ECMA 31380 - Draft Final Project

Fernando Rocha Urbano

Autumn 2024

1 Causal Factor Investing: Can Factor Investing Become Scientific?

Authors of factor models do not identify the causal graph consistent with the observed phenomenon, they justify their chosen model specification in terms of correlation and do not propose experiments for falsifying causal mechanisms.

Absent a causal theory, their findings are likely false, due to rampant backtest overfitting and incorrect specification choices.

Economists subscribe to the view that genuine science must produce refutable implifications.

In the absence of plausible falsifiable theories, researchers must acknowledge that they do not understand why the reported anomalies (risk premia) occur.

1.a Association vs Causation

Two variables are statistically associated if:

$$\mathbb{P}[X = x, Y = y] \neq \mathbb{P}[X = x] \times \mathbb{P}[Y = y]$$

A variable X is said to cause a variable Y when Y is a function of X in the data generating process.

There is a difference between conditioning $X = x$ and setting. The idea of setting is an intervention.

The intervention means that you set the value of X, which is often represented with $do[X = x]$.

If X does not cause Y:

$$\mathbb{P}[Y = y \mid do[X = x]] = \mathbb{P}[Y = y]$$

Even if:

$$\mathbb{P}[Y = y \mid X = x] \neq \mathbb{P}[Y = y]$$

For instance, if Z causes X and Y, Z is considered a confounder, because it is the variable that introduces association between X and Y, even though there is no relationship in the data generating process between X and Y.

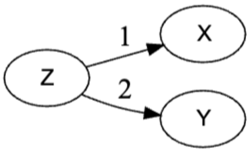


Figure 1: Z confounder for X and Y

Causality is an extra-statistical concept, connected to mechanisms and interventions.

Causation does imply association because setting $X = x$ through an intervention is associated with the $Y = y$.

Causality is directional. The statement "X causes Y" implies:

$$\mathbb{P}[Y = y \mid do[X = x]] \neq \mathbb{P}[Y = y]$$

But does not imply:

$$\mathbb{P}[X = x \mid do[Y = y]] \neq \mathbb{P}[X = x]$$

Association is not directional.

Studies designed to establish causality propose methods to nullify the bias of the ATT (SSB). These are:

1. Intervention Studies (RCT): randomized controlled trials.
2. Natural Experiments: when intervention studies are not possible (unfeasible, unethical). Units are assigned to the treatment and control groups determined randomly by Nature. Common examples of natural experiments include: (1) regression discontinuity design (RDD); (2) crossover studies (COS); and (3) difference-in-differences (DID) studies. The critical assumption behind RDD is that groups (a) and (b) are comparable in everything but the slight difference in the assignment variable, which can be attributed to noise. A COS is a longitudinal study in which the exposure of units to a treatment is randomly removed for a time, and then returned. COS assume that the effect of confounders does not change per unit over time.
3. Simulated Interventions: researchers may still conduct an observational study that simulates a do-operation, with the help of a hypothesized causal graph. The hypothesized causal graph encodes the information needed to remove from observations the SSB introduced by confounders, under the assumption that the causal graph is correct.

1.b Hypothesized Causal Graph

A simulated intervention allows researchers to estimate the strength of a causal effect from observational studies. Second, a simulated intervention may help falsify a hypothesized causal graph, when the strength of one of the effects posited by the graph is deemed statistically insignificant.

In simulated interventions, the causal graph is part of the assumptions, and one cannot prove what one is assuming. The most a simulated intervention can achieve is to disprove a hypothesized causal graph, by finding a contradiction between an effect claimed by a graph and the effect estimated with the help of that same graph.

The latter differs from interventional studies and natural experiments. In those, subject to some assumptions, a researcher can establish or falsify a causal claim without knowledge of the causal graph.

The most a simulated intervention can achieve is to disprove a hypothesized causal graph, by finding a contradiction between an effect claimed by a graph and the effect estimated with the help of that same graph. This power of simulated interventions to falsify causal claims can be very helpful in discovering through elimination the causal structure hidden in the data.

1.c Causal Discovery

Can be defined as the search for the structure of causal relationships, by analyzing the statistical properties of observational evidence.

While observational evidence almost never suffices to fully characterize a causal graph, it often contains information helpful in reducing the number of possible structures of interdependence among variables.

In more recent years, scientists have developed numerous computational methods and algorithms for the discovery of causal relations, represented as directed acyclic graphs.

- constraint-based algorithms
- score-based algorithms
- functional causal models

1.c.1 Constraint-Based Algorithms

The two most widely used are:

- PC Algorithm: assumes that there are no latent (unobservable) confounders, and under this assumption the discovered causal information is asymptotically correct.
- FCI Algorithm: gives asymptotically correct results even in the presence of latent confounders.

1.c.2 Score-based methods

Score-based methods can be used in the absence of latent confounders.

These algorithms attempt to find the causal structure by optimizing a defined score function. An example of a score-based method is the greedy equivalence search (GES) algorithm.

This heuristic algorithm searches over the space of Markov equivalence classes, that is, the set of causal structures satisfying the same conditional independences, evaluating the fitness of each structure based on a score calculated from the data.

1.c.3 Functional Causal Models

Causal graphs can also be derived from non-numerical data. For example, Laudy et al. [2022] apply natural language processing techniques to news articles in which different authors express views of the form $X \to Y$. By aggregating those views, these researchers derive directed acyclic graphs that represent collective, forward-looking, point-in-time views of causal mechanisms.

1.c.4 Machine Learning

With ML, we can decouple the variable search from the specification search.

Examples include mean-decrease accuracy, local surrogate models, and Shapley values (López de Prado [2020, pp. 3-4], López de Prado [2022a]).

1.d Blocked Paths

In a graph with three variables $\{X, Y, Z\}$, the variable Z is:

- Confounder with respect to X and Y: when the causal relation has $X \leftarrow Z \rightarrow Y$
- Collider with respect to X and Y: when the causal relation has $X \rightarrow Z \rightarrow\leftarrow Y$
- Mediator with respect to X and Y: when the causal relation has $X \rightarrow Z \rightarrow Y$

A path is a sequence of arrows and nodes that connect the two variables X and Y.

- Directed path: path in which all arrows point in the same direction
- X is ancestor of Z in the path which starts with X and ends with Z.
- Z is descendant of X in the path which starts with X and ends with Z.
- A path between X and Y is blocked if either:
  - the path traverses collider and the researcher has not conditioned on that collider or its descendants.
  - the researcher conditions on a variable in the path between X and Y, where the conditoned variable is not a collider.
- Causal associations only flows along an unblocked directed path that starts in treatment X and ends in outcome Y, denoted the causal path. Association implies causation only if all non-causal paths are blocked.

1.e Adjustments

1.e.1 Backdoor Adjustment

A backdoor path between X and Y is an unblocked non-causal path that connects those two variables. The term backdoor is inspired by the fact that this kind of paths have an arrow pointing into the treatment (X).

This image has a backdoor path pointed in red (non-causal path) and a causal path in green.

Having this backdoor path is a problem due to association, not allowing to recover the true ATE.

Backdoor paths can be blocked by conditioning on a set of variables S that satisfies the backdoor criterion. Meaning thta we want to control for observable confounders.

A set of variables S satisfies the backdoor criterion with regards to treatment X and outcome Y if the following two conditions are true:

- conditioning on S blocks all backdoor paths between X and Y (blocks all non-causal paths).
- S does not contain any descendants of X (keeps open all causal paths).

Then, S is a sufficient adjustment set, and the causal effect of X on Y can be estimated as

$$\mathbb{P}[Y(x) = y] = \sum_s \mathbb{P}[Y(x) = y \mid X = x, S = s]\mathbb{P}[S = s]$$

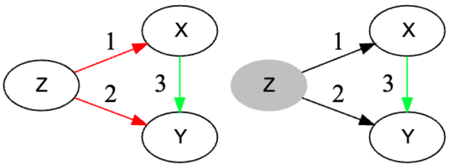


Figure 2 – Example of a causal graph that satisfies the backdoor criterion, before (left) and after (right) conditioning on *Z* (shaded node)

Figure 2: Backdoor Adjustment

1.e.2 Backdoor Adjustment: Another Explanation

`https://www.youtube.com/watch?v=U1S8Rq8IcrY`

We want to block the W_1 and C that are non causal associations.

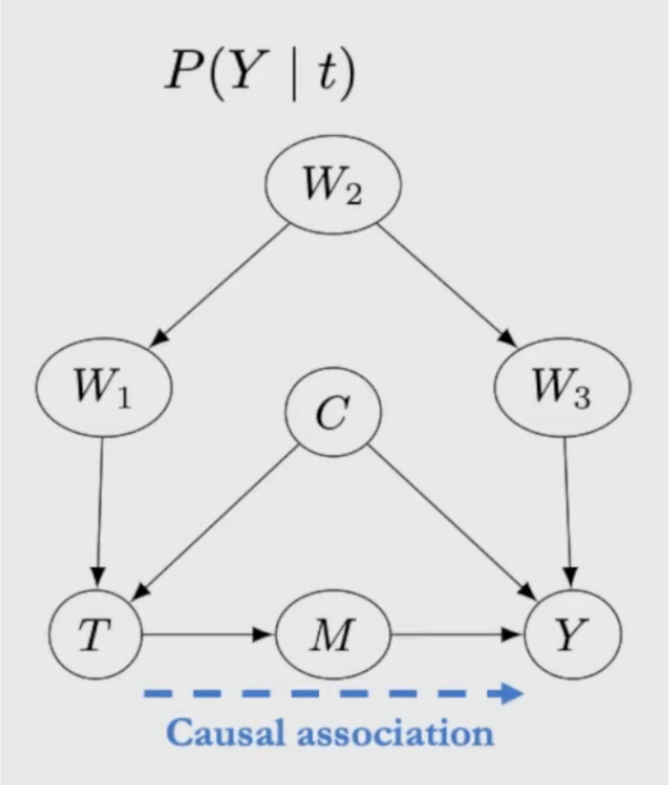


Figure 3: Backdoor Adjustment

Our end goal is to find the interventional distribution of $Y(t)$.

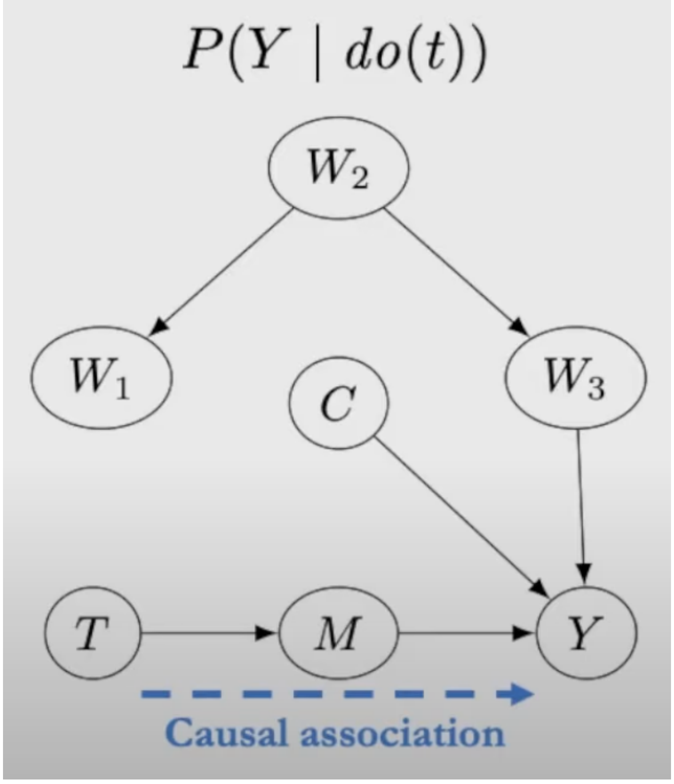


Figure 4: Backdoor Adjustment

Nonetheless, this is not doable in data if, given that it is an intervention.

Thus, we must find $mathbbP[Y \mid t, W_2, c]$ or $mathbbP[Y \mid t, W_1, c]$ or $mathbbP[Y \mid t, W_3, c]$.

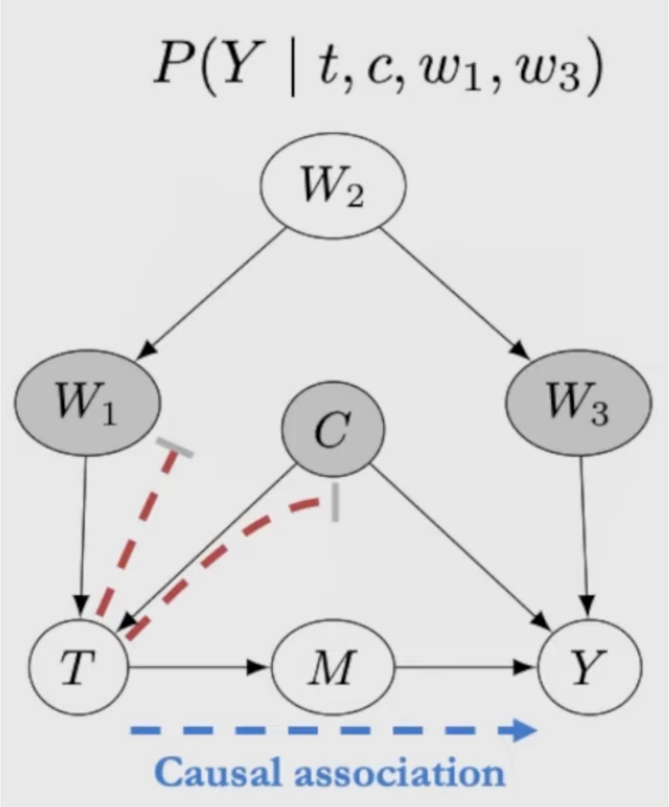
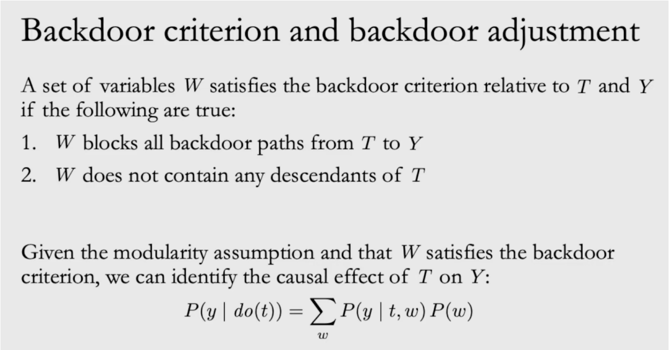


Figure 5: Backdoor Adjustment



Backdoor criterion and backdoor adjustment

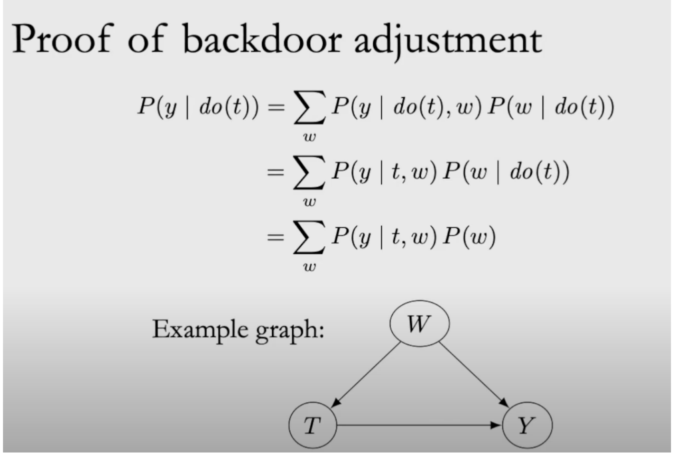
A set of variables W satisfies the backdoor criterion relative to T and Y if the following are true:

1. W blocks all backdoor paths from T to Y
2. W does not contain any descendants of T

Given the modularity assumption and that W satisfies the backdoor criterion, we can identify the causal effect of T on Y:

$$P(y \mid do(t)) = \sum_{w} P(y \mid t, w)\, P(w)$$

Figure 6: Backdoor Adjustment



Proof of backdoor adjustment

$$\begin{aligned} P(y \mid do(t)) &= \sum_w P(y \mid do(t), w)\, P(w \mid do(t)) \\ &= \sum_w P(y \mid t, w)\, P(w \mid do(t)) \\ &= \sum_w P(y \mid t, w)\, P(w) \end{aligned}$$

Example graph:

Figure 7: Backdoor Adjustment

The backdoor criterion is often related to "d-separation".

it is very important that the set of variables used as backdoor blockers do not contain any descendants of the treatment.

Meaning that:

$$Y \perp T|S$$

Where S is again the set of variables used as backdoor blockers.

1.e.3 Front-Door Adjustment

Sometimes researchers may not be able to condition on a variable that satisfies the backdoor criterion. That is such a case when the variable is latent (unobservable).

A causal approach can be achieved with a mediator.

A set of variables S satisfies the front-door criterion with regards to treatment X and outcome Y if the following three conditions are true:

- all causal paths from X to Y go through S
- there is no backdoor path between X and S
- all backdoor paths between S and Y are blocked by conditioning on X.

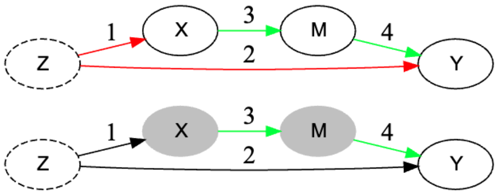


Figure 3 – Example of a causal graph that satisfies the front-door criterion, before (top) and after (bottom) adjustment

Figure 8: Frontdoor Adjustment

Then, S is a sufficient adjustment set, and the causal effect of X on Y can be estimated as:

$$\mathbb{P}[Y(x) = y] = \sum_{s} \mathbb{P}[S = s \mid X = x] \sum_{x'} \mathbb{P}[Y = y \mid X = x', S = s]\mathbb{P}[X = x']$$

1.e.4 Front-Door Adjustment: Another Explanation

Recalling the backdoor adjustment, we use descendent variable to T and Y to correct the bias. Nonetheless, lets say that those variables are not available (latent).

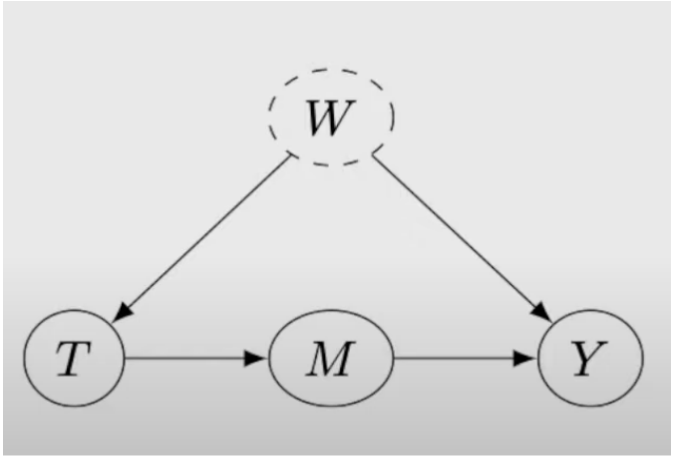


Figure 9: Frontdoor Adjustment

Now, if we only focus on M, we can get a better perspective.

Meaning that we have to understand the causal association of T to M and the causal association of M to Y.

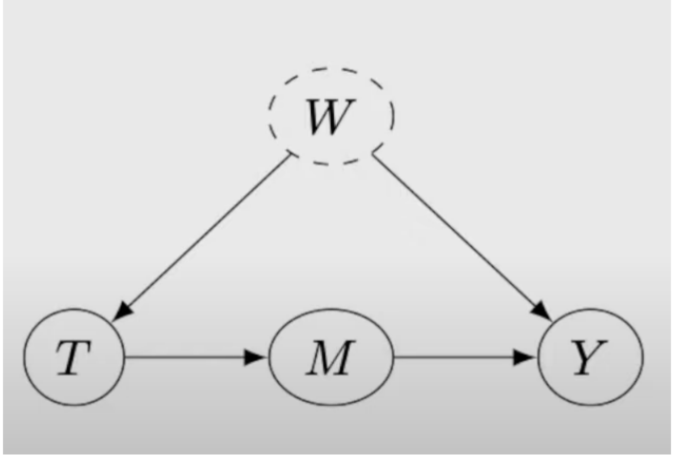
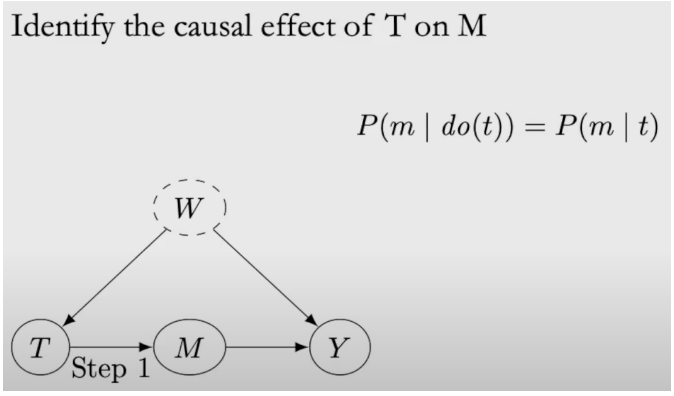


Figure 10: Frontdoor Adjustment

The first step is to identify the causal effect of T on M: here there is no backdoor path, therefore, the step is quite simple.

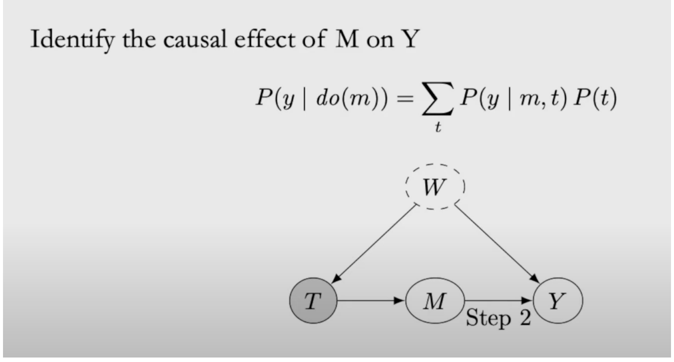


Identify the causal effect of T on M

$$P(m \mid do(t)) = P(m \mid t)$$

Figure 11: Frontdoor Adjustment

The second step is identify the causal effect of M on Y: here there is a backdoor path. Thus, we have to condition of T to account for that.



Identify the causal effect of M on Y

$$P(y \mid do(m)) = \sum_t P(y \mid m, t)\, P(t)$$

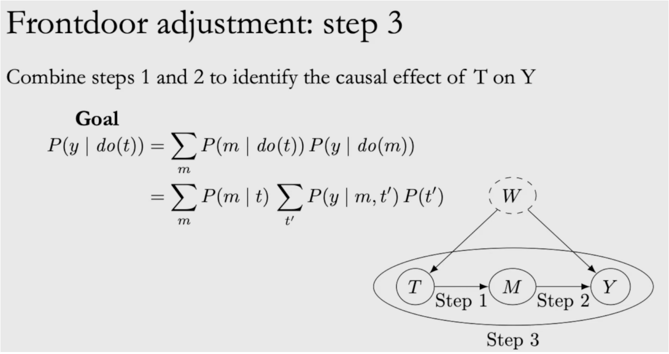
Figure 12: Frontdoor Adjustment

The third step is to combine.

We use the causal association of $\mathbb{P}[m(t)]$ (step 1) and the $\mathbb{P}[y(m)]$.

Thus:

$$\begin{aligned}
\mathbb{P}[y(t)] &= \sum_m \mathbb{P}[m(t)]\mathbb{P}[y(m)] \\
&= \sum_m \mathbb{P}[m \mid t] \sum_{t'} \mathbb{P}[y|m, t']\mathbb{P}[t']
\end{aligned}$$



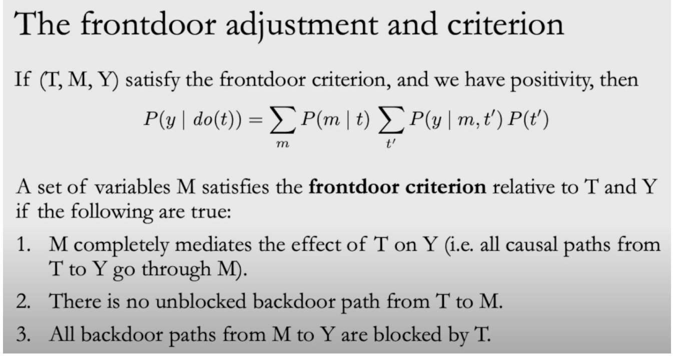
Frontdoor adjustment: step 3

Combine steps 1 and 2 to identify the causal effect of T on Y

Goal

$$\begin{aligned}
P(y \mid do(t)) &= \sum_m P(m \mid do(t))\, P(y \mid do(m)) \\
&= \sum_m P(m \mid t) \sum_{t'} P(y \mid m, t')\, P(t')
\end{aligned}$$

Figure 13: Frontdoor Adjustment



The frontdoor adjustment and criterion

If (T, M, Y) satisfy the frontdoor criterion, and we have positivity, then

$$P(y \mid do(t)) = \sum_{m} P(m \mid t) \sum_{t'} P(y \mid m, t')\, P(t')$$

A set of variables M satisfies the **frontdoor criterion** relative to T and Y if the following are true:

1. M completely mediates the effect of T on Y (i.e. all causal paths from T to Y go through M).
2. There is no unblocked backdoor path from T to M.
3. All backdoor paths from M to Y are blocked by T.

Figure 14: Frontdoor Adjustment

1.e.5 Instrumental Variables

The front-door adjustment controls for latent confounders when a mediator exists.

In the absence of a mediator, the instrumental variables method allows to control for a latent confounder Z, as long as they can find W that turns X into a collider, thus blocking the backdoor path through Z.

A variable W satisfies the instrumental variable criterion relative to the treatment X and outcome Y if:

- there is an arrow $W \to X$;
- the causal effect of W on Y is fully mediated by X;
- there is no backdoor path between W and Y

Intuitively, the first and second condition ensure that W can be used as proxy for X, whereas condition three prevents the need for an additional backdoor adjustment to de-confound the effect of W on Y.

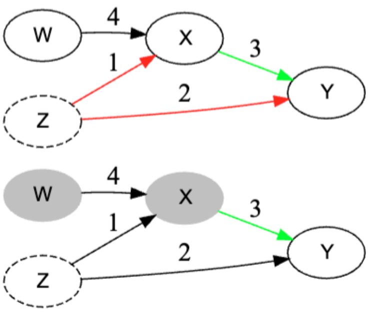


Figure 4 – Example of a causal graph with an instrumental variable W, before (top) and after (bottom) adjustment

Figure 15: Instrumental Variable

1.e.6 Front-Door Adjustment: Another Explanation

When we have unobserved confounding, we have to deal with it in other manner.

If we have the frontdoor adjustment, we can use the frontdoor.

Here, we use Z as the instrumental variable:

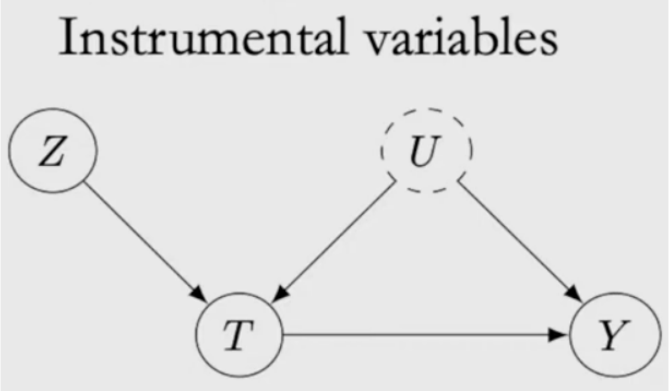


Figure 16: Instrumental Variable

We consider an instrument Z a variable that has a causal effect on the treatment T.

Furthermore, the causal effect of the Z on Y is fully mediated by the treatment T.

We must have a graph where the edge between Z and Y is removed:

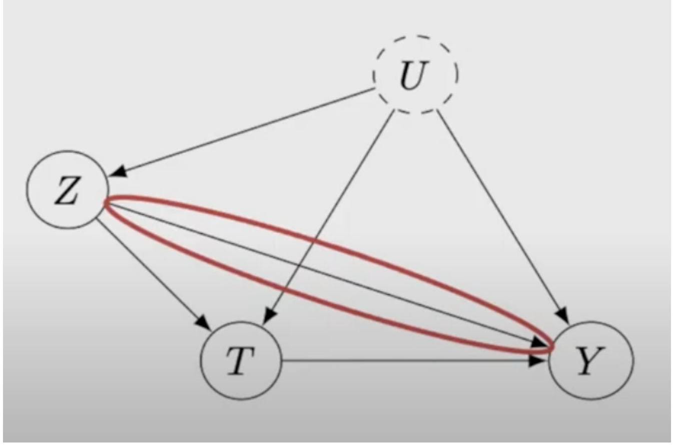


Figure 17: Instrumental Variable

Meaning that we are adding an assumption that we are excluding Z from the causal mechanism that generates Y.

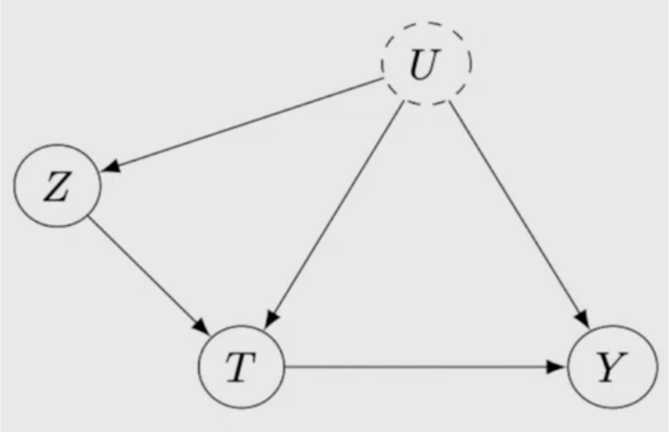


Figure 18: Instrumental Variable

The last assumption for Z to be an instrument is that there is no backdoor path from Z to Y. Meaning that must be no direct path between U and Z:

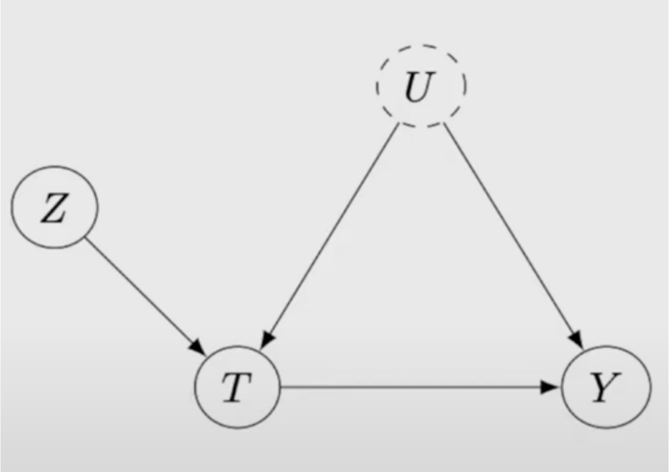


Figure 19: Instrumental Variable

If there is backdoor path between Y and Z, but this backdoor path is observed, we can condition on the backdoor path. In this new case, the backdoor path is the W. If that is the case, Z is called a conditional instrument.

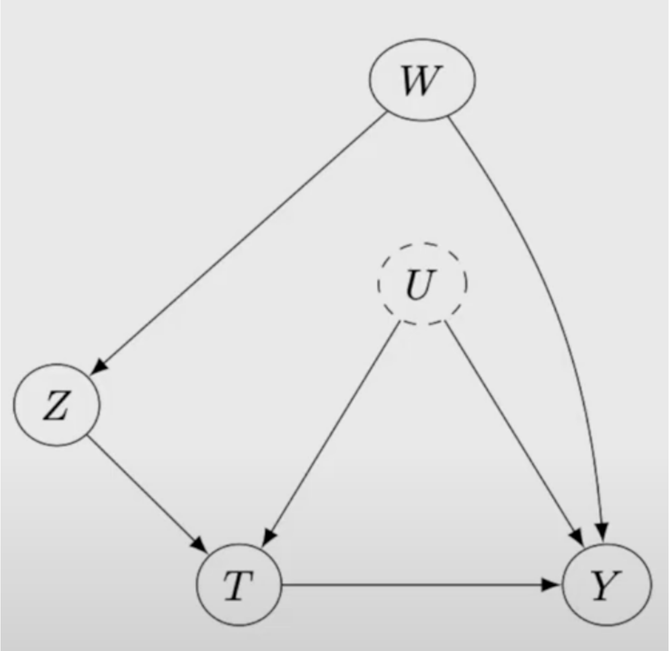


Figure 20: Instrumental Variable

This gives a slightly weaker version of assumption 3, meaning that we unconfoundedness after conditioning on observed variables.

2 Deep Neural Networks for Estimation and Inference

In the 1990s, shallow neural networks with smooth activation functions were shown to have many good theoretical properties.

Stochastic optimization techniques ad more computing power changed the focus from shallow to deep networks and from smooth sigmoid-type activation to rectified linear units (ReLU).

Our bounds immediately yield empirical and population L2 convergence rates.

We follow our main results by applying our nonasymptotic high probability bounds to deliver valid inference on finite-dimensional parameters following first-step estimation using deep learning.

Our work contributes directly to this area of research by showing that deep nets are a valid and useful first-step estimator for semiparametric inference in general. Further, we show that inference after deep learning may not require sample splitting or cross-fitting.

2.a Nonasymptotic Bounds for DNN Estimation

Estimate $f_*(x)$ that relates to the covariates $X \in \mathbb{R}^d$ and $Y \in \mathbb{R}$.

$$f_* = \text{argmin}_f \mathbb{E}[\ell(f, Z)]$$

Where Z is $(Y, X) \in \mathbb{R}^{d+1}$

We allow any loss function that is Lipschitz in f.

Two leading examples of the applications for causal inference are:

- Least squares:
$$f_*(x) := \mathbb{E}[Y|X = x], \quad \ell(f, z) = \frac{1}{2}(y - f(x))^2$$
- Logistic Regression:
$$f_*(x) := \log\left(\frac{\mathbb{E}[Y|X = x]}{1 - \mathbb{E}[Y|X = x]},\right) \quad \ell(f, z) = -yf(x) + \log(1 + e^{f(x)})$$

We call the specific choice of the neural network architecture as $\mathcal{F}_{\text{DNN}}$.

We focus on feedforward neural networks.

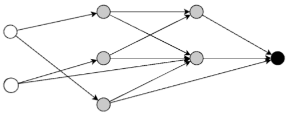


FIGURE 1.—Illustration of a feedforward neural network with $W = 18$, $L = 2$, $U = 5$, and input dimension $d = 2$. The input units are shown in white at left, the output in black at right, and the hidden units in grey between them.

Figure 21: Feed Forward Architecture

In it, we have $X \in \mathbb{R}^d$ in the input and Y in the outcome. U hidden units are present in between, grouped in a sequence of L layers.

The width of the network at a given layer is denoted H_l, meaning the number of neurons in that layer.

The networks is completed with the choice of an activation function $\sigma : \mathbb{R} \mapsto \mathbb{R}$.

An important commonly used subclass of neural networks is a fully connected one. This architecture is known as multilayer perceptron (MLP). We defone it as $\mathcal{F}_{\text{MLP}}$.

We will assume that all the width of all layers share a common asumptotic order H, implying that for this class $UsimeqLH$.

We find a suboptimal rate for MLP case, but our upper bound is still sufficient for semiparametric inference.

The total number of parameters in the network is:

$$W = (d + 1)H + (L - 1)(H^2 + H) + H + 1$$

Where d is the number of covariates.

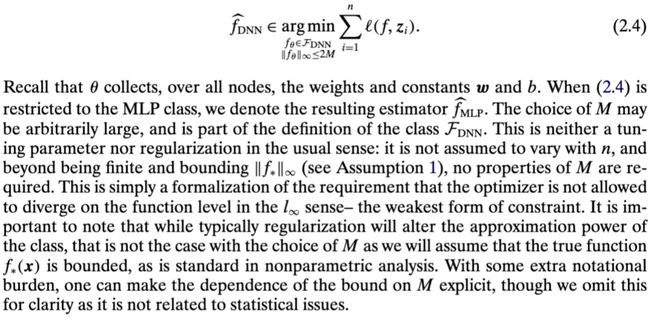
For neural networls, the architecture as a whole are the smoothing parameters while the width and depth play the role of tuning parameters.

Just as for classical nonparametrics, for a fixed architecture it is the tuning parameters that determine the rate of convergence (fixing smoothness of f_*). The recent wave of theoretical study of deep learning is still in its infancy. As such, there is no understanding of optimal architectures or tuning parameters. These choices can be difficult and only preliminary research has been done (see, e.g., Daniely (2017), Telgarsky (2016) and references therein). However, it is interesting that in some cases, results can be obtained even with a fixed width H, provided the network is deep enough; see Corollary 2.

In $\mathcal{F}_{\text{DNN}}$ we have to choose:

- U
- L
- W
- graph structure
- $\sigma(\cdot)$

Why does M need to be bounded? What is M?



$$\widehat{f}_{\text{DNN}} \in \underset{\substack{f_\theta \in \mathcal{F}_{\text{DNN}} \\ \|f_\theta\|_\infty \leq 2M}}{\arg\min} \sum_{i=1}^{n} \ell(f, \boldsymbol{z}_i). \tag{2.4}$$

Recall that θ collects, over all nodes, the weights and constants $\boldsymbol{w}$ and b. When (2.4) is restricted to the MLP class, we denote the resulting estimator $\widehat{f}_{\text{MLP}}$. The choice of M may be arbitrarily large, and is part of the definition of the class $\mathcal{F}_{\text{DNN}}$. This is neither a tuning parameter nor regularization in the usual sense: it is not assumed to vary with n, and beyond being finite and bounding $\|f_*\|_\infty$ (see Assumption 1), no properties of M are required. This is simply a formalization of the requirement that the optimizer is not allowed to diverge on the function level in the l_∞ sense– the weakest form of constraint. It is important to note that while typically regularization will alter the approximation power of the class, that is not the case with the choice of M as we will assume that the true function $f_*(\boldsymbol{x})$ is bounded, as is standard in nonparametric analysis. With some extra notational burden, one can make the dependence of the bound on M explicit, though we omit this for clarity as it is not related to statistical issues.

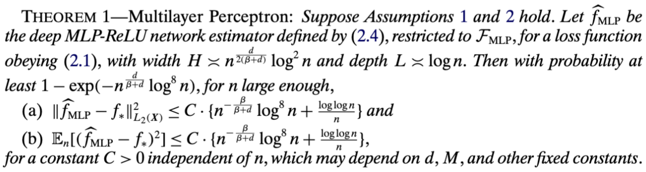
Figure 22: Feed Forward Architecture

Explicit regularization may improve empirical performance in low signal-to-noise ratio problems.

2.a.1 The Nonasymptotic Bonds

Assumptions:

- We use mild assumptions
- The difference is that we also assume that the outcome is bounded.
- Only continous covariates. Discrete covariates would probably show a lower convergence rate. In practice, NN show great performance with discrete covariates.



THEOREM 1—Multilayer Perceptron: *Suppose Assumptions* 1 *and* 2 *hold. Let* $\widehat{f}_{\text{MLP}}$ *be the deep MLP-ReLU network estimator defined by* (2.4), *restricted to* $\mathcal{F}_{\text{MLP}}$, *for a loss function obeying* (2.1), *with width* $H \asymp n^{\frac{d}{2(\beta+d)}} \log^2 n$ *and depth* $L \asymp \log n$. *Then with probability at least* $1 - \exp(-n^{\frac{d}{\beta+d}} \log^8 n)$, *for n large enough,*

(a) $\|\widehat{f}_{\text{MLP}} - f_*\|^2_{L_2(X)} \leq C \cdot \{n^{-\frac{\beta}{\beta+d}} \log^8 n + \frac{\log\log n}{n}\}$ *and*

(b) $\mathbb{E}_n[(\widehat{f}_{\text{MLP}} - f_*)^2] \leq C \cdot \{n^{-\frac{\beta}{\beta+d}} \log^8 n + \frac{\log\log n}{n}\}$,

for a constant $C > 0$ *independent of n, which may depend on d, M, and other fixed constants.*

Figure 23: Feed Forward Architecture

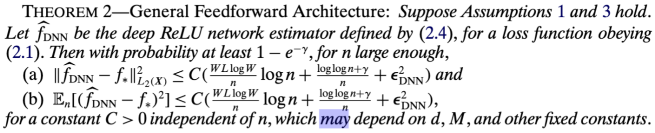
In this paper proof they use scale sensitive localization theory. This has the benefit of:

- not restrict the class of networks architectures to have bounded weights for each unit.
- richer set of approximation possibilities.

They are able to aattain a faster rate on the second term of the bound, order $n{-}1$. This upper bounds inform the trade offs between H and L, and the approximation power, and may point toward optimal architectures for statistical inference.

They note that as is standard in nonparametrics, this result relies on choosing H appropriately given the smoothness of β appearing in H.

Theorem 1 covers only one specific architecture, albeit the most important one for current practice. However, given that this field is rapidly evolving, it is important to consider other possible architectures which may be beneficial in some cases.



THEOREM 2—General Feedforward Architecture: *Suppose Assumptions* 1 *and* 3 *hold. Let* $\widehat{f}_{\text{DNN}}$ *be the deep ReLU network estimator defined by* (2.4), *for a loss function obeying* (2.1). *Then with probability at least* $1 - e^{-\gamma}$, *for n large enough,*

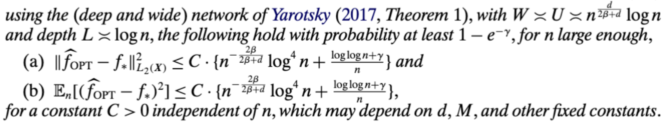
(a) $\|\widehat{f}_{\text{DNN}} - f_*\|^2_{L_2(X)} \leq C(\frac{WL\log W}{n} \log n + \frac{\log\log n + \gamma}{n} + \epsilon^2_{\text{DNN}})$ *and*

(b) $\mathbb{E}_n[(\widehat{f}_{\text{DNN}} - f_*)^2] \leq C(\frac{WL\log W}{n} \log n + \frac{\log\log n + \gamma}{n} + \epsilon^2_{\text{DNN}})$,

for a constant $C > 0$ *independent of n, which may depend on d, M, and other fixed constants.*

Figure 24: Feed Forward Architecture

This result is more general than Theorem 1, covering the general deep ReLU network problem defined in (2.4), general feedforward architectures, and the general class of losses defined by (2.1). The same comments as were made following Theorem 1 apply here as well: the same localization argument is used with the same benefits. We explicitly use this in the next two corollaries, where we exploit the allowed flexibility in controlling DNN by stating results for particular architectures. The bound here is not directly applicable without specifying the network structure, which will determine both the variance portion (through W, L, and U) and the approximation error. With these set, the bound becomes operational upon choosing γ, which can be optimized as desired.



COROLLARY 1—Optimal Rate: *Suppose Assumptions* 1 *and* 2 *hold. Let* $\widehat{f}_{\text{OPT}}$ *solve* (2.4) *using the (deep and wide) network of* Yarotsky (2017, *Theorem* 1), *with* $W \asymp U \asymp n^{\frac{d}{2\beta+d}} \log n$ *and depth* $L \asymp \log n$, *the following hold with probability at least* $1 - e^{-\gamma}$, *for n large enough,*

(a) $\|\widehat{f}_{\text{OPT}} - f_*\|^2_{L_2(X)} \leq C \cdot \{n^{-\frac{2\beta}{2\beta+d}} \log^4 n + \frac{\log\log n + \gamma}{n}\}$ *and*

(b) $\mathbb{E}_n[(\widehat{f}_{\text{OPT}} - f_*)^2] \leq C \cdot \{n^{-\frac{2\beta}{2\beta+d}} \log^4 n + \frac{\log\log n + \gamma}{n}\}$,

for a constant $C > 0$ *independent of n, which may depend on d, M, and other fixed constants.*

Figure 25: Feed Forward Architecture