatics* **29** 765–771.

LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. MR3059084

MAATHUIS, M. H., KALISCH, M. and BÜHLMANN, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Statist.* **37** 3133–3164. MR2549555

MAATHUIS, M. H., COLOMBO, D., KALISCH, M. and BÜHLMANN, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nat. Methods* **7** 247–248.

MEEK, C. (1995). Causal inference and causal explanation with background knowledge. In *UAI* 1995.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 417–473. MR2758523

NANDY, P., HAUSER, A. and MAATHUIS, M. H. (2018). Supplement to "High-dimensional consistency in score-based and hybrid structure learning." DOI:10.1214/17-AOS1654SUPP.

NANDY, P., MAATHUIS, M. H. and RICHARDSON, T. S. (2017). Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *Ann. Statist.* **45** 647–674. MR3650396

OUERD, M., OOMMEN, B. J. and MATWIN, S. (2004). A formal approach to using data distributions for building causal polytree structures. *Inform. Sci.* **168** 111–132. MR2117288

PEARL, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166

PERKOVIC, E., TEXTOR, J., KALISCH, M. and MAATHUIS, M. (2015a). A complete adjustment criterion. In *UAI* 2015.

PERKOVIC, E., TEXTOR, J., KALISCH, M. and MAATHUIS, M. (2015b). Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. Preprint. Available at arXiv:1606.06903.

RAMSEY, J. D. (2015). Scaling up greedy equivalence search for continuous variables. Preprint. Available at arXiv:1507.07749.

RAVIKUMAR, P. K., RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2008). Model selection in Gaussian graphical models: High-dimensional consistency of ℓ_1-regularized MLE. In *NIPS* 2008.

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766

REBANE, G. and PEARL, J. (1987). The recovery of causal poly-trees from statistical data. In *UAI* 1987.

SCHMIDBERGER, M., LENNERT, S. and MANSMANN, U. (2011). Conceptual aspects of large meta-analyses with publicly available microarray data: A case study in oncology. *Bioinform. Biol. Insights* **5** 13–39.

SCHMIDT, M., NICULESCU-MIZIL, A. and MURPHY, K. (2007). Learning graphical model structure using L1-regularization paths. In *AAAI* 2007.

SCHULTE, O., FRIGO, G., GREINER, R. and KHOSRAVI, H. (2010). The IMAP hybrid method for learning Gaussian Bayes nets. In *Canadian AI* 2010.

SCUTARI, M. (2010). Learning Bayesian networks with the bnlearn R package. *J. Stat. Softw.* **35** 1–22.

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

SPIRTES, P., RICHARDSON, T., MEEK, C., SCHEINES, R. and GLYMOUR, C. (1998). Using path diagrams as a structural equation modeling tool. *Sociol. Methods Res.* **27** 182–225.

STEKHOVEN, D. J., MORAES, I., SVEINBJÖRNSSON, G., HENNING, L., MAATHUIS, M. H. and BÜHLMANN, P. (2012). Causal stability ranking. *Bioinformatics* **28** 2819–2823.

TSAMARDINOS, I., BROWN, L. E. and ALIFERIS, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.* **65** 31–78.

UHLER, C., RASKUTTI, G., BÜHLMANN, P. and YU, B. (2013). Geometry of the faithfulness assumption in causal inference. *Ann. Statist.* **41** 436–463. MR3099109

VAN DE GEER, S. and BÜHLMANN, P. (2013). ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Ann. Statist.* **41** 536–567. MR3099113

VERDUGO, R. A., ZELLER, T., ROTIVAL, M., WILD, P. S., MÜNZEL, T., LACKNER, K. J., WEIDMANN, H., NINIO, E., TRÉGOUËT, D.-A., CAMBIEN, F., BLANKENBERG, S. and TIRET, L. (2013). Graphical modeling of gene expression in monocytes suggests molecular mechanisms explaining increased atherosclerosis in smokers. *PLoS ONE* **8** e50888.

VERMA, T. and PEARL, J. (1990). Equivalence and synthesis of causal models. In *UAI* 1990.

ZHANG, K., PETERS, J., JANZING, D. and SCHÖLKOPF, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *UAI* 2011.

ZHAO, T., LIU, H., ROEDER, K., LAFFERTY, J. and WASSERMAN, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* **13** 1059–1062. MR2930633

ZOU, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

P. NANDY
M. H. MAATHUIS
SEMINAR FOR STATISTICS
ETH ZÜRICH
RÄMISTRASSE 101
8092 ZÜRICH
SWITZERLAND
E-MAIL: preetamnandy@gmail.com
       maathuis@stat.math.ethz.ch

A. HAUSER
DEPARTMENT OF ENGINEERING AND
  INFORMATION TECHNOLOGY
BERN UNIVERSITY OF APPLIED SCIENCES
JLCOWEG 1
3400 BURGDORF
SWITZERLAND
(NOW AT GOOGLE)
E-MAIL: alhauser@google.com