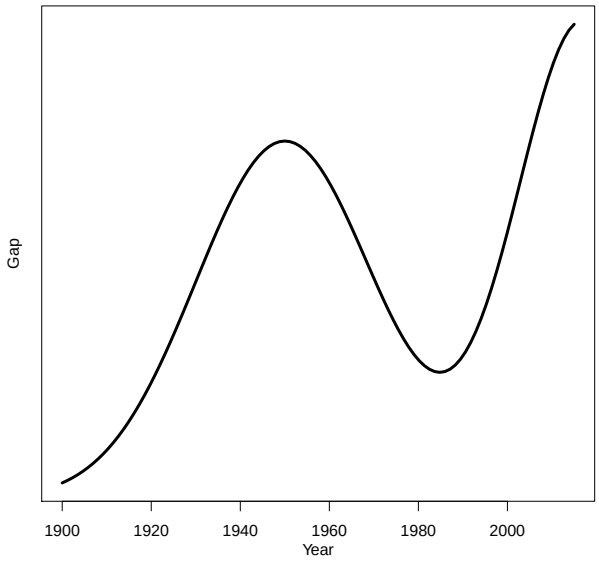
The Role of Assumptions in Machine Learning and Statistics: Don't Drink the Koolaid!

Larry Wasserman

April 12 2015

1 Introduction

There is a gap between the assumptions we make to prove that our methods work and the assumptions that are realistic in practice. This has always been the case, and the size of the gap varies with time. But, due to the ubiquity of high dimensional problems, the gap has become dangerously wide. It looks like this:



The problem is that when we prove theorems about our methods, we make all kinds of assumptions. The assumptions are often highly implausible. The practitioners use the method and they have a false sense of security since some theoretician has proved that works. The theoretician is too detached from real data analysis to realize that his assumptions are bogus. The practitioner does not have the time or background to look closely and see that the assumptions are bogus.

Generally speaking, pure prediction problems don't require many assumptions. But inference does. Inference refers to: estimating parameters, confidence intervals, assessing uncertainty, support recovery, etc.

2 Empirical Indistinguishably and Fragility

We will say that an assumption is *fragile* if:

> the assumption holds for P but the assumption fails for Q where Q is empirically indistinguishable from P.

To make this precise, we need to define what we mean by empirically indistinguishable. Let P be a distribution. We will say that Q is *empirically indistinguishable* from P if there is no reliable hypothesis test for

$$H_0 : X_1, \ldots, X_n \sim P \quad \text{versus} \quad H_1 : X_1, \ldots, X_n \sim Q.$$

More precisely, we have the following from Le Cam (see Donoho 1988). Recall that

$$\text{TV}(P, Q) = \sup_A |P(A) - Q(A)|.$$

Lemma 1 *Fix any small $\epsilon > 0$. Suppose that*

$$\text{TV}(P, Q) \leq 2[1 - (\epsilon^2/8)^{1/n}] \approx \frac{\epsilon^2}{4n}.$$

Then any test for H_0 versus H_1 with type I error α has power at most $\alpha + \epsilon/2$.

In other words, the set of distributions

$$N_n(P) = \left\{ Q : \ \text{TV} \leq \frac{\epsilon^2}{4n} \right\}$$

are indistinguishable from P.

Many of the theorems about the lasso, sparse estimation, matrix completion etc have the following unpleasant property which I call *fragility*. Some theorems show that

$$P^n(A_n) \to 1$$

where A_n is some good event such as: recover all the β_j's that are not 0. But the results are fragile because

$$\inf_{Q \in N_n(P)} Q^n(A_n) \to 0.$$

Examples of fragile assumptions are: linearity, sparsity, constant variance, and design conditions (incoherence, restricted eigenvalue assumptions etc).

3 Regression

This section is adapted from Wasserman (2014), "Discussion of paper by Lockhart, Taylor, Tibshirani and Tibshirani."

3.1 The Assumptions

The assumptions in most theoretical papers on high dimensional regression have several components. These include:

1. X is fixed.
2. The linear model is correct.
3. The variance is constant.
4. The errors have a Normal distribution.
5. The parameter vector is sparse.
6. The design matrix has very weak collinearity. This is usually stated in the form of incoherence, eigenvalue restrictions or incompatibility assumptions.

The first assumption actually makes inference more difficult (the data are no longer iid) and is a hangover from the olden days when regression was applied to designed experiments where X really was fixed. The other assumptions are: (i) not testable, (ii) not likely to hold in practice and (iii) fragile. The regression function $m(x) = \mathbb{E}(Y|X = x)$ can be any function. There is no reason to think it will be close to linear. Design assumptions are also highly suspect. High collinearity is the rule rather than the exception especially in high-dimensional problems. (An exception is signal processing, in particular compressed sensing, where the user gets to construct the design matrix. In this case, if the design matrix is filled with independent random Normals, the design matrix will be incoherent with high probability. But this is a rather special situation.)

3.2 The Assumption-Free Lasso

If we focus only on prediction, then the lasso has a very nice assumption-free interpretation. (Greenshtein and Ritov 2004 and Juditsky and Nemirovski 2000).

Suppose we observe $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$ where $Y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^d$. The regression function $m(x) = \mathbb{E}(Y|X = x)$ is some unknown, arbitrary function. We have no hope to estimate $m(x)$ when d is large nor do we have licence to impose assumptions on m.

Let $\mathcal{L} = \{x^t\beta : \ \beta \in \mathbb{R}^d\}$ be the set of linear predictors. For a given β, define the predictive

risk

$$R(\beta) = E(Y - \beta^T X)^2$$

where (X, Y) is a new pair. Let us define the best, sparse, linear predictor $\ell_*(x) = \beta_*^T x$ (in the ℓ_1 sense) where β_* minimizes $R(\beta)$ over the set $B(L) = \{\beta : ||\beta||_1 \leq L\}$. The lasso estimator $\widehat{\beta}$ minimizes the empirical risk $\widehat{R}(\beta) = \frac{1}{n}\sum_{i=1}^n (Y_i - \beta^T X_i)^2$ over $B(L)$. For simplicity, I'll assume that all the variables are bounded by $C < \infty$ (but this is not really needed). We make no other assumptions: no linearity, no design assumptions, and no models.

Theorem 2 *We have that*

$$R(\widehat{\beta}) \leq R(\beta_*) + \sqrt{\frac{8C^2L^4}{n}\log\left(\frac{2d^2}{\delta}\right)}$$

except on a set of probability at most δ.

Proof. Let $Z = (X, Y)$ let us re-define β as $\beta = (-1, \beta_1, \ldots, \beta_d)$. For simplicity we will assume that all the variables are bounded by a constant B although this is not necessary. Then

$$R(\beta) = \mathbb{E}(Y - \beta^T X)^2 = \beta^T \Sigma \beta$$

where $\Sigma = \mathrm{Var}(Z)$. The training error is

$$R_n(\beta) = \frac{1}{n}\sum_{i=1}^n (Y_i - \beta^T X_i)^2 = \beta^T \widehat{\Sigma} \beta$$

where where $\widehat{\Sigma}$ is the sample covariance matrix of $Z_1 = (X_1, Y_1), \ldots, Z_n = (X_n, Y_n)$. The lasso estimator is

$$\widehat{\beta} = \operatorname*{argmin}_{||\beta||_1 \leq L} R_n(\beta)$$

and the best sparse linear predictor is

$$\beta_* = \operatorname*{argmin}_{||\beta||_1 \leq L} R(\beta).$$

Now, for any β for which $||\beta||_1 \leq L$ we have

$$\begin{aligned} |R_n(\beta) - R(\beta)| &= |\beta^T \widehat{\Sigma} \beta - \beta^T \Sigma \beta| \\ &= |\beta^T (\widehat{\Sigma} - \Sigma)\beta| \leq \sum_{j,k} \beta_j \beta_k \, |\widehat{\Sigma}_{jk} - \Sigma_{jk}| \leq L^2 \, \Delta_n \end{aligned}$$

where $\Delta_n = \max_{j,k} |\widehat{\Sigma}_{jk} - \Sigma_{jk}|$. By Hoeffding's inequality and the union bound,

$$\mathbb{P}(\Delta_n > \epsilon) \leq 2d^2 e^{-n\epsilon^2/(2C^2)}.$$

It follows that, with probability at least $1-\delta$, $\Delta_n \leq \epsilon_n$ where

$$\epsilon_n = \sqrt{\frac{2C^2}{n}\log\left(\frac{2d}{\delta}\right)}$$

which is small as long as $d = o(e^n)$. It follows that, with probability at least $1-\delta$,

$$\sup_{||\beta||_1 \leq L} |R_n(\beta) - R(\beta)| \leq L^2\epsilon_n.$$

So, with probability at least $1-\delta$,

$$R(\beta_*) \leq R(\widehat{\beta}) \leq R_n(\widehat{\beta}) + \epsilon_n \leq R_n(\beta_*) + \epsilon_n \leq R(\beta_*) + 2\epsilon_n.$$

□

This shows that the predictive risk of the lasso comes close to the risk of the best sparse linear predictor. In my opinion, this explains why the lasso "works." The lasso gives us a predictor with a desirable property – sparsity – while being computationally tractable and it comes close to the risk of the best sparse linear predictor.

Remarks: For fixed L, this result shows that

$$R(\widehat{\beta}) - R(\beta_*) = O\left(\sqrt{\frac{\log d}{n}}\right).$$

With no further assumptions on the regression function or design matrix, this rate is optimal: it cannot be improved. A similar bound with the same rate applies to forward stepwise regression (Barron, Cohen, Dahmen, and DeVore 2008). Note that we have risk consistency, i.e. $R(\widehat{\beta}) - R(\beta_*) \xrightarrow{P} 0$ if $L = o(n^{1/4})$.

3.3 What Use Are Models?

When developing new methodology, I think it is useful to consider three different stages of development:

1. Constructing the method.
2. Interpreting the output of the method.
3. Studying the properties of the method.

I also think it is useful to distinguish two types of modeling. In *strong modeling*, the model is assumed to be true in all three stages. In *weak modeling*, the model is assumed to be true for stage 1 but not for stages 2 and 3. In other words, one can use a model to help construct a method. But one does not have to assume the model is true when it comes to interpretation or when studying the theoretical properties of the method.

3.4 Assumption Free Inference: Sample Splitting

If we want to do inference after model selection, we can do so without any assumptions if we use sample splitting.

The idea is to split the data into two halves. $\mathcal{D}_1$ and $\mathcal{D}_2$. For simplicity, assume that n is even so that each half has size $m = n/2$. From the first half $\mathcal{D}_1$ we select a subset of variables S. The method is agnostic about how the variable selection is done. It could be forward stepwise, lasso, elastic net, or anything else. The output of the first part of the analysis is the subset of predictors S and an estimator $\widehat{\beta} = (\widehat{\beta}_j : \ j \in S)$. The second half of the data $\mathcal{D}_2$ is used to provide distribution free inferences for the following questions:

1. **Risk:** What is the predictive risk of $\widehat{\beta}$?
2. **Variable Importance:** How much does each variable in S contribute to the predictive risk?
3. **Projection Parameters:** What is the best linear predictor using the variables in S?

We use $\mathcal{D}_1$ only to produce S and then construct the coefficients of the predictor from $\mathcal{D}_2$.

In more detail, let

$$R = \mathbb{E}|Y - X^T\widehat{\beta}|$$

where the randomness is over the new pair (X, Y); we are conditioning on $\mathcal{D}_1$. In the above equation, it is understood that $\widehat{\beta}_j = 0$ when $j \notin S$. The first question refers to producing a estimate and confidence interval for R (conditional on $\mathcal{D}_1$). The second question refers to inferring

$$R_j = \mathbb{E}|Y - X^T\widehat{\beta}_{(j)}| - \mathbb{E}|Y - X^T\widehat{\beta}|$$

for each $j \in S$, where $\beta_{(j)}$ is equal to $\widehat{\beta}$ except that $\widehat{\beta}_j$ is set to 0. Thus, R_j is the risk inflation by excluding X_j. The third question refers to inferring

$$\beta^* = \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}} \, \mathbb{E}(Y - X_S^T\beta)^2$$

the coefficient of the best linear predictor for the chosen model. We call β^* the *projection parameter*. Hence, $x^T\beta^*$ is the best linear approximation to $m(x)$ on the linear space spanned by the selected variables. The estimate is just the least squares estimator based on $\mathcal{D}_2$.

A consistent estimate of R is

$$\widehat{R} = \frac{1}{m}\sum_{i=1}^{m} \delta_i$$

where the sum is over $\mathcal{D}_2$, and $\delta_i = |Y_i - X_i^T\widehat{\beta}|$. An approximate $1 - \alpha$ confidence interval for R is $\widehat{R} \pm z_{\alpha/2}\, s/\sqrt{m}$ where s is the standard deviation of the δ_i's.

Inference By Sample Splitting

Input: data $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.

1. Randomly split the data into two halves $\mathcal{D}_1$ and $\mathcal{D}_2$.
2. Use $\mathcal{D}_1$ to select a subset of variables S. This can be forward stepwise, the lasso, or any other method.
3. Let $R = \mathbb{E}((Y - X^T\widehat{\beta})^2|\mathcal{D}_1)$ be the predictive risk of the selected model on a future pair (X, Y), conditional on $\mathcal{D}_1$.
4. Using $\mathcal{D}_2$ construct point estimates and confidence intervals for R, $(R_j : \widehat{\beta}_j \neq 0)$ and β_*.

Figure 1: The steps in Sample Splitting Inference.

The validity of this confidence interval is essentially distribution free. In fact, if want to be purely distribution free and avoid asymptotics, we could instead define R to be the median of the law of $|Y - X^T\widehat{\beta}|$. Then the order statistics of the δ_i's can be used in the usual way to get a finite sample, distribution free confidence interval for R. That is $R = [\delta_{(j)}, \delta_{(k)}]$ is a distribution free, finite sample confidence interval for R where $j \approx F^{-1}(\alpha/2)$, $k \approx F^{-1}(1 - \alpha/2)$ and F is the cdf of a Binomial(n,1/2) random variable.

Estimates and confidence intervals for R_j can be obtained from $e_1, \ldots, e_m$ where

$$e_i = |Y_i - X^T\widehat{\beta}_{(j)}| - |Y_i - X^T\widehat{\beta}|.$$

The steps are summarized in Figure 1.

To infer the projection parameters $\beta_*(S)$, we can simply use the bootstrap on $\mathcal{D}_2$. That, we approximate

$$F_n(t) = \mathbb{P}(\sqrt{n}||\widehat{\beta}(S) - \beta_*(S)||_\infty \leq t)$$

with

$$\widehat{F}_n(t) = \frac{1}{B}\sum_{j=1}^{B} I(\sqrt{n}||\widehat{\beta}_j^*(S) - \widehat{\beta}(S)||_\infty \leq t).$$

Here is an example using a dataset about wine. Using the first half of the data we applied forward stepwise selection and used C_p to select a model. The selected variables are Alcohol, Volatile-Acidity, Sulphates, Total-Sulfur-Dioxide and pH. A 95 percent confidence interval for the predictive risk of the null model is (0.65,0.70). For the selected model, the confidence interval for R is (0.46,0.53). The (Bonferroni-corrected) 95 percent confidence intervals for the R_j's are shown in the first plot of Figure 2. The (Bonferroni-corrected) 95 percent

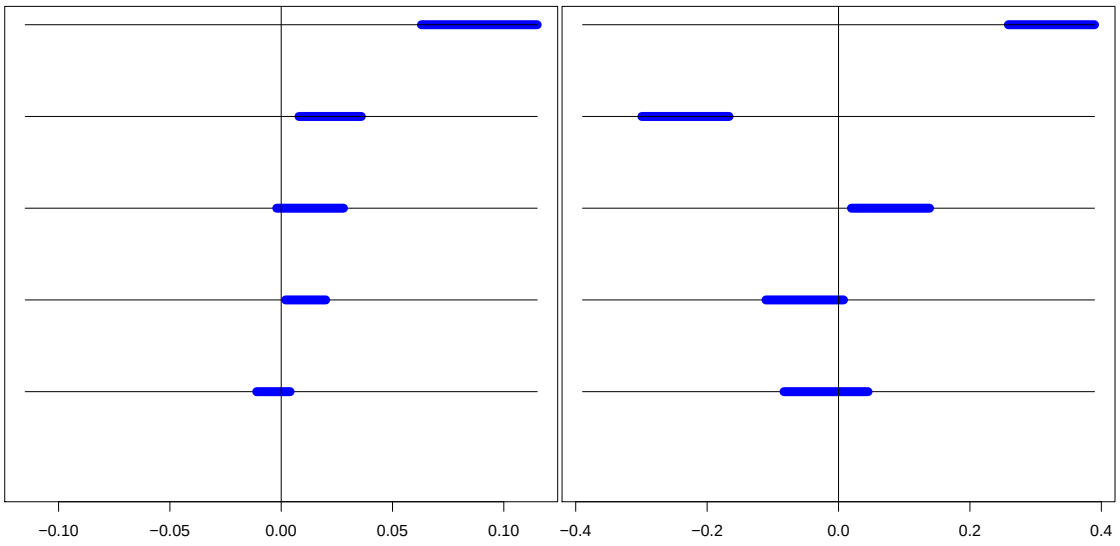


Figure 2: Left plot: Confidence intervals for R_j. Right plot: Confidence intervals for projected parameters. From top down the variables are Alcohol, Volatile-Acidity, Sulphates, Total-Sulfur-Dioxide and pH.

confidence intervals for the parameters of the projected model are shown in the second plot in Figure 2.

Uniform Control Method. An alternative to sample splitting is to use *uniform control*. Berk et al 2013 call this POSI. They still invoke many assumptions but there is a way to do it without assumptions. Let $\mathcal{S}$ be the set of possible models that your model selection method can choose. For example, if you use k steps of forward stepwise, $\mathcal{S}$ is all $\binom{d}{k}$ models of size k. Then we can estimate

$$F(t) = \mathbb{P}(\sqrt{n}\sup_{S\in\mathcal{S}} ||\widehat{\beta}(S) - \beta_*(S)||_\infty \leq t)$$

with the bootstrap. A valid confidence set is $\widehat{\beta} \pm t_\alpha$ where $t_\alpha = F^{-1}(1-\alpha)$. This provides valid, nonparametric inference with splitting. But there are two problems. First, we need to compute $\widehat{\beta}(S)$ for every S. This is computationally infeasible. Second, it leads to large confidence sets.

Theorem 3 *Consider a model selection procedure that chooses k variables. Sample splitting leads to a confidence set of size $O(\sqrt{\log k/n})$ and coverage $1-\alpha-O(\sqrt{\log k/n})$. The uniform method leads to a confidence set of size $O(\sqrt{k\log d/n})$ and coverage $1-\alpha-O(\sqrt{k\log d/n})$.*

Thus, splitting has smaller confidence sets, better accuracy, is fast, and can be applied to any type of regression: linear, nonlinear, nonparametric etc.

Some people are bothered by the fact that the analysis depends on a random split. To assess the sensitivity to the choice of split, one can do many splits and take the union of the confidence intervals. I call this a meta-confidence interval. This met interval will be wide if the model selection procedure is sensitive to the split.

3.5 Conformal Prediction

Recall that we discussed something called conformal prediction. This is a good example of assumption free inference.

Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$ suppose we observe a new X and want to predict Y. Let $y \in \mathbb{R}$ be an arbitrary real number. Think of y as a tentative guess at Y. Form the augmented data set

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X, y).$$

Now we fit a linear model to the augmented data and compute residuals e_i for each of the $n+1$ observations. Now we test $H_0 : Y = y$. Under H_0, the residuals are invariant under permutations and so

$$p(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I(|e_i| \geq |e_{n+1}|)$$

is a distribution-free p-value for H_0.

Next we invert the test: let $C = \{y : \ p(y) \geq \alpha\}$. It is easy to show that

$$\mathbb{P}(Y \in C) \geq 1 - \alpha.$$

Thus C is distribution-free, finite-sample prediction interval for Y. The validity of the method does not depend on the linear model being correct. The set C has the desired coverage probability no matter what the true model is.

One can also look at how the prediction interval C changes as different variables are removed. This gives another assumption free method to explore the effects of predictors in regression. Minimizing the length of the interval over the lasso path can also be used as a distribution-free method for choosing the regularization parameter of the lasso.

4 Some Technical Details About Sparse Estimators

Sparse Estimators Have Poor Risk. Say that $\widehat{\beta}$ is **weakly sparsistent** if, for every β,

$$P_\beta\big(I(\widehat{\beta}_j = 1) \leq I(\beta_j = 1) \ \text{ for all } j\big) \to 1 \tag{1}$$

as $n \to \infty$. In particular, if $\widehat{\beta}_n$ is sparsistent, then it is weakly sparsistent. Suppose that d is fixed. Then the least squares estimator $\widehat{\beta}_n$ is minimax and satisfies

$$\sup_{\beta} E_\beta(n||\widehat{\beta}_n - \beta||^2) = O(1). \tag{2}$$

But sparsistent estimators have much larger risk:

Theorem 4 (Leeb and Pötscher (2007)) *Suppose that the following conditions hold:*

1. *d is fixed.*
2. *The covariates are nonstochastic and $n^{-1}\mathbb{X}^T\mathbb{X} \to Q$ for some positive definite matrix Q.*
3. *The errors ϵ_i are independent with mean 0, finite variance σ^2 and have a density f satisfying*
$$0 < \int \left(\frac{f'(x)}{f(x)}\right)^2 f(x)dx < \infty.$$

If $\widehat{\beta}$ is weakly sparsistent then

$$\sup_{\beta} E_\beta(n||\widehat{\beta}_n - \beta||^2) \to \infty. \tag{3}$$

More generally, if ℓ is any nonnegative loss function then

$$\sup_{\beta} E_\beta(\ell(n^{1/2}(\widehat{\beta}_n - \beta))) \to \sup_s \ell(s). \tag{4}$$

It follows that, if R_n denotes the minimax risk then

$$\sup_{\beta} \frac{R(\widehat{\beta}_n)}{R_n} \to \infty.$$

Proof. Choose any $s \in \mathbb{R}^d$ and let $\beta_n = -s/\sqrt{n}$. Then,

$$\begin{aligned}\sup_{\beta} E_\beta(\ell(n^{1/2}(\widehat{\beta} - \beta)) &\geq E_{\beta_n}(\ell(n^{1/2}(\widehat{\beta} - \beta)) \geq E_{\beta_n}(\ell(n^{1/2}(\widehat{\beta} - \beta))I(\widehat{\beta} = 0)) \\ &= \ell(-\sqrt{n}\beta_n)P_{\beta_n}(\widehat{\beta} = 0) = \ell(s)P_{\beta_n}(\widehat{\beta} = 0).\end{aligned}$$

Now, $P_0(\widehat{\beta} = 0) \to 1$ by assumption. By contiguity, we also have that $P_{\beta_n}(\widehat{\beta} = 0) \to 1$. Hence, with probability tending to 1,

$$\sup_{\beta} E_\beta(\ell(n^{1/2}(\widehat{\beta} - \beta)) \geq \ell(s).$$

Since s was arbitrary the result follows. $\square$

Fragilty of Using the Same Data for Model Selection and Inference. Let $X_1, \ldots, X_n \sim N(\theta, 1)$. We choose between two models:

1. Model 0: $\{N(0,1)\}$.
2. Model 1: $(N(\theta,1);\ \theta \in \mathbb{R})$.

Let $S = \{0,1\}$ index the models. The natural model selector is $I(|\overline{X}| > c_n)$ for some c_n. To make the calculations simpler, suppose we know that $\theta \geq 0$. Then we can use the simpler model selector $S = I(\overline{X} > c_n)$.

The estimate of θ is

$$\widehat{\theta} = \begin{cases} 0 & \text{if } \overline{X}_n < c_n \\ \overline{X}_n & \text{if } \overline{X}_n > c_n. \end{cases}$$

The parameter we are interested in is

$$\psi_n(\theta) = \mathbb{P}(\sqrt{n}(\widehat{\theta}_n - \theta) \leq t)$$

where t is any fixed number.

Claim: There is no uniformly consistent estimator of $\psi_n(\theta)$.

Proof: Let $Z \sim N(0,1)$. We have

$$\begin{aligned} \psi_n(\theta) &= \mathbb{P}(\sqrt{n}(\widehat{\theta}_n - \theta) \leq t) \\ &= \mathbb{P}(\sqrt{n}(0-\theta) \leq t, \overline{X}_n < c_n) + \mathbb{P}(\sqrt{n}(\overline{X}_n - \theta) \leq t, \overline{X}_n > c_n) \\ &= I(-\sqrt{n}\theta \leq t)\mathbb{P}(\overline{X}_n < c_n) + \mathbb{P}(Z \leq t, \overline{X}_n > c_n) \\ &= I(-\sqrt{n}\theta \leq t)\mathbb{P}(Z < \sqrt{n}(c_n - \theta)) + \mathbb{P}(\sqrt{n}(c_n - \theta) \leq Z \leq t) \\ &= I(-\sqrt{n}\theta \leq t)\Phi(\sqrt{n}(c_n - \theta)) + \Phi(t) - \Phi(\sqrt{n}(c_n - \theta)). \end{aligned}$$

Let $\theta_n = a/\sqrt{n}$. Then

$$\begin{aligned} \psi_n(0) &= I(t \geq 0)\Phi(\sqrt{n}c_n) + \Phi(t) - \Phi(\sqrt{n}c_n) \\ \psi_n(\theta_n) &= I(t \geq -a)\Phi(\sqrt{n}c_n - a) + \Phi(t) - \Phi(\sqrt{n}c_n - a). \end{aligned}$$

If $\sqrt{n}c_n$ converges to ∞ or $-\infty$, then there is no model selection, asymptotically. Let us thus suppose that $\sqrt{n}c_n \to c > 0$. Then

$$\begin{aligned} \psi_n(0) &\to I(t \geq 0)\Phi(c) + \Phi(t) - \Phi(c) && \equiv b(0) \\ \psi_n(\theta_n) &\to I(t \geq -a)\Phi(c-a) + \Phi(t) - \Phi(c-a) && \equiv b(a). \end{aligned}$$

Note that $b(0) \neq b(a)$. Suppose that $\widehat{\psi}_n$ is consistent. Then, when $\theta = 0$,

$$\widehat{\psi}_n \xrightarrow{P} b(0).$$

Now consider P_{θ_n}. Since $\theta_n - 0 = O(1/\sqrt{n})$, it follows that

$$P_0^n(A_n) \to 0 \quad \text{implies that } P_{\theta_n}^n(A_n) \to 0.$$

Hence, $P_{\theta_n}^n$ is contiguous to P_0^n. Therefore, by Le Cam's first lemma, $\widehat{\psi}_n$ has the same limit under θ_n as under $\theta = 0$. So, under P_{θ_n} we have

$$\widehat{\psi}_n \xrightarrow{P} b(0) \neq b(a)$$

and so

$$\widehat{\psi}_n(\theta_n) - \psi(\theta_n) \xrightarrow{P} b(0) - b(a) \neq 0$$

and hence is inconsistent. □

5 Individual Sequence Prediction

A good example of a method that makes weak assumptions is individual sequence prediction. The goal is to predict y_t from $y_1, \ldots, y_{t-1}$ with no assumptions on the sequence. [1] The data are not assumed to be iid; they are not even assumed to be random. This is a version of *online learning.* For simplicity assume that $y_t \in \{0, 1\}$.

Suppose we have a set of *prediction algorithms* (or *experts*):

$$\mathcal{F} = \{F_1, \ldots, F_N\}$$

Let $F_{j,t}$ is the prediction of algorithm j at time t based on $y^{t-1} = (y_1, \ldots, y_{t-1})$. At time t:

1. You see y^{t-1} and $(F_{1,t}, \ldots, F_{N,t})$.
2. You predict P_t.
3. y_t is revealed.
4. You suffer loss $\ell(P_t, y_t)$.

We will focus on the loss $\ell(p_t, y_t) = |p_t - y_t|$ but the theory works well for any convex loss. The *cumulative loss* is

$$L_j(y^n) = \frac{1}{n}\sum_{i=1}^{n} |F_{j,t} - y_t| \equiv \frac{1}{n} S_j(y^n)$$

[1] Reference: *Prediction, Learning, and Games.* Nicolò Cesa-Bianchi and Gábor Lugosi, 2006.

where $S_j(y^n) = \sum_{i=1}^n |F_{j,t} - y_t|$. The *maximum regret* is

$$R_n = \max_{y^t \in \{0,1\}^t} \left(L_P(y^n) - \min_j L_j(y^n) \right)$$

and the *minimax regret* is

$$V_n = \inf_P \max_{y^t \in \{0,1\}^t} \left(L_P(y^n) - \min_j L_j(y^n) \right).$$

Let $P_t(y^{t-1}) = \sum_{j=1}^N w_{j,t-1}\ F_{j,t}$ where

$$w_{j,t-1} = \frac{\exp\{-\gamma S_{j,t-1}\}}{Z_t}$$

and $Z_t = \sum_{j=1}^N \exp\{-\gamma S_{j,t-1}\}$. The w_j's are called *exponential weights*.

Theorem 5 *Let* $\gamma = \sqrt{8 \log N/n}$. *Then*

$$L_P(y^n) - \min_{1 \le j \le N} L_j(y^n) \le \sqrt{\frac{\log N}{2n}}.$$

Proof. The idea is to place upper and lower bounds on $\log\left(\frac{Z_{n+1}}{Z_1}\right)$ then solve for $L_P(y^n)$.

Upper bound: We have

$$\begin{aligned} \log\left(\frac{Z_{n+1}}{Z_1}\right) &= \log\left(\sum_{j=1}^N \exp\{-\gamma n L_{j,n}\}\right) - \log N \\ &\ge \log\left(\max_j \exp\{-\gamma n L_{j,n}\}\right) - \log N \\ &= -\gamma n \min_j L_{j,n} - \log N. \end{aligned} \tag{5}$$

Lower bound: Note that

$$\begin{aligned} \log\left(\frac{Z_{t+1}}{Z_t}\right) &= \log\left(\frac{\sum_{j=1}^N w_{j,t-1}\ e^{-\gamma|F_{j,t}-y_t|}}{\sum_{j=1}^N w_{j,t-1}}\right) \\ &= \log \mathbb{E}\left(e^{-\gamma|F_{j,t}-y_t|}\right). \end{aligned}$$

This is a formal expectation with respect to the distribution over j probability proportional to $e^{-\gamma|F_{j,t}-y_t|}$).

Recall Hoeffding's bound for mgf: if $a \leq X \leq b$

$$\log \mathbb{E}(e^{sX}) \leq s\mathbb{E}(X) + \frac{s^2(b-a)^2}{8}.$$

So:

$$\begin{aligned} \log \mathbb{E}\left(e^{-\gamma|F_{j,t}-y_t|}\right) &\leq -\gamma \mathbb{E}|F_{j,t} - y_t| + \frac{\gamma^2}{8} \\ &= -\gamma|\mathbb{E}F_{j,t} - y_t| + \frac{\gamma^2}{8} \\ &= -\gamma|P_t(y^{t-1}) - y_t| + \frac{\gamma^2}{8}. \end{aligned}$$

Summing over t:

$$\log\left(\frac{Z_{n+1}}{Z_1}\right) \leq -\gamma n L_P(y^n) + \frac{n\gamma^2}{8}. \tag{6}$$

Combining (5) and (6) we get

$$-\gamma n \min_j L_j(y^n) - \log N \leq \log\left(\frac{Z_{n+1}}{Z_1}\right) \leq -\gamma n L_P(y^n) + \frac{n\gamma^2}{8}.$$

Rearranging the terms we have:

$$L_P(y^n) \leq \min_j L_j(y^n) + \frac{\log N}{\gamma} + \frac{n\gamma}{8}.$$

Set $\gamma = \sqrt{8\log N/n}$ to get

$$L_P(y^n) - \min_{1\leq j\leq N} L_j(y^n) \leq \sqrt{\frac{\log N}{2n}}.$$

□

Theorem 6 *The above method is close to the minimax risk and hence is optimal.*

The result held for a specific time n. We can make the result uniform over time as follows. If we set $\gamma_t = \sqrt{8\log N/t}$ then we have:

$$L_P(y^n) \leq \min_j L_j(y^n) + \sqrt{\frac{1 + 12n\log N}{8}}$$

for all n and for all $y_1, y_2, \ldots, y_n$.

Now suppose that $\mathcal{F}$ is an infinite class. A set $\mathcal{G} = \{G_1, \ldots, G_N\}$ is an r-covering if, for every F and every y^n there is a G_j such that

$$\sum_{t=1}^{n} |F_t(y^{t-1}) - G_{j,t}(y^{t-1})| \leq r.$$

Let $N(r)$ denote the size of the smallest r-covering.

Theorem 7 (Cesa-Bianchi and Lugosi) *We have that*

$$V_n(\mathcal{F}) \leq \inf_{r>0} \left(\frac{r}{n} + \sqrt{\frac{\log N(r)}{2n}} \right)$$

Cesa-Bianchi and Lugosi also construct a predictor that nearly achieves the bound of the form

$$P_t = \sum_{k=1}^{\infty} a_k P_t^{(k)}$$

where $P_t^{(k)}$ is a predictor based on a finite subset of $\mathcal{F}$.

Using *batchification* it is possible to use online learning for non-online learning. Suppose we are given data: $(Z_1, \ldots, X_n)$ where $Z_i = (X_i, Y_i)$ and an arbitrary algorithm A that takes data and outputs classifier H. We used uniform convergence theory to analyze H but online methods provide an alternative analysis. [2] We apply A sequentially to get classifiers $H_0, H_1, \ldots, H_n$. Let

$$M_n = \frac{1}{n} \sum_{i=1}^{n} \ell(H_{t-1}(X_t), Y_t)$$

To choose a final classifier:

1. usual batch method: use the last one H_n
2. average: $\overline{H} = \frac{1}{n} \sum_{i=1}^{n} H_{t-1}$
3. selection: choose H_t to minimize

$$\frac{1}{t} \sum_{i=1}^{t} \ell(H_t(X_t), Y_t) + \sqrt{\frac{1}{2(n-t)} \log \left(\frac{n(n+1)}{\delta} \right)}$$

Analyzing H_n requires assumptions on A, uniform convergence etc. This is not needed for the other two methods.

[2] Reference: Cesa-Bianchi, Conconi and Gentile (2004).

Theorem 8 *If ℓ is convex:*

$$\mathbb{P}\left(R(\overline{H}) \geq M_n + \sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)}\right) \leq \delta.$$

For any ℓ,

$$\mathbb{P}\left(R(\widehat{H}) \geq M_n + \sqrt{\frac{36}{n}\log\left(\frac{2(n+1)}{\delta}\right)}\right) \leq \delta.$$

6 Causation

At the other end of the scale is causal inference which requires extremely strong assumptions unless we use randomized assignment.

Most of statistics and machine learning is concerned with prediction. A typical question is: what is a good prediction of Y given that I **observe** that $X = x$? Causation is concerned with questions of the form: what is a good prediction of Y given that I **set** $X = x$? The difference between passively observing $X = x$ and actively intervening and setting $X = x$ is significant and requires different techniques and, typically, much stranger assumptions. See Figure 3.

Consider this story. A mother notices that tall kids have a higher reading level than short kids. (This is because the tall kids are older.) The mother puts her small child on a device and stretches the child until he is tall. She is dismayed to find out that his reading level has not changed.

Te mother is correct that height and reading skill are **associated**. Put another way, you can use height to predict reading skill. But that does not imply that height *causes* reading skill. This is what statisticians mean when they say:

> **correlation is not causation.**

On the other hand, consider smoking and lung cancer. We know that smoking and lung cancer are associated. But we also believe that smoking causes lung cancer. In this case, we recognize that intervening and forcing someone to smoke does change his probability of getting lung cancer.

The difference between prediction (association/correlation) and causation is this: in prediction we are interested in

$$\mathbb{P}(Y \in A | X = x)$$

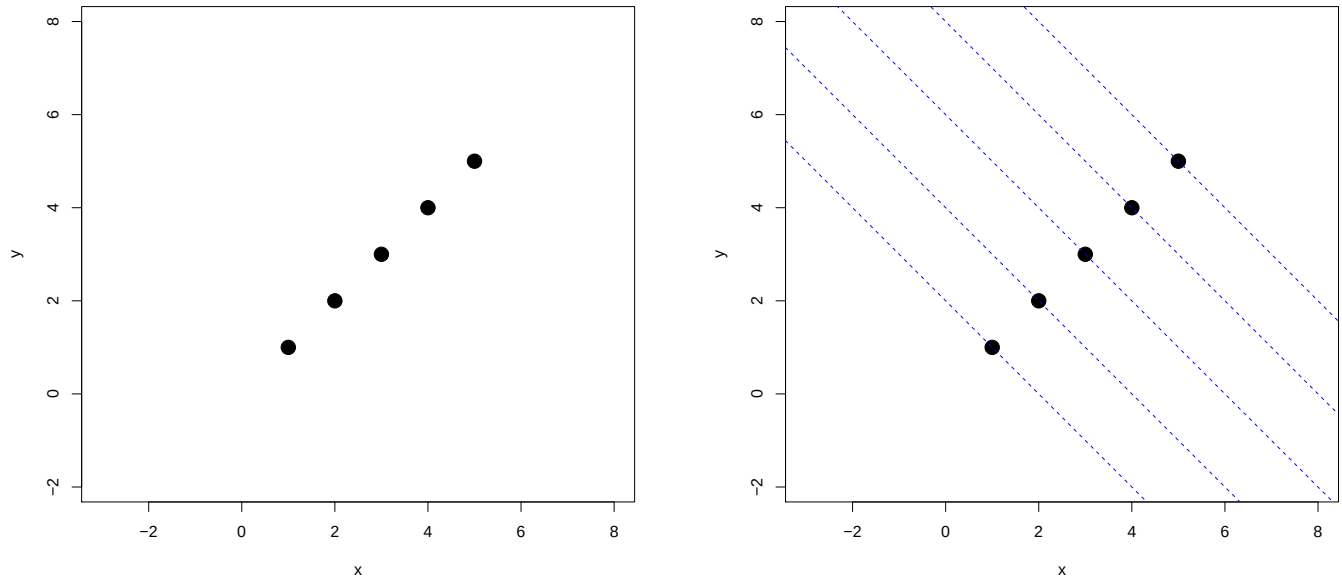


Figure 3: Left: Y is positively associated with X. Right: The lines show the counterfactual data: what would happen if we changed X. Increasing X makes Y decrease. This is the difference between association (left) and causation (right). This discrepancy won't happen if X is randomly assigned.

which means: the probability that $Y \in A$ given that we **observe** that X is equal to x. For causation we are interested in

$$\mathbb{P}(Y \in A | \mathsf{set}\ X = x)$$

which means: the probability that $Y \in A$ given that we **set** X equal to x. Prediction is about passive observation. Causation is about active intervention. Most of statistics and machine learning concerns prediction. But sometimes causation is the primary focus. The phrase **correlation is not causation** can be written mathematically as

$$\mathbb{P}(Y \in A | X = x) \neq \mathbb{P}(Y \in A | \mathsf{set}\ X = x).$$

Despite the fact that causation and association are different, people mix them up all the time, even people trained in statistics and machine learning. On TV recently there was a report that good health is associated with getting seven hours of sleep. So far so good. Then the reporter goes on to say that, therefore, everyone should strive to sleep exactly seven hours so they will be healthy. Wrong. That's confusing causation and association. Another TV report pointed out a correlation between people who brush their teeth regularly and low rates of heart disease. An interesting correlation. Then the reporter (a doctor in this case) went on to urge people to brush their teeth to save their hearts. Wrong!

To avoid this confusion we need a way to discuss causation mathematically. That is, we need someway to make $\mathbb{P}(Y \in A | \mathsf{set}\ X = x)$ formal. There are two common ways to do this. One is to use **counterfactuals**. The other is to use **causal graphs**. These approaches are equivalent. There are two different languages for saying the same thing.

Causal inference is tricky and should be used with great caution. The main messages are:

1. Causal effects can be estimated consistently from randomized experiments.
2. It is difficult to estimate causal effects from observational (non-randomized) experiments.
3. All causal conclusions from observational studies should be regarded as very tentative.

Causal inference is a vast topic. We will only touch on the main ideas here.

Counterfactuals. Consider two variables Y and X. Suppose that X is a binary variable that represents some treatment. For example, $X = 1$ means the subject was treated and $X = 0$ means the subject was given placebo. The response variable Y is real-valued.

We can address the problem of predicting Y from X by estimating $\mathbb{E}(Y|X = x)$. To address causal questions, we introduce *counterfactuals*. Let Y_1 denote the response we observe if the subject is treated, i.e. if we set $X = 1$. Let Y_0 denote the response we observe if the subject is not treated, i.e. if we set $X = 0$. If we treat a subject, we observe Y_1 but we do not observe Y_0. Indeed, Y_0 is the value we would have observed if the subject had been treated. The unobserved variable is called a counterfactual.

We have enlarged our set of variables from (X, Y) to (X, Y, Y_0, Y_1). Note that

$$Y = XY_1 + (1 - X)Y_0. \tag{7}$$

A small dataset might look like this:

X	Y	Y_0	Y_1
1	1	*	1
1	1	*	1
1	0	*	0
1	1	*	1
0	1	1	*
0	0	0	*
0	1	1	*
0	1	1	*

The asterisks indicate unobserved variables. To answer causal questions, we are interested in the distribution $p(y_0, y_1)$. We can interpret $p(y_1)$ as $p(y|\mathsf{set}\ X = 1)$ and we can interpret $p(y_0)$ as $p(y|\mathsf{set}\ X = 0)$. In particular, we might want to estimate the *mean treatment effect* or *mean causal effect*

$$\theta = \mathbb{E}(Y_1) - \mathbb{E}(Y_0) = \mathbb{E}(Y|\mathsf{set}\ X = 1) - \mathbb{E}(Y|\mathsf{set}\ X = 0).$$

The parameter θ has the following interpretation: θ is the mean response if we forced everyone to take the treatment minus mean response if we forced everyone not to take the treatment.

Suppose now that we observe a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. Can we estimate θ? No. In general, there is no consistent estimator of θ. We can estimate $\alpha = \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0)$ but α is not equal to θ.

However, suppose that we did a randomized experiment where we randomly assigned each person to treatment of placebo by the flip of a coin. In this case, X will be independent of (Y_0, Y_1). In symbols:

$$\text{random treatment assignment implies : } \ (Y_0, Y_1) \amalg X.$$

Hence, in this case,

$$\begin{aligned} \alpha &= \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0) \\ &= \mathbb{E}(Y_1|X = 1) - \mathbb{E}(Y_0|X = 0) \quad \text{since } Y = XY_1 + (1 - X)Y_0 \\ &= \mathbb{E}(Y_1) - \mathbb{E}(Y_0) = \theta \quad \text{since } (Y_0, Y_1) \amalg X. \end{aligned}$$

Hence, random assignment makes θ equal to α and α can be consistently estimated. **If X is randomly assigned then correlation = causation.** This is why people spend millions of dollars doing randomized experiments.

In some cases it is not feasible to do a randomized experiment. Smoking and lung cancer is an example. Can we estimate causal parameters from observational (non-randomized) studies? The answer is: sort of.

In an observational study, the treated and untreated groups will not be comparable. Maybe the healthy people chose to take the treatment and the unhealthy people didn't. In other words, X is not independent of (Y_0, Y_1). The treatment may have no effect but we would still see a strong association between Y and X. In other words, α might be large even though $\theta = 0$.

To account for the differences in the groups, we might measure confounding variables. These are the variables that affect both X and Y. By definition, there are no such variables in a randomized experiment. The hope is that if we measure enough confounding variables $Z = (Z_1, \ldots, Z_k)$, then, perhaps the treated and untreated groups will be comparable, conditional on Z. Formally, we hope that X is independent of (Y_0, Y_1) conditional on Z. If this is true, we can estimate θ since

$$\begin{aligned}
\theta &= \mathbb{E}(Y_1) - \mathbb{E}(Y_0) \\
&= \int \mathbb{E}(Y_1|Z=z)p(z)dz - \int \mathbb{E}(Y_0|Z=z)p(z)dz \\
&= \int \mathbb{E}(Y_1|X=1, Z=z)p(z)dz - \int \mathbb{E}(Y_0|X=0, Z=z)p(z)dz \\
&= \int \mathbb{E}(Y|X=1, Z=z)p(z)dz - \int \mathbb{E}(Y|X=0, Z=z)p(z)dz \qquad (8)
\end{aligned}$$

where we used the fact that X is independent of (Y_0, Y_1) conditional on Z in the third line and the fact that $Y = (1-X)Y_1 + XY_0$ in the fourth line. The latter quantity can be estimated by

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(1, Z_i) - \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(0, Z_i)$$

where $\widehat{m}(x, z)$ is an estimate of the regression function $m(x, z) = \mathbb{E}(Y|X=x, Z=z)$. This is known as *adjusting for confounders* and $\widehat{\theta}$ is called the *adjusted treatment effect*.

It is instructive to compare the casual effect

$$\begin{aligned}
\theta &= \mathbb{E}(Y_1) - \mathbb{E}(Y_0) \\
&= \int \mathbb{E}(Y|X=1, Z=z)p(z)dz - \int \mathbb{E}(Y|X=0, Z=z)p(z)dz
\end{aligned}$$

with the predictive quantity

$$\begin{aligned}
\alpha &= \mathbb{E}(Y|X=1) - \mathbb{E}(Y|X=0) \\
&= \int \mathbb{E}(Y|X=1, Z=z)p(z|X=1)dz - \int \mathbb{E}(Y|X=0, Z=z)p(z|X=0)dz
\end{aligned}$$

which are mathematically (and conceptually) quite different.

We need to treat $\widehat{\theta}$ cautiously. It is very unlikely that we have successfully measured all the relevant confounding variables so $\widehat{\theta}$ should be regarded as a crude approximation to θ at best.

Causal Graphs. Another way to capture the difference between $P(Y \in A|X = x)$ and $P(Y \in A|\mathsf{set}\ X = x)$ is to represent the distribution using a directed graph and then we capture the second statement by performing certain operations on the graph.

A Directed Acyclic Graph (DAG) is a graph for a set of variables with no cycles. The graph defines a set of distributions of the form

$$p(y_1, \ldots, y_k) = \prod p(y_j|\mathsf{parents}(y_j)$$

where $\mathsf{parents}(y_j)$ are the parents of y_j. A **causal graph** is a DAG with extra information. A DAG is a causal graph if it correctly encodes the effect of setting a variable to a fixed value.

Consider the graph G in Figure (4). Here, X denotes treatment, Y is response and Z is a confounding variable. To find the causal distribution $p(y|\mathsf{set}\ X = x)$ we do the following steps:

1. Form a new graph G_* by removing all arrow into X. Now set X equal to x. This corresponds to replacing the joint distribution $p(x, y, z) = p(z)p(x|z)p(y|x, z)$ with the new distribution $p_*(y, z) = p(z)p(y|x, z)$. The factor $p(x|z)$ is removed because we know regard x as a fixed number.
2. Compute the distribution of y from the new distribution:

$$p(y|\mathsf{set}\ X = x) \equiv p_*(y) = \int p_*(y, z)dz = \int p(z)p(y|x, z)dz.$$

Now we have that

$$\theta = p(y|\mathsf{set}\ X = 1) - p(y|\mathsf{set}\ X = 0) = \int p(z)p(y|1, z)dz - \int p(z)p(y|0, z)dz$$

This is precisely the same equation as (8). Both approaches lead to the same thing. If there were unobserved confounding variables, then the formula for θ would involve these variables and the causal effect would be non-estimable (as before).

In a randomized experiment, there would be no arrow from Z to X. (That's the point of randomization). In that case the above calculations shows that $\theta = \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0)$ just as we saw with the counterfactual approach.

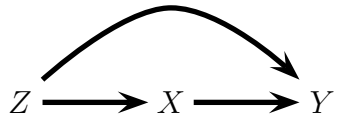


Figure 4: Conditioning versus intervening.

To understand the difference between $p(y|x)$ and $p(y|\text{set } x)$ more clearly, it is helpful to consider two different computer programs. Consider the DAG in Figure 4. The probability function for a distribution consistent with this DAG has the form $p(x,y,z) = p(x)p(y|x)p(z|x,y)$. The following is pseudocode for generating from this distribution.

$$\begin{aligned} \text{For } i &= 1,\ldots,n: \\ x_i &\leftarrow p_X(x_i) \\ y_i &\leftarrow p_{Y|X}(y_i|x_i) \\ z_i &\leftarrow p_{Z|X,Y}(z_i|x_i,y_i) \end{aligned}$$

Suppose we run this code, yielding data $(x_1,y_1,z_1),\ \ldots,(x_n,y_n,z_n)$. Among all the times that we observe $Y=y$, how often is $Z=z$? The answer to this question is given by the conditional distribution of $Z|Y$. Specifically,

$$\begin{aligned} \mathbb{P}(Z=z|Y=y) &= \frac{\mathbb{P}(Y=y,Z=z)}{\mathbb{P}(Y=y)} = \frac{p(y,z)}{p(y)} \\ &= \frac{\sum_x p(x,y,z)}{p(y)} = \frac{\sum_x p(x)\,p(y|x)\,p(z|x,y)}{p(y)} \\ &= \sum_x p(z|x,y)\frac{p(y|x)\,p(x)}{p(y)} = \sum_x p(z|x,y)\frac{p(x,y)}{p(y)} \\ &= \sum_x p(z|x,y)\,p(x|y). \end{aligned}$$

Now suppose we **intervene** by changing the computer code. Specifically, suppose we fix Y at the value y. The code now looks like this:

$$\begin{aligned} \text{set } Y &= y \\ \text{for } i &= 1,\ldots,n \\ x_i &\leftarrow p_X(x_i) \\ z_i &\leftarrow p_{Z|X,Y}(z_i|x_i,y) \end{aligned}$$

Having **set** $Y=y$, how often was $Z=z$? To answer, note that the intervention has changed

the joint probability to be

$$p^*(x,z) = p(x)p(z|x,y).$$

The answer to our question is given by the marginal distribution

$$p^*(z) = \sum_x p^*(x,z) = \sum_x p(x)p(z|x,y).$$

This is $p(z|\mathsf{set}\ Y = y)$.

Example 9 *You may have noticed a correlation between rain and having a wet lawn, that is, the variable "Rain" is not independent of the variable "Wet Lawn" and hence* $p_{R,W}(r,w) \neq p_R(r)p_W(w)$ *where* R *denotes Rain and* W *denotes Wet Lawn. Consider the following two DAGs:*

$$\text{Rain} \longrightarrow \text{Wet Lawn} \qquad \text{Rain} \longleftarrow \text{Wet Lawn}.$$

The first DAG implies that $p(w,r) = p(r)p(w|r)$ *while the second implies that* $p(w,r) = p(w)p(r|w)$ *No matter what the joint distribution* $p(w,r)$ *is, both graphs are correct. Both imply that* R *and* W *are not independent. But, intuitively, if we want a graph to indicate causation, the first graph is right and the second is wrong. Throwing water on your lawn doesn't cause rain. The reason we feel the first is correct while the second is wrong is because the interventions implied by the first graph are correct.*

Look at the first graph and form the intervention $W = 1$ *where 1 denotes "wet lawn." Following the rules of intervention, we break the arrows into* W *to get the modified graph:*

$$\text{Rain} \qquad \boxed{\textbf{set}\ \ \text{Wet Lawn} = 1}$$

with distribution $p^*(r) = p(r)$*. Thus* $\mathbb{P}(R = r \mid W := w) = \mathbb{P}(R = r)$ *tells us that "wet lawn" does not cause rain.*

Suppose we (wrongly) assume that the second graph is the correct causal graph and form the intervention $W = 1$ *on the second graph. There are no arrows into* W *that need to be broken so the intervention graph is the same as the original graph. Thus* $p^*(r) = p(r|w)$ *which would imply that changing "wet" changes "rain." Clearly, this is nonsense.*

Both are correct probability graphs but only the first is correct causally. We know the correct causal graph by using background knowledge.

Learning Casual Structure? We could try to learn the correct causal graph from data but this is dangerous. In fact it is impossible with two variables. With more than two variables there are methods that can find the causal graph under certain assumptions but they are large sample methods and, furthermore, there is no way to ever know if the sample size you have is large enough to make the methods reliable.

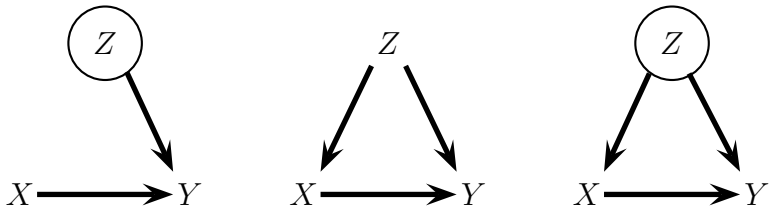


Figure 5: Randomized study; Observational study with measured confounders; Observational study with unmeasured confounders. The circled variables are unobserved.

Randomization Again. We can use DAGs to represent **confounding** variables. If X is a treatment and Y is an outcome, a confounding variable Z is a variable with arrows into both X and Y; see Figure 5. It is easy to check, using the formalism of interventions, that the following facts are true:

In a randomized study, the arrow between Z and X is broken. In this case, even with Z unobserved (represented by enclosing Z in a circle), the causal relationship between X and Y is estimable because it can be shown that $\mathbb{E}(Y|X := x) = \mathbb{E}(Y|X = x)$ which does not involve the unobserved Z. In an observational study, with all confounders observed, we get $\mathbb{E}(Y|X := x) = \int \mathbb{E}(Y|X = x, Z = z)p(z)$ which is just the **adjusted treatment effect.** If Z is unobserved then we cannot estimate the causal effect because $\mathbb{E}(Y|X := x) = \int \mathbb{E}(Y|X = x, Z = z)dF_Z(z)$ involves the unobserved Z. We can't just use X and Y since in this case. $\mathbb{P}(Y = y|X = x) \neq \mathbb{P}(Y = y|X := x)$ which is just another way of saying that causation is not association.

References

Barron, A., Cohen, A., Dahmen, W. and DeVore, R. (2008). Approximation and learning by greedy algorithms. *The annals of statistics*, 64-94.

Berk, Richard and Brown, Lawrence and Buja, Andreas and Zhang, Kai and Zhao, Linda. (2013). Valid post-selection inference. *The Annals of Statistics*, 41, 802-837.

Greenshtein, Eitan and Ritov, Ya'cov. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10, 971-988.

Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric regression. *Ann. Statist.*, 28, 681-712.

Leeb, Hannes and Potscher, Benedikt M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics*, 142, 201-211.

Leeb, Hannes and Potscher, Benedikt M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21, 21-59.

Lei, Robins, and Wasserman (2012). Efficient nonparametric conformal prediction regions. *Journal of the American Statistical Association.*

Lei, Jing and Wasserman, Larry. (2013). Distribution free prediction bands. *J. of the Royal Statistical Society Ser. B.*

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.

Vovk, V., Nouretdinov, I., and Gammerman, A. (2009). On-line predictive linear regression. *The Annals of Statistics*, 37, 1566-1590.

Wasserman (2014), "Discussion of paper by Lockhart, Taylor, Tibshirani and Tibshirani. *Annals of Statistics.*

Wasserman, L., Kolar, M. and Rinaldo, A. (2013). Estimating Undirected Graphs Under Weak Assumptions. arXiv:1309.6933.

Wasserman, Larry and Roeder, Kathryn. (2009). High dimensional variable selection. *Annals of statistics*, 37, p 2178.