8/30/2004 NOTES

Logistics

Logistics, office hour times, the course website, relevant texts, and a list of topics to be covered are given on the course syllabus. This syllabus can be obtained from Mark van der Laan or Alan Hubbard. Course evaluation will be based on attendance, lecture notes (each student will have to transcribe one or more lectures), a midterm, and a final poster project. The final project will be an application of causal inference methodology from the class to real or simulated data, and students will be allowed to work in small groups.

The three big questions

Before a researcher comes to a statistician for help with data analysis, he or she should be able to answer the following questions.

(1) What is my data, and what is my population of interest?
(2) What is my model? (What assumptions can I make about the data generating distribution?)
(3) What is the parameter of interest? (What would I like to know about the population?)

A statistician's job is to then help the researcher estimate the parameter of interest. Causal inference problems fall into this three question framework, but the answers to (1), (2), and (3) are different from what one would find in more traditional fields of statistics, as we describe below.

1. What is the data in causal inference problems?

We will be interested in longitudinal data, where each subject is followed over time, and we define the following variables.

A(t). This denotes the treatment given to a subject at time t.
Y(t). This denotes some outcome of interest, measured at time t.
X(t). This include $Y(t)$, as well as time-dependent (and baseline) covariates measured on the subject.

Note that $A(t)$, $Y(t)$, $X(t)$ can be possibly multivariate. Define $\overline{A}$ as $(A(t) : t \geq 0)$, the process giving the value of $A(t)$ for each t, and define $\overline{Y}$ and $\overline{X}$ similarly. The observed data in causal inference problems is then n i.i.d. copies of $(\overline{A}, \overline{X}) \sim P_0$. If the reader is unfamiliar with the notation i.i.d., feel free to consult any introductory statistics text. Here P_0 is the (unknown) data generating distribution, which assigns probabilities to members of the population of interest.

As an example, consider an AIDS study. Suppose n patients are selected at random from an AIDS registry, and each is followed up for a period of time. Here P_0 is the probability distribution putting equal mass on each sample of n distinct subjects from the registry (this approximates i.i.d. sampling if the registry size is very large compared to n), and the population of interest consists of all members of the registry. We might have $A(t)$ represent the collection of medications being prescribed to the patient at time t, and the outcome $Y(t)$ might represent the viral load at time t or an indicator of whether the patient is still alive at time t. Here $X(t)$ could include baseline measurements such as the patient's sex, age, and income, as well as time-dependent covariates such as the patient's CD4 count at time t.

2. What is the model in causal inference problems?

Causal inference is the study of counterfactuals, which are the outcomes that would have been observed had the treatment somehow been different. Specifically, let $\overline{X}_{\overline{a}}$ be a counterfactual, and represent the process $\overline{X}$ that would have been observed had the treatment been set at $\overline{A} = \overline{a}$. When $\overline{A} = \overline{a}$, we refer to the observed $\overline{X}_{\overline{a}} = \overline{X}_{\overline{A}}$ as the factual. We will assume that $\overline{X}_{\overline{A}} = \overline{X}$, and this is called the consistency assumption. It states that the observed data is equal to what we would have observed in the counterfactual world had the treatment been set to the observed treatment. In the AIDS example, suppose $\overline{a}$ represents

no treatment being given. Then for a given patient, $\overline{X}_{\overline{a}}$ represents the covariate process that would have occured if, contrary to fact, no treatment had been given to that patient. The idea of counterfactuals raises philosophical issues that have been discussed at least since the time of David Hume, because in the real world each subject is only assigned one treatment process. In causal inference problems, our model will assume the existence of counterfactuals. If Θ represents the support of $\overline{A}$, then we refer to $\overline{X}^{FULL} = (\overline{X}_{\overline{a}} : \overline{a} \in \Theta)$ as the full data. For the estimation procedures discussed in this class to be effective, we must make additional assumptions on the distribution of $\overline{X}^{FULL}$ (the full data model), and the conditional distribution of $\bar{A}$, given $\overline{X}^{FULL}$, such as the sequential randomization or no unmeasured confounding assumptions, and these will be formally defined in subsequent lectures. The choice of models are typically heavily driven by the parameter of interest. Our general phylosophy is that one should make model assumptions on the parameter of interest, but try to minimize assumptions on the nuisance parameters.

Marginal structural models, models for direct and indirect effects, and history adjusted marginal structural models (three topics covered in this class) are just different models on (conditional) distributions of counterfactuals, describing how these conditional distributions change with a change in treatment regime $\bar{a}$.

3. What is the parameter of interest in causal inference?

In causal inference problems, parameters of interest are called causal parameters. Typically they are functions of the data generating distribution of $\overline{X}^{FULL}$, but in history adjusted marginal structural models they will be functions of conditional distributions of the counterfactual outcome $Y_{\bar{a}}$, given an observed past. Usually these parameters will be related to how the outcome process $\overline{Y}_{\overline{a}}$ varies with $\overline{a}$, and how this variation is modified by the covariates. For instance, in a marginal structural models our model might assume that $E[Y_{\overline{a}}(t)] = m(t, \overline{a}, \beta)$ for some known function $m(\cdot)$ and unknown Euclidean parameter β. In this case, β would be the causal parameter, and we would have to find a way to estimate it from the observed data. Note that because the observed data $(\overline{A}, \overline{X})$ is a strict subset of the full data $(\overline{A}, \overline{X}^{FULL})$, causal inference can be treated as a missing data problem. This will be heavily exploited in the following lectures. The general estimating function approach for censored/missing data structures as described in van der Laan, Robins (2002), and presented in this course, corresponds with first finding procedures (i.e., full data estimating functions) that can estimate the causal parameter from the full data, and then map these to procedures (i.e., observed data estimating functions) that estimate the causal parameter from the observed data.

Causal graphs, 9/1/2004 Notes.

In point treatment studies we are interested in studying the causal effect of treatment on outcome of interest. Let A denotes the treatment and Y the outcome variables respectively. Other covariates $X_1, X_2, \ldots, X_m$ are also collected from the patients in the study, requiring adjustment of the causal effect for the covariates. In some situations, the set of causal assumptions on the random variables is known before the study begins and can be represented in the form of a *causal graph*.

Before we proceed to give a formal definition of this concept, we introduce the notation using the causal graph in fig. 1 as an example. There are 5 variables in this hypothetical study - X_1, X_2, X_3, A and Y, which are represented in the graph as vertices. Directed edges represent causal dependence (of which the statistical conditional dependence is a special case), thus Y is causally dependent on A, X_1 and X_3; X_2 is dependent on X_1, etc. We define $PA(X)$ to be the set of all random variables Z in the graph which have a direct arrow going into the node X. In other words $PA(X)$ is the set of parent nodes of vertex X, i.e. all those vertices in the causal graph that have a directed edge that ends in X. For example $PA(Y) = \{A, X_1, X_3\}$ and $PA(X_1) = \emptyset$. Loops are not allowed, which is equivalent to requiring that the causal graph is DAG (directed acyclic graph).

Definition 1. A causal graph G for the set of R.V. $(X_1, X_2, \ldots, X_m, A, Y)$ in a point treatment study is a DAG defining a set of causal assumptions:

$$\begin{aligned} X_i &= f_i(PA(X_i), \epsilon_i), \; i = 1, \ldots, m, \\ A &= f_A(PA(A), \epsilon_A), \\ Y &= f_Y(PA(Y), \epsilon_Y), \end{aligned}$$

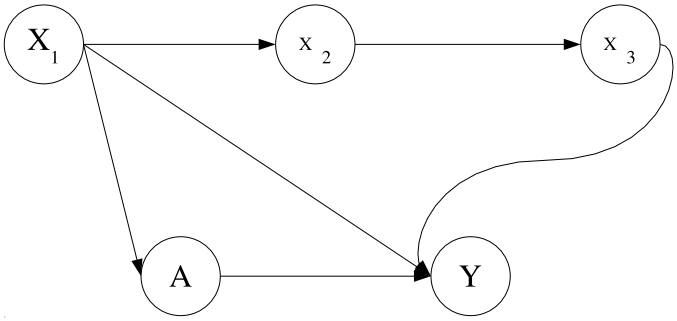

FIGURE 1. Example of a causal graph with three covariates.

for some deterministic functions f_i, $i = 1, \ldots, m$, f_A and f_Y, and ϵ_i, $i = 1, \ldots, m$, ϵ_A and ϵ_Y are random variables called *(exogenous) errors* satisfying the following assumptions $\epsilon_i \perp PA(X_i), \epsilon_A \perp PA(A), \epsilon_Y \perp PA(Y)$.

Note that the causal graphs just assumes that nodes are certain functions of parents-nodes and exogenous error variables, but it assumes nothing about the functional form of these deterministic functions. However, in practice, if one aims to estimate these unknown deterministic functions, then, by the curse of dimensionality, one will parameterize each of these functions with a set of parameters such as coefficients of linear/logistic regression models. The causal graph can involve unmeasured variables: that is, A, Y or $X_1, \ldots, X_m$ can be subject to missingness or (right-)censoring. In this case one views the causal graph as a set of assumptions on the full-data random vector being the collection of nodes in the causal graph: thus, in this case the causal graph does not make any assumptions about the conditional distribution of censoring/missingness variables, given the full data $X^{FULL} \equiv (X_1, \ldots, X_m, A, Y)$. Specifying the whole DAG is usually hard in practice. Given such a DAG it is now possible to identify a *causal* effect of one node on another node in the graph from the distribution/density of the full-data $X^{FULL} = (X_1, \ldots, X_m), A, Y$. In addition, if the observed data is $O = \Phi(C, X^{FULL})$ for some known function Φ of a censoring variable C and X^{FULL} (this is the most general definition of a censored/missing data structure), and one assumes that the conditional distribution of C, given X^{FULL}, satisfies *coarsening at random* (see e.g., van der Laan, Robins, 2002, for literature overview and definitions), then one can often identify from the observed data distribution the full data distribution X^{FULL}, and thereby the wished causal effects. To understand what variables in the graph can be completely missing (i.e., they are not observed on any subject in the sample) while still having coarsening at random and (thereby) identification of the wished causal effect is an interesting problem, and area of research.

In class we we will present an alternative *counterfactual* approach exists that does not require full specification of a causal graph, but, does only require knowing which variables are pre-treatment, but does also need to assume that there are no unmeasured confounders.

Once we have the causal graph we can identify from the distribution of X^{FULL} the counterfactual distribution $P(Y_a = y)$ which is the marginal distr. of Y when treatment is set at level $A = a$. If the interest is only in the effect of A on Y, all covariates connected to A only through an undirected path that includes Y should be ignored. One important issue in the study of causal graphs is what's the minimal subset of covariates that is sufficient to identify the counterfactual distribution of Y_a; related to this is the question of what's the minimal set of confounders that needs to be stratified upon in a point treatment study.

Definition 2. A set of edges connecting two vertices A and Y is called a *back-door path* if $\exists X$ s.t. $X \in PA(A)$ and there is a undirected path between X and Y. A vertex in a back-door path is called a *collider* if the path edges incident with this vertex are incoming (i.e. having their direction towards the vertex). A *confounding* between A and Y is present if $\exists$a back-door path with no colliders connecting A and Y.

For example, there are two back-door paths present in fig. 1 - $A \to X_1 \to Y$ and $A \to X_2 \to X_3 \to Y$.

How to find the likelihood of the data given a causal graph? Using the chain rule for factoring the joint likelihood of discrete R.V. $(Z_1, Z_2, \ldots, Z_d)$:

$$P(Z_1, Z_2, \ldots, Z_d) = P(Z_1) \prod_{i=2}^{d} P(Z_i | Z_1, \ldots, Z_{i-1}), \tag{1}$$

together with the conditional dependence between variables implied by the causal graph $P(Z_i|Z_1, \ldots, Z_{i-1}) = P(Z_i|PA(Z_i))$, we obtain:

$$P(Z_1, Z_2, \ldots, Z_d) = \prod_{i=1}^{d} P(Z_i|PA(Z_i)).$$

When applying this to a single observation in a point-treatment study with discrete R.V. $(X_1, X_2, \ldots, X_m, A, Y)$ we obtain:

$$\begin{aligned} P(X_1 = x_1, \ldots, X_m = x_m, A = a, Y = y) &= \\ &= \prod_{j=1}^{m} P(X_j = x_j | PA(X_j)) P(A = a | PA(A)) P(Y = y | PA(Y)). \end{aligned} \tag{2}$$

Assuming that each of the functional relationships in definition 1 is parameterized as a (e.g., linear) regression in the parent nodes with known (or up till a finite dimensional parameter) conditional distribution of the error-term, given the parent-nodes, we can express the joint probability in 2 as a function of an unknown parameter vector (e.g., the collection of node-specific regression coefficients), which represents now a parametric model and for the density/likelihood of X^{FULL}.

One can estimate the unknown parameters with the maximum likelihood estimator obtained by maximizing the likelihood $\prod_{i=1}^{n} P(X^{FULL} = x_i^{FULL})$ of an observed sample x_i^{FULL}, $i = 1, \ldots, n$. Typically, this can be carried out with standard software (e.g. implementations of generalized linear regression).

Given the causal graph and its estimated functional relations, one can now define the distribution of Y_a as the distribution one obtains by fixing the treatment variable $A = a$ in the system of equations defined by the node-specific equations, and generating all nodes accordingly.

Example 1. The causal graph in this example is specified in fig. 2.

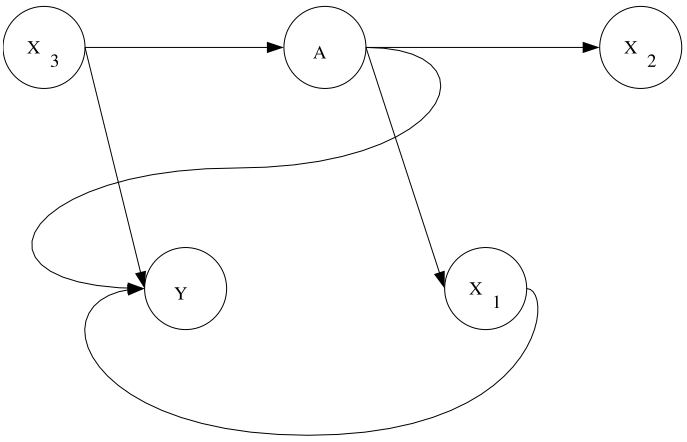

FIGURE 2. Another example of a causal graph with three covariates.

Assuming continuous distr. for all R.V in the causal graph, the density for a single observation can be written as follows:

$$f(X_1, X_2, X_3, A, Y) = f(X_3) f(A|X_3) f(X_2|A) f(Y|A, X_1, X_3) f(X_1|A). \tag{3}$$

Now let's find the counterfactual distribution for $A = a$:

(1) Erase $f(A|PA(A))$ from 3. This is equivalent to performing an incision to the graph in fig. 2 that removes both vertex A and the edges incident to A (fig. 3);
(2) Set $A = a$ in all functions where A belongs to the parents' set;
(3) Define $f_a(X_1, X_2, X_3, Y) = f(X_3) f(X_2|A = a) f(Y|A = a, X_1, X_3) f(X_1|A = a)$;
(4) Integrate out X_1, X_2 and X_3 in $f_a(X_1, X_2, X_3, Y)$ to get the counterfactual density $f_a(Y)$.

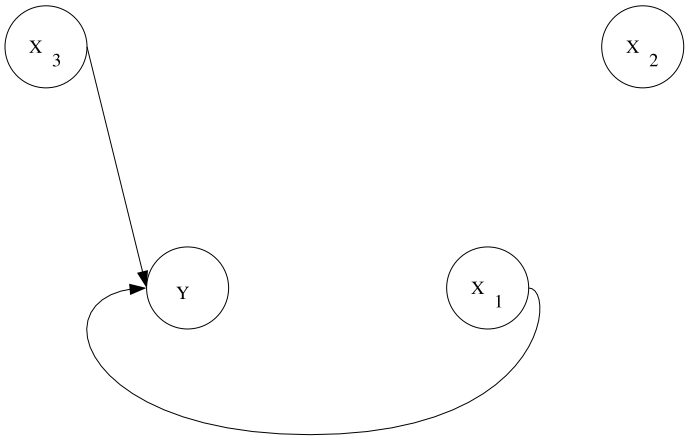

FIGURE 3. Graph from example 1 after incision of A.

Example 2. This is how Pearl defined the counterfactual probability distribution $f_a(Y)$. Hence, if A has two levels - treatment ($a = 1$) and control ($a = 0$), and the causal parameter (or effect) of interest is the treatment difference, one needs to calculate the quantity $Ef_1(Y) - Ef_0(Y)$.

The above method for doing causal inference can be summarized as follows. Using standard software we can do maximum likelihood estimation to fit the functional forms of each node in the causal graph, thus estimating the parameters in each functional relationships in def. 1. Subsequently, given the maximum likelihood estimator of the unknown parameters, we can identify corresponding estimates of the treatment specific disstribution of $(Y_a, X_{1a}, \ldots, X_{ma})$ by Monte-Carlo simulation for each choice of treatment level a.

If the functions in def. 1 are parameterized using flexible regression functions, one can employ data-adaptive model selection methods such as cross-validation or penalized likelihood methods. Such methods provide tools to decide data adaptively how flexible the parametrization should be. Clearly, if the number of parameters is larger than the sample size n, then the maximum likelihood estimator of the unknown parameter vector becomes too variable or ill defined, and thereby our estimate of the treatment specific distribution is too variable as well. That is, the size/dimension/complexity of the model needs to be data dependent. An important research area is the development of methods which data adaptively selects models for the purpose of estimating a particular parameter (such as in our case, the causal effect of treatment on the outcome Y) of interest.

To estimate the variability of the estimates of the causal parameters, (non-) parametric bootstrap is usually employed. This involves resampling repeatedly n observations from the actual sample (nonparametric bootstrap) or from a fit of the true probability distribution of the data (parametric bootstrap).

Is it possible to make a choice between different causal graphs based exclusively on the data from a study? Short answer is, "No". Consider a simple causal graph with 2 R.V. X_1, X_2, where the true causal graph and data generating distribution is: $X_1 \rightarrow X_2$, $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 = X_1 + c$. If we would choose between the only two possible causal graphs $X_1 \rightarrow X_2$ and $X_1 \leftarrow X_2$ the one with the largest corresponding fitted likelihood (with the second model assuming $X_1 = X_2 + c^{'}$, $X_2 \sim N(\mu_2, \sigma_2^2)$, and μ_2, σ_2 estimated from the data generated using the true data distribution) we'll get that approximately 50% of the time we pick the wrong causal graph. In general, if all nodes $(A, Y, X_1, \ldots, X_m)$ are discrete valued, then all possible causal graphs of X^{FULL} give the same corresponding fitted maximum likelihood estimate of the distribution of

X^{FULL}, *if* we use for all causal graphs the nonparametric model (that is, do not assume any parametric form for the functional relations). In this case, the causal-graph specific maximum likelihood is completely flat in the choice of causal graph. Consequently, any variability in the causal-graph specific maximum likelihood values across causal-graphs is NOT due to changes in the causal graphs, but it is due to the fact that different *parametrized* causal graphs result in different statistical models for X^{FULL}, and one might approximate the true distribution of X^{FULL} better than another. For example, if in truth $X_1 \rightarrow X_2$, and the conditional mean of X_1, given X_2 happens to be linear in X_2, then a wrong causal graph $X_2 \rightarrow X_1$ with corresponding linear normal regression assumption $X_1 \sim N(beta X_2, \sigma^2)$, will likely give a higher maximum likelihood value then a correct causal graph $X_1 \rightarrow X_2$ with corresponding assumption X_2 is exponential with $\lambda(X_1) = \beta X_1$.