Causal Inference Project: A Data-driven Approach to Prognostic Scores

Sohuen Yi, Heewoong Choi
Seoul National University
{lsdluis, chw0501}@snu.ac.kr

Abstract

A prognostic score is a measure that summarizes the information about the potential outcomes for the control group, $Y(0)$, that is relevant to the covariates. It is useful for estimating the average treatment effect on the treated (ATT). However, the one-dimensional prognostic score proposed by Hansen B. B. [1] may not be sufficient in certain situations where $Y(0)$ consisted of two independent components. To address this issue, we propose a multi-dimensional prognostic score and an effective method for extracting the prognostic score under a specific simulated model. We use independent component analysis and mutual information regression to obtain the prognostic score. Finally, we also demonstrate that our method is superior to using a propensity score with inverse probability weighting (IPW) for estimating ATT.

1 Introduction

In an observational setting, where X are covariates and $Y(0)$, $Y(1)$ are potential outcomes for the control and treated groups, respectively, the higher the dimension of X, the greater the likelihood that some of the information it contains will be irrelevant to the potential outcomes. This is known as the curse of dimensionality. To overcome this issue, one can use a propensity score or prognostic score. A propensity score summarizes the information of receiving treatment into a scalar ($P(Z = 1|X)$), while a prognostic score summarizes the information of $Y(0)$. These scores are useful for calculating the average treatment effect on the treated (ATT). The prognostic score was first proposed in [1], but the authors focused on the one-dimensional prognostic score. However, there are many cases in which all the information on the potential outcomes of controls cannot be compressed into a single real number.

In this paper, we explore the case where the distribution of potential outcomes of the controls cannot be characterized by a scalar-valued prognostic score. To address this, we extend the one-dimensional prognostic score to a multi-dimension. We use independent component analysis (ICA) and mutual information (MI) regression to uncover the multi-dimensional prognostic score of observational data samples. We show that ICA and MI regression can recover the most informative components of $Y(0)$ in our simulated data samples and consider these components as a multi-dimensional prognostic score. Furthermore, we validate the quality of the estimated prognostic score in estimating the average treatment effect of the treated (ATT) using a three-layer neural network and Random Forest. Our results show that estimating ATT using our proposed method achieves higher accuracy than using a propensity score and inverse probability weighting (IPW). We attach the code of the Python implementation of the proposed method.

Fall 2022 Casual Inference Final Project

2 Backgrounds

2.1 Prognostic Score and Propensity Score

Definition 1. *$\psi(X)$ is a prognostic score if*

$$Y(0) \perp\!\!\!\perp X \mid \psi(X)$$

where X denotes covariates and $Y(0)$ denotes the potential outcome for those not receiving treatment[1].

Definition 1 states that we call $\psi(X)$ a prognostic score when $\psi(X)$ has enough information for $Y(0)$. A prognostic score can be posed equivalently as the following:

Proposition 1. *If $\psi(X)$ is a prognostic score, it satisfies*

$$P(Y(0) \,|\, X) = P(Y(0) \,|\, \psi(X)),$$

which roughly means $\psi(X)$ contains all information of X that is relevant to Y.

This proposition renders the following proposition, which we can exploit to calculate the ATT:

Proposition 2. *The ATT can be estimated with a prognostic score, $\psi(X)$:*

$$\mathbb{E}[\,Y(1) - Y(0) \mid Z = 1] = \mathbb{E}[\,Y \mid Z = 1] - \mathbb{E}[\,\mathbb{E}[\,Y \mid \psi(X), Z = 0] \mid Z = 1]$$

Proof.

$$\begin{aligned}
\mathbb{E}[\,Y(1) - Y(0) \mid Z = 1] &= \mathbb{E}[\,Y(1) \mid Z = 1] - \mathbb{E}[\,Y(0) \mid Z = 1] \\
&= \mathbb{E}[\,Y \mid Z = 1] - \mathbb{E}[\,\mathbb{E}[Y(0) \mid X] \mid Z = 1] \\
(\because \textit{Proposition } 1) &= \mathbb{E}[\,Y \mid Z = 1] - \mathbb{E}[\,\mathbb{E}[Y(0) \mid \Psi(X)] \mid Z = 1] \\
&= \mathbb{E}[\,Y \mid Z = 1] - \mathbb{E}[\,\mathbb{E}[Y(0) \mid \Psi(X), Z = 0] \mid Z = 1] \\
&= \mathbb{E}[\,Y \mid Z = 1] - \mathbb{E}[\,\mathbb{E}[Y \mid \Psi(X), Z = 0] \mid Z = 1].
\end{aligned}$$

□

On the other hand, a more standard approach to ATT estimation is utilizing a propensity score, as presented in the following proposition.

Proposition 3. *Assuming a propensity score $\rho(X)$ satisfies $0 < \rho(X) < 1$, the ATT can be estimated through Inverse Probability Weighting (IPW):*

$$\mathbb{E}[\,Y(1) - Y(0) \mid Z = 1] = \mathbb{E}[\,Y \mid Z = 1] - \frac{P(Z=0)}{P(Z=1)}\mathbb{E}\left[\,\frac{\rho(X)}{1-\rho(X)}Y \mid Z = 0\right]$$

Proof.

$$\begin{aligned}
\mathbb{E}[\,Y(1) - Y(0) \mid Z = 1] &= \mathbb{E}[\,Y(1) \mid Z = 1] - \mathbb{E}[\,Y(0) \mid Z = 1] \\
&= \mathbb{E}[\,Y \mid Z = 1] - \mathbb{E}[\,Y(0) \mid Z = 1]
\end{aligned}$$

The second term can be calculated with IPW,

$$\begin{aligned}
\mathbb{E}[\,Y(0) \mid Z = 1] &= \sum_X P(X \mid Z = 1)\mathbb{E}[\,Y(0) \mid X, Z = 1] \\
&= \sum_X P(X \mid Z = 1)\mathbb{E}[\,Y(0) \mid X, Z = 0] \\
&= \sum_X P(X \mid Z = 1)\mathbb{E}[\,Y \mid X, Z = 0] \\
&\overset{(i)}{=} \sum_X P(X \mid Z = 0)\frac{P(Z=0)}{P(Z=1)}\mathbb{E}\left[\,\frac{\rho(X)}{1-\rho(X)}Y \mid X, Z = 0\right] \\
&= \frac{P(Z=0)}{P(Z=1)}\mathbb{E}\left[\,\frac{\rho(X)}{1-\rho(X)}Y \mid Z = 0\right]
\end{aligned}$$

which completes the proof. Note that (i) is due to

$$\frac{P(X \mid Z=1)}{P(X \mid Z=0)} = \frac{P(Z=0)P(X, Z=1)}{P(Z=1)P(X, Z=0)} = \frac{P(Z=0)}{P(Z=1)} \frac{\rho(X)}{1-\rho(X)}.$$

□

A prognostic score has two similarities with a propensity score. First, both summarize the information into a scalar and satisfy the independence conditions $Z \perp\!\!\!\perp X \mid \rho(X)$ for a propensity score and $Y(0) \perp\!\!\!\perp X \mid \psi(X)$ for a prognostic score. Second, both are useful for estimating ATT by proposition 2 and 3.

While a prognostic score contains valuable information in data samples, a one-dimensional form has a significant limitation. For example, consider a scenario where the true $Y(0)$ is composed of the product of two independent covariates X_1 and X_2 including a noise term:

$$Y(0) = (X_1 + \epsilon_1)(X_2 + \epsilon_2). \tag{1}$$

Both X_1 and X_2 must be specified to determine the distribution of $Y(0)$. This indicates that a one-dimensional prognostic score is insufficient to fully capture the information of $Y(0)$ in specific situations. Therefore, we propose a multi-dimensional prognostic score to apply better the concept of a prognostic score to the real world. We will use Independent Component Analysis and Mutual Information regression to find a multi-dimensional prognostic score given data samples.

2.2 Independent Component Analysis (ICA)

Independent component analysis (ICA) is a method to extract independent components from mixed observational data, assuming the unmixed components are non-Gaussian distributions [2]. To informally present the problem, the ICA finds the mixing matrix A and the unmixed components U satisfying $X = UA$, where X is given observational data.

The non-Gaussianity assumption of each component of U is crucial; the ICA tries to find A that maximizes 'non-Gaussianity' of U, so the method will not work if U follows Gaussian distribution. One of the widely used ICA methods is FastICA [2], which maximizes the kurtosis of U to obtain non-Gaussian components. This method is implemented in `Python` package `scikit-learn` [5], and we use this implementation throughout our experiments.

2.3 Mutual Information (MI)

Informally speaking, the mutual information between two random variables X, Y is defined as

$$\mathcal{I}(X; Y) = D_{\mathrm{KL}}(P_{(X,Y)} \| P_X \otimes P_Y)$$

where D_{KL} denotes the KL divergence, and P_X and P_Y are distributions of X and Y respectively, and $P_{(X,Y)}$ is the joint distribution of (X, Y).

$\mathcal{I}(X; Y) = 0$ implies the joint distribution $P_{(X,Y)}$ and $P_X \otimes P_Y$ are identical thus X and Y are independent. In the same sense, smaller $\mathcal{I}(X; Y)$ means $P_{(X,Y)}$ and $P_X \otimes P_Y$ are 'closer'. Roughly speaking, this indicates X and Y share less information. $\mathcal{I}(X; Y)$ can be estimated from samples [4], and it is implemented in `scikit-learn` [5].

3 Method

3.1 Problem Statement

We consider a scenario where the observed covariates $X \in \mathbb{R}^d$ and potential outcomes $Y(0)$, $Y(1) \in \mathbb{R}$ are influenced by unmeasured confounders $U \in \mathbb{R}^d$. We assume that the relationship between X and U is linear, such that $X = UA$ for some mixing matrix $A \in \mathbb{R}^{d \times d}$. Under these assumptions, we aim to solve the following problems:

1. With prior knowledge of the minimum dimension of prognostic scores, we find (estimate) an adequate prognostic score.
2. Using the estimated prognostic score, we calculate the ATT from the given data.

3.2 Model Specification

In this section, we provide a more detailed specification of the data-generating model that was described in the previous section. Consider

$$U_i = [u_{i1}, u_{i2}, \ldots, u_{id}] \tag{2}$$
$$X_i = U_i A \tag{3}$$
$$Y_i(0) = f_0(u_1, u_2, \ldots, u_k, \epsilon_1, \epsilon_2, \ldots, \epsilon_k) \tag{4}$$
$$Y_i(1) = f_1(u_1, u_2, \ldots, u_k, \ldots, u_d, \xi_1, \xi_2, \ldots, \xi_k, \ldots, \xi_d) \tag{5}$$
$$Z_i = \text{Bernoulli}(p_i), \quad p_i = f(X_i) \tag{6}$$
$$Y = (1 - Z)Y_i(0) + ZY_i(1) \tag{7}$$

where U, X, $Y(0)$, and $Y(1)$ are defined as before, Z is a treatment vector depend on X, and ϵ_i, $\xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$ are random noises.

According to the model specification, the potential control outcome $Y(0)$ is determined by the unmeasured confounders $u_1, \ldots, u_k$. This means that a prognostic score must include all information about these confounders. If $u_1, \ldots, u_k$ are independent, then a prognostic score should be at least a k-dimensional vector in order to capture all of this information. For example, if $k = 2$ and u_1, u_2 are independent, then a prognostic score should be at least a 2-dimensional vector.

3.3 Finding a Prognostic Score

The data generation procedure implies that a prognostic score cannot be one-dimensional. The lowest possible dimension of a prognostic score is k in our model: we assume this knowledge is accessible. Given this knowledge, we propose the following procedure for estimating a prognostic score,

1. Given X, perform ICA to recover U (up to scalar multiplication)
2. Estimate the MI between $Y(0)$ and each component of recovered $\hat{U}$
3. Sort the components of $\hat{U}$ in descending order according to the MI estimates of step 2 and select the top k components.

We provide more details about the three-step procedure for prognostic score estimation that was introduced earlier. The first step of this procedure is to recover the unmeasured confounders U. The recovered version of U is denoted as $\hat{U}$. If we know that the observed covariates X are a linear combination of U with a noise term, then we can use ICA to recover U. We assume that the number of unmeasured confounders is known and ICA may not be able to recover U if the number of confounders is incorrect. The second step of the procedure is to apply mutual information (MI) regression between $Y(0)$ and the estimated confounders $\hat{U}$. This step is motivated by the idea that the MI between two variables decreases as they become more independent, and that zero MI indicates that the two variables are perfectly independent. To implement this step, we use the MI regression method by [4] as implemented in [5]. The final step is to select the informative components of $\hat{U}$ that are relevant to $Y(0)$. We expect that if we choose the high MI estimation components of $\hat{U}$ that generate Y, they will be an appropriate prognostic score. Since we assume that we have access to the value of k, we select the top-k components of the MI regression outcome as the prognostic score.

3.4 Evaluating a Prognostic Score via Estimation of the ATT

With the prognostic score $\psi(X)$ obtained in the previous section, we can now estimate ATT using Proposition 2 as:

$$\mathbb{E}[\,Y(1) - Y(0) \mid Z = 1] = \mathbb{E}[\,Y \mid Z = 1] - \mathbb{E}[\,\mathbb{E}[\,Y \mid \psi(X), Z = 0] \mid Z = 1] \tag{8}$$

To evaluate the right-hand side of equation 8, we need to estimate both $\mathbb{E}[\,Y \mid Z = 1]$ and $\mathbb{E}[Y \,|\, \psi(X), Z = 1]$. The former can be estimated directly from the data, but the latter requires a more complex estimation procedure. We use machine learning (ML) approaches, specifically neural networks (NN) and random forests (RF), to estimate $\mathbb{E}[\,\mathbb{E}[\,Y \mid \psi(X), Z = 0] \mid Z = 1]$.

4 Experiments

We use the ATT as a metric to evaluate the accuracy and effectiveness of our proposed method for extracting a prognostic score from the data. This value can be estimated using Propostion 2, and we train ML models to estimate the second term of the equation 8. Once we have obtained the prognostic score using ICA and MI regression, we train a model to predict $Y(0)$ under the control group using the prognostic score as input. To analyze the data and make predictions, we use two ML models: a three-layer neural network (NN) and a random forest (RF) which 100 estimators and a maximum depth of 5. We also utilize a propensity score to estimate the ATT using Proposition 3. We use a neural network to estimate the propensity score in this case. This will allow us to compare the results of the original method that uses a propensity score.

4.1 Simulated Data

In this section, we provide the result of a toy experiment. We now specify the parameters we used regarding the model (2)–(7). We set $d = 100$, $k = 2$, and $\sigma = 0.1$ in the model. We sample $N = 2000$ units through the following:

$$
\begin{aligned}
u_{i1} &= \sigma_1(\sin(2i/M) + \epsilon_{i1}) && \in \mathbb{R}^{2000} \\
u_{i2} &= \sigma_2(\mathrm{sign}(\sin(2i/M)) + \epsilon_{i2}) && \in \mathbb{R}^{2000} \\
u_{i3} &= \sigma_3(2\mathrm{frac}(i/M) - 1 + \epsilon_{i3}) && \in \mathbb{R}^{2000} \\
u_{i3} &= \sigma_4(2\mathrm{frac}(3i/2M) - 1 + \epsilon_{i4}) && \in \mathbb{R}^{2000} \\
X &= UA \\
&= \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ u_{i1} & u_{i2} & u_{i3} & u_{i4} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} A && \in \mathbb{R}^{2000 \times 100} \\
Z_i &\sim \mathrm{Bernoulli}(p_i), \quad p_i = |\sin(\textstyle\sum_j x_{ij})|
\end{aligned}
$$

where $M = 250$, $\mathrm{frac}(x) \coloneqq x - \lfloor x \rfloor$, $\epsilon_{il} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 0.2^2)$, A is sampled from the standard normal distribution, and σ_l's are positive numbers that standardizes $U_{:,l}$.

4.2 Experiment Results

Recovering $\hat{U}$ via ICA

We perform ICA to recover $\hat{U}$. The horizontal axis of Figure 1 represents each unit. Figure 1 depicts that ICA recovers U very well; The only difference is scaling.

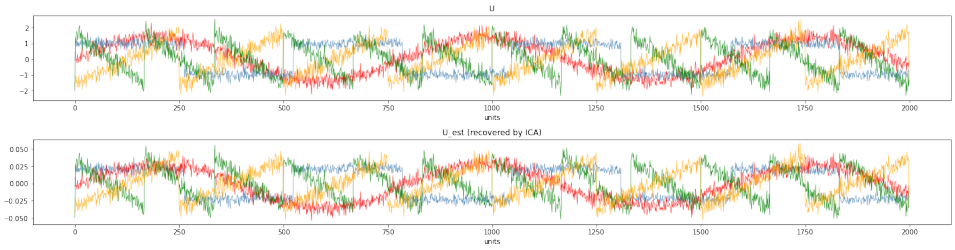

Figure 1: Ground truth U and estimated $\hat{U}$

Estimating ATT

We present the estimated ATT obtained using our proposed method with both the NN and RF and the results of a propensity score-based ATT estimation which is denoted as IPW. We further provide a boxplot in Figure 2.

Our proposed method for extracting the prognostic score relies on the assumption that we know the number of dimensions required to predict the prognostic score accurately. Meanwhile, as shown in

Table 1 and Figure 2, the performance of our method may change if this assumption is not met (*i.e.*, we do not know the minimum possible dimension of a prognostic score). If we assume a prognostic score to be a scalar (*i.e.*, $k = 1$), the estimated ATT is significantly biased upwards. As we assume a prognostic score to have a dimension greater than 1, the bias gets reduced.

As shown in the figures, our proposed method using a random forest presents far smaller bias than the IPW method; our procedure works better than the standard IPW approach.

k	NN	RF	IPW	Ground Truth
1	-0.1320±0.0314	-0.1199±0.0023	-0.1368±0.0036	-0.1667
2	-0.1634±0.0137	-0.1628±0.0016		
3	-0.1527±0.0068	-0.1636±0.0015		
4	-0.1641±0.0114	-0.1635±0.0015		

Table 1: Estimated ATT. Mean ± Std for 10 repetitions. k stands for a prognostic score dimension.

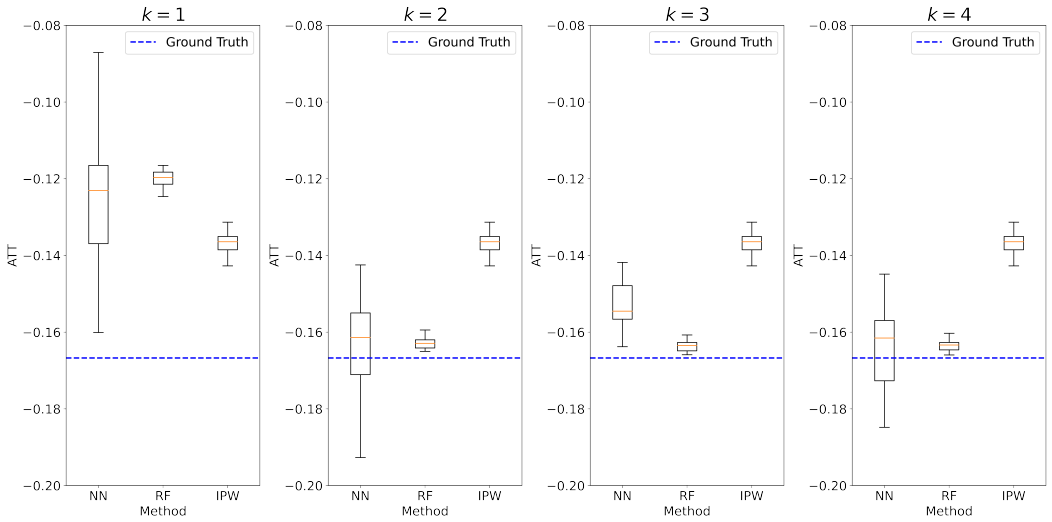

Figure 2: The boxplot of the ATT estimation results. The blue dashed line represents the ground truth ATT, k stands for a prognostic score dimension, and x labels indicate a method used. NN and RF mean our method using a prognostic score and IPW means a method using a propensity score.

5 Conclusion

In this study, we introduced the concept of a multi-dimensional prognostic score and an effective, data-driven method to extract it, which exploits mutual information and the independent component analysis. We also demonstrated its advantages over a propensity score when applied to the ATT estimation problem. Using a multi-dimensional prognostic score resulted in more accurate and precise estimates of ATT compared to the IPW method. However, the approach relies on the assumption that the minimum possible dimension of the prognostic score is known. Also, it only works for linear mixtures of unobserved confounders when using the linear ICA technique. Future research should address these limitations by determining the lowest dimension of prognostic scores and applying non-linear ICA techniques [3] to identify unobserved confounders in non-linear mixtures.

References

[1] B. B. Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, February 2008.

[2] Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[3] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

[4] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004.

[5] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12:2825–2830, 2011.