# Research Discussion on DNN

May 6, 2024

# Introduction

- Consider a nonparametric imputation of $\theta = E(Y)$ using DNN model

$$\hat{\theta}_I = \frac{1}{n}\sum_{i=1}^{n}\{\delta_i y_i + (1-\delta_i)\hat{y}_i\}$$

where $\hat{y}_i$ is the predictor of $y_i$ using DNN model.

- Using the augmentation technique (including $\hat{\pi}_i^{-1}$ into the final layer of the DNN), we obtain

$$\hat{\theta}_I = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{y}_i + \frac{\delta_i}{\hat{\pi}_i}(y_i - \hat{y}_i)\right\}. \tag{1}$$

Unfortunately, $\hat{\theta}_I$ is biased when the nuisance parameters are estimated using DNN.

- Let $m(\mathbf{x}) = E(Y \mid \mathbf{x})$ and $\pi(\mathbf{x}) = P(\delta = 1 \mid \mathbf{x})$. Define

$$\Psi_i = m(\mathbf{x}_i) + \frac{\delta_i}{\pi(\mathbf{x}_i)} \{y_i - m(\mathbf{x}_i)\} \tag{2}$$

and

$$\widehat{\Psi}_i = \hat{m}(\mathbf{x}_i) + \frac{\delta_i}{\hat{\pi}(\mathbf{x}_i)} \{y_i - \hat{m}(\mathbf{x}_i)\}$$

- Farrell et al. (2021) showed (in Theorem 3) that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\Psi}_i - \frac{1}{n}\sum_{i=1}^{n}\Psi_i\right) = o_p(1) \tag{3}$$

which implies that

$$\sqrt{n}\left(\hat{\theta}_I - \frac{1}{n}\sum_{i=1}^{n}\Psi_i\right) = o_p(1)$$

- By CLT, we obtain

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\Psi_i - E(\Psi)\right) \xrightarrow{\mathcal{L}} N(0, V(\Psi)) \tag{4}$$

- Now, since $E(\Psi) = E(Y) = \theta$, two results (3) and (4) imply that

$$\sqrt{n}\left(\hat{\theta}_I - \theta\right) \xrightarrow{\mathcal{L}} N(0, V(\Psi))$$

# Remark

- To understand (3), note that, writing $\bar{\Psi}_n = n^{-1}\sum_{i=1}^n \Psi_i$,

$$\begin{aligned}
\hat{\theta}_I - \bar{\Psi}_n &= \frac{1}{n}\sum_{i=1}^n \left[\hat{m}(\mathbf{x}_i) + \frac{\delta_i}{\hat{\pi}(\mathbf{x}_i)}\{y_i - \hat{m}(\mathbf{x}_i)\}\right] \\
&- \frac{1}{n}\sum_{i=1}^n \left[m(\mathbf{x}_i) + \frac{\delta_i}{\pi(\mathbf{x}_i)}\{y_i - m(\mathbf{x}_i)\}\right] \\
&= -\frac{1}{n}\sum_{i=1}^n \{\hat{m}(\mathbf{x}_i) - m(\mathbf{x}_i)\}\left(\frac{\delta_i}{\pi(\mathbf{x}_i)} - 1\right) \\
&+ \frac{1}{n}\sum_{i=1}^n \delta_i\left(\frac{1}{\hat{\pi}(\mathbf{x}_i)} - \frac{1}{\pi(\mathbf{x}_i)}\right)(m(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i)) \\
&+ \frac{1}{n}\sum_{i=1}^n \delta_i\left(\frac{1}{\hat{\pi}(\mathbf{x}_i)} - \frac{1}{\pi(\mathbf{x}_i)}\right)\{y_i - m(\mathbf{x}_i)\} \\
&:= D_{n1} + D_{n2} + D_{n3}
\end{aligned}$$

- If

$$\sup_x |\hat{m}(\mathbf{x}) - m(\mathbf{x})| = o_p(1) \tag{5}$$

holds, then we can easily show that

$$\sqrt{n} D_{1n} = o_p(1). \tag{6}$$

- Instead of showing (5), Farrell et al. (2021) showed that

$$\frac{1}{n} \sum_{i=1}^{n} \{\hat{m}(\mathbf{x}_i) - m(\mathbf{x}_i)\}^2 = o_p(1) \tag{7}$$

holds under some choices of tuning parameters in the DNN model.

- It is not clear how to show (6) from (7).

- One way to obtain (6) from assumption (7) is to use the sample split technique (Chernozhukov et al., 2018).
- In the sample split technique, we randomly partition the sample into two parts, $S = S_1 \cup S_2$, and estimate $m(\mathbf{x})$ using $S_2$ only to get $\hat{m}^{(2)}(\mathbf{x})$.
- Using $\hat{m}^{(2)}(\mathbf{x})$, we can compute

$$D_{n1}^{(1)} = -\frac{1}{n_1} \sum_{i \in S_1} \left\{ \hat{m}^{(2)}(\mathbf{x}_i) - m(\mathbf{x}_i) \right\} \left( \frac{\delta_i}{\pi(\mathbf{x}_i)} - 1 \right)$$

- Now, we can obtain

$$E \left[ \left\{ \hat{m}^{(2)}(\mathbf{x}_i) - m(\mathbf{x}_i) \right\} \left( \frac{\delta_i}{\pi(\mathbf{x}_i)} - 1 \right) \mid X, \hat{m}^{(2)} \right] = 0$$

and

$$\begin{aligned} E \left[ \left\{ D_{n1}^{(1)} \right\}^2 \right] &= n_1^{-1} E \{ \pi^{-1}(X) - 1 \} \{ \hat{m}^{(2)}(X) - m(X) \}^2 \\ &\leq C n_1^{-1} E \{ \hat{m}^{(2)}(X) - m(X) \}^2 \end{aligned}$$

if $\pi^{-1}(X)$ is uniformly bounded over $X$.

- Therefore, $E\{(D_{n1}^{(1)})^2\} = o(n^{-1})$ if $E\{\hat{m}^{(2)}(X) - m(X)\}^2 = o(1)$, namely, the prediction norm of $\hat{m}^{(2)}(x)$ converges. (This is weaker than the uniform convergence of $\hat{m}^{(2)}(x)$ over $x$.)
- Thus, under (7), we can establish that $\sqrt{n}D_{n1} = o_p(1)$ if we use the sample split technique.
- A similar argument can be made for $D_{n3}$.

- To show $\sqrt{n}D_{n2} = o_p(1)$, note that

$$D_{n2}^2 \leq \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{\hat{\pi}(\mathbf{x}_i)} - \frac{1}{\pi(\mathbf{x}_i)} \right)^2 \right\} \left\{ \frac{1}{n} \sum_{i=1}^{n} (m(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i))^2 \right\}$$

- Farrell et al. (2021) showed that

$$\left\{ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{\hat{\pi}(\mathbf{x}_i)} - \frac{1}{\pi(\mathbf{x}_i)} \right)^2 \right\} \left\{ \frac{1}{n} \sum_{i=1}^{n} (m(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i))^2 \right\} = o_p(n^{-1}) \tag{8}$$

- Thus, we obtain $\sqrt{n}D_{n2} = o_p(1)$.
- Therefore, the main result (3) is established.

# Conclusion

- Using the theory of Farrell et al. (2021), we can establish the CLT for the DNN imputation estimator. But, we need to include two techniques.
  1. Augmented regression in the final layer of the DNN model.
  2. Sample split method
- In the sample split method, the imputation model is trained from the other part of the half-sample.
- Instead of using the imputation context, we can develop the same theory in the context of propensity weighting. That is, we can use the DNN model to create the propensity score weights for handling missing data. (I can provide more details later.)

# Updates (on 3/7)

- Seonjin pointed out that we do not have to use the sample split to achieve (3).
- The idea is to use the MAR assumption and obtain independence between $\hat{m}(\mathbf{x}_i) - m(\mathbf{x}_i)$ and $(\delta_i/\pi(\mathbf{x}) - 1)$ to show (6) from (7).
- I think it is a great idea, but showing it rigorously is not straightforward (as the independence between the two terms holds asymptotically).
- I would like to see its proof. If we can prove it, it will be a good selling point as the existing literature relies on the sample split method.

# Updates (on 5/1)

- If we use the sample split method to estimate $\theta$, we can obtain

$$\hat{\theta}_{I,\text{ss}} = \frac{1}{2}\left(\hat{\theta}_I^{(1)} + \hat{\theta}_I^{(2)}\right) \tag{9}$$

where

$$\hat{\theta}_I^{(1)} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{m}^{(1)}(\mathbf{x}_i) + \frac{2\delta_i I_i^{(2)}}{\hat{\pi}^{(1)}(\mathbf{x}_i)}\left(y_i - \hat{m}^{(1)}(\mathbf{x}_i)\right)\right\}$$

and

$$\hat{\theta}_I^{(2)} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{m}^{(2)}(\mathbf{x}_i) + \frac{2\delta_i I_i^{(1)}}{\hat{\pi}^{(2)}(\mathbf{x}_i)}\left(y_i - \hat{m}^{(2)}(\mathbf{x}_i)\right)\right\}$$

and $I_i^{(1)} \sim \text{Bernoulli}(0.5)$ with $I_i^{(1)} + I_i^{(2)} = 1$.

- Here, the set with $I_i^{(1)} = 1$ is the training sample for $\hat{m}_i^{(1)}$ and $\hat{\pi}_i^{(1)}$.

- The theory of Farrell et al. (2021) can be used to show that

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\psi}_i^{(1)} - \frac{1}{n}\sum_{i=1}^{n}\tilde{\psi}_i^{(1)}\right) = o_p(1) \tag{10}$$

where

$$\tilde{\psi}_i^{(1)} = m(\mathbf{x}_i) + \frac{\delta_i I_i^{(2)}}{\pi(\mathbf{x}_i)/2}\{y_i - m(\mathbf{x}_i)\}$$

and

$$\widehat{\psi}_i^{(1)} = \hat{m}^{(1)}(\mathbf{x}_i) + \frac{\delta_i I_i^{(2)}}{\hat{\pi}^{(1)}(\mathbf{x}_i)/2}\left\{y_i - \hat{m}^{(1)}(\mathbf{x}_i)\right\}$$

- Similarly, we can establish

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\psi}_i^{(2)} - \frac{1}{n}\sum_{i=1}^{n}\tilde{\psi}_i^{(2)}\right) = o_p(1) \tag{11}$$

- Combining (10) and (11) together, we can obtain

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\Psi}_i - \frac{1}{n}\sum_{i=1}^{n}\tilde{\Psi}_i\right) = o_p(1) \tag{12}$$

where

$$\begin{aligned}\hat{\Psi}_i &= \left(\hat{\Psi}_i^{(1)} + \hat{\Psi}_i^{(2)}\right)/2 \\ &= \hat{m}(\mathbf{x}_i) + \frac{\delta_i I_i^{(2)}}{\hat{\pi}^{(1)}(\mathbf{x}_i)}\left\{y_i - \hat{m}^{(1)}(\mathbf{x}_i)\right\} + \frac{\delta_i I_i^{(1)}}{\hat{\pi}^{(2)}(\mathbf{x}_i)}\left\{y_i - \hat{m}^{(2)}(\mathbf{x}_i)\right\}\end{aligned} \tag{13}$$

with $\hat{m}(\mathbf{x}_i) = \left(\hat{m}^{(1)}(\mathbf{x}_i) + \hat{m}^{(2)}(\mathbf{x}_i)\right)/2$ and

$$\begin{aligned}\tilde{\Psi}_i &= \left(\tilde{\Psi}_i^{(1)} + \tilde{\Psi}_i^{(2)}\right)/2 \\ &= m(\mathbf{x}_i) + \frac{\delta_i I_i^{(2)}}{\pi(\mathbf{x}_i)}\left\{y_i - m(\mathbf{x}_i)\right\} + \frac{\delta_i I_i^{(1)}}{\pi(\mathbf{x}_i)}\left\{y_i - m(\mathbf{x}_i)\right\} \\ &= m(\mathbf{x}_i) + \frac{\delta_i}{\pi(\mathbf{x}_i)}\left\{y_i - m(\mathbf{x}_i)\right\}.\end{aligned}$$

- Note that $\tilde{\Psi}_i$ is equal to $\Psi_i$ in (2).
- That is, the sample-split estimator $\hat{\theta}_{I,\mathrm{SS}}$ in (9), which is algebraically equivalent to $n^{-1}\sum_{i=1}^{n}\hat{\Psi}_i = n^{-1}\sum_{i=1}^{n}(\hat{\Psi}_i^{(1)} + \hat{\Psi}_i^{(2)})/2$, is $\sqrt{n}$-consistent for $\theta = E(Y)$ and has asymptotic variance $n^{-1}V(\Psi)$.
- Furthermore, we can use $\hat{\Psi}_i$ in (13) as the estimated influence function to estimate the variance of $\hat{\theta}_{I,\mathrm{SS}}$.
- Therefore, while the original claim in (3) is not fully justified, we can establish (12), which is good enough to achieve our purpose.
- If I am not mistaken, this result is promising and worthy of rigorous investigation. It is consistent with the claim of Farrell et al. (2021) and Chernozhukov et al. (2018), but our example gives more specific details on the proposed method.

# Check Point

- The critical result is Equation (12), which can also be expressed as

$$\sqrt{n}\left(\hat{\theta}_{I,\mathrm{ss}} - \hat{\theta}_{\mathrm{oracle}}\right) = o_p(1)$$

where $\hat{\theta}_{I,\mathrm{ss}}$ is defined in (9) and

$$\hat{\theta}_{\mathrm{oracle}} = \frac{1}{n}\sum_{i=1}^{n}\left\{m(\mathbf{x}_i) + \frac{\delta_i}{\pi(\mathbf{x}_i)}\left(y_i - m(\mathbf{x}_i)\right)\right\}$$

- Therefore, using a Toy simulation, we may make a histogram of the following statistics:

$$T_1 = \sqrt{n}\left(\hat{\theta}_{I,\mathrm{ss}} - \theta\right)$$

$$T_2 = \sqrt{n}\left(\hat{\theta}_{\mathrm{oracle}} - \theta\right)$$

- Also, for comparison, we may also compupte the Monte Carlo sampling distribution of the following statistics:

$$T_3 = \sqrt{n}\left(\hat{\theta}_I - \theta\right)$$

  where $\hat{\theta}_I$ is defined in (1).
- If our theory is correct, then the sampling distribution of $T_1$ should be very similar to that of $T_2$. But, the sampling distribution of $T_3$ should be different. It would be worthwhile to check this using a samll simulation.

# Additional remark

- Chernozhukov et al. (2018) considered two different sample-split double machine learning estimators of $\theta$. They called two estimators DLM1 and DLM2 (in page 23 of the paper).
- In our context, the DML1 estimator can be expressed as

$$\hat{\theta}_{\mathrm{DML1}} = \frac{1}{2}\left(\hat{\theta}_1 + \hat{\theta}_2\right)$$

where

$$\hat{\theta}_1 = \frac{1}{n_1}\sum_{i \in S_1}\left\{\hat{m}^{(2)}(x_i) + \frac{\delta_i}{\hat{\pi}^{(2)}(x_i)}\left(y_i - \hat{m}^{(2)}(x_i)\right)\right\}$$

and

$$\hat{\theta}_2 = \frac{1}{n_2}\sum_{i \in S_2}\left\{\hat{m}^{(1)}(x_i) + \frac{\delta_i}{\hat{\pi}^{(1)}(x_i)}\left(y_i - \hat{m}^{(1)}(x_i)\right)\right\}$$

- Also, the DML2 estimator can be expressed as

$$\begin{aligned}\hat{\theta}_{\mathrm{DML2}} &= \frac{1}{n}\sum_{i\in S_1}\left\{\hat{m}^{(2)}(x_i)+\frac{\delta_i}{\hat{\pi}^{(2)}(x_i)}\left(y_i-\hat{m}^{(2)}(x_i)\right)\right\}\\ &+ \frac{1}{n}\sum_{i\in S_2}\left\{\hat{m}^{(1)}(x_i)+\frac{\delta_i}{\hat{\pi}^{(1)}(x_i)}\left(y_i-\hat{m}^{(1)}(x_i)\right)\right\}\end{aligned}$$

- Thus,

$$\hat{\theta}_{\mathrm{DML2}} = \frac{n_1}{n}\hat{\theta}_1 + \frac{n_2}{n}\hat{\theta}_2.$$

- Therefore, our proposed estimator in (9) is different from the DML estimators of Chernozhukov et al. (2018). It would be interesting to compare the original DML estimators with ours in the simulation study.

# REFERENCES

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey and J. M. Robins (2018), 'Double/debiased machine learning for treatment and structural parameters', *The Econometrics Journal* **21**(1), C1–C68.

Farrell, Max H, Tengyuan Liang and Sanjog Misra (2021), 'Deep neural networks for estimation and inference', *Econometrica* **89**(1), 181–213.