Summary of Losses in NOTEARS

Mengqi Liu

Aug 7, 2023

1 Notations

- d: number of nodes
- $s0$: expected number of edges
- n: number of samples
- Data matrix: $X = [X_{ij}]_{i,j} = \begin{pmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{pmatrix} = \begin{pmatrix} \tilde{x}_1, & \tilde{x}_2, & ..., & \tilde{x}_d \end{pmatrix} \in \mathbb{R}^{n\times d}$, $i = 1, 2, ..., n$, $j = 1, 2, ..., d$.
- Weighted adjacent matrix: $W = [W_{ij}]_{i,j} = \begin{pmatrix} w_1, & w_2, & \cdots, & w_d \end{pmatrix} \in \mathbb{R}^{d\times d}$, $i = 1, 2, ..., d$, $j = 1, 2, ..., d$.

2 Review on losses in NOTEARS[1]

- loss_est: $\ell(W; X) = \frac{1}{2n}\|X - XW\|_F^2$
- loss_l1: $F(W) = \ell(W; X) + \lambda\|W\|_1 = \frac{1}{2n}\|X - XW\|_F^2 + \lambda\|W\|_1$
- constraint function: $h(W) = \mathrm{tr}(e^{W\circ W}) - d = 0$
- obj_new: $A(W, \rho) = F(W) + \frac{\rho}{2}|h(W)|^2 = \frac{1}{2n}\|X - XW\|_F^2 + \lambda\|W\|_1 + \frac{\rho}{2}[\mathrm{tr}(e^{W\circ W}) - d]^2$
- obj_dual: $L^\rho(W, \alpha) = A(W, \rho) + \alpha h(W) = \frac{1}{2n}\|X - XW\|_F^2 + \lambda\|W\|_1 + \frac{\rho}{2}[\mathrm{tr}(e^{W\circ W}) - d]^2 + \alpha[\mathrm{tr}(e^{W\circ W}) - d]$

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

Two facts:

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

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

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

4 $\mathbf{d > n}$

For any $j \in \{1, 2, ..., d\}$, consider the linear system with W as unknown:

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

According to Rouché–Capelli theorem, when $d > n$, this system has infinitely many solutions. With any solution, we will have $\ell(W; X) = 0$.

5 Drawbacks

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

- The algorithm actually tells estimates from parsimonious linear SEMs rather than causal discoveries[2]. 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.

- The algorithm is not robust to rescaling data. Marcus Kaiser and Maksim Sipos[2] 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[3].

- Ignavier Ng, etc.[4] 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.

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.

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