Notes

Xingyu Chen[1]

January 15, 2026

Contents

1 Tpyst symbols

`Tpyst symbols`.

2 Random effects model for election polls

Andrew Gelman's recent blog post, `Polls & Betting Odds & Nonsampling Errors & Win Probabilities & Vote Margins`, discusses how to estimate winning probabilities based on polling data. A ChatGPT summary is available here: `chatgpt-history`.

He presents three examples: the New Jersey governor race, the Virginia governor race, and the New York City mayoral race. Individual polls are often unreliable, so systematic differences between polls must be considered. A simple random-effects model can capture this phenomenon:

$$Y_{i,t} = \theta + \eta_{i,t} + \varepsilon_i$$

Here, $Y_{i,t}$ is the t-th poll for the i-th media, θ is the true underlying support level, ε_i represents the systematic error for poll i, and $\eta_{i,t}$ represents the sampling error of the poll.

When focusing on a single media's polls (fixing $i = i_0$), Gelman uses historical information to estimate ε_t, assuming $\varepsilon_t \sim \mathcal{N}(0, 2\%)$. He then uses polling data of media i_0 to estimate the sample variance $\sigma_{i_0}^2$ of η_{i,i_0}.

Let $\overline{Y_{i_0}} = \frac{1}{n}\sum_{i=1}^{n} Y_{i,i_0}$. Under his Bayesian viewpoint, the posterior distribution of θ is approximated by:

$$\theta \mid Y \sim \mathcal{N}\left(\overline{Y_{i_0}}, \sqrt{\sigma_{i_0}^2 + 2\%}\right)$$

The winning probability can then be computed as:

$$\Pr(\theta > 0.5 \mid Y).$$

3 Writing

3.1 Good title: The unreasonable effectiveness of XX in/of YY

Good title form `The unreasonable effectiveness of mathematics in the natural sciences` by Eugene Wigner in 1960 [1].

Richard Hamming in computer science, "The Unreasonable Effectiveness of Mathematics".

Arthur Lesk in molecular biology, "The Unreasonable Effectiveness of Mathematics in Molecular Biology"

Peter Norvig in artificial intelligence, "The Unreasonable Effectiveness of Data"

Vela Velupillai in economics, "The Unreasonable Ineffectiveness of Mathematics in Economics"

Terrence Joseph Sejnowski in Artificial Intelligence: The Unreasonable Effectiveness of Deep Learning in Artificial Intelligence".

3.2 Good title: Rebel With a Cause

Form the famous movie `Rebel Without a Cause`.

The special issue Volume 8, Issue 2, 2022 Issue of *Observational Studies* titleed `Rebel With a Cause`

3.3 Words

"Nonparametric identification and estimation of the ATE with non-binary IVs are more involved." in [2]. *involved* means complicated, difficult, complex.

3.4 Fun example

3.4.1 On overparameterized models

Form comment in Impossible statistical problems of Andrew Gelman by Phil, November 14, 2024.

"I'm imagining a political science student coming in for statistical advice:
Student: I'm trying to predict the Democratic percentage of the two-party vote in U.S. Presidential elections, six months before Election Day. I want to use just the past ten elections because I think the political landscape was too different before that.
Statistician: Sounds interesting. What predictive variables do you have?
Student: I've got the Democratic share in the last election, and the change in unemployment rate over the past year and the past three years, and the inflation rate over the past year and the past three years, and the change in median income over the past year and past three years.
Statistician: That's a lot of predictors for not many elections, we are going to have some issues, but maybe we can use lasso or a regularization scheme or something. Let's get started.
Student: I also own an almanac.
Statistician: Oh. Sorry, I can't help you, your problem is impossible."

With only 10 data points and 7 predictors, there is still some room for analysis. However, when using an almanac with over 1,000 predictors, the problem becomes unsolvable: the model is overparameterized and loses all predictive power for future observations.

Therefore, in scenarios with extremely small sample sizes, an excess of irrelevant predictors can contaminate the data—rather than enriching it—and render meaningful analysis impossible.

But it now an empirical observation, can we theoretically explain this phenomenon?

***Question* 3.1.** For sample siez $n = 200$, outcome $Y \in \mathbb{R}$ and predictors $X \in \mathbb{R}^p$, $\frac{p}{n} = c \in (0, \infty)$, Y is independent of X, under what condition? We will see that a machine learning algorithm can still predict Y well from X?.

Well, this is just the global null hypothesis testing problem in high dimension models. We can use a nonparametric regression view to see this problem.

$$Y = f(X) + \varepsilon, f \in \mathcal{H}$$
$$\mathrm{H}_0 : f = 0, \mathrm{H}_1 : f \neq 0$$

Using a base function of $\mathcal{H}$ and truncating at k terms, it should be answered well as a hypothesis testing problem, the signal noise ratio, sparsity and the constant $\frac{p}{n}$ will give the detection boundary, also the local minimax rate.

Well, that's asymptotic theory, there still is the question for tiny n, say $n = 10$ or 20. Can we give any answer for this? Can we know anything useful form so tiny sample size? This case may be called Knightian uncertainty?

"In economics, Knightian uncertainty is a lack of any quantifiable knowledge about some possible occurrence, as opposed to the presence of quantifiable risk (e.g., that in statistical noise or a parameter's confidence interval). The concept acknowledges some fundamental degree of ignorance, a limit to knowledge, and an essential unpredictability of future events."

4 On the undistinguishable or identification of statistical models

Based on talk with Ruiqi Zhang, Lin Liu and Chatgpt.

The undistinguishable means you can not distinguish two models form data. Which will corresponds to 3 cases.

One and two are both in the modeling stage. When you consider one model P_θ, this corresponds to non-identifiable model. When you consider two models, this corresponds to two models sharing some common distributions.

The third case is in the hypothesis testing stage, two models can not be distinguished by data since they are too close.

In a word, two models $\mathcal{P}_1$ and $\mathcal{P}_2$ as two distribution calsses are undistinguishable if $\mathcal{P}_1 \cap \mathcal{P}_2 \neq \emptyset$ or they are too close under some metric.

4.1 On latent structure models

The potential outcome model is an example of latent structure model. The observed random variable is determined by some unobservable/latent variable is in this calss.

Definition 4.1. *The observed random variable X is determined by a high dimensional latent variable Z by a map $X = f(Z)$.*

***Example* 4.2.** The observed random variable is (A, Y) determined by three latent variable $(A, Y(0), Y(1)), A \in \{0, 1\}, Y = AY(1) + (1 - A)Y(0)$, consider two submodels:

- $\mathcal{P}_1$: $Y(1) - Y(0) = 0, Y(1) \perp A, Y(0) \perp A$
- $\mathcal{P}_2$: $Y(1) - Y(0) = Z \neq 0$, but $Y(1) \overset{d}{=} Y(0), Y(1) \perp A, Y(0) \perp A$.

The model 2 is not empty, take

$$Y(0), \varepsilon \sim \mathcal{N}(0, 1), Y(0) \perp \varepsilon, Z = -\frac{1}{2}Y(0) + \frac{\sqrt{3}}{2}\varepsilon$$

then $Y(1) \sim \mathcal{N}(0, 1)$ and $Y(1) \overset{d}{=} Y(0)$.

On the observed data level, we can not distinguish these two models since they both have the same conditional distribution of $Y|A$, therefore they are undistinguishable in modeling stage.

This example is the reason why the sharp null hypothesis can not be tested in randomized experiment, and also the joint distribution of $(Y(0), Y(1))$ is not identifiable.

[3] talk about the identification of joint distribution of potential outcomes under some assumptions.

5 On the cluster analysis

6 Causal Inference

6.1 Causal Example

6.1.1 Yule–Simpson's Paradox

6.1.1.1 Healty worker effect

6.1.2 Smoke cause lung cancer

6.1.2.1 Mediation analysis

6.1.2.2 Sensitivity analysis

6.2 Identification

6.2.1 G-formula

6.2.2 ATT ATU

6.2.3 Mediation

[4], [5] (X,D,M,Y), D and Y are binary, $X \in \mathbb{R}^p$ is high-dimensional covariates, $M \in \mathbb{R}$ is mediator, estimating the natural direct effect(NDE) and natural indirect effect(NIE) can be reduced to estimate $\theta(d, d') = \mathbb{E}[Y(d, M(d')]$ for $d, d' \in \{0, 1\}$.

The influence function of $\theta(d, d')$ is:

$$
\begin{aligned}
\mathrm{EIF}_{\theta(d,d')} &= \Phi(d, d') - \theta(d, d') \\
\Phi(d, d') &= \underbrace{\frac{\mathbb{1}\{D = d\} f(M|X, d')}{a(d|X) f(M|X, d)} (Y - \mu(X, d, M))}_{A(d,d')} \\
&\quad + \underbrace{\left(1 - \frac{\mathbb{1}\{D = d'\}}{a(d' \mid X)}\right) \int_{m \in \mathcal{M}} \mu(X, d, m) f(m|X, d') \, \mathrm{d}m}_{B(d,d')} \\
&\quad + \underbrace{\frac{\mathbb{1}(D = d')}{a(d'|X)} \mu(X, d, M)}_{C(d,d')}.
\end{aligned}
$$

Let's verify that $\mathbb{E}\Big[\mathrm{EIF}_{\theta_{d,d'}}\Big] = 0$.

$$
\begin{aligned}
\mathbb{E}[A(d, d')] &= \mathbb{E}[\#(X, M) \mathbb{1}\{D = d\}(Y - \mu(X, d, M))] \\
&= \mathbb{E}[\mathbb{E}[\#(X, M) \mathbb{1}\{D = d\}(Y - \mu(X, d, M))| \ X, D = d, M]] \\
&= \mathbb{E}[\#(X, M) \mathbb{E}[\mathbb{1}\{D = d\}(Y - \mu(X, d, M))| \ X, D = d, M]] \\
&= \mathbb{E}\left[\#(X, M) \left(\int_D \mathbb{1}\{D = d\} \int_Y Y \frac{p(X, D, M, Y)}{p(X, D, M)} \, \mathrm{d}D \, \mathrm{d}Y - \mu(X, d, M)\right)\right] \\
&= \mathbb{E}[\#(X, M) \cdot 0] \\
&= 0.
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[B(d,d')] &= \mathbb{E}\left[\left(1-\frac{\mathbb{1}\{D=d'\}}{a(d' \mid X)}\right)\#(X)\right] \\
&= \mathbb{E}\left[\left(1-\frac{\mathbb{1}\{D=d'\}}{a(d' \mid X)}\right)\#(X) \mid X\right] \\
&= \mathbb{E}\left[\#(X)\mathbb{E}\left[\left(1-\frac{\mathbb{1}\{D=d'\}}{a(d' \mid X)}\right) \mid X\right]\right] \\
&= \mathbb{E}\left[\#(X)\left(1-\frac{1}{a(d' \mid X)}\int_D \mathbb{1}\{D=d'\}\frac{p(D,X)}{p(X)}\right)\right] \\
&= \mathbb{E}\left[\#(X)\left(1-\frac{a(d' \mid X)}{a(d' \mid X)}\right)\right] \\
&= 0. \\
\mathbb{E}[C(d,d')] &= \mathbb{E}\left[\frac{\mathbb{1}(D=d')}{a(d'|X)}\mu(X,d,M)\right] \\
&= \int_{X,D,M} \frac{\mathbb{1}(D=d')}{a(d'|X)}\mu(X,d,M)p(X,D,M)\,\mathrm{d}(X,D,M) \\
&= \int_{X,M} \frac{1}{a(d'|X)}\mu(X,d,M)p(X,d',M)\,\mathrm{d}(X,M) \\
&= \int_{X,M} \mu(X,d,M)f(M|X,d')p(X)\,\mathrm{d}(X,M) \\
&= \theta(d,d').
\end{aligned}
$$

6.2.4 Joint distribution of potential outcomes

[3] consider a case with multiple randomized controlled trials(RCTs), where data are (G, A, Y), G is the indicator of RCTs, A is the treatment, Y is the outcome.

Under consistency, positivity, and exchangeability, Adding one assumption called "transportability":

$$Y(1) \perp G \mid Y(0)$$

We can then identify the conditional distribution $Y(1) \mid Y(0)$.

$$
\begin{aligned}
\Pr(Y(1)=b \mid G=g) &= \sum_a \Pr(Y(1)=b, Y(0)=a \mid G=g)\Pr(Y(0)=a \mid G=g) \\
&= \sum_a \Pr(Y(1)=b \mid Y(0)=a)\Pr(Y(0)=a \mid G=g)
\end{aligned}
$$

Here $\Pr(Y(1)=b \mid G=g)$ and $\Pr(Y(0)=a \mid G=g)$ can be identified form data by the consistency, positivity and unconfounder assumption, using them to solve the above equation system, we can identify $\Pr(Y(1)=b \mid Y(0)=a)$.

6.2.5 Instrumental variable

[6]

- data are (X, A, Z, Y) only assume consistency, positivity, unconfounder, exculusion, no monotonicity, provide a bound estimation on ATE.
- The identification assumption of IV:

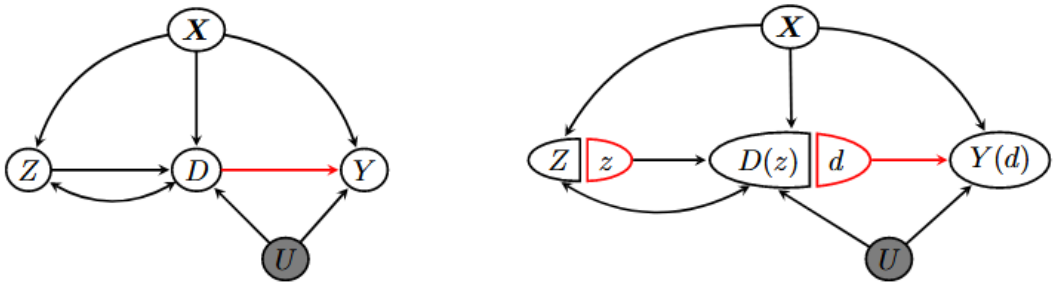
"Critically, under the four assumptions introduced in the previous section, the ATE is not point identified. Analysts typically take one of two approaches for point identification. The first approach invokes some type of homogeneity assumptions and places various restrictions

on how the effects of A and Z vary from unit to unit in the study population. See Hernan and Robins (2019) and Wang and Tchetgen Tchetgen (2018) for prominent examples. However, homogeneity assumptions are often implausible or difficult to verify in specific applications. The second approach invokes an assumption known as monotonicity, which has the following form: $A(z = 1) \geq A(z = 0)$, i.e., if $A(z = 0) = 1$ then $A(z = 1) = 1$ (Imbens and Angrist, 1994). Under monotonicity, the target estimand is no longer the ATE, but instead is the local average treatment effect (LATE):"

- The lower bound and upper bound is not a differentiable functional, thus an assumption is invoked to make the bound functional differentiable and thus have inference function to faster convergence rate.

[2]

- talk about the indenfication of ATE with continuous or multiple-category IVs with binary treatment.
- data are (X, D, Z, Y)



(a) A DAG with a bi-directed arrow. (b) A SWIG with a bi-directed arrow.

Figure 1: Causal graphs for an IV model defined by Assumptions 1′ and 2–4. The bi-directed arrow between Z and D represents potential unmeasured common causes; $\boldsymbol{X}$, Z, D, and Y are observed, and U is unobserved. (a) Causal directed acyclic graph (DAG) (Pearl, 2009) with a bi-directed arrow; (b) Single-world intervention graph (SWIG) (Richardson and Robins, 2013).

- The identification assumption:
  (i) Stable Unit Treatment Value Assumption (SUTVA) for potential outcomes:
    - Consistency and no interference between units:

$$Y = Y(D) = DY(1) + (1 - D)Y(0)$$
$$D = D(Z)$$

  (ii) IV relevance (version 1): $Z \not\perp D \mid X$ almost surely.
  (iii) IV independence : $Z \perp U \mid X$
  (iv) IV exclusion restriction : $Z \perp Y \mid D, X$
  (v) Unconfounderness/d-separation : $(Z, D) \perp Y(d) \mid X, U$ for $d = 0, 1$

As [6] mentioned, under these assumptions, the ATE is not point identified, homogeneity assumptions are :

(i) Version 1, for binary Z : Either

$$\mathbb{E}[D \mid Z = 1, X, U] - \mathbb{E}[D \mid Z = 0, X, U]$$

or

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

does not depend on U.

- "Assumption 5′ rules out additive effect modification by U of the $Z - D$ relationship or $d - Y(d)$ relationship within levels of X. A weaker alternative is the no unmeasured common effect modifier assumption (Cui and Tchetgen Tchetgen, 2021, Hartwig et al., 2023), which stipulates that no unmeasured confounder acts as a common effect modifier of both the additive effect of the IV on the treatment and the additive treatment effect on the outcome:"

(ii) Version 2, weaker alternative for binary Z, following equation holds almost surely:

$$\text{Cov}(\mathbb{E}(D|\ Z = 1, X, U) - \mathbb{E}(D|Z = 0, X, U), \mathbb{E}(Y(1) - Y(0) \mid X, U) \mid X) = 0$$

(iii) Final version, for continuous or multiple-category Z, for any z in the support of Z, following equation holds almost surely:

$$\text{Cov}(\mathbb{E}(D|\ Z = z, X, U) - \mathbb{E}(D|\ X, U), \mathbb{E}(Y(1) - Y(0) \mid X, U) \mid X) = 0$$

for any z, z' in the support of Z.

- The real-data applicationis combine many genetic variants as weak IVs to a strong and continuous IV to solve the "obesity paradox" in oncology.
  - "Obesity is typically associated with poorer oncology outcomes. Paradoxically, however, many observational studies have reported that non-small cell lung cancer (NSCLC) patients with higher body mass index (BMI) experience lower mortality, a phenomenon often referred to as the "obesity paradox" (Zhang et al., 2017)."

- Using the ratio of conditional weighted average treatment effect, for multiple-category (CWATE) or conditional weighted average derivative effect (CWADE) to identify the ATE.
  - Using semiparametric theory to provide the efficient influence function and build a triply robust estimator.
  - The tangent space is strange such that second-order parametric submodels are needed to validate the efficient influence function.

[7]

6.3 The equivalence between DAG and potential outcome framework

[8]

6.3.1 The equivalence between nonparametric structural equation model(NPSEM) and potential outcome framework

6.3.2 The equivalence between SWIG and FFRCISTG

6.4 ADMG : the inequility constraints

6.4.1 Bell inequality

7 Semiparametric theory

7.1 Parametric theory

7.1.1 Efficiency

7.1.1.1 Convolution theorem

7.2 nonparametric theory

7.3 Regularity

7.4 Influence function

7.4.1 The smoothness needed for influence function

Formulation in [9]

7.4.2 Numerical calcaulation of influence function

[10], [11], [12]

Automatic differentiation is remarkable! By building in all basic differentiable operations such as addition, multiplication, sine, and exponential functions, it calculates every derivative value and uses the chain rule to combine them, yielding derivatives equivalent to those computed by exact formulas. [13], [14]

A conversation with Qwen. The delta method used to estimate the variance of coefficients in our work [15] requires computing numerical derivatives of a complex mapping involving nonlinear function solving and integration. This can be made differentiable using "AD"-friendly functions in the implementation. Therefore, our R code could be replaced with more powerful Python code.

7.4.3 Von mise representation

7.4.4 Tangent space

S8 in [16]

7.4.4.1 Nuisance tangent space

7.5 Neyman orthogonality

[17] used a little different neyman orthogonality. Their problem can be summarized by following:

When the model is $X \sim \mathbb{P}_{\theta,\overline{\eta}}$ where $\overline{\eta}$ is the (nuisance) parameter and θ is the finite dimensional parameter of interest and $\theta = R(\overline{\eta}) = \max_{\eta} R(\eta)$ where $R(\eta) = \mathbb{E}_X L(X;\eta)$ and L is a loss function.

$$\theta = \mathbb{E}_X L(X;\overline{\eta}) = \max_{\eta} \mathbb{E}_X L(X;\eta)$$

Then $\psi(X;\eta) := L(X;\eta)$ naturally satisfies that the Gâteaux derivative of η is always zero in $\overline{\eta}$:

$$\frac{\partial \mathbb{E}_X[\psi(X;\eta_0 + t(\eta - \eta_0))]}{\partial t} \mid_{t=0} = 0, \forall \eta.$$

The parameterization need to check, if in above setting, θ is totally determined by $\overline{\eta}$.

Their paper mentioned the indenfication of θ, need to check.

7.5.1 Higher order influence function

8 U statistics

8.1 Computation

8.1.1 Motif counts

- For GPU:
  - Gunrock [18]
  - Cugraph [19]
- For CPU:
  - Peregrine [20]
  - Automine [21]

8.2 Occurrence

8.2.1 Energy function in N particle system

The interaction energy function in N particle system can be written as a U statistic.

9 Applications

9.1 Umbrella review

See `Belief in the law of small numbers as a way to understand the replication crisis and silly researchers who continue to cite discredited behavioral research` by Andrew Gelman, but I just see the paper involved a title "umbrella review" [22], so I search more about umbrella review.

Umbrella review consider the evidence form multiple meta-analysis and system review studies on the same topic.

using [23](10K+ citation!) to assess the quality of included meta-analysis/system review studies, but they did not provide a way to aggregate the quality scores and just give a subjective criteria, the overall confidence is rating as "high", "moderate", "low", "critically low". Produce an conclusion such that "In the 60 meta-analyses/systematic reviews included on the treatment for type 2 diabetes, treatments A have 5 strong evidence to support its effectiveness, while treatment B has only 1 moderate evidence to support its effectiveness…"

10 multiple testing

10.1 Varibale selection

A talk given by Zhong Wei: `A unified stability approach to false discovery rate control`

Knockoff is a randomized method for variable selection controlling false discovery rate(FDR) [24]. The key idea is to construct knockoff variables $\tilde{X} \in R^p$ that mimic the correlation structure of the original variables $X \in R^p$, such that for any subset $S \subset \{1, ..., p\}$, swapping the variables

in S with their knockoffs does not change the joint distribution of $(X, \tilde{X})$, while also ensuring that the knockoff variables are conditionally independent of the response variable Y given the original variables X.

randomized method have a problem in reproduciblity, e-value developed by [25] is a way to solve this problem, the e-value is a function of data that has an expected value at most 1 under the null hypothesis, thus can be used to control type I error in multiple testing, it can combine evidence from multiple independent tests easily, by running the knockoff procedure multiple times and averaging the resulting e-values for each variable, we can obtain a more stable measure of evidence against the null hypothesis for each variable, that's the work in [26].

Zhong Wei proposed a general framework of stability for FDR control, which includes derandomized knockoff as a special case, and provide theoretical guarantee for FDR control under this framework.

Their another work is consider the inference after variable selection, using the same sample for variable selection and inference will lead to selection bias [27].

***Question* 10.1.** A quesition is deose there have some minimax optimal method or other criteria for variable selection controlling FDR? e-value BH method or p-value BH method? or here derandomized knockoff BH method? Which is best?

Rina Foygel Barber 是 2025 年新任美国科学院院士，今年统计方向的两位美国科学院（National Academy of Sciences）院士，另一位是刘军，Rina Foygel Barber 是 2020 年 COPSS 总统奖（12 年 phd 毕业 at U of Chicago，postdoc under Emmanuel Candès）获得者，颁奖理由是：

"For fundamental contributions to statistical sparsity and selective inference in high-dimensional problems, for the creative and novel knockoff filter to cope with correlated coefficients, for contributions to compressed sensing, the jackknife, and conformal predictive inference; for the encouragement and training of graduate and undergraduate students."

e-valuede 的王若度（U of Waterloo， Chair profess） 曾经是星际争霸职业选手。

11 Machine Learning

11.1 Why named black box model?

Introduction of *Limitations of ML Interpretability* give a good review to ML Interpretability. Black box model is because the model is based algorithm not as generalized linear model have a simple representation. As [28] saying, two cultures of modeling.

11.2 Varibale importance

11.2.1 Leave-One-Covariate-Out(LOCO)

[29] give a measure named loco:

$$I_x = l(y, f(x, z)) - l(y, f(z))$$

to measure the importance of variable x by comparing the loss when including x versus excluding x.

[17] proposed an extension of LOCO under multiple source data and using semiparametric theory to provide the inference of their measure.

11.3 Ensemble learning

Ensemble learning, as the name suggests, combines multiple basde models to improve prediction performance. Common ensemble learning methods include bagging, boosting, and stacking.

11.3.1 Bagging

BoostrapAggregating (bagging) is proposed by Leo Breiman [30] (4 w + citations). The key idea is to generate multiple boostrap samples from the original data and train the base model on each boostrap sample then aggragate the predictions from all models, such as by averaging for regression or majority voting for classification.

11.3.1.1 Theory properties of bagging

- [30] Bagging can reduce the variance of unstable base models.
- Peter Bühlmann and Bin Yu [31] (1.2 k + citations) give some convergence rate analysis for bagging.
- In the 2000s, many works on the theoretical understanding of bagging, seems no more work needed now.

11.3.1.2 Random forest

- Leo Breiman [32] (17 w + citations) proposed random forest, an ensemble learning method that builds multiple decision trees and merges their results to improve accuracy.

11.3.1.3 Causal forest

- [33](2.5 k + citations) proposed causal tree to estimate heterogeneous treatment effects namely the conditional average treatment effect (CATE) by extending decision tree, then [34](4.3 k + citations) using random forest to improve the estimation accuracy and provide asymptotic normality for inference.
- [35] establishes an inconsistency lower bound on the point wise convergence rate of causal tree, and challenges the α-regularity condition (each split leaves at least a fraction α of available samples on each side) needed to establish the convergence rate in [34].

11.3.2 Boosting

- XGBoost

11.3.3 Stacking

Bibliography

[1] E. P. Wigner and others, "The unreasonable effectiveness of mathematics in the natural sciences," *Mathematics and science*, vol. 13, pp. 1–14, 1990.

[2] M. Dong, L. Liu, D. Tang, G. Liu, W. Xu, and L. Wang, "Marginal Causal Effect Estimation with Continuous Instrumental Variables," *arXiv preprint arXiv:2510.14368*, 2025.

[3] P. Wu and X. Mao, "The Promises of Multiple Experiments: Identifying Joint Distribution of Potential Outcomes," *arXiv preprint arXiv:2504.20470*, 2025.

[4] S. Xu, L. Liu, and Z. Liu, "DeepMed: Semiparametric causal mediation analysis with debiased deep learning," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022, pp. 28238–28251.

[5] E. J. Tchetgen Tchetgen and I. Shpitser, "Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis," *The Annals of Statistics*, vol. 40, no. 3, pp. 1816–1845, 2012.

[6] A. W. Levis, M. Bonvini, Z. Zeng, L. Keele, and E. H. Kennedy, "Covariate-assisted bounds on causal effects with instrumental variables," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, p. qkaf28, 2025.

[7] S. Chen, P. Zhang, and Y. Cui, "Identification and Debiased Learning of Causal Effects with General Instrumental Variables," *arXiv preprint arXiv:2510.20404*, 2025.

[8] L. Wang, T. Richardson, and J. Robins, "Causal Inference: A Tale of Three Frameworks," *arXiv preprint arXiv:2511.21516*, 2025.

[9] A. van der Vaart, "On differentiable functionals," *The Annals of Statistics*, vol. 19, no. 1, pp. 178–204, 1991.

[10] Y. Mukhin, "Kernel von Mises Formula of the Influence Function," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*,

[11] M. Jordan, Y. Wang, and A. Zhou, "Empirical Gateaux derivatives for causal inference," 2022, pp. 8512–8525.

[12] R. Agrawal, S. Witty, A. Zane, and E. Bingham, "Automated efficient estimation using monte carlo efficient influence functions," *Advances in Neural Information Processing Systems*, vol. 37, pp. 16102–16132, 2024.

[13] A. Paszke *et al.*, "Automatic differentiation in pytorch," 2017.

[14] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind, "Automatic differentiation in machine learning: a survey," *Journal of machine learning research*, vol. 18, no. 153, pp. 1–43, 2018.

[15] X. Chen, L. Liu, and R. Mukherjee, "Method-of-Moments Inference for GLMs and Doubly-Robust Functionals under Proportional Asymptotics," *arXiv preprint arXiv:2408.06103*, 2024.

[16] E. Graham, M. Carone, and A. Rotnitzky, "Towards a Unified Theory for Semiparametric Data Fusion with Individual-Level Data," *arXiv preprint arXiv:2409.09973*, 2024.

[17] Z. Wang, N. Si, Z. Guo, and M. Liu, "Multi-source stable variable importance measure via adversarial machine learning," *arXiv preprint arXiv:2409.07380*, 2024.

[18] Y. Wang, A. Davidson, Y. Pan, Y. Wu, A. Riffel, and J. D. Owens, "Gunrock: A high-performance graph processing library on the GPU," in *Proceedings of the 21st ACM SIGPLAN symposium on principles and practice of parallel programming*, 2016, pp. 1–12.

[19] A. Fender, B. Rees, and J. Eaton, "Rapids cugraph," *Massive Graph Analytics*. Chapman, Hall/CRC, pp. 483–493, 2022.

[20] K. Jamshidi, R. Mahadasa, and K. Vora, "Peregrine: a pattern-aware graph mining system," in *Proceedings of the Fifteenth European Conference on Computer Systems*, 2020, pp. 1–16.

[21] D. Mawhirter and B. Wu, "Automine: harmonizing high-level abstraction and high performance for graph mining," in *Proceedings of the 27th ACM Symposium on Operating Systems Principles*, 2019, pp. 509–523.

[22] H. Lin, M. D. de Barcellos, and H. De Steur, "The role of nudges in food choices: An umbrella review," *Food Quality and Preference*, p. 105679, 2025.

[23] B. J. Shea *et al.*, "AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both," *bmj*, vol. 358, 2017.

[24] R. F. Barber and E. J. Candès, "Controlling the false discovery rate via knockoffs," 2015.

[25] V. Vovk and R. Wang, "E-values: Calibration, combination and applications," *The Annals of Statistics*, vol. 49, no. 3, pp. 1736–1754, 2021.

[26] Z. Ren and R. F. Barber, "Derandomised knockoffs: leveraging e-values for false discovery rate control," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 86, no. 1, pp. 122–154, 2024.

[27] T. Zrnic and M. I. Jordan, "Post-selection inference via algorithmic stability," *The Annals of Statistics*, vol. 51, no. 4, pp. 1666–1691, 2023.

[28] L. Breiman, "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," *Statistical science*, vol. 16, no. 3, pp. 199–231, 2001.

[29] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1094–1111, 2018.

[30] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, pp. 123–140, 1996.

[31] P. Bühlmann and B. Yu, "Analyzing bagging," *The Annals of Statistics*, vol. 30, no. 4, pp. 927–961, 2002.

[32] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[33] S. Athey and G. Imbens, "Recursive partitioning for heterogeneous causal effects," *Proceedings of the National Academy of Sciences*, vol. 113, no. 27, pp. 7353–7360, 2016.

[34] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018.

[35] M. D. Cattaneo, J. M. Klusowski, and R. R. Yu, "The honest truth about causal trees: Accuracy limits for heterogeneous treatment effect estimation," *arXiv preprint arXiv:2509.11381*, 2025.