Inference of Causal Effects when Control Variables are Unknown Supplementary Material

Ludvig Hult[1] **Dave Zachariah**[1]

[1]Systems and Control, Uppsala Universitet, Uppsala, Sweden

1 SUPPLEMENTARY MATERIAL

1.1 LEMMAS AND PROOFS

Lemma 1. *For the matrix $M = (I - ZW^\top)^{-1}$, the element M_{1i} is equal to the Kronecker delta δ_{1i}, for $W \in \mathbb{R}^{d\times d}$ and Z from equation* (7)*.*

Proof of Lemma 1. Using Cramers rule $M_{1i} = \frac{1}{\det(I-ZW^\top)}C_{i1}$, where C is the cofactor matrix of $(I - ZW^\top)$.

By definition of a cofactor as plus/minus a minor, and that the first row of $(I - ZW^\top)$ is zero for all but the first element, C_{i1} is zero for $i > 1$, so $C_{i1} = \delta_{i1}C_{11}$

By Laplace expansion of $\det(I - ZW^\top)$ along the first row

$$\det(I - ZW^\top) = \sum_{k=1}^{d}(I - ZW^\top)_{1k}C_{1k} = C_{11}$$

We conclude $M_{1i} = \frac{1}{C_{11}}\delta_{i1}C_{11} = \delta_{1i}$ □

Proof of lemma 1. We need to show the result of equation (9). Introduce $M = (I - ZW)^{-1}$.

The proof follows by a direct computation, using Lemma 1. The noise covariance under the interventional distribution $\widetilde{\Sigma}$ is diagonal by assumption, which is also key.

$$\begin{aligned}
\gamma(W) &= \frac{\widetilde{\mathrm{Cov}}_W[x, y]}{\widetilde{\mathrm{Var}}_W[x]} && (1)\\
&= \frac{\widetilde{\mathrm{Cov}}_W[v, v]_{1,2}}{\widetilde{\mathrm{Cov}}_W[v, v]_{1,1}} && (2)\\
&= \frac{\sum_{i,j=1}^{d} M_{1j}M_{2i}\widetilde{\Sigma}_{ij}}{\sum_{i,j=1}^{d} M_{1j}M_{1i}\widetilde{\Sigma}_{ij}} && (3)\\
&= \frac{\sum_{i=1}^{d} M_{2i}\widetilde{\Sigma}_{i1}}{\widetilde{\Sigma}_{11}} && (4)\\
&= \frac{M_{21}\widetilde{\Sigma}_{11}}{\widetilde{\Sigma}_{11}} && (5)\\
&= M_{21} && (6)
\end{aligned}$$

Supplement for the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021).

This completes the proof. □

We notice that there is nothing in the proofs of Lemma 1 and Lemma 1 specific about the first and second component - redefining the matrix Z accordingly, it is straight forward to generalize the result if needed. To keep the notation simple, we do stay with the convention that the first component is the one we intervene on, and that the second is the outcome of interest.

Lemma 2. *The function h of Zheng et al. [2018] has a closed form matrix gradient. It is $\nabla h(W) = 2W \circ (\exp[W \circ W])^\top$.*

This formula is reported by Zheng et al. [2018], but without derivation. The result follows from liberal application of the chain rule.

Proof of Lemma 2. $\frac{\partial}{\partial A_{i,j}} \operatorname{tr} A^k = k(A^{k-1})^\top_{i,j}$ by the product rule for derivation, and cyclicity of traces.

By series expansion and using the equation above $\frac{\partial}{\partial A_{i,j}} \operatorname{tr} \exp[A] = (\exp[A])^\top_{i,j}$

We have that $\frac{\partial (W \circ W)_{k,l}}{\partial W_{i,j}} = 2W_{i,j}\delta_{i,k}\delta_{j,l}$ using the Kronecker delta symbol.

The chain rule for differentiation now says $\frac{\partial}{\partial W_{i,j}} \operatorname{tr} \exp[W \circ W] = \sum_{k,l} \frac{\partial \operatorname{tr} \exp[W \circ W]}{\partial (W \circ W)_{k,l}} \frac{\partial (W \circ W)_{k,l}}{\partial W_{i,j}} = 2W_{i,j} \frac{\partial \operatorname{tr} \exp[W \circ W]}{\partial (W \circ W)_{i,j}} = 2W_{i,j}(\exp[W \circ W])^\top_{i,j}$

The rest is a matter of notation and diffrentiating a constant. □

Lemma 3. *The set of all* DAG*:s, $\mathcal{W}_0$ in* (6)*, has the following properties*

1. *All points of $\mathcal{W}_0$ are boundary points (i.e., empty interior)*
2. *$\mathcal{W}_0$ is a direct sum of linear subspaces, so it is a unbounded set, and a cone*
3. *$\mathcal{W}_0$ is nonconvex. The convex hull of $\mathcal{W}_0$ is the set of all real $d \times d$-matrices.*
4. *$h(W) = 0$ iff $\nabla h(W) = 0$.*

Proof of Lemma 3. Only point four is a nontrivial result, as the others have a direct geometrical interpretation.

The first point follows from the fact that for q being any matrix with a nonzero on the diagonal, $h(W + \varepsilon q) > 0 \quad \forall \varepsilon > 0$, even when $W \in \mathcal{W}$

The second point follows from the fact that $h(W) = 0$ iff W is the weighted directed adjacency matrix of a DAG, and positive scaling that matrix will not affect the cyclicity structure.

The third point: Consider the example $w = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. Then, $w, w^\top \in \mathcal{W}$, but $(w + w^\top)/2 \notin \mathcal{W}_0$, so $\mathcal{W}_0$ is nonconvex. Consider also an arbitrary matrix $W = \sum_{ij=1}^{d} w_{ij} E^{ij}$. It is a convex combination of the matrices E^{ij}, which all belong to $\mathcal{W}_0$. Since W was arbitrary, all matrices are in the convex hull of $\mathcal{W}_0$.

The last point needs some more work, and is detailed below.

We start with the forward implication. Since any DAG W is permutation similar to a strictly upper triangular matrix, $(\exp[W \circ W])^\top$ is permutation similar to a strictly lower triangular matrix, with the same similarity transformation. $\nabla h(W)$ is therefore permutation similar to the elementwise product between a strictly upper and a strictly lower triangular matrix, which must be the zero matrix.

For the the reverse implication, assume W is not a DAG, so it has some cycle of length K, and $1 \leq K \leq d$. Select i and j such that node i and j lies on that cycle. Now $W_{i,j} \neq 0$. One can go from node i to node j in 1 step, so one must be able to go from node j to node i in $K - 1$ steps. Therefore $(W \circ W)^{K-1}_{j,i} \neq 0$. This makes sure that the exponential factor in $\nabla h(W)$ has a nonzero i, j-entry.

$$\left[(\exp[W \circ W])^\top\right]_{i,j} = \sum_{k=0}^{\infty} \frac{[(W \circ W)^k]_{j,i}}{k!} \neq 0$$

$$\nabla h(W)_{i,j} = 2W_{i,j}\left[(\exp[W \circ W])^\top\right]_{i,j}$$

Since this is a product of two positive real numbers, we can conclude that $\nabla h(W) \neq 0$. □

This result supplements the discussion of Zheng et al. [2018, p.7]. Not only is the DAG:s the global minima of h, but they are also the zeroes of ∇h.

The fourth point in Lemma 3 has during the time of writing this being reported in Wei et al. [2020, lemma 4], but with a more different derivation technique valid for a slightly broader class of h-functions. It has also been reported in Ng et al. [2019, proposition 1], with a proof technique very similar to ours.

Lemma 4. *The least-squares objective, and its derivatives are*

$$\ell_\theta(v) = \frac{1}{2}(L\theta - \operatorname{vec}(I))^\top \left[\Sigma^{-1} \otimes vv^\top\right](L\theta - \operatorname{vec}(I)) \tag{7}$$

and its gradient and hessian is

$$\nabla \ell_\theta(v) = L^\top \left[\Sigma^{-1} \otimes vv^\top\right](L\theta - \operatorname{vec}(I)) \tag{8}$$
$$\nabla^2 \ell_\theta(v) = L^\top \left[\Sigma^{-1} \otimes vv^\top\right] L$$

The proof is direct computation, after using the formula $\operatorname{tr}(A^\top Y^\top BX) = (\operatorname{vec}(Y))^\top [A \otimes B] \operatorname{vec}(B)$.

Proof of Lemma 4. Use the vec-trick $\operatorname{tr}(A^\top Y^\top BX) = \operatorname{vec}(Y)^\top [A \otimes B] \operatorname{vec}(B)$, and find the objective.

$$\ell_\theta(v) = \frac{1}{2}\|\Sigma^{-1/2}\left(I - \operatorname{mat}(L\theta)^\top\right)v\|^2 \tag{9}$$
$$= \frac{1}{2}\operatorname{tr}\left[\Sigma^{-1}\left(\operatorname{mat}(L\theta) - I\right)^\top vv^\top \left(\operatorname{mat}(L\theta) - I\right)\right] \tag{10}$$
$$= \frac{1}{2}(L\theta - \operatorname{vec}(I))^\top \left[\Sigma^{-1} \otimes vv^\top\right](L\theta - \operatorname{vec}(I)) \tag{11}$$

The rest is differentiation of a quadratic. □

Lemma 5. *The quantities of Lemma 3 can be computed to be*

$$K_n = L^\top \left[\Sigma^{-1} \otimes \mathbb{E}_n\left[vv^\top\right]\right] L$$

$$\Pi_n = I - (qq^\top)/(q^\top q)$$
$$q = L^\top \operatorname{vec}(2W_n \circ (\exp[W_n \circ W_n])^\top)$$
$$J_n = L^\top \tilde{J}_n L$$

$$(\tilde{J}_n)_{d(j-1)+i,d(l-1)+k} = \sum_{q,r,o,p=1}^{d} \Big\{\big(\mathbb{E}_n\left[v_i v_q v_o v_k\right] - \mathbb{E}_n\left[v_i v_q\right]\mathbb{E}_n\left[v_o v_k\right]\big)\Sigma^{-1}_{j,r}\Sigma^{-1}_{p,l}(W-I)_{q,r}(W-I)_{o,p}\Big\} \tag{12}$$

Proof of Lemma 5. The expression for K_n follows from Lemma 4.

$$K_n = \mathbb{E}_n[\nabla^2 \ell_\theta(v)] = \mathbb{E}_n\left[L^\top \left[\Sigma^{-1} \otimes vv^\top\right] L\right] = L^\top \left[\Sigma^{-1} \otimes \mathbb{E}_n\left[vv^\top\right]\right] L \tag{13}$$

Π_n is a projection matrix with respect to the orthogonal complement of $q \coloneqq \nabla_\theta h(\operatorname{mat}(L\theta_n))$. Since q is a vector, projection on the orthogonal complement is $\Pi_n = I - (qq^\top)/(q^\top q)$. The expression $q = L^\top \operatorname{vec}(2W_n \circ (\exp[W_n \circ W_n])^\top)$ follows from Lemma 2, and $W_n = \operatorname{vec} L\theta_n$.

The derivation of J_n is an mostly tracking indices. Start with $J_n = \mathbb{E}_n[\nabla \ell_{\theta_n}(v) \nabla \ell_{\theta_n}(v)^\top] - \mathbb{E}_n[\nabla \ell_{\theta_n}(v)] \mathbb{E}_n[\nabla \ell_{\theta_n}(v)]^\top$ and apply to Lemma 4. First factor out the L matrix of (8), and then covert the rest into indices. Apply the index conversion for vectorizations $\operatorname{vec} A_{d(j-1)+i} = A_{i,j}$ and for kronecker products $[A \otimes B]_{d(i-1)+j,d(k-1)+l} = A_{i,k} B_{j,l}$ when A and B are $d \times d$ sized. □

The next lemma collects the assumption verification for applying Corollary 6 in proof of Lemma 3. Herein we use the redundant norm-constraint, that is in some parts skipped.

Lemma 6. *Using the loss function* (13)*, and the parameter set* $\Theta := \{\theta \mid |h\left(\operatorname{mat}(L\theta) - \epsilon = 0 \wedge \|\theta\| \leq B\right\}$*, we see that*

1. *The techincal conditions for M-estimation [Wooldridge, 2010, Theorem 12.2] holds.*
2. *The loss function* $\ell_\theta(v)$ *is two times continously diffrentiable in* v*.*
3. $\Theta := \{\theta \in \mathbb{R}^p \mid g(\theta) = 0\}$ *for some vector-valued constraint function* g *such that* Θ *is bounded.*
4. *The Jacobian matrix* $\nabla g(\theta_n)$ *has full rank for all* n*.*
5. $\mathbb{E}_n\left[\nabla^2 \ell_\theta(v)\right]$ *is invertible for all* θ*.*
6. $\theta_\circ$ *is the unique minimizer of* $\mathbb{E}[\ell_\theta(v)]$

Proof. First notice that (13) is quadratic in θ, but also in v, which is more clearly seen in (11).

1. The technical conditions are (a) that Θ is compact, which follows from closed and boundedness (b) that $\ell_\theta(v)$ is borel measurable in v for each θ, which follow from being quadratic, (c) that $\ell_\theta(v)$ is continuous in θ for each v, which follows from being a quadratic and (d) that there is a dominating function $d(v) \geq |\ell_\theta(v)|$ for all θ so that $\mathbb{E}[d(v)] < \infty$, which needs a few steps to prove. Observe

$$|\ell_\theta(v)| = \frac{1}{2} \|\Sigma^{-1/2}(I - \operatorname{mat}(L\theta))v\|_2^2 \tag{14}$$
$$\leq \frac{1}{2} \sigma_1(\Sigma^{-1/2})^2 \sigma_1(I - \operatorname{mat}(L\theta))^2 \|v\|^2 \tag{15}$$
$$\leq C\|v\|^2 =: d(v), \tag{16}$$

   where σ_1 denotes the largest singular value and

$$C := \frac{1}{2} \sigma_1(\Sigma^{-1/2})^2 \max_{\theta \in \Theta} \sigma_1(I - \operatorname{mat}(L\theta))^2,$$

   utilizing compactness of Θ. Finally $\mathbb{E}[d(v)] = C\mathbb{E}[\|v\|^2] = C \operatorname{tr}\left[(I - W^\top)^{-1} \Sigma (I - W)^{-1}\right] \leq \infty$, using the assumed data generating process (5).
2. $\ell_\theta(v)$ is two times continously diffrentiable in v, since it is a quadratic in v
3. The form of $\Theta := \{\theta \mid |h\left(\operatorname{mat}(L\theta) - \epsilon = 0 \wedge \|\theta\| \leq B\right\}$ can be transformed into equality form by introduction of a slack variable s, so that $\Theta := \{\theta, s \mid |h\left(\operatorname{mat}(L\theta) - \epsilon = 0 \wedge \|\theta\| + s^2 - B = 0\right\}$, so $g(s, \theta) = \begin{bmatrix} h\left(\operatorname{mat}(L\theta) - \epsilon \right. \\ \|\theta\| + s^2 - B \end{bmatrix}$.
4. By lemma 3, $\nabla g(\theta_n)$ is nonzero over Θ, but the gradient with respect to the slack is zero. Furthermore $\nabla_s[\|\theta\| + s^2 - B] = 2s$, which is zero only for $s = 0$, but we know from 2 that $s \neq 0$. So the two components of g must have linerarly independent gradients, and the jacobian has full rank. Do note that the slack-formulation used here is supressed from the formalism in the rest of the article, since it is an inactive constraint, making the proofs and text less clear with no gain.
5. $\mathbb{E}_n\left[\nabla^2 \ell_\theta(v)\right] = L^\top \left[\Sigma^{-1} \otimes \mathbb{E}_n[vv^\top]\right] L$, which almost surely has full rank. We ignore the measure zero case.
6. The unicity of $\theta_\circ$ we have to take by assumption, as discussed elsewhere in this article.

□

Lemma 7. *The gradient of the causal effect* γ *with respect to the parameter* θ *is*

$$[\nabla_\theta \gamma(\theta)]_k = -\left([MZ \otimes I]\, L\right)_{d+1,k} \tag{17}$$

Proof of Lemma 7. Start from Lemma 1. Apply derivation rules for matrix inverses, and utilize the unit basis matrices $E^{i,j}$ which zero in every entry, except the i, j-entry.

$$\frac{\partial(\gamma(W))}{\partial W_{i,j}} = \frac{\partial(M_{21})}{\partial W_{i,j}} \tag{18}$$

$$= \sum_{k,l=1}^{d} M_{2k} \frac{\partial(I - ZW^\top))_{kl}}{\partial W_{i,j}} M_{l1} \tag{19}$$

$$= - \sum_{k,l=1}^{d} M_{2k} Z_{km} E^{ij}_{lm} M_{l1} \tag{20}$$

$$= -(MZ)_{2j} M_{i1} \tag{21}$$

$$= - \left[MZ \otimes M^\top\right]_{d+1,d(j-1)+i} \tag{22}$$

$$\tag{23}$$

As an aside, we can note that the matrix with these entries has a compact definition, $-(\left[MZ \otimes M^\top\right]) = \frac{\partial \operatorname{vec}(M^T)}{\partial \operatorname{vec} W}$. Armed with this expression and

$$\frac{\partial W_{i,j}}{\partial \theta_k} = L_{d(j-1)+i,k} \tag{24}$$

we can compute

$$[\nabla_\theta \gamma(\theta)]_k = \sum_{i,j=1}^{d} \frac{\partial(\gamma(W))}{\partial W_{i,j}} \frac{\partial W_{i,j}}{\partial \theta_k} \tag{25}$$

$$= - \left([MZ \otimes I]\, L\right)_{d+1,k} \tag{26}$$

□

1.2 NUMERICAL EXPERIMENTS

1.2.1 Detailed sensitivity study

In section 4.5 we studied the impact of ϵ in relation to our causal effect measure $\gamma_\circ$. In this section, we provide additional results (in Figure 1) that shed more light on the behavior of the solution.

The computations are performed as in in section 4.5, but with 20 random graphs instead of 10, and a wider range of ϵ

Comparing Figures 1d and 1b, we note that while setting $\epsilon > \epsilon_\star$ yields an inaccurate non-DAG matrix $W_\circ(\epsilon)$, it may occasionally produce accurate $\hat{\gamma_\circ}(\epsilon)$ depending on the unknown data-generating process and the nonlinear mapping in (9).

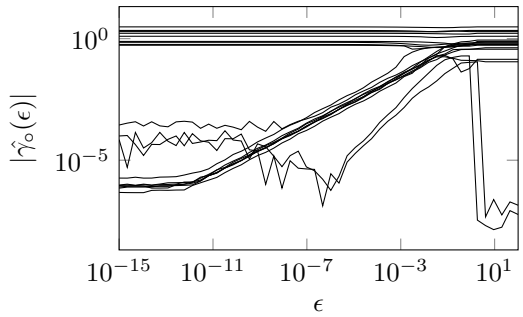
In Figure 1c we see that to improve the DAG-fidelity (quantified by $h(W)$), we need to reduce η. However, in the numerical runs, we could see that required raising ρ_{max} further, which may lead to numerical inaccuracies.

1.2.2 Linearity assumptions violations

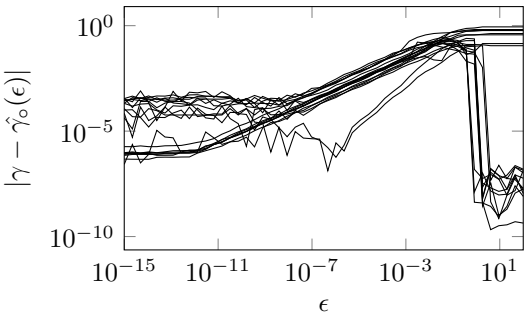
All numerical experiments above been performed using data drawn from *linear* SCMs. We now consider the behavior of the method when the data-generating process is non-linear and study the coverage of the target quantity $\gamma_\circ$. It is still defined in (10a) as the average causal effect of the optimal linear SCM (although it will diverge from the unknown distribution parameter γ depending on the type of nonlinearity).

We use the same models as Yu et al. [2019]:

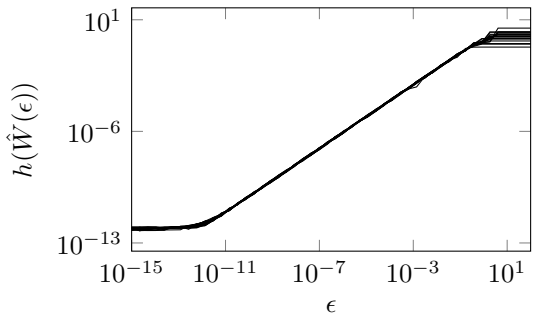
1. Linear: $v = W^\top v + e$ where
2. Nonlinear 1: $v = W^\top \cos(v + \mathbf{1}) + e$,



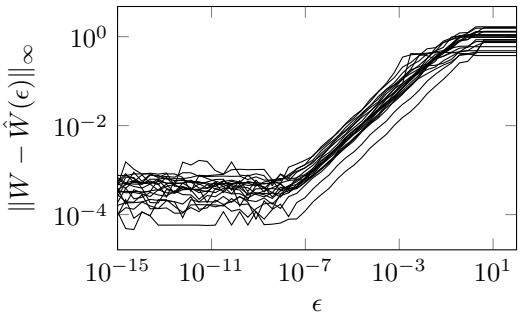
(a) The average causal effect estimated for various ϵ. Absolute value imposed to allow log-log-plot.



(b) The absolute error in the estimate of the causal effect. Smilar to figure 5.



(c) The constraint function h at the numerical approximation of the ϵ-almost DAG $W_\circ$. If the numerical solver is good and $\epsilon \leq \epsilon_\star$, we should have $h(\hat{W(\epsilon)}) \approx \epsilon$, which is what we observe down to circa $10^{-12} = \eta$, the tolerated constraint violation of Algorithm 1. We can also see that when $\epsilon > \epsilon_\star$, the solution does not depend on ϵ.



(d) The maximum error in the point estimate of the adjacency matrix W. The results indicate $\epsilon \to 0$ is a necessary condition to retrieve the true DAG-matrix W, but numerical precision limits this convergence.

Figure 1: Detailed graphs for the extended sensitivity analysis. We conclude that $\epsilon \to 0$ is a strong indication that $W_\circ(\epsilon) \to W_\circ(0)$..

3. Nonlinear 2: $v = 2\sin(W^\top(v + 0.5 \cdot \mathbf{1})) + W^\top(v + 0.5 \cdot \mathbf{1}) + e$

The coefficient matrix W is generated as in section 4 and the random elements of e are drawn independently as $\mathcal{N}(0, 1)$. Let $\mathbf{1}$ denote a vector of ones, and $\cos(\cdot)$ and $\sin(\cdot)$ on vectors be defined entry-wise. For each of these models $n = 10^3$ data points are generated.

We performed 200 Monte Carlo runs and report the empirical coverage rate CR of $\Gamma_{\alpha,n}$ in Table 1, d is the number of nodes in the SCM and k denotes the number of number of expected edges per node. We find that in all cases the empirical coverage rate exceeds the target $1 - \alpha = 95\%$, in accordance with the theory, but the confidence interval is more conservative in the nonlinear cases than the linear case.

1.2.3 Misspecified latent covariance structure

One of the major challenges of the method is the assumption of an approximately known latent covariance Σ. This section explores the sensitivity to misspecification in this parameter.

First, we restate Loh and Bühlmann [2014, Theorem 9]. Let $W_1 \gg W_0$ if the directed graph encoded by W_1 is a supergraph of W_0. *I.e.* for all indices i, j, $[W_0]_{i,j} \neq 0$ implies $[W_1]_{i,j} \neq 0$. The converse, $W_1 \not\gg W_0$ means that there is some component of W_1 that is zero, even though the corresponding component of W_0 is not. Define the *additive gap* ξ to be

Table 1: Empirical coverage rates of $\Gamma_{n,\alpha\%}$ from numerical experiment on linear assumption violation. Nominal coverage set to $1-\alpha = 95\%$.

d	k	linear	nonlinear1	nonlinear2
5	1	98.0%	97.0%	99.5%
5	2	97.5%	96.5%	100.0%
10	1	96.0%	98.5%	99.5%
10	2	95.5%	96.5%	100.0%

the difference in expected squared loss between the optimal DAG adjacency matrix and the second best one among the non-supergraph-models. Compare the following with (10b). Define

$$\texttt{score}(W) := \mathbb{E}\left[\|\Sigma^{-1/2}\left(I - W^\top\right)v\|^2\right] \tag{27}$$

$$W_0 := \arg\min_{W\in\mathcal{W}_0} \texttt{score}(W) \tag{28}$$

$$\xi := \min_{\substack{W\in\mathcal{W}_0 \\ W\not\gg W_0}} \{\texttt{score}(W)\} - \texttt{score}(W_0) \tag{29}$$

This gap is defined from the data generating process uniquely, and can only be computed if the the data generating latent covariance Σ is known - at least up to a scale factor. When this is not known, we assume some latent variance structure $\widehat{\Sigma}$, and quantify our misspecification by the condition number $\kappa\left(\widehat{\Sigma}^{-1}\Sigma\right)$.

Lemma 8 (Loh Bühlmann, Lemma 9)**.** *If*

$$\kappa\left(\widehat{\Sigma}^{-1}\Sigma\right) \leq 1 + \frac{\xi}{d}$$

then $W_0 \in \arg\min_{W\in\mathcal{W}_0} \mathbb{E}\left[\|\widehat{\Sigma}^{-1/2}\left(I - W^\top\right)v\|^2\right]$*. If the inqeuality is strict, then* W_0 *is the unique minimizer.*

If the structure is correctly assumed, *i.e.* $\Sigma = s\widehat{\Sigma}$ for some scaling factor s, then

$$\min_{W\in\mathcal{W}_0} \mathbb{E}\left[\|\widehat{\Sigma}^{-1/2}\left(I - W^\top\right)v\|^2\right] = sd$$

so we can estimate the scale factor s from data, assuming that we have the correct latent covariance structure $\widehat{\Sigma}$.[Loh and Bühlmann, 2014, Corollary 8] Denote this empirical estimate $\hat{s}$.

How does these results affect the confidence interval of Theorem 4? We replace Σ in (17) with $\hat{s}\widehat{\Sigma}$ using the biased estimate of the scale s. [1] We conducted numerical studies aiming to illustrate that the confidence interval is good when $\kappa\left(\widehat{\Sigma}^{-1}\Sigma\right)$ is small enough.

We generate data as in 4.3, but with a random latent noise matrix Σ. The matrix is diagonal, with entries drawn uniformly iid from from the interval $[1-\Delta, 1+\Delta]$, and $\Delta = \frac{1-\kappa_{max}}{1+\kappa_{max}}$. We use $\widehat{\Sigma} = I$ as before. This guarantees that $\kappa\left(\widehat{\Sigma}^{-1}\Sigma\right) \leq \kappa_{max}$.

For each draw of n data points, compute $\kappa\left(\widehat{\Sigma}^{-1}\Sigma\right)$, as well as $\gamma_\circ$ and Γ as described in section 4.

[1] The estimate is most likely biased since most likely $\widehat{\Sigma}$ is not proportional to the true data generating Σ.

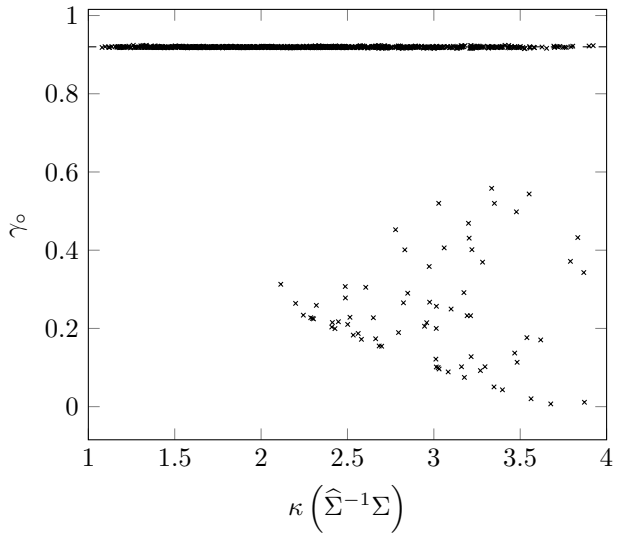


Figure 2: The average causal effect $\gamma_\circ$ is in general close to the true value, except when the condition number $\kappa\left(\widehat{\Sigma}^{-1}\Sigma\right)$ becomes larger than some threshold value. This computation is not dependant on the number of data points drawn. Every run is marked with an x, and the true average causal effect is denoted with a dashed hosrizontal line, mostly occluded by the x-marks.

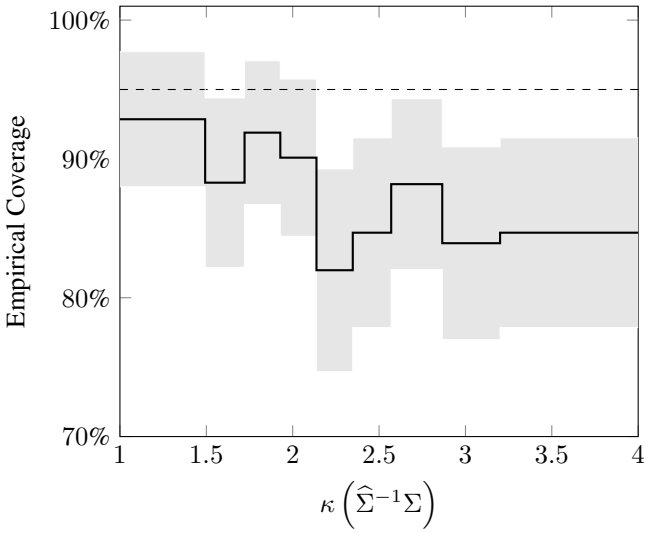


Figure 3: For $n = 100$. Empirical coverage, as the misspecification is increased. 1000 runs with random noise matrices Σ run. For each run, we have computed if $\gamma_\circ \in \Gamma$ or not. The runs have been binned in groups of $n_b = 100$, and each bin b has an empirical coverage rate $\hat{p}_b$ computed. The shaded area represent $\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n_b}}$. In general, the misspecification voids the guarantee for the coverage rate, but as long as the misspecification is small, the coverage rate is close to the promised one.

.

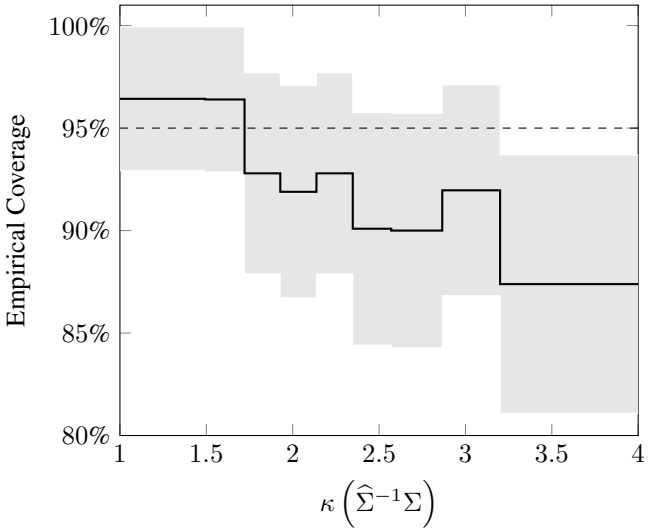


Figure 4: Setup as in Figure 3, but $n = 10000$.

References

Po-Ling Loh and Peter Bühlmann. High-Dimensional Learning of Linear Causal Networks via Inverse Covariance Estimation. *Journal of Machine Learning Research*, 15(88):3065–3105, 2014. URL http://jmlr.org/papers/v15/loh14a.html.

Ignavier Ng, Shengyu Zhu, Zhitang Chen, and Zhuangyan Fang. A Graph Autoencoder Approach to Causal Structure Learning, November 2019. arXiv: 1911.07420, presented at NeurIPS 2019 Workshop "Do the right thing".

Dennis Wei, Tian Gao, and Yue Yu. Dags with no fears: A closer look at continuous optimization for learning bayesian networks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS 2020)*, 2020.

Jeffrey M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT Press, Cambridge, Mass, 2nd ed edition, 2010. ISBN 978-0-262-23258-6.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7154–7163. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/yu19a.html.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32 proceedings (NeurIPS 2018)*, 2018.