Validating Causal Models

Dustin Tran
Columbia University

Francisco J. R. Ruiz
Columbia University

Susan Athey
Stanford University

David M. Blei
Columbia University

Abstract

The goal of causal inference is to understand the outcome of alternative courses of action. However, all causal inference requires assumptions—more so than for standard tasks in probabilistic modeling—and testing those assumptions is important to assess the validity of a causal model. We develop Bayesian model criticism for causal inference, building on the idea of posterior predictive checks to assess model fit. Our approach involves decomposing the problem, separately criticizing the model of treatment assignments and the model of outcomes. Further we discuss how and when we can check the central assumption of unconfoundedness, which enables causal statements from observational data. Our approach provides a foundation for diagnosing causal inferences from observational data.

1 Introduction

A causal model of the world reflects its true scientific mechanisms. It helps us learn about how things work and predict what happens when certain things change. Consider the problem of understanding the "treatment effect" of an intervention, such as giving a drug to patients with a given disease. In the language of Neyman (1923); Rubin (1974), each individual has a potential outcome when given the drug and a potential outcome when not given the drug. One measurement of the causal effect is the average difference (over individuals) between those potential outcomes. In the language of graphical models (Pearl, 2000), this is framed as evaluating the impact of an intervention on random variables in a probabilistic graph. Holland's fundamental problem of causal inference is that we do not observe both potential outcomes for any individual at the same time (Holland, 1986); thus we need to make additional assumptions about the generating process of the data to estimate the causal effect.

Assumptions are important to all of causal inference, but especially when analyzing observational studies. In such studies, we make strong assumptions about how treatments are assigned to individuals and each individual's distribution of potential outcomes. There are myriad methods for working with observational data, given the assumptions, to make causal inferences (Dawid, 2000; Bottou et al., 2013; Mooij et al., 2014; Peters et al., 2015).

We propose a collection of methods for checking assumptions in causal inference, a suite of tools for the applied researcher to understand if and to what degree her assumptions hold when analyzing observational data. Relative to prediction, the implicit goal of standard probabilistic modeling, evaluating the performance of causal models is more subtle. A good causal model predicts well in settings different from the one that generated the data, but without data from such settings there is no observable test set with which to directly evaluate performance. Crucial assumptions for drawing inferences from observational data are not testable without additional assumptions.

Despite these challenges, we develop Bayesian model criticism for causal inference. We build on the idea of *posterior predictive check*s (PPCS) to adapt goodness-of-fit style calculations to causal modeling. Our approach involves decomposing the problem, separately criticizing the two components that make up a causal model: the model of treatment assignments and the model of outcomes. Conditioned on the assumption of *unconfoundedness*—that the treatments are assigned independently of the potential outcomes—we show how to check the assumed models of who is given the treatment and how well it works. Finally, we discuss how to check unconfoundedness in a causal model. This requires additional assumptions, namely that either the outcome model or treatment model is correctly specified and is validated independent of the data set under study.

We begin by describing model-based causal inference with potential outcomes. We then develop methods for checking the models. Finally, we study our methods on simulated and real data.

2 Causal Models

We describe causal models in terms of potential outcomes (Imbens and Rubin, 2015). Variables $\boldsymbol{y}(0)$, $\boldsymbol{y}(1)$ are

a set of *potential outcomes* (Neyman, 1923; Rubin, 1974) with binary treatment assignment $a \in \{0, 1\}$; the variable $y_i(a)$ is the outcome when an individual i is assigned to treatment a. Let $\boldsymbol{a}$ denote a set of treatment assignments; and let $\boldsymbol{x}$ denote a set of observed covariates. For each individual we only observe $y_i(a_i)$, the potential outcome for the assigned treatment; the other outcome is necessarily unobserved.

A *causal model* is a joint distribution of the potential outcomes, assignments, and governing parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, conditional on the covariates. Through the structure of the joint, it encodes assumptions about the underlying process. Here we focus on this model,

$$p(\boldsymbol{y}(0), \boldsymbol{y}(1), \boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{x}) = \tag{1}$$
$$\left(p(\boldsymbol{\theta}) \prod_{i=1}^{n} p(y_i(0), y_i(1) \mid x_i, \boldsymbol{\theta})\right)\left(p(\boldsymbol{\phi}) \prod_{i=1}^{n} p(a_i \mid x_i, \boldsymbol{\phi})\right)$$

The causal model has two components: the *outcome model* $p(\boldsymbol{\theta})p(\boldsymbol{y}(0), \boldsymbol{y}(1) \mid \boldsymbol{x}, \boldsymbol{\theta})$ and the *assignment model* $p(\boldsymbol{\phi})p(\boldsymbol{a} \mid \boldsymbol{x}, \boldsymbol{\phi})$. The outcome model determines the distribution of potential outcomes, i.e., the distribution of each observed outcome $y_i(a_i)$ and what would have happened in an alternative world where a_i is a different value. The assignment model determines the distribution of treatments. Equation 1 assumes "unconfoundedness" (Rubin, 1974); the conditional treatment and potential outcomes are conditionally independent given the covariates, $p(\boldsymbol{a} \mid \boldsymbol{y}(0), \boldsymbol{y}(1), \boldsymbol{x}) = p(\boldsymbol{a} \mid \boldsymbol{x})$.[1]

Given a causal model, causal inference concerns the outcomes $\boldsymbol{y}(0), \boldsymbol{y}(1)$. For example, the *average treatment effect (*ATE*)* is the expected difference in outcomes

$$\text{ATE} = \mathbb{E}\left[Y(0)\right] - \mathbb{E}\left[Y(1)\right],$$

where the expectation is taken across the population of individuals.

One goal of causal modeling is to infer the ATE from observational data. Observational data contains per-individual treatment assignments and the corresponding outcome $\mathcal{D}^{\text{obs}} = \{a_i, y_i(a_i)\}_{i=1}^{n}$. In this data set, the observed quantities are assumed drawn from Equation 1 and the counterfactual outcome is a latent variable. The "fundamental problem of causal inference" (Holland, 1986) is that only one of the outcomes is observed for each data point. Using observational data, naive estimates of the ATE, such as the empirical estimate, are biased by the assignment mechanism.

Causal inference tries to find estimators which remove this bias. Ideally, we would collect data from an experiment where each assignment a_i is set according to known assignment parameters $\boldsymbol{\phi}^*$ (e.g., random assignment). This gives an alternative joint on the variables (Pearl, 2000),

$$p(\boldsymbol{y}(0), \boldsymbol{y}(1), \boldsymbol{a}, \boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{\phi}^*) = \tag{2}$$
$$\left(p(\boldsymbol{\theta}) \prod_{i=1}^{n} p(y_i(0), y_i(1) \mid x_i, \boldsymbol{\theta})\right)\left(\prod_{i=1}^{n} p(a_i \mid x_i, \boldsymbol{\phi}^*)\right).$$

We call this the *do model*. It is also called an "intervention" or "mutilation."

The challenge of causal inference is to use observational data to estimate an underlying causal model, and then to use that causal model to estimate quantities under alternative assignment processes. Widely used methods, such as stratification, matching, and inverse propensity weighting, all use data collected from Equation 1 to estimate causal quantities from Equation 2.

These methods rest on the same assumption—the conditional independence of Equation 1, unconfoundedness. Unobserved confounders that affect both treatment assignment and potential outcomes (e.g., sicker patients are more likely to receive a drug) invalidate the methods no matter how many patients are observed. Further, in real-world analyses with finite samples, these methods also require assumptions on the outcome model $p(\boldsymbol{\theta})p(\boldsymbol{y}(0), \boldsymbol{y}(1) \mid \boldsymbol{x}, \boldsymbol{\theta})$ and the assignment model $p(\boldsymbol{\phi})p(\boldsymbol{a} \mid \boldsymbol{x}, \boldsymbol{\phi})$. We now describe when and how we can check these assumptions.

3 Validating Causal Models

Model criticism measures the degree to which our model falsely describes the data (Gelman and Shalizi, 2012). Following falsificationists such as Popper and Box, we can never validate whether a model is true—no model will be true in practice—but we can seek to uncover where the model goes wrong. Model criticism can help justify the model as an approximation or point to good directions for model revision.

The central tool of model criticism is the PPC. It quantifies the degree to which data generated from the model deviate from data generated from the true distribution (Box, 1980; Rubin, 1984; Gelman et al., 1996). The procedure is:

1. Design a discrepancy function, a statistic of the data and hidden variables. A "targeted" discrepancy summarizes a specific component of the data, such as a quantile. An "omnibus" discrepancy is an overall summary of the data, such as the χ^2 goodness of fit.
2. Form the *realized discrepancy*, which is the statistic applied to observed data (along with posterior samples of hidden variables).
3. Form the *reference distribution*, the distribution of the

[1] In general, the true data generating mechanism need not satisfy this conditional independence. But this is the class of models that is most commonly used in practice.

discrepancy applied to many replicated data sets from the posterior predictive distribution.

4. Check if the realized discrepancy is unlikely to have come from the reference distribution (e.g., by calculating the tail probability). If so, then the model poorly describes the data according to this function; revise the model. Otherwise, this provides evidence that the model is justified.

Model criticism is typically applied to validating non-causal models, especially for exploratory and unsupervised tasks (Mimno and Blei, 2011; Krafft et al., 2012; Mimno et al., 2015; Lloyd and Ghahramani, 2015). Here we aim to extend it to validating causal models. Define a *causal discrepancy* to be a scalar function of the form,

$$T((\boldsymbol{y}(0), \boldsymbol{y}(1)), \boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\phi}). \tag{3}$$

This is possibly a function of realizations from all variables in a causal model: the potential outcomes $\boldsymbol{y}(0), \boldsymbol{y}(1)$, the treatment assignment $\boldsymbol{a}$, the outcome model parameters $\boldsymbol{\theta}$, and the assignment model parameters $\boldsymbol{\phi}$. Depending on the check, it is a function of a subset of these variables.

There are two ingredients to a causal check: the *reference distribution* and the *realized discrepancy*. In a classical PPC the reference distribution is simply the posterior predictive. This is the distribution that the data would have come from if the model were true. But this approach does not work for a causal check—the posterior predictive of the potential outcomes $(y(0), y(1))$ should come from a causal inference, rather than the observed data.

We define the posterior predictive distribution of the potential outcomes as conditioned on a second (hypothetical) data set $\mathcal{D}^{\text{do}}$, which comes from the do-model in Equation 2. (We can think of this data as being an imaginary experiment that we would have performed if we could.) Below, in implementing the check, we use inverse propensity weighting to approximate this data from the observed data. But, for defining the reference distribution, assume that we see both data sets.

Let $\boldsymbol{y}(0)^{\text{rep}}, \boldsymbol{y}(1)^{\text{rep}}, \boldsymbol{a}^{\text{rep}}$ denote replicated data, i.e., if the process that produced the observational data were replicated to produce new data. Consider calculating the discrepancy over replications from the causal model, fitted to observational data $\mathcal{D}^{\text{obs}}$:

$$\begin{aligned} \boldsymbol{\phi} &\sim p(\boldsymbol{\phi} \mid \mathcal{D}^{\text{obs}}), \\ \boldsymbol{a}^{\text{rep}} \mid \boldsymbol{\phi} &\sim p(\boldsymbol{a}^{\text{rep}} \mid \boldsymbol{x}, \boldsymbol{\phi}), \\ \boldsymbol{\theta} &\sim p(\boldsymbol{\theta} \mid \mathcal{D}^{\text{do}}), \\ \boldsymbol{y}(0)^{\text{rep}}, \boldsymbol{y}(1)^{\text{rep}} \mid \boldsymbol{\theta} &\sim p(\boldsymbol{y}(0)^{\text{rep}}, \boldsymbol{y}(1)^{\text{rep}} \mid \boldsymbol{x}, \boldsymbol{\theta}). \end{aligned} \tag{4}$$

The reference distribution of the discrepancy is governed by this distribution of its arguments.

The realized discrepancy is evaluated on observed data. When $T(\cdot)$ depends on latent variables—either assignment parameters, outcome parameters, or alternative outcomes—we replicate them from the reference distribution. Following Gelman et al. (1996), the observed data are always held fixed at their observed values; only latent variables are replicated. Note this is in contrast to the reference distribution, which resamples all of the variables.

We described the general form of the discrepancy, the causal model reference distribution, and the realized discrepancy. We use these constructions to criticize causal models. Following the decomposition of a causal model in Equation 1, and motivated by the clear separation of replications in Equation 4, we separate criticism into three components: criticizing the assignment model, criticizing the outcome model, and criticizing their implicit independence relationship.

3.1 Criticizing the assignment model

The gold standard for validating causal models is a held-out experiment, where we have access to the assignment mechanism when validating against held-out outcomes (Rubin, 2008). In observational studies, however, the assignment mechanism is unknown; we must model it. The goal is to capture the true distribution of the assignments. Thus we can apply a standard PPC for non-causal models. We criticize the assignment model with discrepancies of the form $T(\boldsymbol{a}, \boldsymbol{\phi})$.

Algorithm 1 describes the procedure. It isolates the components of the model and data relevant to the assignment mechanism. We calculate the realized discrepancy $T(\boldsymbol{a}, \boldsymbol{\phi}^{\text{rep}})$ and compare against the reference distribution $T(\boldsymbol{a}^{\text{rep}}, \boldsymbol{\phi}^{\text{rep}})$. The reference distribution is simply the posterior predictive; inferences about $\boldsymbol{a}$ and $\boldsymbol{\phi}$ in Equation 4 do not require data other than the observations.[2]

Example. Consider the average log-likelihood of assignment,

$$T(\boldsymbol{a}, \boldsymbol{\phi}) = \frac{1}{n} \sum_{i=1}^{n} \log p(a_i \mid x_i, \boldsymbol{\phi}). \tag{5}$$

The reference distribution is this discrepancy evaluated over posterior draws of $\boldsymbol{\phi}$ and replications of $\boldsymbol{a}$ from the likelihood given posterior draws. The realized discrepancy is the discrepancy evaluated on the observed set of assignments $\boldsymbol{a}$ and over the same posterior draws of $\boldsymbol{\phi}$. Figure 1a show examples of this check. In the third panel we have simulated data with a misspecified assignment model. The realized discrepancy is far away from the reference distribution.

[2]The realized discrepancy can also be evaluated on held-out assignments to avoid criticizing the model with the same observational data that is used to train it (Bayarri et al., 2007). We use this approach in our study.

Algorithm 1: Criticism of the assignment model

Input: Assignment model $p(\boldsymbol{\phi} \mid \mathcal{D}^{\text{obs}})\, p(\boldsymbol{a}^{\text{rep}} \mid \boldsymbol{x}, \boldsymbol{\phi})$,
discrepancy $T(\boldsymbol{a}, \boldsymbol{\phi})$.
Output: Reference distribution $p(T)$ and realized test statistic T^{obs}.
for $s = 1, \ldots, S$ replications **do**
  Draw assignment parameters $\boldsymbol{\phi}^s \sim p(\boldsymbol{\phi} \mid \mathcal{D}^{\text{obs}})$.
  Draw assignments $\boldsymbol{a}^{\text{rep},s} \sim p(\boldsymbol{a}^{\text{rep}} \mid \boldsymbol{x}, \boldsymbol{\phi}^s)$.
  Calculate discrepancy $T^{\text{rep},s} = T(\boldsymbol{a}^{\text{rep},s}, \boldsymbol{\phi}^s)$.
  Calculate discrepancy $T^{\text{obs},s} = T(\boldsymbol{a}, \boldsymbol{\phi}^s)$.
end
Form reference distribution $p(T)$ from replications $\{T^{\text{rep},s}\}$.
Form realized discrepancy T^{obs} from replications $\{T^{\text{obs},s}\}$.

3.2 Criticizing the outcome model

The second component of a causal model is the outcome model, $p(\boldsymbol{\theta})\, p(\boldsymbol{y}(0), \boldsymbol{y}(1), \boldsymbol{\theta} \mid \boldsymbol{x})$. The outcome model represents the causal phenomenon, that is, the outcomes $\boldsymbol{y}$ caused by setting $\boldsymbol{A} = \boldsymbol{a}$. The outcome model is inherently difficult to infer and criticize. It involves inferences about a distribution of counterfactuals, but with data only available from one counterfactual world.

One solution is to use discrepancies that are a function only of the observed outcomes, rather than all potential outcomes. The realized discrepancy is feasible, and the reference distribution is easy to form. However, this severely limits the class of discrepancies. Another solution is to use the reference distribution to impute the counterfactual outcomes. This strategy has been studied for missing data analysis (Gelman et al., 2005), but it makes the discrepancy less useful for criticizing the outcome model; ideally we compare the outcome model to imputed outcomes not relying on the outcome model.

We show how to calculate a realized discrepancy over all outcomes $T(\boldsymbol{y}(0), \boldsymbol{y}(1), \boldsymbol{\theta})$ using inverse propensity weighting (Rosenbaum and Rubin, 1983). Consider a proxy to the realized discrepancy,

$$T\left(\left\{\frac{\delta_{A_i=0}(a_i)}{p(a_i \mid x_i)} y_i(0)\right\}, \left\{\frac{\delta_{A_i=1}(a_i)}{p(a_i \mid x_i)} y_i(1)\right\}, \boldsymbol{\theta}\right), \tag{6}$$

where $\delta_{A_i=0}$ denotes a Dirac delta distribution with all probability mass at $A_i = 0$, and $p(a_i \mid x_i)$ is the assignment model's probability of observing assignment a_i given individual covariates x_i. With inverse propensity weighting, the observed outcome $y_i(a_i)$ is re-weighted as if it came from the model with the intervention δ. The unobserved outcome $y_i(1-a_i)$ is weighted by zero so it does not explicitly appear during computation. Note that Equation 8 is a valid realized discrepancy: it is an asymptotically unbiased estimator of the true realized discrepancy calculated over both sets of potential outcomes. See Appendix for details.

Algorithm 2: Criticism of the outcome model

Input: Causal model $p(\boldsymbol{\theta} \mid \mathcal{D}^{\text{do}})\, p(\boldsymbol{y}(0)^{\text{rep}}, \boldsymbol{y}(1)^{\text{rep}} \mid \boldsymbol{x}, \boldsymbol{\theta})$
$p(\boldsymbol{\phi} \mid \mathcal{D}^{\text{obs}})\, p(\boldsymbol{a}^{\text{rep}} \mid \boldsymbol{x}, \boldsymbol{\phi})$,
discrepancy $T((\boldsymbol{y}(0), \boldsymbol{y}(1)), \boldsymbol{\theta})$.
Output: Reference distribution $p(T)$ and realized test statistic T^{obs}.
for $s = 1, \ldots, S$ replications **do**
  Draw outcome parameters $\boldsymbol{\theta}^s \sim p(\boldsymbol{\theta} \mid \mathcal{D}^{\text{obs}})$.
  Draw outcomes
  $\boldsymbol{y}(0)^{\text{rep},s}, \boldsymbol{y}(1)^{\text{rep},s} \sim p(\boldsymbol{y}(0)^{\text{rep}}, \boldsymbol{y}(1)^{\text{rep}} \mid \boldsymbol{\theta}^s)$.
  Calculate discrepancy
  $T^{\text{rep},s} = T((y(0)^{\text{rep}}, y(1)^{\text{rep}}), \boldsymbol{\theta}^s)$.
  Calculate discrepancy
  $T^{\text{obs},s} = T\left(\left\{\frac{\delta_{A_i=0}(a_i)}{p(a_i \mid x_i)} y_i(0)\right\}, \left\{\frac{\delta_{A_i=1}(a_i)}{p(a_i \mid x_i)} y_i(1)\right\}, \boldsymbol{\theta}^s\right)$.
end
Form reference distribution $p(T)$ from replications $\{T^{\text{rep},s}\}$.
Form realized discrepancy T^{obs} from replications $\{T^{\text{obs},s}\}$.

Algorithm 2 describes the procedure. It criticizes outcome models using discrepancies of the complete counterfactuals. For the inverse propensity weighting to work, the procedure assumes that the assignment model is correctly specified. While this is never true in practice, we can separately criticize the assignment model (Section 3.1) to make sure it is sufficient. As with non-causal modeling, our goal is to recover a good enough approximation by iteratively refining the model (Rubin, 1984).

Example. Following Athey and Imbens (2015), consider the mean squared error of average treatment effect,

$$T((\boldsymbol{y}(0), \boldsymbol{y}(1)), \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} ((y_i(1) - y_i(0) - \widehat{\tau}(x_i))^2, \tag{7}$$

and where $\widehat{\tau}(x_i)$ is the model's estimate of the conditional average treatment effect, $\mathbb{E}[y_i(1) - y_i(0) \mid x_i]$. This test statistic involves the complete set of counterfactuals. The reference distribution is directly in Equation 4. To calculate the realized discrepancy,

$$\tau_i^* = \frac{\delta_{A_i=1}(a_i)}{p(a_i \mid x_i)} y_i(1) - \frac{\delta_{A_i=0}(a_i)}{p(a_i \mid x_i)} y_i(0)$$

is an unbiased estimate of the conditional average treatment effect, where we use the assignment model to predict

$p(a_i \mid x_i)$. We can use the realized discrepancy,

$$T\left(\left\{\frac{\delta_{A_i=0}(a_i)}{p(a_i \mid x_i)} y_i(0)\right\}, \left\{\frac{\delta_{A_i=1}(a_i)}{p(a_i \mid x_i)} y_i(1)\right\}, \boldsymbol{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n}(\tau_i^* - \widehat{\tau}(x_i)^2,$$

and compare it to the reference distribution. Figure 1c shows an example. In the second panel, we have misspecified the outcome model. Though we evaluate it from purely observational data, we can detect this misspecification. (We discuss these figures in detail in Section 4.)

3.3 Criticizing unconfoundedness

Causal inference often hinges on one primary assumption: the distribution of potential outcomes and treatment assignment are conditionally independent given covariates, $\boldsymbol{Y}(0), \boldsymbol{Y}(1) \perp\!\!\!\perp \boldsymbol{A} \mid \boldsymbol{x}$. This is known by many terms—unconfoundedness, strong ignorability, selection on observables, no self-selection, no hidden variable bias, or no backdoor paths. It says that the assignment mechanism cannot depend on unobserved counterfactuals, which serve as missing data. Otherwise, it is impossible to model the assignment mechanism correctly and thus to make valid causal statements. This assumption is implicit in the causal model of Equation 1, $p(\boldsymbol{a} \mid \boldsymbol{y}(0), \boldsymbol{y}(1), \boldsymbol{x}) = p(\boldsymbol{a} \mid \boldsymbol{x})$.

Unlike the causal modeling assumptions which we checked above, the unconfoundedness assumption is untestable without making additional (untestable) assumptions. Ultimately, we require insight from the scientific domain. Consider criticizing unconfoundedness through the most general discrepancies of Equation 3, $T((\boldsymbol{y}(0), \boldsymbol{y}(1)), \boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\phi})$. Suppose further that the causal model is correctly specified. If the probability of observing the realized discrepancy is low under the reference distribution, then either the causal model is a poor fit or the assumptions made to enable causality are wrong. The correct model forces us to conclude that the assumptions are wrong.

While mathematically simple, this is core to the literature on conditional hypothesis testing. In a conditional test, the causal model is typically trivial, such as no treatment effects in Fisher's randomization inference ($\boldsymbol{y}(0) = \boldsymbol{y}(1)$); see, e.g., Angrist and Kuersteiner (2011); Zhang et al. (2012); Huber (2013). The null hypothesis asserts that unconfoundedness holds and p-values determine the probability that we can reject the null.

In practice, we cannot additionally assume that the causal model is correctly specified; this can be difficult to support from domain knowledge. However, we can assume that a component of the causal model is correctly specified, which may be easier to support. Such an assumption is ultimately necessary to check unconfoundedness.[3]

Example. As an example of a test of unconfoundedness, define $T((\boldsymbol{y}(0), \boldsymbol{y}(1)), \boldsymbol{a})$ to be a correlation calculated between $\boldsymbol{y}(0)$, $\boldsymbol{y}(1)$ and $\boldsymbol{a}$ given covariates $\boldsymbol{x}$. E.g., perform a linear regression of $\boldsymbol{y}(0)$ on $\boldsymbol{x}$ and $\boldsymbol{a}$; then report the estimate of $\boldsymbol{a}$'s coefficient. This measures the linear dependence of two variables after controlling for other variables. Also regress $\boldsymbol{y}(1)$ on $\boldsymbol{x}$ and $\boldsymbol{a}$.

Form the reference distribution by replicating data according to Equation 4. For the realized discrepancy, we can impute the unobserved outcomes using one of two approaches: direct draws from the outcome likelihood, i.e., the reference distribution; or inverse propensity weighting as described in Section 3.2. The choice of imputation method depends on whether we want to assume the functional form of the outcome model is correct (reference distribution) or the functional form of the assignment model is correct (propensity weighting). The imputation provides a filled-in data set, enabling us to examine whether the realized correlation likely arose from the reference distribution.

4 Empirical Study

We illustrate how to validate causal models with both synthetic and real data. With synthetic data we compare our conclusions to the true mechanism that generated the data. With real data we demonstrate how to apply our approach to criticize causal models in practice. In all our studies, we apply the test statistics from Sections 3.1 and 3.2, i.e., the average log-likelihood of the assignments and the mean squared error of the ATE; in future work we are analyzing experiments with the unconfoundedness tests. Note we are interested in the insights of the criticisms, not the insights of the causal models.

Synthetic data. We generate observations using a simple linear model (detailed below) to showcase the results of different predictive checks. We first infer the model parameters using the correct model specification and then introduce different types of misspecifications, either on the outcome model or the assignment model. Finally, we consider a scenario with a confounded model.

We generate $10,000$ data points, each a 10-dimensional covariate x_i, a binary treatment a_i, and a set of potential outcomes $(y_i(0), y_i(1))$,

$$x_i \sim \text{Uniform}(x_i \mid [0,1]^{10}),$$

[3] Note that we cannot use separate criticisms of the assignment or outcome model as in previous sections to support this additional assumption. These criticisms require unconfoundedness, making the reasoning circular.

$$
\begin{aligned}
a_i \mid x_i &\sim \text{Bernoulli}\left(a_i \mid \text{logistic}(x_i^\top \phi)\right), \\
y_i(0) \mid x_i &\sim \mathcal{N}\left(y_i(0) \mid x_i^\top \theta^{(0)}, \sigma^2\right), \\
y_i(1) \mid x_i &\sim \mathcal{N}\left(y_i(1) \mid x_i^\top \theta^{(1)}, \sigma^2\right).
\end{aligned}
$$

We place a standard normal prior over the model parameters ϕ, $\theta^{(0)}$, and $\theta^{(1)}$, and a Gamma prior with unit shape and rate on the variance σ^2.

We study two scenarios: (i) In the "science-fiction" scenario, we have access simultaneously to $y_i(0)$ and $y_i(1)$. (This is, of course, not possible in the real world.) (ii) In the "fiction" scenario, we only have access to one counterfactual outcome, $y_i = a_i y_i(1) + (1 - a_i) y_i(0)$. In each scenario, we check four causal models: (a) the correct model as specified above; (b) the "wrong outcomes" model, in which we mis-specify the distribution over $y_i(1)$ by using a t-distribution (with 3 degrees of freedom) instead of a Gaussian; (c) the "wrong assignment" model, in which the treatment likelihood $p(a_i \mid x_i)$ is misspecified; and (d) the "confounded" model, which only sees 7 out of the 10 covariates that constitute x_i.

We approximate the posterior with Markov chain Monte Carlo in Stan (Carpenter et al., 2016), obtaining 1000 samples. In the fiction scenario, we weigh the observations according to their inverse propensity scores. We also weigh the observations to form the outcome test statistic (Section 3). In the science-fiction scenario, we do not weigh the observations because we see both counterfactual outcomes.

Figure 1a illustrates criticism of the assignment model (Equation 5) for the science-fiction scenario (the plots for the fiction scenario are similar to these). As expected, when we use the correct model, the realized test statistic is approximately in the center of the reference distribution (panel a); this indicates that the model is correctly specified. The same for the misspecified outcomes model (panel b), because the assignment mechanism is still correctly specified. However, if the assignment model is wrong then the realized test statistic against the reference distribution (panel c) suggests that the model is misspecified. Regarding the confounded model, the result suggests that the model misspecification is not enough to make the test fail (panel d), although we will still detect confoundedness via the other test (discussed below).

We now turn to criticizing the outcome model. Figures 1b and 1c illustrate the test of the mean squared error of the ATE (Equation 7) for the science-fiction and fiction scenarios, respectively. As expected, the test for the correct model does not suggest any issue (panel a). When the outcome model is wrong, the test fails (panel b). This correctly indicates that we should revise the model. In the science-fiction scenario, the misspecification on the assignment model does not affect the outcome model (panel c), and thus the test indicates correctness. In the fiction scenario, however, the misspecified assignment model affects the inverse propensity score weighting and thus the p-value for the test is smaller (than for science-fiction). The confounded model (panel d) can also be detected in both scenarios.

Cockroaches. We analyze a real observational study of the effect of pest management on reducing cockroach levels in urban apartments (Gelman and Hill, 2006). Each apartment i was set up with traps for several days, with the goal of measuring the number of cockroaches trapped during a certain period.

Let t_i be the number of trap-days and y_i the number of cockroaches for each apartment. We use two additional covariates: the pre-treatment roach level r_i and an indicator s_i of whether the apartment is a "senior" building, restricted to the elderly. We model the data as Poisson,

$$
y_i \mid a_i, x_i \sim \text{Poisson}\left(y_i \mid t_i \exp\{\theta_0 + \theta_1 s_i + \theta_2 r_i + \theta_3 a_i\}\right),
$$

where a_i is the treatment indicator, θ is the outcome model, and $x_i = \{t_i, s_i, r_i\}$. We posit a logistic assignment model,

$$
a_i \mid x_i \sim \text{Bernoulli}\left(a_i \mid \text{logistic}(\phi_0 + \phi_1 s_i + \phi_2 r_i)\right).
$$

We place standard normal priors over the parameters and draw 1000 posterior samples, weighing observations by inverse propensity.

We first evaluate the assignment model with the average log-likelihood of the assignments. Figure 2a illustrates the test. The realized discrepancy seems plausible. Next we evaluate the outcome model, again with the mean squared error of the ATE. Figure 2b (left) illustrates the outcome test for the Poisson model. The test fails: the model lacks the overdispersion needed to capture the high variance of the data (Gelman and Hill, 2006). This is typical with the Poisson; the variance is equal to the mean.

We propose two alternative models. Model (b) replaces the Poisson likelihood with a negative binomial distribution, $\mu_i = t_i \exp\{\theta_0 + \theta_1 s_i + \theta_2 r_i + \theta_3 a_i\}$. It has the same mean as the Poisson but its variance increases quadratically. The variance is $\mu_i + \mu_i^2/\theta_4$, where θ_4 is a dispersion parameter. We place a gamma prior with unit shape and rate over θ_4. Model (c) has similar considerations, but the variance is now a linear function of the mean $\theta_4 \mu_i$.[4] Figure 2b (center and right) illustrates the causal tests for models (b) and (c). These results suggest that model (c) is the most plausible.

The Electric Company. We now consider an educational experiment performed around 1970 on a set of elementary

[4] This can be achieved with a quasi-Poisson regression (Gelman and Hill, 2006), but this is not a proper probabilistic model. Rather, we use a heteroscedastic Gaussian distribution with the same mean and variance.

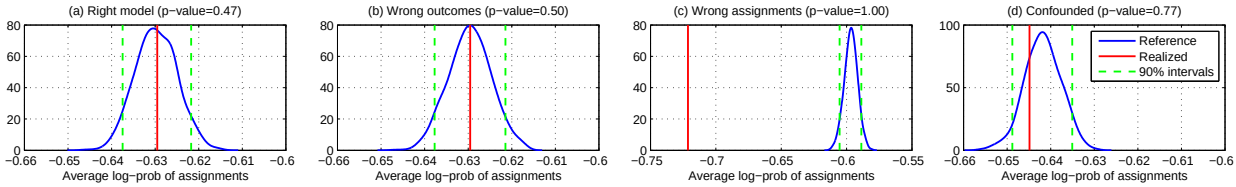
school classes. The treatment in this experiment was exposure to a new educational television show called The Electric Company. In each of four grades, the classes were completely randomized into treated and control groups. At the end of the school year, students in all the classes were given a reading test, and the average test score within each class was recorded. Our analysis is based at the classroom level, as we do not have access to the individual student scores.

Two classes from each grade were selected from each school. Let y_{i1} and y_{i2} be the scores of each class for the treatment and control groups, respectively. Let p_{i1} and p_{i2} be their pre-treatment scores at the beginning of the year. We also introduce the notation $g(i)$ to denote the grade (from 1 to 4) of the two classes from the i-th pair. We first use a Gaussian likelihood model of the form

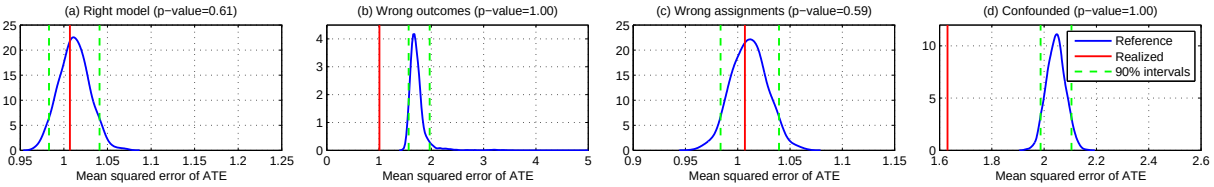
$$y_{i1} \sim \mathcal{N}(y_{i1} \mid b_i + m_{g(i)} p_{i1} + \theta_{g(i)}, \sigma^2_{g(i)}),$$
$$y_{i2} \sim \mathcal{N}(y_{i2} \mid b_i + m_{g(i)} p_{i2}, \sigma^2_{g(i)}),$$

where the model parameters are: b_i, which models the intercept term that depends on the specific pair i; $m_{1:4}$, which denotes the weight of p_{i1} for each grade; $\theta_{1:4}$, which models the treatment effect; and $\sigma^2_{1:4}$, which represents the variance for each grade. We place a Gaussian distribution over the intercept terms as $b_i \sim \mathcal{N}(\mu_{g(i)}, \tau^2_{g(i)})$. We also place Gaussian priors with zero mean and variance 10^4 over μ_g, θ_g, and m_g, as well as gamma priors with shape 10 and rate 1 over σ_g and τ_g. We refer to this model as model (a), and we also test two simplified models. Model (b) assumes that θ and m do not depend on the specific grade. Model (c) assumes instead that the intercept b is shared for all pairs. Since we know that this is a completely randomized experiment by design, we do not posit any assignment model.

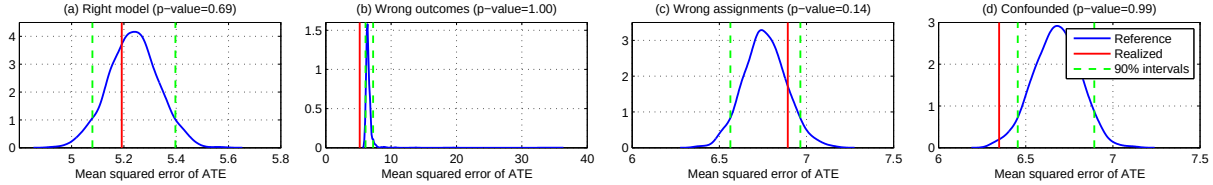
We plot in Figure 3 the results of the outcome test, which is based on the mean squared error of the average treatment effect. Model (a), which is the most flexible, seems to provide a sensible fit of the data. However, models (b) and (c) are too simplistic, and thus they clearly fail the test. If we had started from any of these models, the test would suggest the need to revise them.



(a) Results of the assignment test in the science-fiction scenario, in which we have access to both counterfactual outcomes. Model (c), which has a wrong assignment mechanism, fails the test. The plots for the fiction scenario (not shown) are similar to these ones, as the assignments are assumed independent of the outcomes.

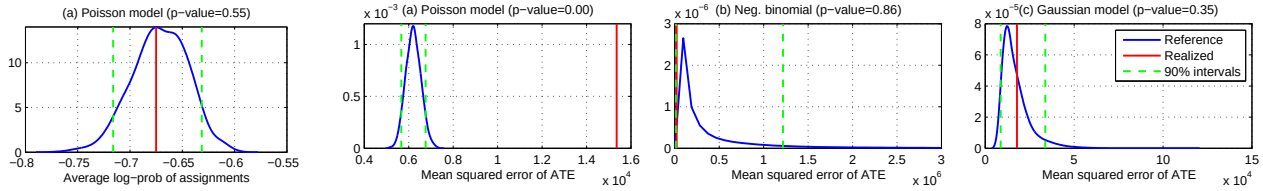


(b) Results of the outcome test in the science-fiction scenario. The test fails for the model in which the outcome model is mis-specified and for the confounded model.



(c) Results of the outcome test in the fiction scenario. The test fails for the model in which the outcome model is mis-specified and for the confounded model, and it also seems to suggest a flaw for the model in which the assignment mechanism is wrong.

Figure 1: Results of the tests for the synthetic experiments.



(a) Assignment test. **(b)** Outcome test.

Figure 2: Results for the cockroaches infestation study. The assignment test does not suggest any assignment model flaw. The outcome tests suggest that the variance is a linear function of the mean.

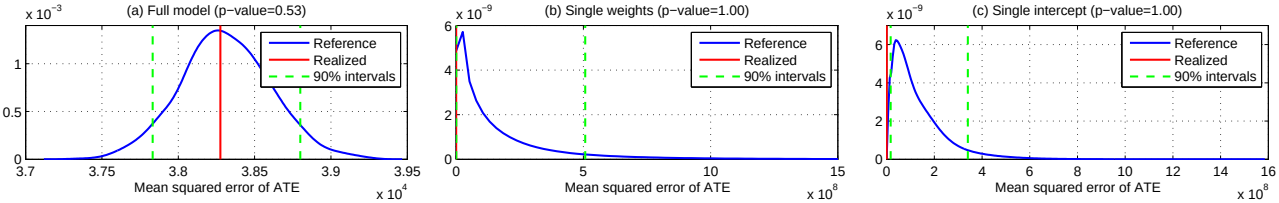


Figure 3: Results of the outcome test for the television show study. The models with a single weight for all grades (b) or a single intercept for all pairs (c) are too simple and fail the test.

References

Angrist, J. D. and Kuersteiner, G. M. (2011). Causal effects of monetary shocks: Semiparametric conditional independence tests with a multinomial propensity score. *Review of Economics and Statistics*, 93.

Athey, S. and Imbens, G. (2015). Machine Learning Methods for Estimating Heterogeneous Causal Effects. *arXiv preprint arXiv:1504.01132*.

Bayarri, M., Castellanos, M., et al. (2007). Bayesian checking of the second levels of hierarchical models. *Statistical Science*, 22(3):322–343.

Bickel, P. J. and Doksum, K. A. (1977). Mathematical statistics: Ideas and concepts.

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. (2013). Counterfactual reasoning and learning systems: the example of computational advertising. *The Journal of Machine Learning Research*, 14:3207–3260.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A. General*, 143(4):383–430.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424.

Gelman, A. and Hill, J. L. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4):733–760.

Gelman, A. and Shalizi, C. R. (2012). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38.

Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D. F., and Meulders, M. (2005). Multiple Imputation for Model Checking: Completed-Data Plots with Missing and Latent Data. *Biometrics*, 61(1):74–85.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Stat. Association*.

Huber, M. (2013). A simple test for the ignorability of noncompliance in experiments. *Econ. Letters*.

Imbens, G. and Rubin, D. B. (2015). *Causal Inference*. Cambridge University Press.

Krafft, P., Moore, J., Desmarais, B., and Wallach, H. M. (2012). Topic-partitioned multinetwork embeddings. In *Neural Information Processing Systems*.

Lloyd, J. R. and Ghahramani, Z. (2015). Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems*.

Mimno, D. and Blei, D. M. (2011). Bayesian checking of topic models. In *EMNLP*.

Mimno, D., Blei, D. M., and Engelhardt, B. (2015). Posterior predictive checks to quantify lack-of-fit in admixture models of latent population structure. *Proceedings of the National Academy of Sciences*, 112(26).

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. (2014). Distinguishing cause from effect using observational data: methods and benchmarks. *arXiv preprint arXiv:1412.3773*.

Neyman, J. (1923). On the application of probability theory to agricul- tural experiments. essay on principles. section 9. *Roczniki Nauk Rolniczych Tom X*.

Pearl, J. (2000). *Causality*. Cambridge University Press.

Peters, J., Bühlmann, P., and Meinshausen, N. (2015). Causal inference using invariant prediction: identification and confidence intervals. *arXiv preprint arXiv:1501.01332*.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*.

A Notation

We provide a table describing the notation we use in this paper. See Tables 1 to 3.

B Inverse propensity weighted estimator

Consider the proxy to the realized discrepancy for criticizing the outcome model,

$$T\left(\left\{\frac{\delta_{A_i=0}(a_i)}{p(a_i \mid x_i)} y_i(0)\right\}, \left\{\frac{\delta_{A_i=1}(a_i)}{p(a_i \mid x_i)} y_i(1)\right\}, \boldsymbol{\theta}\right), \quad (8)$$

where $\delta_{A_i=0}$ denotes a Dirac delta distribution with all probability mass at $A_i = 0$, and $p(a_i \mid x_i)$ is the assignment model's probability of observing assignment a_i given individual covariates x_i.

We explain in more detail that this is a valid realized discrepancy. First note that the two proxies for the outcomes are unbiased estimates to the true observed outcomes with respect to the population distribution:

$$\begin{aligned}
&\mathbb{E}_X[\mathbb{E}_{Y \mid X}[Y(0) \mid X = x]] = \\
&\qquad \mathbb{E}_X\left[\mathbb{E}_{p(A \mid X=x)}\left[\frac{\delta_{A=0}(0)}{p(A \mid X = x)} Y(0)\right]\right] \\
&\qquad \approx \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_{A_i=0}(a_i)}{p(a_i \mid X = x_i)} y_i(0),
\end{aligned}$$

where $\mathcal{D}^{\text{obs}} = \{a_i, y_i(a_i)\}$ represent observational data such that a_i are samples from the distribution of the density $p(A = a_i \mid X = x)$ and along with covariates $x_i \sim p(X)$ for some population distribution of covariates. The same applies for $\mathbb{E}[Y(1)]$.

Because the discrepancy of Equation 8 is a function (possibly non-linear) of unbiased estimates, by the delta method (Bickel and Doksum, 1977), it is an asymptotically unbiased estimator of the true realized discrepancy calculated over both sets of potential outcomes. Recall that as we observe infinite data $\{a_i, y_i(a_i)\}$, the posterior of $\boldsymbol{\theta}$ collapses to a point. Thus the realized discrepancy and the reference distribution both collapse to a point. Asymptotic unbiasedness ensures us that the predictive check always says the model fits the data poorly unless it is the true data generating process. Only in that case do the realized discrepancy and reference distribution collapse to the same point.

Note this proxy to the realized discrepancy is only an aymptotically unbiased estimator of the true realized discrepancy if its function evaluation indeed changes with respect to an increasing number of data points. This is true in the example we provide. However, this does not hold for example if the discrepancy is simply $T((\boldsymbol{y}(0), \boldsymbol{y}(1)), \boldsymbol{\theta}) = \boldsymbol{y}_4(1)$, the outcome of the 4th individual when assigned to treatment $a_4 = 1$.

Symbol	Description
A_i	Treatment assignment of individual i (random variable)
$Y_i(0), Y_i(1)$	Potential outcomes of individual i (random variable)
$\boldsymbol{A} = (A_1, \ldots, A_n)^\top$	Set of treatment assignments (random variable)
$\boldsymbol{Y}(0), \boldsymbol{Y}(1)$ $= (Y_1(0), \ldots, Y_n(0))^\top, (Y_1(1), \ldots, Y_n(1))^\top$	Set of potential outcomes (random variable)
a_i	Treatment assignment of individual i
$y_i(a_i)$	Outcome of individual i when assigned to treatment a_i
x_i	Observed covariates of individual i
$\boldsymbol{a}$	Set of treatment assignments
$\boldsymbol{y}(0), \boldsymbol{y}(1)$	Set of potential outcomes
$\boldsymbol{y}(\boldsymbol{a}) = (y_1(a_1), \ldots, y_n(a_n))^\top$	Set of outcomes when assigned to set of treatments
$\boldsymbol{x}$	Set of observed covariates
$\mathcal{D}^{\text{obs}} = \{a_i, y_i(a_i)\}$	Observed data set
$\mathcal{D}^{\text{do}} = \{a_i^{\text{do}}, y_i(a_i^{\text{do}})\}$	Hypothetical data set from an intervention

Table 1: Notation for observational data.

Symbol	Description
$\boldsymbol{\phi}$	Parameters of the assignment model
$p(a_i \mid x_i, \boldsymbol{\phi})$	Assignment likelihood for individual i
$p(\boldsymbol{a} \mid \boldsymbol{x}, \boldsymbol{\phi}) = \prod_{i=1}^n p(a_i \mid x_i, \boldsymbol{\phi})$	Assignment likelihood
$p(\boldsymbol{a} \mid \boldsymbol{x}, \boldsymbol{\phi}) p(\boldsymbol{\phi})$	Assignment model
$\boldsymbol{\theta}$	Parameters of the outcome model
$p(y_i(0), y_i(1) \mid x_i, \boldsymbol{\theta})$	Outcome likelihood for individual i
$p(\boldsymbol{y}(0), \boldsymbol{y}(1) \mid \boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i=1}^n p(y_i(0), y_i(1) \mid x_i, \boldsymbol{\theta})$	Outcome likelihood
$p(\boldsymbol{y}(0), \boldsymbol{y}(1) \mid \boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta})$	Outcome model
$p(\boldsymbol{y}(0), \boldsymbol{y}(1) \mid \boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) p(\boldsymbol{a} \mid \boldsymbol{x}, \boldsymbol{\phi}) p(\boldsymbol{\phi})$	Causal model
$p(\boldsymbol{a}^{\text{rep}} \mid \boldsymbol{x}, \boldsymbol{\phi}) p(\boldsymbol{\phi} \mid \boldsymbol{a})$	Assignment model (a posteriori)
$p(\boldsymbol{y}(0)^{\text{rep}}, \boldsymbol{y}(1)^{\text{rep}} \mid \boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y})$	Outcome model (a posteriori)
$p(\boldsymbol{y}(0)^{\text{rep}}, \boldsymbol{y}(1)^{\text{rep}} \mid \boldsymbol{x}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) p(\boldsymbol{a}^{\text{rep}} \mid \boldsymbol{x}, \boldsymbol{\phi}) p(\boldsymbol{\phi} \mid \boldsymbol{a})$	Causal model (a posteriori)

Table 2: Notation for causal models.

Symbol	Description
$(\boldsymbol{y}(0)^{\text{rep}}, \boldsymbol{y}(1)^{\text{rep}}), \boldsymbol{a}^{\text{rep}}$	Replicated data set of outcomes and assignments
$T((\boldsymbol{y}(0), \boldsymbol{y}(1)), \boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\phi})$	Causal discrepancy (over realizations)
$T^{\text{rep,s}} = T((\boldsymbol{y}(0)^{\text{rep},s}, \boldsymbol{y}(1)^{\text{rep},s}), \boldsymbol{a}^{\text{rep}}, \boldsymbol{\theta}^s, \boldsymbol{\phi}^s)$	Discrepancy over replication s
$T^{\text{obs,s}} = T((\boldsymbol{y}(0), \boldsymbol{y}(1)), \boldsymbol{a}, \boldsymbol{\theta}^s, \boldsymbol{\phi}^s)$	Realized discrepancy over replication s
$p(T) = \{T^{\text{rep,s}}\}$	Reference distribution
$T^{\text{obs}} = \{T^{\text{obs,s}}\}$	Realized discrepancy

Table 3: Notation for model criticism.