Supplementary material to "Structure Learning of Linear Gaussian Structural Equation Models with Weak Edges"

1 PRELIMINARIES

Two vertices X_i and X_j are *adjacent* if there is an edge between them. A *path* between X_i and X_j is a sequence $(X_i, \dots, X_j)$ of distinct vertices in which all pairs of successive vertices are adjacent. A *directed path* is a path between X_i and X_j where all edges are directed towards X_j, i.e., $X_i \to \cdots \to X_j$. A directed path from X_i to X_j together with the edge $X_j \to X_i$ forms a *directed cycle*. If $X_i \to X_j \leftarrow X_k$ is part of a path, then X_j is a *collider* on this path.

A vertex X_j is a *child* of the vertex X_i if $X_i \to X_j$. If there is a directed path from X_i to X_j, X_i is a *descendant* of X_j, otherwise it is a non-descendant. We use the convention that X_i is also a descendant of itself.

A DAG encodes conditional independence constraints through the concept of d-separation (Pearl, 2009). For three pairwise disjoint subsets of vertices A, B, and S of X, A is d-separated from B by S, $A \perp B|S$, if every path between a vertex in A and a vertex in B is *blocked* by S. A path between two vertices X_i and X_j is said to be *blocked* by a set S if a non-collider vertex on the path is present in S or if there is a collider vertex on the path for which none of its descendants is in S. If a path is not blocked it is *open*.

The set of d-separation constraints encoded by a DAG G is denoted by $\mathcal{I}(G)$. All DAGs in a Markov equivalence class encode the same set of d-separation constraints. Hence, for a CPDAG C, we let $\mathcal{I}(C) = \mathcal{I}(G)$, where G is any DAG in C. A DAG G_1 is an independence map (I-map) of a DAG G_2 if $\mathcal{I}(G_1) \subseteq \mathcal{I}(G_2)$, with an analogous definition for CPDAGs. A DAG G_1 is a perfect map of a DAG G_2 if $\mathcal{I}(G_1) = \mathcal{I}(G_2)$, again with an analogous definition for CPDAGs.

For the proof of Theorem 3.2 of the main paper we make use of two lemmas of Nandy et al. (2015).

Lemma 1.1. *(cf. Lemma 9.5 of the supplementary material of Nandy et al. (2015)) Let $G = (X, E)$ be a DAG such that $X_i \to X_j \in E$. Let $G' = (X, E \setminus \{X_i \to X_j\})$. If G is an I-map of a DAG G_1 but G' is not, then $X_i \not\perp_{G_1} X_j \mid \mathrm{Pa}_{G'}(X_i)$.*

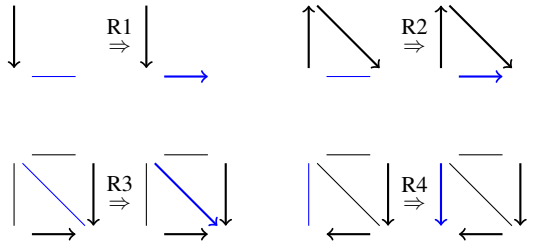


Figure 1: The four orientation rules from Meek (1995). If a PDAG contains one of the graphs on the left-hand-side of the four rules, then orient the blue edge as shown on the right-hand-side.

Lemma 1.2. *(cf. Lemma 5.1 of Nandy et al. (2015)) Let $G = (X, E)$ be a DAG such that X_i is neither a descendant nor a parent of X_j. Let $G' = (X, E \cup \{X_i \to X_j\})$. If the distribution of X is multivariate Gaussian, then the ℓ_0-penalized log-likelihood score difference between G' and G is*

$$S_\lambda(G', X^{(n)}) - S_\lambda(G, X^{(n)}) = \frac{1}{2}\log(1 - \hat{\rho}^2_{X_i, X_j \mid \mathrm{Pa}_G(X_j)}) + \lambda.$$

The last step of Algorithm 1 of the main paper consists of MeekOrient. This step applies iteratively and sequentially the four rules depicted in Figure 1. These orientation rules can lead to some additional orientations, and the resulting output is a maximally oriented PDAG (Meek, 1995). For an example of its utility see Example 3.1 and Figure 3 in the main paper.

2 PROOFS

2.1 PROOF OF THEOREM 3.2 OF THE MAIN PAPER

We first establish the following Lemma.

Lemma 2.1. *Consider two CPDAGs C_1 and C_2 where C_1 is an I-map of C_2. If C_1 and C_2 have the same skeleton, then C_1 is a perfect map of C_2.*

Proof. Let G_1 and G_2 be arbitrary DAGs in the Markov equivalence classes described by C_1 and C_2, respectively. Then C_1 is a perfect map of C_2 if and only if G_1 and G_2 have the same skeleton and the same v-structures (Verma and Pearl, 1990). Since G_1 and G_2 have the same skeleton by assumption, we only need to show that they have identical v-structures.

Suppose first that there is a v-structure $X_i \to X_j \leftarrow X_k$ in G_1 that is not present in G_2. Since G_1 and G_2 have the same skeleton, this implies that X_j is a non-collider on the path (X_i, X_j, X_k) in G_2.

We assume without loss of generality that X_i is a non-descendant of X_k in G_1. Then, $X_i \perp_{G_1} X_k | \operatorname{Pa}_{G_1}(X_k)$, where $X_j \notin \operatorname{Pa}_{G_1}(X_k)$. On the other hand, we have $X_i \not\perp_{G_2} X_k | \operatorname{Pa}_{G_1}(X_k)$, since the path (X_i, X_j, X_k) is open in G_2, since $X_j \notin \operatorname{Pa}_{G_1}(X_k)$. This contradicts that C_1 is an I-map of C_2.

Next, suppose that there is a v-structure $X_i \to X_j \leftarrow X_k$ in G_2 that is not present in G_1. Since G_1 and G_2 have the same skeleton, this implies that X_j is a non-collider on the path (X_i, X_j, X_k) in G_1.

We again assume without loss of generality that X_i is a non-descendant of X_k in G_1. Then the path (X_i, X_j, X_k) has one of the following forms: $X_i \to X_j \to X_k$ or $X_i \leftarrow X_j \to X_k$. In either case, $X_j \in Pa_{G_1}(X_k)$. Hence, $X_i \perp_{G_1} X_k | \operatorname{Pa}_{G_1}(X_k)$, where $X_j \in \operatorname{Pa}_{G_1}(X_k)$. But X_i and X_k are d-connected in G_2 by any set containing X_j. This again contradicts that C_1 is an I-map of C_2. □

Proof of Theorem 3.2 of the main paper. We need to prove that the CPDAGs in Step S.1 of the main paper and the CPDAGs in Step S.3 of the main paper coincide, i.e., $\mathcal{C} = \tilde{\mathcal{C}}$. We prove this result for one of the CPDAGs. Take for instance $\tilde{C}_\ell$, $1 \leqslant \ell \leqslant k$, the CPDAG of $G_\ell = (V, E_\ell)$. Note that G_ℓ is not a perfect map of the distribution of X, and therefore we cannot directly use the proof of Chickering (2002). We can still use the main idea though, in combination with Lemma 1.2.

Consider running GES with penalty parameter $\lambda = -1/2 \log(1 - \delta_\ell^2)$ and denote by C^f and C^b the output of the forward and backward phase, respectively.

Claim 1: C^f is an I-map of $\tilde{C}_\ell$ i.e., all d-separation constraints true in C^f are also true in $\tilde{C}_\ell$.

Proof of Claim 1:
Assume this is not the case, then there are two vertices $X_i, X_j \in X$ and a DAG $G^f \in C^f$ such that $X_i \perp_{G^f} X_j | \{\operatorname{Pa}_{G^f}(X_j) \setminus X_i\}$ but $X_i \not\perp_{C_\ell} X_j | \{\operatorname{Pa}_{G^f}(X_j) \setminus X_i\}$. Because of the δ_ℓ-strong faithful condition, $|\rho_{X_i, X_j | \operatorname{Pa}_{G^f}(X_j)}| > \delta_\ell$. Thus, adding this edge would improve the score. This is a contradiction to the GES algorithm stopping here.

Claim 2: C^b is an I-map of $\tilde{C}_\ell$ i.e., all d-separation constraints true in C^b are also true in $\tilde{C}_\ell$.

Proof of Claim 2: By Claim 1 the backward phase starts with an I-map of $\tilde{C}_\ell$. Suppose it ends with a CPDAG that is not an I-map of $\tilde{C}_\ell$. Then, at some point there is an edge deletion which turns a DAG G that is an I-map of G_ℓ into a DAG G' that is no longer an I-map of G_ℓ. Suppose the deleted edge is (X_i, X_j). By Lemma 1.1, we have $X_i \not\perp_{G_\ell} X_j | \{\operatorname{Pa}_{G'}(X_j)\}$. Hence, again because of the $\delta_\ell-$strong faithfulness condition, $|\rho_{X_i, X_j | \operatorname{Pa}_{G'}(X_j)}| > \delta$. Thus, deleting this edge would worsen the score. This is a contradiction to the GES algorithm deleting this edge.

Claim 3: $C^b = \tilde{C}_\ell$, i.e., C^b is a perfect map of $\tilde{C}_\ell$.

This claim follows from Lemma 2.1 since we know from the previous claim that C^b is an I-map of $\tilde{C}_\ell$ and by construction the skeletons of C^b and $\tilde{C}_\ell$ are the same.

It follows from $\mathcal{C} = \tilde{\mathcal{C}}$ that $\operatorname{AggregateCPDAGs}(\mathcal{C}) = \operatorname{AggregateCPDAGs}(\tilde{\mathcal{C}})$. □

2.2 PROOF OF THEOREM 3.3 OF THE MAIN PAPER

Recall that AGES combines a collection of CPDAGs obtained in the solution path of GES, where the largest CPDAG corresponds to the BIC penalty with $\lambda = \log(n)/(2n)$. In the consistency proof of GES with the BIC penalty, Chickering (2002) used the fact that the penalized likelihood scoring criterion with the BIC penalty is locally consistent as $\log(n)/(2n) \to 0$. We note that the other penalty parameters involved in the computation of the solution path of GES do not converge to zero. This prevents us to obtain a proof of Theorem 3.3 of the main paper by applying the consistency result of Chickering (2002). A further complication is that the choices of the penalty parameters in the solution path of GES depend on the data.

In order to prove Theorem 3.3 of the main paper, we rely on the soundness of the oracle version of AGES (Theo-

rem 3.2 of the main paper). In fact, we prove consistency of AGES by showing that the solution path of GES coincides with its oracle solution path as the sample size tends to infinity. Since the number of variables is fixed and the solution path of GES depends only on the partial correlations (see Lemma 1.2 and Section 3.5 of the main paper), the consistency of AGES will follow from the consistency of the sample partial correlations.

Proof of Theorem 3.3 of the main paper. Given a scoring criterion, each step of GES depends on the scores of all DAGs on p variables through their ranking only, where each step in the forward (backward) phase corresponds to improving the current ranking as much as possible by adding (deleting) a single edge. Let $\hat{\boldsymbol{\rho}}$ denote a vector consisting of the absolute values of all sample partial correlations $\hat{\rho}_{X_i,X_j|S}$, $1 \leq i \leq j \leq p$ and $S \subseteq X \setminus \{X_i, X_j\}$, in some order. It follows from Lemma 1.2 that the solution path of GES (for $\lambda \geq \log(n)/(2n)$) solely depends on the ranking of the elements in $\hat{\boldsymbol{\gamma}}$, where $\hat{\boldsymbol{\gamma}}$ contains the elements of $\hat{\boldsymbol{\rho}}$ appended with $(1 - n^{-1/n})^{1/2}$, where the last element results from solving $-\log(1-\rho^2)/2 = \log(n)/(2n)$ for ρ.

Similarly, an oracle solution path of GES solely depends on a ranking of the elements in $\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ contains the elements of $\boldsymbol{\rho}$ appended with the value 0, and $\boldsymbol{\rho}$ denotes a vector consisting of the absolute values of all partial correlations in the same order as in $\hat{\boldsymbol{\rho}}$. Note that there can be more than one oracle solution paths of GES depending on a rule for breaking ties. We will write $\text{rank}\,(\hat{\boldsymbol{\gamma}}) = \text{rank}(\boldsymbol{\gamma})$ if $\text{rank}\,(\hat{\boldsymbol{\gamma}})$ equals a ranking of $\boldsymbol{\gamma}$ with some rule for breaking ties.

Finally, we define

$$\epsilon = \min\left\{\left||\rho_{X_{i_1},X_{j_1}|S_1}| - |\rho_{X_{i_2},X_{j_2}|S_2}|\right| : |\rho_{X_{i_1},X_{j_1}|S_1}| \neq |\rho_{X_{i_2},X_{j_2}|S_2}|\right\},$$

where the minimum is taken over all $1 \leq i_1 < j_1 \leq p$, $S_1 \subseteq X \setminus \{X_{i_1}, X_{j_1}\}$, $1 \leq i_2 < j_2 \leq p$ and $S_2 \subseteq X \setminus \{X_{i_2}, X_{j_2}\}$. Therefore, it follows from Theorem 3.2 of the main paper and the consistency of the sample partial correlations that

$$\begin{aligned}
&\mathbb{P}\left(\text{AGES}(X^{(n)}) \neq A_0\right) \\
&\leq \mathbb{P}\left(\text{rank}\,(\hat{\boldsymbol{\gamma}}) \neq \text{rank}(\boldsymbol{\gamma})\right) \\
&\leq \sum_{\substack{1 \leq i < j \leq p, \\ S \subseteq X \setminus \{X_i, X_j\}}} \mathbb{P}\left(\left||\hat{\rho}_{X_i,X_j|S}| - |\rho_{X_i,X_j|S}|\right| \geq \epsilon/2\right)
\end{aligned}$$

converges to zero as the sample size tends to infinity. □

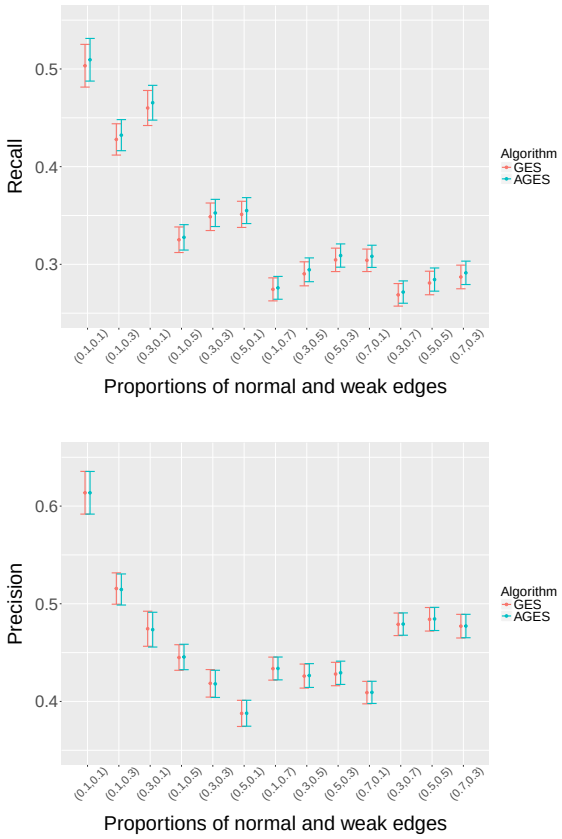


Figure 2: Mean precision and recall of GES and AGES over 500 simulations for all combinations of $(q_s, q_w) \in \{0.1, 0.3, 0.5, 0.7\}$ such that $q_s + q_w \leqslant 1$, for $p = 10$, $\lambda = \log(n)/(2n)$ and $n = 100$ (see Section 3). The bars in the plots correspond to $\pm$ twice the standard error of the mean.

3 ADDITIONAL SIMULATION RESULTS WITH $p = 10$

We also ran AGES on the settings described in the main paper but with smaller sample sizes. Figures 2 and 3 show the results for $n = 100$ and $n = 1000$, respectively, based on 500 simulations per setting.

For the larger sample size, $n = 1000$, we see that we still gain in recall and that the precision remains roughly constant. For the smaller sample size, $n = 100$, the differences become minimal. In all cases AGES performs at least as good as GES.

With a sample size of 100 we expect to detect only partial correlations with an absolute value larger than 0.21. This can be derived solving $1/2\log(1-\rho^2) = -\log(n)/(2n)$ for ρ. This limits the possibility of detecting weak edges, and if an edge is not contained in the output of GES it is also not contained in the output of AGES. This

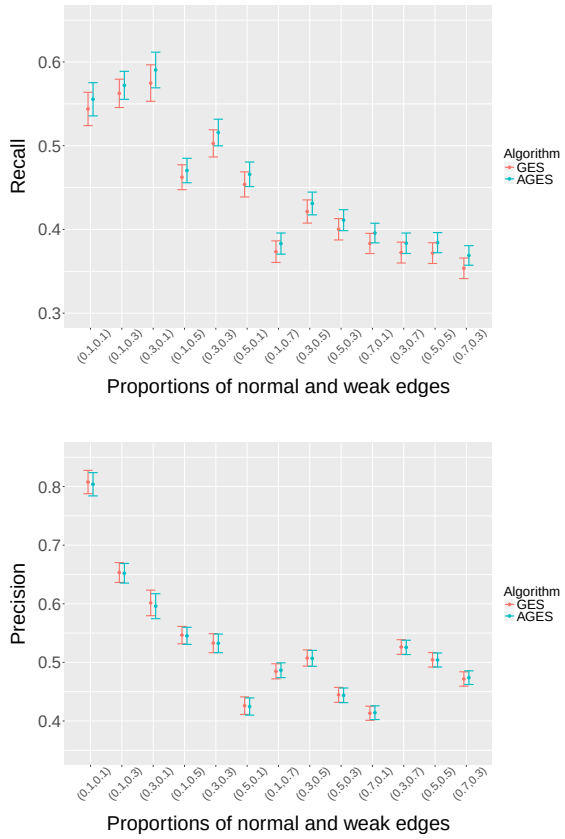


Figure 3: Mean precision and recall of GES and AGES over 500 simulations for all combinations of $(q_s, q_w) \in \{0.1, 0.3, 0.5, 0.7\}$ such that $q_s + q_w \leqslant 1$, for $p = 10$, $\lambda = \log(n)/(2n)$ and $n = 1000$ (see Section 3).The bars in the plots correspond to $\pm$ twice the standard error of the mean.

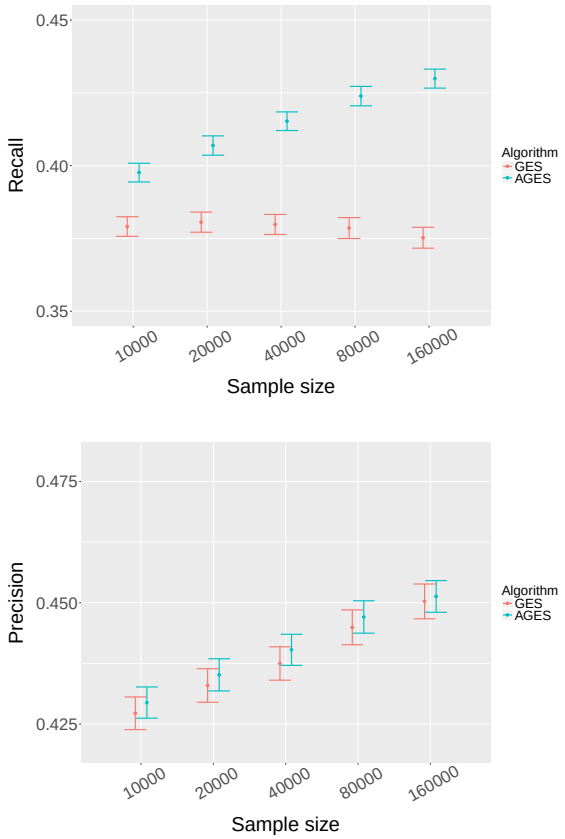


Figure 4: Mean precision and recall of GES and AGES with ARGES-skeleton over 500 simulations for $(q_s, q_w) = (0.3, 0.7)$, $p = 100$, $\lambda = \log(n)/(2n)$, and varying sample sizes (see Section 4). The bars in the plots correspond to $\pm$ twice the standard error of the mean.

explains why we do not see a large improvement with smaller sample sizes. However, AGES then simply returns an APDAG which is very similar, or identical, to the CPDAG returned by GES.

4 FURTHER SIMULATION RESULTS WITH $p = 100$

We randomly generated 500 DAGs consisting of 10 disjoint blocks of complete DAGs, where each block contains strong and weak edges with concentration probabilities $(q_s, q_w) = (0.3, 0.7)$. The absolute values of the strong and weak edge weights are drawn from Unif(0.8,1.2) and Unif(0.1,0.3), respectively. The sign of each edge weight is chosen to be positive or negative with equal probabilities. The variance of the error variables are drawn from Unif(0.5,1.5).

This setting leads more often to a violation of the skeleton condition of Algorithm 3 of the main paper, i.e., the skeleton of the output of GES with $\lambda > \log(n)/(2n)$ is not a subset of the skeleton of the output of GES with $\lambda = \log(n)/(2n)$. This results in almost identical outputs of GES and AGES. In order to alleviate this issue, in each step of AGES with $\lambda > \log(n)/(2n)$, we replace GES with the ARGES-skeleton algorithm of Nandy et al. (2015), based on the skeleton of the output of GES with $\lambda = \log(n)/(2n)$. ARGES-skeleton based on an estimated CPDAG is a hybrid algorithm that operates on a restricted search space determined by the estimated CPDAG and an adaptive modification. The adaptive modification was proposed to retain the soundness and the consistency of GES and it can be easily checked that our soundness and consistency results continue to hold if we replace GES by ARGES-skeleton in each step of AGES with $\lambda > \log(n)/(2n)$. An additional advantage of using ARGES-skeleton is that it leads to a substantial improvement in the runtime of AGES.

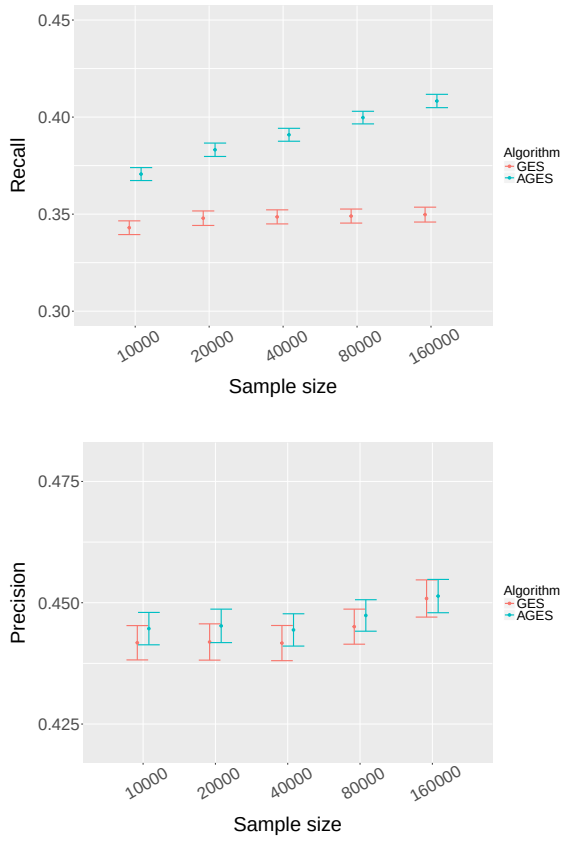


Figure 5: Mean precision and recall of GES and AGES with ARGES-skeleton over 500 simulations for $(q_s, q_w) = (0.3, 0.7)$, $p = 100$, $\lambda = \log(n)/(2n) + \log(p)$, and varying sample sizes (see Section 4). The bars in the plots correspond to $\pm$ twice the standard error of the mean.

Figure 4 shows that AGES (based on ARGES-skeleton) achieves higher recall than GES for estimating the true directions while retaining a similar precision as GES. Unsurprisingly, the difference in the recalls of AGES and GES becomes more prominent for larger sample sizes. We obtain a similar relative performance by using the extended BIC penalty $\lambda = \log(n)/(2n) + \log(p)$ (e.g., Foygel and Drton, 2010) instead of the BIC penalty (Figure 5).

5 PATH STRONG FAITHFULNESS

To produce Figure 5 of the main paper we started by determining the possible APDAGs A_0 for each choice of the edge weights. This is done by considering the four steps in Section 3.1 of the main paper. In Step S.1 we can obtain many CPDAGs (3 with one edge, 6 with two edges, and 1 with three edges). However, once we proceed to Step S.2, we note that only one DAG contains a v-structure. Hence, the orientations in the CPDAGs in Step S.3 are limited to this v-structure. Therefore, the only two possible APDAGs are given in Figures 4a and 4b of the main paper.

Now we consider possible outputs of the oracle version of AGES for every choice of the edge weights. To compute them we have to compute all marginal and partial correlations. Then, we select the in absolute value largest marginal correlation. This corresponds to the first edge addition. Now, we consider the four remaining marginal and partial correlations between non-adjacent vertices. If the in absolute value largest partial correlation is actually a marginal correlation, then we do not obtain a v-structure and the output of AGES is Figure 4b of the main paper. Otherwise, AGES recovers an APDAG with a v-structure.

With these results, we can compute the different areas depicted in Figure 5 of the main paper.

6 AGES δ-STRONG FAITHFULNESS

The path strong faithfulness assumption in Theorem 3.2 of the main paper is sufficient but not necessary for the theorem.

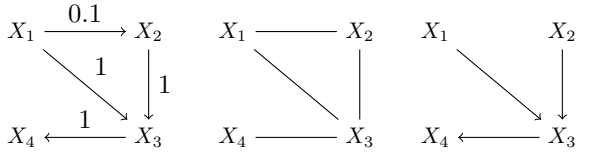
We now present an alternative strong faithfulness assumption which is weaker than path strong faithfulness. This new assumption is necessary and sufficient for Theorem 3.2 of the main paper.

In a CPDAG $C = (V, E)$ we say that $S \subseteq V \setminus \{X_i\}$ is a *possible parent set* of X_i in C if there is a DAG G in the Markov equivalence class represented by C such that $\mathrm{Pa}_G(X_i) = S$.

Definition 6.1. *A multivariate Gaussian distribution is said to be* AGES δ-strong faithful *with respect to a DAG G if it holds that $X_i \not\perp_G X_j | S \Rightarrow |\rho_{X_i,X_j|S}| > \delta$ for every triple (X_i, X_j, S) belonging to at least one of the following two sets:*

1. *Consider the output of the forward phase of oracle GES with penalty parameter $\lambda = -1/2 \log(1-\delta^2)$. The first set consists of all triples (X_i, X_j, S) such that, in this forward phase output, S is a possible parent set of X_j, X_i is a non-descendant of X_j in the DAG used to define S, and X_i and X_j are not adjacent.*

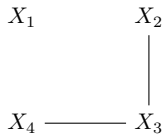
2. *Consider the backward phase of oracle GES when ran with penalty parameter λ and starting from the output of the forward phase. The second set consists of all triples (X_i, X_j, S) such that the edge between X_i and X_j has been deleted during the backward phase using S as conditioning set.*



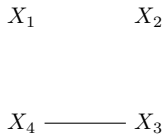
(a) True DAG G_0.

(b) CPDAG $C_0 = \mathrm{GES}_{\lambda_0}(f)$.

(c) CPDAG $C_1 = \mathrm{GES}_{\lambda_1}(f)$.



(d) CPDAG $C_2 = \mathrm{GES}_{\lambda_2}(f)$.



(e) CPDAG $C_3 = \mathrm{GES}_{\lambda_3}(f)$.

Figure 6: Graphs corresponding to Example 6.2. Figure 6a shows the true underlying DAG G_0. Figure 6b - 6e show the sub-CPDAG oracle GES found.

The need for a different condition becomes clear when we think about how GES operates. In Example 6.2, we show why path strong faithfulness is too strong.

Example 6.2. *Consider the distribution f generated from the weighted DAG G_0 in Figure 6a with $\varepsilon \sim N(0, I)$. The solution path of oracle GES is shown in Figures 6b-6e. Note that all sub-CPDAGs found by oracle GES coincide with the CPDAGs constructed as described in Step S.3 of the main paper, i.e., $C_i = \tilde{C}_i$ for $0 \leqslant i \leqslant 3$.*

Intuitively, we would like a condition that is satisfied if and only if the two CPDAGs coincide. However, this is not necessarily the case for the path strong faithfulness condition.

For the CPDAG in Figure 6e, path strong faithfulness imposes δ_3-strong faithfulness with respect to C_3, i.e., $|\rho_{X_3,X_4|S}| > \delta_3$ for all sets S not containing X_3 or X_4. However, the forward phase of GES only checks the marginal correlation between X_3 and X_4. The same is true for the backward phase.

Consider now Figure 6d. Path strong faithfulness imposes δ_2-strong faithfulness with respect to C_2. For instance, it requires that $|\rho_{X_2,X_4|X_1}| > \delta_2$. However, this partial correlation does not correspond to a possible edge addition. Hence, this constraint is not needed, and it is not imposed by AGES δ-strong faithfulness.

In this example, $|\rho_{X_2,X_4|X_1}| < \delta_2 = |\rho_{X_1,X_3|X_2}|$. Hence, f does not satisfy δ_2-strong faithfulness with respect to C_2, but it does satisfy AGES δ_2-strong faithfulness with respect to C_2.

The following lemma states that the AGES δ-strong faithfulness assumption is necessary and sufficient for Claim 1 and Claim 2 in the proof of Theorem 3.2 of the main paper.

Lemma 6.3. *Given a multivariate Gaussian distribution f and a CPDAG C on the same set of vertices, $\mathrm{GES}(f, \lambda)$ with $\lambda = -1/2 \log(1 - \delta^2)$ is an I-map of C if and only if f is AGES δ-strong faithful with respect to C.*

Proof. For a CPDAG C, we use the notation $X_i \perp_C X_j | S$ to denote that $X_i \perp_G X_j | S$ in any DAG G in the Markov equivalence class described by C.

We first prove the "if" part. Thus, assume that f is AGES δ-strong faithful with respect to C. We consider running oracle GES with $\lambda = -1/2 \log(1 - \delta^2)$, and denote by C^f and C^b the output of the forward and backward phase, respectively.

Claim 1: C^f is an I-map of C, i.e., all d-separation constraints true in C^f are also true in C.

Proof of Claim 1:

For each triple (X_i, X_j, S) contained in the first set of Definition 6.1, we have $|\rho_{X_i,X_j|S}| < \delta$, since otherwise there would have been another edge addition. From AGES δ-strong faithfulness, it follows that $X_i \perp_C X_j | S$. Since this set of triples characterizes the d-separations that hold in C^f, all d-separations that hold in C^f also hold in C.

Claim 2: C^b is an I-map of C, i.e., all d-separation constraints true in C^b are also true in C.

Proof of Claim 2: By Claim 1 the backward phase starts with an I-map of C. Suppose it ends with a CPDAG that is not an I-map of C. Then, at some point there is an edge deletion which turns a DAG G that is an I-map of C into a DAG G' that is no longer an I-map of C. Suppose the deleted edge is (X_i, X_j). By Lemma 1.1, we have $X_i \not\perp_C X_j | \mathrm{Pa}_{G'}(X_j)$. Since the edge has been deleted, the corresponding triple $(X_i, X_j, \mathrm{Pa}_{G'}(X_j))$ is contained in the second set of Definition 6.1. Hence, by AGES δ-strong faithfulness, we obtain $|\rho_{X_i,X_j|\,\mathrm{Pa}_{G'}(X_j)}| > \delta$. Thus, deleting this edge would worsen the score. This is a contradiction to the GES algorithm deleting this edge.

We now prove the "only if" part. Thus, suppose there is a triple (X_i, X_j, S) in one of the sets in Definition 6.1 such that $|\rho_{X_i,X_j|S}| < \delta$ and $X_i \not\perp_C X_j | S$.

Suppose first that this triple concerns the first set. Since all triples in the first set characterize the d-separations that hold in C^f, we know that $X_i \perp_{C^f} X_j | S$. Therefore, C^f is not an I-map of C. Hence, C^b is certainly not an I-map of C.

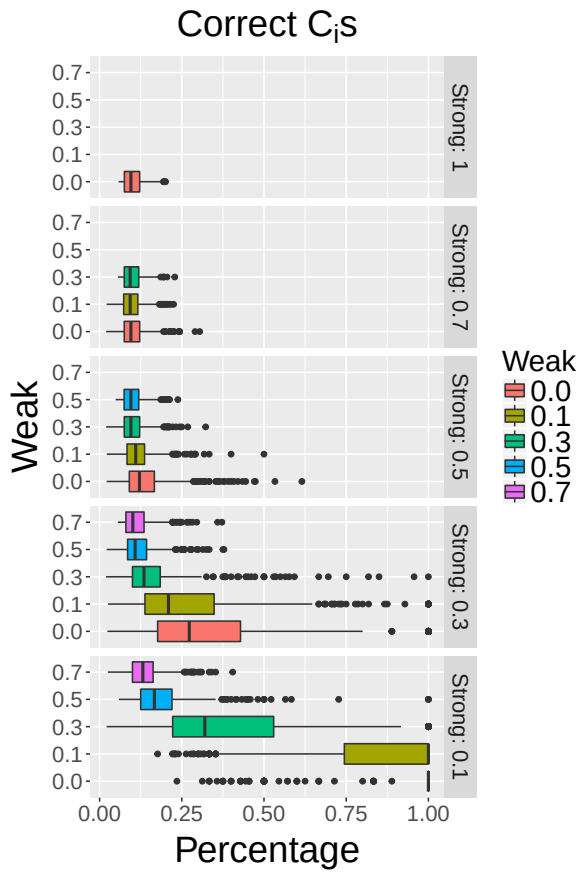


Figure 7: Boxplots of the proportion of correct sub-CPDAGs $\tilde{C}_1, \ldots, \tilde{C}_k$ (as defined in Step S.3 of Section 3.1 of the main paper) found by oracle AGES in each solution path (see Section 6). The different colors represent the different proportions of weak edges. The plots are grouped by the proportion of strong edges.

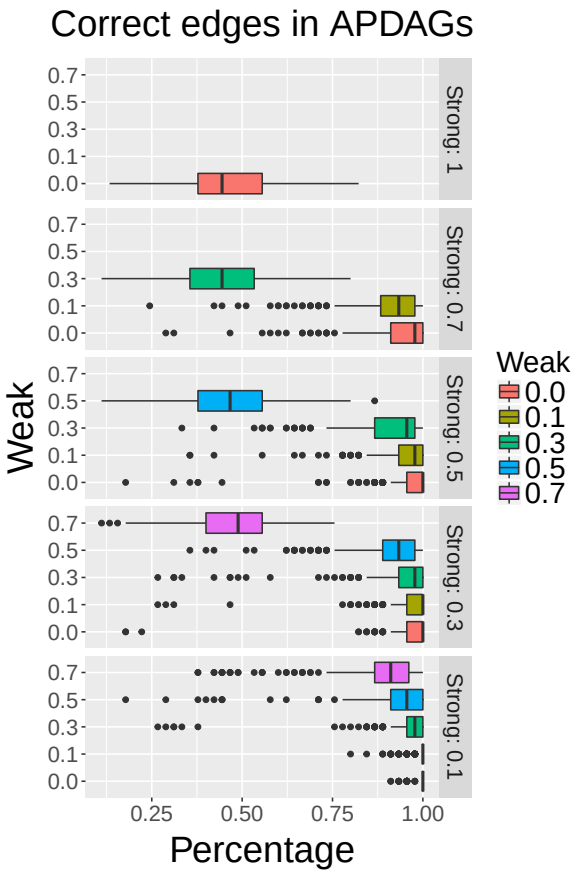


Figure 8: Boxplots of the proportion of edge orientations in the APDAGs found by oracle AGES that are equal to the edge orientations in the true APDAGs (see Section 6). The different colors represent the different proportions of weak edges. The plots are grouped by the proportion of strong edges.

Next, suppose the triple concerns the second set. This means that at some point there is an edge deletion which turns a DAG G into a DAG G' by deleting the edge $X_i \to X_j$, using S as conditioning set. This means that $S = \mathrm{Pa}_G(X_j) \setminus \{X_i\} = \mathrm{Pa}_{G'}(X_j)$. In the resulting DAG G', X_i and X_j are therefore d-separated given S. But we know that $X_i \not\perp_C X_j | S$. Hence, C^b is not an I-map of C. □

We analysed how often the AGES δ-strong faithfulness assumption is met in the simulations presented in the main paper, as well as how often oracle AGES is able to find the correct APDAG. Lemma 6.3 provides a necessary and sufficient condition for the equality of the CPDAGs of Theorem 3.2 of the main paper. For the equality of the APDAGs this condition is only sufficient.

Figure 7 shows the proportion of correct sub-CPDAGs $\tilde{C}_1, \ldots, \tilde{C}_k$ (as defined in Step S.3 of Section 3.1 of the main paper) found by oracle AGES in each solution path and for all simulated settings. We can see that the sparsity of the true underlying DAG plays an important role in the satisfiability of the assumption. We can also see that for the same total sparsity, the settings with more weak edges produce better results.

Even though the AGES δ-strong faithfulness assumption is not very often satisfied for denser graphs, it is much weaker than the classical δ-strong faithfulness assumption. Indeed, we verified that the δ-strong faithfulness assumption is rarely satisfied even for single sub-CPDAGs C_i.

Figure 8 shows the proportion of edge orientations in the APDAGs found by oracle AGES that are equal to the edge orientations in the true APDAGs. With equal edge orientations, we mean that the edges have to be ex-

actly equal. For example, an edge that is oriented in the APDAG found by oracle AGES, but oriented the other way around or unoriented in the true APDAG counts as an error. We see that in many settings AGES can correctly find a large proportion of the edge orientations.

7 APPLICATION TO DATA FROM SACHS ET AL., 2005

We log-transformed the data because they were heavily right skewed. Based on the network provided in Figure 2 of Sachs et al. (2005), we produced the DAG depicted in Figure 9 that we used as partial ground truth. In the presented network, only two variables are connected by a bi-directed edge, meaning that there is a feedback loop between them. To be more conservative, we omitted this edge.

For the comparison of GES and AGES we need to account for the interventions done in the 14 experimental conditions. Following Mooij and Heskes (2013), we distinguish between an intervention that changes the abundance of a molecule and an intervention that changes the activity of a molecule. Interventions that change the abundance of a molecule can be treated as do-interventions (Pearl, 2009), i.e., we delete the edges between the variable and its parents. Activity interventions, however, change the relationship with the children, but the causal connection remains. For this reason, we do not delete edges for such interventions. We also do not distinguish between an activation and an inhibition of a molecule. All this information is provided in Table 1 of Sachs et al. (2005).

The only abundance intervention done in the six experimental conditions we consider in Table 1 of the main paper is experimental condition 5. This intervention concerns $PIP2$. For this reason, when comparing the outputs of GES and AGES we need to consider the DAG in Figure 9 with the edge $PLC_\gamma \rightarrow PIP2$ deleted. For the other five experimental conditions we used the DAG depicted in Figure 9 as ground truth.

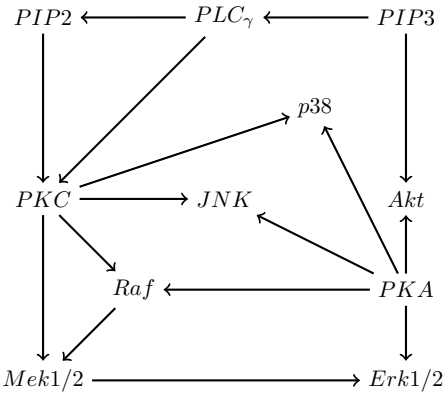


Figure 9: The DAG used as partial ground truth derived from the conventionally accepted network (Sachs et al., 2005).

References

Chickering, D. M. (2002). Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3:507–554.

Foygel, R. and Drton, M. (2010). *Extended Bayesian information criteria for Gaussian graphical models*.

Meek, C. (1995). Causal inference and causal explanation with background knowledge. In *Proceedings of UAI 1995*, pages 403–410.

Mooij, J. M. and Heskes, T. (2013). Cyclic causal discovery from continuous equilibrium data. In *Proceedings of UAI 2013*, pages 431–439.

Nandy, P., Hauser, A., and Maathuis, M. H. (2015). High-dimensional consistency in score-based and hybrid structure learning. arXiv:1507.02608v4.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, 2nd edition.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308:523–529.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of UAI 1990*, pages 255–270.