Supplemental Notes for Causal Link Discovery with Unequal Edge Error Tolerance

Joni Shaska
University of Southern California
shaska@usc.edu

Urbashi Mitra
University of Southern California
ubli@usc.edu

***Abstract*—This paper proposes a novel framework for causal discovery with asymmetric error control, called *Neyman-Pearson causal discovery*. Despite the importance of applications where different types of edge errors may have different importance, current state-of-the-art causal discovery algorithms do not differentiate between the types of edge errors, nor provide any finite-sample guarantees on the edge errors. Hence, this framework seeks to minimize one type of error while keeping the other below a user-specified tolerance level. Using techniques from information theory, fundamental performance limits are found, characterized by the Rényi divergence, for Neyman-Pearson causal discovery. Furthermore, a causal discovery algorithm is introduced for the case of linear additive Gaussian noise models, called $\epsilon - CUT$, that provides finite-sample guarantees on the false positive rate, while staying competitive with state-of-the art methods.**

I. Generalized Likelihood Ratio Test (GLRT)

We briefly describe the modified implementation of the GLRT used for the numerical results. The algorithm is extensively described in [1], so we only briefly overview the method. Recall that we wish to compute the test statistic

$$\Lambda_{i,j} = 2(\sup_{\mathcal{G}\in\mathcal{M}} \ell(\mathcal{G}) - \sup_{\mathcal{G}\in\mathcal{M}^0_{i,j}} \ell(\mathcal{G})), \tag{1}$$

where $\mathcal{M}$ denotes the set of all DAGS and $\mathcal{M}^0_{i,j}$ the set of all DAGS with no edge between vertices i and j, and the likelihood $\ell(\mathcal{G})$ of a graph $\mathcal{G}$ with adjacency matrix $\boldsymbol{A}$ is given as

$$\ell(\mathcal{G}) = -\frac{np}{2}\log(2\pi\sigma^2) - \frac{n}{2\sigma^2}\text{Trace}\big((I-\boldsymbol{A})^\top (I-\boldsymbol{A})\hat{\Sigma}\big). \tag{2}$$

To find the unrestricted maximum likelihood estimate $\sup_{\mathcal{G}\in\mathcal{M}} \ell(\mathcal{G})$, we use the method described in [1]. First, we find the causal ordering using the method from [2]. Briefly, a *causal ordering* is an ordering of the nodes such that if for two nodes k and l, $k < l$ implies there exists a directed path from node k to node l. Upon finding the estimated causal ordering, for a given node l, we regress l on all nodes that precede it in the causal ordering, i.e., regress l on all nodes k such that $k < l$ in the causal ordering. This regression gives us the lth row of $\boldsymbol{A}$. Hence, regressing all nodes l on those that precede it in the causal ordering yields the matrix $\boldsymbol{A}$. Notice that this procedure yields a maximally connected DAG, i.e., a DAG with the maximum possible number of edges (though the weights for certain edges may be small). Indeed, the maximum likelihood estimate (with no sparsity constraints) can be found by considering only maximally connected DAGs (Proposition 3.2 in [1]), simplifying the process.

To find the restricted maximum likelihood estimate $\sup_{\mathcal{G}\in\mathcal{M}^0_{i,j}} \ell(\mathcal{G})$, we unfortunately need to test all possible causal orderings, of which there are $d!$ (although it is still sufficient to consider only complete DAGs). For each causal ordering, we once again regress a given node on those that precede it in the causal ordering while also enforcing the edge constraint on nodes i and j. That is, if we consider node j and have that $i < j$ in the causal ordering we regress j on all nodes that precede it in the causal ordering *except* i, similarly if we consider node i and $j < i$. This ensures that $\boldsymbol{A}_{i,j} = \boldsymbol{A}_{j,i} = 0$. Then, we select the causal ordering and corresponding matrix that achieves the maximum likelihood over all possible ordering. This allows us to compute the test statistic $\Lambda_{i,j}$.

A. Computational Complexity

To obtain the unrestricted maximum likelihood estimate, we perform $d-1$ linear regressions, each with complexity at most $\mathcal{O}(d^2(n+d))$. Then, we must also compute the likelihood given in (2). This can be done (naively) with complexity $\mathcal{O}(d^3)$ (optimizing matrix multiplication will not significantly reduce the computational complexity, as we will shortly see). Then, to compute the maximum likelihood estimate over $\mathcal{M}^0_{i,j}$, we consider only complete DAGS over all possible causal orderings, which implies we need to search over $d!$ possible graphs. For each graph, we compute $d-1$ linear regressions. Moreover, for each graph, we must compute the likelihood, and perform a simple threshold test. Hence, the computational complexity of the GLRT for a single edge pair is

$$\begin{aligned}&\mathcal{O}((d-1)(1+d!)d^2(n+d) + (d!+1)d^3)\\ &= \mathcal{O}((d-1)(1+d!)d^2(n+d)).\end{aligned} \tag{3}$$

This work has been funded in part by one or more of the following grants: ARO W911NF1910269, ARO W911NF2410094, DOE DE-SC0021417, Swedish Research Council 2018-04359, NSF CCF-2008927, NSF RINGS-2148313, NSF CCF-2200221, NSF CCF-2311653, ONR 503400-78050, ONR N00014-22-1-2363, and NSF A22-2666-S003.

Portions of this paper were presented at the 2024 International Symposium on Information Theory (ISIT) and the 2024 Asilomar Conference on Signals, Systems, and Computers.

Parameter values for Fig. 2				
Graph Paramters	NOTEARS (λ)	LASSO (λ)	GLRT (τ)	Optimal (τ)
$d = 4$, $a = 1$	$\{5.375, 4.789, 4.203, 3.616, 3.03, 2.444, 1.858, 1.271, 0.685, 0.099\}$	$\{9.5, 8.754, 8.009, 7.264, 6.518, 5.773, 5.027, 4.282, 3.536, 2.791, 2.045, 1.3\}$	$\{1.5, 2.722, 3.944, 5.167, 6.389, 7.611, 8.833, 10.056, 11.278, 12.5\}$	$\{0.01, 0.231, 0.452, 0.673, 0.894, 1.116, 1.337, 1.558, 1.779, 2\}$

TABLE I

Parameter values for Fig. 3					
Graph Paramters	NOTEARS (λ)	DAGMA (λ)	LASSO (λ)	GLRT (τ)	$\epsilon - CUT$ (τ)
$d = 5$	$\{1000, 492, 243, 119, 59, 29, 14, 7, 3, 2, .8, .4, .2, .1, .05\}$	$\{80000, 20000, 8000, 2000, 800, 400, 100, 50, 25, 17, 10, 5, 3, 2, 1, .5\}$	$\{4000, 2211.92, 1223.15, 676.37, 374.02, 206.83, 114.37, 63.25, 34.97, 19.34, 10.69, 5.91, 3.27, 1.80, 1\}$	$\{1, 2.6, 4.1, 5.7, 7.2, 8.8, 10.3, 11.9, 13.4, 15\}$	$\{21, 19.5, 18.1, 16.6, 15.2, 13.7, 12.3, 10.8, 9.3, 7.9, 6.4, 4.9, 3.5, 2.1, 0.6\}$
$d = 7$	$\{12000, 10000, 8000, 400, 200, 100, 50, 25, 17, 12, 7, 5, 3, 2, 1\}$	$\{400000, 200000, 100000, 10000, 1000, 500, 100, 50, 25, 17, 10\ 7, 5, 3, 2, 1, .5\}$	$\{60000, 32231.7, 17314.72, 9301.4, 4996.66, 2684.18, 1441.93, 774.6, 416.1, 223.5, 120.1, 64.5, 34.7, 18.6, 10\}$	NA	$\{10, 7.2, 5.2, 3.7, 2.7, 1.9, 1.4, 1, 0.72, 0.52, 0.37, 0.27, 0.19, 0.14, 0.1\}$

TABLE II

Since we have $\frac{d(d-1)}{2}$ possible edge pairs, the total computational complexity of the GLRT is

$$\mathcal{O}((1 + d!)d^3(d-1)^2(n+d)). \tag{4}$$

II. Computational Complexity of Linear Regression

We briefly review the computational complexity of ordinary least squares (OLS). We analyze the naive approach. That is, suppose we have the system model

$$y = A^\top X + W, \tag{5}$$

where y is a scalar, A and X are $d \times 1$ vectors, and W is a $d \times 1$ noise vector. Then, if we have n measurements of y and X, where the kth measurement is denoted as y_k, and X_k, and define the vector $Y = [y_1, ..., y_n]^\top$ and the matrix $\boldsymbol{X} = [X_1, ..., X_n]$, the OLS estimate is given by

$$\hat{A} = (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}Y. \tag{6}$$

we understand that there are more efficient OLS methods than the naive approach (we use numpy.linalg.lstsq to implement OLS). Still, for analysis and to obtain a benchmark for the computational complexity of $\epsilon - CUT$ and the GLRT, we analyze the naive approach. Then, transposing $\boldsymbol{X}$ takes $\mathcal{O}(nd)$ time. $\boldsymbol{X}\boldsymbol{X}^\top$ takes $\mathcal{O}(d^2n)$ and produces a $d \times d$ matrix. Then, inverting a $d \times d$ matrix takes $\mathcal{O}(d^3)$. Multiplying the inverted matrix by $\boldsymbol{X}$ takes $\mathcal{O}(d^2n)$. The final matrix multiplication with Y takes $\mathcal{O}(dn)$. Hence, the final complexity is $\mathcal{O}(nd + d^2n + d^3 + d^2n + dn) = \mathcal{O}(d^2(d+n))$.

III. Numerical Algorithms

We give the regularizer values for NOTEARS and LASSO, and the threshold values for the GLRT and the optimal detector used to generate Fig. 2 in Table I. We give the regularizer values used for NOTEARS, DAGMA, and LASSO, in addition to the threshold values used by $\epsilon - CUT$ and the GLRT to generate the plots in Fig. 3 in Table II.

References

[1] D. Strieder and M. Drton, “Confidence in causal inference under structure uncertainty in linear causal models with equal variances,” *Journal of Causal Inference*, vol. 11, no. 1, p. 20230030, 2023. [Online]. Available: https://doi.org/10.1515/jci-2023-0030

[2] W. Chen, M. Drton, and Y. S. Wang, “On causal discovery with an equal-variance assumption,” *Biometrika*, vol. 106, no. 4, p. 973–980, Sep. 2019. [Online]. Available: http://dx.doi.org/10.1093/biomet/asz049