Exploration of Losses in NOTEARS

Mengqi Liu

Aug 30, 2023

1 Notations

NOTEARS[1] solves the problem of structure learning for directed acyclic graph(DAG) under the assumption that data is generated by generalized linear SEM. Consider a vector of random variables of our interest $X = (X_1, X_2, ..., X_d)$. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be the data matrix, where (i, j)-th entry X_{ij} represents the i-th i.i.d observation of the j-th variable X_j.

Under the assumption that the data fits the generalized linear SEM, we could model X through a weighted adjacency matrix $W = (w_1, w_2, \cdots, w_d) \in \mathbb{R}^{d\times d}$ by $\mathbb{E}(X_j) = f(w_j^\top X)$ for $j = 1, 2, ..., d$.

Note $A(W) \in \{0, 1\}^{d\times d}$ as the adjacency matrix with weights in W removed. Specifically, let $[A(W)]_{ij}$ be 0 if $W_{ij} = 0$ and 1 otherwise. Let $G(W)$ denote the graph corresponding to adjacency matrix $A(W)$. Also let $\mathbb{D}$ denote the discrete space of DAGs and $G = (V, E)$ denote a graph with vertices V and edges E. Define the number of edges $s0 = |E|$. Our causal discovery task is to learn a DAG $G \in \mathbb{D}$ given $\mathbf{X}$. Under settings of Bayesian networks, we could model X in this way: $\mathbb{E}(X_j|\text{pa}(X_j)) = f(w_j^\top X)$ for $j = 1, 2, ..., d$. And notice $W_{ij} = 0$ for any $i \in \{1, 2, ..., d\}\backslash\text{pa}(X_j)$.

2 Review on losses in NOTEARS

2.1 Continuous linear model

For Gaussian linear SEM, $X_j = w_j^\top X + \epsilon_j$ for $j = 1, 2, ..., d$. Here we do not restrict ϵ_j's to Gaussian noises.

In the paper, Xun Zheng et al. use L2 loss to score DAGs under linear SEM. The benefits are as follows: (1) The minimizer of the L2 loss has been shown to recover a true DAG with high probability on finite-samples and in high-dimensions. (2) The estimator through the L2 loss is consistent for both Gaussian SEM and non-Gaussian SEM. (3) No faithfulness assumption needed in the setting of linear SEM and L2 loss. The L2 loss under linear SEM is

$$\ell(W; \mathbf{X}) = \frac{1}{2n}\|\mathbf{X} - \mathbf{X}W\|_F^2. \tag{1}$$

Due to the interest in learning sparse DAG, authors introduced a $\ell 1$-regularization $\|W\|_1 = \|\text{vec}(W\|_1)$ in the regularized score function:

$$F(W) = \ell(W; \mathbf{X}) + \lambda\|W\|_1 = \frac{1}{2n}\|\mathbf{X} - \mathbf{X}W\|_F^2 + \lambda\|W\|_1,\ \lambda \geq 0. \tag{2}$$

They put forward a smooth function $h(W)$ to characterize the topological property of $G(W)$ and have proven that $h(W) = 0$ if and only if W is acyclic (i.e. $G(W) \in \mathbb{D}$).

$$h(W) = \text{tr}(e^{W\circ W}) - d = 0 \tag{3}$$

Then the causal discovery question becomes a equality-constrained program(ECP) as

$$\begin{aligned} &\min_{W\in\mathbb{R}^{d\times d}} F(W) \\ &\text{subject to } h(W)=0. \end{aligned} \tag{4}$$

In order to solve this ECP, they take advantage of augmented Lagrangian method to solve the original problem by a approximate program:

$$\begin{aligned} &\min_{W\in\mathbb{R}^{d\times d}} O(W,\rho) \\ &\text{subject to } h(W)=0, \end{aligned} \tag{5}$$

where $\rho > 0$ and

$$\begin{aligned} O(W,\rho) &= F(W) + \frac{\rho}{2}|h(W)|^2 \\ &= \frac{1}{2n}\|\mathbf{X}-\mathbf{X}W\|_F^2 + \lambda\|W\|_1 + \frac{\rho}{2}[\mathrm{tr}(e^{W\circ W})-d]^2. \end{aligned} \tag{6}$$

And the dual problem is

$$\max_{\alpha\in\mathbb{R}} \min_{W\in\mathbb{R}^{d\times d}} D(W,\rho,\alpha) \tag{7}$$

where

$$\begin{aligned} D(W,\rho,\alpha) &= O(W,\rho) + \alpha h(W) \\ &= \frac{1}{2n}\|\mathbf{X}-\mathbf{X}W\|_F^2 + \lambda\|W\|_1 + \frac{\rho}{2}[\mathrm{tr}(e^{W\circ W})-d]^2 + \alpha[\mathrm{tr}(e^{W\circ W})-d]. \end{aligned} \tag{8}$$

And the problem was numerically solved by dual ascent method (iteratively update α and ρ) and L-BFGS ($\lambda = 0$) / PQN ($\lambda > 0$, find approximation of descent direction for non-smooth function).

2.2 Discrete logistic model

For logistic SEM, $X_j \sim \text{Bernoulli}(\sigma(w_j^\top X))$ for $j = 1, 2, ..., d$, where $\sigma(t) = \frac{e^t}{1+e^x}$ is sigmoid function. Accordingly, the logistic loss is

$$\begin{aligned} \ell(W;\mathbf{X}) &= -\frac{1}{n}\mathbb{1}_n^\top[\mathbf{X}\odot\log(\sigma(\mathbf{X}W) + (\mathbb{1}_n\mathbb{1}_n^\top - \mathbf{X})\odot\log(\mathbb{1}_n\mathbb{1}_n^\top - \sigma(\mathbf{X}W))]\mathbb{1}_n \\ &= \frac{1}{n}\mathbb{1}_n^\top[\log(\mathbb{1}_n\mathbb{1}_n^\top + e^{\mathbf{X}W}) - \mathbf{X}\odot\mathbf{X}W]\mathbb{1}_n, \end{aligned} \tag{9}$$

where $\mathbb{1}_n$ is a n-length vector with all elements of 1 and $\odot$ stands for Hadamard product (a.k.a. element-wise product).

For poisson SEM, $X_j \sim \text{poisson}(e^{w_j^\top X})$ for $j = 1, 2, ..., d$ the empirical poisson loss is

$$\ell(W;\mathbf{X}) = \frac{1}{n}\mathbb{1}_n^\top[e^{\mathbf{X}W} - \mathbf{X}\odot\mathbf{X}W]\mathbb{1}_n. \tag{10}$$

Similarly, we will have

$$\begin{aligned} F(W) &= \ell(W;\mathbf{X}) + \lambda\|W\|_1,\ \lambda \geq 0 \\ O(W,\rho) &= F(W) + \frac{\rho}{2}|h(W)|^2 \\ D(W,\rho,\alpha) &= O(W,\rho) + \alpha h(W) \end{aligned} \tag{11}$$

and thus we could solve the problem the same way as in continuous case.

3 Adversarial relationship between $\ell(W;\mathbf{X})$ and $h(W)$

Notice two facts:

(1) $W = I_d$ is always a solution of $\ell(W;\mathbf{X}) = 0$. But acyclicity requires that diagonal of W should be 0.

(2) If there's deterministic relationship between $\tilde{x}_i$ and $\tilde{x}_j$, W_{ij} and W_{ji} should play the same role in $\ell(W;\mathbf{X})$ while acyclicity needs at least one of the two to be 0.

Therefore intuitively minimizing $L^\rho(W,\alpha)$ will lead to a trade-off between $\ell(W;\mathbf{X})$ and $h(W)$.

4 Theoretical and empirical global minimum for L2 loss

4.1 Propositions

Lemma 4.1 *Assume W is lower triangular with zero diagonal w.o.l.g. (i.e. variables have been arranged according to their topological order),*

$$\min_{G(W)\in DAG} \ell(W;\mathbf{X}) = \min_{G(W)\in DAG} \frac{1}{2n} \sum_{j=\max\{d-n+1,1\}}^{d} \|\mathbf{X}_j - \mathbf{X}_j w_j\|_2^2.$$

Proof: When $d \le n$, $\max\{d-n+1,1\} = 1$ and the statement holds naturally. When $d > n$, for any $j \in \{1,2,...,d-n\}$ (here we choose $j > 1$ w.l.o.g.), consider the inhomogeneous(probably) linear system with W as unknown:

$$\begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1,j-1} & X_{1,j+1} & \cdots & X_{1d} \\ X_{21} & X_{22} & \cdots & X_{2,j-1} & X_{2,j+1} & \cdots & X_{2d} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{n,j-1} & X_{n,j+1} & \cdots & X_{nd} \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ W_{j+1,j} \\ \vdots \\ W_{dj} \end{pmatrix} = \begin{pmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{nj} \end{pmatrix}.$$

The linear system can be truncated to

$$\begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1,d-j-1} \\ X_{21} & X_{22} & \cdots & X_{2,d-j-1} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{n,d-j-1} \end{pmatrix} \begin{pmatrix} W_{j+1,d} \\ W_{j+2,d} \\ \vdots \\ W_{d,j} \end{pmatrix} = \begin{pmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{nj} \end{pmatrix},$$

where $d-(j+1)+1 \ge n$ and $W_{j+1,j}, W_{j+2,j}, ..., W_{d,j}$ can take any value in $\mathbb{R}$.

Because rows of $\mathbf{X}$ stand for n i.i.d observations, we know that $\text{rank}(\mathbf{X}) = n$ and the corresponding augmented matrix also should have rank of n. According to Rouché–Capelli theorem, as the rank of its coefficient matrix $\mathbf{X}$ is equal to the rank of its augmented matrix, this system has at least one solution. Therefore $\min_{G(W)\in\text{DAG}} \frac{1}{2n}\|\mathbf{X}_j - \mathbf{X}_j w_j\|_2^2 = 0$ for all $1 \le j \le d-n$.

As a result,

$$\begin{aligned}
\min_{G(W)\in\text{DAG}} \ell(W;\mathbf{X}) &= \min_{G(W)\in\text{DAG}} \frac{1}{2n}\|\mathbf{X}-\mathbf{X}W\|_F^2 \\
&= \min_{G(W)\in\text{DAG}} \frac{1}{2n}\sum_{j=1}^{d}\|\mathbf{X}_j-\mathbf{X}_j w_j\|_2^2 \\
&= \frac{1}{2n}\sum_{j=1}^{d}\min_{w_j}\|\mathbf{X}_j-\mathbf{X}_j w_j\|_2^2 \\
&= \min_{G(W)\in\text{DAG}} \frac{1}{2n}\sum_{j=\max\{d-n+1,1\}}^{d}\|\mathbf{X}_j-\mathbf{X}_j w_j\|_2^2.
\end{aligned}$$

□

Lemma 4.2 *Let π denote a permutation and P is the corresponding permutation matrix. Theoretically,*

$$\min_{G(W)\in DAG}\mathbb{E}\|X-W^\top X\|_2^2 \geq \min_{P}\min_{G(W)\in DAG}\mathbb{E}\|PX-W^\top PX\|_2^2.$$

Empirically,

$$\begin{aligned}
\min_{G(W)\in DAG}\ell(W;\mathbf{X}) &= \min_{G(W)\in DAG}\frac{1}{2n}\sum_{j=1}^{d}\|\mathbf{X}_j-\mathbf{X}_j w_j\|_2^2 \\
&\geq \min_{P}\min_{G(W)\in DAG}\frac{1}{2n}\sum_{j=1}^{d}\|\mathbf{X}_j P^\top-\mathbf{X}P^\top w_j\|_2^2 \\
&= \min_{P}\min_{G(W)\in DAG}\ell(W;\mathbf{X}P^\top).
\end{aligned}$$

Remark. *Notice that* $\arg\min_{G(W)\in DAG}\mathbb{E}\|PX-W^\top PX\|_2^2 \neq \arg\min_{G(W)\in DAG}\mathbb{E}\|X-W^\top X\|_2^2$ *in general. Consider the linear regression $X=W^\top X+\epsilon$, where $X\in\mathbb{R}^d$, $W\in\mathbb{R}^{d\times d}$, $\epsilon\in\mathbb{R}^d$, $cov(X)=\Sigma$ and $cov(\epsilon)=D$. Assume W is lower triangular, we could conduct Cholesky decomposition as*

$$\Sigma=(I-W^\top)^{-1}D(I-W^\top)^{-1}=[(I-W^\top)^{-1}D^{\frac{1}{2}}][D^{\frac{1}{2}}(I-W^\top)^{-1}]=L(X)L(X)^\top.$$

After permutation, the decomposition becomes

$$P\Sigma P^\top=PL(X)L(X)^\top P^\top=L(PX)L(PX)^\top,$$

where $L(PX)$ corresponds to new $\tilde{W}$ with respect to $PX=\tilde{W}^\top PX+P\epsilon$. Generally, $\tilde{W}=I-\{L(PX)[PDP^\top]^{-\frac{1}{2}}\}^{-1}$ is different from original $W=I-\{L(X)D^{-\frac{1}{2}}\}^{-1}$.

4.2 Empirical results

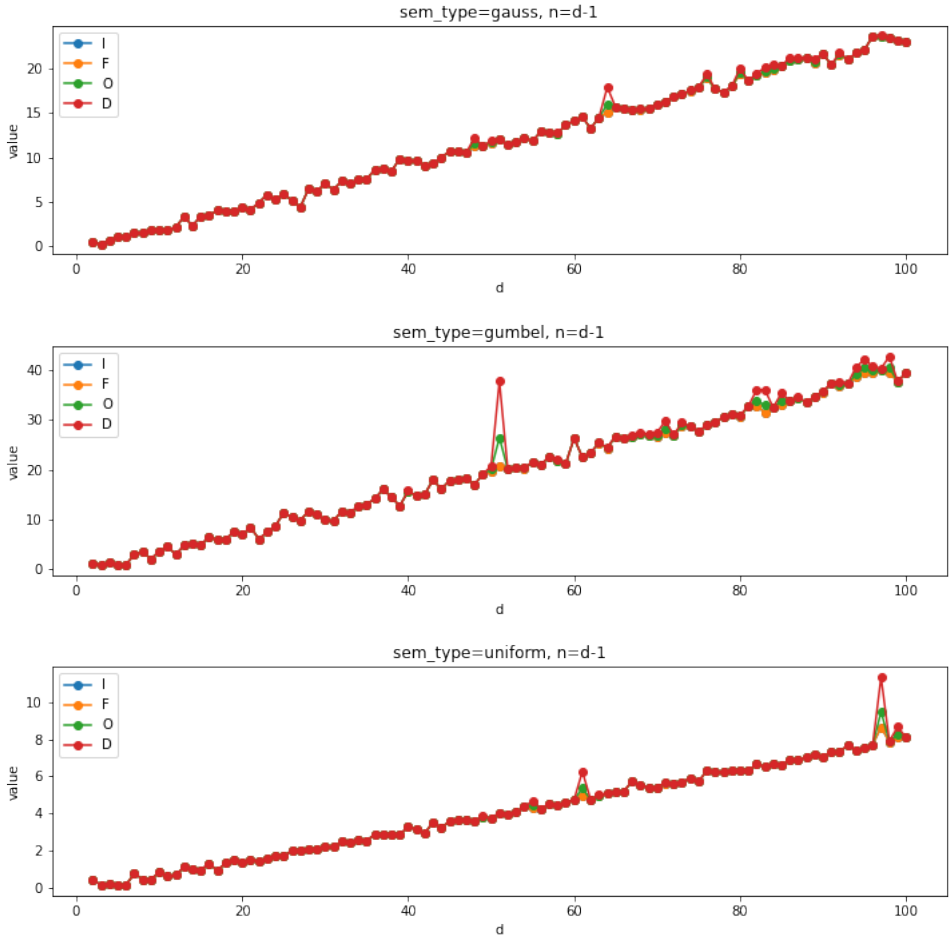

Figure 1: $d - n = 1$

To check the performance of empirical L2 loss, we only train DAGs with data generated with noises of continuous distributions (Gaussian, Uniform, and Gumbel in total) and set $\lambda = 0$.

First, we let $d - n = 1$, $s0 = d - 1$ and train with $d \in \{2, 3, ..., 100\}$ respectively. The empirical losses are shown in Figure 1. From the results, we could find an increasing trend in L2 loss with increasing d(or n). This is reasonable because larger d will introduce more noise and and the right side of Lemma 4.1 will increase accordingly.

Second, zooming in and observing losses when $d \leq 20$ in Figure 2, where different lines refer to repeated experiments with different random seeds. random fluctuations when d is relatively small are not catastrophic to subvert the trend for now.

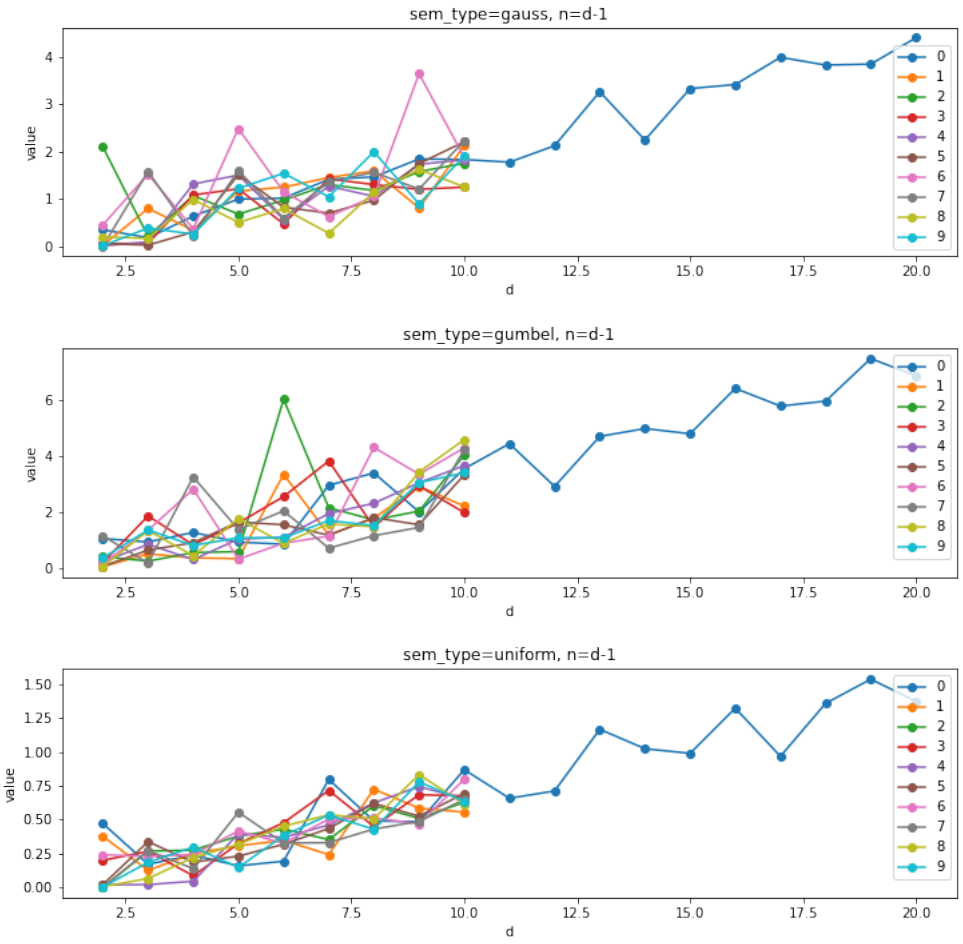

Figure 2: $d \leq 20$

Third, we try to permute the columns of data matrix $\mathbf{X}$ to construct a lower bound of empirical L2 loss and to verify if NOTEARS has attained the empirical global minimum. The losses optimized by L-BFGS are showed in Figure 3. We could find that all empirical minimized L2 losses are almost the same (although they're not exactly the same after zooming in). But it is worth noting that the gap between permutations become larger when d increases. Similarly, we could find the learned losses are very close with permutations in Figure 4. It is worth noting that for SGD, I set the learning rate to be $\frac{1}{\rho}$ avoid infinite loss that occurred in the previous experiments. Also I only test for Gaussian distributions and fewer d's in SGD due to the time-consuming learning process.

This can be showed as an initial trial to verify that NOTEARS has performed good optimization with respect to L2 loss through L-BFGS as it originally put forward.

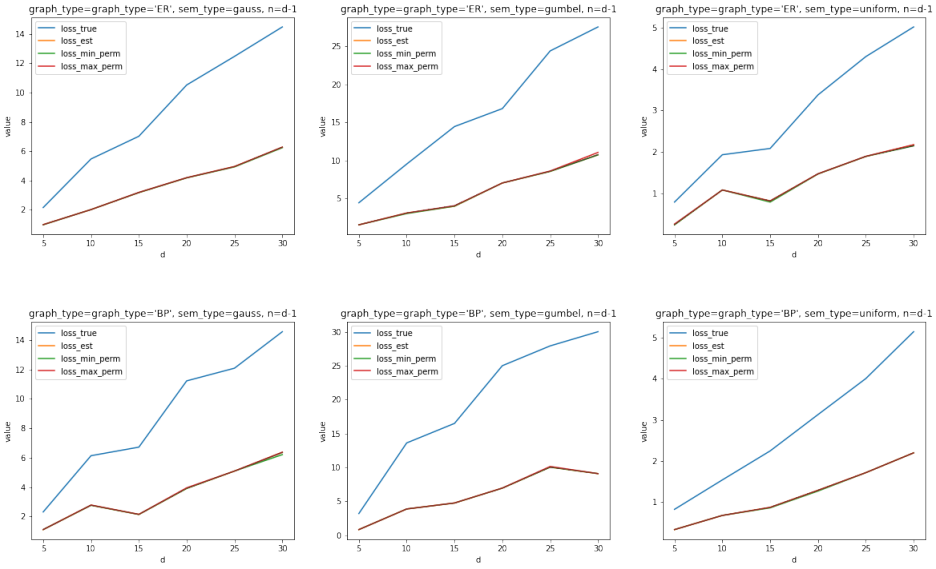

Figure 3: L-BFGS

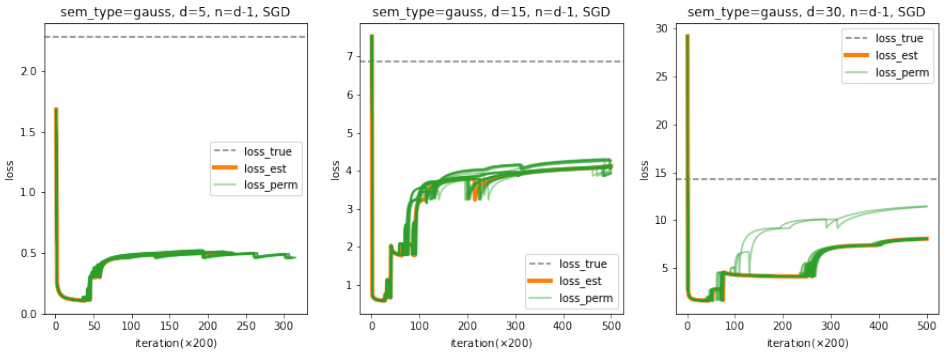

Figure 4: SGD

4.2.1 Inability to distinguish between graphs with Markov equivalence

The nature question is that why NOTEARS is quite robust under permutations. Hence we use a three-node template to test if the algorithm could identify the actual edge direction under permutations and figure out how do the losses perform correspondingly. From Figure 5 and Figure 6, we could find that the algorithm does fail to identify the true direction sometimes but always could identify the true positives up to Markov equivalence. More specifically, always four of six permuted results will identify the structure up to Markov equivalence with respect to the true adjacency matrix. This pattern appears across different methods (NOTEARS, NOTEARS combined with SGLD, DAGMA & DAGMA combined with SGLD) as well as a variety of settings including different continuous distributions (including uniform, gumbel, exponential distributions) and under constant and non-constant variance. Notably Y. S. Wang, etc.[2] proposed a method to identify the exact structure with non-Gaussian noise. However, regardless of accuracy of structure learning, Figure 9 shows that L2 losses are optimized to the similarly same level under permutations.

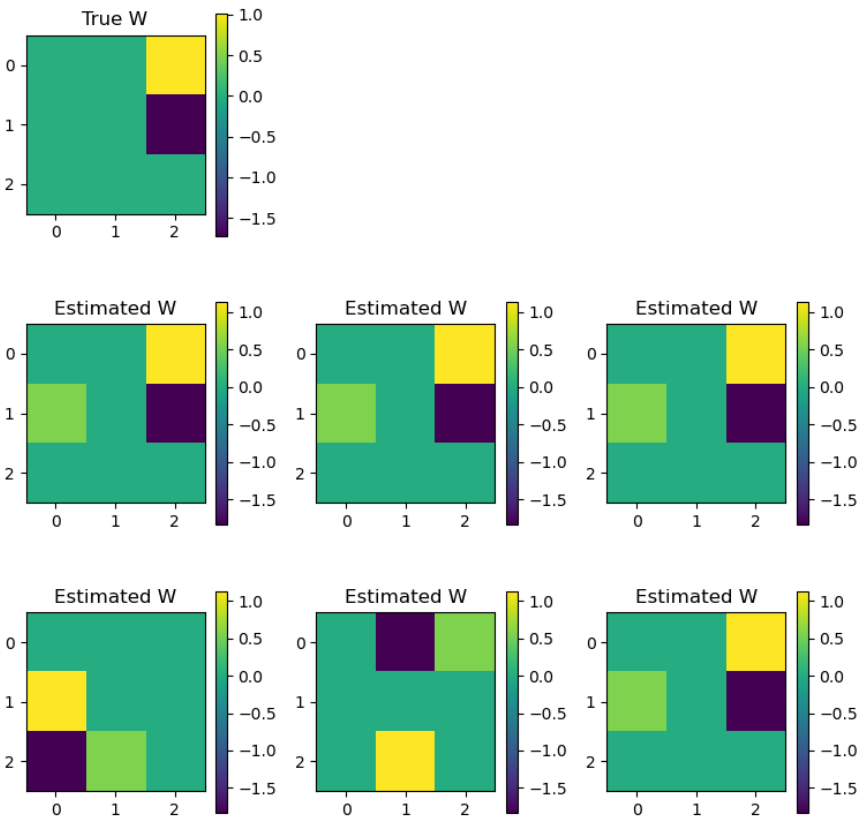

Figure 5: 3-node template

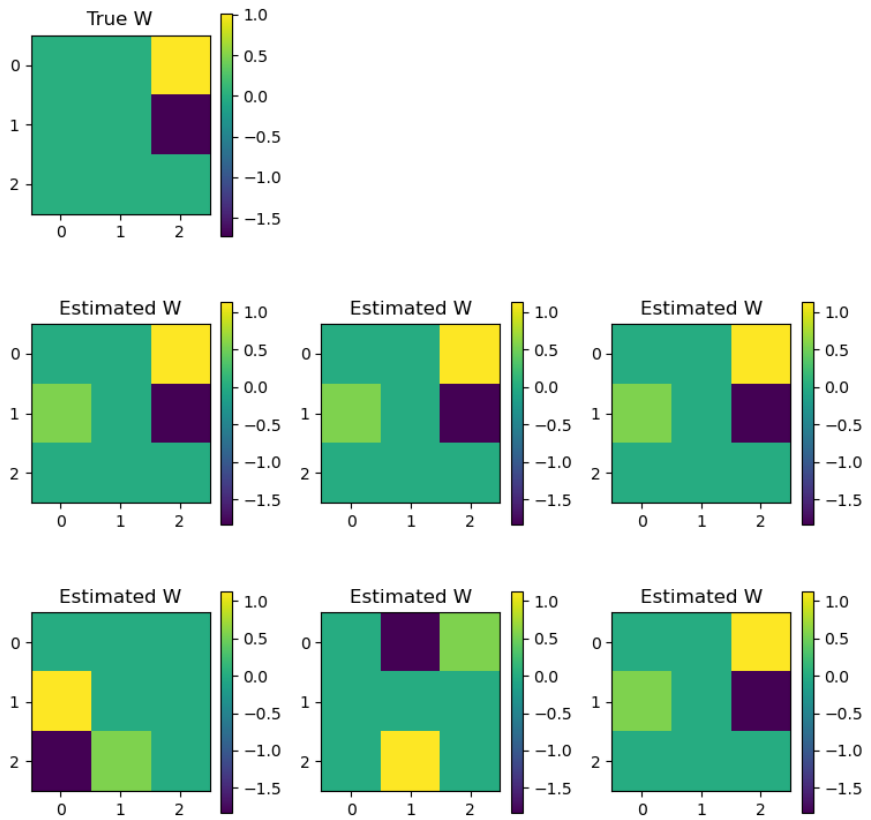

Figure 6: 3-node template

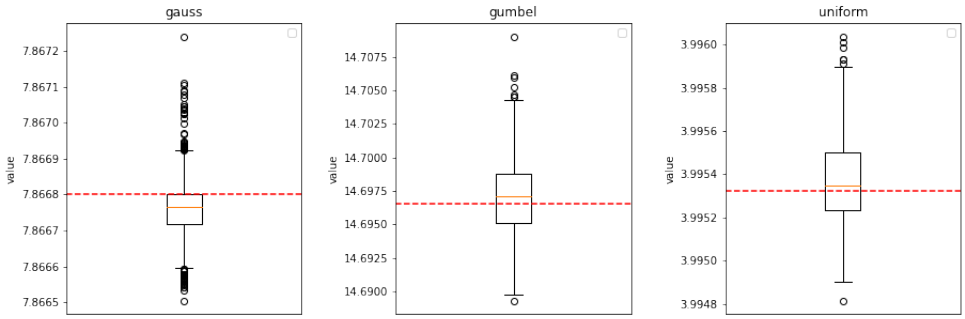

Figure 7: 3-node template repeated for 4 times with noise with variance of Gaussian distributed size (uniform one behaves similarly)

Actually, this phenomenon does not come forth only when we use some well-designed examples, the algorithm fails to distinguish between structures with Markov equivalence even for constant variance and normal ER graphs as shown in Figure 8. The

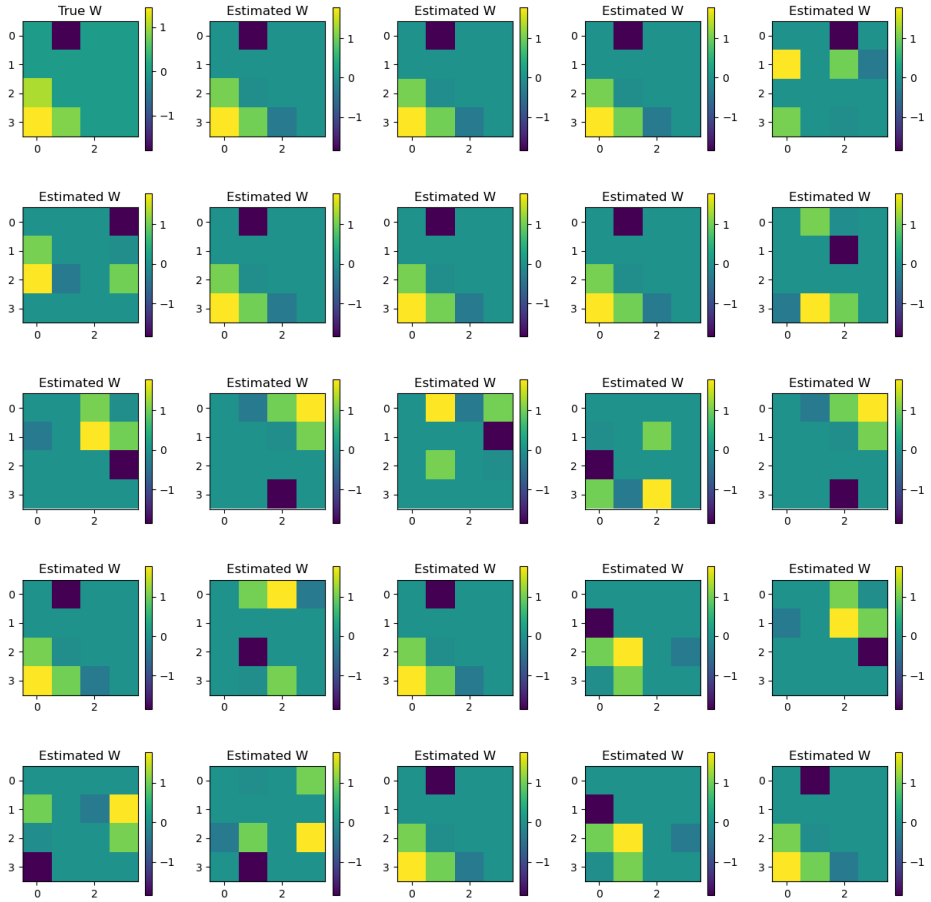

Figure 8: 4-node ER graph

As a supplement, DAGMA performs the same way in terms of failure in orientation identification in Figure 9 and 10. in Also it attains similar L2 loss after optimization under permutations as shown in Figure 11.

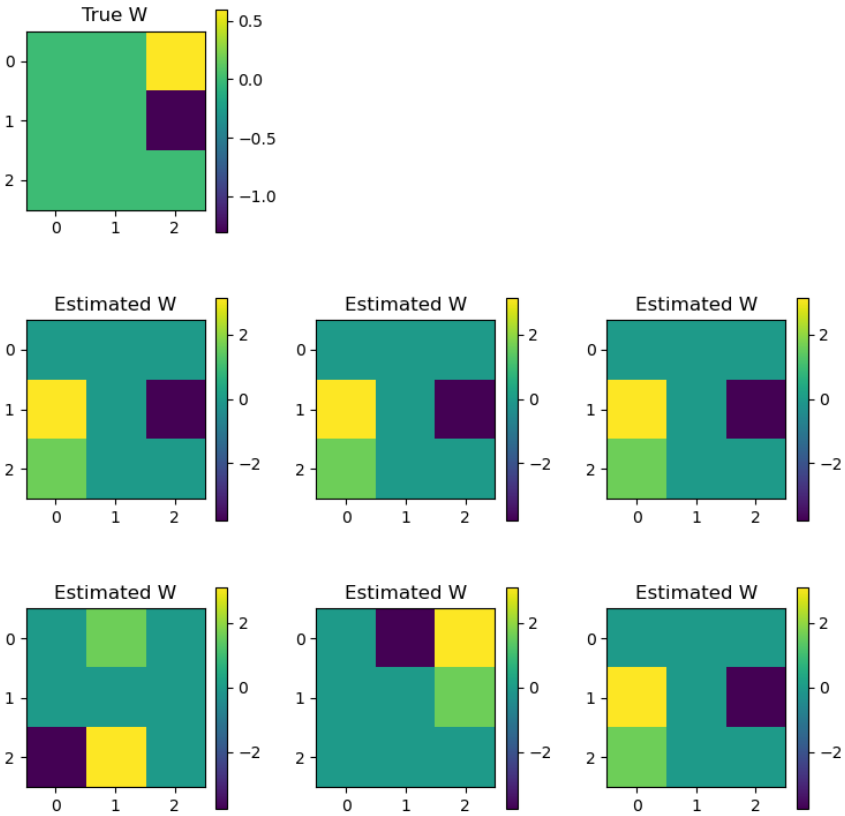

Figure 9: 3-node template by DAGMA

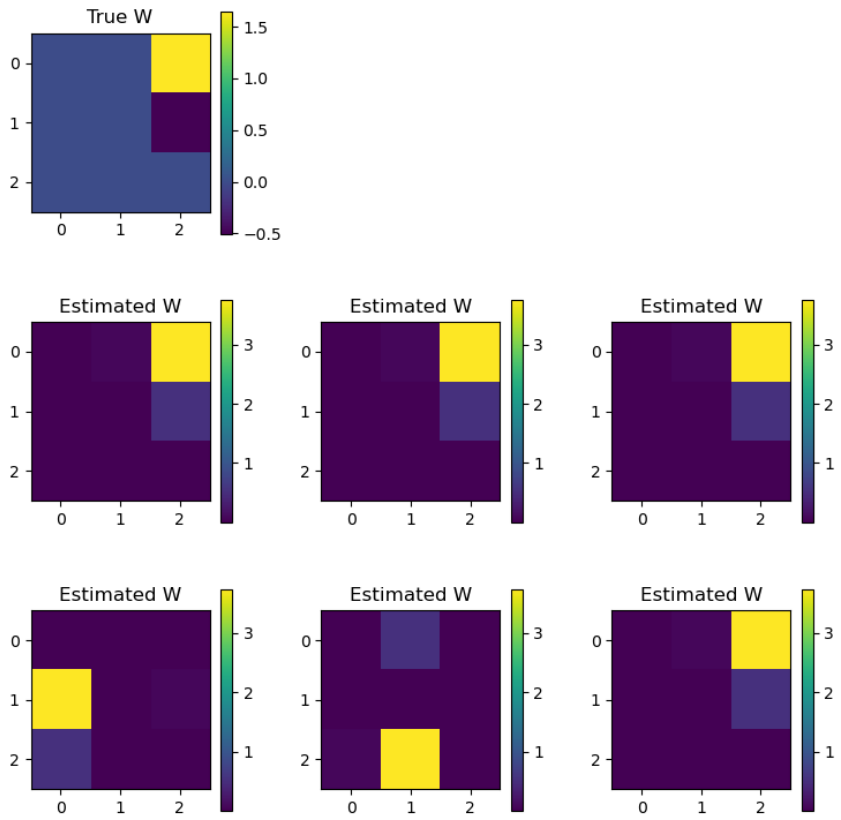

Figure 10: 3-node template by DAGMA

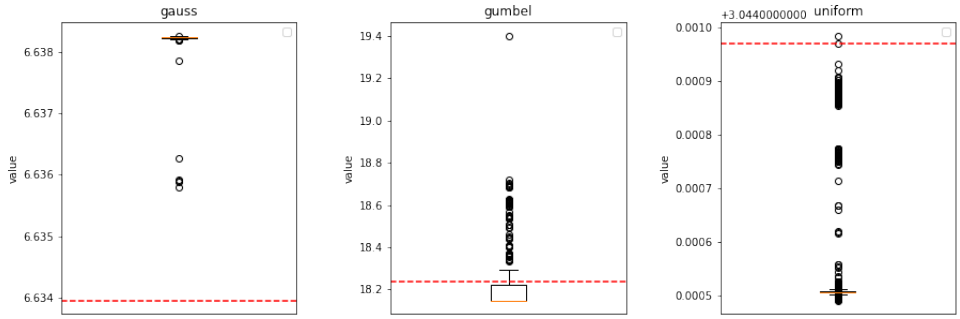

Figure 11: 3-node template repeated for 4 times with noise with variance of Gaussian distributed size (uniform one behaves similarly) by DAGMA

4.2.2 Non-constant variance

Then I also explored if NOTEARS is robust towards data generated with non-constant variance and found that NOTEARS performs not so robust on data generated by noise with variance of Gaussian distributed size. The results are shown in Figure 12. Notably, previous

papers have talked about this problem and Ignavier Ng, etc.[3] suggested using log-likelihood to avoid rescaling problem and proposed GOLEM. However, we find that GOLEM performs even worse than NOTEARS in terms of robustness under settings of heterogeneous variance from Figure 13.

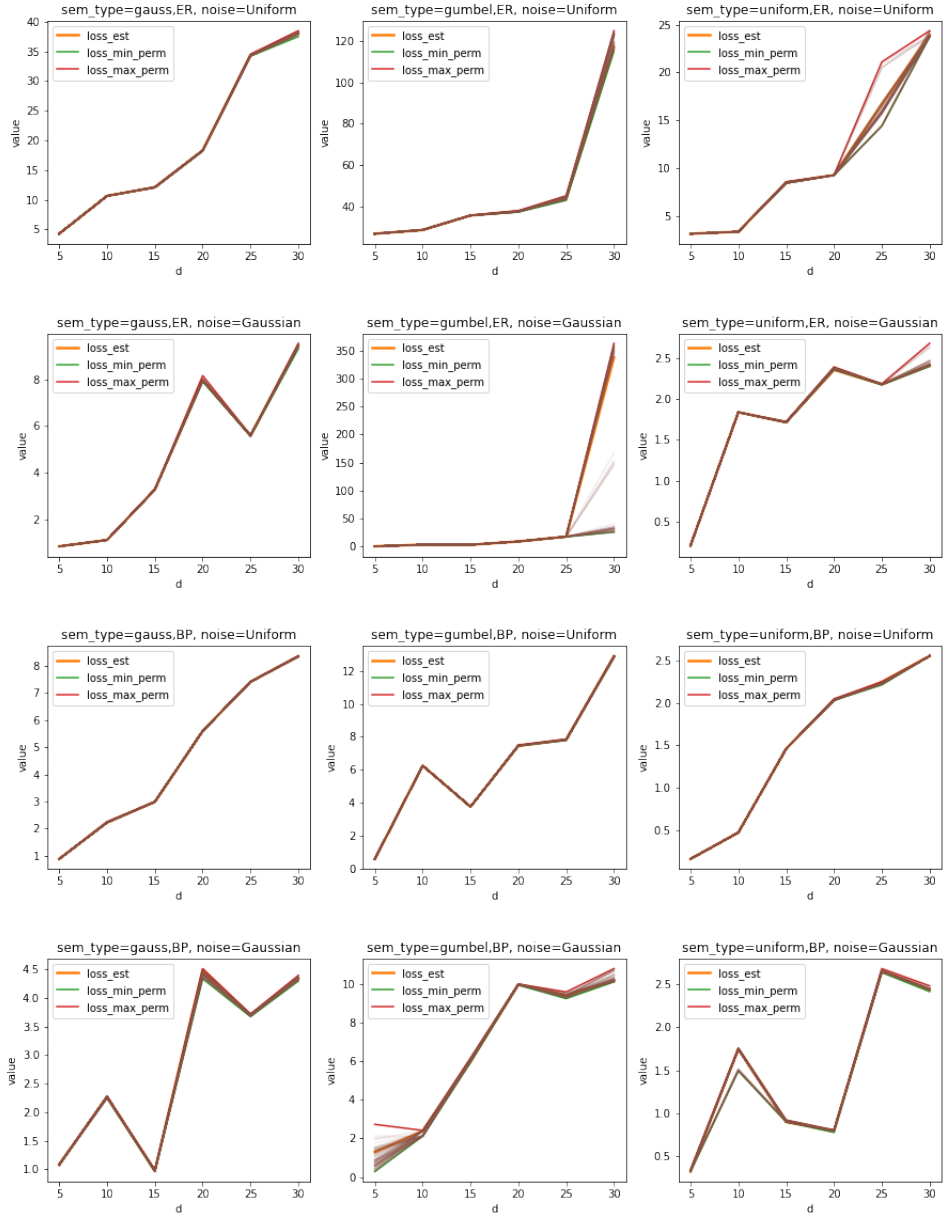

Figure 12: NOTEARS with non-constant variance

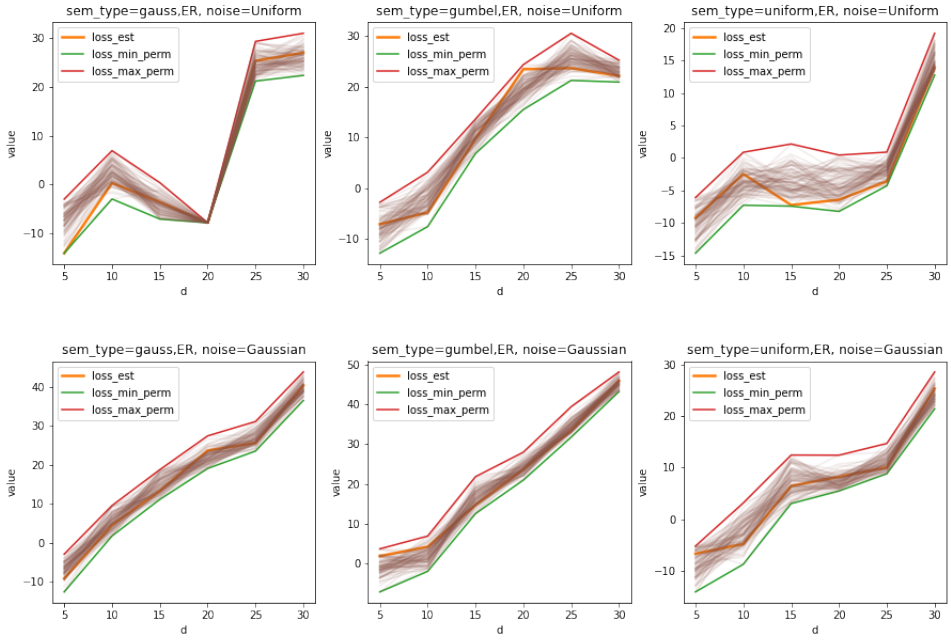

Figure 13: GOLEM with non-constant variance

Meanwhile, DAGMA also shows signs of instability when variance is not constant in Figure 14.

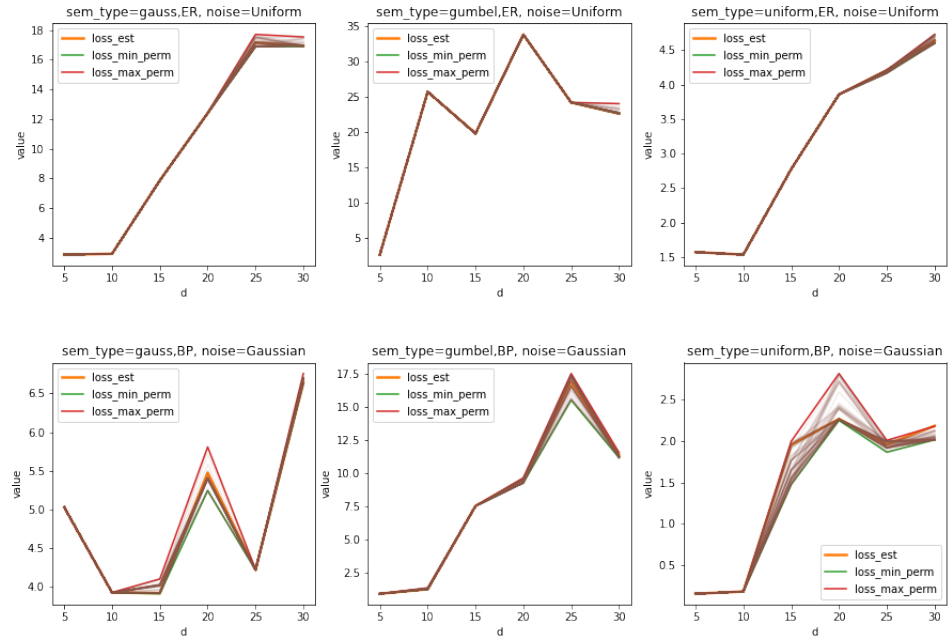

Figure 14: DAGMA with non-constant variance

4.2.3 SGLD

From the above results, we can preliminarily infer that neither NOTEARS, DAGMA nor GOLEM has reached the global minimum loss especially under non-constant Gaussian noises. So here we make attempts to introduce some randomness to the gradient in the process of minimizing loss, thereby helping them escape the local minimum trap. This method refers to Stochastic Gradient Langevin Dynamics (SGLD)[4]. This time we also plot the difference between the estimated and the true adjacency matrices, which is measured by Frobenius distance. The experiments results for NOTEARS and DAGMA are shown in the Figure 15 - Figure 19. It can be seen that whether for a relatively stable or volatile setting for permutations, the Frobenius distance between the estimated and the true adjacency matrices has not reached the global minimum (i.e. it is not relatively robust to permutation).

For NOTEARS, we add some Gaussian noise into gradients in Adam to optimize the loss. Figure 15 and Figure 16 show results under constant and non-constant Gaussian variance respectively. From two figures, we could discover that added noise introduces greater uncertainty into the estimated losses, sometimes making it extremely unstable and somewhat reducing the lower bound of minimized L2 loss. Surprisingly, SGLD greatly helps diminish the Frobenious distance between the estimated adjacency matrix and the true one.

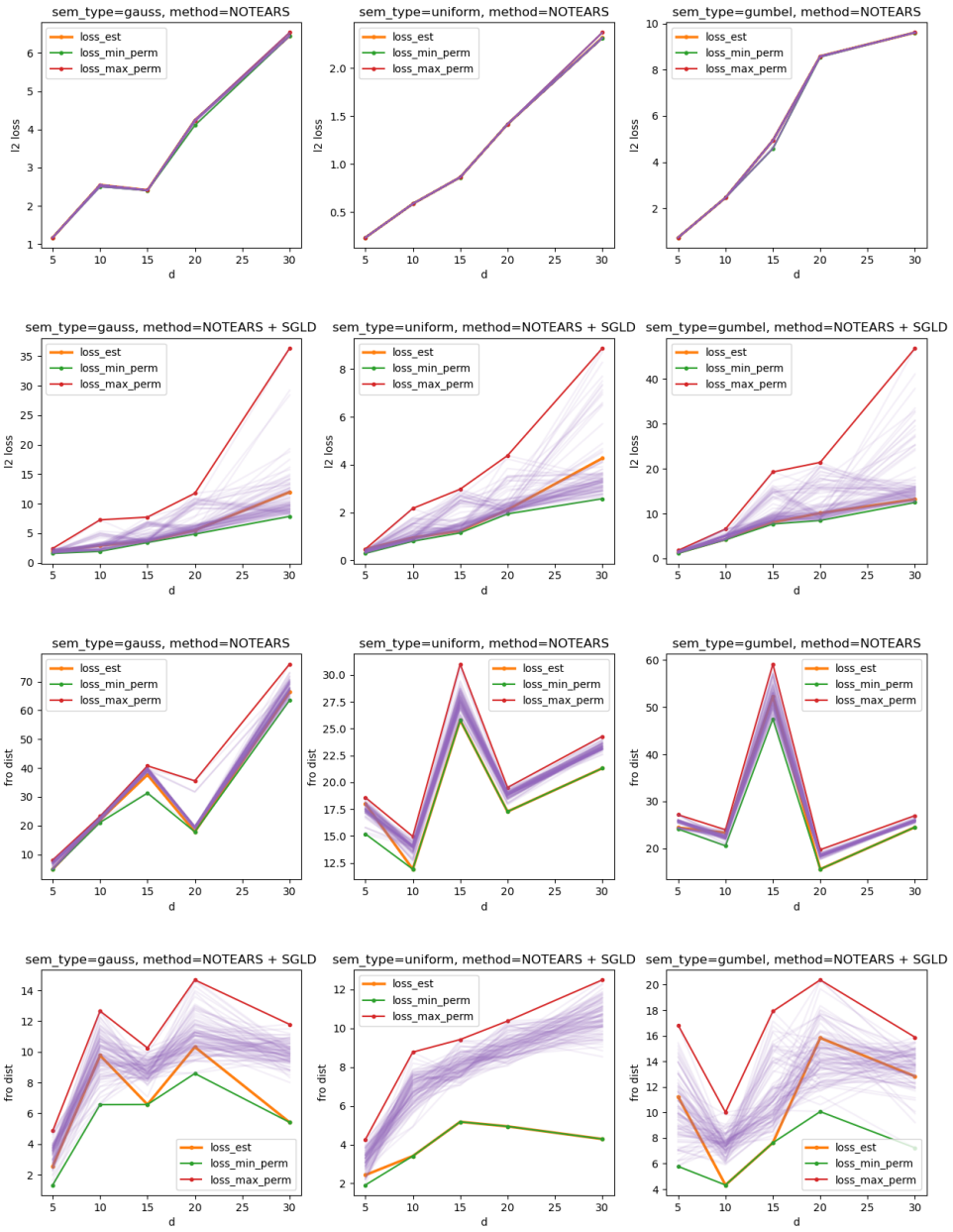

Figure 15: NOTEARS combined with SGLD under constant variance

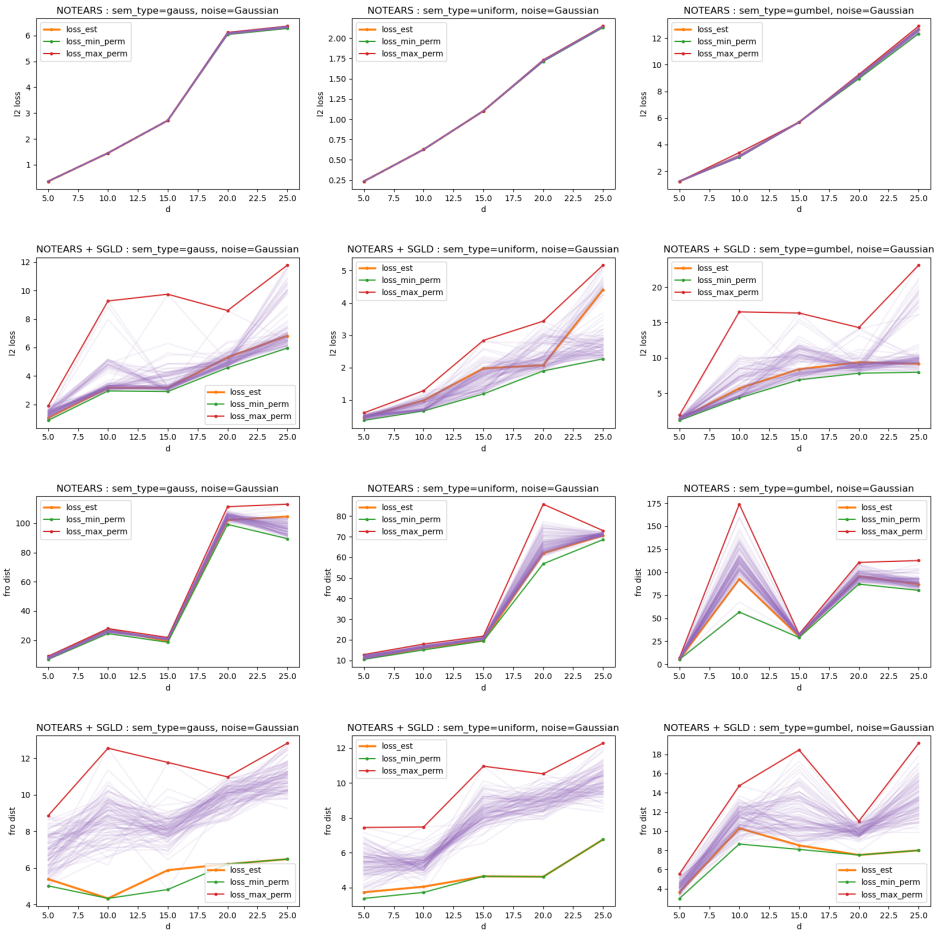

Figure 16: NOTEARS combined with SGLD under non-constant variance

As for DAGMA, results are displayed in Figure 17, Figure 18 and Figure 19. The difference compared to NOTEARS is that SGLD does not bring more fluctuation to estimated L2 losses along with permutations, but it also has little effect in reducing minimized L2 loss. Noteworthily SGLD also significantly diminishes the Frobenious distance both under constant and non-constant variance.

Figure 17: DAGMA combined with SGLD under constant variance

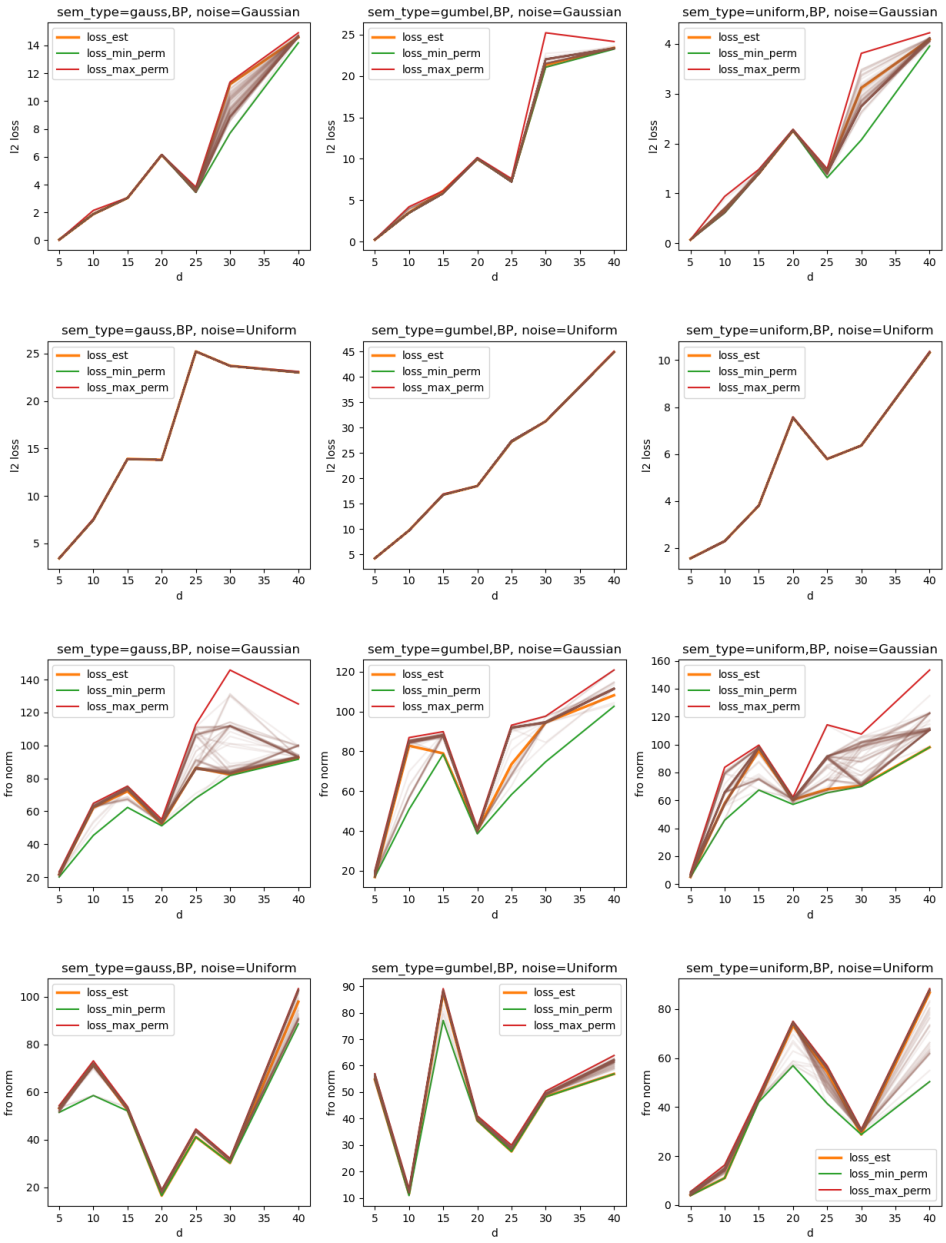

Figure 18: DAGMA non-constant variance

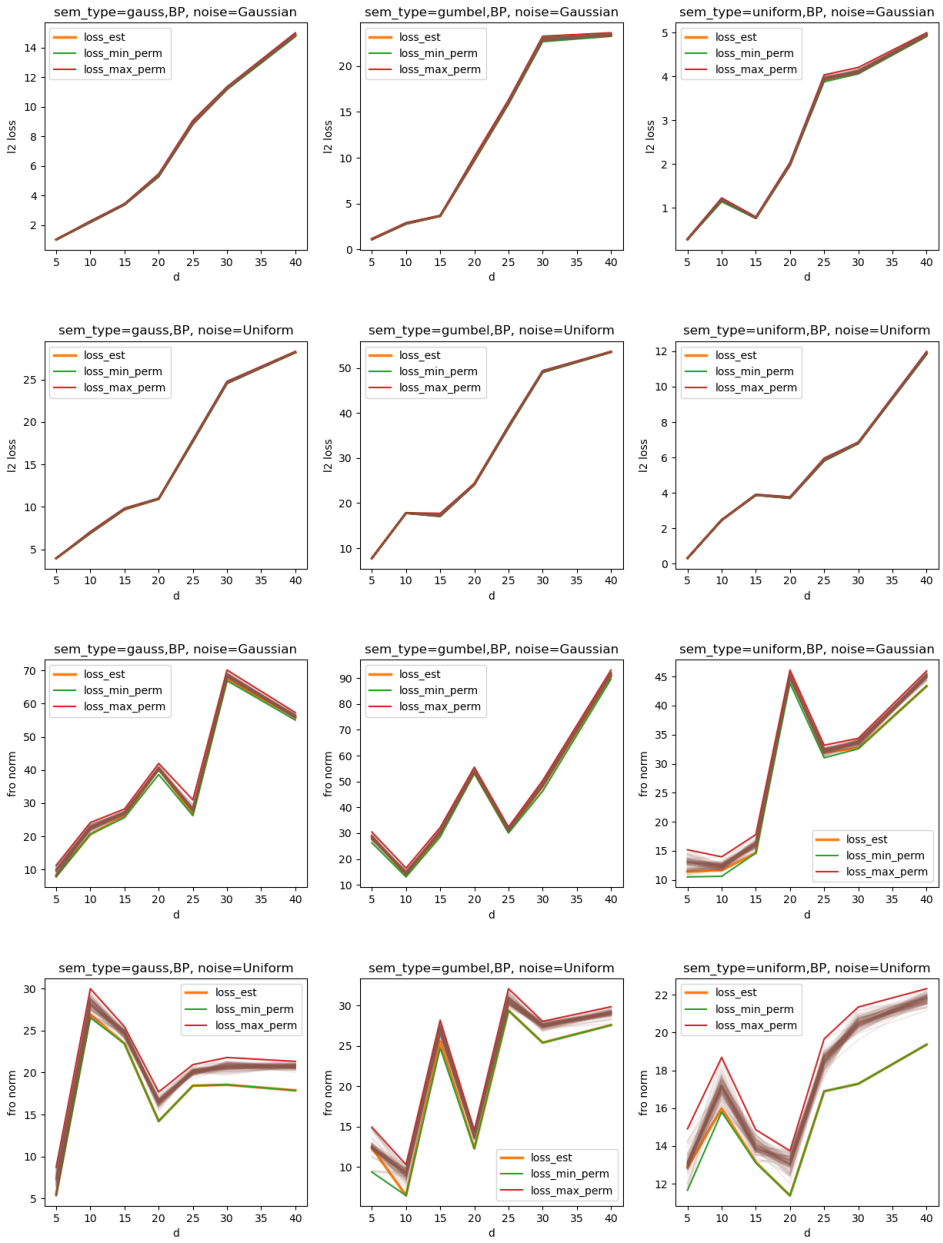

Figure 19: DAGMA combined with SGLD under non-constant variance

4.2.4 Discrete cases

Moreover, we wonder how does NOTEARS perform if data is discrete distributed. Surprisingly, from Figure 20, NOTEARS still perform well in terms of estimated losses in discrete cases (data generated by logistic distribution) even when loss function is not a perfect match. However, it shows that the Frobenius distance has achieved global minimum similarly as in

continuous case. Moreover, from Figure 21, SGLD (i.e. introduced noise in gradients) will attenuate the problem of large distance and instability of distance estimated.

It is worth noting that I did not introduce the heteroskedasticity problem here, because the general data generation function does not have relevant instructions, but perhaps this is an issue worth exploring.

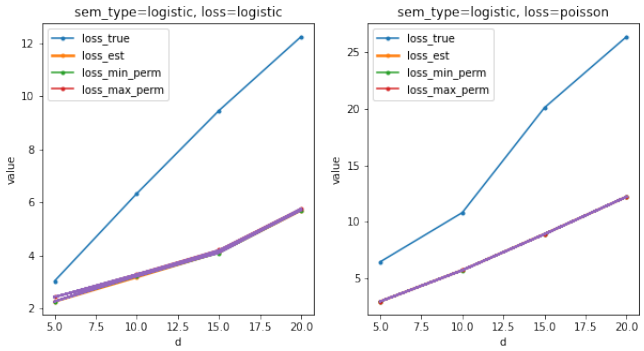

Figure 20: NOTEARS with data from logistic distribution

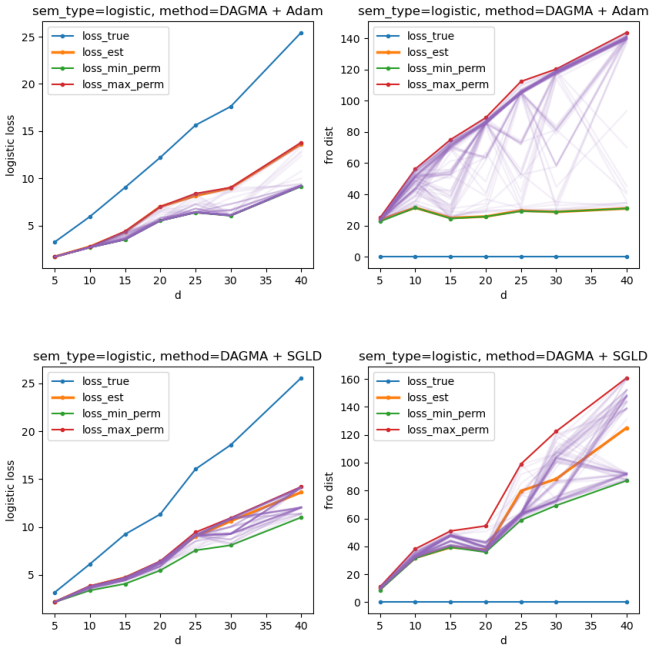

Figure 21: DAGMA

5 Limitations

First, notice the constraint function $h(W) = \text{tr}(e^{W \circ W}) - d = \sum_{k=0}^{\infty} \frac{1}{k!}(W \circ W)^k - d$ is non-convex, leading to a non-convex objective function $O(W, \rho)$. This contributes to multiple minima and possibility of getting stuck in local minimum during optimization.

Second, the algorithm actually tells estimates from parsimonious linear SEMs rather than causal discoveries[5]. Sometimes the magnitude of estimates of coefficients in SEMs doesn't

imply the yes or no problem. Especially, the estimated W will be truncated by an arbitrary threshold.

Third, the algorithm is not robust to rescaling data. Marcus Kaiser and Maksim Sipos[5] show that first update step largely determines the final structure of estimated DAG. Also nodes with high variance are preferred to be sinks as opposed to sources. Moreover, simulated data are generated from additive noise models where child nodes have a higher variance than their parents[6].

Later, Ignavier Ng, etc.[3] suggest using maximized log-likelihood (consider covariance matrix in the denominator to adapt to various data scales) and using soft DAG constraint (directly minimizing $\ell + \lambda_1 \|W\|_1 + \lambda_2 h(W)$ without hard constraints).

References

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Dags with no tears: Continuous optimization for structure learning. 2018.

Y. Samuel Wang and Mathias Drton. Causal discovery with unobserved confounding and non-gaussian data, 2021.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. 2021.

Max Welling and Yee Teh. Bayesian learning via stochastic gradient langevin dynamics. pages 681–688, 01 2011.

Marcus Kaiser and Maksim Sipos. Unsuitability of notears for causal graph discovery. 2021.

Alexander G. Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. 2021.