TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

Uncertainty Propagation in Causal Effects Estimation

Citation for published version (APA):
Kaptein, M. C. (2025). *Uncertainty Propagation in Causal Effects Estimation*. Technische Universiteit Eindhoven.

Document status and date:
Published: 06/02/2025

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

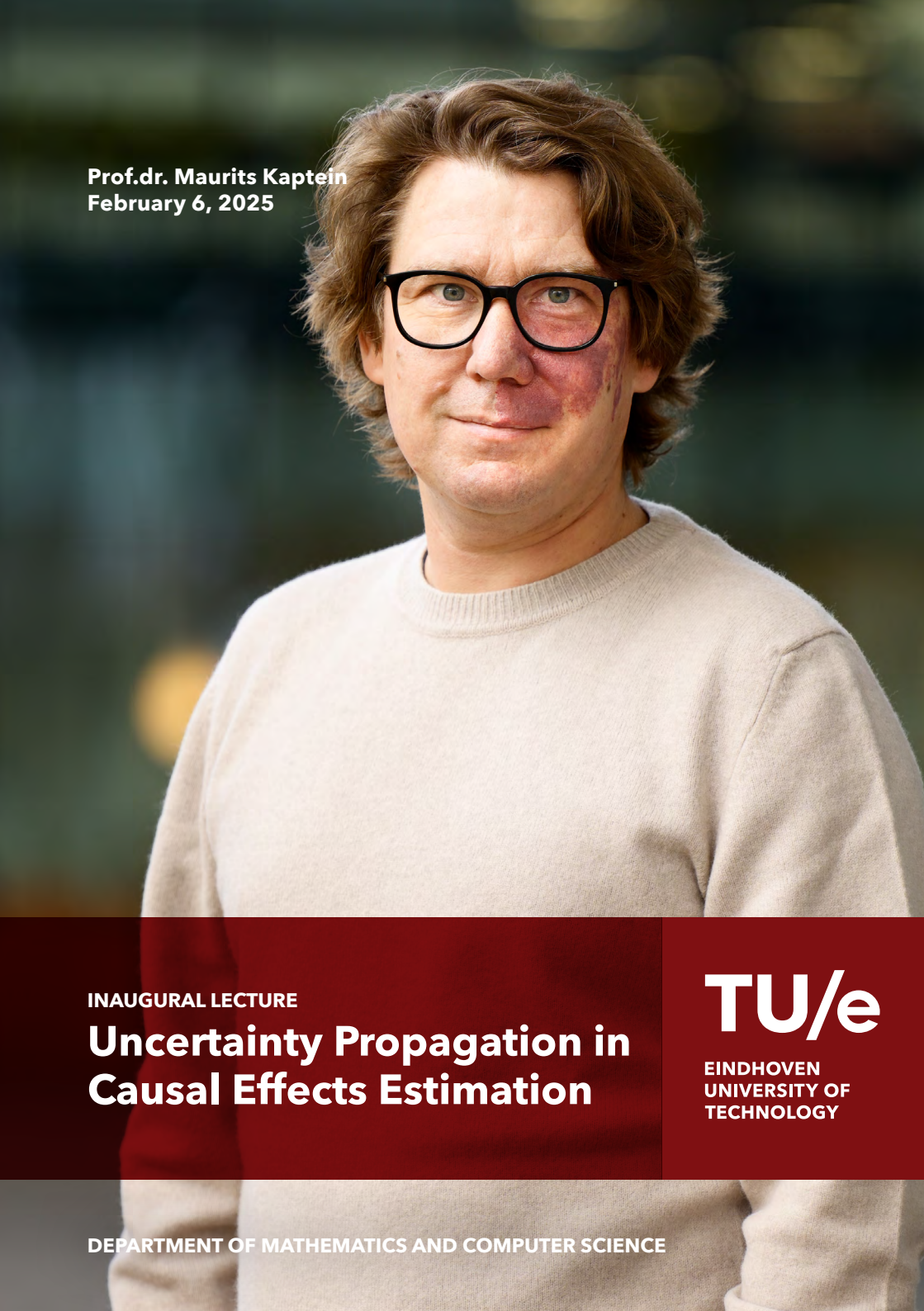
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Download date: 28. mars. 2025



Prof.dr. Maurits Kaptein
February 6, 2025

INAUGURAL LECTURE

Uncertainty Propagation in Causal Effects Estimation

TU/e
EINDHOVEN UNIVERSITY OF TECHNOLOGY

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

INAUGURAL LECTURE PROF.DR. MAURITS KAPTEIN

Uncertainty Propagation in Causal Effects Estimation

February 6, 2025
Eindhoven University of Technology

0. Introduction

"I wonder what happens when I press this button" is a common and practical question, but it can be surprisingly difficult to answer. Such questions belong to a broad category we call *causal* questions.[1] Colloquially, causal questions ask how actions (causes) lead to outcomes (effects). For example, pressing a button (cause) makes a light turn on (effect).

Causal questions arise in all scientific fields.[2] For example, in my own work, we considered a question about treating colon cancer: what effect does adjuvant chemotherapy (chemotherapy after surgery) have on the five-year survival rate? At first glance, this seems simple: take patients with colon cancer who have undergone surgery and ask, does giving chemotherapy after surgery improve their chances of being alive five years later compared to not giving it?[3]

While easy to ask, such causal questions are notoriously hard to answer. The core of the challenge is summed up by what is known as the *fundamental problem of causal inference*: we can never observe both the outcome that occurs under one action, for example providing chemotherapy, and an alternative action *simultaneously*. Hence, in one way or another, we always have to infer what would have happened if we had taken another course of action. Perhaps the light would have also turned on without pressing the button.[4]

Before proceeding, let me formally introduce the causal question of comparing the outcomes of two potential actions. For example, when studying the effect of chemotherapy on patient survival rates, we often focus on the average treatment effect (ATE). The ATE measures the difference in the expected outcome (*i.e.,* the

[1] In this written version of the originally spoken lecture, I will include **remarks** (1, 2, . . .), **technical details** (a, b, . . .), and **code examples** (i, ii, . . .). References to related literature will appear in the remarks and technical details, not in the main text.

[2] To highlight the growing interest in causality, here is a list of recent books on the topic: [Imbens and Rubin, 2015, MA and JM, 2020, Cunningham, 2021, Didelez and Pigeot, 2001, Peters et al., 2017, Pearl et al., 2016, Pearl and Mackenzie, 2018].

[3] We examined this specific question in earlier work, where we developed methods to compare causal effect estimates from experimental (Randomized Controlled Trial, RCT) data and observational data; see [Shen et al., 2024].

[4] The fundamental problem of causal inference has long been known and described in various ways by various authors. A compelling account, and also one of the starting points of the development of a formal treatment of the estimation of causal effects, is given in [Rubin, 1978].

average survival rate) between those who receive the treatment and those who do not. It is defined as:[a]

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

where $Y(0)$ denotes the potential outcome when no treatment is received, $Y(1)$ denotes the potential outcome when the treatment is received, and $\mathbb{E}$ denotes the expected value. Sometimes the dependence on a treatment, X, is made specific by denoting $Y(1) = Y(X =$ or $Y(1) = Y \mid X = 1$. Whatever the notation used, in practice we would like to be able to compute the ATE and we would like to be able to assess how certain, given only limited data, we are about our computed answer.[b] In this public lecture, I will discuss how we can use empirical *data* to estimate the ATE.[5] More broadly, I will explore how scientific *models* help answer questions like: should we give chemotherapy to maximize patient survival? To do this, I will first explain how we learn from observed data and how models summarize this knowledge to answer specific questions. I will define what scientific models are and emphasize their subjective nature: models are not 'true', but they might be useful. Next, I will highlight what I believe is missing in current model-building approaches for answering causal questions. Finally, I will conclude this lecture with a brief explanation of the title, "Uncertainty Propagation in Causal Effects Estimation," which many have, apparently, found difficult to grasp.

[5] I use the term *estimate* instead of *compute* intentionally. When we calculate the ATE using (limited) sample data, we obtain a quantity that serves as an estimate of the ATE in the population.

[a] This definition of the ATE uses the potential outcomes framework [Rubin, 2005]. The outcome of interest, Y (a random variable, denoted by capitals), differs for the potential outcomes $Y(1)$ (treatment received) and $Y(0)$ (treatment not received). $\mathbb{E}$ is the expectation operator. The fundamental theorem of causal inference highlights that for any individual i, $Y_i(1)$ and $Y_i(0)$ cannot be observed simultaneously.

[b] Admittedly, in the main body text, the notation used is a bit informal, but hopefully easy to follow. More technically, the notation $Y(1) = Y \mid X = 1$ seems like it would permit the following expression of the ATE: $\text{ATE} = \mathbb{E}(Y \mid X = 1) - \mathbb{E}(Y \mid X = 0)$. However, the latter implies that simply conditioning on the treatment is sufficient to compute the causal effect, which is very often not the case and depends heavily on the assumptions made and the data-generating mechanism or the underlying causal structure; I will return to these points later in the text.

1. Creating Scientific Models

One of science's main activities is to construct, based on data, models that we can use to answer specific questions regarding the world surrounding us. To make precise what I mean by the term model and to provide a first example, let us consider the situation which we are interested in understanding (or perhaps I should say 'modeling') the behavior of a very simple machine, $\mathcal{M}_1$. The machine works as follows: it has a single button[6] and each time the button is pressed, a single digit, 0 or 1, is printed onto a stretch of paper. A sketch of the machine is provide in Figure 1.

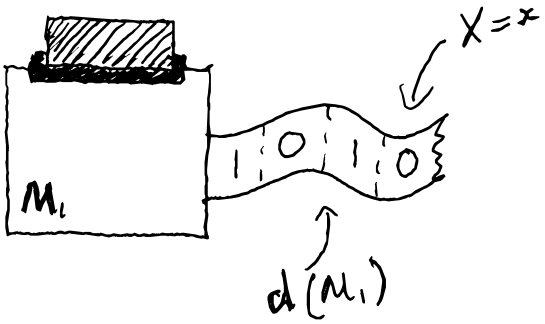


Figure 1. Machine $\mathcal{M}_1$ printing data (single digits) to its tape.

In an attempt to understand the machine, we press the button ten times to get the following data:[i]

$$d^1(\mathcal{M}_1) = (0, 0, 1, 0, 0, 0, 1, 1, 0, 0,).$$

What does it mean for us to create a scientific model of the machine based on such data? I would argue that our ability to properly model the machine is simply reflected by our ability to answer questions about the machine. One could imagine questions such as: "If we press the button once more, what will the next printed digit be?" Or, "How likely is it for the machine to print ten 0s in a row?" Obviously,

[6] Given the introduction, the effect of the button press might be a causal question in its own right. However, at this point I will consider the effect of the button a given.

not all meaningful questions might be known a priori, but I hope these examples provide a gist of the types of questions a scientist might want to answer about the machine.
Given our machine and the first batch of data from our machine, I can now clarify what I mean by a scientific model – being fully aware that this specific description of a model is a bit unorthodox.[7] In my view, for the purpose of the types of causal questions set out above, a scientific model consists of four components:

1. A set of *assumptions* regarding the process or object we intend to model. Often, regretfully, many of these assumptions are left implicit.
2. A (quantitive) representation of our knowledge derived from our observed data. I will call this the *data summary*.[8]
3. A *set of rules to update our data summary* based on newly observed data (*i.e.,* these rules dictate how we update (or potentially discard) our model given new observations).
4. A *set of rules to, given our assumptions and data summary, answer specific questions* (*i.e.,* these rules dictate how to, given our assumptions and data summary, answer specific scientific questions that the model aims to answer).[9]

Let me clarify each component by providing an example. For Machine *M*1, we could consider the following model:[10]

Model 1. *Parity of Pi.*

1. ***Assumptions:*** *We assume that the machine prints out the parity, the even or oddness, of consecutive digits or Pi using* 0 *for odd and* 1 *for even.*[11]
2. ***Data summary:*** *After the tenth button press, the last printed digit is understood to be the parity of the tenth digit of pi, and hence we summarize the data by simply using the index* $i = 10$.

[7] The definitions and uses of models in the sciences are hotly debated, see for example [Giere, 1999].
[8] This is admittedly a bit cryptical; it might be the full dataset, but often it is some function of the data, i.e., the computed parameters of some parametric model.
[9] In many cases in which we try to answer causal questions, for example by estimating the ATE, not all four components are made explicit or are complete. For example, in many applied papers answering causal questions, a number of more subtle assumptions – such as the Stable Unit Treatment Value Assumption [Rubin, 1980] – are left unmentioned. Additionally, when adopting a frequentist approach to modeling and answering questions – especially when answering questions using commonly known p-values – the update rules are very opaque: in the frequentist framework it is rare to make explicit how future observations will be incorporated into our models.
[10] Even in this description of the model, numerous details are left implicit for the sake of brevity of the lecture.
[11] . . and we assume we are the first one to ever press the button.

3. ***Update rules:*** *Our update rule would be to increment our index every time we print a new digit:* $i := i + 1$.[12]
4. ***Answer rules:*** *Our set of rules to answer questions like "What is the next printed number?" or "What is the probability that the machine will print out ten* 1*s in a row?" will be based on our understanding of the progression of the digits of Pi and are – once the digits of Pi can be computed – relatively easily answered. I won't detail them for sake of brevity.*[13]

From the above model construction, several points stand out. First, the accuracy of our model in answering questions depends heavily on the correctness of our assumptions. This reliance on assumptions is a common feature of all scientific reasoning. Often, the number of assumptions we could make – and thus the number of potential models we could build – is extremely large. Moreover, if two different models yield the same answers to the same questions using the same data, we cannot determine which of the two models is more valid based on the data alone. This makes scientific models inherently *subjective*.

Of course, we can improve models over time by discarding those that fail to answer our questions satisfactorily, but there is no guarantee that our models reflect the true underlying process.[14]
If history teaches us anything, it is that scientists do not seek 'the truth', despite claims to the contrary. Instead, we seek models that are useful.[15] A second thing to note, based on our above model construction, is the fact that – at least in theory – it is possible to evaluate competing models that provide different answers; the differing answers can be checked empirically. It is often possible to quantify the accuracy of our model by comparing the model's answer with the empirically observed answer. To be concrete, if the 11^{th} printed digit by Machine $\mathcal{M}_1$ is a 1, this would fully contradict our model as the 11^{th} digit of Pi is odd.

[12] I would like to stress here that I am personally very much a proponent of sequential learning approaches, i.e., new data becomes available as we go. Together with my PhD students I have worked on a number of different ways of dealing with sequential data in model estimation [Ippel et al., 2016, 2019], sequential decision-making [Provodin et al., 2022, Eckles and Kaptein, 2019, Kruijswijk et al., 2020], and sequential causal learning [Puha et al., 2020].

[13] Actually, computing consecutive digits of Pi once was an active area of study. Interesting references are [Bailey et al., 1997, Borwein and Borwein, 1987].

[14] What "true" means here is intentionally left vague. I will treat it as a broadly understood but poorly defined concept.

[15] I strongly oppose presenting science as a search for objective truth. This perspective has harmed public trust in science, as people struggle to reconcile the idea of truth with useful models that are eventually disproven. A more modest view of science, like Richard Feynman's description of science as "the pleasure of finding things out", is far more appealing. For a detailed discussion, see [Kaptein, 2021].

The above model construction mainly aimed to illustrate the specific components of a model and highlight the role of the underlying assumptions and the data summary. In the remainder of this lecture, I will focus more on *probabilistic* models: These are models that utilize the amazing richness of mathematical theory developed around random variables and random processes. These are not only models that are close to my heart; due to the existence of a rich mathematical theory for such models, the processes that are modeled using this approach benefit from established rules for updating and answering questions.[16] To illustrate a probabilistic model, let's examine an alternative model for Machine $\mathcal{M}_1$:

Model 2. *Realizations of a Bernoulli random variable.*

1. ***Assumptions:*** *We assume that our machine prints out realizations of a Bernoulli random variable.* i.e., *we assume* $P(x_i = k) = \theta^k(1 - \theta)^{(1-k)}$ *where the parameter* θ *is the probability of a* 1. *Note that a Bernoulli random variable – which might sound a bit intimidating for those uninitiated – can simply be interpreted as modeling the outcomes of a coin toss with a potentially biased coin (where* θ *is the probability of heads).*[c]
2. ***Data summary:*** *Given the data at hand consisting of observations* $x_1, x_2, \ldots x_n$, *we can build on a rich theory to estimate* θ *given the observed data. A simple maximum likelihood estimate is given by* $\hat{\theta} = \frac{\sum_{i=1}^{n} \mathbb{1}\{x_i=1\}}{n}$. *The full data summary would consist of the set* $\{\hat{\theta}, n\}$.[d]
3. ***Update rules:*** *Given an additional* m *observations, we can compute our updated* $\hat{\theta}$ *using* $\hat{\theta}' = \frac{\sum_{i=1}^{n} \mathbb{1}\{x_i=1\}+\sum_{i=n+1}^{m} \mathbb{1}\{x_i=1\}}{n+m}$.
4. ***Answer rules:*** *Within in this framework many rules to answer questions are available and the performance of a variety of rules is well-studied and understood. For example, our estimated probability of ten* 1*s in a row, given our stated assumptions is, given by* $P(10 \text{ consecutive } 1s) = \hat{\theta}^{10}$.[ii]

[16] Effectively, a number of the properties of the answers provided by probabilistic models are well-known. For example, we know the error rates of decision-making using frequentist p-values, as long as all our assumptions are correct. They never are.

[c] Or, more succinctly, we could denote a random variable $X \overset{\text{i.i.d.}}{\sim} \text{Bern}(\theta)$ and interpret each x_i as a realization of this random variable. This is a notation I will use later in this text.

[d] Here, as is the convention, I am using the $\hat{.}$ notation to denote that θ is estimated based on data. The maximum likelihood estimator presented in the body text – which can be interpreted as choosing the value of θ for which the complete dataset $d^n = x_1, \ldots, x_n$ is most likely under the assumed model – is just another subjective choice in our model-building exercise. For other estimation methods and a broader discussion see for example [Kaptein and Van den Heuvel, 2022].

I hope by now I have provided a first illustration of how we can build scientific models based on data and how we can evaluate the performance of our models based on new observations. We will need these ingredients in our quest to build models that allow us to answer causal questions. However, before extending the above model to facilitate a broader use, I want to briefly highlight that very often it is possible to provide visual summaries of a model. While admittedly not all too impressive, Figure 2 provides a visual summary of Model 2: we see the possible values of θ running from 0 to 1 and the actual (estimated) value of $\hat{\theta}$ denoted by c in the figure. As we will see later, for more complex models, such visualizations often become increasingly useful.

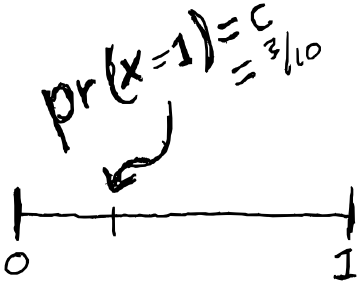


Figure 2. Visualization of the data summary of Model 2. The number line between 0 and 1 indicates the possible values of θ *and* c marks the estimated value $\hat{\theta}$.

2. Modeling uncertainty

Model 2 can be used to answer questions such as: "What is the probability of a sequence of length n of alternating digits?" Or, "What is the most likely number of 1s in a sequence of m new observations?" This model is surprisingly powerful, in a large part because, by assumption, it builds on the theory of probability and hence we know how to compute a large number of answers.[17]
However, I think Model 2 lacks a very important property. If you closely inspect Model 2, you will notice that the rules to update our knowledge, *i.e.*, our rules to compute $\hat{\theta}$, provide the exact same summary based on vastly different observed data. I think this is an undesirable property.[18] Consider our original data:

$$d^1(\mathcal{M}_1) = (0, 0, 1, 0, 0, 0, 1, 1, 0, 0),$$

leading to $\hat{\theta} = {}^3/_{10}$. Now, compare this to a dataset collected by pressing the button 50 times:

$$\begin{aligned} d^2(\mathcal{M}_1) \quad = \quad & (0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, \\ & 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 0). \end{aligned}$$

[17] Additionally, we often know how well these answers are supposed to hold up in empirical practice if our assumption(s) regarding the data-generating process are correct. Clearly, Model 2 would give erroneous answers if the true data-generating model was accurately captured by Model 1.

[18] Given that we have already surrendered to the fact that scientific models are subjective, I permit myself the liberty to motivate model choices in subjective terms such as "undesirable".

[e] One could argue at this point that using the tuple $\{\hat{\theta}, n\}$ allows the user of the model to use n specifically in answering questions and hence include the number of observations. However, it is currently much more common to solely use $\hat{\theta}$ in the computation of any answer. This changes when adopting a Bayesian framework, which is exactly what is proposed in Model 3.

If we estimate θ based on both datasets we regrettably obtain the exact same answer: $\hat{\theta} = {}^{3}/_{10} = {}^{15}/_{50}$.[iii]
Thus, our constructed Model 2 does not include a common-sense notion of gaining more certainty when collecting more observations: independent of the number of data points, the resulting estimate of θ turns out to be the exact same. Effectively, our assumptions – and our resulting data interpretation of $\hat{\theta}$ – do not provide the flexibility to treat dataset $d^1(\mathcal{M}_1)$ differently from dataset $d^2(\mathcal{M}_1)$, despite the fact that the two datasets are clearly different.[e]
In my view, if we want to take our scientific activity of developing useful models seriously, we should see whether can alter Model 2 to facilitate the above difference: we should find a way in which we can explicitly model the increased certainty regarding our computation of $\hat{\theta}$ that is provided by observing a larger dataset.
Luckily, building on centuries of development in probability theory, it is relatively easy to provide our model with the flexibility to actively model uncertainty and to distinguish the two distinct cases. By assuming that θ itself is uncertain and needs to be learned based on the observed data – where more data leads to more certainty – we can formulate a model in the following way:

Model 3. *Beta prior for a Bernoulli RV*

1. ***Assumptions:*** *We still assume that our machine prints out realizations of a Bernoulli random variable,* i.e.*, we assume* $P(x_i = k) = \theta^k(1 - \theta)^{(1-k)}$ *where the parameter* θ *is the probability of a* 1 *(as in Model 2). However, this time we further assume that, given the data,* $\theta \mid d(\mathcal{M}_1) \sim Beta(\alpha_n, \beta_n)$*. Thus,* θ *itself is a random variable whose likely values we learn based on the observed data.*

2. ***Data summary:*** *Whereas in Model 2 the core of our data summary was a single number ($\hat{\theta}$), we now use two numbers, α_n and β_n, to summarize our data. We compute $\alpha_n = \sum_{i=1}^{n} \mathbb{1}\{x_i = 1\} + \alpha_0$ and $\beta = \sum_{i=1}^{n} \mathbb{1}\{x_i = 0\} + \beta_0$.*[f]
3. ***Update rules:*** *Within this framework, α_n and β_n are easily updated when new data is observed. Given $m > n$ observations, we compute: $\alpha_{n+m} = \sum_{i=n+1}^{m} \mathbb{1}\{x_i = 1\} + \alpha_n$ and $\beta_{n+m} = \sum_{i=n+1}^{m} \mathbb{1}\{x_i = 0\} + \beta_n$.*
4. ***Answer rules:*** *Given that the above model builds on known assumptions, the answers to many questions are readily computed.*[19,iv]

Admittedly, at this point the story might become a bit tricky to follow for those unfamiliar with the larger mathematical theory that underlies Models 2 and 3. However, using the analogy of a biased coin, the functioning of Model 3 can easily be understood: Model 3 assumes the observations are generated by a biased coin whose probability of heads we are uncertain about. This uncertainty regarding the

[19] I will not detail these methods. However, any introductory Bayesian textbook will provide numerous examples; see for example [Gelman et al., 1995].

[f] A full treatment is as follows: let X be a Bernoulli random variable with parameter θ, where $\theta \in [0, 1]$. Assume a Beta prior distribution for the parameter θ:

$$\theta \sim \text{Beta}(\alpha, \beta),$$

with the probability density function:

$$p(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where α and β are the shape parameters of the Beta distribution. Given n observations $\{x_1, x_2, \ldots, x_n\}$ from a Bernoulli distribution, the likelihood function is:

$$p(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

Let $S_n = \sum_{i=1}^{n} x_i$ be the sum of the observed successes. Then, the likelihood can be rewritten as:

$$p(\mathbf{x} \mid \theta) = \theta^{S_n}(1-\theta)^{n-S_n}.$$

By applying Bayes' theorem, the posterior distribution of θ given the observations is:

$$p(\theta \mid \mathbf{x}) \propto p(\mathbf{x} \mid \theta)p(\theta).$$

Substituting the prior and the likelihood, we have:

$$p(\theta \mid \mathbf{x}) \propto \theta^{S_n+\alpha-1}(1-\theta)^{n-S_n+\beta-1}.$$

The posterior distribution is thus a Beta distribution:

$$\theta \mid \mathbf{x} \sim \text{Beta}(\alpha + S_n, \beta + n - S_n).$$

Thus, for the complete model, we have:

- Prior: $p(\mathbf{x} \mid \theta) = \theta^{S_n}(1-\theta)^{n-S_n}$
- Likelihood: $\theta \mid \mathbf{x} \sim \text{Beta}(\alpha + S_n, \beta + n - S_n)$
- Posterior: $\theta \sim \text{Beta}(\alpha_0, \beta_0)$

probability of success, and its dependence on the observed data, is visualized in Figure 3. On the left-hand side of Figure 3 we see what is often called 'our belief' regarding the value of θ resulting from our first set of observations, d^1. Our belief of the true value of θ – leaving aside philosophical discussions regarding the meaning of the word 'true' in this context – lies, by assumption, between 0 and 1 and the higher the line in the plot, the more probable that specific value. Thus, we see that $^3/_{10}$ is the most likely value, but other values are also very plausible. On the right-hand side we see how our belief is updated when observing the second dataset, d^2: The most likely value is still $^3/_{10}$; however, other values, for example the value $^8/_{10}$, are deemed much less probable.

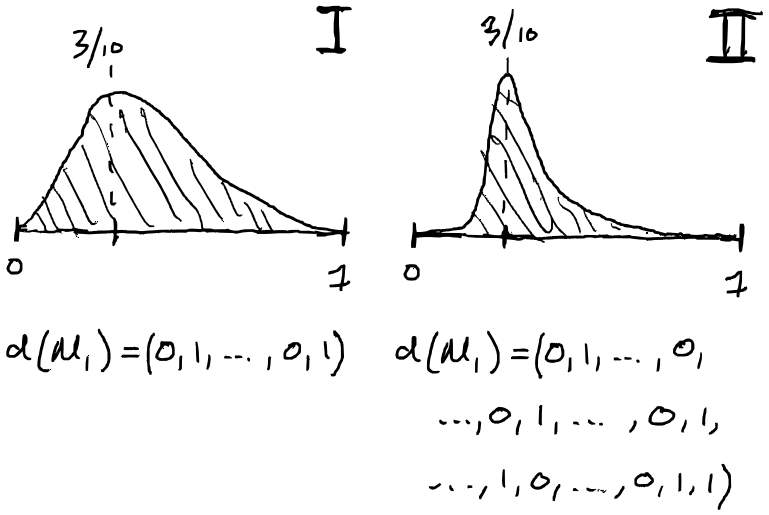


Figure 3. Two potential visual summaries of Model 3 given two possible datasets. Note that the curve indicates – given our observed data – the plausible values of θ. Given limited data (left-hand side), a large range of true values of θ is deemed likely. When more data is observed (right-hand side), values of θ that are unlikely to have generated these data are deemed implausible.

Model 3 can be thought of as an extension of Model 2 to include the flexibility of dealing with datasets of different sizes. Given Model 3 or, more broadly, models of the *type* of Model 3, we also know how to compute the answers to various questions. These answers are not always the same as those coming from Model 2, which means that the performance of these models can be compared empirically.

Model 3 is known as a Bayesian model; Bayesian statistics is a somewhat separate branch of probability theory which is very well-developed and, as well as having a strong philosophical and decision theoretic motivation, can simply be perceived as a mature mathematical theory that is superbly capable of generating models that are reasonably easy to compute, have consistent rules for updating knowledge based on observed data, and have a mature set of rules to answer various questions of which the validity is well-understood.[20] Bayesian approaches inherently include our *uncertainty* regarding our data summary in the models we use. This is a property which I deem essential to meaningfully answering applied causal questions. Hence, Bayesian approaches are my 'go-to' choice to further advance our understanding of causal questions.

[20] I will not be able to provide a full introduction to Bayesian thinking or estimation during this lecture. For a compelling account as to why Bayesian approaches are superior to alternative (frequentist) approaches please see [Jaynes, 2003]. For a more practically-oriented discussion of Bayesian estimation see [Gelman et al., 1995], and for a great introduction into Bayesian decision-making – *i.e.*, extending Bayesian thinking up to the point of seriously developing and evaluating answer rules – see [Robert et al., 2007].

3. Extending our models

I am aware of the fact that I still have not delivered on my promise to discuss causal models. Bear with me, I will. However, before I do so I will introduce one more model type – tailored to modeling more complex processes of interest – to set up the transition to causal models.
Consider a second machine that we are interested in understanding, which is shown in Figure 4. This second machine is virtually identical to our first machine in the fact that it has a button and, when pressed, the machine prints digits onto a stretch of paper. However, this time, at each button press, two digits are printed. After ten presses we obtain the following dataset:[v]

$$d^3(\mathcal{M}_2) = \left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\0\end{pmatrix}\right).$$

Hence, the machine prints two digits every time the button is pressed. We call this multivariate data, as opposed to univariate.
To model the behavior of this machine, we can simply extend the univariate (i.e., single variable) Bayesian model, Model 3, to include two variables. In this case we assume at each button press a realization from a so-called *bivariate Bernoulli random variable* is generated. I will present the Bayesian model for such data here without delving into the details:

Model 4. *A Bayesian Bivariate Bernoulli model*

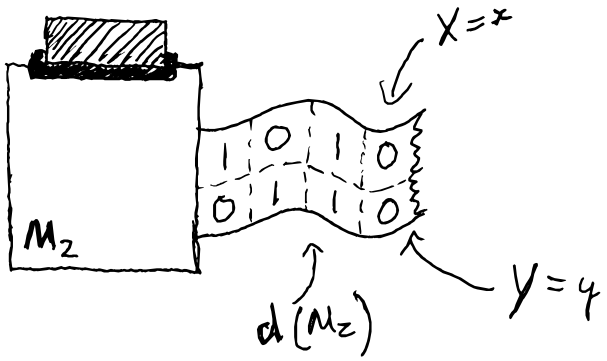


Figure 4. Machine $\mathcal{M}_2$ printing data (pairs of digits) to its tape.

1. ***Assumptions:*** *We still assume that our machine prints out realizations of a Bernoulli random variable, this time a Bivariate Bernoulli. Thus, we assume the likelihood of the data* $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$*, to be:*

$$P(\{(x_i, y_i)\}|\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11}) = \theta_{00}^{n_{00}}\theta_{01}^{n_{01}}\theta_{10}^{n_{10}}\theta_{11}^{n_{11}}$$

   where n_{00}*,* n_{01}*,* n_{10} *and* n_{11} *are the counts of the tuples* (0, 0)*,* (0, 1)*,* (1, 0)*, and* (1, 1) *respectively.*[g]
2. ***Data summary:*** *Given the prior distribution*

$$(\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11}) \sim Dirichlet(\alpha_{00}, \alpha_{01}, \alpha_{10}, \alpha_{11})$$

   our data summary (or posterior) is given by:

$$(\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11})|\{(x_i, y_i)\} \sim Dirichlet(\alpha_{00} + n_{00}, \alpha_{01} + n_{01}, \alpha_{10} + n_{10}, \alpha_{11} + n_{11})$$

3. ***Update rules:*** *Similar to Model 3, given new data, the data summary is easily updated by including all* $m > n$ *data points in the computation of the data summary.*
4. ***Answer rules:*** *Again, given that the above model builds on known assumptions, the answers to many questions are readily computed.*[vi]

While Model 4 looks even more unwieldy than Model 3, it is useful too not get to distracted by the mathematical details; effectively, we are modeling two coins, each with *a)* their own probability of success and *b)* a potential relationship between them.

As before, we can summarize the model by looking at a visualization. Figure 5 presents a visualization of the model's parameters (i.e., the summary of the data) after observing the $d^1(\mathcal{M}_2)$. The Figure is a bit harder to read as it is a so-called contour plot; we are looking at the strength of our belief, quantified using probabilities, of a 1 for each of the two digits, $P(X)$ and $P(Y)$ in the plot. We can see that the most likely value of $P(X) \approx .6$ and the most likely value of $P(Y) \approx .25$.[h]

[g] Here, we are using the notation $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ to denote the observations from both the first row of data $x_1, \ldots, x_n$ and the second row $y_1, \ldots, y_n$.

[h] Note that given the relatively small sample used, the posterior means are actually quite far away from the true data-generating process. Please see code snippet **v** and **vi** for the generation of the data and the construction of the Figure.

Also note that there is a dependency between the two variables: if the first of the two digits is a 1, it is more likely for the second of the two digits to also be a 1.[i]

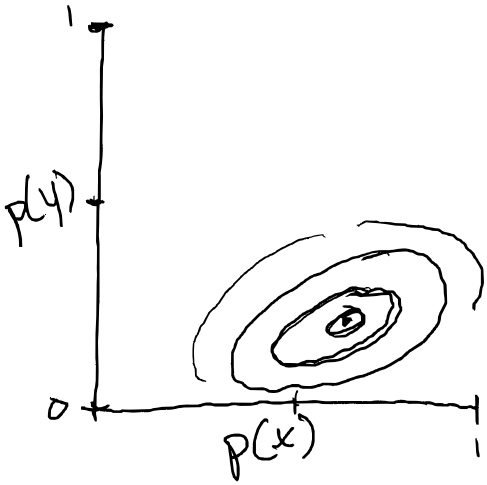


Figure 5. Visual summary of Model 4 using a contour plot (which can be interpreted as an elevation map of our belief in the probability of a 1 for each of the two digits given our observed data.

Before transitioning into our causal question, I would like to note that we have effectively used this Bayesian Bivariate Bernoully model to model – and make decisions – in meaningful applications. In work that I supervised together with Prof.dr. Joris Mulder, one of my former PhD students, Xynthia Kavelaars, used this model to understand the outcomes of brain radiation treatment.[21] In this case the outcomes (measurements of individual patients) are measured in terms of both cognitive decline (the first digit) and reoccurrence of the brain tumor (the second digit). We used the exact same Bayesian Bivariate Bernoulli model as presented here to summarize the data and we completed our contribution by adding novel answer rules to the literature.

[21] In a sequence of papers Dr. Kavelaars developed statistical methods to deal with multivariate data and make meaningful decisions for clinical trials. See [Kavelaars et al., 2020, 2023, 2024].

[i] In the actual contour plot generated using the code in code snippet **vi**, the dependency is hardly visible due to the small number of observations. In the illustrations in the body text I have exaggerated the dependency for illustrative purposes.

4. The causal data-generating machine

Now, let's finally turn to causal models. I will start by constructing a new machine which generates data that are interesting in terms of informing our thinking about causal models. Note that I have been using these machines simply as simple analogies to focus our thinking solely on how we deal with observed data, as opposed to being distracted by the larger, applied context. To carry on this experiment, suppose we have a machine which consists of our by now familiar button, our tape, and, as a novel addition, two switches (see Figure 6). Each switch has three positions which are marked '1', 'Ø', and '0'. Upon a first glance, the machine seems to perform the same as machine 2: pressing the button prints two digits to our tape. However, when fiddling a bit with the switches on the machine we quickly learn that setting the first switch to 1 leads to the first digit being a 1 and similarly for the second digit. Hence we can now represent the data as follows:

$$d^1(\mathcal{M}_3) = \begin{pmatrix} 1_1, 1_1, 1_1, 0_\varnothing, 1_\varnothing \dots \\ 0_\varnothing, 1_1, 1_\varnothing, 0_0, 0_0 \dots \end{pmatrix}.$$

In the above representation I use the subscripts of each printed digit to indicate the position of the associated switch at the time the digit was printed. It is clear that each time a switch is set to a specific value, that value is printed. If the switch is set to Ø, the machine prints 0s and 1s similar to a coin toss, as before.[22,vii]

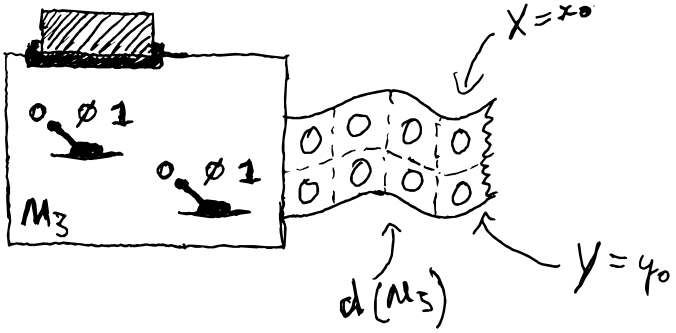


Figure 6. Machine $\mathcal{M}_3$ printing data (pairs of digits) to its tape with switches set to 0 for X and 0 for Y.

[22] I will use this subscript notation both on the realized values – to indicate the position of the lever when sampled – as well as on the random variable(s).

Note that this new machine maps to our causal question(s) in the following way: if we consider the first printed digit, X, to represent the chemotherapy treatment, with 0 representing 'no treatment' and 1 representing 'treatment', and if we further consider the second digit, Y to represent the five-year survival rate, we can interpret the first switch as indicating whether or not we *explicitly* administered chemotherapy (in the cases of 00 or 11) or whether we simply received data with yes or no chemotherapy, but we are uncertain about the mechanism that led to the administration of the chemotherapy (i.e., the switch was set to Ø). Next, given the data, we aim to compute – eventually, once we have developed a proper model to do so – the ATE that we saw in the introduction (Eq. 1). Making the dependence on the switches explicit, we can now denote the ATE as follows:[j]

$$\mathrm{ATE} \quad = \quad \mathbb{E}(Y \mid X_1 = 1) - \mathbb{E}(Y \mid X_0 = 0).$$

Thus, we are interested in the difference in the expected outcome of Y, the second digit, when we explicitly set the switch to the first digit, X, to fix its values to 1 or to 0.

[j] I am choosing this population notation here for simplicity: to compute an estimate of the ATE given sample data one could imagine using the following estimator:

$$\hat{\mathrm{ATE}} = \frac{1}{n_1} \sum_{i \in \mathcal{T}} Y_i - \frac{1}{n_0} \sum_{i \in \mathcal{C}} Y_i$$

where:

- $\mathcal{T}$: The observations under X_1, i.e., the switch on the first digit set to 1.
- $\mathcal{C}$: The observations under X_0. i.e., the switch on the first digit set to 0.
- $n_1 = |\mathcal{T}|$: Number of observations under X_1.
- $n_0 = |\mathcal{C}|$: Number of observations under X_0.
- Y_i: Observed outcomes, the second digits, for each button press i.

5. The ways of understanding causal machines

Interestingly, we now have two meaningful paths in front of us to estimate the ATE given data from Machine $\mathcal{M}_3$. One way to go is to simply focus solely on the data that was generated under the data-generating mechanism of interest. Thus, we discard all instances when the first switch is set to Ø.[23] Since the observed data when the switch is set explicitly to 0 or 1 represents the exact case(s) we are interested in, we can relatively safely assume that the data generated under these settings is representative for the future data and can thus be used to provide answers.[24]

Before turning to the second possibility, let's quickly discuss what this approach implies in real-world applications. If we are interested in the effect of chemotherapy on patient survival rate, the above implies that we should not look at data collected 'in the field', *i.e.*, data for which we do not know whether a switch was set explicitly. We should focus solely on data for which we know exactly why and how it was generated, i.e., we know how the switch was set. This path leads us towards what is called the randomized controlled trial (RCT): if we randomly assign the treatment or control to patients, we are in full control (*i.e.*, we set the switch) and we know that

[23] And obviously also those where the second switch is set to 0 or 1.

[24] Note the "relatively safely assume" here: even for this to hold we are making assumptions. For example, we will have to assume that the X switch setting on button press i does not influence the outcome (Y) of button press $i + 1$. Effectively, the latter is the well-known SUTVA assumption [Imbens and Rubin, 2015].

the data generated under such a scheme is representative for the situation we are actually interested in.[k]

I would like to provide an intuitive example of the problems that occur when we do use data collected 'in the field' in practical applications, i.e., data for which we do not know the data-generating process. Consider the applied case of chemotherapy treatment for breast cancer patients. When naively looking at the difference in survival rates between patients who receive chemotherapy versus those who don't using data collected 'in the field' – i.e., when the switch is set to Ø – one would be inclined to conclude that chemotherapy *decreases* the life expectancy. However, in clinical practice, both chemotherapy and a decreased survival rate have a common cause: the severity of the tumor affects both the probability of chemotherapy (i.e., a severe tumor is less likely to be surgically removed in full), and the survival rate (patients with a severe tumor are less likely to survive). Looking solely within the severe tumor group, we find that survival rates are higher for those patients who receive chemotherapy and thus should conclude that chemotherapy is effective.

I hope this example illustrates that it is essential to know 'how the switch was set'. Looking at the data generated under the mechanism of interest – i.e., knowing how the switch was set – is luckily not the only way to meaningfully answer causal questions: a meaningful second approach exists. I say "luckily" here because in some applied cases it is absolutely unethical to generate data under the mechanism of interest: if there is doubt regarding the negative effects of smoking, no ethical committee would approve a study that randomly assigns a sample of people to be forced to smoke for a few years. We are currently experiencing a similar problem in trying to decisively understand the causal effect(s) of various pesticides.

[k] Technically, we need to only know the so-called *propensity score*, which is the probability of receiving the treatment conditional on any co-variates. In our running example of the machines, this would be akin to the knowledge that the switch on X is set randomly (with some known probability) to 0 or 1 for each i (and thus never to Ø). To subsequently provide an unbiased estimate of the ATE using propensity scores, especially when considering more complex systems consisting of more variables, we will need to make a few common assumptions: we assume *conditional independence* given the observed co-variates (i.e., all co-variates are observed), *overlap*, meaning that for all values of the treatment X the probability of being selected is neither 0 or 1, and $SUTVA$, which we saw in an earlier note. Given these assumptions, the propensity score can be used in various ways to compute an (unbiased) estimate of the ATE, for example using inverse propensity score weights (IPW):

$$\hat{\mathrm{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i Y_i}{e(X_i)} - \frac{(1-Z_i)Y_i}{1-e(X_i)}\right)$$

where:

- Z_i: Treatment indicator ($Z_i = 1$ for treated, $Z_i = 0$ for control).
- Y_i: Observed outcome for individual i.
- $e(X_i) = P(Z = 1 \mid X_i)$: Propensity score for individual i.
- n: Total number of individuals in the sample.

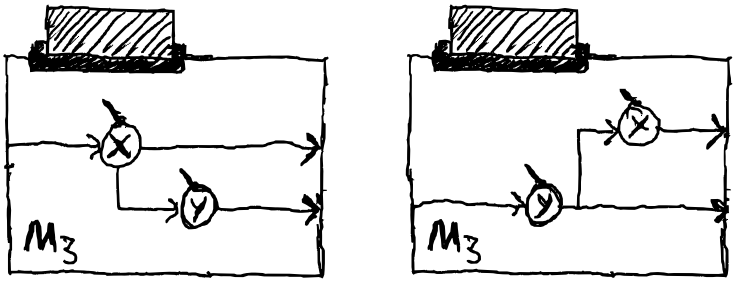


Figure 7. Two potential wirings inside Machine $\mathcal{M}_3$.

When working with this second approach, we are less concerned with the procedure by which the data was generated, but we are interested in the inner workings of the machine. Consider Figure 7, which abstractly represents two possible inner wirings of the machine. On the left-hand side we see – reading from left to right – that the button press causes a (noisy) generation of the value for X, the first digit. Next, the *value of X directly affects the (noisy) value of Y*. Once we know that this is the case, we can easily deduce that the setting of the switch on X is irrelevant: we can simply compute the causal effect we are interested in by computing:

$$\text{ATE} \quad = \quad \mathbb{E}(Y \mid X_{1 \vee \varnothing} = 1) - \mathbb{E}(Y \mid X_{0 \vee \varnothing} = 0),$$

where I am using the subscript notation $1 \vee \varnothing = 1$ to mean any data generated while the switch was set to 1 or to $\varnothing$.[1]

This second method potentially gives us more data to use than the first method. Effectively, *given that* our assumption regarding the inner workings of the machine is correct, we can use data under multiple data-generating mechanisms to properly estimate the causal effect. However, what if the inner workings of the machine in actuality are those depicted on the right-hand side of Figure 7? In this case the button press first generates Y, only after which X is generated. The switch on X

[1] Again, for sake of readability, I am using a population specification; the sample version would be:

$$\hat{\text{ATE}} = \frac{1}{n_1} \sum_{i \in \mathcal{T}} Y_i - \frac{1}{n_0} \sum_{i \in \mathcal{C}} Y_i$$

where:

- $\mathcal{T}$: The occurrences of $x_i = 1$ under both X_1 or $X_\varnothing$.
- $\mathcal{C}$: The occurrences of $x_i = 0$ under both X_0 or $X_\varnothing$.
- $n_1 = |\mathcal{T}|$: Number of occurrences of $x_i = 1$ under both X_1 or $X_\varnothing$.
- $n_0 = |\mathcal{C}|$: Number of occurrences of $x_i = 0$ under both X_0 or $X_\varnothing$.
- Y_i: Observed outcome for each button press i.

thus affects the value of X, but, due to the wiring, cannot directly affect the value of Y. Hence, in this case, including data collected when the switch on X is set to Ø would potentially provide an erroneous answer: we would mistakingly confuse any correlation between X and Y for causation.[m]

Abstractly, we often represent the so-called causal structure underlying the machine using a graphical representation called a (causal) directed acyclic graph (DAG).[25] Figure 8 provides three more abstract examples of potential underlying causal structures when two variables are concerned: either X causes Y, Y causes X, or the two have a common, potentially unknown, cause.

To complete our more abstract description of the inner workings of the machine we often specify what is called a causal structural model.[26] For example, the structural causal model,

$$X \sim U_x$$
$$Y \sim f(X, U_y),$$

specifies that X is generated through some unknown random process U_x – for example, the biased coin flip we saw earlier – and *Y is a function of X and some additional unknown "noise", U_y*. Interestingly, once we know the underlying structural causal model – i.e., the inner workings of the machine – we do not need to know the data-generating mechanism anymore and we have, over the last decade, developed accurate models to answer causal questions irrespective of the data-generating mechanism.[27] In its bare essence, looking at Figure 8, when X and Y have a common cause (and no direct relationship) or when Y causes X, we simply know that, by the design of the machine, the $\mathrm{ATE} = 0$. If we are fully confident X causes Y (scenario iii) we can use any pair of digits, whether the switch on X was set to 0, 1, or Ø, to compute the ATE.

[25] This second approach to answering causal questions has been spearheaded in recent decades by Pearl (see [Pearl, 2000, Pearl et al., 2016, Spirtes et al., 2000]) although it builds on earlier thinking in various fields. A lot of this literature focuses on either a) determining which variables to statistically correct for given a known underlying structure (see [Robins and Greenland, 2001, Richardson and Spirtes, 2013]) or b) discovering a structure based either on conditional independencies or other assumptions (see [Pearl, 1993, Meek, 1995, Spirtes and Zhang, 2000, Kalisch and Buhlmann, 2007, Peters et al., 2016, Chickering, 2002]).

[26] For an introduction, see [Pearl, 2000] or [Peters et al., 2017].

[27] A very nice introduction to the various ways we can use assumptions regarding the structural causal model – assumptions beyond just the DAG – is given in [Peters et al., 2017].

[m] To be more precise, in this case the ATE we have been discussing, $\mathrm{ATE} = \mathbb{E}(Y \mid X_{1 \vee \varnothing} = 1) - \mathbb{E}(Y \mid X_{0 \vee \varnothing} = 0)$, is simply 0. However, there might still be a dependency between X and Y, and thus a naive estimate, $\mathbb{E}(Y \mid X) - \mathbb{E}(Y \mid X)$, that disregards the causal structure might provide a non-zero estimate.

6. Building a causal model

As this whole endeavor is getting somewhat complex by now, let me collect some of the pieces:

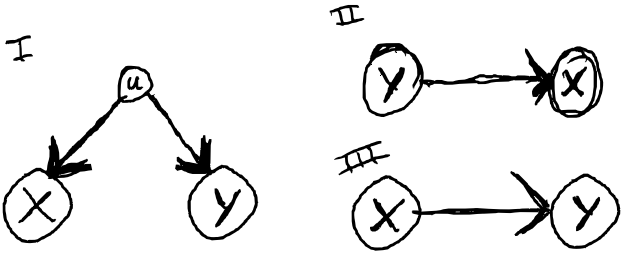


Figure 8. A more abstract representation of a number of potential wirings of Machine $\mathcal{M}_3$ using (causal) directed acyclic graphs (DAGs).

- We are interested in causal questions. For example, we are interested in computing the average treatment effect (ATE).[28]
- We build models to understand processes based on data. A model is composed of assumptions, a data summary, update rules, and answer rules. Given an appropriate model, fair assumptions, and sufficient data, we should be able

[28] I would like to stress here that this is just one of the potential causal questions we might be interested in. Ideally, a developed causal model allows us to generate answers to multiple questions given a single set of assumptions, data summary, and updating rules.

to provide meaningful answers. The term meaningful can be made precise by studying properties of the answers.[n]

- Probabilistic models are useful because they are well-suited to modeling processes that involve randomness or noise. Additionally, probabilistic models build on hundreds of years of theory development, which makes it possible for us to evaluate properties of the answers which such models provide (at least given some specific assumptions).[o]
- Bayesian models – a specific subset of probabilistic models – enable us to meaningfully model certainty caused by the fact that we have limited data at our disposal. Using Bayesian probabilistic models, the expected accuracy of our answers naturally varies with the amount of observed data.[29]
- Causal questions are particularly challenging: due to the fundamental problem of causal inference, we need additional assumptions to compute causal answers. More precisely, we often need an understanding of either the data-generating mechanism (i.e., which switch was set?) or the underlying causal model (i.e., the inner workings of the machine) to generate valid answers to causal questions.[30]

[29] I do not want to stress the philosophical superiority of Bayesian approaches over frequentist approaches too much in this lecture; however, for those interested (or not yet convinced), see [Jaynes, 2003]. For more details on how to use Bayesian approaches to generate meaningful answers using Bayesian decision-theory, see [Robert et al., 2007].

[30] At this point it is good to note that besides the potential outcome approach to causal inference (which leads to estimators using propensity scores) and the structural causal model approach (which leads to sufficient adjustment sets), there are other approaches to causality. Notably there is a lot of work on causality using time series. See for example [Granger, 1969, Hamilton, 1994, Lütkepohl, 2005, Broszeit and Hansen, 2019].

[n] Considering desirable properties of estimators is common practice in statistical research. To make the types of properties I would consider here more concrete, consider quantifying, using simulation, the Type 1 error rate of a two-sample t-test under the null hypothesis (i.e., we assume that there is no difference between the means of two groups). Such a simulation would consider

- Null Hypothesis: $H_0 : \mu_1 = \mu_2$ (no difference in means between the two groups).
- Alternative Hypothesis: $H_1 : \mu_1 \neq \mu_2$.

We would simulate data for two groups, each following a normal distribution with mean 0 and standard deviation 1, under the null hypothesis (i.e., $\mu_1 = \mu_2 = 0$). For each simulated sample, we perform a two-sample t-test to test whether the means are significantly different. The p-value from the t-test is compared to a significance level $\alpha = 0.05$. If $p < \alpha$, we reject the null hypothesis. The Type 1 error rate is calculated as the proportion of simulations where the null hypothesis was rejected, despite the fact that it is true. Thus, the Type 1 error rate is computed as the proportion of simulations where we incorrectly reject the null hypothesis:

$$\text{Type 1 Error Rate} = \frac{\text{Number of rejections of } H_0}{\text{Total number of simulations}}.$$

Since the null hypothesis is true (i.e., $\mu_1 = \mu_2$), we expect the Type 1 error rate to be close to the significance level $\alpha = 0.05$ in this case.

[o] Just to follow up on the remark regarding Type 1 errors: one of the main advantages of using probabilistic models is the fact that many properties of answers provided by such models – if the assumptions hold – are well-known (even analytically in many cases).

Given all of the above, my quest in the years to come will be to develop causal models, with explicit assumptions, data summaries, update rules, and answer rules, that we can use to answer causal questions. I think these models should properly incorporate what we know about the data-generating mechanism *and* what we know about the underlying causal model. These models should also actively include our uncertainty: in causal models, uncertainty arises in many places, including uncertainty due to small samples (as encountered earlier), uncertainty regarding the data-generating mechanism (i.e., 'in-the-field' data collection), and uncertainty regarding the underlying causal structure.
The development of causal models will rely on a number of active strands of the scientific literature which are, regrettably, mostly disjoint: there is a very active literature on the development of Bayesian estimation methods to ensure that we can properly model uncertainty. However, this literature hardly interacts with the mostly methodological literature on understanding data-generating mechanisms[31], which is an active literature in its own right. In turn, both of these strands of the literature hardly interact with the literature regarding structural causal models, which has been booming in recent decades. In my work in the coming years, I aim to combine all of these ideas into a single type of causal model and study the properties of the answers provided by such models.
Let me provide a first attempt at specifying such a causal model and the necessary components. This is admittedly not finished – I aim to study this topic for the next few years – but the contours are relatively clear. Let me call this class of models *Compound Causal Models* and provide a specific example:

Model 5. *Compound Causal Models*

1. ***Assumptions:*** *While admittedly it is extremely hard to provide a full overview of all implicit and explicit assumptions that go into a model, I suggest that, building on the current literature, any bivariate data — such as those originating from Machine $\mathcal{M}_3$ — are treated as realizations of potentially related random variables, and hence we can describe these data in the binary case using the Bayesian bivariate Bernoulli model introduced earlier (used in Model 4). Additionally, we assume that we either know the data-generating mechanism (i.e., we know that the data was or wasn't generated through explicit manipulation) or we are able to enumerate the possible underlying causal structures and quantify our belief that one of the potential structures is the true structure.*

[31] This despite the fact that early work in this space was inherently Bayesian, see [Rubin, 1978].

2. ***Data summary:*** *Our complete data summary, given our assumptions, should contain:*
   - *A description of the distribution of Y when X is manipulated, i.e., $P(Y \mid X_x = 0$ and $P(Y \mid X_x = 1)$. To actively model our uncertainty regarding this quantity we use a Bayesian approach and, in the binary outcome case, we assume likelihood $Y \mid X \sim$ Bernoulli(θ) and prior distribution $\theta \sim$ Beta(α_x, β_x). The resulting posterior distribution, akin to Model 3, completes our summary of the data under the condition that $X_0 = 0 \vee X_1 = 1$.*
   - *A description of the distribution of X when Y is manipulated, which is similar to the one described above, simply reversing X and Y.*
   - *A description of the joint distribution of X and Y when the manipulation is unknown, i.e., the switches are set to Ø. We use Model 4 to model $P(X, Y)_\varnothing$ for all data generated under this data-generating scheme. Here, the posterior distribution $(\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11}) \mid \{(x_i, y_i)\}$ provides the data summary for this data-generating mechanism.*
   - *Finally, the probability that the various possible data-generating structures described in Figure 8 are true. Thus, we need $P(\mathcal{G}) = P(G_k = c_k), \forall k$, the probability that graph Gk is truly the underlying graph given our observed data.*[p]
3. ***Update rules:*** *Given that the distribution of Y given X, X given Y, and the joint of X and Y are all known objects, we know the update rules under various data-generating mechanisms (see earlier model descriptions). For $P(\mathcal{G})$, regrettably, we currently do not have any well-defined update rules: we do not know how to compute the probability that a certain underlying causal structure, represented as a graph, is the true underlying structure given the data that we have.*[32]
4. ***Answer rules:*** *Given the assumptions and provided the data summary, we end up with a set of distributions describing X and Y under various data-generating mechanisms and a description of the probability of each underlying causal structure. Jointly, these summaries should suffice to create rules that allow us to compute meaningful causal answers – such as the ATE – whose properties we can study.*[viii]

[32] A notable direction in this context is working on scoring different DAGs, see for example [Gonzales and Valizadeh, 2024, Montagna et al., 2023].

[p] In a true Bayesian fashion we should not just have individual probabilities for each graph, but rather, given our data, a posterior for each graph. For the sake of illustrating how to compute the ATE later on I will stick to probabilities. However, a fully Bayesian framework is one of the targets of my future work.

Before providing a gist of how the ATE could potentially be computed using a Compound Causal Model, let me be explicit about the components of the model I just described that are known and the ones that we do not know. A number of components are currently reasonably well-understood:

- Assumptions: Given the combined study of causal questions in multiple strands of the literature, we know a lot of the assumptions we need to make to ensure that knowing the data-generating mechanism or the causal structure is sufficient for answering causal questions.
- Data summary: We know how to store all the needed parameters for a compound causal model as proposed above (i.e., we know how to story the various values of α and β, θ, and c_k).
- Update rules: For the (joint) distributions of X, Y, and (X, Y), the update rules are well-known. This is even true for much more complex, multivariate structures. The Bayesian statistics literature has matured to such a point that the proposed updating is a routine actively.
- Answer rules: In extreme cases, for example when the probability of a specific underlying causal structure is 1 or when we have data that is collected under one specific and known data-generating mechanism, we know how to compute the answers to specific questions and we even know many of the properties of these answers.

However, we are currently in the dark about the following:

- Assumptions: As far as I know, there is no work examining how assumptions regarding the data-generating mechanism can interact with or substitute assumptions regarding the causal structure. These branches of the literature are mostly disjoint. Especially when considering more complex systems of variables, does it suffice to understand the data-generating mechanism of some and the underlying causal structure of others in order to identify all causal effects?
- Data summary: My specification of $P(\mathcal{G})$ is admittedly a bit vague at this point and not truly Bayesian (in the sense that we derive a point estimate for each structure, as opposed to i.e., putting a prior on the probability of each structure). The exact way of (compactly) summarizing our knowledge regarding the plausibility of various causal structures given specific observations is, as far as I know, not known.
- Update rules: As is clear in the model description, we do not have good ways of updating $P(\mathcal{G})$. Let me call $P(G_k = c_k)$ the *causal structure likelihood*: developing the causal structure likelihood and a fully Bayesian approach to working with $P(\mathcal{G})$ is one of the prime targets of my upcoming research.

- Answer rules: While we know specific answer rules in specific cases (for example, we know how to compute the ATE when the data assignment mechanism is known and randomized), we are yet to develop answer rules for mixed data-generating mechanisms or uncertain causal structures (i.e., $\forall k$, $0 < P(G_k) < 1$).

7. Answering causal questions using compound causal models

It is clear from this first sketch of the Compound Causal Model that there is a lot of work to be done still: we are missing a number of pieces and we don't understand the properties of the answers provided by the – to-be-developed – answer rules.[33] Admittedly, at this moment in time, I am not even sure whether the Compound Causal Model will provide meaningful answers.[34]

Why then, do I go through all this trouble?

Well, the Compound Causal Model as described above provides, in my perception, the most feasible answer to causal questions in a very simple situation. Suppose we have data generated under the unknown, or Ø, setting. Thus, we assume we have data observed 'in the field'. Next, we suppose $G_1 : X \rightarrow Y$, $G_2 : Y \rightarrow X$, $G_3 : U \rightarrow (X, Y)$, and $P(G_k) = {}^1/_3$, $k = 1, 2, 3$. Thus, we believe the data to be generated under either scenario i, ii or iii as shown in Figure 8 and we deem every possible underlying structure equally probable. In such a case – which turns out to be surprisingly common in practice as we often do not know the data-generating mechanism and we are often highly uncertain about the true underlying causal structure – I would think it is feasible to state that the ATE is 0 with probability ${}^2/_3$ or $ATE = \mathbb{E}(Y \mid X = 1) - \mathbb{E}(Y \mid X = 0)$ with probability ${}^1/_3$.[35]

Given the model, and in a proper Bayesian fashion, we should be computing the posterior distribution over all plausible values of the ATE. In this case we would end up with a visualization like Figure 9: given our assumptions and our

[33] We can thus also not yet investigate the properties of the provided answers. We will need to start with very simple or extreme cases to develop answer rules and work using analytical or computer simulation methods to evaluate the provided answers.

[34] As discussed, surely, the answers will not be 'true'; the best we can hope for is that they are useful.

[35] Again, as in earlier cases, I am adopting population notation for simplicity. In actuality, as is also shown in the code examples in the Appendix, I would adopt a Bayesian approach and sample from the posterior ATE. Subsequently, I would approach the uncertainty over the underlying causal mechanisms by treating our data summary as a mixture model (see [Frühwirth-Schnatter, 2006, Richardson and Green, 1997, McLachlan and Peel, 2000] or, approached more from a model averaging perspective [Hoeting et al., 1999, Wasserman, 2000]) over the various options. Similarly, we could potentially include alternative data-generating mechanisms – i.e., the switches set to other values – by also approaching the final answer as a mixture of multiple posterior distributions. How to compute the (posterior) probability of each mixture component, both in the case of the underlying causal structure and in the case of the data-generating mechanism, is an open research question.

data, we have some belief that the ATE is close to, say, $^{8}/_{10}$.[36] However, given our uncertainty regarding the underlying causal structure(s), we also have a lot of reason to believe that the ATE is 0. In my view, given all the uncertainties involved and given the complexity of answering causal questions, it is desirable in our answers to make explicit that *we do not know for sure whether there is an effect for X to Y, but if there is, we have an idea of the plausible range of the size of the effect*. Hence, showing the uncertainty on all levels, in my subjective view, provides the most honest answer possible given the data that we have.[ix]

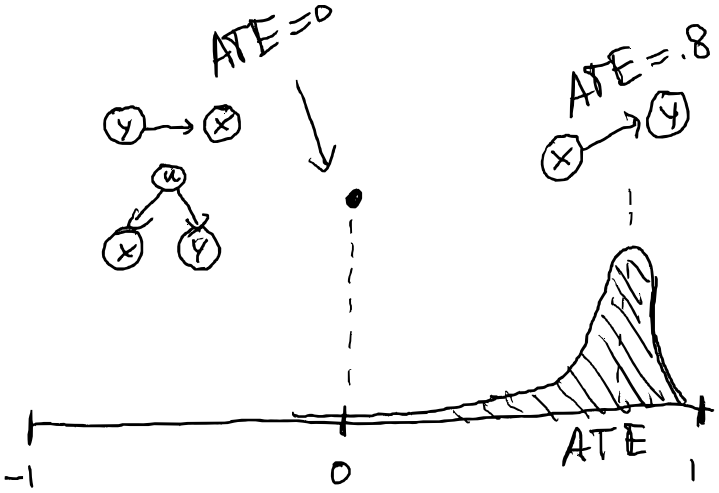


Figure 9. A visual summary of a possible answer provided by a Compound Causal Model (Model 4).

[36] This would be the case of the specific numerical example chosen in the code provided in the Appendix. Note that actually providing a meaningful point estimate of the ATE in this case is challenging: choosing the posterior expectation as the point estimate would provide an ATE for which there is virtually no empirical support, while choosing the posterior mode would lead to a point estimate of 0, which also seems to not properly reflect our current belief. For more on decision theoretic approaches to computing point estimates, see [Robert et al., 2007]).

8. Challenges and next steps

By now, I hope I have introduced some of the challenges in answering applied causal questions. My proposed Compound Causal Model aims to address some of these issues, though it still needs much refinement. However, in this lecture, I've drafted the initial implementation. I've also explained why I believe models that account for uncertainty in the data due to sampling or measurement uncertainty, uncertainty regarding the data-generating process, and uncertainty regarding the causal structure are practical and feasible. Such models offer the most accurate summary of our knowledge based on the data we have.[37]

Before closing, and besides the unknowns already mentioned such as the causal structure likelihood and the properties of various (novel) answer rules, I would like to highlight two more challenges that I foresee in further developing compound causal models. The first one is of a technical nature and concerns the (computational) complexity of Compound Causal Models. It is well-known that the number of possible DAGs, as a function of the number of variables n, grows super exponentially, Number of DAGs = $\sum_{k=1}^{n}(-1)^{k+1}\binom{n}{k}2^{k(n-k)}a(n-k)$.[q]

In compound causal models this will affect both the complexity of $P(G)$ and the number of possible data-generating mechanisms (as each variable could potentially be manipulated independently). It will be no simple feat to scale Compound Causal Models to deal with applied cases which often involve many variables.

A second challenge is more societal: the general audience seems to not like uncertain answers to seemingly simple questions. And because the general audience doesn't like such answers, scientists prefer to provide exact and definitive answers. We have jointly come to believe that the best answer to "How effective is chemotherapy in increasing five-year survival rate?" is "It increases the five-year survival rate by 20%." Neither the general public likes to hear nor do scientists like to answer: "At this moment we think it is very likely that the five-year survival rate increases by 20%, but there is also a substantial probability that the effect of

[37] I want to emphasize that any method ignoring one or more of these uncertainties tends to make stronger assumptions, which are likely incorrect. Methods that ignore uncertainty in the data-generating process, like simply calculating the ATE as the difference in means between control and treatment groups, risk misrepresenting the possible values of the ATE. Similarly, methods that fix one causal structure and compute a point estimate for the ATE may give an overly confident and inaccurate result.

[q] Where n is the number of variables (or nodes) in the DAG. To illustrate, we get the following for $n = 1, 2, \ldots, 5$: 1, 3, 25, 543, 29281.

chemotherapy is zero." This answer preference regrettably exists even when we have to admit that we are truly uncertain.
If anything, I hope that my future work on Compound Causal Models can play a role in addressing this 'need for unfounded certainty'. I hope my future work allows both scientists using Compound Causal Models, and the general audience consuming their answers to embrace uncertainty. Again, our scientific models are never 'true'; they are simply useful. Finally, I hope that by now I have clarified the title of this public lecture: Compound Causal Models help us to properly propagate – or 'include' – our uncertainty regarding the data-generating mechanism and the underlying causal structure into our computation and reporting of causal effects. Hence, I have argued that we should "propagate uncertainty in causal effect estimates."

References

D. H. Bailey, P. B. Borwein, and S. Plouffe. On the rapid computation of various polylogarithmic constants. *Mathematics of Computation*, 66(218):903–913, 1997. doi: 10.1090/S0025-5718-97-00810-2.

J. M. Borwein and P. B. Borwein. *Pi and the AGM: A Study in Analytic Number Theory and Computational Complexity*. Wiley, 1987. ISBN 9780471912084.

S. Broszeit and P. Hansen. *Causality and Time Series Analysis: Theories and Methods*. Springer, 2019.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–533, 2002.

S. Cunningham. *Causal inference: The mixtape*. Yale University Press, 2021.

V. Didelez and I. Pigeot. Causality: models, reasoning, and inference, 2001.

D. Eckles and M. Kaptein. Bootstrap thompson sampling and sequential decision problems in the behavioral sciences. *Sage Open*, 9(2):2158244019851675, 2019.

S. Frühwirth-Schnatter. *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer, New York, NY, 2006. ISBN 978-0387357683. doi: 10.1007/978-0-387-35768-3.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

R. Giere. *Science without laws*, volume 285. University of Chicago Press, 1999.

C. Gonzales and A.-H. Valizadeh. A full DAG score-based algorithm for learning causal bayesian networks with latent confounders. *arXiv preprint arXiv:2408.11181*, 2024.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

J. D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401, 1999. doi: 10.1214/ss/1009212519.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

L. Ippel, M. Kaptein, and J. Vermunt. Dealing with data streams. *Methodology*, 2016.

L. Ippel, M. Kaptein, and J. K. Vermunt. Estimating multilevel models on data streams. *psychometrika,* 84:41–64, 2019.

E. T. Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

M. Kalisch and P. Buhlmann. *Estimating Causal Effects Using Graphical Models*, volume 3. De Gruyter, 2007.

M. Kaptein. The sciences during the new common: A missed opportunity? *The New Common: How the COVID-19 Pandemic is Transforming Society*, pages 111–116, 2021.

M. Kaptein and E. Van den Heuvel. *Statistics for Data Scientists*. Springer, 2022.

X. Kavelaars, J. Mulder, and M. Kaptein. Decision-making with multiple correlated binary outcomes in clinical trials. *Statistical Methods in Medical Research*, 29(11):3265–3277, 2020.

X. Kavelaars, J. Mulder, and M. Kaptein. Bayesian multilevel multivariate logistic regression for superiority decision-making under observable treatment heterogeneity. *BMC Medical Research Methodology*, 23(1):220, 2023.

X. Kavelaars, J. Mulder, and M. Kaptein. Bayesian multivariate logistic regression for superiority and inferiority decision-making under observable treatment heterogeneity. *Multivariate Behavioral Research*, pages 1–24, 2024.

J. Kruijswijk, R. van Emden, P. Parvinen, and M. Kaptein. StreamingBandit: Experimenting with bandit policies. *Journal of Statistical Software*, 94(9), 2020.

H. Lütkepohl. New introduction to multiple time series analysis. 2005.

H. MA and R. JM. *Causal Inference: What If*. Boca Raton: Chapman and Hall/CRC, 2020.

G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley-Interscience, New York, NY, 2000. ISBN 978-0471006268. doi: 10.1002/0471721182.

C. Meek. Causal inference and causal explanation with probabilistic graphical models. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI 1995)*, pages 411–418, 1995.

F. Montagna, N. Noceti, L. Rosasco, K. Zhang, and F. Locatello. Scalable causal discovery with score matching. In *Conference on Causal Learning and Reasoning*, pages 752–771. PMLR, 2023.

J. Pearl. Learning causal structures from data. *Proceedings of the 8th Annual Conference on Uncertainty in Artificial Intelligence (UAI 1993)*, pages 268–276, 1993.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

J. Pearl and D. Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

J. Pearl, M. Glymour, and N. P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal discovery with continuous and discrete variables. *Journal of Machine Learning Research*, 17:1–52, 2016.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

D. Provodin, P. Gajane, M. Pechenizkiy, and M. Kaptein. The impact of batch learning in stochastic linear bandits. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 1149–1154. IEEE, 2022.

Z. Puha, M. Kaptein, and A. Lemmens. Batch mode active learning for individual treatment effect estimation. In *2020 International Conference on Data Mining Workshops (ICDMW)*, pages 859–866. IEEE, 2020.

S. Richardson and P. J. Green. Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B (Methodological)*, 59(4): 731–792, 1997. doi: 10.1111/1467-9868.00095.

T. J. Richardson and P. Spirtes. Efficient computation of adjustment sets in causal inference. *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence (UAI 2013)*, page 577–586, 2013.

C. P. Robert et al. *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer, 2007.

J. M. Robins and S. Greenland. The identification of causal effects using graphical models. *Journal of the American Statistical Association*, 89(425):637–658, 2001.

D. B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58, 1978.

D. B. Rubin. Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

L. Shen, E. Visser, F. van Erning, G. Geleijnse, and M. Kaptein. A two-step framework for validating causal effect estimates. *Pharmacoepidemiology and Drug Safety*, 33(9):e5873, 2024.

P. Spirtes and K. Zhang. The causal discovery algorithm: A formal review. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI 2000)*, pages 453–459, 2000.

P. Spirtes, C. Glymour, and R. Scheines. Causation, prediction, and search. *MIT Press*, 2000.

L. Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1):92–107, 2000. doi: 10.1006/jmps.1999.1278.

9. Code examples

All code examples are provided using the [R] language. The following examples are provided:

1. i. Generating $d^1(\mathcal{M}_1)$.
2. ii. Computing theta.
3. iii. Generating d^2 and comparing estimates of θ.
4. iv. Model $\mathcal{M}_3$: A Bayesian treatment of the Bernoulli observations.
5. v. Generation of Bivariate Bernoulli Data.
6. vi. Bayesian treatment of a Bivariate Bernoulli RV.
7. vii. Generating data from Machine 3.
8. viii. Computing the data summary for Model 5.
9. ix. Computing the ATE given Model 5 (example).

i. Generating $d^1(\mathcal{M}_1)$. Generation of the $d^1(\mathcal{M}_1)$. The random draws correspond – after some searching for a correct seed – to the parity of the first ten digits of Pi.

```
set.seed(100225254)
d1 <- rbinom(10, 1, .5)
# [1] 0 0 1 0 0 0 1 1 0 0
```

For reference, the first ten digits of Pi:

```
sprintf("%.10f", pi)
# [1] "3.1415926536"
```

ii. Computing theta. Computing $\hat{\theta}$ and other quantities mentioned in the description of Model $\mathcal{M}_2$.

```
# Given d1, compute theta_hat:
theta_hat <- sum(d1) / length(d1)

# Given additional observations, update theta_hat
d_add <- rbinom(5, 1, .5)
print(d_add)
# [1] 1 0 0 0 1

# Update rule
theta_hat_updated <- (sum(d1) + sum(d_add)) / (length(d1) + length(d_add))

# Compute answer to "observing 1s in ar row based on theta_hat_updated
n_seq <- 10
theta_hat_updated^n_seq
# [1] 1.693509e-05
```

Additionally, some code to provide a Figure akin to Figure 2.

```
# Create a simple number line from -10 to 10
plot(0:1, rep(0, 2), type = "n", xlab = "P(X = 1)", ylab = "", axes = FALSE
    )
axis(1, at = 0:1)  # Draw the number line

# Plot the number with a point and label
points(theta_hat_updated, -1, col = "red", pch = 25)
text(theta_hat_updated, -.8, labels = theta_hat_updated, col = "red")
```

iii. Generating d_2 and comparing estimates of θ. Demonstration of the exact same maximum likelihood estimate for θ for two different datasets:

```
# Generating d2 by simply replicating d1:
d2 <- rep(d1, 5)
print(d2)

# Compute theta based on d1 and d2:
theta_hat_d1 <- sum(d1) / length(d1)
theta_hat_d2 <- sum(d2) / length(d2)
theta_hat_d1 == theta_hat_d2
# [1] TRUE
```

iv. Model $\mathcal{M}_3$: A Bayesian treatment of the Bernoulli observations. Bayesian treatment of d^1 and d^2 including resulting figure(s) (Figure 3).

```r
# Improper prior:
alpha_0 = 0
beta_0 = 0

# d1 posterior:
alpha_n_d1 <- alpha_0 + sum(d1)
beta_n_d1 <- beta_0 + (length(d1) - sum(d1))

# d2 posterior:
alpha_n_d2 <- alpha_0 + sum(d2)
beta_n_d2 <- beta_0 + (length(d2) - sum(d2))

# Setup to plot 2 figures:
par(mfrow = c(1, 2))

# Plot d1 posterior
theta <- seq(0, 1, length.out = 500)
posterior_density <- dbeta(theta, alpha_n_d1, beta_n_d1)

plot(theta, posterior_density, type = "l", lwd = 2,
     xlab = expression(theta),
     ylab = "Density",
     main = "Posterior Distribution of Theta",
     col = "blue")

# Add MLE estimate
theta_map <- sum(d1) / length(d1)
abline(v = theta_map, col = "red", lty = 2)


# Plot d2 posterior
theta <- seq(0, 1, length.out = 500)
posterior_density <- dbeta(theta, alpha_n_d2, beta_n_d2)

plot(theta, posterior_density, type = "l", lwd = 2,
     xlab = expression(theta),
     ylab = "Density",
     main = "Posterior Distribution of Theta",
     col = "blue")

# Add MLE estimate
theta_map <- sum(d2) / length(d2)
abline(v = theta_map, col = "red", lty = 2)

# Reset the plotting layout to default
par(mfrow = c(1, 1))
```

v. Generation of Bivariate Bernoulli Data. Generation of bivariate Bernoulli observations.

```r
# Function to generate Bivariate Bernoulli Observations.
# p_X and p_Y are the marginal success probabiltiites, rho is the
    dependency.

generate_bivariate_bernoulli <- function(n, p_X, p_Y, rho) {
  if (p_X < 0 | p_X > 1 | p_Y < 0 | p_Y > 1 | rho < -1 | rho > 1) {
    stop("Probabilities must be between 0 and 1, and correlation between -1
        and 1.")
  }

  cov_term <- rho * sqrt(p_X * (1 - p_X) * p_Y * (1 - p_Y))
  p11 <- p_X * p_Y + cov_term
  p10 <- p_X * (1 - p_Y) - cov_term
  p01 <- (1 - p_X) * p_Y - cov_term
  p00 <- (1 - p_X) * (1 - p_Y) + cov_term

  if (abs(p00 + p01 + p10 + p11 - 1) > 1e-6) {
    stop("Calculated probabilities must sum to 1.")
  }

  outcomes <- sample(1:4, size = n, replace = TRUE, prob = c(p00, p01, p10,
      p11))
  data <- data.frame(
    X = as.integer(outcomes %in% c(3, 4)),
    Y = as.integer(outcomes %in% c(2, 4))
  )
  return(data)
}

# Next, I use this function to generate d3(M2)
n <- 10
p_X <- 0.7    # Marginal probability P(X = 1)
p_Y <- 0.4    # Marginal probability P(Y = 1)
rho <- 0.3    # Correlation parameter

# Generate data
d3 <- generate_bivariate_bernoulli(n, p_X, p_Y, rho)
print(t(d3))
```

.

vi. Bayesian treatment of a Bivariate Bernoulli RV. Code to compute the posterior as detailed in Model 4. The code includes the creation of a contour plot.

```
# First, we count the occurances of each tuple in d3 (generated in previous snippet)
n00 <- sum(d3$X == 0 & d3$Y == 0)
n01 <- sum(d3$X == 0 & d3$Y == 1)
n10 <- sum(d3$X == 1 & d3$Y == 0)
n11 <- sum(d3$X == 1 & d3$Y == 1)

# Prior parameters for the Dirichlet distribution
# This time selecting a relatively uninformative prior:
alpha_prior <- c(1, 1, 1, 1)

# Compute posterior parameters (effectively this summarizes the model)
alpha_posterior <- alpha_prior + c(n00, n01, n10, n11)

# For plotting: sample from the posterior and create a contour plot:
library(ggplot2)

# Function to sample from Dirichlet distribution
rdirichlet <- function(n, alpha) {
  samples <- matrix(0, nrow = n, ncol = length(alpha))
  for (i in 1:n) {
    samples[i, ] <- rgamma(length(alpha), shape = alpha, rate = 1)
    samples[i, ] <- samples[i, ] / sum(samples[i, ])
  }
  return(samples)
}

# Sample from the posterior distribution
posterior_samples <- rdirichlet(10000, alpha_posterior)
colnames(posterior_samples) <- c("p00", "p01", "p10", "p11")
head(posterior_samples)

# Compute the marginal probabilities P(X) and P(Y) given the samples
posterior_samples <- as.data.frame(posterior_samples)
posterior_samples$pX <- posterior_samples$p10 + posterior_samples$p11    # Computing the marginals
posterior_samples$pY <- posterior_samples$p01 + posterior_samples$p11

contour_plot <- ggplot(posterior_samples, aes(x = pX, y = pY)) +
  stat_density_2d(aes(), geom = "contour", color = "black", bins = 5) +
  theme_minimal() +
  labs(title = "Posterior Distribution of P(X) and P(Y)",
       x = "P(X)",
       y = "P(Y)") +
  scale_x_continuous(limits = c(0, 1)) + # Set x-axis limits between 0 and 1
  scale_y_continuous(limits = c(0, 1))   # Set y-axis limits between 0 and 1

# Display the plot
print(contour_plot)
```

vii. Generating data from Machine 3. I will provide several functions to generate data from Machine 3 with potentially different internal wirings. The code is heavily annotated and should speak for itself.

```
# Causal data generation akin to Model 3.
# I will detail four version of the machine:
#      X -> Y   (Note: the only one used)
#      Y -> X
#      U -> (X,Y)
#      X _|_ Y

# I start with the machine on the left-hand side of Fig 7.
# X -> Y
gen_M3_XcauseY <- function(n, x_switch, y_switch, pX = .5, pYX0 = .1, pYX1
    = .9){

  # n is the number of generated samples under this scheme
  # x_switch is the position of the switch on X (0, 1, v (for void))
  # y_switch is the position of the switch on Y
  # pX is the marginal probablity of X
  # pYX0 is the (conditional) probability of Y given X=0
  # pYX1 is the (conditional) probability of Y given X=1

  # X ~ Ux
  # Y ~ f(X, Uy)

  # Check switches
  if(! x_switch %in% c(0, 1, "v")){stop("No valid value for x_switch");}
  if(! y_switch %in% c(0, 1, "v")){stop("No valid value for y_switch");}

  # Sample X
  if(x_switch == 0){ X <- rep(0, n)}
  if(x_switch == 1){ X <- rep(1, n)}
  if(x_switch == "v"){ X <- rbinom(n, 1, pX)}

  # Sample Y
  if(y_switch == 0){ Y <- rep(0, n)}
  if(y_switch == 1){ Y <- rep(1, n)}
  if(y_switch == "v"){
    Y0 = rbinom(n, 1, pYX0)
    Y1 = rbinom(n, 1, pYX1)
    Y <- ifelse(X == 0, Y0, Y1)
  }

  data <- data.frame(        # adding switch to the data
    x_switch = x_switch,
    y_switch = y_switch,
    X = as.integer(X),
    Y = as.integer(Y)
  )
```

```r
  return(data)
}

# Example usage of the first machine:
gen_M3_XcauseY(20, "v", 1)



# Y -> X
gen_M3_YcauseX <- function(n, y_switch, x_switch, pY = .5, pXY0 = .1, pXY1
    = .9){

  # n is the number of generated samples under this scheme
  # y_switch is the position of the switch on Y (0, 1, v (for void))
  # x_switch is the position of the switch on X
  # pY is the marginal probablity of Y
  # pXY0 is the (conditional) probability of Y given X=0
  # pXY1 is the (conditional) probability of Y given X=1

  # Y ~ Uy
  # X ~ f(Y, Ux)

  # Check switches
  if(! y_switch %in% c(0, 1, "v")){stop("No valid value for y_switch");}
  if(! x_switch %in% c(0, 1, "v")){stop("No valid value for x_switch");}

  # Sample Y
  if(y_switch == 0){ Y <- rep(0, n)}
  if(y_switch == 1){ Y <- rep(1, n)}
  if(y_switch == "v"){ Y <- rbinom(n, 1, pY)}

  # Sample Y
  if(x_switch == 0){ Y = rep(0, n)}
  if(x_switch == 1){ Y = rep(1, n)}
  if(x_switch == "v"){
    X0 = rbinom(n, 1, pXY0)
    X1 = rbinom(n, 1, pXY1)
    X <- ifelse(Y == 0, X0, X1)
  }

  # Keep the order the same in every machine:
  data <- data.frame(        # adding switch to the data
    x_switch = x_switch,
    y_switch = y_switch,
    X = as.integer(X),
    Y = as.integer(Y)
  )
  return(data)
}

```

```
95
96  # U -> (X, Y)
97  gen_M3_CommonCause <- function(n, x_switch, y_switch, pXU0 = .2, pXU1 = .8,
        pYU0 = .4, pYU1 = .9){
98
99    # n is the number of generated samples under this scheme
100   # y_switch is the position of the switch on Y (0, 1, v (for void))
101   # x_switch is the position of the switch on X
102   # pXU0 is the probability for X=1 given U=0 (idem for the others)
103
104   # Y ~ f(U, Uy)
105   # X ~ g(U, Ux)
106
107   # Check switches
108   if(! x_switch %in% c(0, 1, "v")){stop("No valid value for x_switch");}
109   if(! y_switch %in% c(0, 1, "v")){stop("No valid value for y_switch");}
110
111   # Sample U
112   U = rbinom(n, 1, .5)
113
114   # Sample Y
115   if(x_switch == 0){ X <- rep(0, n)}
116   if(x_switch == 1){ X <- rep(1, n)}
117   if(x_switch == "v"){
118     X0 = rbinom(n, 1, pXU0)
119     X1 = rbinom(n, 1, pXU1)
120     X <- ifelse(U == 0, X0, X1)
121   }
122
123   # Sample Y
124   if(y_switch == 0){ Y <- rep(0, n)}
125   if(y_switch == 1){ Y <- rep(1, n)}
126   if(y_switch == "v"){
127     Y0 = rbinom(n, 1, pYU0)
128     Y1 = rbinom(n, 1, pYU1)
129     Y <- ifelse(U == 0, Y0, Y1)
130   }
131
132   # Keep the order the same in every machine:
133   data <- data.frame(        # adding switch to the data
134     x_switch = x_switch,
135     y_switch = y_switch,
136     X = as.integer(X),
137     Y = as.integer(Y)
138   )
139   return(data)
140 }
141
142 # X _|_ Y
143 gen_M3_Ind <- function(n, x_switch, y_switch, pX = .4, pY = .8){
```

```
144
145   # n is the number of generated samples under this scheme
146   # x_switch is the position of the switch on X (0, 1, v (for void))
147   # y_switch is the position of the switch on Y
148   # pXU0 is the probability for X=1 given U=0 (idem for the others)
149
150   # X ~ Ux)
151   # Y ~ Uy)
152
153   # Check switches
154   if(! x_switch %in% c(0, 1, "v")){stop("No valid value for x_switch");}
155   if(! y_switch %in% c(0, 1, "v")){stop("No valid value for y_switch");}
156
157   # Sample U
158   U = rbinom(n, 1, .5)
159
160   # Sample Y
161   if(x_switch == 0){ X <- rep(0, n)}
162   if(x_switch == 1){ X <- rep(1, n)}
163   if(x_switch == "v"){
164     X <- rbinom(n, 1, pX)
165   }
166
167   # Sample Y
168   if(y_switch == 0){ Y <- rep(0, n)}
169   if(y_switch == 1){ Y <- rep(1, n)}
170   if(y_switch == "v"){
171     Y <- rbinom(n, 1, pY)
172   }
173
174   # Keep the order the same in every machine:
175   data <- data.frame(       # adding switch to the data
176     x_switch = x_switch,
177     y_switch = y_switch,
178     X = as.integer(X),
179     Y = as.integer(Y)
180   )
181   return(data)
182 }
```

The four machines can easily be used to generate data under different settings of the switches of the machine and different internal wirings of the machine(s).

viii. Computing the data summary for Model 5. The code below provides an R S4 (object-oriented) implementation of Model 5. I provide the initialization, data summary, update rules, and one answer rule (for computing the posterior ATE, see below).

```
1
2  # An R S4 implementation of Model 5.
3
4  # Define an S4 class including data summary.
5  # Note that assumptions are not encoded in the object.
6  setClass(
7    "CCM",                    # Compound causal model class
8    slots = list(             # List containing summary data
9      XY = "numeric",         # a00, a01, a10, a11 (joint posterior)
10     XfY0 = "numeric",        # a,b (beta posterior given Y fixed to 0)
11     XfY1 = "numeric",        # a,b (...)
12     YfX0 = "numeric",        # a,b
13     YfX1 = "numeric",        # a,b
14     G = "numeric")           # Probabilities of Graphs X->Y, Y->X, U->(X,Y)
15 )
16
17 # Create an S4 object with "uninformative" priors (start state of model
       data)
18 Mod5 <- new("CCM",
19             XY   = c(1,1,1,1),
20             YfX0 = c(1,1),
21             YfX1 = c(1,1),
22             XfY0 = c(1,1),
23             XfY1 = c(1,1),
24             G    = c(1/3,1/3,1/3))
25
26 # Method to *update* data summary based on new data
27 # I.e., example of rules to update the data summary.
28 setGeneric("dataSummary", function(object, data, ...) {
29   standardGeneric("dataSummary")
30 })
31 setMethod(
32   "dataSummary",
33   "CCM",
34   function(object, data, ...) {
35
36     # Update of the respective posteriors:
37     # an ugly way to deal with mixed data, simply run through the data row
           by row
38     for (i in 1:nrow(data)) {
39
40       # Simply list all cases
41       if(data$x_switch[i] == 0 && data$y_switch[i] == 0){ } # ignore
42       if(data$x_switch[i] == 0 && data$y_switch[i] == 1){ } # ignore
43       if(data$x_switch[i] == 1 && data$y_switch[i] == 0){ } # ignore
44       if(data$x_switch[i] == 1 && data$y_switch[i] == 1){ } # ignore
```

```
45
46        # manipulate X
47        if(data$x_switch[i] == 0 && data$y_switch[i] == "v"){
48          ifelse(data$Y[i] == 1, object@YfX0[1] <- object@YfX0[1] +1 ,
                object@YfX0[2] <- object@YfX0[2] + 1)
49        }
50        if(data$x_switch[i] == 1 && data$y_switch[i] == "v"){
51          ifelse(data$Y[i] == 1, object@YfX1[1] <- object@YfX1[1] +1 ,
                object@YfX1[2] <- object@YfX1[2] + 1)
52        }
53
54        # manipulate Y
55        if(data$y_switch[i] == 0 && data$x_switch[i] == "v"){
56          ifelse(data$X[i] == 1, object@XfY0[1] <- object@XfY0[1] +1 ,
                object@XfY0[2] <- object@XfY0[2] + 1)
57        }
58        if(data$y_switch[i] == 1 && data$x_switch[i] == "v"){
59          ifelse(data$X[i] == 1, object@XfY1[1] <- object@XfY1[1] +1 ,
                object@XfY1[2] <- object@XfY01[2] + 1)
60        }
61
62        # Free
63        if(data$x_switch[i] == "v" && data$y_switch[i] == "v"){
64          n00 <- ifelse(data$X[i] == 0 && data$Y[i] == 0, 1, 0)
65          n01 <- ifelse(data$X[i] == 0 && data$Y[i] == 1, 1, 0)
66          n10 <- ifelse(data$X[i] == 1 && data$Y[i] == 0, 1, 0)
67          n11 <- ifelse(data$X[i] == 1 && data$Y[i] == 1, 1, 0)
68
69          # Compute posterior parameters (effecitvely this summarizes the
                model)
70          object@XY  <- object@XY + c(n00, n01, n10, n11)
71        }
72
73      }
74
75      # Update of the graph structure G
76      # Simply overwrite the probabilities if this data is passed (for now)
77      call <- match.call()
78      if( "graph" %in% names(call)){
79        object@G <- eval(call$graph)
80      }
81
82      return(object)
83
84    }
85  )
86
87
88  # Method to *compute the ATE* given the current state of the data
89  # I.e., an example of an answer rule.
```

```
# Note that the function assumes the ATE concerns X -> Y,
#  and is only partially implemented.
setGeneric("answerATE", function(object, prior = c(1,1), samples = 1000) {
  standardGeneric("answerATE")
})
setMethod(
  "answerATE",
  "CCM",
  function(object, prior = c(1,1), samples = 1000) {
    cat("Note: This function is only partially implemented. \n")

    # Check if unimplemented cases contain data and alert
    if(!identical(object@XfY0, prior)){ print("XfY0 is not included. \n")}
    if(!identical(object@XfY1, prior)){ print("XfY1 is not included. \n")}
    if(!identical(object@YfX0, prior)){ print("YfX0 is not included. \n")}
    if(!identical(object@YfX1, prior)){ print("YfX1 is not included. \n")}

    # Recall: object@G = Probabilities of Graphs X->Y, Y->X, U->(X,Y)
    # Sample from posterior as a mixture step by step

    # First compute the number of samples from each DGP (X->Y, Y->X, U->(X,
        Y))
    g_samples <- sample(c(1,2,3), size = samples, replace = TRUE, prob =
        object@G)

    # Note that the Dirichlet posterior in @XY presents prior + counts for
      # Count of (X = 0, Y = 0)
      # Count of (X = 0, Y = 1)
      # Count of (X = 1, Y = 0)
      # Count of (X = 1, Y = 1)
    # Also, we will use rdirichlet defined on line 190.

    # Generate samples from the posterior XY
    posterior_samples <- rdirichlet(samples, object@XY)

    # Extract components of the posterior samples
    p00 <- posterior_samples[, 1]
    p01 <- posterior_samples[, 2]
    p10 <- posterior_samples[, 3]
    p11 <- posterior_samples[, 4]

    # Compute marginal p(Y)
    p_y0 <- p00 + p10
    p_y1 <- p01 + p11

    # Compute conditionals p(Y | X = 0) and p(Y | X = 1)
    p_y_given_x0 <- cbind(p00 / (p00 + p01), p01 / (p00 + p01))
    p_y_given_x1 <- cbind(p10 / (p10 + p11), p11 / (p10 + p11))

    return_samples <- c()    # Empty vector
```

```
138
139     # Sample ATE under X->Y, i.e., sample under X=0, X=1, and select
140     # Note that the [,2 selects for Y=1]
141     ATE_XY <- p_y_given_x1[,2] - p_y_given_x0[,2]
142     return_samples <- c(return_samples, ATE_XY[g_samples == 1])
143
144     # Sample ATE under Y->X, i.e., just sample marginal of Y
145     # These are just an ATE of 0
146     ATE_YX <- p_y1 - p_y1
147     return_samples <- c(return_samples, ATE_YX[g_samples == 2])
148
149     # Sample ATE under U->(X,Y), i.e., just sample marginal of Y
150     # These are just an ATE of 0
151     ATE_U <- p_y1 - p_y1
152     return_samples <- c(return_samples, ATE_U[g_samples == 3])
153
154     return(return_samples)
155   }
156 )
157
158
159 # Usage example
160 # Generate some data:
161 test_data <- gen_M3_CommonCause(10, "v", "v")
162
163 # Update the model (initialized on line 433)
164 Mod5 <- dataSummary(Mod5, test_data)
165
166 # Inspect the model
167 Mod5
```

An example of how to use the answerATE method to sample from the posterior and visualize the results is provided in the next code snippet. Please note that the implementation of the answerATE method is limited and only covers the case discussed in the main text.

ix. Computing the ATE given Model 5 (example). Example of data generation and usage of Model 5 to compute (posterior samples of) the ATE and generate a plot akin to Figure 9.

```
1
2  # Initialize a new model,
3  # create 100 observations under "v" and X->,
4  # and finally update the model
5  ModPlot <- new("CCM",
6                 XY    = c(1,1,1,1),
7                 YfX0  = c(1,1),
8                 YfX1  = c(1,1),
9                 XfY0  = c(1,1),
10                XfY1  = c(1,1),
11                G     = c(1/3,1/3,1/3))
12
13 obs <- gen_M3_XcauseY(100, "v", "v")
14
15 ModPlot <- dataSummary(ModPlot, obs)
16
17 # Generate posterior samples of the ATE
18 ATE_samples <- answerATE(ModPlot, samples = 1000)
19
20 # Fit density
21 density_estimate <- density(ATE_samples)
22
23 # Plot the density
24 plot(density_estimate,
25      main = "Density Plot of Samples",
26      xlab = "Sample Values",
27      ylab = "Density",
28      col = "blue",
29      lwd = 2)
30
31
32 # The density plot does not very clearly depict the point mass at 0
33 # The following plot might be cleaner:
34
35 # Create histogram
36 hist(ATE_samples, breaks = 30, freq = FALSE, col = rgb(0, 0, 1, 0.5),
37      main = "Density of Mixture of Continuous RV and Point Mass at 0",
38      xlab = "Value", xlim = c(-1, 1), ylim = c(0, 15))
39
40 # Add a kernel density estimate
41 lines(density(ATE_samples), col = "red", lwd = 2)
42
43 # Add a vertical line at 0 to show the point mass
44 abline(v = 0, col = "green", lwd = 2, lty = 2)
45
46 # Optionally, you can add a rug at the bottom to indicate individual points
47 rug(ATE_samples, col = "black")
```

Curriculum Vitae

Prof.dr. Maurits Clemens Kaptein, was appointed part-time Professor of Applied Causality at the Department of Mathematics and Computer Science, Eindhoven University of Technology (TU/e) on October 1, 2023.

Prof.dr. Maurits Kaptein is a professor of Applied Causality at Eindhoven University of Technology and serves as Chief Data Scientist at Network Optix Inc. After obtaining his PhD cum laude in Industrial Design from TU Eindhoven, Maurits held academic positions at Stanford University (Distinguished Visiting Scholar, Communication), Aalto School of Economics (Post-doc, Marketing), Radboud University (Assistant Professor, Artificial Intelligence), and Tilburg University (Assistant, Associate, and Full Professor, Statistics). He has authored or co-authored four books and over 100 academic papers, many published in high-impact journals or conferences across fields such as communications, marketing, statistics, healthcare, and computer science. Additionally, Maurits co-founded Scailable BV, a Dutch edge-AI deeptech startup acquired by Network Optix Inc. in early 2024. His recent research focuses on developing novel statistical methods for sequential learning and causal inference.

Colophon

Production
Communication Expertise Center TU/e

Cover photography
Bart van Overbeeke Photography, Eindhoven

Design
Grefo Prepress, Eindhoven

Digital version:
research.tue.nl/en/

Visiting address
Building 1, Auditorium
Groene Loper, Eindhoven
The Netherlands

Navigation
De Zaale, Eindhoven

Postal address
PO Box 513
5600 MB Eindhoven
The Netherlands
Tel. +31 (0)40 247 9111
www.tue.nl/map

TU/e
EINDHOVEN
UNIVERSITY OF
TECHNOLOGY