Explainable Automated Graph Representation Learning with Hyperparameter Importance

Supplementary File

Proof

Lemma 1 *If the number of covariates p_1 and p_2 is fixed, then there exists a sample weight $\gamma \succeq 0$ such that*

$$\lim_{n\to\infty} \mathcal{L}_{Deco} = 0 \tag{1}$$

with probability 1. In particular, a solution γ to Eq (1) is $\gamma_i^\star = \frac{\Pi_{j=1}^p \hat{f}(\mathbf{X}_{i,j})}{\hat{f}(\mathbf{X}_{i,1},\cdots,\mathbf{X}_{i,p})}$, where $\hat{f}(x_{\cdot,j})$ and $\hat{f}(x_{\cdot,1},\cdots,x_{\cdot,p})$ are the Kernel density estimators.[1]

Proof 1 *From [1], if $h_j \to 0$ for $j = 1,\cdots,p$ and $nh_1\cdots h_p \to \infty$,*

$$\hat{f}(x_{i,j}) = f(x_{i,j}) + o_p(1)$$

and

$$\hat{f}(x_i) = f(x_i) + o_p(1)$$

Note that for any j,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbf{X}_{i,j}\gamma_i &= \frac{1}{n}\sum_{i=1}^n \mathbf{X}_{i,j}\frac{\Pi_{j=1}^p f(\mathbf{X}_{i,j})}{f(\mathbf{X}_{i,1},\cdots,\mathbf{X}_{i,p})} + o_p(1)\\ &= \mathbb{E}\left[\mathbf{X}_{i,j}\frac{\Pi_{j=1}^p f(\mathbf{X}_{i,j})}{f(\mathbf{X}_{i,1},\cdots,\mathbf{X}_{i,p})}\right] + o_p(1)\\ &= \int\cdots\int \mathbf{X}_{i,j}\Pi_{l=1}^p f(\mathbf{X}_{i,l})d\mathbf{X}_{i,1}\cdots d\mathbf{X}_{i,p} + o_p(1)\\ &= \int \mathbf{X}_{i,j} f(x_{i,j})d\mathbf{X}_{i,j} + o_p(1)\end{aligned}$$

Similarly, for any j and k, $j \neq k$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbf{X}_{i,j}\mathbf{X}_{i,k}\gamma_i &= \frac{1}{n}\sum_{i=1}^n \mathbf{X}_{i,j}\mathbf{X}_{i,k}\frac{\Pi_{j=1}^p f(\mathbf{X}_{i,j})}{f(\mathbf{X}_{i,1},\cdots,\mathbf{X}_{i,p})} + o_p(1)\\ &= \mathbb{E}\left[\mathbf{X}_{i,j}\mathbf{X}_{i,k}\frac{\Pi_{j=1}^p f(\mathbf{X}_{i,j})}{f(\mathbf{X}_{i,1},\cdots,\mathbf{X}_{i,p})}\right] + o_p(1)\\ &= \int\int \mathbf{X}_{i,j}\mathbf{X}_{i,k} f(x_{i,j})f(x_{i,k})d\mathbf{X}_{i,j}d\mathbf{X}_{i,k}\\ &= \int \mathbf{X}_{i,j} f(x_{i,j})d\mathbf{X}_{i,j}\cdot\int \mathbf{X}_{i,k} f(x_{i,k})d\mathbf{X}_{i,k}\end{aligned}$$

[1] In detail, $\hat{f}(x_{i,j}) = \frac{1}{nh_j}\sum_{i=1}^n k\left(\frac{\mathbf{X}_{i,j}-x_{i,j}}{h_j}\right)$, where $k(u)$ is a kernel function and h_j is the bandwidth parameter for covariate $\mathbf{X}_j$; and $\hat{f}(x_i) = \frac{1}{n|H|}\sum_{i=1}^n K\left(H^{-1}(\mathbf{X}_i - x_i)\right)$, where $K(u)$ is a multivariate kernel function, $H = diag(h_1,\cdots,h_p)$ and $|H| = h_1\cdots h_p$.

Thus, for any $j \neq k$, we have

$$\lim_{n\to\infty} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_{i,j}\mathbf{X}_{i,k}\gamma_i - \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_{i,j}\gamma_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_{i,k}\gamma_i\right)\right)^2 = 0.$$

Hence, for any $\mathbf{A}_{,j} \neq \mathbf{X}_{,k}$, we have

$$\lim_{n\to\infty} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{A}_{i,j}\mathbf{X}_{i,k}\gamma_i - \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{A}_{i,j}\gamma_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_{i,k}\gamma_i\right)\right)^2 = 0.$$

Finally,

$$\lim_{n\to\infty} \sum_{j=1}^{p_1} \left\|\mathbf{A}_{,j}^T \mathbf{\Sigma}_\gamma \mathbf{X}_{,-j}/n - \mathbf{A}_{,j}^T\gamma/n \cdot \mathbf{X}_{,-j}^T\gamma/n\right\|_2^2 = 0.$$

But the solution γ that satisfies Eq (1) in Lemma 1 is not unique. To address this problem, we propose to simultaneously minimize the variance of γ and restrict the sum of γ in our regularizer as follows:

$$\hat{\gamma} = \arg\min_{\gamma \in \mathcal{C}} \mathcal{L}_{Deco} + \frac{\lambda_3}{n}\sum_{i=1}^{n}\gamma_i^2 + \lambda_4(\frac{1}{n}\sum_{i=1}^{n}\gamma_i - 1)^2, \tag{2}$$

where $\mathcal{C} = \{\gamma : |\gamma_i| \leq c\}$ for some constant c.

Then, we have following theorem on our hyperparameter decorrelation regularizer in Eq (2).

Theorem 1 *The solution $\hat{\gamma}$ defined in Eq (2) is unique if $\lambda_3 n \gg p^2 + \lambda_4$, $p^2 \gg \max(\lambda_3, \lambda_4)$ and $|\mathbf{X}_{i,j}| \leq c$ for some constant c.*

Proof 2 *For simplicity, we let $\mathcal{L}_1 := \frac{1}{n}\sum_{i=1}^{n}\gamma_i^2$, $\mathcal{L}_2 := \left(\frac{1}{n}\sum_{i=1}^{n}\gamma_i - 1\right)^2$, and $\mathcal{J}(\gamma) := \mathcal{L}_{Deco} + \lambda_3\mathcal{L}_1 + \lambda_4\mathcal{L}_2$.*

First, we calculate Hessian of $\mathcal{J}(\gamma)$, denoted as $\mathbf{H}$, as follows:

$$\mathbf{H} = \frac{\partial^2 \mathcal{L}_B}{\partial \gamma^2} + \lambda_3 \frac{\partial^2 \mathcal{L}_1}{\partial \gamma^2} + \lambda_4 \frac{\partial^2 \mathcal{L}_2}{\partial \gamma^2}.$$

With some algebra, we have

$$\begin{aligned} \frac{\partial^2 \mathcal{L}_1}{\partial \gamma^2} &= \frac{1}{n}\mathbf{I}, \\ \frac{\partial^2 \mathcal{L}_2}{\partial \gamma^2} &= \frac{1}{n^2}\vec{\mathbf{1}}\vec{\mathbf{1}}^T, \end{aligned}$$

where $\mathbf{I} \in \mathbb{R}^{n\times n}$ is identity matrix, and $\vec{\mathbf{1}} = [1, \cdots, 1]^T \in \mathbb{R}^{n\times 1}$.

For the term $\mathcal{L}_{Deco}$, when $|\mathbf{X}_{i,j}| \leq c$, for any j and k, we have

$$\begin{aligned} \frac{\partial^2}{\partial \gamma^2}(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{i,j}\mathbf{X}_{i,k}\gamma_i)^2 &= \mathcal{O}\left(\frac{1}{n^2}\right), \\ \frac{\partial^2}{\partial W^2}((\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{i,j}\gamma_i)(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{i,k}\gamma_i))^2 &= \mathcal{O}\left(\frac{1}{n^2}\right). \end{aligned}$$

and

$$\frac{\partial^2}{\partial\gamma^2}((\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{i,j}\mathbf{X}_{i,k}\gamma_i)(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{i,j}\gamma_i)(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{i,k}\gamma_i)) = \mathcal{O}\left(\frac{1}{n^2}\right).$$

Then

$$\frac{\partial^2}{\partial\gamma^2}(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{i,j}\mathbf{X}_{i,k}\gamma_i - (\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{i,j}\gamma_i)(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{i,k}\gamma_i))^2 = \mathcal{O}\left(\frac{1}{n^2}\right).$$

$\mathcal{L}_{Deco}$ *is sum of* $p(p-1)$ *such terms, then*

$$\frac{\partial^2\mathcal{L}_B}{\partial\gamma^2} = \mathcal{O}\left(\frac{p^2}{n^2}\right).$$

Thus,

$$\mathbf{H} = \mathcal{O}\left(\frac{p^2}{n^2}\right) + \frac{\lambda_3}{n}\mathbf{I} + \frac{\lambda_4}{n^2}\vec{\mathbf{1}}\vec{\mathbf{1}}^T = \frac{\lambda_3}{n}\mathbf{I} + \mathcal{O}\left(\frac{p^2+\lambda_4}{n^2}\right).$$

Therefore, $\mathbf{H}$ *is an almost diagonal matrix when* $\frac{\lambda_3}{n} \gg \frac{p^2+\lambda_4}{n^2}$*, equivalent to* $\lambda_3 n \gg p^2 + \lambda_4$*. From the relative Weyl theorem [2],* $\mathbf{H}$ *is positive definite. Then the loss function* $\mathcal{J}(\gamma)$ *in Eq (2) is convex on* $\mathcal{C}$*, and has a unique optimal solution* $\hat{\gamma}$*.*

We further want $\mathcal{L}_{Deco}$ *to dominate the regularization terms* $\lambda_3\mathcal{L}_1$ *and* $\lambda_4\mathcal{L}_2$*. On* $\mathcal{C}$*,* $\mathcal{L}_1 = \mathcal{O}(1)$ *and* $\mathcal{L}_2 = \mathcal{O}(1)$*. Moreover,*

$$(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{i,j}\mathbf{X}_{i,k}\gamma_i - (\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{i,j}\gamma_i)(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_{i,k}\gamma_i))^2 = \mathcal{O}(1).$$

and then

$$\mathcal{L}_{Deco} = \mathcal{O}(p^2).$$

As long as $p^2 \gg \max(\lambda_3, \lambda_4)$*,* $\mathcal{L}_{Deco}$ *dominates the regularization terms* $\mathcal{L}_1$ *and* $\mathcal{L}_2$*.*

References

[1] Bruce E Hansen. Lecture notes on nonparametrics. *Lecture notes*, 2009.

[2] Yuji Nakatsukasa. Absolute and relative weyl theorems for generalized eigenvalue problems. *Linear Algebra and its Applications*, 432(1):242–248, 2010.