Topics in Causal Inference and the Law

A dissertation presented

by

Thomas Howland Ferriss

to

The Department of Statistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Statistics

Harvard University

Cambridge, Massachusetts

April 2018
Abstract

Randomized experiments are a fundamental tool for estimating the causal effects of proposed interventions. While analysis of some experiments can be quite straightforward, other experiments may present difficult analytical choices. For a variety of reasons, causal inference problems may call for complex modeling of the potential outcomes by, for example, posing a likelihood function and a prior distribution on a parameter vector. Additionally, analytical challenges might arise when the randomized treatment assignments are not obeyed (i.e. when there is “noncompliance”), particularly when there is substantial missing data about whether the units complied with their assignments. After a brief introduction in Chapter 1, Chapter 2 investigates the consequences of model misspecification in model-based Bayesian causal inference settings via a full factorial simulation experiment. We use Bayesian evaluation criteria for evaluating the severity of the misspecification. We find that posterior inferences are particularly sensitive to a failure to adequately transform the outcome variable, especially in conjunction with a poor choice of prior distribution. Chapter 3 develops a methodology for analyzing experiments with two-way noncompliance when there is substantial missing data about whether units took the treatment to which they were assigned. We validate this methodology with simulation studies and apply it to a randomized evaluation of Philadelphia divorce court. The results of this analysis, we argue, suggest that Philadelphia divorce court procedures are unconstitutional under the Due Process Clause of the Fourteenth Amendment.
Contents

Contents iv

Acknowledgements vii

1 Introduction 1

1.1 The Rubin Causal Model ................................................. 3

2 Validity of Bayesian Causal Inference under Model Misspecification: A Simulation Experiment 5

2.1 Summary ................................................................. 5

2.2 Introduction ............................................................. 6

2.3 Previous Work ........................................................... 9

2.4 Background: Bayesian Model-Based Approaches to Causal Inference .... 12

2.5 Methods ................................................................. 14

2.6 Results and Discussion .................................................. 40

2.7 The Importance of Checking Model Assumptions in the Age of Big Data ... 68

2.8 Conclusion ............................................................... 70

3 Estimating Complier Average Causal Effects in the Presence of Missing Compliance Data: A Randomized Evaluation of Divorce Court Accessibility 73

3.1 Introduction ............................................................... 73
3.2 The Rubin Causal Model and Previous Work ........................................ 77
3.3 The Philadelphia Divorce Study ............................................................. 78
3.4 Notation, Assumptions and Data ............................................................ 83
3.5 Intention-to-Treat Analysis of the Philadelphia Divorce Study ................. 88
3.6 Analysis Methods ....................................................................................... 91
3.7 Model for The Philadelphia Divorce Study .............................................. 101
3.8 Simulation Studies .................................................................................... 115
3.9 Results for The Philadelphia Divorce Study ............................................ 136
3.10 Remarks on Model Misspecification ....................................................... 140
3.11 Future Extensions .................................................................................... 141
3.12 Conclusion ............................................................................................... 142

4 Conclusion .................................................................................................... 144

Appendices ....................................................................................................... 146

A Simulation Study: ANOVA Tables for All Outcomes .................................. 147
  A.1 ANOVA for the PC(Yobs, X, W) ................................................................. 147
  A.2 ANOVA for the KL Divergence ................................................................. 149
  A.3 ANOVA for the Normalized Difference in Posterior Means .................. 150
  A.4 ANOVA for the Ratio of Posterior Standard Deviations ....................... 152
  A.5 ANOVA for the Ratio of Credible Interval Widths ................................. 154
  A.6 ANOVA for the KS Statistic .................................................................. 156
  A.7 ANOVA for False Positives .................................................................. 157
  A.8 ANOVA for False Negatives .................................................................. 159
  A.9 ANOVA for Skew Ratio ......................................................................... 159

B Divorce Study: Sequential Binomial Modeling for the Compliance Type 161
  B.1 Likelihood ............................................................................................... 161
Acknowledgements

I owe thanks to many people who have supported me along the way. A handful were so tremendously influential, generous and supportive that I should thank them again here:

- Professor Donald B. Rubin, for his brilliance, generosity, vision and his advocacy. His ideas have profoundly changed my life for the better and I am deeply grateful.

- Professor D. James Greiner, for his mentorship, for encouraging me to pursue a doctorate in Statistics, for teaching me much about the law, for countless hours spent advising me on my scholarship and career, and much more.

- Professor Mary Beth Landrum for her guidance, collaboration and generosity.

- My mother, father and brother, for all of their support, love, advice, and faith.
Chapter 1

Introduction

Many scientific research questions are fundamentally causal in nature, rather than predictive or descriptive. Randomized experiments are the gold standard for estimating the causal effects of proposed interventions. Causal inference is thus a core area of statistics and science, particularly the design and analysis of experiments.

However, the analysis of randomized experiments can often yield difficult analytical questions. For example, the analysis methods may, for a wide range of reasons, call for posing a complex model. For example, one must often pose a likelihood function that is appropriate to the collected data as well as a prior distribution on the parameter vector. Additionally, randomized experiments might encounter problems that challenge or even break the usual assumptions that we wish held for our experiment. For example, an experiment might experience “noncompliance,” where units that are assigned to one treatment end up receiving a different treatment, notwithstanding the randomized assignment. Such problems seemingly raise doubts about the benefits of randomization, when the randomization itself is not adhered to.

This dissertation presents two extensive studies that explore such issues. The first study investigates the consequences of model misspecification for model-based Bayesian approaches to the causal analysis of randomized experiments. The second study proposes a new method-
ology for analyzing randomized experiments with two-way noncompliance that additionally are missing data about which treatment the units actually received. The first study shows that misspecification of Bayesian models in the causal inference setting can carry important risks. The second study shows that model-based Bayesian models are sometimes necessary, insofar as they can solve problems that appear intractable under other approaches.

The first study is a full factorial simulation experiment where each factor in the experiment is a different form of Bayesian model misspecification. We compare the resultant misspecified posterior distribution to the true posterior distribution using Bayesian evaluation criteria. The results suggest that causal analyses of randomized experiments can be particularly sensitive to failures to adequately transform the outcome variable. Additionally, when the outcome is not properly transformed, the inferences can also be quite sensitive to the choice of prior. In some scenarios, therefore, it may be best to avoid assumptions about the form of the likelihood or the prior if other approaches are available (for example, if Fisher’s exact test will suffice).

The second study develops a methodology for analyzing experiments where such simple approaches simply do not appear to apply. Indeed, the results suggest that the flexibility and power of the Bayesian model-based approach dominate competing approaches to the problem. The study concerns an experiment where potential divorce court litigants were randomized to either receive or not receive a free attorney. The goal is to estimate the causal effect of having an attorney on successfully filing for, and obtaining, a divorce in Philadelphia. Using the proposed methodology, we estimate that the effects of having an attorney are very large. These results call into question the constitutionality of the current implementation of divorce court in Philadelphia (in particular, because litigants do not have a right to an attorney and because the right to divorce is a constitutionally protected right). We remark on the applicable legal doctrine and briefly present the constitutional implications of the study results.

To handle the two-way non-compliance and missing data in the divorce study, we propose
a form of model-based Bayesian analysis that augments the data space with latent, partially observed compliance statuses. The posterior predictive distribution for the missing outcome data can be recovered via an application of the Data Augmentation algorithm by using the complete data likelihood. We end with some remarks for why the model misspecification problems that arise in the first study likely do not apply to the second study. For example, since the outcomes are binary in the divorce study, we argue that the problem of appropriately transforming the outcome does not pose the same risks as cases where the outcome is continuous.

All of the work below draws heavily on the Rubin Causal Model framework, which we briefly summarize in the next section.

1.1 The Rubin Causal Model

In the simplest randomized experiments, experimental units are probabilistically assigned to receive either an active treatment or a control treatment. The estmand is the causal effect of the active treatment on some outcome of interest, which may be denoted $Y^{obs}$.

Let $W_i$ be the indicator for whether the $i^{th}$ unit was randomly assigned to receive treatment ($W_i = 1$) or control ($W_i = 0$). Under the SUTVA assumption, each experimental unit, $i$, has two potential outcomes ($Y_i(1), Y_i(0)$), which correspond to the outcomes that the unit would experience if it received treatment or control, respectively. Only one of these potential outcomes can be observed for each unit. Correspondingly, at least one potential outcome will always be missing. Due to this missingness, we can never directly observe the unit-level causal effect of the treatment for any unit. The project of imputing these missing potential outcomes is a central challenge for estimating causal effects.

The explicit use of potential outcomes in the above framework, often referred to as the Rubin Causal Model (see e.g. Holland (1986) and only quickly summarized here), clarifies that causal inference is fundamentally a missing data problem (Rubin, 1978). The concept of
potential outcomes was originally introduced by Neyman (1923) for randomized experiments and was extended by Rubin (1974, 1977, 1978) to apply to a broad array of causal inference questions, including analysis of both randomized experiments and observational studies.

There are approaches to imputing missing potential outcomes that require few unit-level modeling assumptions, such as Fisher’s exact test and Neyman’s estimator for the super-population causal estimand. There are also approaches which allow for greater flexibility at the expense of more assumptions, allowing for complex modeling of the relationships between the outcomes, covariates and parameters. Such approaches include likelihood-based methods, such as the Bayesian methods that we will consider in Chapter 2.
Chapter 2

Validity of Bayesian Causal Inference under Model Misspecification: A Simulation Experiment

2.1 Summary

We present the results of a full factorial simulation experiment in order to explore the consequences of model misspecification on causal inferences that are obtained with model-based Bayesian methods. Each factor in the experiment corresponds to either a form of likelihood misspecification, choice of prior, or experimental design choice. We evaluate the results of each combination of factors by comparing the results of the misspecified model with the results that would be obtained from a correctly specified model. The evaluation criteria include $PC(Y_{obs}, X, W)$, which we define as the coverage under the true posterior of the 95% credible interval that is constructed with the misspecified model as well as the KL divergence from the misspecified posterior distribution to the true posterior distribution. Other evaluation criteria are also considered.

The results of the experiment suggest that, of the forms of misspecification considered,
Bayesian model-based causal inferences are most sensitive to failure to adequately transform the outcome variable as well as the choice of prior distribution. Models that fail to properly transform the outcome, even when using the true prior, can lead to posterior credible intervals that are very poorly centered. The use of flat priors appears to mitigate this problem in the examples considered.

Furthermore, a failure to properly transform the outcome variable can produce posterior credible intervals that are much too narrow (or, sometimes, much too wide). This problem occurs for all of the prior distributions that are considered. These problems remain even when the transformations are very mild and even when the sample size of the experiment is very large.

Considering these sensitivities, it may be better to avoid strong modeling assumptions about the form of the likelihood unless such assumptions are truly necessary or unless the researcher has reliable knowledge about the true form of the likelihood function. Similarly, strong priors should be used with care. For example, if Fisher’s exact test suffices to answer the question at hand, it may be safest to choose this test for analysis, since it is unnecessary to explicitly pose a likelihood for such a test. In the next chapter, however, we will see an example where such simple tests do not apply (due to experimental complications such as noncompliance and data collection problems that result in missing data) and where model-based Bayesian approaches perform better than the considered alternatives.

### 2.2 Introduction

Bayesian model-based inference is commonly used for assessing treatment effects in randomized experiments. These Bayesian models typically model the potential outcomes as dependent on some unknown model parameters. Choosing a model can be thought of as analogous to specifying a likelihood function, which characterizes the relationship between the data and the parameters, with the further step of selecting a prior distribution for the
parameters.

We investigate the sensitivity of causal inferences to misspecification of the likelihood function, the choice of prior, as well as interactions with experimental design choices, such as the sample size or the assignment mechanism. We explore the influence of these various forms of model misspecification on causal inferences via a full factorial simulation experiment. Each factor in the experiment represents a different form of model misspecification or an experimental design choice. We can then investigate which factors (i.e. which forms of model misspecification) or interactions of factors most affect the causal estimates via ANOVA. Once we have identified the most interesting factors, we focus analysis on these most consequential factors. Of course, we caveat these results with the observation that the relative influence of each factor depends on the levels that we choose to include in the experiment.

One difference between this study and prior studies is the Bayesian nature of our evaluation criteria. We do not, for example, consider the frequency coverage of posterior intervals for fixed parameter values. Rather, we start by treating our parameters as random variables and by using the data generating prior as our true prior. With the true prior and the true likelihood function, we can construct the true posterior distribution of the causal estimand.

We can then compare the true posterior to a misspecified posterior (a posterior whose likelihood function or prior distribution is incorrectly chosen) via a variety of measures. One measure that we consider is the coverage (under the true posterior) of a credible interval that is constructed with a misspecified posterior. Another measure is the Kullback-Leibler divergence from the misspecified posterior distribution to the true posterior distribution.

We begin by first summarizing the full Bayesian approach for estimating finite population causal estimands from data generated by randomized experiments.

We then describe the full factorial simulation experiment, first discussing the simulation algorithm and then discussing each of the experimental factors in some detail; particular emphasis is paid to the prior, the transformation of the outcome variable, and the assignment mechanism, as these are the least straightforward (and, often, the most interesting) factors
considered in the experiment.

Next, we discuss the outcomes of interest for the full factorial experiment. As mentioned above, this primarily involves the coverage of misspecified intervals under the true posterior distribution and the KL divergence from the misspecified distribution to the true distribution. We also consider a variety of secondary outcomes in order to help generate intuition about the results for the primary outcomes. In particular, we consider outcomes that indicate whether the misspecified posterior distribution is poorly centered, whether its standard deviation is too large or small, and so forth. In other words, these secondary outcomes are mostly used as diagnostics to identify the manner in which the misspecified posterior is failing (or not).

We next consider the results of the experiment. The ANOVA tables for each outcome universally agree that the choice of prior and the failure to transform the outcome are by far the most important factors that are considered in the experiment. By “failure to transform” we mean that the outcome should have been, for example, log transformed and then modeled with a multiple linear regression model on the transformed scale. Several types of Box-Cox transformations are studied in this experiment.

Misspecified posterior inferences can be especially misleading when the data generating prior is used in conjunction with a failure to transform the outcome variable appropriately. Such modeling mistakes can lead to posterior intervals that, on average, have nearly 0 coverage under the true posterior distribution. The use of a flat prior can help to mitigate undercoverage problems, although undercoverage may still occur.

We then look at particular examples to illustrate why these misspecified models underperform. The choice of prior can dramatically shift the mean of the posterior distribution of the estimand, even for relatively large sample sizes. The use of a flat prior, by contrast, tends to lead to true and misspecified posterior means that are quite similar (when averaged over many simulated data sets). The posterior intervals obtained with a flat prior, however, can be quite a bit wider than posterior intervals obtained with the true prior. Perhaps more concerning, the misspecified posterior intervals can be much too narrow when the true posterior
distribution is skewed and the misspecified posterior distribution is relatively symmetric.

We consider examples where even very mild incorrect transformations of the outcome variable and very large sample sizes exhibit the same inferential problems. Lastly, even factors such as the choice between two ignorable assignment mechanisms can have a meaningful impact on the average quality of a misspecified posterior distribution of the causal estimand.

In short, posterior inferences for causal estimands can be quite sensitive to the choice of prior, especially in conjunction with a failure to properly transform the outcome variable. Causal inferences in these scenarios can be quite poor even with flat priors. Nevertheless, the use of a flat prior does mitigate these problems somewhat, although often at the cost of overly wide posterior intervals. In short, it may be wise to avoid strong modeling assumptions if there are alternative approaches (for example, Fisher’s exact test) that do not require as many assumptions.

2.3 Previous Work

This work is inspired by previous efforts that describe the need for “calibrated Bayes.” “Calibrated Bayes” is the idea that inferences from a particular model should be Bayesian but that model evaluation and model selection can include frequentist ideas (see e.g. Little 2006 and Little 2011). As Little (2006) says, frequentist ideas are “a set of concepts for assessing properties of inference procedures rather than an inference system per se.”

The idea of calibrated Bayes traces back at least to Welch (1965) but Box (1980) and Rubin (1984) are cited as “seminal” contributions (see, for example, Little 2006). Box proposed “prior predictive checks” for model assessment whereas Rubin proposed, among other ideas, posterior predictive checks, an idea which was further developed by Gelman, Meng, and Stern (1996).

More generally, Rubin (1984) suggests that inferences, conditional on a model specification, should be Bayesian but that the task of model selection and model evaluation should
be informed by frequency calculations that are both “Bayesianly relevant” and “Bayesianly justifiable.” A calculation is “Bayesianly justifiable” if it conditions on observed data and model specifications to calculate, via Bayes’s Theorem, the posterior distribution of the unknowns of interest (in particular, it treats unknown values as unobserved random variables). A calculation is “Bayesianly relevant” if it aids the process of model validation and model selection. A process such as standard goodness-of-fit tests, for example, may be Bayesianly relevant insofar as they can help sanity check model candidates. However, such tests are not obviously Bayesianly justifiable (although some such procedures may possibly be reformulated in Bayesianly justifiable ways).

Why do we need Bayesianly relevant and justifiable frequency calculations? An opposing view is that Bayesian probability statements could be justified merely as a rational way to update one’s personal, subjective beliefs in light of new evidence. However, we also desire to use Bayesian models for the advancement of science. In order to be scientific, Bayesian statements should be falsifiable on the basis of empirical observation.

With falsification based on empirical observation in mind, one may use frequency calculations to question whether the model was adequately posed in the first place. Generally speaking, this is the idea of “calibration.” Rubin argues that Bayesian probability statements should be calibrated to “real world events” and that a “Bayesian is calibrated if his probability statements have their asserted coverage in repeated experience” (Rubin, 1984). For example, a sequence or subsequence of Bayesian 95% credible intervals are calibrated if they actually cover their unknowns 95% of the time.

One way to investigate the calibration of a misspecified Bayesian model is by reference to the true Bayesian model (setting aside, for now, questions about what it means for a Bayesian model to be “true”). If we accept that Bayesian inferences simply flow from the laws of probability once we have correctly specified the model (via the application of Bayes’s Theorem), then it is reasonable to judge the adequacy of inferences from suggested models against the gold standard of inferences from the correctly specified Bayesian model.
The use of the true posterior as a benchmark is a key idea in Rosenbaum and Rubin (1984), where the performance of a misspecified Bayesian model was evaluated based on its performance relative to the correctly specified model over many simulated datasets. Their simulation study investigated how the amount of undercoverage due to model misspecification depends on data-dependent stopping rules. The study was motivated by the claim that Bayesians need not worry about data-dependent stopping rules because such rules are ignorable.

The datasets were simulated via the true Bayesian model. Rosenbaum and Rubin then calculated credible intervals with the misspecified model (using the true likelihood but incorrectly using a flat prior) and computed the coverage of this interval under the true posterior. The study found that, even when the true model and the misspecified model are fixed, the undercoverage of the misspecified interval can vary dramatically based on which ignorable stopping rule was (or wasn’t) used. Specifically, undercoverage was much more likely to occur with an extreme stopping rule (e.g. “stop when the interim effect size estimate exceeds some large threshold”) than in designs that had no such stopping rule (or had milder stopping rules). Ignorability, in some sense, depends on correct model specification.

In many ways, this current study is a natural extension of Rosenbaum and Rubin. Unlike Rosenbaum and Rubin, which mainly considered the interaction between prior misspecification and the use of ignorable stopping rules, this simulation study also considers various forms of likelihood misspecification. Much has been written about the subjective process of choosing a prior, although the choice of likelihood also depends on the subjective judgment of the analyst. Indeed, misspecification of the likelihood may be more consequential than misspecification of the prior in certain cases. While we do not consider the effects of using different ignorable stopping rules, we do consider the effects of using different ignorable assignment mechanisms.

Similar to Rosenbaum and Rubin, in this study there is no question about whether there is a true prior and a true likelihood. Indeed, we constructed a true prior and true likelihood
and then used them to generate the data (below we describe what we mean, in general, by the phrase “true prior”). As an evaluation metric, Rosenbaum and Rubin principally considered the posterior coverage of misspecified credible intervals. We will additionally consider other evaluation criteria for comparing the true posterior distribution to the misspecified posterior distributions (for example, the Kullback-Leibler divergence from the misspecified posterior distribution to the true posterior distribution).

2.4 Background: Bayesian Model-Based Approaches to Causal Inference

In the Bayesian modeling approaches we will consider, we begin with the random vectors \((Y(1), Y(0), X, W)\). We can factor the joint distribution as

\[
p(Y(1), Y(0), X, W) = p(W | Y(1), Y(0), X)p(Y(1), Y(0), X)
\]

This simple factorization is convenient when the assignment mechanism is ignorable (for a definition of ignorability, see the discussion of assignment mechanisms in Section 2.5.3)– we may simply ignore the \(p(W | Y(1), Y(0), X)\) term and focus on the joint distribution of the data \(p(Y(1), Y(0), X)\) when the model is correctly specified.

Specifying the joint distribution of these quantities, \(p(Y(1), Y(0), X)\), can be a daunting task. However, assuming row exchangeability and with an appeal to de Finetti’s theorem, we model this joint distribution as the integral of the product of IID unit-level distributions.

\[
p(Y(1), Y(0), X) = \int \prod_{i=1}^{N} [p(Y_i(1), Y_i(0), X_i | \theta)] p(\theta) d\theta
\]

The unknown parameter vector \(\theta\) has the prior distribution \(p(\theta)\). It is often convenient to factor the unit-level joint conditional distributions \(p(Y_i(1), Y_i(0), X_i | \theta)\) like so:

\[
p(Y_i(1), Y_i(0), X_i | \theta) = p(Y_i(1), Y_i(0) | X_i, \theta_{Y|X})p(X_i | \theta_X)
\]
θ_{Y|X} and θ_X are functions of θ and are often modeled as a priori independent. With the above reasoning, we have reduced our main task to the specification of the joint conditional distribution \( p(Y_i(1), Y_i(0) \mid X_i, \theta_{Y|X}) \) (a presumably easier task than specifying \( p(Y(1), Y(0), X) \)) and specification of the prior distribution \( p(\theta) \).

The prior distribution that appears in de Finetti’s theorem can be thought of as the true prior. It is synonymous with the data generating prior. This is the prior that we use in our simulation studies when we compute the true posterior distribution of the causal estimand.

When the assignment mechanism is ignorable (see below), our primary task, aside from fitting the model, is to correctly specify the prior \( p(\theta) \) and the likelihood function,

\[
\prod_{i=1}^{N} p(Y_i(1), Y_i(0) \mid X_i, \theta)
\]

In practice, we don’t usually expect to get this perfectly right. The rest of this study, after a short description of model-based Bayesian inference, explores the consequences of misspecification of the true prior and/or the true likelihood during analysis.

**Bayesian Inference on the Finite Population Causal Estimand**

In model-based Bayesian causal inference the primary goal is to derive the posterior of some estimand of interest, which we denote \( \tau \). For example, we might choose the average finite population causal effect of the treatment as the estimand \( \tau \).

We obtain the posterior for \( \tau \) by first obtaining the posterior predictive distribution of the missing potential outcomes given the observed data, \( p(Y^{mis} \mid Y^{obs}, X, W) \). The estimand, \( \tau \), is typically just a simple function of the observed potential outcomes and the missing potential outcomes such that obtaining the posterior predictive distribution of the missing potential outcomes immediately yields the posterior for \( \tau \).

We perform the following steps to obtain the posterior distribution of \( \tau \), in the case where the assignment mechanism is ignorable (Imbens and Rubin, 2015):

1. First, specify the joint distribution \( p(Y_i(1), Y_i(0), X_i \mid \theta) \) and the prior distribution \( p(\theta) \)
2. Derive \( p(Y^{\text{mis}} \mid Y^{\text{obs}}, X, W, \theta) \), the conditional distribution of the missing potential outcomes.

3. Derive \( p(\theta \mid Y^{\text{obs}}, X, W) \), the posterior distribution of \( \theta \).

4. Use these two previous results to obtain the posterior predictive distribution of the missing data, \( p(Y^{\text{mis}} \mid Y^{\text{obs}}, X, W) \), by integrating out \( \theta \).

5. Since the estimand \( \tau \) is merely a function of the observed data and the missing potential outcomes, use the posterior predictive distribution of the missing data and the observed data to derive the posterior distribution of \( \tau \) given the observed data, \( p(\tau \mid Y^{\text{obs}}, X, W) \).

### 2.5 Methods

#### 2.5.1 Simulation Experiment

We ran a \( 5 \times 3 \times 3 \times 3 \times 3 \times 3 \times 2 \times 2 \) full factorial simulation experiment to investigate the effect of model misspecification on the validity of posterior inferences for the average finite population causal estimand. Table 2.1 lists the various factors and their levels.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Size, N</td>
<td>100, 400, 1000</td>
</tr>
<tr>
<td>Number of Covariates, P</td>
<td>1, 2, 3</td>
</tr>
<tr>
<td>Assignment Mechanism</td>
<td>CR, RR, FSM</td>
</tr>
<tr>
<td>True Likelihood</td>
<td>Normal, ( t_7 ), ( t_{15} )</td>
</tr>
<tr>
<td>Analysis Likelihood</td>
<td>Normal, ( t_7 ), ( t_{15} )</td>
</tr>
<tr>
<td>Analysis Prior</td>
<td>Correct, Flat</td>
</tr>
<tr>
<td>Analysis covariates</td>
<td>All covariates, No covariates</td>
</tr>
<tr>
<td>Transformation</td>
<td>Identity, Log, ( 1 - \frac{1}{y} ), ( \frac{y^2-1}{3} ), ( \frac{1-y^{-3}}{3} )</td>
</tr>
</tbody>
</table>

Each experimental setting corresponds to a data generating process (which varies), experimental design (specifically, the choice of assignment mechanism and sample size), and modeling assumptions (which are typically incorrect). For each combination of factor levels,
we simulate data sets and, for example, construct posterior intervals for \( \tau \), the average finite population treatment effect, using the misspecified model, and then evaluate the coverage of these intervals using the true posterior for \( \tau \). Aside from the true coverage of the constructed intervals, we consider other measures of similarity between the true posterior distribution and the misspecified posterior distribution (see Section 2.5.4).

The true model includes the assignment mechanism, the model for the potential outcomes conditional on all relevant covariates and parameters (the likelihood function), and the prior distribution for the parameters. The misspecified model, by contrast, typically includes a misspecified likelihood function or an “incorrect” prior. The prior is “incorrect” in the sense that it is not the true, data generating prior that appears in de Finetti’s theorem.

In order to evaluate the performance of the misspecified models, we consider a variety of metrics that measure the divergence from the misspecified posterior distribution to the true posterior distribution, as well as other quantifications of how well the misspecified model performs. We discuss these evaluation criteria in detail below in Section 2.5.4.

### 2.5.2 Simulation Algorithm

For each combination of factor levels we do the following:

1. Sample an observed data set \((Y_{\text{obs}}, X, W)\)
   a) Draw all covariates, \(X\). First draw any necessary parameters \(\theta_X \sim p(\theta_X)\). Then draw the covariates \(X \sim p(X \mid \theta_X)\).
   b) Draw \(\theta_{Y \mid X} \sim p(\theta_{Y \mid X})\).
   c) Draw the potential outcomes \((Y(1), Y(0)) \sim p(Y(1), Y(0) \mid X, \theta_{Y \mid X})\).
   d) Sample an assignment vector \(W\) according to the chosen assignment mechanism
   e) Using the assignment vector \(W\) and the potential outcomes \((Y(1), Y(0))\), obtain the observed outcome data \(Y_{\text{obs}}\) where \(Y_{i,\text{obs}} = W_i Y_i(1) + (1 - W_i)Y_i(0)\) for each unit \(i\)
2. Derive the posterior distribution of the estimand $\tau$ under the true model:

$$p(\tau \mid Y^{obs}, X, W)$$

3. Derive the posterior distribution of the estimand $\tau$ under the misspecified model:

$$q(\tau \mid Y^{obs}, X, W)$$

4. Calculate the outcomes of interest that compare the true posterior to the misspecified posterior

- For example, calculate the posterior probability coverage of $I_q(Y^{obs}, X, W)$ under the true posterior. $PC(Y^{obs}, X, W) = p(\tau \in I_q(Y^{obs}, X, W) \mid Y^{obs}, X, W)$

where $I_q(Y^{obs}, X, W)$ is the 95% posterior interval for the causal estimand that is constructed using the misspecified posterior $q$ (see Section 2.5.4)

5. Repeat the above steps for 250 simulated data sets for each experimental condition in order to obtain a distribution for each of the outcomes over repeated samples of $(Y^{obs}, X, W)$.

### 2.5.3 Factors

Refer to Table 2.1 for a list of the factors used in the experiment along with their levels.

**Sample Size**

The number of units, $N$, in any particular simulated dataset can be either 100, 400, 1000.

**True Number of Covariates**

The true number of covariates, $P$, in the data generating process can take three levels: 1, 2 or 3. All covariates are IID from a standard normal distribution. I.e. for $\forall i \in \{1, \ldots, n\}$, $\forall j \in \{1, 2, 3\}$:
Assignment Mechanism

We take the time here to review some properties of assignment mechanisms. An assignment vector, denoted $W$, is a vector indicating each unit’s assignment to either the active treatment or the control treatment (in the case of a binary treatment). An assignment mechanism is a function that gives the probability of all possible assignment vectors, conditional on all covariates and potential outcomes.

An advantage of some randomized assignment mechanisms is that they achieve balance (in expectation) on the background covariates. This helps avoid any systematic, confounding differences in the distribution of covariates between treatment and control.

There are several important properties that assignment mechanisms may have. An assignment mechanism is *individualistic* if the probability of assignment does not depend on the covariates or potential outcomes of other units. An assignment mechanism is *probabilistic* if each unit has a non-zero probability of assignment to each possible treatment. An assignment mechanism is *unconfounded* if it is independent of the potential outcomes.

An assignment mechanism is *strongly ignorable* if it is individualistic, probabilistic and unconfounded (Rosenbaum and Rubin, 1983). More generally, an assignment mechanism is *ignorable* if:

1. The distribution of the assignment vector does not depend on the missing potential outcomes (the missing potential outcomes are missing at random (MAR)):

   $$p(W | Y_{obs}, Y_{mis}, X) = p(W | Y_{obs}, Y_{mis}^*, X) \text{ for } \forall Y_{mis}, Y_{mis}^*$$

   $X$ denotes the matrix of covariates, $Y_{obs}$ is the vector of the observed outcomes for the $N$ experimental units, and $Y_{mis}$ is the vector of missing potential outcomes. $Y_{mis}^*$ is a vector of other values that the missing potential outcomes could have assumed.
2. The parameters that parameterize the distribution of the assignment mechanism are a priori independent of the parameters for the distribution of the potentially observable data.

If the assignment mechanism depends on recorded covariates, then any dependency of the outcome variable on those same covariates, \( p(Y_{\text{obs}} | X, \theta) \), must be modeled.

When the assignment mechanism is ignorable, then likelihood-based estimation methods (including many Bayesian methods) may be applied without explicitly accounting for the assignment mechanism. Specifically, there is no need to model the distribution of the assignment mechanism when deriving the posterior distribution of a causal estimand of interest; it can simply be ignored during the analysis. Similarly, ignorable assignment mechanisms do not affect maximum-likelihood estimates of causal estimands, since contributions to the likelihood from the assignment mechanism can be absorbed into a proportionality constant.

These likelihood-based approaches to causal inference may be contrasted with other causal inference techniques (such as Fisher’s exact test), which must account for the assignment mechanism.

In this study we consider three ignorable assignment mechanisms for randomized experiments: complete randomization, rerandomization and the finite selection model.

**Complete Randomization**

Suppose there are \( N \) total units, with \( N_t \) assigned to treatment and \( N_c \) assigned to control. In a completely randomized design, all assignment vectors that assign \( N_t \) units to treatment and \( N_c \) units to control have positive probability and, indeed, all of these assignment vectors have the same probability of being drawn. In other words,

\[
P(W | X, Y(0), Y(1)) = \frac{1}{\binom{N}{N_t}} I \left\{ \sum_{i=1}^{N} W_i = N_t \right\}
\]

\( Y(1) \) and \( Y(0) \) denote the vectors of potential outcomes under treatment and control, respectively. Note that the complete randomization assignment mechanism does not depend
on the potential outcomes, covariates, or any unknown parameters.

Rerandomization

While complete randomization achieves balance between treatment groups in expectation, rerandomization is a design that directly constrains the set of possible assignment vectors in order to improve actual balance across many covariates. A criterion for balance must first be specified and this criterion must be calculable as a function of the observed covariates and some proposed assignment vector. One then randomly generates an assignment vector in the same manner as one would in the complete randomization scheme. However, instead of automatically accepting this proposed assignment vector, the assignment vector is rejected if it fails to satisfy the balance criterion. If the proposed assignment is rejected, new assignment vectors are proposed until one finally satisfies the balance criterion and is therefore accepted.

Specifically, we consider rerandomization with a Mahalanobis distance criterion, which was proposed and investigated by Morgan and Rubin (2012). We compute the following balance metric:

$$M = (\bar{X}_t - \bar{X}_c)'[\text{Cov}(\bar{X}_t - \bar{X}_c)]^{-1}(\bar{X}_t - \bar{X}_c)$$

where $\text{Cov}(\bar{X}_t - \bar{X}_c)$ is the sample covariance matrix of the difference in covariate means between the treatment and control groups. $M$ is the Mahalanobis distance between the vector of covariate means for the treatment group, $\bar{X}_t$, and the corresponding vector of covariate means for the control group, $\bar{X}_c$, for the set of covariates, $X$, and is thus a univariate measure of covariate balance between the treatment and control groups. A randomization will be accepted if $M$ falls below some specified constant $\alpha$, which we can choose such that assignment vectors are accepted with some desired probability $p_\alpha$.

In this experiment, we choose an acceptance threshold $\alpha$ such that $p_\alpha = P(M \leq \alpha) = 0.001$, where $M$ is the Mahalanobis difference between the vectors of a treatment and control covariate means.
Finite Selection Model

The Finite Selection Model (FSM) is an ignorable, although not individualistic assignment mechanism introduced in Morris (1979). Like rerandomization, it is a covariate-dependent assignment mechanism that attempts to construct randomizations that are balanced on the available covariates.

The FSM assigns units to treatment or control sequentially in a way that attempts to optimize covariate balance at each step. The method can be described as follows:

1. Assign a randomly selected unit to either treatment or control. Select treatment or control randomly with equal probability.

2. Take turns allocating units between the treatment and control groups from the pool of remaining unallocated units. The experimental unit selected at each round depends on a specified selection function that is a function of the covariates that we wish to balance.

3. Continue until all units are allocated.

As a result of using the same selection function, the treatment groups are likely to be better balanced on the observed covariates, if a reasonable selection function is used. Whether a unit is selected at a particular stage is typically dependent on the covariates of the units already allocated to the target treatment group. Specifically, the selection function typically involves the covariate matrix of the already allocated units. Suppose treatment \( w \) is selecting its next unit at round \( j \). Let \( \mathbf{X}_{w,i}^{(j)} \) be the covariate matrix of units already assigned to treatment \( w \) up to step \( j \) and including and for some proposed, unselected unit \( i \). Then our selection function chooses the unit that minimizes the trace (or, perhaps, the determinant) of \((\mathbf{X}_{w,i}^{(j)})'(\mathbf{X}_{w,i}^{(j)})^{-1}\) over \( i \), which indexes the unallocated units. Such FSM selection functions are chosen for their close relationship to the theory of optimal experimental designs.
In this experiment, we use a selection function that minimizes the determinant of the
covariate matrix of the already-selected units and the proposed new unit.

True Likelihood

For simplicity, we choose a true data generating process with a linear relationship between
the (untransformed) potential outcomes and the covariates and treatment indicator.

\[ Z_i(W_i) = \beta_0 + \beta_1 X_{i,1} + \ldots + \beta_p X_{i,p} + \beta_W W_i + \epsilon_i \]

The untransformed potential outcomes are independent, conditional on the covariates
and parameters. The distribution of the noise term \( \epsilon_i \) always has mean 0 and variance \( \sigma_y^2 \)
(i.e. \( \epsilon_i \sim [0, \sigma_y^2] \)). The probability distribution of the errors is a factor in the experiment and
follows one of three possible distributions: a normal distribution, a \( t_7 \) distribution (which is
close to the logistic) and the \( t_{15} \) distribution.

Typically, a transformation is applied to the untransformed potential outcomes in order
to obtain the observed data (see below).

Analysis Likelihood

The distribution of the errors in the analysis model is also a three level factor with these levels:
normal, \( t_7 \) and \( t_{15} \). Another form of likelihood misspecification – whether the covariates are
omitted from the model, is discussed below. The outcome variable is never transformed in
the misspecified model; the potential outcomes are always modeled as a linear function of
the covariates and the assignment, with additive, zero mean noise.

Transformation of the Response

The transformation factor is a five level factor that indicates the transformation that is
necessary for proper data analysis. In the simulation experiment, the true posterior is cal-
culated by applying the correct transformation. The misspecified posterior is calculated by mistakenly failing to transform the outcome variable.

The possible required transformations are all Box-Cox transformations, where the Box-Cox transformation parameter $\lambda$ can take these values: -3, -1, 0, 1, 3. We consider the factor level $\lambda = 1$ to represent the identity transformation. A value of 0 indicates that a logarithmic transformation is necessary. Other values of $\lambda$ indicate that the following transformation is required for proper analysis:

$$h_\lambda(y) = \frac{y^\lambda - 1}{\lambda}$$

The inverse Box-Cox transformation is:

$$g_1(z) = z$$

for $\lambda = 1$

$$g_0(z) = e^z$$

for $\lambda = 0$ and

$$g_\lambda(z) = (\lambda z + 1)^{\frac{1}{\lambda}}$$

for other values of $\lambda$. The levels of $\lambda$ include 0 and odd integer values to ensure that the transformation is invertible on the real line.

In other words, the true data generating model is:

$$(Y_i(1), Y_i(0)) = (g_\lambda(Z_i(1)), g_\lambda(Z_i(0)))$$

$$(Z_i(1), Z_i(0)) \mid \theta \sim f(Z_i(1), Z_i(0) \mid \theta)$$

$$\theta \sim \pi(\theta)$$
In this notation, $\theta$ is the parameter (including the scale parameter $\sigma^2_y$ and $\beta$, the regression coefficients), $f(\cdot | \theta)$ indicates the distribution of the untransformed potential outcomes and is parameterized by $\theta$ (corresponding to either a normal or a t distribution), and $\pi(\theta)$ is the prior distribution of $\theta$. Again, $g_{\lambda}$ is the relevant inverse Box-Cox transformation.

This factor will be the most important factor in our experiment and so it is worth giving an example. Suppose that $Y^{obs}$ is the observed outcome of interest in some experiment. For example, it could be the settlement amounts for asbestos lawsuits. Suppose that a law firm that specializes in this practice typically can choose whether to file in state or federal court. The firm randomizes its choice of jurisdiction in an experiment in an effort to measure whether settlement amounts are greater (or lesser) because of the decision to file in state court. Our estimand is $\tau = \frac{1}{n} \sum_{i=1}^{n} (Y_i(1) - Y_i(0))$.

Let’s suppose that the true data generating model is as follows:

$$Y_i(W_i) = e^{Z_i(W_i)}$$

$$Z_i(W_i) = \log(Y_i(W_i)) \mid \beta, \sigma^2 \sim N(\beta_0 + \beta_1X_1 + \beta_2W_i, \sigma^2)$$

$$\beta \sim N(\mu_0, \Sigma_0)$$

$$\sigma^2 \sim Inv-Gamma(\alpha_0, \beta_0)$$

Here we define $Z_i(W_i)$ to be the log transform of the potential outcome $Y_i(W_i)$. For analysis, we would thus first transform $Y^{obs}$ to the log scale (a Box-Cox transformation with $\lambda = 0$), then we would obtain posterior draws of our parameters, $\beta$ and $\sigma^2$, using the usual MCMC methods, and then we would obtain draws of the posterior of the transformed potential outcomes, $Z(1)$ and $Z(0)$. For each iteration of the MCMC algorithm this would yield a “science table” (see Table 2.2):

To obtain the potential outcomes of interest, $Y(1)$ and $Y(0)$, we then back transform the outcomes $Z(1)$ and $Z(0)$.
Table 2.2: Science Table for Transformed Outcome

<table>
<thead>
<tr>
<th>$Z(1) = \log(Y(1))$</th>
<th>$Z(0) = \log(Y(0))$</th>
<th>$W$</th>
<th>$Z_{\text{obs}}$</th>
<th>$X$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_{11}$</td>
<td>$z_{01}$</td>
<td>1</td>
<td>$z_{11}$</td>
<td>$x_1$</td>
</tr>
<tr>
<td>$z_{12}$</td>
<td>$z_{02}$</td>
<td>0</td>
<td>$z_{02}$</td>
<td>$x_2$</td>
</tr>
<tr>
<td>$z_{13}$</td>
<td>$z_{03}$</td>
<td>1</td>
<td>$z_{13}$</td>
<td>$x_3$</td>
</tr>
<tr>
<td>$z_{14}$</td>
<td>$z_{04}$</td>
<td>0</td>
<td>$z_{04}$</td>
<td>$x_4$</td>
</tr>
<tr>
<td>$z_{15}$</td>
<td>$z_{05}$</td>
<td>1</td>
<td>$z_{15}$</td>
<td>$x_5$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 2.3: Science Table for Outcome

<table>
<thead>
<tr>
<th>$(Y(1))$</th>
<th>$(Y(0))$</th>
<th>$W$</th>
<th>$Y_{\text{obs}}$</th>
<th>$X$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_{11} = \exp(z_{11})$</td>
<td>$y_{01} = \exp(z_{01})$</td>
<td>1</td>
<td>$y_{11}$</td>
<td>$x_1$</td>
</tr>
<tr>
<td>$z_{12} = \exp(z_{12})$</td>
<td>$z_{02} = \exp(z_{02})$</td>
<td>0</td>
<td>$y_{02}$</td>
<td>$x_2$</td>
</tr>
<tr>
<td>$z_{13} = \exp(z_{13})$</td>
<td>$z_{03} = \exp(z_{03})$</td>
<td>1</td>
<td>$y_{13}$</td>
<td>$x_3$</td>
</tr>
<tr>
<td>$z_{14} = \exp(z_{14})$</td>
<td>$z_{04} = \exp(z_{04})$</td>
<td>0</td>
<td>$y_{04}$</td>
<td>$x_4$</td>
</tr>
<tr>
<td>$z_{15} = \exp(z_{15})$</td>
<td>$z_{05} = \exp(z_{05})$</td>
<td>1</td>
<td>$y_{15}$</td>
<td>$x_5$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td></td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

We can now calculate our estimand of interest:

$$
\tau = \frac{1}{n} \sum_{i=1}^{n} (Y_i(1) - Y_i(0)) = \sum_{i=1}^{n} (e^{Z_i(1)} - e^{Z_i(0)})
$$

It is interesting to note that, in the log-normal case, the parameters do not obviously correspond to any causal estimand. Suppose that the true model, rather than requiring a log transformation of the outcome, required no such transformation. I.e. suppose the following data generating model:

$$
Y_i(W_i) \mid \beta, \sigma^2 \sim N(\beta_0 + \beta_1 X_i + \beta_2 W_i, \sigma^2) \\
\beta \sim N(\mu_0, \Sigma_0) \\
\sigma^2 \sim Inv-Gamma(\alpha_0, \beta_0)
$$

In this case, $E(Y_i(1) - Y_i(0) \mid \beta, \sigma^2) = \beta_2$. Here, the expectation is taken over the data and the $\beta_2$ parameter does indeed correspond to a superpopulation causal estimand. By
contrast, let’s again consider the lognormal model that we considered above. Recalling basic properties of the lognormal, we see that:

$$E(Y_i(1) - Y_i(0) \mid \beta, \sigma^2) = e^{X_0 + \beta_1 X_1 + \beta_2 + \frac{\sigma^2}{2}} - e^{X_0 + \beta_1 X_1 + \beta_2 + \frac{\sigma^2}{2}}$$

$$= (e^{X_0 + \beta_1 X_1 + \frac{\sigma^2}{2}})(e^{\beta_2} - 1)$$

Here, $\beta_2$ no longer has the same interpretation. This simple example illustrates that the scientific interpretations of the model parameters depends on the specification of the model. Partly, this illustrates the danger of using the same notation for parameters in different models. It also underscores the danger of equating causal estimands with model parameters, given that such an equivalence will depend on correct specification of the model. We prefer to focus on estimands such as $\sum_{i=1}^{n} (Y_i(1) - Y_i(0))$ and treat the model parameters as tools for obtaining posterior draws of the actual estimand. Unfortunately, as we will see, this approach can also be strongly sensitive to correct specification of the model.

**Analysis Covariates**

Whether the analysis model contains the covariates or not is a two level factor of the experiment. The analysis model may contain all of the covariates in the true model, i.e.

$$Z_i(W) = \beta_0 + \beta_1 X_{i,1} + ... + \beta_p X_{i,p} + \beta_W W_i + \epsilon_i$$

Alternatively, the analysis model may omit all of the covariates and only includes the treatment indicator:

$$Z_i(W) = \beta_0 + \beta_W W_i + \epsilon_i$$
**True Prior**

The true, data generating prior for $\sigma_y^2$ is an inverse-gamma with shape parameter $\nu$ and scale $\nu s^2$ i.e.

$$\sigma_y^2 \sim Inv - Gamma(\nu = 10, \nu s^2 = 100)$$

The true prior on the regression coefficients ($\beta_0, \beta_1, ..., \beta_p, \beta_W$) is normal such that the coefficients are a priori independent:

$$(\beta_0, \beta_1, ..., \beta_p, \beta_W) \sim N((\beta_0^0, \beta_1^0, ..., \beta_p^0, \beta_W^0), 0.0025I_{p+2})$$

The $\beta$ parameters are also a priori independent of $\sigma_y^2$. The hyperparameter values are fixed as $\beta_0^0 = 2$, $\beta_1^0 = \ldots = \beta_p^0 = 3$ and $\beta_W^0 = 4$. These hyperparameters are selected to reflect a strong dependence on the covariates and a strong treatment effect. $I_{p+2}$ is an identity matrix of appropriate dimension. Recalling our earlier, generalized notation, $\theta_{Y|X} = (\beta_0^0, \beta_1^0, ..., \beta_p^0, \beta_W^0, \sigma_y^2)$.

**Analysis Prior**

The analysis prior can take two levels: a flat prior or the true prior. The flat prior is as follows:

$$p(\beta_0, \beta_1, ..., \beta_p, \beta_W, \sigma_y^2) \propto \frac{1}{\sigma_y^2}$$

These choices of prior are meant to either be correct (to the extent that a true prior can be “correct” even when the likelihood is misspecified) or to reflect the common and often convenient choice of a flat prior.

In the event that covariates are omitted from the model, the “true” prior consists of the components of the prior for the parameter components that are included in the model (the variance parameter, the coefficient for the treatment indicator, and the intercept). Since
each of the parameters in the parameter vector are a priori independent, we can easily omit the components that correspond to any omitted covariates, if covariates are omitted in the misspecified model.

Although we use the phrase “true prior” interchangeably with “data generating prior,” it isn’t clear what it means to have a “true” prior when the likelihood is incorrect. We will see below that using the data generating prior can be a suboptimal choice when the likelihood is incorrect. For this reason, we often put the word “true” in scare quotes when we discuss the “true” prior in a model with a misspecified likelihood. For example, in the case of omitted variables, we might interpret the parameters for the omitted covariates as having a dogmatic prior centered on 0. Such a choice would lead to an identical posterior distribution as the misspecified model. However, this is clearly not the correct prior.

As another example, suppose again that the true data generating model for the potential outcomes is as follows:

\[
Y_i(W_i) | \beta, \sigma^2 \sim \text{LogNormal}(\beta_0 + \beta_1 X_1 + \beta_2 W_i, \sigma^2)
\]

\[
\beta \sim N(\mu_0, \Sigma_0)
\]

\[
\sigma^2 \sim \text{Inv-Gamma}(\alpha_0, \beta_0)
\]

Now, suppose again that the misspecified model is as follows:

\[
Y_i(W_i) | \beta, \sigma^2 \sim \text{Normal}(\beta_0 + \beta_1 X_1 + \beta_2 W_i, \sigma^2)
\]

\[
\beta \sim N(\mu_0, \Sigma_0)
\]

\[
\sigma^2 \sim \text{Inv-Gamma}(\alpha_0, \beta_0)
\]

At first glance, it appears that the prior is correct in the misspecified model and that the likelihood is incorrect. However, in some sense, the prior and likelihood cannot be
disentangled so easily. In our appeal to DeFinetti’s theorem, we introduced the parameter as an aid in modeling the joint distribution of \((Y(1), Y(0), X, W)\). Suppose the correctly specified distribution of these vectors is \(p_{\text{true}}(Y(1), Y(0), X, W)\) and that the misspecified distribution is \(p_{\text{false}}(Y(1), Y(0), X, W)\). If we misspecify the likelihood function, there may be some prior other than the true data generating prior that minimizes the distance (for some definition of distance, perhaps the KL divergence) between the distributions \(p_{\text{true}}\) and \(p_{\text{false}}\). Suppose the true likelihood is \(Q_{N_{i=1}} p(Y_i(1), Y_i(0), X_i | \theta)\) and the misspecified likelihood is \(\prod_{i=1}^N q(Y_i(1), Y_i(0), X_i | \phi)\). Now, suppose the true prior is denoted \(f(\theta)\). Ideally, the misspecified model would use a prior \(g(\phi)\) such that

\[
\int \prod_{i=1}^N [q(Y_i(1), Y_i(0), X_i | \phi)] g(\phi) d\phi = \int \prod_{i=1}^N [p(Y_i(1), Y_i(0), X_i | \theta)] f(\theta) d\theta
\]

This idea will become more clear in the Results (Section 2.6), where we show that the data generating prior dramatically underperforms a flat prior when the likelihood function is misspecified in certain ways.

A Note on Computation

These choices for the likelihood and prior allow for straightforward simulation from the posterior of \(\theta\) using a Gibbs sampler. When the likelihood is \(t\) distributed, it is convenient to represent the \(t\) hierarchically. Suppose \(Y \sim t_{\nu}(\mu, \Sigma)\). Then we have the following representation for \(Y\):

\[
Y | \lambda \sim N_n(\mu, \lambda\Sigma)
\]

\[
\lambda \sim \text{Inv-Gamma}(\alpha = \frac{\nu}{2}, \beta = \frac{\nu}{2})
\]

We verify that all implemented Gibbs samplers work as intended by using the diagnostic methods described in Cook, Gelman and Rubin (2000) for testing Bayesian model-fitting software.
2.5.4 Evaluation Criteria

There are multiple possible evaluation criteria for evaluating a misspecified posterior distribution for the causal estimand by comparing it to the true posterior distribution.

A frequentist evaluation of a statistical procedure would call for analysis of the performance of that procedure over repeated sampling. For example, to evaluate a procedure by which confidence intervals are constructed, one could compare the nominal coverage of the intervals to the actual coverage of such intervals over repeated sampling, where coverage is defined relative to a fixed parameter value. This sort of frequentist evaluation can also be applied to a Bayesian procedure, such as the model-based Bayesian procedure that we use in the simulations to construct posterior intervals. For example, we could assume a fixed parameter value and investigate how often posterior intervals cover that parameter value over repeated data sets.

By contrast, we will emphasize Bayesian evaluation criteria for evaluating our misspecified posterior distributions for the causal estimand. For example, one of our primary outcomes for our full factorial experiment is the coverage (under the true posterior distribution) of the 95% credible intervals that are constructed using the misspecified posterior distribution. In other words, we do not consider how often an interval covers a “true,” fixed parameter value over repeated sampling; indeed our data sets are not generated from some fixed parameter but rather a parameter that is randomly drawn from some true prior distribution. Rather, we consider the coverage of an interval (constructed via the misspecified posterior distribution) under the true posterior. All of our other criteria similarly quantify the similarity of the misspecified posterior distribution to the true posterior distribution (for example, the KL divergence between these two distributions). We consider these criteria below.

Coverage Under the True Posterior

Suppose we have some outcome of interest $Y$. Further suppose that $f(Y \mid \theta)$ is the “true” model of the data conditional on the model parameters and that $p(\theta)$ is the “true” prior
distribution of the model parameters. As mentioned above, these constituent model components are “true” in the sense that they constitute the true data generating process (recall our appeal to de Finetti’s Theorem); we generate the data by first drawing $\theta$ from the prior and then drawing the potential outcomes conditional on $\theta$ and the covariates. To make inferences about the estimand using data generated by this process, one would obviously want to use the true model. The true posterior distribution of $\theta$ is:

$$p(\theta \mid Y) = \frac{f(Y \mid \theta)p(\theta)}{\int_{\theta} f(Y \mid \theta)p(\theta)d\theta}$$

Of course, the true data generating model is typically unknown. One must instead pose some other model for analysis, which is often chosen, at least in part, for convenient computational or theoretical properties. We call this the “misspecified” or “analysis” model. The analysis model has a likelihood defined by $g(Y \mid \theta)$ and prior $q(\theta)$, which yields the misspecified posterior $q(\theta \mid Y)$. Using the analysis model, one can construct $(1 - \alpha)$ posterior interval for $\theta$, $I_q(Y)$. For $\alpha \in (0, 1)$, we construct $I_q(Y)$ such that it satisfies

$$\frac{\int_{\theta \in I_q(Y)} g(Y \mid \theta)q(\theta)d\theta}{\int_{\theta} g(Y \mid \theta)q(\theta)d\theta} = 1 - \alpha$$

Following Rubin and Rosenbaum (1984), for some fixed $Y$, the posterior coverage, $PC(Y)$, of $I_q(Y)$ over the correct posterior distribution is then

$$PC(Y) = \frac{\int_{\theta \in I_q(Y)} f(Y \mid \theta)p(\theta)d\theta}{\int_{\theta} f(Y \mid \theta)p(\theta)d\theta}$$

In other words, $PC(Y)$ measures how much of the true posterior probability mass lies within the proposed interval $I_q(Y)$. Depending on how bad the misspecification is, a posterior interval with nominal posterior probability coverage 0.95 (for example) might have true posterior coverage that falls dramatically below (or above) 0.95. Indeed, we will observe experimental conditions where the average coverage of such intervals is 0. Of course, if the analysis model is well specified (i.e. $g(Y \mid \theta) = f(Y \mid \theta)$ and $q(\theta) = p(\theta)$), then, by construction, $PC(Y) = 1 - \alpha$. 

30
For each set of simulation conditions we do the following:

1. Generate the data from the true model. First, draw $\theta \sim p(\theta)$ and then draw $Y \sim f(Y | \theta)$

2. Using the misspecified analysis model $g(Y | \theta)q(\theta)$, construct a $1 - \alpha$ posterior interval $I_q(Y)$

3. Compute $PC(Y)$, the posterior probability of the event $\{\theta \in I_q(Y)\}$ over the true posterior $f(Y | \theta)p(\theta)$

Of course, $PC(Y)$ has some distribution that results from the randomness of $Y$, the dataset that we generate in the first step. Ideally, one would pose an analysis model such that $PC(Y)$ is tightly distributed around the nominal level $1 - \alpha$ for a range of plausible candidates for the true data generating model. In other words, ideally the analysis model would be chosen such that it is relatively robust to misspecification, if possible.

For each generated dataset $(Y^{obs}, X, W)$ we obtain one draw of $PC(Y^{obs}, X, W)$. We approximate the distribution of $PC(Y^{obs}, X, W)$ via Monte Carlo, generating many draws of $(Y^{obs}, X, W)$ and, via application of the simulation algorithm, many draws of $PC(Y^{obs}, X, W)$. The resulting distribution of $PC(Y^{obs}, X, W)$ would ideally be nearly a point mass centered on the interval’s nominal value (in our examples, 0.95).

**Kullback-Leibler Divergence**

Another natural measure of how well the misspecified posterior matches the true posterior is the Kullback-Leibler (KL) divergence between the two distributions.

The KL divergence is a common measure of distance between two distributions. Although it is not truly a metric, since it is not symmetric, this isn’t a problem here since we have a non-symmetric problem. Our problem is non-symmetric in the sense that we know that one of the distributions is the correctly specified distribution and that the other is the misspecified distribution.
The KL divergence (also known as relative entropy) of two densities $p$ and $q$ is defined as:

$$D(p \| q) = \int p(x) \log \left( \frac{p(x)}{q(x)} \right) dx$$

$D(p \| q)$ is always non-negative (an inequality often referred to as Gibbs’ inequality), it is 0 if and only if $p = q$, and it is finite whenever $p$ is absolutely continuous with respect to $q$.

The KL divergence arises naturally in the discussion of the maximum-likelihood estimates obtained from analysis of misspecified models. For example, suppose data is generated from some true distribution $p$. Now suppose that, for analysis, we choose a misspecified model $M_\theta$ ($M_\theta$ is misspecified in the sense that $p$ does not belong to the family of distributions that are elements of $M_\theta$). For example, suppose our data is generated by $p$, which is $\text{Gamma}(4,1)$ and we choose $M_\theta = \text{Normal}(\theta_1, \theta_2)$ as our model, which is here misspecified since $p$ is a normal distribution. If we estimate $\theta_1$ and $\theta_2$ with their maximum likelihood estimates (using the misspecified normal model to define the likelihood function), then we estimate the true data generating distribution as $q_\hat{\theta} = \text{Normal}(\hat{\theta}_1, \hat{\theta}_2)$. (See Pawitan (2013)).

White (1982) shows that, under mild regularity conditions, the maximum-likelihood estimator $\hat{\theta}_{\text{MLE}}$ (using the misspecified likelihood) is a strongly consistent estimator for the parameter vector $\theta_0$ that minimizes $D(p \| q_\theta)$, where $q_\theta \in M_\theta$. Returning to our example above, $(\theta_0, \theta_1) = (4, 4)$ is the parameter that minimizes the KL divergence between the $\text{Gamma}(4, 1)$ and the $\text{Normal}(\theta_1, \theta_2)$ model. Therefore, the MLE that is obtained using the misspecified normal model is consistent for $(\theta_0, \theta_1) = (4, 4)$.

For a correctly specified model (i.e. when $p \in M_\theta$), $D(p \| q_\theta) \to 0$ as $n \to \infty$, which is the classic result of the consistency of the MLE in a well-specified model. Similarly, for a misspecified model (i.e. $p \notin M_\theta$), $D(p \| q_\theta)$ will converge to some positive constant. We can thus think of $D(p \| q_\theta)$ as a measure of model misspecification. The KL divergence is often thought of as a measure of model misspecification in the Bayesian setting as well, as in, for
example, Kleijn and Van Der Vaart (2006).

In our simulations, computing the KL divergence between the true posterior and the mis-specified posterior is non-trivial because we typically do not have closed form representations of the densities. Typically, we can only simulate from these distributions. We therefore first calculate kernel density estimates of these densities from our Monte Carlo samples. Methods have been proposed for estimating the KL divergence between two continuous distributions via Monte Carlo samples without the intermediate step of constructing kernel density estimates (Fernando Perez-Cruz, 2008). However, we found this method to be unreliable for certain simulation conditions.

**Kolmogorov-Smirnov Statistic**

The Kolmogorov-Smirnov statistic is the test statistic used in the Kolmogorov-Smirnov test, which is a commonly used non-parametric hypothesis test for testing the equality of continuous, univariate distributions.

It is defined as

\[
KS = \sup_x | \hat{F}_{P,n}(x) - \hat{F}_{Q,m}(x) |
\]

where \( \hat{F}_{P,n}(x) \) is the empirical cumulative distribution function for a probability distribution \( p \) based on \( n \) data points and \( \hat{F}_{Q,m}(x) \) is the empirical distribution for a probability distribution \( q \) based on \( m \) data points. It can be thought of as a measure of distance between two empirical distributions.

One reason to introduce the Kolmogorov-Smirnov statistic as an evaluation criterion is that it is explicitly defined as a function of the empirical distributions, unlike the KL divergence. As discussed above, estimating the KL divergence using just posterior draws (i.e. without the normalized densities) can be challenging. It is therefore convenient that the KS is explicitly defined in terms of the ECDFs, which are determined by the draws from the posterior distributions.
Additionally, due to the Kolmogorov-Smirnov statistic’s ubiquitous use in testing the identity of two univariate distributions, we can also motivate the use this statistic as a measure of distance between the misspecified and true posteriors for this reason. The results for this outcome will be very similar to the results for the KL divergence.

**Normalized Difference in Posterior Means**

The remaining metrics are not so much measures of distance between two distributions as they are diagnostic outcomes that help generate intuition about the consequences of model misspecification in a particular experimental setting.

For example, suppose that the posterior coverage under a certain form of model misspecification is at the nominal coverage, but the KL divergence between the two distributions is large. How might we diagnose why the divergence is so large but the coverage is correct? Is it because the posterior credible interval under the misspecified model is much wider than the corresponding interval constructed with the true posterior? Or perhaps the intervals tend to be the same size but the interval for the misspecified model is always poorly centered. The diagnostics that we define below can help answer these sorts of questions.

The first such criterion that we consider is the normalized difference in posterior means. We define the normalized difference in posterior means as the following:

\[
\frac{\mu_Q - \mu_P}{\sigma_P}
\]

\(\mu_Q\) is the posterior mean of the causal estimand under the misspecified model, \(\mu_P\) is the posterior mean under the true model, and \(\sigma_P\) is the posterior standard deviation of the causal estimand under the true model. The true posterior distributions can be quite different, depending on the experimental conditions under consideration, and so we standardize by the true standard deviation in order to facilitate comparisons across these scenarios.

This measure can be thought of as an analogue to the concept of standardized bias. Whereas large differences in coverage probabilities or KL divergence indicate that two dis-
tributions are quite different, measures like the normalized difference in posterior means gives greater intuition for exactly how the misspecified posterior is failing. In particular, this measure will give us a sense of whether misspecification causes the posterior to be poorly centered relative to the true posterior.

**Ratio of Credible Interval Widths**

The ratio of the credible interval widths is the ratio of the width of the credible interval constructed using the misspecified model to the width of the credible interval constructed using the correctly specified model. I.e. it is the following quantity:

\[
\frac{F_Q^{-1}(0.975) - F_Q^{-1}(0.025)}{F_P^{-1}(0.975) - F_P^{-1}(0.025)}
\]

\(F_Q^{-1}(0.975)\) is the 0.975 quantile of the misspecified posterior distribution, \(q\), of the causal estimand. \(F_P^{-1}(0.025)\) is the 0.025 quantile of the true posterior, \(p\).

This metric provides a multiplier for how much wider or narrower the misspecified credible interval is relative to the credible interval that results from the true posterior. As a default, we will only consider 95% credible intervals. Often, the misspecified interval is much wider or narrower than the true interval, which can lead to significant undercoverage or overcoverage of the misspecified interval. It may also cause the KL divergence to explode.

Whereas the previous outcome attempts to capture some measurement of the bias of the misspecified posterior mean, this outcome aims to diagnose how the true and misspecified posteriors differ in terms of variance. As mentioned, the posterior coverage of the misspecified interval might achieve the nominal coverage but only at the expense of a very wide credible interval. Such intervals might be so wide that they are functionally useless, from any practical perspective.
Ratio of Posterior Standard Deviations

In the spirit of comparing the variance of the true and misspecified posterior distributions, we can directly compare the posterior standard deviations. The ratio of the posterior standard deviations gives us a sense of whether the form of model misspecification in any particular experiment setting causes the variance of the misspecified posterior distribution to change, relative to the variance of the correctly specified posterior. We define the ratio of the posterior standard deviations as follows:

\[
\frac{\sigma_Q}{\sigma_P}
\]

In this expression \(\sigma_P\) denotes the standard deviation of the true posterior, \(p\), and \(\sigma_Q\) denotes the standard deviation of the misspecified posterior, \(q\).

Ratio of Skews

We are also interested in whether the misspecified posterior distribution of the estimand is much more skewed than the correctly specified posterior. To that end, we can calculate the skew for each distribution and calculate the ratio of the two. In other words, if \(\tau_i\) denotes the \(i^{th}\) sample from the posterior of the estimand and if \(\sigma\) denotes the posterior standard deviation, we calculate the following:

\[
\gamma = \frac{\sum_{i=1}^{N}(\tau_i - \bar{\tau})^3/N}{\sigma^3}
\]

The ratio of the skews is then calculated as \(\frac{\gamma_Q}{\gamma_P}\), where \(\gamma_P\) is the skew for the true posterior, \(p\), and \(\gamma_Q\) is the skew for the misspecified posterior, \(q\).

This outcome sheds some light on which forms of model misspecification fail to capture the skew of the true posterior distribution. Additionally, model misspecification may also introduce excess skew that doesn’t exist in the true posterior distribution.
Rates of False Negatives and False Positives

We do not encourage using posterior credible intervals as a tool for making definitive yes/no declarations about whether a treatment effect exists or not in a particular study. Rather, when we calculate a posterior distribution of a causal effect, it is more appropriate to report the posterior, or summaries thereof, complete with the uncertainty that still remains about the size and sign of the treatment effect.

However, we also recognize that, in practice, researchers are often quite preoccupied with whether their credible intervals for a causal estimand cover 0. It is not unheard of for decisions to be made on the basis of such coverage or non-coverage of 0. Such decisions might include whether a research paper gets published or not, whether researchers report that some intervention has an impact (or not) on the outcome of interest, or whether more capital is devoted to development of a commercial product, for example. Ultimately, people conduct experiments in order to make decisions, in addition to advancing scientific knowledge.

With this in mind, we also keep track of the rate of “false negatives” and “false positives” that result from model misspecification. A “false negative” occurs when the true credible interval does not cover 0 but the misspecified credible interval does cover 0. More precisely, a false negative rate can be defined as follows:

\[ E(1_{0 \notin I_p(\tau)} \times 1_{0 \in I_q(\tau)}) \]

A “false positive” occurs in the opposite scenario, when the credible interval constructed using the true posterior covers 0 but the credible interval that corresponds with the misspecified posterior fails to cover 0. I.e. the false positive rate is:

\[ E(1_{0 \in I_p(\tau)} \times 1_{0 \notin I_q(\tau)}) \]

We investigate the false positive rate and false negative rate as two separate outcomes in the simulation study. The intervals that we construct from both the true and misspecified
posterior s are always constructed using the 0.975 and 0.025 quantiles.

**Summaries of the Evaluation Metrics**

For each of the outcomes described above, we can consider some summary of that outcome’s distribution over the 250 simulated datasets that are generated for each simulation condition. For each experimental setting, we have many possible outcomes of interest for each of our evaluation metrics. Typically, we will consider the mean of the evaluation metric, where, for each experimental setting, the mean is taken over the 250 simulated datasets. We can also consider the quantiles of the outcome, such as the 0.1 or 0.9 quantile. For example, for some experimental settings, the mean coverage of the misspecified interval might be very close to the nominal coverage. However, it may be close to the mean because it undercovers sometimes and overcovers sometimes, rather than always being close to the nominal coverage. The quantiles of the evaluation metrics will uncover these sorts of phenomena.

Occasionally, we will consider the median outcome, in cases where the mean is highly unrepresentative of a typical outcome. There were one or two example datasets where the ratio metrics took very large, unrepresentative values, usually because the standard deviation of the misspecified posterior exploded relative to the standard deviation of the true posterior. In such cases the medians or even the 0.9 quantiles are more representative of commonly observed values for these metrics. This issue only arose for the lognormal model (i.e. when a Box-Cox transformation of $\lambda = 0$ was required).

**2.5.5 Motivating example**

It is helpful to look at an example to motivate the various different outcomes that are listed above.

Figure 2.1 plots the true posterior and the misspecified posterior for an example data set of 100 observations. The true likelihood is defined by a $t_7$ distribution, where the mean is additive in three covariates and with an additive treatment effect. The true prior is the same
true prior as described above. The misspecified model uses a normal likelihood, rather than the correct $t_7$, and additionally uses a flat prior. Complete randomization was used as the assignment mechanism. No transformation of the outcome was necessary. The misspecified model includes all of the relevant covariates in the model.

We can see that the posterior distribution of the causal effect is quite different, depending on whether one uses the correct model or the misspecified model. However, the posterior coverage of the misspecified credible interval is 0.95, identical to the nominal coverage. Of course, the posteriors are not the same, as reflected by the fact that $D(p \| q) = 0.73$. Clearly, the posterior under the misspecified model is also poorly centered. The posterior mean is 4.3 true standard deviations away from the true posterior mean (i.e. the “standardized bias” is 4.3). The reason that this misspecified model still achieves the nominal coverage is because its credible interval is so wide. The credible interval ratio is 1.7, indicating that the interval from the misspecified model is 70% wider than the correctly specified interval. Both of the posterior intervals only cover positive values and therefore there is no “false negative” or “false positive” in this example.
2.6 Results and Discussion

There are $5 \times 3 \times 3 \times 3 \times 3 \times 3 \times 2 \times 2 = 4860$ experimental settings. We use ANOVA to analyze which factors account for the most variation in the mean or quantiles of each evaluation metric. We can therefore identify which factors lead to the greatest dissimilarity between the misspecified posterior distribution of the estimand and the true posterior distribution of the estimand, conditional on the factor levels that are considered. Once we have identified this subset of factors, we can focus analysis more deeply on these factors and their interactions.

We will see that the ANOVA tables indicate that, of the factors and levels considered, a failure to transform the outcome and the choice of prior are by far the most consequential forms of model misspecification with respect to every evaluation metric.

We first consider the mean and quantiles of $PC(Y^{obs}, X, W)$.

2.6.1 ANOVA for Coverage of the Misspecified Credible Interval Under the True Posterior

The ANOVA table for the mean of $PC(Y^{obs}, X, W)$ for the full factorial experiment is presented in Table 2.4. We use the ANOVA table to identify which factors are most likely to result in poorly constructed posterior intervals. We truncate the table to only display the top 10 factor combinations, ranked by mean squared error.

For some experimental settings, the analysis model is correctly specified. As a sanity check, we can calculate the mean $PC(Y^{obs}, X, W)$ for these conditions. A typical mean $PC(Y^{obs}, X, W)$ is 0.950 and the 0.10 quantile is 0.946, indicating that there is some very small amount of Monte Carlo error. In the absence of Monte Carlo error, this quantity would always be 0.950.

By far, the factors in this experiment that account for the most variance of the mean of $PC(Y^{obs}, X, W)$ are the prior and whether the outcome was properly transformed. Beyond that, the assignment mechanism and whether the likelihood omits important covariates
(and how many it omits) are distant runners-up. The ANOVA tables for the quantiles of $PC(Y^{obs}, X, W)$ are very similar in the sense that the analysis prior and missing transformations are the dominant factors. ANOVA tables for the quantiles are included in Appendix A for reference.

Table 2.4: ANOVA for the mean of $PC(Y^{obs}, X, W)$: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>332.3</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>50.6</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation</td>
<td>4</td>
<td>45.0</td>
</tr>
<tr>
<td>Assignment Mechanism</td>
<td>2</td>
<td>0.6</td>
</tr>
<tr>
<td>Analysis Prior * Assignment Mechanism</td>
<td>2</td>
<td>0.5</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>0.4</td>
</tr>
<tr>
<td>Missing Covariates</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>Analysis Prior * P</td>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>Analysis Likelihood * Missing Transformation</td>
<td>8</td>
<td>0.2</td>
</tr>
<tr>
<td>Analysis Likelihood</td>
<td>2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

We can look at a table that computes the average coverage for each combination of prior and required transformation. This table helps provide additional insight into the influence of these factors.

Table 2.5: 2 x 5 table for the mean of $PC(Y^{obs}, X, W)$

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^{-\frac{1}{3}}}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>log(y)</th>
<th>$y$</th>
<th>$\frac{y^{\frac{1}{3}}-1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>0.917</td>
<td>0.738</td>
<td>0.612</td>
<td>0.966</td>
<td>0.917</td>
<td>0.830</td>
</tr>
<tr>
<td>&quot;True&quot; Prior</td>
<td>0.000</td>
<td>0.102</td>
<td>0.460</td>
<td>0.973</td>
<td>0.000</td>
<td>0.307</td>
</tr>
<tr>
<td>Total</td>
<td>0.458</td>
<td>0.420</td>
<td>0.536</td>
<td>0.969</td>
<td>0.459</td>
<td>0.568</td>
</tr>
</tbody>
</table>

In the 2 x 5 table we can see that undercoverage results from both the choice of prior and omission of a necessary transformation of the outcome. When no transformation is required, coverage is close to the nominal coverage for both the true prior and the flat prior. There is some overcoverage on average for these cases. This overcoverage occurs when the intervals are correctly centered (which is more likely to occur with the true prior or a covariate balancing assignment mechanism) but the interval is wider than it would be under a correctly specified model (e.g. because precision enhancing covariates are missing from the model).
Interestingly, the posterior coverage is often much worse when the “true” prior is used, rather than a flat prior. Using the “true” prior in the correct model centers the causal effect in the appropriate place. Using the “true” prior when the likelihood is misspecified can shift the center of the misspecified posterior distribution to incorrect locations on the real line, as we will see in some examples below.

To get a sense of which factors result in undercoverage and which result in overcoverage (as well as the magnitude of these effects), we can look at the regression coefficients for a regression of the mean outcome on the experimental factors (see Table 2.6). The regression includes all of the main effects as well as the interactions between the analysis prior and the required transformation of the outcome.

Table 2.6: Regression Coefficients for Main Effects

<table>
<thead>
<tr>
<th>Factor</th>
<th>Level</th>
<th>ReferenceLevel</th>
<th>Estimate</th>
<th>SE</th>
<th>t-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>NA</td>
<td>NA</td>
<td>0.974</td>
<td>0.004</td>
<td>263</td>
</tr>
<tr>
<td>N</td>
<td>400</td>
<td>100</td>
<td>-0.001</td>
<td>0.002</td>
<td>-0.6</td>
</tr>
<tr>
<td>N</td>
<td>1000</td>
<td>100</td>
<td>-0.003</td>
<td>0.002</td>
<td>-1.9</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>1</td>
<td>-0.025</td>
<td>0.002</td>
<td>-12.5</td>
</tr>
<tr>
<td>P</td>
<td>3</td>
<td>1</td>
<td>-0.025</td>
<td>0.002</td>
<td>-12.8</td>
</tr>
<tr>
<td>True Likelihood</td>
<td>t15</td>
<td>Normal</td>
<td>0.007</td>
<td>0.002</td>
<td>3.8</td>
</tr>
<tr>
<td>True Likelihood</td>
<td>t7</td>
<td>Normal</td>
<td>0.020</td>
<td>0.002</td>
<td>10.1</td>
</tr>
<tr>
<td>Analysis Likelihood</td>
<td>t15</td>
<td>Normal</td>
<td>0.001</td>
<td>0.002</td>
<td>0.3</td>
</tr>
<tr>
<td>Analysis Likelihood</td>
<td>t7</td>
<td>Normal</td>
<td>0.001</td>
<td>0.002</td>
<td>0.4</td>
</tr>
<tr>
<td>Missing Covariates</td>
<td>Missing All</td>
<td>Missing None</td>
<td>0.014</td>
<td>0.002</td>
<td>8.5</td>
</tr>
<tr>
<td>Analysis Prior</td>
<td>flat prior</td>
<td>correct prior</td>
<td>-0.006</td>
<td>0.004</td>
<td>-1.7</td>
</tr>
<tr>
<td>Transformation</td>
<td>-3</td>
<td>1</td>
<td>-0.973</td>
<td>0.004</td>
<td>-267</td>
</tr>
<tr>
<td>Transformation</td>
<td>-1</td>
<td>1</td>
<td>-0.870</td>
<td>0.004</td>
<td>-239</td>
</tr>
<tr>
<td>Transformation</td>
<td>0</td>
<td>1</td>
<td>-0.513</td>
<td>0.004</td>
<td>-141</td>
</tr>
<tr>
<td>Transformation</td>
<td>3</td>
<td>1</td>
<td>-0.972</td>
<td>0.004</td>
<td>-267</td>
</tr>
<tr>
<td>Assignment Mechanism</td>
<td>FSM</td>
<td>CR</td>
<td>-0.030</td>
<td>0.002</td>
<td>-15</td>
</tr>
<tr>
<td>Assignment Mechanism</td>
<td>RR</td>
<td>CR</td>
<td>-0.003</td>
<td>0.002</td>
<td>1.6</td>
</tr>
<tr>
<td>Analysis Prior * Transformation</td>
<td>flat * -3</td>
<td>correct * 1</td>
<td>0.923</td>
<td>0.005</td>
<td>179</td>
</tr>
<tr>
<td>Analysis Prior * Transformation</td>
<td>flat * -1</td>
<td>correct * 1</td>
<td>0.641</td>
<td>0.005</td>
<td>124</td>
</tr>
<tr>
<td>Analysis Prior * Transformation</td>
<td>flat * 0</td>
<td>correct * 1</td>
<td>0.159</td>
<td>0.005</td>
<td>31</td>
</tr>
<tr>
<td>Analysis Prior * Transformation</td>
<td>flat * 3</td>
<td>correct * 1</td>
<td>0.923</td>
<td>0.005</td>
<td>179</td>
</tr>
</tbody>
</table>

The regression table confirms how important the prior and the required transformation are, as well as their interactions. As for the factors with lesser effects, while they are still interesting, they tend to have a much smaller impact.
2.6.2 ANOVA for the KL Divergence

The ANOVA table for the mean of the KL divergence is presented below. We truncate the table to only display the top 10 factor combinations, ranked by mean squared error.

Table 2.7: ANOVA for the mean of the KL divergence: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Prior</td>
<td>4</td>
<td>693881</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation</td>
<td>4</td>
<td>115195</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>2</td>
<td>97422</td>
</tr>
<tr>
<td>N * Analysis Prior</td>
<td>2</td>
<td>2108</td>
</tr>
<tr>
<td>N</td>
<td>1</td>
<td>1769</td>
</tr>
<tr>
<td>Analysis Prior * Missing Covariates</td>
<td>2</td>
<td>1355</td>
</tr>
<tr>
<td>Missing Covariates</td>
<td>2</td>
<td>1271</td>
</tr>
<tr>
<td>Missing Covariates * N</td>
<td>2</td>
<td>1237</td>
</tr>
<tr>
<td>Missing Covariates * Analysis Prior * N</td>
<td>1</td>
<td>1111</td>
</tr>
<tr>
<td>N * Missing Transformation</td>
<td>2</td>
<td>761</td>
</tr>
<tr>
<td>Analysis Prior * P</td>
<td>2</td>
<td>408</td>
</tr>
</tbody>
</table>

Again the prior and whether the outcome was properly transformed dominate the other factors that were considered in the experiment. The sample size, N, also affects the KL divergence, although to a much lesser extent than the dominant factors. Together, these three factors account for seven of the top ten effects. The ANOVA tables for the quantiles of the KL divergence are qualitatively very similar and are included in Appendix A.

We can again look at a 2 x 5 table where we tabulate the KL divergence for each combination of the prior versus the required transformation of the outcome. Again, within each combination of prior and transformation, we average over all of the other simulation factors.

Table 2.8: 2 x 5 table for the mean of the KL Divergence

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>( \frac{1-y^{-3}}{3} )</th>
<th>( 1 - \frac{1}{y} )</th>
<th>log(y)</th>
<th>( y )</th>
<th>( \frac{y^{p-1}}{3} )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>0.8</td>
<td>7.9</td>
<td>10.1</td>
<td>0.7</td>
<td>1.2</td>
<td>4.1</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>46.5</td>
<td>35.4</td>
<td>13.5</td>
<td>0.1</td>
<td>44.8</td>
<td>28.0</td>
</tr>
<tr>
<td>Total</td>
<td>23.7</td>
<td>21.6</td>
<td>11.8</td>
<td>0.4</td>
<td>23.0</td>
<td>16.1</td>
</tr>
</tbody>
</table>

The KL divergence is especially large when the “true” prior is used but the misspecified model incorrectly fails to transform the outcome. This often occurs because the use of the
informative prior mistakenly centers the posterior distribution of the estimand in the wrong location, as we will see below.

### 2.6.3 ANOVA for the Normalized Difference in Posterior Means

One question we may ask is whether undercoverage or large KL divergence results from a poor centering of the misspecified posterior credible interval. The metric that we use to diagnose poor centering is the difference between the misspecified posterior mean and the correctly specified posterior mean, normalized by the true posterior standard deviation.

Similar to the ANOVA tables for the other outcomes, the ANOVA tables for the normalized difference in posterior means suggest that the failure to transform the outcome and the choice of prior are the largest sources of miscentering of the misspecified posterior.

Table 2.9: ANOVA for the mean of the log absolute standardized bias: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>6759</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior</td>
<td>4</td>
<td>2897</td>
</tr>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>1529</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>57</td>
</tr>
<tr>
<td>Missing Covariates</td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>Analysis Prior * N</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>Missing Transformation * Missing Covariates</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior * Missing Covariates</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Assignment Mechanism</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Analysis Prior * Assignment Mechanism</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>

We can look at tables that calculate the mean standardized bias for all of the different combinations of levels of the prior factor versus levels of the missing transformation. When there is no missed transformation of the outcome, the true and misspecified posterior tend to have the same mean in these experimental settings.

When a transformation of the outcome is missing, the use of the “true” prior often places the center of the distribution in the incorrect place, which results in undercoverage of the misspecified posterior interval under the true posterior and large KL divergence between
the misspecified posterior distribution and the true posterior distribution. This problem is most pronounced for the simulation settings where Box-Cox transformations with \( \lambda = -3 \) or \( \lambda = 3 \) are required. By contrast, when a flat prior is used, the true posterior and misspecified posterior tend to have the same posterior mean on average, averaged over many datasets. We can see this by looking at the 2 x 5 table that displays the median standardized posterior bias for the prior versus the missed transformation. As a side note, the mean tables show large mean bias for the data generated via the lognormal model. This is due to a single simulated data set with large posterior standardized bias.

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>( \frac{1-y^{-\frac{2}{3}}}{3} )</th>
<th>( 1 - \frac{1}{y} )</th>
<th>( \log(y) )</th>
<th>( \frac{y}{3} )</th>
<th>( \frac{y^{1/3}}{3} )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>-0.02</td>
<td>0.00</td>
<td>-0.05</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.01</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>117</td>
<td>0.09</td>
<td>-0.07</td>
<td>0.00</td>
<td>28</td>
<td>29</td>
</tr>
<tr>
<td>Total</td>
<td>58</td>
<td>0.05</td>
<td>-0.06</td>
<td>0.00</td>
<td>14</td>
<td>14</td>
</tr>
</tbody>
</table>

It is not surprising that the use of a flat prior often leads to well-centered posterior intervals on average. To gather intuition for why this is so, we return briefly to non-model based causal inference techniques. If we view the potential outcomes as fixed, it is easy to see that the estimator \( \hat{\tau}_{\text{diff}} = \frac{1}{N_c} \sum_{i: W_i=1} Y_i^{\text{obs}} - \frac{1}{N_c} \sum_{i: W_i=0} Y_i^{\text{obs}} \) is unbiased for the finite population causal effect (where the estimand is the difference in means of the potential outcomes). In showing that this estimator is unbiased, the expectation is taken over the randomization distribution; no distributional assumptions about the potential outcomes (which, again, are considered fixed) are required. This unbiasedness is simply a product of the use of randomization.

Returning to the Bayesian model with a flat prior, the posterior difference in potential outcome means will be similar to the simple difference in means estimator. Rather than simply computing the difference in observed means, we impute all of the missing outcomes and then compute the difference in means of the fully imputed potential outcomes. When we use a flat prior, all of the information in this imputation process comes directly from the data. Without much influence from the prior, the missing potential outcomes will tend to
have similar means to the observed potential outcomes, for the choices of likelihood that we
consider in the experiment. In repeated sampling, therefore, we should expect the posterior
mean to be close to the finite population causal effect. The price that we pay for this model
misspecification will lie in the estimation of the variance and higher moments, such as the
skew.

2.6.4 ANOVA for the Ratio of the Widths of the CI intervals

Looking at the ANOVA tables for the quantiles of the ratio of the widths of the CI intervals,
the failure to correctly transform the outcome and the choice of prior by far account for most
of the variance in this outcome.

Table 2.11: ANOVA for the 0.9 quantile of the CI width ratio: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Transformation * Analysis Prior</td>
<td>4</td>
<td>5088</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>4662</td>
</tr>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>2096</td>
</tr>
<tr>
<td>True Likelihood</td>
<td>2</td>
<td>1375</td>
</tr>
<tr>
<td>Missing Transformation * True Likelihood</td>
<td>8</td>
<td>1370</td>
</tr>
<tr>
<td>Missing Transformation * P</td>
<td>8</td>
<td>854</td>
</tr>
<tr>
<td>Missing Transformation * N</td>
<td>8</td>
<td>785</td>
</tr>
<tr>
<td>Missing Transformation * True Likelihood * P</td>
<td>16</td>
<td>617</td>
</tr>
<tr>
<td>True Likelihood * P</td>
<td>4</td>
<td>614</td>
</tr>
<tr>
<td>Missing Transformation * True Likelihood * N</td>
<td>16</td>
<td>607</td>
</tr>
</tbody>
</table>

The failure to use an appropriate prior or the failure to transform the outcome variable
appropriately can lead to misspecified posterior distributions whose 95% credible intervals
are dramatically too large or too small, when compared to the size of the true posterior
interval.

For example, when a log transformation of the outcome is required and the analysis fails
to transform the outcome, the median posterior interval constructed under the misspecified
model is 10 times smaller than the true posterior interval. Similarly, when the required trans-
formation is of the form \( f(y) = 1 - \frac{1}{y} \), the misspecified interval is 5 times smaller than the
true posterior interval. When these intervals are too small, as they often are, they often result in severe undercoverage, as we will see in some examples. Often, the dramatic difference in the widths of the intervals suggests very different shapes of the posterior distributions, as is reflected in the large KL divergences that accompany these scenarios.

Table 2.12: 2 x 5 table for the median of the ratio of CI Widths

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^{-3}}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>log$(y)$</th>
<th>$y$</th>
<th>$\frac{y^2-1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>1.8</td>
<td>0.2</td>
<td>0.1</td>
<td>2.2</td>
<td>2.2</td>
<td>1.3</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>11</td>
<td>0.2</td>
<td>0.1</td>
<td>1.3</td>
<td>2.8</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>6.2</td>
<td>0.2</td>
<td>0.1</td>
<td>1.8</td>
<td>2.5</td>
<td>2.2</td>
</tr>
</tbody>
</table>

Intervals may also sometimes be too large, which can result in significant over coverage, coverage that is close to 1. For the most part, the narrowness of the intervals is a problem for the simulation settings where Box-Cox transformations with $\lambda = -1$ or $\lambda = 0$ are required.

Table 2.13: 2 x 5 table for the 0.9 quantile of the ratio of CI Widths

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^{-3}}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>log$(y)$</th>
<th>$y$</th>
<th>$\frac{y^2-1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>2.1</td>
<td>1.2</td>
<td>4.5</td>
<td>2.4</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>11</td>
<td>0.7</td>
<td>2.4</td>
<td>1.4</td>
<td>3.2</td>
<td>3.8</td>
</tr>
<tr>
<td>Total</td>
<td>6.7</td>
<td>1.0</td>
<td>3.4</td>
<td>1.9</td>
<td>2.8</td>
<td>3.2</td>
</tr>
</tbody>
</table>

Table 2.14: 2 x 5 table for the 0.1 quantile of the ratio of CI Widths

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^{-3}}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>log$(y)$</th>
<th>$y$</th>
<th>$\frac{y^2-1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>1.6</td>
<td>0.09</td>
<td>0.09</td>
<td>2.0</td>
<td>2.0</td>
<td>1.1</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>10</td>
<td>0.18</td>
<td>0.01</td>
<td>1.2</td>
<td>2.5</td>
<td>2.8</td>
</tr>
<tr>
<td>Total</td>
<td>5.9</td>
<td>0.13</td>
<td>0.01</td>
<td>1.6</td>
<td>2.2</td>
<td>2.0</td>
</tr>
</tbody>
</table>

2.6.5 ANOVA Results for other outcomes

We have generated ANOVA tables for the mean, median, 0.9 quantile and 0.1 quantile of all of the various evaluation metrics that we discussed above. These tables are available in Appendix A, for the most part (some of the ANOVA tables for some quantiles are omitted, if they add little value). We move most of these tables to Appendix A because they all lead to
the same basic conclusion: the choice of the prior and the failure to transform the outcome appropriately are by far the most consequential forms of model misspecification according to every outcome. We will spend some time discussing some of the other experimental factors (such as the omission of covariates and the choice of assignment mechanism) but, due to their importance, we will tend to focus on the choice of prior and the failure the appropriately transform the outcome.

2.6.6 Discussion of Missed Transformations and the True Prior

Broadly speaking, we have seen that missing a required transformation of the outcome and the choice of prior are the most important factors for all of the outcomes, including the posterior coverage of the misspecified interval and the KL divergence.

Poor coverage and poor KL divergence can be caused by poor centering of the misspecified posterior distribution, a phenomenon that can easily occur with a suboptimal choice of prior (even when using the data generating prior). Such “bias” in the misspecified posterior is close to 0 for the median dataset in these simulation settings when a flat prior is used, even when the hypothetical analyst fails to properly transform the outcome.

We have also seen that poor coverage can result from misspecified posterior distributions that yield credible intervals that are too narrow. These intervals fail to capture skew in the true posterior or otherwise fail to encapsulate the magnitude of uncertainty under the true posterior. This problem may occur for either choice of prior that we consider in this experiment.

In the subsections that follow, we plot both the true posterior distribution of the causal estimand and the misspecified posterior distribution of the causal estimand for example data sets where each possible missed transformation of the outcome and choice of analysis prior is plotted separately. The examples underscore, and provide intuition for, the results above.
No transformation required

Figure 2.2 plots the correct posterior and the misspecified posterior for an example data set of 100 observations, where no transformation is required. A normal likelihood is used in both the true model and in the analysis model. The true data generating model is a function of three covariates and these covariates are included in the misspecified model. The only form of misspecification is that a flat prior is used instead of the correct prior.

![True and Misspecified Posterior Densities](image)

Figure 2.2: \( \lambda = 1 \), flat prior

As one can see, the posterior interval in the misspecified case is much wider than the true posterior interval (the ratio of the credible interval widths is 1.8). Indeed, the coverage of this interval under the true posterior is 0.999. This is expected since a flat prior represents greater prior uncertainty than the true prior and therefore this uncertainty should propagate into the posterior in the form of wider intervals. These observations are consistent with the results of the simulation study, which are averaged over 250 synthetic datasets for each combination of factors. Note that there is some small difference in the posterior means (the misspecified posterior mean is about -0.6 true standard deviations away from the true posterior mean), due to the failure to use the correct prior, which would tend to center the causal effect in
the correct location.

When no transformation is required, the intervals constructed with the misspecified models considered in this experiment tend to achieve or exceed the nominal coverage. However, these intervals often satisfy the nominal coverage in an inefficient, conservative way. To the extent that overly conservative results are undesirable, these forms of misspecification can be problematic.

**Required Transformation:** \( g(y) = \log(y) \)

The following plots are examples of cases where a log transformation was required but omitted. We consider the case where the misspecified model uses both the “correct” prior and a flat prior. In Figure 2.3, we plot both posterior distributions based on an example simulated data set of 100 observations.

![True and Misspecified Posterior Densities](image)

Figure 2.3: \( \lambda = 0 \), true prior (truncated)

The true model is normal and linear in the three covariates on the transformed outcome. The misspecified model is also normal and linear in the three covariates, but on the untransformed outcome. The misspecified model in Figure 2.3 uses the “true” prior, as does Figure
The true posterior is incredibly skewed, with a very long right tail. The kurtosis of the correct posterior is $2 \times 10^4$ larger than it is for the misspecified posterior, and it has $8 \times 10^4$ more skew. The posterior coverage of the misspecified interval under the correct posterior is only 0.3. The credible interval width is 78 times larger for the true posterior than for the misspecified posterior. Indeed, Figure 2.3 is truncated for the purpose of easing visual comparison. When plotting on a scale that includes the entire credible interval for the true posterior, we obtain Figure 2.4.

![True and Misspecified Posterior Densities](image)

Figure 2.4: $\lambda = 0$, true prior (untruncated)

On this latter scale, the misspecified posterior almost looks like a point mass in comparison to the long tailed true posterior. The plot underscores just how different these distributions are, especially with respect to the posterior variance and skew of the estimand.

We can also look at misspecified posteriors where a flat prior is used. Figure 2.5 plots the true posterior and the misspecified posterior for another example dataset. The true model is the same as the true model in the previous plot. The only difference between the misspecified model in Figure 2.3 and the misspecified model in Figure 2.5 is the substitution
of a flat prior in place of the “correct” prior. In this case, the coverage of the misspecified interval is somewhat better (0.54), although it is still pretty poor. The true and misspecified means have the same sign, at least, when the flat prior is used (this was not the case with the “true” prior). However, the misspecified posterior interval is still much too narrow; the true posterior interval is 44 times wider.

Figure 2.5: \( \lambda = 0 \), flat prior (truncated)

The x-axis scale of Figure 2.5 is truncated so that the modes of the two distributions can be easily displayed. Figure 2.6 widens the x-axis so that the entire credible interval for the correct model can be plotted. As can be seen, the true posterior is very heavy tailed and the misspecified distribution almost resembles a point mass in comparison.
Figure 2.6: $\lambda = 0$, true prior (untruncated)

**Required Transformation:** $g(y) = 1 - \frac{1}{y}$

The example plots in this subsection involve misspecified models where a Box-Cox transformation with $\lambda = -1$ was required but no such transformation of the outcome was used in analysis. Similar to the other example plots that are presented, the example data sets contain 100 observations and were generated with a normal likelihood, where the mean of the distribution depends on three covariates and the treatment indicator. Complete randomization is used in the simulated experiment. The data generating prior is the same as described above. The only difference between these examples and the examples where a log transformation was required is the nature of the required transformation of the outcome variable. The misspecified model is only misspecified insofar as the outcome was not properly transformed and, possibly, the prior is incorrect.

Figure 2.7 plots the correct posterior and the misspecified posterior, where the misspecified posterior uses the “correct” data generating prior. By contrast, the misspecified prior in Figure 2.8 uses a flat prior.

The misspecified posterior with the informative prior performs especially poorly. The
coverage is close to 0 (0.03). Even though the mean of the true posterior is close to 0, the 95% credible interval under the misspecified model does not cover 0.

Figure 2.7: $\lambda = -1$, true prior

The misspecified posterior with the flat prior, by contrast, is centered much closer to the true posterior mean and, for that reason, has much better coverage (0.52). In this case, it is
much safer to use the flat prior, although serious problems remain even with the use of a flat prior. The posterior intervals, as in the cases where a log transformation were required, are simply too narrow under the misspecified model and result in undercoverage for this reason, even when the posterior means of the two distributions are similar.

**Required Transformation:** \( g(y) = \frac{1-y^{-3}}{3} \)

Here we present example plots where the true model required a Box-Cox transformation with \( \lambda = -3 \) transformation of the outcome variable. Aside from the required transformation, the true data generating process is identical to the above examples. The misspecified model, as in the other examples, is correct except for the omission of the required transformation of the outcome and, potentially, the prior. Again, we consider the case where the misspecified model uses the “true” prior (Figure 2.9) and the case where the misspecified model uses a flat prior (Figure 2.10)

![True and Misspecified Posterior Densities](image)

**Figure 2.9:** \( \lambda = -3 \), true prior

Similar to what we have observed in the last section, the use of the true prior with the misspecified likelihood centers the posterior in the completely wrong place. The width
of the misspecified posterior credible interval is much greater than the width of the true posterior credible interval. Even though the posterior credible interval is quite wide under the misspecified model, the coverage of the interval is 0 under the true model, due to the poor centering of the posterior that results from the use of the informative prior.

Even worse, we reach highly misleading inferences with the misspecified model with true prior. The posterior mean under the true posterior is negative and the 95% credible interval does not cover 0. By contrast, the posterior mean under the misspecified model is positive, and the misspecified credible interval does not cover 0. This suggests that the true causal effect is negative with high probability but that this form of model misspecification might lead an analyst to conclude the exact opposite, that the causal effect is positive with high probability.

![True and Misspecified Posterior Densities](image)

Figure 2.10: $\lambda = -3$, flat prior

The misspecified model with the flat prior performs better. In this example, the posterior means are relatively close to each other (the posterior mean of misspecified model is 1.3 true standard deviations away from the true posterior mean). However, the posterior interval under the misspecified model is too wide, about 1.8 times wider than the true posterior
interval. This results in posterior overcoverage of the misspecified interval. The misspecified credible interval covers zero, whereas the true credible interval does not cover zero. To the extent that applied statisticians make decisions based on whether their interval does or does not cover zero, a researcher with the true model might make a different decision from the decision made by a researcher who uses the misspecified model.

**Required Transformation:** \( g(y) = \frac{y^3 - 1}{3} \)

Examples where the true model requires a \( \lambda = 3 \) Box-Cox transformation of the outcome variable are similar in some ways to the case where a \( \lambda = -3 \) transformation is required. Again, the simulation conditions are identical to the previous example except for the required transformation of the outcome variable. Figure 2.11 plots the misspecified model that uses the “true” prior and Figure 2.12 plots the misspecified model that uses the flat prior.

![True and Misspecified Posterior Densities](image)

**Figure 2.11:** \( \lambda = 3 \), true prior

In the example with the “true” prior, the location of the posterior is again in the wrong place and the posterior coverage of the misspecified interval is again 0. Additionally, the credible interval is much wider than the true posterior interval. When using the “true”
prior, the misspecified posterior interval is confidently in the wrong place. This is in keeping with the simulation results, where the average coverage was zero for this type of model misspecification.

![True and Misspecified Posterior Densities](image)

Figure 2.12: $\lambda = 3$, flat prior

The misspecified model that uses a flat prior performs better; the misspecified and correct posterior distributions are more or less centered in similar places. However, the misspecified credible interval is still much wider than what is obtained with the correctly specified prior. The posterior coverage of the misspecified interval is 1 in this example.

**Subtle Transformations Are Also Problematic**

We may try to brush away these concerns by hoping that a competent data analyst will always notice that the outcome variable needs to be transformed and will succeed in transforming it appropriately. Perhaps the required transformations in these examples were somehow “obvious” such that we needn’t worry in practice about these forms of misspecification. However, similar inferential problems can arise even in cases where the required transformation is quite subtle and, we would argue, easy to omit.
In the following examples, we consider misspecified models where a transformation of the form \( g(y) = y^{(19/21)} \) was required. This transformation was chosen such that, on its face, it appears to be quite mild and such that the function is invertible on the real line. Beyond these criteria, the exact value of the exponent is arbitrary and similar results can be obtained with different values. The simulation conditions, aside from the required transformation, are otherwise identical to the preceding examples.

![True and Misspecified Posterior Densities](image)

**Figure 2.13:** Subtle transformation, true prior

Figure 2.13 shows the true posterior and a misspecified posterior that uses the “true” prior. Even though the required transformation is mild (so mild, in fact, that one might hope it isn’t truly required at all to get approximately correct inferences), the posteriors are quite different. The posterior coverage of the misspecified credible interval is 0.5 and the posterior mean for the misspecified model lies outside of the true posterior interval.

Figure 2.14 shows the true posterior and a misspecified posterior that uses a prior for the beta coefficients in the regression model that is centered at 0. Choosing a prior that is centered at zero is sometimes used in applied practice as a “conservative choice” in the face of scant prior knowledge. Common procedures such as ridge regression can be interpreted
as providing an estimator based on the posterior mode for a normal-normal model with a prior centered on 0. In this particular example, a zero-centered prior performs very poorly, with a posterior coverage of 0.

Note that the full factorial experiment does not consider priors that are centered at 0, nor does it consider this kind of mild transformation. Future studies could include such simulation conditions in a more systematic way. We only include them here to highlight examples where even very subtle omitted transformations of the response and priors that are sometimes considered “safe”, “conservative”, or otherwise desirable as default priors (such as zero-centered prior on the mean parameters), can lead to very misleading inferences.

Lastly, Figure 2.15 plots the posterior for the causal estimand for a misspecified model that uses a flat prior. In this particular example, the posterior interval is relatively well centered although, again, the posterior interval is much wider than the true posterior interval.
2.6.7 Discussion of Other Factors

The other experimental factors are relatively unimportant in comparison to the influence of the prior and the influence of a failure to transform the outcome of interest. However, we take the opportunity here to remark on some of the other factors. In particular, we discuss the effects of the sample size, the omission of covariates, and of the assignment mechanism.

We highlight these three factors for a variety of reasons. Aside from the prior and omitted transformations, the assignment mechanism is the greatest source of variation for the posterior coverage of the misspecified interval. Similarly, the sample size and missing covariates are the greatest sources of variability for the KL divergence, aside from the prior and the omitted transformation.

More conceptually, we might hope that problems associated with model misspecification fade as the sample size increases, if asymptotic results begin to “kick in”. If we omit covariates, we might hope that such a simplified model (which is also expected to have higher posterior variance) will result in posterior intervals with less undercoverage, especially when combined with the flat priors, which tend to result in little posterior bias. Lastly, prior work
has shown that an “ignorable” missing data mechanism is only guaranteed to be truly ignorable when the model is correctly specified, and thus this phenomenon is worth revisiting here via discussion of the assignment mechanism.

Large N Doesn’t Always Help

First, we consider an example where N is truly large. We see that, even in this example, the posterior distribution of the causal effect can be quite misleading when we fail to transform the outcome. The posterior means are relatively similar, but the posterior intervals are vastly different.

Figure 2.16 plots the correct posterior and the misspecified posterior for an example data set of 100,000 observations. A log transformation of the outcome is performed to obtain the true posterior distribution. The misspecified model fails to transform the outcome variable. It was not feasible to include simulated datasets with such a large number of observations in the actual simulation study, due to computational limitations, even with the fast, parallel computing resources that we had available (each simulated data set required two hours of computing time). However, even looking at a single example can provide insight.

![True and Misspecified Posterior Densities](image)

Figure 2.16: λ = 0, flat prior, large N
The true, data generating likelihood follows a normal distribution (on the transformed outcome) and a normal likelihood was used in the misspecified model (on the untransformed outcome). Both the true and misspecified models include three covariates. The assignment mechanism is complete randomization. The only sources of misspecification are (a) the failure to log transform the outcome and (b) the use of a flat prior.

The use of a flat prior is meant to be a conservative choice, since we have seen that the use of a “correct” prior, even with moderately misspecified likelihood functions, can shift the center of the posterior in undesirable ways. Additionally, the use of a flat prior reflects great prior uncertainty and thus should err on the side of producing conservatively wide posterior intervals.

A sample size of 100,000 is far beyond the constraints of many social science and medical experiments; it is unfeasibly large for many contexts. Even with such a large sample size, the inferences that we obtain from these two posterior distributions are quite different. The incredibly long tail on the true posterior distribution implies a very wide posterior credible interval. The posterior means are almost identical from the two models (the standardized “bias” is basically 0), but the posterior intervals are radically different. The ratio of the width of the misspecified posterior interval to the width of the correctly specified posterior interval is 0.11, indicating that the true credible interval is 9 times larger than the credible interval constructed with the misspecified model. In other words, the misspecified model appears to be quite overconfident and its posterior interval only has a coverage of 0.72 under the true model, even with such a large sample size. To the extent that a decision maker would make a determination based on whether the posterior interval covers 0 or not, the wrong decision would be reached in this case, since the true posterior does not cover 0, although the misspecified interval does cover 0.

In short, even in the case where we have a very large sample size, we may have very similar posterior means for the true and misspecified models, but the credible intervals can be very different, to the point of yielding different inferences. This is true even when a
flat prior is used. Indeed, a flat prior may yield an interval that is much too narrow and undercovers (even though we hope that a flat prior might err on the side of overcoverage) if one fails to transform the outcome correctly.

**Missing Covariates**

Next, we consider the cases where covariates are missing from the model. In general, this widens the posterior interval. The failure to include covariates leads to a loss of precision, since these covariates explain some of the variance in the observed outcome. Figures 2.17 and 2.18 are typical examples of how the posterior interval widens for a misspecified model that omits important covariates. Both of the models are based on example data sets of 100 observations where both the true and misspecified models use a normal likelihood. No transformation of the outcome is required in either of these example data sets. The main source of misspecification is that the true data generating process involves three covariates. The misspecified model omits all of these covariates from the misspecified likelihood.

The difference between Figure 2.17 and Figure 2.18 is that, in the former, the misspecified model employs a flat prior whereas in the latter, the misspecified model uses the “true” prior. By “true” prior, we mean that the prior on the variance parameter and the beta parameter for the treatment indicator are correct. The beta parameters for the covariates are, of course, missing from the model, and therefore there is no prior on these parameters.

As expected, the posterior interval with a flat prior and no covariates is much too wide (and is wider than the misspecified interval that uses the “true” prior). Indeed, the misspecified interval is over three times wider. The misspecified interval is also no longer perfectly centered, as the modeling of the covariates and the use of a correct prior both tend to center the interval closer to the center of the true posterior interval. The coverage of the misspecified interval under the true posterior is 1.

Even when using the “true” prior, the resultant posterior interval is still quite a bit wider, as can be seen in Figure 2.18. The use of the correct prior produces a misspecified posterior
Figure 2.17: $\lambda = 1$, flat prior, no covariates

Figure 2.18: $\lambda = 1$, true prior, no covariates
mean that is nearly identical to the true posterior mean. However, the posterior interval is about 75% too large; its coverage under the true posterior is 0.999. In short, the omission of covariates tends often results in “conservative” posterior intervals, in the sense that the intervals are too wide.

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$1 - \frac{y^2}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>log($y$)</th>
<th>$y$</th>
<th>$\frac{y^2 - 1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing All Covariates</td>
<td>1.8</td>
<td>0.2</td>
<td>0.09</td>
<td>2.8</td>
<td>2.5</td>
<td>1.5</td>
</tr>
<tr>
<td>Not Missing Covariates</td>
<td>1.7</td>
<td>0.2</td>
<td>0.09</td>
<td>1.7</td>
<td>1.8</td>
<td>1.1</td>
</tr>
<tr>
<td>Total</td>
<td>1.8</td>
<td>0.2</td>
<td>0.09</td>
<td>2.2</td>
<td>2.2</td>
<td>1.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$1 - \frac{y^2}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>log($y$)</th>
<th>$y$</th>
<th>$\frac{y^2 - 1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing All Covariates</td>
<td>0.92</td>
<td>0.74</td>
<td>0.62</td>
<td>0.98</td>
<td>0.94</td>
<td>0.84</td>
</tr>
<tr>
<td>Not Missing Covariates</td>
<td>0.91</td>
<td>0.74</td>
<td>0.61</td>
<td>0.95</td>
<td>0.89</td>
<td>0.82</td>
</tr>
<tr>
<td>Total</td>
<td>0.92</td>
<td>0.74</td>
<td>0.61</td>
<td>0.97</td>
<td>0.92</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Turning to the tables of coverage and credible interval width ratios, we see that credible intervals are indeed wider when covariates are excluded from the model. However, the posterior coverage of the misspecified interval is always lower when covariates are incorporated into the model (when using a flat prior). Since undercoverage is known to be a problem, it is unclear that use of the covariates in the model, even when those covariates are known to be important, adds value. One may be better off avoiding covariates altogether if we cannot correctly specify the relationship between the covariates and the outcomes.

“Ignorable” Assignment Mechanisms are not Always Ignorable

In likelihood-based analyses such as the model-based Bayesian methods that we consider in this study, ignorable assignment mechanisms need not be modeled when the model is correctly specified. However, the freedom to forget the assignment mechanism may not be warranted when the modeling assumptions are incorrect. Ignorability is defined relative to some correctly specified model. When the model is misspecified, including when an
inappropriate prior is used, posterior inferences may be incorrect. Such mistakes may be more problematic for some ignorable assignment mechanisms than for others. In this sense, it may be unwise to “ignore” the assignment mechanism used in the experimental design.

In this simulation experiment we observe that, when the model is misspecified, the choice of ignorable assignment mechanism can have an impact on the average coverage of the misspecified interval under the true posterior distribution. The effects are relatively small, especially in comparison to some of the other experimental factors that we have considered. In particular, complete randomization and rerandomization yield similar results. The finite selection model, in particular, appears to occasionally yield suboptimal results for some forms of model misspecification (see Table 2.17).

Table 2.17: 3 x 5 table for the mean of the posterior coverage

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>( \frac{1-y^{-1}}{3} )</th>
<th>( 1 - \frac{1}{y} )</th>
<th>( \log(y) )</th>
<th>( y )</th>
<th>( \frac{y^{-1}-1}{3} )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete Randomization</td>
<td>0.95</td>
<td>0.74</td>
<td>0.62</td>
<td>0.96</td>
<td>0.96</td>
<td>0.84</td>
</tr>
<tr>
<td>Rerandomization</td>
<td>0.96</td>
<td>0.74</td>
<td>0.62</td>
<td>0.97</td>
<td>0.98</td>
<td>0.82</td>
</tr>
<tr>
<td>Finite Selection Model</td>
<td>0.84</td>
<td>0.73</td>
<td>0.61</td>
<td>0.97</td>
<td>0.81</td>
<td>0.82</td>
</tr>
<tr>
<td>Total</td>
<td>0.92</td>
<td>0.74</td>
<td>0.61</td>
<td>0.97</td>
<td>0.92</td>
<td>0.83</td>
</tr>
</tbody>
</table>

One can find particular examples where rerandomization and complete randomization produce credible intervals with different average posterior coverages under the true posterior. For example, Figure 2.19 contains two graphs that illustrate the difference. The top graph plots the misspecified (and true) posterior when the assignment mechanism is rerandomization. The bottom graph plots the misspecified (and true) posterior when the assignment mechanism is complete randomization. No transformation of the outcome is required in this example. The sample size of the data set is 400. Two covariates are involved in the true data generating process (although the misspecified model omits both of these covariates). The true likelihood and the analysis likelihood are both \( t_7 \). A flat prior is used for both misspecified models.

The complete data for both examples are identical – they only differ in their observed data, due to the fact that they use different assignment vectors drawn from different assign-
ment mechanisms. In this example, the coverage of the posterior interval is close to 1 for the rerandomization example and close to 0 for the complete randomization examples. This is an extreme example; when averaging the difference in coverage over 250 datasets, the difference in coverage between rerandomization and complete randomization is 0.05. Nevertheless, this is a meaningful difference considering that the difference is ignorable for a correctly specified model. As mentioned above, previous work exploring a similar phenomenon includes Rosenbaum and Rubin’s (1984) investigation of the sensitivity of posterior inferences to model misspecification when data is obtained from data-dependent, yet ignorable, stopping rules.

![Figure 2.19: Example differences between re-randomization and complete randomization](image)

2.7 The Importance of Checking Model Assumptions in the Age of Big Data

Today, data is recorded in greater volume than ever before. Companies and other organizations are increasingly sophisticated in the analysis of their data, including the use of random-
ized experiments. It is no secret, for example, that sophisticated organizations constantly run many simultaneous randomized experiments in order to optimize their operations and products (Tang et al., 2010). In light of these trends, philosophical and practical questions about causal inference from randomized experiments are more important than ever.

As randomized experiments become more prevalent, model misspecification is an increasingly important topic; data analysts may choose to avoid complicated models (Bayesian or frequentist) because they do not believe the model assumptions. Additionally, they may avoid complex modeling because they do not understand the consequences of breaking the assumptions. Rather, especially when large amounts of data are available, such analysts may prefer “nonparametric” estimation methods. This study aims to directly address the consequences of breaking assumptions so that analysts have a better sense of when they actually need to be worried.

Modeling will be unavoidable at times. Randomized experiments can encounter unavoidable analytical difficulties such as non-compliance, truncation due to death, and others. Experiments often “break” in these ways, as we will see below in the Philadelphia Divorce Study. Such difficulties often require constructing a model that leverages the conceptual framework provided by the Rubin Causal Model. Therefore, understanding the sensitivity of such models to choices of likelihood and prior is practically important.

Even when analysts have large amounts of data and even if their experiment exhibits none of the difficulties mentioned above, a model-based approach may be needed for detecting very small effects (for example, by leveraging available covariates). Additionally, even within the context of “Big Data,” organizations may choose to run a large number of relatively small, simultaneous experiments rather than sequentially running one big experiment after another. When data is divided up very finely to accommodate many simultaneous experiments, analysts may once again find themselves dealing with smaller sample sizes. In such scenarios, analysts will be greatly tempted to take a model-based approach to causal inference.
Model-based causal inference from randomized experiments is here to stay. In fact, it is more prevalent than ever. Whether we turn to such models to handle problems like non-compliance or because we desire increased precision in detecting small effects or whether we turn to them for some other reason, it is paramount that we understand the consequences of model misspecification.

2.8 Conclusion

In this full factorial simulation experiment we investigated the influence of several different forms of model misspecification on causal inferences when using model-based Bayesian methods. We considered misspecification of the likelihood, misspecification of the prior and experimental design choices such as the sample size and the assignment mechanism.

We found that, for the factors and factor levels considered, the choice of prior and a failure to properly transform the outcome variable created the greatest differences between a misspecified posterior distribution and the correct posterior distribution. To measure the differences between a true posterior distribution and a misspecified posterior distribution we primarily relied on two measures: (1) the coverage (under the true posterior) of a credible interval that is constructed with a misspecified posterior and (2) the Kullback-Leibler divergence from the misspecified posterior distribution to the true posterior distribution.

We saw that posterior inferences can be greatly misleading when the data generating prior is used along with a failure to adequately transform the outcome variable. Indeed, the use of misspecified posteriors of this sort can lead to credible intervals that are completely non-overlapping with the true posterior credible interval, even on average. By contrast, the use of a flat prior tended to produce posterior distributions with posterior means that were close to the true posterior mean, even when the correct transformation of the outcome was not applied. Nevertheless, when a necessary transformation of the outcome was omitted, the posterior standard deviation and credible interval width could be quite incorrect regardless
of the choice of prior, often with overly narrow or overly wide intervals. We considered a couple of examples where the needed transformation of the outcome was very mild and where the sample size was very large. Some problems that we observed in the experiment still persisted even in these cases.

In summary, model misspecification can be a troubling problem in a causal inference setting, even when $N$ is large and even when the required transformation of the outcome is mild. Perhaps even more troubling, when the likelihood is mildly misspecified, posterior inferences can be quite sensitive to the choice of prior. Even the data generating prior can lead to wildly incorrect results when the likelihood is misspecified. A flat prior often appeared to be a safer choice for the models and forms of model misspecification considered in this study.

Misspecification of the likelihood can also lead to credible intervals that are much too narrow. Often we will make modeling assumptions in order to model the relationship between the covariates and the outcome, so that we can increase precision and, accordingly, reduce the width of our posterior credible interval for the causal estimand. However, when we misspecify the likelihood function, it is unclear whether we can trust these posterior intervals. It may be better to avoid complex modeling altogether, unless it is truly necessary.

Clearly, care must be given to the choice of the likelihood function. It may be most safe to use relatively non-parametric forms of inference when the opportunity is available and reserve likelihood based inference for difficult cases that necessitate complex modeling, such as experiments with non-compliance, dose-response relationships, truncation due to death, and other difficult cases. This simulation study only begins to probe the consequences (using Bayesian evaluation criteria) of model misspecification for Bayesian model-based causal inference and only for a relatively small class of true data generating models. The results of this study are limited by the choices made in the design of the simulation experiment, such as the factors and the levels of the factors. The results, while preliminary, shed light on the (sometimes stark) consequences of model misspecification for causal inference. The subject
clearly merits follow-up investigation.
Chapter 3

Estimating Complier Average Causal Effects in the Presence of Missing Compliance Data: A Randomized Evaluation of Divorce Court Accessibility

3.1 Introduction

Randomized experiments are considered the gold standard for estimating the causal effects of proposed interventions. However, randomized experiments can manifest unexpected challenges that complicate the estimation of such effects. One such challenge is noncompliance with assignment: in some experiments, the actual treatment received by some experimental units is different from the treatment to which the units were randomly assigned. In other words, even though the assignment itself is randomized, the actual receipt of treatment is not fully controlled by the investigator.
There are multiple flavors of noncompliance. One-way noncompliance occurs when some units assigned to the active treatment group actually receive the control treatment, but none of the units in the control group receive the active treatment. Two-way noncompliance occurs when control group units sometimes take active treatment and, additionally, some active treatment group units take the control treatment. Similarly, sometimes units only take a fractional dose, rather than a full dose, of their treatment. For example, participants in a drug trial may stop taking their pills halfway through the experiment if they experience side effects (see, for example Jin and Rubin (2008) and Efron and Feldman (1991)).

When noncompliance is present in an experiment, often an intention-to-treat analysis (ITT) is performed. Such an analysis estimates the causal effect of assignment to treatment (rather than actual receipt of treatment). Rather than measuring the effect of encouragement to take treatment, we often wish to estimate the causal effect of the treatment itself. One estimand of interest is the complier average causal effect (CACE), the average causal effect for units that comply with their treatment assignment. Methodologies have been developed to estimate the complier average causal effect in the presence of both one-way and two-way noncompliance (e.g. Hirano et al. 2000).

These existing methodologies apply to scenarios where both the assignment vector and the vector of treatment indicators are fully observed. In other words, for each unit, we know both what treatment it was assigned to and which treatment it actually received. This study extends these methodologies to the case where the treatment indicators for many of the experimental units are unobserved.

The motivating example for this work is the Harvard Law School Access to Justice Lab Philadelphia Divorce Study ("The Philadelphia Divorce Study"). The Philadelphia Divorce Study was a randomized experiment where potential Philadelphia divorce court litigants were randomized to either receive a free attorney from a local legal aid organization (the active treatment) or, alternatively, a control treatment that did not include a free attorney. One goal of the Philadelphia Divorce Study was to estimate the causal effect of having a free
attorney versus no attorney on various outcomes of interest, such as whether the litigant succeeded in filing for divorce or whether the litigant succeeded in obtaining a divorce.

Two-way noncompliance is observed in the study. Some potential litigants in the treatment group declined the offer of free legal representation. Conversely, some potential litigants in the control group managed to obtain a lawyer on their own. The main challenge of this data set is that, for most people assigned to the control group, whether or not the person received treatment (obtained an attorney) is unobserved. There is therefore substantial missing data in the vector of treatment indicators.

We begin by describing previous work on non-compliance problems. We then describe the legal background to the Philadelphia Divorce Study and the study’s implications.

Next, we discuss the ITT analysis of the Philadelphia Divorce Study data. We show that the ITT effects are quite large and that the results are highly statistically significant. Effects are on the order of 0.28 to 0.44, depending on the estimand.

We then discuss in generality two possible approaches to estimation of the CACE, a maximum likelihood approach and a fully Bayesian approach. We discuss the form that the observed data likelihood and complete data likelihood take and advocate for exploiting the relative simplicity of the complete data likelihood via the EM algorithm and the DA algorithm.

Next, we develop an approach for estimating the CACE in the presence of missing data in the treatment vector. We discuss the simplest model that could be used to estimate the CACE under these circumstances, where we model compliance type and success probabilities without using background covariates. We discuss various possible modeling choices, especially for the modeling of compliance type. Once this model is fully specified, we derive the E-Step and M-Step for the EM algorithm. We then derive Gibbs sampling steps for the DA Algorithm. We discuss the choices of prior that can be made for this Bayesian approach.

We then conduct simulation studies to demonstrate the efficacy and proper implementation of the proposed algorithms for calculating the maximum likelihood estimates and the
full posteriors for the estimands of interest. We demonstrate that these algorithms work as intended for simulated data when the treatment indicators are missing completely at random, even for large amounts of missing data. The MLE on average recovers the true parameter values, as does the posterior mean. Moreover, we show that the posterior credible intervals have the correct coverage.

We then show that, when the missing data mechanism is the exact missing data mechanism that we observe in the Philadelphia Divorce Study, the MLE begins to fail but that the posterior mean of the estimand still recovers the true value, on average, and that the posterior intervals for the estimands still have the proper coverage. We use an example simulated data set to explain why the MLE fails and, conversely, why the Bayesian approach succeeds. We further note that the missing data mechanism that we observe in the Philadelphia Divorce Study causes multimodality in the likelihood and substantial missing data that frustrates both conventional instrumental variables approaches and maximum-likelihood approaches. We explain that Bayesian methods do not suffer from this problem but rather still have appropriate coverage and are reasonably well centered.

Finally, we apply the method to the Philadelphia Divorce Study. We find that the complier average causal effects are quite large and the credible intervals for these effects are all far from 0. We discuss the legal significance of the results. In particular, we argue that the results of the Philadelphia Divorce Study suggest that current divorce court procedures in Philadelphia county are unconstitutional under the Due Process Clause of the Fourteenth Amendment to the United States Constitution.

We conclude by discussing possible future extensions of this work, including an extension to include background covariates, sensitivity to model misspecification, and further simulation studies.
3.2 The Rubin Causal Model and Previous Work

3.2.1 Previous Work

There is important previous work on analyzing randomized experiment with two-sided non-compliance using the Rubin Causal Model framework. For example, Hirano, Imbens, Rubin and Zhou (2000) provide a framework and example for computing the posterior distribution of the CACE using the Data Augmentation algorithm introduced by Tanner and Wong (1987). The notion of potential outcomes is crucial for modeling complex non-compliance behavior, especially in the face of substantial missing data about compliance, as we will see below.

The seminal paper by Hirano et al. built on earlier efforts by Imbens and Angrist (1994) and Angrist et al. (1996) which formulated an RCM interpretation of the instrumental variable (IV) methods that are used in the econometrics literature to estimate causal effects in experiments with noncompliance. This interpretation showed that these methods estimate a well-defined causal effect under the RCM. With the RCM interpretation of IV techniques, Imbens and Rubin (1997) were further able to improve on IV methods by introducing maximum likelihood and fully Bayesian methods to estimate causal effects in the presence of noncompliance.

The additional challenged posed by the Philadelphia Divorce Study is that the treatment indicator (indicating whether treatment was actually received by the unit) is often missing in the dataset. This problem did not arise in the influenza data set. To our knowledge, it has not been analyzed elsewhere. We extend the approach used in Hirano, et al. to the cases with missing treatment indicator data, specifically to cases where data is missing at random (MAR).

Our approach considers the case where the outcome of interest is binary, although the methodology could be extended to continuous outcomes.

The Philadelphia Divorce Study is one contribution to the growing empirical legal studies
movement in legal academia. In particular, there is a nascent movement in the legal academy to use randomized controlled trials to estimate causal effects that are of interest to legal scholars, courts and, ultimately, the public. Many of these recent studies have taken place within the context of an oversubscribed legal services provider, such that the organization had no choice but to decline legal representation to potential clients due to constrained resources (see, for example, Greiner and Pattanayak (2012) and Greiner et al. (2013)). In such contexts, it is feasible to randomize potential clients to either receive or not receive legal representation, for example, and thus measure the causal effects of having an attorney in legal proceedings.

Such studies often encounter a methodological difficulty insofar as the experimental units (the potential clients) can obtain legal representation when assigned to the control arm and can decline legal representation when randomized to the treatment arm. This persistent problem of two-way non-compliance in such studies is one motivation for developing a more flexible framework for handling such problems.

3.3 The Philadelphia Divorce Study

3.3.1 Legal background

In the United States, courts may generally impose a non-waivable filing fee. If such a filing fee exists, prospective litigants must pay a fee before they are able to file and proceed with their cases.

A trio of Supreme Court cases in the 1970s, taken together, interpreted the Due Process Clause of the Fourteenth Amendment to the United States Constitution as requiring some exceptions to a court’s general ability to impose non-waivable filing fees (Boddie v. Connecticut (1971), United States v. Kras (1973), Ortwein v. Schwab (1973)). Specifically, these cases state that the Constitution requires the availability of an in forma pauperis process for cases involving constitutional rights that can only be effectuated via the courts. An in forma
pauperis process provides for the waiver of filing fees for litigants who are unable to pay the fees. In particular, the Due Process Clause requires waiver of filing fees for indigent litigants who seek divorce. Married couples have a constitutional right to end their marriages. Furthermore, this right can only be effectuated through the courts because all jurisdictions in the United States require one spouse to sue the other in order to terminate a marriage. By contrast, it is trivially easy to form a marriage by simply forming a marriage contract (usually, via signing a marriage license).

More recently, the Supreme Court clarified when civil litigants have a right to an attorney. The Sixth Amendment of the Constitution (and the Supreme Court’s interpretation of this amendment in *Gideon v. Wainwright* (1963)) guarantees indigent criminal defendants a right to an attorney in both federal and state courts. However, litigants are not always guaranteed an attorney in civil (as opposed to criminal) cases. In certain civil cases, litigants are constitutionally guaranteed an attorney, but the law is by no means settled on all of the cases in which such a right exists. For example, as recently as the Supreme Court’s decision in *Turner v. Rogers* (2011), the Court held that a litigant has a civil right to an attorney in cases where there is significant procedural complexity, at least for cases where crucial rights were at issue.

One question of interest is when (if ever) is a person entitled to an attorney in a divorce proceeding. The right to a divorce qualifies as a “crucial” right under the reasoning of *Turner v. Rogers*. Indeed, the Court has already recognized that the right to divorce is a constitutionally protected right in the *in forma pauperis* cases.

Divorce court procedures in Philadelphia can be arcane and difficult. For example, some forms must be filled out via typewriter – neither handwritten nor computer printed forms are allowed. There are myriad different forms and documents required, and the required documentation depends on many factors, such as the grounds for divorce, whether the defendant has an attorney, how notice is provided to the defendant, and so forth. The litigant must know or figure out how to navigate these documents or else hire a competent attorney.
Mistakes, such as sending an original document when a copy should have been sent, result in delays. If a litigant fails to file a necessary document, the case will not move forward and neither will the court take action (such as dismissing the case or contacting the litigant, both of which might alert the litigant of the problem).

The Philadelphia Divorce Study proposed to indirectly test the procedural complexity of divorce court in Philadelphia county in a bid to determine whether divorce court procedures in this jurisdiction pass constitutional muster. Whether the procedural complexity of Philadelphia divorce court is so onerous that it requires the right to an attorney for potential litigants is, in part, an empirical question. The Philadelphia Divorce Study is a randomized controlled trial which intended to measure the causal effect of having an attorney on several outcomes, including a litigant’s ability to file for divorce and to obtain a divorce, both within a reasonable time frame. If the effect is large, then this is evidence that the process is too complicated for many litigants to navigate by themselves. If this problem is widespread and the state takes no steps to simplify its procedures, then these litigants are entitled, under the reasoning in *Turner*, to state provided legal representation.

Ultimately, the intention-to-treat effect was quite large. However, the effect of encouragement isn’t the direct estimand of interest; this could merely reflect the efficacy of the legal services provider with whom we partnered in the Philadelphia Divorce Study. A more direct estimand of interest is the complier average causal effect, which represents the causal effect of having an attorney on the primary outcomes of interest for an important subpopulation (the compliers). As we shall see, the CACE effect sizes are very large. Indeed, the complier average causal effects are so large that they seriously call into question the constitutionality of divorce court procedures in Philadelphia county.

The main analytical challenge for the study was estimating the complier average causal effect in the presence of severe missing data in the vector of indicators of which units actually received an attorney. A novel solution to this analytical challenge is the main focus of this study.
3.3.2 Philadelphia Divorce Study

Participants

The Access to Justice Lab is a research organization at the Harvard Law School that, *inter alia*, designs, implements and analyzes randomized experiments related to access to justice issues in the United States, often in partnership with legal aid organizations.

The Philadelphia Divorce Study was a partnership between the Access to Justice Lab and Philadelphia VIP, a pro-bono legal services provider in Philadelphia. Philadelphia VIP provides a variety of legal services for low-income residents of Philadelphia. Those services include a family law practice which, among other services, provides low-income residents with pro-bono legal representation in divorce proceedings.

Like many such legal services organizations, Philadelphia VIP does not have enough resources to help every potential client that seeks its representation. Since they must decline representation to some proportion of potential clients in the divorce practice, one way to decide which potential clients receive representation (perhaps the “fairest” way, in some sense of the word) is via randomization.

The Access to Justice Lab partnered with Philadelphia VIP to administer a randomized experiment in its divorce practice. The benefits of running such an experiment are manyfold including: (1) Helping to analyze and, potentially, optimize Philadelphia VIP’s family law practice and (2) analyze whether the Philadelphia divorce court system complies with constitutionally mandated procedural due process requirements. With these objectives in mind, the Philadelphia Divorce Study was conceived with the experimental design described below.

Design of the Philadelphia Divorce Study

Over a period of about two and a half years, 311 potential clients were randomized within blocks to either the treatment group or a control group. Enrollment in the study opened in
January 2011 and closed in July 2013. After randomization, it takes a minimum of three years to collect outcome data, such that the full length of the data collection process was over five years. Block randomization was used because not all of the potential clients are available at the same time. Rather, potential clients approach Philadelphia VIP one at a time with time sensitive requests for help. Once enough potential clients have asked for help (the average block size was approximately 17 units), the block is completely randomized to the treatment and control groups. After a block is randomized, the process of waiting for the next block begins. The number of units assigned to treatment varied somewhat by block but, overall, 74 of the 311 units (about 24%) were assigned to treatment.

In the treatment group, potential clients were offered pro bono representation by Philadelphia VIP divorce attorneys. In the control group, potential clients received a referral to existing self-help resources and an offer to have questions answered over the telephone.

The two outcomes of primary interest are both binary outcomes. The primary outcome was whether the unit filed for divorce within 18 months. The secondary outcome was whether the unit obtained a divorce within 36 months. The time periods associated with each outcome are in part practical (because we cannot wait forever to observe whether the unit obtained a divorce in the distant future) and in part a reflection of the true estimand; the constitutionality of these divorce procedures will hinge on whether they overly burden people who wish to obtain a divorce within a reasonable time frame.

All units that received divorces in the study had mutual consent or two-year separation as grounds for divorce. In other words, they were all no-fault divorces.

Noncompliance and Missing Data

Two-way noncompliance was observed in the study. Not all potential litigants accepted the offer of free representation (following the noncompliance literature, these study participants are referred to as “nevertakers”). Additionally, some people in the control group were able to obtain treatment (the “alwaystakers”). The alwaystakers were able to obtain legal
representation for their divorce court proceeding despite randomization to the control group.

The missing data mechanism for the treatment indicator is highly structured and fully known to the Philadelphia Divorce Study investigators. In the treatment group, we always observe the treatment indicator (whether the potential litigant has a lawyer or not). In the control group, by contrast, we only observe the treatment indicator if the unit manages to file for divorce. Recall that the filing of such a lawsuit is the primary outcome of interest. The investigators lacked the resources to follow the control group and thus could only determine whether control group study participants obtained a lawyer by reading public filings. Such filings exist if and only if the study participant manages to file for divorce. These filings always indicate whether or not the study participant is represented by an attorney in the divorce court proceeding. Of course, whenever the treatment indicator is missing, we do not know the compliance type of the individual. We need to know what treatment the individual would receive under both assignment to active treatment and assignment to control treatment in order to know an individual’s compliance type.

As mentioned above, the missing data mechanism is known exactly: \( W_i \) (whether unit \( i \) has a lawyer) is missing if and only if \( Z_i = 0 \) (unit \( i \) was assigned to the control group) and \( Y_{i,\text{obs}} = 0 \) (unit \( i \) failed to file for divorce). Before we dive into the data analysis methods used to estimate the ITT and, ultimately, the CACE, we define our notation and assumptions and describe the data using this notation.

### 3.4 Notation, Assumptions and Data

#### 3.4.1 Notation

For each experimental unit \( i \), we have the following notation:

1. \( Z_i \) is an indicator for the assignment of unit \( i \) to the active treatment. \( Z_i = 1 \) if the unit is assigned to the active treatment and is 0 if assigned to the control treatment.
2. $W_i(Z_i = 1)$ and $W_i(Z_i = 0)$ (often abbreviated as $W_i(1)$ and $W_i(0)$) are the potential outcomes for the intermediate outcome of whether the unit actually obtained an attorney or not. Essentially, these indicators encode whether unit $i$ complies with his or her treatment assignment. For example, $W_i(0) = 1$ if unit $i$ was assigned to control but obtained an attorney anyway (otherwise, $W_i(0) = 0$). These potential outcomes define the principal strata. For example, unit $i$ is an always-taker if $(W_i(1), W_i(0)) = (1, 1)$.

3. $W_{i}^{obs}$ and $W_{i}^{mis}$. $W_{i}^{obs}$ is the observed potential outcome for actual receipt of treatment. We often refer to these as the treatment indicators in this study. Analogously, $W_{i}^{mis}$ is the missing potential outcome. For example, if the unit is assigned to treatment, then $W_{i}^{obs} = W_i(1)$ and $W_{i}^{mis} = W_i(0)$. Calling these outcomes “missing” or “observed” is somewhat of a misnomer because a key challenge of this data set is that both are often missing. In other words, $W_{i}^{obs}$ is often missing, which we represent as $W_{i}^{obs} = \text{?}$. When we say that the treatment indicators are missing we mean that many components of the vector $W_{i}^{obs}$ are missing.

4. $G_i$ is the compliance status of unit $i$. Under the monotonicity assumption (see below), there are three possible compliance types: complier (co), always-taker (at), or never-taker (nt). This is typically an unobserved latent variable.

5. $Y_i(W_i = 1)$ and $Y_i(W_i = 0)$ are the potential outcomes for the outcomes of interest for unit $i$ under treatment and control, respectively. For example, these potential outcomes might represent whether a unit, $i$, manages to file for divorce within 18 months (or, secondarily, whether unit $i$ obtained a divorce). These potential outcomes are a well-defined function of $W_i$ under SUTVA and our exclusion restriction assumptions (see below).

6. $Y_{i}^{obs}$ and $Y_{i}^{mis}$ are the observed and missing potential outcome for unit $i$, respectively.

7. $X_i$ represents a vector of observed covariates for unit $i$. We may or may not use
covariates in the actual models that we construct, but it is nevertheless useful to define
$X_i$ for the purposes of describing our estimation procedures in their generality.

8. Our estimand will be denoted by $\tau_{fp} = \frac{1}{N_{co}} \sum_{i:G_i = co} (Y_i(1) - Y_i(0))$. In English, our
estimand is the average finite population treatment effect for compliers. Additionally,
we sometimes will refer to the super population estimand $\tau_{sp} = E(Y(1) - Y(0) \mid G = co)$. Occasionally, we will abbreviate $\tau_{fp}$ as $\tau$ for the sake of avoiding overly
cumbersome notation.

9. The bold-faced $Z$, $G$, etc refer to vectors (except for $X$, which denotes a matrix).

### 3.4.2 Assumptions

We explicitly make these following assumptions:

- **SUTVA** The Stable Unit Treatment Value Assumption (SUTVA) has two components.
  First is the assumption that the potential outcomes for any unit do not depend on
  the treatment assignments of other units (assumption of no interference). The second
  component is that there are no different versions of each treatment level such that those
different versions correspond to different potential outcomes (assumption of no hidden
variations of treatments). This assumption makes our use of notation like $Y_i(1)$ and
$Y_i(0)$ well defined, as this notation suggests that there is a one-to-one correspondence
between treatment received and potential outcomes for each unit (we don’t require a
multiplicity of potential outcomes that depend on the treatments of other units, for
example).

- **Monotonicity**: $W_i(Z_i = 1) \geq W_i(Z_i = 0)$. There are no “defiers.” This assumption
  seems quite reasonable: it would be strange for a potential litigant to seek out help
from a legal aid organization, reject the aid of an attorney if offered free legal services,
and yet hire an attorney if not offered free legal services. We have no reason to believe
that this occurs and, furthermore, it seems to defy common sense. We cannot directly test this assumption.

- **Exclusion Restrictions:** \( Y_i(Z_i = 0, W_i = w) = Y_i(Z_i = 1, W_i = w) \) for all units. This assumption seems reasonable for alwaystakers – one’s assignment should not affect the ability of one’s lawyer to file a divorce proceeding, which is a bare minimum of legal competence. It seems reasonable that the outcomes for nevertakers and compliers are likewise unaffected by randomization, conditional on the actual receipt of treatment. These assumptions can be easily relaxed to their stochastic versions, in which case they become additional quantities to impute in either the EM algorithm or the DA algorithm (see below).

### 3.4.3 Philadelphia Divorce Study Data

The observed data for the Philadelphia Divorce Study (aside from the covariates) can be summarized in Table 3.1. There are recorded covariates from surveys of the Philadelphia Divorce Study participants, which are useful for checking balance between the treatment and control groups with respect to these background covariates. They are also useful, although not necessary, for model-based estimation of the various causal estimands of interest. There are 7 patterns of missing and observed data in this dataset. Each pattern corresponds to a set of units that exhibit that pattern, which will be denoted by \( S(z, w, y) = \{ i \mid Z_i = z, W_i = w, Y_i^{obs} = y \} \). Table 3.1 summarizes the 7 distinct groups of units that are observed in the Philadelphia Divorce Study.

The compliance status, \( G_i \), for each unit \( i \) is often missing in the Philadelphia Divorce Study. Indeed, in any study where noncompliance is possible, the compliance statuses of the units are often unknown. When \( Z_i = 1 \) and \( W_i = 0 \), then the unit must be a nevertaker (\( G_i = nt \)). Similarly, when \( Z_i = 0 \) and \( W_i = 1 \), then the unit must be an alwaystaker (\( G_i = at \)). By contrast, when \( Z_i = W_i = 0 \), then \( G_i \in \{ co, nt \} \). When \( Z_i = W_i = 1 \), then \( G_i \in \{ co, at \} \). In short, when \( Z_i = W_i^{obs} \), then \( G_i \) is always missing. This missingness
in $G$ always occurs in randomized experiments with two-sided noncompliance. For such problems, Bayesian model-based inference on the CACE estimand can be performed using the procedure outlined in Imbens and Rubin, Ch. 25.

As previously mentioned, the added difficulty in the Philadelphia Divorce Study is that the treatment indicators, $W_i$, are often missing. This situation corresponds to units in the set $S(0, ?, 0)$. Note that about half of the units fall into this set, seemingly representing a large analytical roadblock. If the experimental unit does not file a lawsuit, then we cannot record whether they obtained a lawyer or not. In other words, $(Y_i = 0$ and $Z_i = 0) \implies W_i$ is missing. When $W_i$ is missing in this fashion, $G_i$ could conceivably take any value $\{\text{co, at, nt}\}$. This additional uncertainty in the compliance statuses $G$ creates extra uncertainty in the CACE.

Table 3.1: Patterns of Missing and Observed Data

<table>
<thead>
<tr>
<th>$S(z, w, y)$</th>
<th>$G_i$</th>
<th>$Z_i$</th>
<th>$W_i$</th>
<th>$Y_i$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>S(0, ?, 0)</td>
<td>${\text{co, nt, at}}$</td>
<td>0</td>
<td>?</td>
<td>0</td>
<td>150</td>
</tr>
<tr>
<td>S(0, 0, 1)</td>
<td>${\text{co, nt}}$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>60</td>
</tr>
<tr>
<td>S(0, 1, 1)</td>
<td>at</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>27</td>
</tr>
<tr>
<td>S(1, 0, 0)</td>
<td>nt</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>S(1, 1, 0)</td>
<td>${\text{co, at}}$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>S(1, 0, 1)</td>
<td>nt</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>S(1, 1, 1)</td>
<td>${\text{co, at}}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>42</td>
</tr>
</tbody>
</table>

In all, 74 participants were randomized to the active treatment and 237 were randomized to the control treatment.

To get a sense of the study participants, we can summarize some of the primary covariates. Beginning with demographic information, most of the participants were female (80%). The mean age of a study participant was 42 years. Of the study participants, 59% were black, 22% were Hispanic, and 15% were white (the corresponding racial breakdown in the Philadelphia populations are 43%, 11%, and 37%, respectively). Interpreters were required for 16% of study participants (the rest were English speakers).

The participants reported substantial financial distress. The majority had 0 income
with 95% earning less than $23,000 annually. Only 35% were employed. Two thirds received public assistance. The median income for employed participants was $12,000 per year. Three quarters of experimental units had less than $500 in the bank. Interestingly, despite the lack of income and savings, about one third were homeowners.

Turning to the marriages themselves, the mean marriage length was 12 years. Only about 8% of study participants still lived with their spouses. With respect to minor children, 43% had a minor child and 38% wanted custody of the child. However, the cases that are observed seem relatively simple. Over 80 percent of the cases that were actually filed were simple terminations of marriage that required no division of property, child custody, or child support.

3.5 Intention-to-Treat Analysis of the Philadelphia Divorce Study

3.5.1 Fisher’s Exact Test

Beginning with the observed data, Table 3.2 shows the counts of successes for each outcome, out of the 311 total units. In parentheses, we have the success rate for that outcome.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Number of Successes</th>
<th>Number of Treatment Successes</th>
<th>Number of Control Successes</th>
<th>Difference in Mean Success Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>File for Divorce (PA)</td>
<td>130 (0.42)</td>
<td>45 (0.61)</td>
<td>85 (0.36)</td>
<td>0.25</td>
</tr>
<tr>
<td>File for Divorce (PHI)</td>
<td>73 (0.24)</td>
<td>40 (0.54)</td>
<td>33 (0.14)</td>
<td>0.40</td>
</tr>
<tr>
<td>Obtain Divorce (PA)</td>
<td>95 (0.31)</td>
<td>36 (0.49)</td>
<td>59 (0.25)</td>
<td>0.24</td>
</tr>
<tr>
<td>Obtain Divorce (PHI)</td>
<td>53 (0.17)</td>
<td>33 (0.45)</td>
<td>20 (0.08)</td>
<td>0.36</td>
</tr>
</tbody>
</table>

As noted earlier, the two outcomes of interest are whether the unit filed for divorce and whether the unit obtained a divorce. These outcomes are further split out by whether the unit succeeded in filing/obtaining a divorce in Philadelphia County, or whether they succeeded in filing/obtaining a divorce in Philadelphia and/or one of seven neighboring counties in Pennsylvania that were known to occasionally grant divorces to Philadelphians.
These courts would process mutual consent cases by mail and had lower filing fees than Philadelphia (about $85 versus $325 in Philadelphia, although the Philadelphia courts do have an *in forma pauperis* process for pro se litigants). By law, the Philadelphia court system is the proper venue for divorce proceedings for the study participants and these outlying counties are improper venues. If an improper venue is chosen and the defendant objects, the case will be dismissed. However, to account for the practicality that these venue rules are sometimes bent (or broken) we include these other outcomes.

Depending on the outcome, the success rate might be double or up to five times larger in the treatment group than in the control group. We can formally test for statistical significance of these differences by using Fisher’s exact test, using the difference in means as a test statistic and taking care to account for the block randomization scheme.

Table 3.3 shows the results of these Fisher exact tests, as well as point estimates for the ITT and confidence intervals. We avoid specifying a likelihood function at this stage of analysis, as the assumption of a likelihood is not yet necessary. Point estimates and confidence intervals can be constructed using a variety of methods. The point estimates and intervals in Table 3.3 are constructed using Horvitz-Thompson inverse probability weight methods (other approaches yielded similar results).

<table>
<thead>
<tr>
<th>Outcome</th>
<th>P-value</th>
<th>Pt. Estimate</th>
<th>Std. Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>File for Divorce (PA)</td>
<td>$4.3 \times 10^{-5}$</td>
<td>0.31</td>
<td>0.056</td>
</tr>
<tr>
<td>File for Divorce (PHI)</td>
<td>0</td>
<td>0.44</td>
<td>0.058</td>
</tr>
<tr>
<td>Obtain Divorce (PA)</td>
<td>$2.8 \times 10^{-5}$</td>
<td>0.28</td>
<td>0.060</td>
</tr>
<tr>
<td>Obtain Divorce (PHI)</td>
<td>0</td>
<td>0.39</td>
<td>0.059</td>
</tr>
</tbody>
</table>

Figure 3.1 is an example plot that graphically analogizes Fisher’s exact test. The vertical red line is the point estimate and the histogram represents the values of the test-statistic under counterfactual randomizations under the null hypothesis. We can see the observed value is quite unlikely under the null. Indeed, it is the most extreme value that was observed over 100,000 randomizations, which is why we estimate the p-value as 0. Similar plots for
the other outcomes look very similar.

![Fisher Randomization Distribution](image)

**Figure 3.1: Fisher Randomization Distribution for File in Pennsylvania Outcome**

The ITT effects all seem quite significant and the differences in success probabilities are large. If we think of the intention to treat effect as the effect of the Philadelphia VIP divorce practice on its clients, we see that the Philadelphia VIP divorce practice is highly effective in aiding many clients. The result also raises suspicion about the accessibility of divorce court for Philadelphians who cannot obtain an attorney for their divorce case.

With regards to balance on the background covariates, we observed no covariate imbalance that would raise concerns for the study. One of the core benefits of a randomized controlled experiment is that, due to randomization, the active and control treatment arms are balanced on background covariates, in expectation. Therefore, in expectation, we have no confounding variables and we can much more easily attribute differences between the groups to the actual treatments that were applied. Of course, we only achieve balance *in expectation* and it is therefore worth checking whether balance is *actually* achieved.
We recorded over 400 covariates for the 311 units. Philly VIP conducts a 45-60 minute interview for each potential client. The covariates come from records from these interviews. While some imbalance was observed for some of these covariates, as would be expected with so many covariates, these were typically uninteresting covariates that were not expected to have a large impact on study results. For example, one covariate that was unbalanced was the average balance of non-car, non-home loans (the treated group had an average balance of about $4000 vs $2000 in the control group). However, since we didn’t observe property division issues in the divorce proceedings and since this imbalance, if it had any effect at all, would likely hinder the treatment group more than the control group, this imbalance seems, in essence, irrelevant. Thus, broadly speaking, no problems were raised by the balance checks.

3.6 Analysis Methods

We consider both maximum-likelihood and full Bayesian approaches to analysis. These methods have some similarity insofar as both will require assumptions about the form of the likelihood function. One main analysis challenge is the large amount of missing data in both $W_t$, the treatment vector, and $G_t$, the vector of compliance statuses. The compliance statuses are, in fact, (mostly) unobserved latent variables. The missing compliance statuses result in an observed-data likelihood function with inconvenient mixture components.

Our approach will be to augment the data space to include these latent variables. These data augmentation methods handle the missing data by exploiting the relative simplicity of the complete data likelihood. We consider two approaches:

1. Use the EM Algorithm to calculate the MLE of the model parameters

2. Use the Data Augmentation algorithm to get draws from the posterior distribution of the estimand
These methods have many similarities and, perhaps unsurprisingly, produce similar inferences under certain conditions, as we will see below. However, results from these two methods can also diverge, as we will also see.

Both methods employ the same complete data likelihood function. We make note of the observed data likelihood to underscore its complexity when compared to the complete data likelihood. However, the observed data likelihood is also useful for debugging purposes, for example, to verify that the observed data likelihood is increasing at each iteration of our implementation of the EM algorithm. Additionally, it is sometimes useful to evaluate the observed data likelihood at various points, such as at the MLE.

On the face of it, the Philadelphia Divorce Study data set and data sets like that which was analyzed in Hirano et al. (2000) seem very different: the former has a tremendous amount of missing data in a key variable (the indicators for whether treatment was actually received) whereas the latter kind of data set, if you simply loaded into one’s favorite statistical software, would appear as if it had no missing data. The missing data in the former would frustrate traditional instrumental variable approaches (without further accounting for the missing data), whereas the latter would be a trivial application of these methods.

However, by first thinking about the problems in terms of the RCM, we note that there is considerable missing data in both data sets, notably in the potential outcomes and in the vector of compliance statuses. Due to the two-way noncompliance, the primary treatment in the Philadelphia Divorce Study is only latently unconfounded. That is, unconfoundedness only holds conditional on a partially observed covariate (specifically, it holds for compliers, yet we don’t know exactly which units comprise the set of compliers). This key observation suggests that principal stratification will be a fruitful approach to analysis, and so we recognize the need to introduce the compliance statuses as a partially observed latent variable.

By introducing the compliance status vector, \( G \), which is only partially observed, we can more easily handle missingness in the vector of treatment actually received, \( W \). \( W \) is determined by the compliance statuses \( G \) and the random assignments \( Z \). Therefore, we
only need to impute the missing values in $G$ in order to deal with missing values in $W$. In some sense, we have substituted one missing data problem with another. The missingness in $W$ still causes difficulty by increasing the amount of missing data in $G$ and, additionally, making the imputation of missing values in $G$ trickier. Yet the introduction of the machinery associated with the RCM framework allows for a tractable extension of existing analysis methods to data sets with substantial missingness in $W$. We will exploit this fact in our models below.

### 3.6.1 Maximum Likelihood Approach

We can calculate maximum likelihood estimates for the super population complier average causal effect, if we are willing to make parametric assumptions about the form of the likelihood. In particular, we use the EM algorithm to calculate the MLE. The EM algorithm is a natural choice due to the presence of the compliance statuses in the complete data likelihood (see below). The compliance statuses are latent variables that we typically do not know, and whose introduction dramatically simplifies the complete data likelihood, relative to the observed data likelihood. The EM algorithm was introduced by Dempster, Laird and Rubin (1977) and the original paper contains many illustrative examples that are similar in spirit to the application of EM to this data set.

One benefit of using the EM Algorithm is that the maximum-likelihood estimators are widely used and broadly understood. An added advantage of maximum likelihood over Bayesian alternatives is that there is no need to choose a prior distribution on the parameter vector.

The MLE performs at its best when its asymptotic properties approximately hold but it can fail badly when this is not the case – we will see examples of this later. It can fail especially badly when we have a multimodal likelihood function. An additional downside of the maximum likelihood approach is that we cannot clearly estimate the finite population estimand since it is not a function of the model parameters.
In contrast to the maximum-likelihood approach, Bayesian methods allow for estimation of the finite population estimand, they yield the full posterior distribution (making credible interval estimation straightforward), and only require the additional assumption of a prior. In general, the choice of prior in Bayesian modeling can be challenging especially when, as here, we have no real pre-existing expert knowledge. For this reason, we focus on flat (proper) priors and on weak pseudo observation priors as relatively innocuous choices.

Experience suggests that the stronger assumptions lie in the likelihood, rather than in the prior. Therefore, the most likely failure point lies with misspecification of the likelihood – this is a problem for the maximum likelihood approach as well as for Bayesian approaches (this statement is consistent with the findings of the previous chapter). For these reasons, we will emphasize results from the DA Algorithm, although we also present maximum-likelihood results.

### 3.6.2 Bayesian Approach

The finite population causal estimands that we consider are of the form \( \tau(Y, G, X, Z) \). In particular, they are usually some comparison of the potential outcomes under treatment, \( Y(1) \), with the potential outcomes under control, \( Y(0) \), perhaps restricted to some subpopulation defined by the compliance statuses, \( G \) (for example, the complier average causal effect) or some subpopulation defined by the covariates, \( X \).

Any causal estimate of this form can be written as \( \tau(Y_{\text{obs}}, Y_{\text{mis}}, W_{\text{obs}}, W_{\text{mis}}, X, Z) \). This quantity is unknown and random because the quantity \( (Y_{\text{mis}}, W_{\text{mis}}) \) is unknown and random. Our objective is therefore to derive the posterior predictive distribution of \( (Y_{\text{mis}}, W_{\text{mis}}) \) conditional on the observed data \( (Y_{\text{obs}}, W_{\text{obs}}, X, Z) \). The causal estimand, \( \tau \), is simply a deterministic function of these observed and predicted quantities and thus, by obtaining draws from the posterior predictive distribution of \( (Y_{\text{mis}}, W_{\text{mis}}) \), we immediately obtain draws from the posterior distribution of \( \tau \).
We don’t directly specify the posterior predictive distribution of the missing data,

\[ f(\mathbf{Y}^{\text{mis}}, \mathbf{W}^{\text{mis}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z}) \]

because, in general, this is too difficult. Rather, we will obtain the posterior predictive distribution of the missing data by introducing some parametric assumptions that allow us to obtain the posterior distribution of the parameter, \( \theta \), given the observed data as well as the posterior distribution of the missing data, given the observed data and \( \theta \). Integrating out \( \theta \), we obtain the desired posterior predictive distribution of the missing data. More formally,

\[
    f(\mathbf{Y}^{\text{mis}}, \mathbf{W}^{\text{mis}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z}) = \\
    \int_{\theta} f(\mathbf{Y}^{\text{mis}}, \mathbf{W}^{\text{mis}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z}, \theta) p(\theta | \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z}) d\theta
\]

This expression is the integral over \( \theta \) of the product of two terms: (1) \( f(\mathbf{Y}^{\text{mis}}, \mathbf{W}^{\text{mis}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z}, \theta) \), the conditional distribution of the missing potential outcomes conditional on the parameter and the observed data and (2) \( p(\theta | \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z}) \), the posterior distribution for \( \theta \), both of which are derived in the following sections.

**Derivation of** \( f(\mathbf{Y}^{\text{mis}}, \mathbf{W}^{\text{mis}} | \mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Z}, \theta) \)

We begin by specifying three components that together can be used to construct the posterior predictive distribution.

First, we model the potential outcomes given the compliance status, covariates and parameter: \( f(Y_i(0), Y_i(1) | G_i, X_i, \theta) \). We assume unit exchangeability and appeal to De Finetti’s theorem to justify conditional independence conditional on the parameter: \( f(\mathbf{Y} | \mathbf{G}, \mathbf{X}, \theta) = \prod_{i=1}^{N} f(Y_i(0), Y_i(1) | G_i, X_i, \theta) \)

Second, we model the compliance statuses conditional on the covariates and the parameter: \( f(G | \mathbf{X}, \theta) = \prod_{i=1}^{N} f(G_i | X_i, \theta) \). Note that we will have already taken these first two
steps in writing down the complete data likelihood in order to use the EM algorithm to find
the maximum likelihood estimates of the parameters.

The third building block for the Bayesian approach is the choice of prior. We will aim
to choose uninformative or weakly informative priors so that the posterior is dominated by
the data, rather than by the prior, and so, if done appropriately, the choice of prior is a
relatively innocuous choice when compared to the choice of the likelihood.

The specification of \( f(Y \mid G, X, \theta) \) and \( f(G \mid X, \theta) \) together determine the joint distri-
bution \( f(Y(1), Y(0), W(1), W(0) \mid X, \theta) \), which, due to the ignorability of the assignment
mechanism, is proportional to \( f(Y(1), Y(0), W(1), W(0) \mid X, \theta, Z) \). In effect, we have
derived the joint density \( f(Y^{mis}, Y^{obs}, W^{mis}, W^{obs} \mid X, Z, \theta) \) because these missing and ob-
served vectors are a function of the assignments, \( Z \) and the potential outcomes \((Y(1), Y(0),
W(1), W(0))\). An invocation of Bayes’ Rules provides us with the conditional distribution
of the missing potential outcomes conditional on the parameter and the observed data:

\[
f(Y^{mis}, W^{mis} \mid Y^{obs}, W^{obs}, X, Z, \theta) = \frac{f(Y^{mis}, Y^{obs}, W^{mis}, W^{obs} \mid X, Z, \theta)}{\int \int f(Y^{mis}, Y^{obs}, W^{mis}, W^{obs} \mid X, Z, \theta) dY^{mis} dW^{mis}}
\]

**Derivation of \( p(\theta \mid Y^{obs}, W^{obs}, X, Z) \)**

The derivation of the posterior distribution of \( \theta \) given observed data is, in the abstract, a
familiar task in Bayesian inference. The posterior distribution of \( \theta \), \( p(\theta \mid Y^{obs}, W^{obs}, X, Z) \),
is proportional to the product of the likelihood and the prior:

\[
p(\theta \mid Y^{obs}, W^{obs}, X, Z) \propto p(\theta) L(\theta \mid Y^{obs}, W^{obs}, X, Z)
\]

If we were to analytically write down the posterior, we would want to find a normalizing
constant such that the product of the normalizing constant and the expression above inte-
grated to 1. However, we will ignore the normalizing constant since we will use simulation
methods to sample from all of the necessary components of the posterior predictive distri-
bution of the missing potential outcomes and the normalizing constant will not be required for these simulation methods.

Noncompliance and the DA Algorithm

Thus far we have generally described Bayesian inference in the causal inference setting. The integrals in the above derivation are usually intractable and we would therefore typically use Markov chain Monte Carlo (MCMC) algorithms to obtain samples from the target posterior distributions.

It would be relatively easy to get draws from the posterior of $\theta$ (and draws from the posterior predictive distribution of the missing potential outcomes) if we could ignore non-compliance or if we knew the compliance statuses for each of the units. Using the complete data likelihood with compliance status latent variables, we can use the Data Augmentation algorithm to sample from the posterior of the model parameters. The Data Augmentation algorithm is an MCMC algorithm and a Bayesian analog of the EM algorithm by which one can obtain samples from a posterior distribution when operating in a data space that is augmented with latent variables. See Section 3.7 below for a brief explanation of the algorithm as well as the specific instantiation of the algorithm that we use in our simulations and in analysis of the Philadelphia Divorce Study data set.

3.6.3 Complete Data Likelihood

In both the maximum likelihood and Bayesian approaches, we start by writing down the complete data likelihood. The complete data likelihood is the likelihood that would result if we observed all of the compliance statuses. These compliance statuses are typically unobserved latent variables and thus we need to use missing data methods such as the EM algorithm or the Data Augmentation algorithm in order to exploit the complete data likelihood. Here we write the complete likelihood as a function of the unobserved (or partially observed) compliance statuses:
\[ L_{\text{comp}} = \prod_{i:G_i = nt} f(Y_i(0) \mid G_i = nt, X_i, Z_i, \beta_{nt}) \]
\[ \times \prod_{i:G_i = at} f(Y_i(1) \mid G_i = at, X_i, Z_i, \beta_{at}) \]
\[ \times \prod_{i:G_i = co, Z_i = 0} f(Y_i(0) \mid G_i = co, X_i, Z_i, \beta_{co,c}) \]
\[ \times \prod_{i:G_i = co, Z_i = 1} f(Y_i(1) \mid G_i = co, X_i, Z_i, \beta_{co,t}) \]
\[ \times \prod_{i:G_i = nt} p(G_i = nt \mid X_i, Z_i, \gamma) \]
\[ \times \prod_{i:G_i = at} p(G_i = at \mid X_i, Z_i, \gamma) \]
\[ \times \prod_{i:G_i = co} p(G_i = co \mid X_i, Z_i, \gamma) \]

This can be rewritten as:
\[ L_{\text{comp}} = \prod_{i=1}^{n} f(Y_i(0) \mid G_i = nt, X_i, Z_i, \beta_{nt}) I(G_i = nt) \]
\[ \times f(Y_i(1) \mid G_i = at, X_i, Z_i, \beta_{at}) I(G_i = at) \]
\[ \times f(Y_i(0) \mid G_i = co, X_i, Z_i, \beta_{co,c}) I(G_i = co) I(Z_i = 0) \]
\[ \times f(Y_i(1) \mid G_i = co, X_i, Z_i, \beta_{co,t}) I(G_i = co) I(Z_i = 1) \]
\[ \times p(G_i = nt \mid X_i, Z_i, \gamma) I(G_i = nt) \]
\[ \times p(G_i = at \mid X_i, Z_i, \gamma) I(G_i = at) \]
\[ \times p(G_i = co \mid X_i, Z_i, \gamma) I(G_i = co) \]

Given the relative simplicity of the complete data likelihood (relative to the observed data likelihood, which is described in the next section), it is worth using the data augmentation algorithm if we wish to generate draws from the posterior distribution of the parameters.
This is essentially a Gibbs sampler where we serially draw the prior, the compliance statuses and the potential outcomes at each iteration (see below). For the EM algorithm, $\mathcal{L}_{\text{comp}}$ can be rewritten as one product with indicators for the compliance statuses (such as $I_{i=nt}$) and the values of these indicators are the missing data that get averaged over in the E-step.

### 3.6.4 Observed Data Likelihood

The observed data likelihood is based only on the observed data (rather than including the unobserved latent variables). We include it to emphasize that it is unwieldy when compared with the complete data likelihood. This comparison provides motivation for using data augmentation methods, which can initially seem like unnecessary complication. However, Bayesian inferences using the observed data likelihood, with its numerous mixture components, would require a (possibly) high dimensional Metropolis Hastings or Hamiltonian Monte Carlo step in a (possibly) high dimensional space (the dimension can be very large when the model incorporates covariates) within a Gibbs sampler, which could cause poor computational performance.

This isn’t to say that the observed data likelihood is useless – it can be used, for example, to verify that the EM algorithm increases the likelihood at each step. It is therefore useful for checking that the algorithm is implemented as intended.

Before we write down the observed data likelihood functions, let’s again define $S(z, w, y)$ to be the set of units $\{i \mid Z_i = z, W_i = w, Y_i = y\}$. We consider the case where $Z_i$ and $Y_i$ are fully observed (with two possible values each) but $W_i$ is potentially missing (three possible states: $W_i = 1$, $W_i = 0$, or $W_i =$ ?). Therefore, the set of all units can be decomposed into $2 \times 2 \times 3 = 12$ distinct disjoint sets. Note that not all 12 of these sets exist in the Philadelphia Divorce Study dataset. We include them because, in general, they can exist and, in fact, do exist in data sets that are generated in the simulation studies below.

1. $S(1,?,1) = \{i \mid Z_i = 1, W_i = ?, Y_i = 1\}$
2. \(S(1,?,0) = \{i \mid Z_i = 1, W_i = ?, Y_i = 0\}\)

3. \(S(0,?,1) = \{i \mid Z_i = 0, W_i = ?, Y_i = 1\}\)

4. \(S(0,?,0) = \{i \mid Z_i = 0, W_i = ?, Y_i = 0\}\)

5. \(S(1,1,1) = \{i \mid Z_i = 1, W_i = 1, Y_i = 1\}\)

6. \(S(1,1,0) = \{i \mid Z_i = 1, W_i = 1, Y_i = 0\}\)

7. \(S(0,1,1) = \{i \mid Z_i = 0, W_i = 1, Y_i = 1\}\)

8. \(S(0,1,0) = \{i \mid Z_i = 0, W_i = 1, Y_i = 0\}\)

9. \(S(1,0,1) = \{i \mid Z_i = 1, W_i = 0, Y_i = 1\}\)

10. \(S(1,0,0) = \{i \mid Z_i = 1, W_i = 0, Y_i = 0\}\)

11. \(S(0,0,1) = \{i \mid Z_i = 0, W_i = 0, Y_i = 1\}\)

12. \(S(0,0,0) = \{i \mid Z_i = 0, W_i = 0, Y_i = 0\}\)

The observed data likelihood can be written as a product of these twelve constituent parts. We can write the observed data likelihood as follows:
\[ L_{\text{obs}} = \prod_{i \in S(0,1,y)} p(G_i = at \mid X_i, Z_i, \gamma) \cdot f(Y_i(1) = y \mid G_i = at, X_i, Z_i, \beta_{at}) \]
\[ \times \prod_{i \in S(1,0,y)} p(G_i = nt \mid X_i, Z_i, \gamma) \cdot f(Y_i(0) = y \mid G_i = nt, X_i, Z_i, \beta_{nt}) \]
\[ \times \prod_{i \in S(1,1,y)} [p(G_i = at \mid X_i, Z_i, \gamma) \cdot f(Y_i(1) = y \mid G_i = at, X_i, Z_i, \beta_{at}) \]
\[ + p(G_i = co \mid X_i, Z_i, \gamma) \cdot f(Y_i(1) = y \mid G_i = co, X_i, Z_i, \beta_{co,t})] \]
\[ \times \prod_{i \in S(0,0,y)} [p(G_i = nt \mid X_i, Z_i, \gamma) \cdot f(Y_i(0) = y \mid G_i = nt, X_i, Z_i, \beta_{nt}) \]
\[ + p(G_i = co \mid X_i, Z_i, \gamma) \cdot f(Y_i(0) = y \mid G_i = co, X_i, Z_i, \beta_{co,c})] \]
\[ \times \prod_{i \in S(1,?,y)} [p(G_i = at \mid X_i, Z_i, \gamma) \cdot f(Y_i(1) = y \mid G_i = at, X_i, Z_i, \beta_{at}) \]
\[ + p(G_i = nt \mid X_i, Z_i, \gamma) \cdot f(Y_i(0) = y \mid G_i = nt, X_i, Z_i, \beta_{nt}) \]
\[ + p(G_i = co \mid X_i, Z_i, \gamma) \cdot f(Y_i(1) = y \mid G_i = co, X_i, Z_i, \beta_{co,t})] \]
\[ \times \prod_{i \in S(0,?,y)} [p(G_i = at \mid X_i, Z_i, \gamma) \cdot f(Y_i(1) = y \mid G_i = at, X_i, Z_i, \beta_{at}) \]
\[ + p(G_i = nt \mid X_i, Z_i, \gamma) \cdot f(Y_i(0) = y \mid G_i = nt, X_i, Z_i, \beta_{nt}) \]
\[ + p(G_i = co \mid X_i, Z_i, \gamma) \cdot f(Y_i(0) = y \mid G_i = co, X_i, Z_i, \beta_{co,c})] \]

Note the complicated mixture density components, for example, the terms in the product over \( S(0, ?, 0) \). We would rather deal with these mixtures gracefully in an augmented space.

### 3.7 Model for The Philadelphia Divorce Study

#### 3.7.1 Model of the Potential Outcomes Conditional on Compliance Type

We model the potential outcomes conditional on the compliance status and treatment received as Bernoulli random variables.
For compliers:

\[
f(Y_i(1) = y \mid G_i = \text{co}, X_i, Z_i, p_{\text{co}, t}) = \Pr(Y_i(1) = y \mid x_i, G_i = \text{co}, \theta) = p_{\text{co}, t}^{y}(1 - p_{\text{co}, t})^{1-y}
\]

\[
f(Y_i(0) = y \mid G_i = \text{co}, X_i, Z_i, p_{\text{co}, c}) = \Pr(Y_i(0) = y \mid x_i, G_i = \text{co}, \theta) = p_{\text{co}, c}^{y}(1 - p_{\text{co}, c})^{1-y}
\]

For always takers:

\[
f(Y_i(1) = y \mid G_i = \text{at}, X_i, Z_i, p_{\text{at}}) = \Pr(Y_i(1) = y \mid x_i, G_i = \text{at}, \theta) = p_{\text{at}}^{y}(1 - p_{\text{at}})^{1-y}
\]

For never takers:

\[
f(Y_i(0) = y \mid G_i = \text{nt}, X_i, Z_i, p_{\text{nt}}) = \Pr(Y_i(0) = y \mid x_i, G_i = \text{nt}, \theta) = p_{\text{nt}}^{y}(1 - p_{\text{nt}})^{1-y}
\]

Recall that in our notation \( Y_i(1) = Y_i(W_i = 1) \)

### 3.7.2 Model of Compliance Type

**Multinomial Model of Compliance Type**

We model compliance type with a multinomial logit model.

\[
p(G_i = \text{nt} \mid X_i, Z_i, q) = q_{\text{nt}}
\]

\[
p(G_i = \text{at} \mid X_i, Z_i, q) = q_{\text{at}}
\]

\[
p(G_i = \text{co} \mid X_i, Z_i, q) = 1 - q_{\text{nt}} - q_{\text{at}}
\]

There are multiple choices we could make to model the compliance type, such as using sequential binomials (see the next subsection). However, we found that results are not very sensitive to this choice. Results for other possible choices are available in Appendix B.
Sequential Binomial Model of Compliance Type

As mentioned above, we could alternatively model the compliance status in two stages: We could first infer whether the unit is a nevertaker or not. If the unit is not a nevertaker, then we could infer whether it is a complier or an alwaystaker. In some scenarios, this sort of model may be more interpretable. For example, it’s possible that nevertakers are simply quite different from other units. These are potential clients who refuse legal representation even if it is offered to them for free and they are therefore less likely to be serious about getting divorced in the first place. Perhaps they approached Philadelphia VIP while still angry or upset after a domestic dispute but were never really serious about following through with a divorce. Some such story would be possible motivation for modeling each unit in two stages – first determining whether they are nevertakers and then, if the unit is not a nevertaker, further inferring its compliance status.

When we omit covariates from the model, this two stage process requires rewriting the likelihood in terms of the following probabilities:

\[ p(G_i = nt | X_i, \theta) = \phi_{nt} \]

\[ p(G_i = co | X_i, G_i \in \{co, at\}, \theta) = \phi_{co} \]

This modeling choice will change several of the terms in our likelihood function. In particular, these terms take new forms:

\[ p(G_i = co | X_i, \theta) = Pr(G_i = co | X_i, G_i \in \{co, at\}, \theta) * Pr(G_i \in \{co, at\} | X_i, \theta) = \phi_{co}(1 - \phi_{nt}) \]

\[ p(G_i = at | X_i, \theta) = (1 - \phi_{co})(1 - \phi_{nt}) \]

This is basically a partial reparameterization of the case where we model the compliance statuses via a multinomial:

\[ (q_{nt}, q_{co}, 1 - q_{nt} - q_{co}) \to (\phi_{nt}, (1 - \phi_{nt})\phi_{co}, (1 - \phi_{co})(1 - \phi_{co})) \]
Choosing a new prior in this reparameterized space can change our inferences. For the choice of prior on $\phi_{nt}$ and $\phi_{nt}$, we considered again both a flat prior and prior of synthetic observations. Again, inferences obtained under the sequential binomial model for compliance type were very similar to those obtained under the multinomial model for compliance type. We present results for the multinomial model below.

3.7.3 Complete Data Likelihood

We begin by rewriting the complete data likelihood as follows, using the multinomial model for compliance type:

$$L_{comp} = \prod_{i=1}^{n} \left[ p_{nt}^{y_i} (1 - p_{nt})^{1-y_i} \right] I(G_i=nt)$$

$$\times \left[ p_{at}^{y_i} (1 - p_{at})^{1-y_i} \right] I(G_i=at)$$

$$\times \left[ p_{co,t}^{y_i} (1 - p_{co,t})^{1-y_i} \right] I(G_i=co) I(Z_i=1)$$

$$\times \left[ p_{co,c}^{y_i} (1 - p_{co,c})^{1-y_i} \right] I(G_i=co) I(Z_i=0)$$

$$\times q_{nt}^{I(G_i=nt)} q_{at}^{I(G_i=at)} (1 - q_{nt} - q_{at}) I(G_i=co)$$

The complete data log-likelihood is therefore:

$$\ell_{comp}(\theta) = \sum_{i=1}^{n} \left[ I(G_i = nt) [ y_i \log(p_{nt}) + (1 - y_i) \log(1 - p_{nt}) ] \right]$$

$$+ I(G_i = at) [ y_i \log(p_{at}) + (1 - y_i) \log(1 - p_{at}) ]$$

$$+ I(G_i = co) z_i [ y_i \log(p_{co,t}) + (1 - y_i) \log(1 - p_{co,t}) ]$$

$$+ I(G_i = co)(1 - z_i) [ y_i \log(p_{co,c}) + (1 - y_i) \log(1 - p_{co,c}) ]$$

$$+ I(G_i = nt) \log(q_{nt}) + I(G_i = at) \log(q_{at}) + I(G_i = co) \log(1 - q_{nt} - q_{at})$$
Please see Appendix B for the corresponding complete data likelihood with the sequential binomial model of compliance type.

### 3.7.4 EM Algorithm

We will apply the EM algorithm to find the MLE of the model parameters. The missing data in the complete data likelihood consist of many of the entries of the compliance vector \( G \). We may apply EM when the missing data is MAR (Missing at Random), as it is in all examples considered in this study.

**EM Algorithm**

In the E-step, at each iteration we take the expectation of the log-likelihood conditional on all the observed data and conditional on \( \theta = \theta^{(t)} \), where \( \theta^{(t)} \) is the result of the previous iteration’s M-step approximation to \( \theta \), the vector of model parameters. This expectation is taken over the missing data \( G \). We take these expectations conditional on all of the observed data; note that sometimes we will condition on \( W_i \) and other times we will not, depending on whether it is observed. For this particular model, this task requires calculating the following expectations, for each \( i \):

- \( E(I(G_i = nt \mid Z_i, W_i, Y_i, \theta^{(t)}) = Pr(G_i = nt \mid Z_i, W_i, Y_i, \theta^{(t)}) \)
- \( E(I(G_i = at \mid Z_i, W_i, Y_i, \theta^{(t)}) = Pr(G_i = at \mid Z_i, W_i, Y_i, \theta^{(t)}) \)
- \( E(I(G_i = co \mid Z_i, W_i, Y_i, \theta^{(t)}) = Pr(G_i = co \mid Z_i, W_i, Y_i, \theta^{(t)}) \)

\( G_i \) is known when \( Z_i \neq W_i \): these units are either nevertakers or alwaystakers, depending on their assignment. This calculation only requires real work when \( G_i \) is actually missing. Note that we will re-use these probability calculations later when describing the DA algorithm.

\( G_i \) is missing in the following scenarios:
• $Z_i = 1$ and $W_i = 1$

• $Z_i = 0$, $W_i = 0$ and $Y_i = 1$

• $Z_i = 0$, $W_i = ?$ and $Y_i = 0$. This is the unusual case.

**Case 1:** $Z_i = 1$, $W_i = 1$ and $Y_i = y_i$

Note that $I(G_i = at)$ is not deterministic when $Z_i = 1$ and $W_i = 1$. We therefore need to derive the probability that $G_i = at$ for this case.

\[
Pr(G_i = at \mid Z_i = 1, W_i = 1, Y_i, \theta^{(t)}) = \frac{Pr(Y_i \mid G_i = at, Z_i = 1, W_i = 1, Y_i, \theta^{(t)})}{Pr(Y_i \mid Z_i = 1, W_i = 1)}
\]

\[
= \frac{Pr(Y_i \mid G_i = at, Z_i = 1, W_i = 1, Y_i, \theta^{(t)})Pr(G_i = at \mid Z_i = 1, W_i = 1)}{\sum_{g \in \{nt, at, co\}} Pr(Y_i \mid G_i = g, Z_i = 1, W_i = 1, Y_i, \theta^{(t)})Pr(G_i = g \mid Z_i = 1, W_i = 1)}
\]

\[
= \frac{p_{at}^{yi}(1 - p_{at})^{1 - y_i}q_{at}q_{co} + 0 + p_{co,t}^{yi}(1 - p_{co,t})^{1 - y_i}q_{at}q_{co}}{p_{at}^{yi}(1 - p_{at})^{1 - y_i}q_{at} + p_{co,t}^{yi}(1 - p_{co,t})^{1 - y_i}q_{co}}
\]

Note that, in the above, we take advantage of the fact that the conditional probability that the unit is a nevertaker is 0 when both $Z_i = 1$ and $W_i = 1$. Furthermore, we use calculations like the following:

\[
Pr(G_i = at \mid Z_i = 1, W_i = 1) = \frac{Pr(G_i = at, W_i = 1 \mid Z_i = 1)}{Pr(W_i = 1 \mid Z_i = 1)}
\]

\[
= \frac{Pr(W_i = 1 \mid G_i = at, Z_i = 1)Pr(G_i = at \mid Z_i = 1)}{\sum_g Pr(W_i = 1 \mid G_i = g, Z_i = 1)Pr(G_i = g \mid Z_i = 1)}
\]

\[
= \frac{q_{at}}{q_{at} + q_{co}}
\]

Terms such as $Pr(W_i = 1 \mid G_i = at, Z_i = 1)$ are either 0 or 1 and terms such as $Pr(G_i = g \mid Z_i = 1)$ can be simplified due to the independence of $G$ and $Z$ that is a consequence of our randomization scheme.

Compliers are analogous: for $Z_i = 1$ and $W_i = 1$ we have:
\[ Pr(G_i = co \mid Z_i = 1, W_i = 1, Y_i = y_i) = \frac{p_{co,t}(1 - p_{co,t})^{1-y_i}q_{co}}{p_{at}(1 - p_{at})^{1-y_i}q_{at} + p_{co,t}(1 - p_{co,t})^{1-y_i}q_{co}} \]

**Case 2:** \( Z_i = 0, W_i = 0, Y_i = y_i \)

Case 2 is completely analogous to Case 1 and the calculations do not bear repeating. Recall that, in the Philadelphia Divorce Study, Case 2 only occurs when \( Y_i = 1 \) since, in that data set, \( Y_i = 0 \) and \( Z_i = 0 \) implies that \( W_i \) is missing.

With computations similar to those above we obtain:

\[ Pr(G_i = nt \mid Z_i = 0, W_i = 0, Y_i = y_i) = \frac{p_{nt}(1 - p_{nt})^{1-y_i}q_{nt}}{p_{nt}(1 - p_{nt})^{1-y_i}q_{nt} + p_{co,t}(1 - p_{co,t})^{1-y_i}q_{co}} \]

and

\[ Pr(G_i = co \mid Z_i = 0, W_i = 0, Y_i = y_i) = \frac{p_{co,t}(1 - p_{co,t})^{1-y_i}q_{co}}{p_{nt}(1 - p_{nt})^{1-y_i}q_{nt} + p_{co,t}(1 - p_{co,t})^{1-y_i}q_{co}} \]

**Case 3:** \( Z_i = 0, W_i =?, Y_i = y_i \)

The cases where \( W_i \) is missing are the more interesting cases. In our actual dataset, Case 3 should only occur when \( Y_i = 0 \). However, we derive the general case here, in part because we also wish to simulate datasets whose missingness for \( W_i \) may be different from what is actually observed.

Since \( W_i \) is missing we don’t condition on it in our expectation calculations. These expressions end up being a bit more complex, since the compliance status takes all possible values. However, the spirit of these calculations is similar to the above.

Since we no longer may condition on \( W \), we start with an expression like the following:

\[ Pr(G_i = nt \mid Z_i = 0, Y_i, \theta(t)) = \frac{Pr(G_i = nt, Y_i \mid Z_i = 0, \theta(t))}{Pr(Y_i \mid Z_i = 0, \theta(t))} \]
Suppressing \( \theta \) in our expressions to avoid clutter, we have the following:

\[
Pr(G_i = nt \mid Z_i = 0, Y_i) = \frac{Pr(G_i = nt, Y_i \mid Z_i = 0)}{Pr(Y_i \mid Z_i = 0)}
\]

\[
= \frac{Pr(Y_i \mid G_i = nt, Z_i = 0)Pr(G_i = nt \mid Z_i = 0)}{\sum_{g \in \{nt, at, co\}} Pr(Y_i \mid G_i = g, Z_i = 0)Pr(G_i = g \mid Z_i = 0)}
\]

\[
= \frac{p_{nt}Y_i(1 - p_{nt})^{(1-Y_i)}q_{at}}{p_{nt}(1 - p_{nt})^{(1-Y_i)}q_{nt} + p_{at}(1 - p_{at})^{(1-Y_i)}q_{at} + p_{co,c}(1 - p_{co,c})^{(1-Y_i)}q_{co}}
\]

To obtain the expression for the denominator note that, by the law of total probability and since \( Pr(W_i = w \mid G_i = g, Z = 0) \) is always 0 or 1, we have:

\[
Pr(Y_i \mid Z_i = 0) = \sum_{w \in \{0,1\}} \sum_{g \in \{nt, at, co\}} Pr(Y_i \mid W_i = w, G_i = g, Z = 0)Pr(W_i = w, G_i = g \mid Z = 0)
\]

\[
= \sum_{w \in \{0,1\}} \sum_{g \in \{nt, at, co\}} Pr(Y_i \mid W_i = w, G_i = g, Z = 0)\times Pr(W_i = w \mid G_i = g, Z = 0)Pr(G_i = g \mid Z = 0)
\]

\[
= p_{nt}Y_i(1 - p_{nt})^{(1-Y_i)}q_{nt} + p_{at}(1 - p_{at})^{(1-Y_i)}q_{at} + p_{co,c}(1 - p_{co,c})^{(1-Y_i)}q_{co}
\]

Similarly,

\[
Pr(G_i = at \mid Z_i = 0, Y_i) = \frac{p_{at}Y_i(1 - p_{at})^{(1-Y_i)}q_{at}}{p_{nt}(1 - p_{nt})^{(1-Y_i)}q_{nt} + p_{at}(1 - p_{at})^{(1-Y_i)}q_{at} + p_{co,c}(1 - p_{co,c})^{(1-Y_i)}q_{co}}
\]

and

\[
Pr(G_i = co \mid Z_i = 0, Y_i) = \frac{p_{co,c}(1 - p_{co,c})^{(1-Y_i)}q_{co}}{p_{nt}(1 - p_{nt})^{(1-Y_i)}q_{nt} + p_{at}(1 - p_{at})^{(1-Y_i)}q_{at} + p_{co,c}(1 - p_{co,c})^{(1-Y_i)}q_{co}}
\]

**Case 4:** \( Z_i = 1, W_i = ?, Y_i = y_i \)

Case 4 does not occur at all in the Philadelphia Divorce Study. We include it for completeness and so that we can include datasets that experience Case 4 in simulation, if we so
desire. The derivations are completely analogous to Case 3. We could combine these cases,
but to avoid complicating the expressions (and since one occurs in the real dataset and the
other does not), we break them out separately:

\[
P_r(G_i = nt \mid Z_i = 1, Y_i) = \frac{p_{nt}^Y (1 - p_{nt}) (1 - Y_i) q_{nt}}{p_{nt}^Y (1 - p_{nt}) (1 - Y_i) q_{nt} + p_{at}^Y (1 - p_{at}) (1 - Y_i) q_{at} + p_{co,t}^Y (1 - p_{co,t}) (1 - Y_i) q_{co}}
\]

and

\[
P_r(G_i = at \mid Z_i = 1, Y_i) = \frac{p_{at}^Y (1 - p_{at}) (1 - Y_i) q_{at}}{p_{nt}^Y (1 - p_{nt}) (1 - Y_i) q_{nt} + p_{at}^Y (1 - p_{at}) (1 - Y_i) q_{at} + p_{co,t}^Y (1 - p_{co,t}) (1 - Y_i) q_{co}}
\]

and

\[
P_r(G_i = co \mid Z_i = 1, Y_i) = \frac{p_{co,t}^Y (1 - p_{co,t}) (1 - Y_i) q_{co}}{p_{nt}^Y (1 - p_{nt}) (1 - Y_i) q_{nt} + p_{at}^Y (1 - p_{at}) (1 - Y_i) q_{at} + p_{co,t}^Y (1 - p_{co,t}) (1 - Y_i) q_{co}}
\]

**E-Step**

The E-step merely requires taking the expectation of the complete data likelihood at
each step, conditional on the observed data and the current iteration value of \(\theta\). Since the
missing data appears linearly in the log-likelihood this amounts to replacing terms such as,
for example, \(I(G_i = nt)\) with its expectation \(E(I(G_i = nt \mid Z_i, W_i, Y_i, \theta^{(t)}) = Pr(G_i = nt \mid Z_i, W_i, Y_i, \theta^{(t)})\) and so forth for each i. Conditional on knowing \(\theta\), i.e. conditional on
\(\theta^{(t)}\), these expectations are all just numerical constants. We therefore denote \(E(I(G_i = nt \mid Z_i, W_i, Y_i, \theta^{(t)}) = k_{nt,i}^{(t)}\). The E-Step thus yields the following expectation for the complete
data log likelihood:
\( E(\ell_{\text{comp}}(\theta) \mid Y, W, Z, \theta^{(t)}) = \sum_{i=1}^{n} k_{nt,i}^{(t)} [y_i \log(p_{nt}) + (1 - y_i) \log(1 - p_{nt})] \\
+ k_{at,i}^{(t)} [y_i \log(p_{at}) + (1 - y_i) \log(1 - p_{at})] \\
+ k_{co,i}^{(t)} [y_i \log(p_{co,t}) + (1 - y_i) \log(1 - p_{co,t})] \\
+ k_{co,i}^{(t)} (1 - z_i) [y_i \log(p_{co,c}) + (1 - y_i) \log(1 - p_{co,c})] \\
+ k_{nt,i}^{(t)} \log(q_{nt}) + k_{at,i}^{(t)} \log(q_{at}) + k_{co,i}^{(t)} \log(1 - q_{nt} - q_{at}) \)

**M-Step**

In the M-step, we find the parameter value \( \theta \) that maximizes E-step expression in order to obtain the next iteration value of \( \theta^{(t+1)} \). This is a basic calculus problem where we set the gradient to 0 and solve simultaneously for each component of the parameter vector. We omit most of the calculations here except for a quick example for the case of nevertakers. Note that only a couple of terms survive when we take the partial derivative with respect to \( p_{nt} \), the rest of the terms being constant in \( p_{nt} \):

\[
0 = \frac{\partial}{\partial p_{nt}} E(\ell_{\text{comp}}(\theta) \mid Y, W, Z, \theta^{(t)}) = \frac{\partial}{\partial p_{nt}} \sum_{i=1}^{n} k_{nt,i}^{(t)} [y_i \log(p_{nt}) + (1 - y_i) \log(1 - p_{nt})] + C
\]

\[
0 = \sum_{i=1}^{n} k_{nt,i}^{(t)} \left[ \frac{y_i}{p_{nt}} - \frac{(1 - y_i)}{(1 - p_{nt})} \right]
\]

\[
p_{nt} \sum_{i=1}^{n} k_{nt,i}^{(t)} (1 - y_i) = (1 - p_{nt}) \sum_{i=1}^{n} k_{nt,i}^{(t)} y_i
\]

\[
p_{nt} \sum_{i=1}^{n} k_{nt,i}^{(t)} = \sum_{i=1}^{n} k_{nt,i}^{(t)} y_i
\]

\[
\hat{p}_{nt} = \frac{\sum_{i=1}^{n} k_{nt,i}^{(t)} y_i}{\sum_{i=1}^{n} k_{nt,i}^{(t)}}
\]

This makes intuitive sense. The usual estimate for a binomial parameter is the number of successes over the number of trials. In this case, we have the expected number of successes
over the expected number of trials. The M-step results for the other compliance types are completely analogous.

### 3.7.5 Data Augmentation Algorithm

Taking a Bayesian approach, we can also use the DA algorithm to sample from the posterior distribution of the missing potential outcomes conditional on the observed data.

The Data Augmentation algorithm has a very similar flavor to EM, except it additionally allows us to calculate the entire posterior distribution of interest, rather than simply the MLE point estimate. We will again augment the observed data with the latent compliance statuses so that we still use the complete data likelihood. Rather than taking expectations over the missing data at each step, we will sample/impute the missing data using the conditional marginal posterior, as in a Gibbs sampler.

The algorithm does the following at the $j^{th}$ step:

1. Begin with $g_j(\theta)$, the approximation of the posterior for $\theta$ at the $j^{th}$ step.

2. For each unit, $i$, impute the missing data, $G_i$, from the posterior predictive distribution of $G_i$ given the observed data $Y_i, W_i, Z_i$. In practice, to sample $G_i(j) \sim p(G_i \mid Y_i, W_i, Z_i)$ we first sample $\theta(j) \sim g_j(\theta)$ and then sample $G_i(j) \sim p(G_i \mid Y_i, W_i, Z_i, \theta(j))$

3. Draw the missing potential outcomes from their posterior predictive distribution. In the previous step we sampled $G_i(j) \sim p(G_i \mid Y_i, W_i, Z_i)$ for each $i$. We now sample $(Y_i(1), Y_i(0)) \sim f(Y_i(1), Y_i(0) \mid G_i(j), Y_i, W_i, Z_i)$.

4. Compute the $j^{th}$ draw of $\tau$, the causal estimand, which is typically a simple deterministic function of the potential outcomes. For example, $\tau_{fp}^{(j)} = \frac{1}{N_{co}} \sum_{i \mid G_i(j) = co} Y_i(1)^{(j)} - Y_i(0)^{(j)}$

5. Update the posterior: $g_{j+1} = p(\theta \mid Y^{(j)}, W^{(j)}, Z^{(j)}, G^{(j)})$.

6. In the next iteration, we will draw $\theta^{(j)} \sim g_{j+1}(\theta)$, and so on.
At the first iteration, we can just start with some arbitrary values for $\theta$ or draw $\theta$ from its prior. We now explain these various steps in some detail below for the no-covariates model.

**The Prior**

We select the prior on the parameter vector $\theta = (p_{co,t}, p_{co,c}, p_{at}, p_{nt}, q)$ such that the prior components are a priori independent, flat, and conjugate. This simplifies the Gibbs sampler. In this case we have:

$$p(p_{co,t}, p_{co,c}, p_{at}, p_{nt}, q) = p(p_{co,t})p(p_{co,c})p(p_{at})p(p_{nt})p(q)$$

i.e. prior independence of the binomial parameters and the multinomial parameter. We choose flat conjugate priors for these various parameters:

- $p_{co,t} \sim Beta(1, 1)$
- $p_{co,c} \sim Beta(1, 1)$
- $p_{at} \sim Beta(1, 1)$
- $p_{nt} \sim Beta(1, 1)$
- $q \sim Dirichlet(1, 1, 1)$

Alternatively, one can choose a pseudo observations prior, which essentially adds some number of pseudo-observations to the likelihood. We have implemented both priors and results are not sensitive to this choice. We present the results of the model with the flat prior for simplicity of exposition. In a model without covariates such as this, a pseudo observations prior will be slightly more informative than a flat prior but the parameter vector components will still have a beta-binomial and dirichlet-multinomial conjugacy. We add pseudo-observations that correspond to each possible combination of the treatment assignment, compliance statuses, outcomes and covariates. In the case of no covariates, this...
corresponds to $2 \times 2 \times 3$ combinations, since there are two possible treatment assignments, two possible outcome values and three possible compliance statuses. Especially when there are covariates that take many possible combinations of values, the number of pseudo-observations can grow quickly. We therefore set the total weight of the pseudo-observations such that they are equivalent to $N_a$ additional data points.

In a dataset of this size where there are only 311 observations, we set $N_a = 5$, so as to limit the influence of the prior.

For greater detail on the derivation of the pseudo observations prior, see Appendix C.

**Imputing $G$ conditional on the observed data and the current value of $\theta^{(i)}$**

Conditional on the parameter values and the observed data, we can impute each $G_i$.

We have already derived the necessary conditional distribution for $G_i$ in our derivation of the E-step in the EM algorithm. We did it separately for each possible observed (and missing) values of $Z_i$, $W_i$, and $Y_i$. For example, recall that, for the case where $Z_i = 0$ and missing $W_i$ we derived the following:

\[
Pr(G_i = nt | Z_i = 0, Y_i) = \frac{p_{nt}^{Y_i}(1 - p_{nt})^{(1-Y_i)}q_{nt}}{p_{nt}^{Y_i}(1 - p_{nt})^{(1-Y_i)}q_{nt} + p_{at}^{Y_i}(1 - p_{at})^{(1-Y_i)}q_{at} + p_{co,c}^{Y_i}(1 - p_{co,c})^{(1-Y_i)}q_{co}}
\]

\[
Pr(G_i = at | Z_i = 0, Y_i) = \frac{p_{at}^{Y_i}(1 - p_{at})^{(1-Y_i)}q_{at}}{p_{nt}^{Y_i}(1 - p_{nt})^{(1-Y_i)}q_{nt} + p_{at}^{Y_i}(1 - p_{at})^{(1-Y_i)}q_{at} + p_{co,c}^{Y_i}(1 - p_{co,c})^{(1-Y_i)}q_{co}}
\]

\[
Pr(G_i = co | Z_i = 0, Y_i) = \frac{p_{co,c}^{Y_i}(1 - p_{co,c})^{(1-Y_i)}q_{co}}{p_{nt}^{Y_i}(1 - p_{nt})^{(1-Y_i)}q_{nt} + p_{at}^{Y_i}(1 - p_{at})^{(1-Y_i)}q_{at} + p_{co,c}^{Y_i}(1 - p_{co,c})^{(1-Y_i)}q_{co}}
\]

These probabilities define a multinomial distribution for each $G_i$ that we can use to impute $G_i$, conditional on $Z_i$, $Y_i$ and $\theta^{(i)}$. We have an analogous multinomial distribution for the case when $Z_i = 1$. When $W_i$ is observed, then $G_i$ is either determined or we impute it from the appropriate binomial distribution, again reusing the conditional distributions for $G_i$ that we derived for the EM algorithm.
Appendix B has details on the Gibbs sampling draws of the compliance status for the sequential binomial model.

**Imputing the missing potential outcomes, \(Y_{mis}\), and drawing the causal estimand, \(\tau\)**

Once we have sampled \(\theta^{(i)}\) and the compliance statuses, imputing \(Y_{i}^{mis}\) is trivial. We can impute \(Y_{i}^{mis}\) with a simple Bernoulli draw. For example, \(Y_{i}^{mis} | G_i = nt, p_{nt}^{(i)} \sim Bern(p_{nt}^{(i)})\).

To draw the causal estimand \(\tau\), the complier average causal effect, we directly calculate it from the observed and missing potential outcomes for compliers in the sample, \(\tau = \frac{1}{N_{co}} \sum_{i: G_i = co} Y_i(1) - Y_i(0)\), where \(N_{co}\) is the number of compliers in the sample.

**Updating the Posterior**

Finally, we update the parameter vector, i.e. we obtain a new draw from the posterior of \(\theta\).

We update the posterior from \(g_i(\theta)\) to \(g_{i+1}(\theta)\) using standard conjugacy rules. The posterior for components of \(\theta\) will be either beta or dirichlet.

Specifically, at iteration \(i\), once we have imputed the compliance statuses, let’s denote the following:

- Let \(n_{1,co}^{(i)}\) be the number of successes for compliers who receive control in the imputed dataset and \(n_{0,co}^{(i)}\) the number of failures
- Let \(n_{1,co}^{(i)}\) be the number of successes for compliers who receive treatment in the imputed dataset and \(n_{0,co}^{(i)}\) the number of failures
- Let \(n_{1,at}^{(i)}\) be the number of successes for alwaystakers in the imputed dataset and \(n_{0,at}^{(i)}\) the number of failures
- Let \(n_{1,nt}^{(i)}\) be the number of successes for nevertakers in the imputed dataset and \(n_{0,nt}^{(i)}\) the number of failures
Then we update our posterior $g_i(\theta)$ by drawing the next iteration values as follows:

$$
\begin{align*}
  p_{i+1}(co, c) & \sim \text{beta}(1 + n_{1, co}^{(i)}, 1 + n_{0, co}^{(i)}) \\
  p_{i+1}(co, t) & \sim \text{beta}(1 + n_{1, co}^{(i)}, 1 + n_{0, co}^{(i)}) \\
  p_{i+1}(at) & \sim \text{beta}(1 + n_{1, at}^{(i)}, 1 + n_{0, at}^{(i)}) \\
  p_{i+1}(nt) & \sim \text{beta}(1 + n_{1, nt}^{(i)}, 1 + n_{0, nt}^{(i)}) \\
  q & \sim \text{Dirichlet}(1 + n_{1, co}^{(i)} + n_{0, co}^{(i)}, 1 + n_{1, at}^{(i)} + n_{0, at}^{(i)}, 1 + n_{1, nt}^{(i)} + n_{0, nt}^{(i)})
\end{align*}
$$

Posterior updates when assuming a pseudo observations prior are similar.

### 3.8 Simulation Studies

We wish to investigate the performance of the proposed maximum likelihood and Bayesian methods for different levels of missingness in the treatment vector $W$. For each method, we can simulate datasets where components of $W$ are missing and observe how well we estimate the estimand on average. In particular, we can check the average bias of the MLE for the maximum likelihood approach. Similarly, we can check various summaries of the posterior distribution of the estimand for the Bayesian approach. Such summaries of the posterior include the bias in the posterior mean, median or mode, the frequency coverage of the posterior intervals, the average width of the posterior intervals, and so forth.

We check the performance of these algorithms under two different missing data mechanisms. We want to check the performance in the simplest case, which is when the elements of $W$ are missing completely at random (MCAR). We call these the MCAR simulations. The data are missing completely at random when the missingness does not depend on the missing or observed data (see generally Little & Rubin 1987).

Secondly, we wish to investigate the performance of these methods when the elements of $W$ are missing as they are in the Philadelphia Divorce dataset, since this is the motivating example for development of this methodology. In other words, we want to investigate per-
formance for the case when \( W_i \) is missing if and only if \( Z_i = 0 \) and \( Y_i = 0 \). We call these the Philadelphia Divorce Study simulations.

We present simulation results for the following data generating parameter value.

\[
\theta = (p_{co,t}, p_{co,c}, p_{at}, p_{nt}, q_{co}, q_{at}, q_{nt}) = (0.32, 0.29, 0.4, 0.1, 0.6, 0.2, 0.2)
\]

These values were chosen to see if we could recover a relatively mild treatment effect when the success rates between the compliance groups were meaningfully different. The success probabilities reflect the realistic possibility that alwaystakers fare better than compliers, on average, and that compliers fare better than nevertakers. The parameters that govern the relative prevalence of the compliance statuses are chosen so that compliers are quite common, as we would expect them to be in experiments such as the Philadelphia Divorce Study.

As we will see below, both methods perform very well when data is missing completely at random from the vector of treatment indicators.

### 3.8.1 MCAR Missingness Simulation Results

The MCAR simulations are conducted as follows: First, select true values for the data generating parameters. Next, for \( n = 1, ..., 300 \) and for \( i = 1, ..., K \), do the following:

- Generate the \( i^{th} \) dataset of 300 units, \( D_i \). In each dataset \( D_i \), \( n \) values of the treatment vector \( W \) are missing completely at random.

- Perform the algorithm (either EM or DA) on \( D_i \) and store the relevant statistics (e.g. the MLE or summaries of the posterior distribution)

Lastly, we average the results for each value of \( n \) over the \( K \) datasets. The choice of \( K \) will depend on the choice of algorithm, since the DA Algorithm is much more computationally intensive than the EM algorithm for this particular problem. We choose \( K = 100 \) for the DA Algorithm and \( K = 1000 \) for the EM algorithm.
We will see that both the MLE estimates found by the EM algorithm and the posterior mean calculated by the DA algorithm are unbiased for the causal estimand, even for very large amounts of missing data. Additionally, the credible intervals produced by the DA algorithm achieve the nominal coverage, except when the proportion of missing data gets very close to 1, in which case the intervals overcover. Additionally, the MLE and the posterior mean also tend to have similarly nice bias properties for the nuisance parameters. The posterior intervals also nice coverage properties for the nuisance parameters.

EM Algorithm

On average, the MLE recovers the data generating parameters until n, the amount of missing data, becomes too large, at which point all compliance information vanishes (see Figures 3.2, 3.3 and 3.4).

![MLE Performance vs Amount of Missing Data (MCA)](image)

Figure 3.2: Average MLE of super population estimand
Figure 3.3: Average MLE of success probabilities

Figure 3.4: Average MLE of Compliance Probabilities
In Figures 3.2, 3.3 and 3.4, the red line represents the true value of the data generating parameter. The x-axis is the value of n, the number of units with missing treatment values, and the y-axis is the average MLE for the parameter, averaged over the simulated datasets. The MLE does remarkably well in recovering the true parameter values, on average, even up to the point when around 250 of the 300 treatment values are missing. As the proportion of missing values approaches 1, we lose all compliance information and the estimates are essentially random – their averages values are largely a function of the random initialization values in the EM algorithm and the overall success rate.

DA Algorithm

We can begin our evaluation of the DA Algorithm’s performance by looking at the coverage of our posterior intervals for the finite population and super population causal estimands. The coverage is very close to the nominal coverage (see Figure 3.5).

Figure 3.5: Average Coverage of Posterior CIs for causal effects
The intervals tend to overcover as our compliance information vanishes (as the proportion of missing compliance data approaches 1). This overcoverage is due to the fact that the flat prior dominates. The choice of such a diffuse prior was conservative and our posterior interval essentially covers the parameter space. Regardless of the amount of missing compliance data, we tend to achieve the nominal coverage and, at worst, we have slight overcoverage. Overcoverage can be avoided by not attempting to estimate the CACE when there is a vanishingly small amount of compliance information.

In Figure 3.6, we plot the average posterior mean and average posterior interval endpoints for the causal estimands.

![Figure 3.6: Average Posterior Mean/Quantiles for causal effects](image)

The posterior intervals are quite wide even when there is no missing data. This suggests that, for small causal effects, it may make practical sense to leverage covariates for increased precision.

Leaving the causal estimand to the side, momentarily, we can also look at the frequency
coverage of posterior intervals for the parameters. The parameter vector consists of nuisance parameters that we integrate out when we perform inference on the causal estimands of interest (the exception is when the super population $\tau_{sp} = p_{co,t} - p_{co,c}$ is the causal estimand but, even in this case, we don’t care about the parameters $p_{co,t}$ and $p_{co,c}$ so much as we care about their difference). Nevertheless, it can be interesting to understand the properties of the posterior distribution of the parameters. We separately consider the success probabilities, $p$, and the compliance status probabilities $q$.

Looking at Figure 3.7, we see that, on average, the frequency coverage for success parameters in $p$ is the nominal coverage when the proportion of missing data is not too large. Similar to the posterior intervals for the estimand, these intervals tend to get large and over-cover when the proportion of missingness in the treatment vector, $W$, approaches 1.

![Figure 3.7: Average Coverage for Posterior Intervals for $p_i$](image)

Turning to the coverage properties of posterior intervals for components of $q$, the parameter that controls the relative frequency of compliers, always-takers, and nevers-takers, we
see in Figure 3.8 that the posterior interval for $q_{co}$ tends to undercover when we lose all compliance info.

![Figure 3.8: Average Coverage for Posterior Intervals for $q_i$](image)

For these particular parameter values, most units are expected to be compliers but there is no way to know this without any compliance information. The intervals are overly confident, whereas intervals for $q_{at}$ and $q_{nt}$ are relatively well-behaved.

The posterior means for the success probability parameters, $p_i$, on average recover the true parameter values for reasonable amounts of missing data (see Figure 3.9). The credible intervals tend to be wide, due to large amount of missing data. Again, we could increase precision by leveraging covariates if one of these parameters were the actual estimand of interest. The exception is the posterior mean for the nevertaker success probability. The posterior mean tends to be a biased estimator of the parameter, due to skew in this distribution. For these parameters, the posterior median or mode might be a better choice as a measure of centrality of the posterior, if we wish to reduce the posterior distribution to a single numerical summary. It isn’t necessary to agonize over the choice of, e.g. the posterior mean versus the posterior median since (a) the MCMC algorithm gives us the full posterior, avoiding the need to choose a univariate statistic as an estimator and (b) for our purposes, we need only check that the proposed algorithm behaves as claimed.
Figure 3.9: Average Posterior Means and CI Endpoints for $p_i$

Figure 3.10 reveals a similar problem with the posterior mean for the compliance proportions, $q_i$ – there is some bias for large amounts of missing data due to skew. The credible intervals are very wide, especially as the compliance data disappears. When all of the compliance information vanishes, the posterior distribution for each component essentially represents random guessing in the parameter space, since the prior is so weak.

Figure 3.10: Average Posterior Means and CI Endpoints for $q_i$
In conclusion, for MCAR simulations, both the EM and DA algorithm perform well in estimating the CACE. The credible intervals for the CACE obtained from the DA algorithm achieve the nominal coverage on average. All methods perform poorly when all or almost all of the data is missing, although the intervals still do not undercover. Of course, no method will perform well when all of the compliance data is missing.

### 3.8.2 Philadelphia Divorce Study Missingness Simulation Results

Simulation results are encouraging when the treatment indicators are missing completely at random (MCAR). However, we wish to know how these methods perform when the treatment indicators are missing in the highly structured way in which the data is missing in the Philadelphia Divorce Study. In this set of simulations, the missing data mechanism is the same mechanism that we observe in the Philadelphia Divorce Study, i.e. $W_i$ is missing if and only if $Y_i = 0$ and $Z_i = 0$.

These simulations are somewhat different from the MCAR simulations in the sense that we can’t simply pick the desired amount of missingness, $n$, and then select (at random) which $n$ components of the $W$ vector will be missing. Rather, the number of missing observations is completely determined by the values that are realized in the $Z$ and $Y$ vectors. Missingness in $W$ occurs when $Y_i = 0$ and $Z_i = 0$. The amount of missingness is therefore a random variable that depends on the data generating process. Note, in particular, that the number of missing observations can never exceed $N_c$, the number of observations that are randomized to the control group of the experiment.

For simulations that mirror the missing data mechanism in the Philadelphia Divorce Study, we first fix some parameter values and then simulate 20,000 datasets using these fixed parameters.

We will see that maximum likelihood methods begin to fail, in large part due to multimodality in the likelihood function. The DA algorithm still performs well – the posterior tends to be centered on the true value of the estimand. Even in the case where there are
highly skewed or multimodal posteriors, the intervals tend to have correct coverage on average.

**EM Algorithm**

Table 3.4 contains the MLE average results for simulations that mimic the missingness pattern of the Philadelphia Divorce Study. The MLE tends to have some bias for datasets of this size, which indicates that the assumptions that are usually relied on for maximum likelihood estimation are failing here. In particular, the MLE performs poorly because of the multimodality of the likelihood function (see below). In particular, $q_{co}$ and $q_{at}$ are poorly identified when small numbers of alwaystakers are observed, and this phenomenon generates multimodality in their respective profile likelihood functions.

Table 3.4: MLE Simulation Results for Simulated Data with Philadelphia Divorce Study Missingness

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Value</th>
<th>MLE (mean value)</th>
<th>MLE (modal value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_{sp}$</td>
<td>0.030</td>
<td>0.040</td>
<td>0.039</td>
</tr>
<tr>
<td>$p_{co,c}$</td>
<td>0.290</td>
<td>0.455</td>
<td>0.282</td>
</tr>
<tr>
<td>$p_{co,t}$</td>
<td>0.320</td>
<td>0.495</td>
<td>0.319</td>
</tr>
<tr>
<td>$p_{at}$</td>
<td>0.400</td>
<td>0.319</td>
<td>0.161</td>
</tr>
<tr>
<td>$p_{nt}$</td>
<td>0.100</td>
<td>0.010</td>
<td>0.070</td>
</tr>
<tr>
<td>$q_{co}$</td>
<td>0.600</td>
<td>0.450</td>
<td>0.251</td>
</tr>
<tr>
<td>$q_{at}$</td>
<td>0.200</td>
<td>0.350</td>
<td>0.541</td>
</tr>
<tr>
<td>$q_{nt}$</td>
<td>0.200</td>
<td>0.200</td>
<td>0.200</td>
</tr>
</tbody>
</table>

**DA Algorithm**

Table 3.5 contains average simulation results across 20,000 simulated datasets where the missing data mechanism is identical to the missing data mechanism in the Philadelphia Divorce Study. We use the same parameter values as in the MCAR simulations. In general, the credible intervals either achieve the nominal coverage or overcover. On average, the posterior mean for $\tau_{fp}$ is the same as its true (expected) value. Note that, in Table 3.5, the true value of $\tau_{fp}$ is starred to indicate that this is the expected value over many datasets.
The nuisance parameters, while they have relatively well behaved posterior intervals, have posterior means, medians and modes that can differ significantly from the true parameter values, even when averaged over many datasets. This is related to the problems that we see for the MLE and underscores the benefits of using the entire posterior distribution, rather than reducing our inference to a single summary statistic and the behavior of that statistic. For these reasons, these statistics probably have undesirable properties as point estimators for their associated parameters.

Table 3.5: DA Simulation Results for Simulated Data with Philadelphia Divorce Study Missingness

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Value</th>
<th>Mean</th>
<th>CI lower</th>
<th>CI upper</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_{fp}$</td>
<td>0.03$^*$</td>
<td>0.03</td>
<td>-0.25</td>
<td>0.31</td>
<td>0.95</td>
</tr>
<tr>
<td>$\tau_{sp}$</td>
<td>0.03</td>
<td>0.03</td>
<td>-0.24</td>
<td>0.31</td>
<td>0.95</td>
</tr>
<tr>
<td>$p_{co,c}$</td>
<td>0.29</td>
<td>0.47</td>
<td>0.19</td>
<td>0.90</td>
<td>0.98</td>
</tr>
<tr>
<td>$p_{co,t}$</td>
<td>0.32</td>
<td>0.50</td>
<td>0.21</td>
<td>0.94</td>
<td>0.98</td>
</tr>
<tr>
<td>$p_{at}$</td>
<td>0.40</td>
<td>0.30</td>
<td>0.10</td>
<td>0.88</td>
<td>1.00</td>
</tr>
<tr>
<td>$p_{nt}$</td>
<td>0.10</td>
<td>0.13</td>
<td>0.04</td>
<td>0.26</td>
<td>0.95</td>
</tr>
<tr>
<td>$q_{co}$</td>
<td>0.60</td>
<td>0.42</td>
<td>0.18</td>
<td>0.72</td>
<td>0.997</td>
</tr>
<tr>
<td>$q_{at}$</td>
<td>0.20</td>
<td>0.37</td>
<td>0.09</td>
<td>0.62</td>
<td>0.999</td>
</tr>
<tr>
<td>$q_{nt}$</td>
<td>0.20</td>
<td>0.21</td>
<td>0.15</td>
<td>0.27</td>
<td>0.95</td>
</tr>
</tbody>
</table>

In order to gain intuition for why the MLE performs poorly but why the DA algorithm still performs well, we can look at an example data set.

3.8.3 Philadelphia Divorce Study Missingness – Example

Simulated Dataset

We can look at an example dataset, generated with the same parameter values as in the simulation studies above, to assess the performance of the DA Algorithm and EM on this particular dataset. Table 3.6 summarizes the example dataset.

The true finite population complier average causal effect in this dataset is $\tau_{fp} = 0.01$. 
Table 3.6: Example Dataset

<table>
<thead>
<tr>
<th>$S(z, w, y)$</th>
<th>$G_i$</th>
<th>$Z_i$</th>
<th>$W_i$</th>
<th>$Y_i$</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S(0, ?, 0)$</td>
<td>{co, nt, at}</td>
<td>0</td>
<td>?</td>
<td>0</td>
<td>108</td>
</tr>
<tr>
<td>$S(0, 0, 1)$</td>
<td>{co, nt}</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>28</td>
</tr>
<tr>
<td>$S(0, 1, 1)$</td>
<td>at</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>$S(1, 0, 0)$</td>
<td>nt</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>34</td>
</tr>
<tr>
<td>$S(1, 0, 1)$</td>
<td>nt</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>$S(1, 1, 0)$</td>
<td>{co, at}</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>71</td>
</tr>
<tr>
<td>$S(1, 1, 1)$</td>
<td>{co, at}</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>38</td>
</tr>
</tbody>
</table>

**EM Algorithm – Example Dataset**

Table 3.7: Mean Simulation Results for Simulated Data with Philadelphia Divorce Study Missingness

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True Value</th>
<th>MLE</th>
<th>Posterior Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_{sp}$</td>
<td>0.030</td>
<td>0.068</td>
<td>0.039</td>
</tr>
<tr>
<td>$p_{co,c}$</td>
<td>0.290</td>
<td>0.478</td>
<td>0.273</td>
</tr>
<tr>
<td>$p_{co,t}$</td>
<td>0.320</td>
<td>0.547</td>
<td>0.307</td>
</tr>
<tr>
<td>$p_{at}$</td>
<td>0.400</td>
<td>0.215</td>
<td>0.201</td>
</tr>
<tr>
<td>$p_{nt}$</td>
<td>0.100</td>
<td>0.171</td>
<td>0.172</td>
</tr>
<tr>
<td>$q_{co}$</td>
<td>0.600</td>
<td>0.293</td>
<td>0.202</td>
</tr>
<tr>
<td>$q_{at}$</td>
<td>0.200</td>
<td>0.434</td>
<td>0.509</td>
</tr>
<tr>
<td>$q_{nt}$</td>
<td>0.200</td>
<td>0.273</td>
<td>0.274</td>
</tr>
</tbody>
</table>

The MLE estimates of the data generating parameters are pretty poor, as we might expect after observing the simulation results above. The MLE for the super population causal effect is reasonably close to the true value, when we consider its standard error. A 95% confidence interval for $\tau_{sp}$ is [-0.295, 0.392], which comfortably contains the true parameter value.

**DA Algorithm – Example Dataset**

The posterior median, mean, and mode for the finite populate causal estimand are similar, taking the values 0.047, 0.047 and 0.046, respectively. The credible interval for $\tau_{fp}$, (-0.29, 0.38) is quite large, reflecting the large amount of uncertainty in the estimate. Figure 3.11 is a histogram plot of draws from the marginal posterior distribution of $\tau_{fp}$.

The green line marks the true value of $\tau_{fp}$, the red line is the median, the purple line
is the mean, and the blue line is the mode (note that they are quite similar to each other). In this dataset, 108 out of 300 of the treatment indicators are missing (and 245 of the 300 compliance statuses), and the posterior distribution of $\tau_{fp}$ is a nice, unimodal bell-shaped distribution. The posterior median or mode would both be reasonable point estimates.

The marginal posterior for the super population causal estimand is quite similar to the finite population estimand (see Figure 3.12), as might be expected.

The marginal posteriors of the success probabilities are less well-behaved, in part because they can exhibit significant skew, as can be seen in Figure 3.13. Again the credible intervals are quite wide and again the intervals cover the true value, in this one example dataset.

The marginal posteriors for the proportions of compliers, alwaystakers and nevertakers are still less well-behaved (see Figure 3.14).

We observe multimodal posteriors for alwaystaker and complier probabilities, and the posterior mode seems to find the wrong mode for each in this example. It is difficult to know
Figure 3.12: Marginal posterior for $\tau_{sp}$

Figure 3.13: Marginal posterior for $p_i$
the relative prevalence of compliers and always-takers because of the structure of the missing data mechanism. The treatment indicator is missing whenever \( Y_i = 0 \) and \( Z_i = 0 \) and we therefore can’t observe which units are always-takers in the control group whenever \( Y_i = 0 \). Estimating the relative prevalence of each compliance type is difficult for this reason. Indeed, our ability to estimate the CACE via an IV approach is subverted because we can’t construct the usual estimate for the proportion of always-takers in the sample (see Section 3.8.4). At least in the Bayesian approach, however, we can integrate out these nuisance parameters and still get nice estimates for the posterior distributions of the causal estimands.

We see that many of the nuisance parameters have skewed or multimodal marginal posteriors, which reflects skew and multimodality in the likelihood function. We also see the credible intervals appear to perform relatively well. In the following section we explain why instrumental variables and maximum likelihood methods fail whereas the full Bayesian method succeeds.
3.8.4 Why the Bayesian Approach Works Well for the Philadelphia Divorce Study

We have seen that maximum-likelihood and full Bayesian methods work well when data is missing completely at random from the vector of treatment indicators. A variant of the instrumental variables approach would also work for MCAR missingness—while inefficient, you could always use an IV approach on the complete cases. However, for Philadelphia Divorce Study missingness, a MAR missing data mechanism, both the IV and maximum-likelihood approaches fail, whereas the full Bayesian method that employs the DA algorithm succeeds.

We first detail the failure of IV and maximum likelihood and then explain the success of the DA algorithm.

Failure of the Instrumental Variables Approach

One common approach to estimating the complier average causal effect (valid under the monotonicity assumption and exclusion restriction assumptions for nevertakers and always-takers) is an instrumental variables approach. Under the assumptions just mentioned, one can show that $\frac{ITT_Y}{ITT_W} = \text{CACE}$. Here, $ITT_Y$ denotes the intention-to-treat effect of the random assignment on the outcome and $ITT_W$ denotes the intention-to-treat effect of the random assignment on treatment.

The usual estimator is a simple method-of-moments estimator that plugs in unbiased estimators for $ITT_Y$ and $ITT_W$. Namely, $\hat{\tau}^{iv} = \frac{ITT_Y}{ITT_W}$. Now, the estimator $\hat{ITT}_W = W_{1}^{obs} - W_{0}^{obs}$ is unavailable in the case when we have missing entries in the $W$ vector, the vector of treatment indicators, as is the case in the Philadelphia Divorce Study. In the case where the missing $W_i$ values are missing completely at random, we could simply ignore the missing data and still have an unbiased estimator $\hat{ITT}_W$ based solely on the observed values. We would thus be able to salvage the IV approach in this case.

Let’s now consider the case where we are unwilling to assume that the data is MCAR.
Indeed, we are unwilling to assume anything about the missing data mechanism. In this case, we would be able to calculate bounds on $\text{ITT}_W$ but not much else.

For example, suppose for the sake of simplicity that the Philadelphia Divorce Study had been a completely randomized experiment (rather than block randomized). Using this simplified version of the data set as an example, and using the primary outcome (filing a divorce lawsuit) as our outcome, we have:

$$\text{ITT}_Y = \bar{Y}_1^{\text{obs}} - \bar{Y}_0^{\text{obs}} = 0.249$$

However, we cannot calculate the estimate for $\text{ITT}_W = \bar{W}_1^{\text{obs}} - \bar{W}_0^{\text{obs}}$ due to missing values. Indeed, the usual instrumental variables estimate for the proportion of alwaystakers, $\frac{1}{N_e} \sum_{i:Z_i=0} I_{W_i=1}$, can’t be computed as usual. When values are missing, this quantity can take a range of values depending on which of the missing $W_i$ values are zeros and which are ones.

It is easy to check that for this example, the range of possible values for $\text{ITT}_W$ lie in the interval $[0.083, 0.724]$. This in turn puts bounds on the possible values for the instrumental variables CACE estimator, $\hat{\tau}^{iv}$. Such bounds for the Philadelphia Divorce Study are $[0.344, 3.02]$. Note that these bounds can be nonsensical, as the upper bound is in this case. Indeed, if the bounds for $\text{ITT}_W$ have different signs then the bounds for $\hat{\tau}^{iv}$ are $(-\infty, \infty)$, which tells us nothing. Moreover, we shouldn’t interpret the bounds to mean that $\tau_{CACE}$ is at least 0.344. We only have bounds for the estimate, not the true, unknown value of the estimand. Constructing a confidence interval for $\tau_{CACE}$ when we only have a range of values for $\hat{\tau}_{CACE}$ would present an additional challenge which we do not attempt to solve here.

This problem is even worse in the actual Philadelphia Divorce Study with its block randomized structure. Some of the blocks have only missing values in the $W_0$ vector, the vector of treatment indicators for control units. This unfortunate scenario means that all logically possible values of $\bar{W}_0$ (and, hence, $\text{ITT}_W$) are possible within the block, such that the best bounds we can obtain are $(-\infty, \infty)$. Other blocks are nearly as troublesome.
Modeling assumptions seem required, therefore, to even bound the CACE for this particular data set.

In fact, we can do much better than simply writing down bounds. The bounds for an estimator simply define the range of possible values, not the relative probability of those values. If all possible values for the missing data were equally likely, we would be stuck with an estimate at least as imprecise as the bounds on the instrumental variables estimator. Leaving aside the method-of-moments approach and the $\text{ITT}_W$, we can say more about the quantity of direct interest, the CACE, by using the rest of the information at our disposal, rather than limiting ourselves to $\text{ITT}_Y$ and the bounds on $\text{ITT}_W$.

In particular, we have complete knowledge of the missing data mechanism. Additionally, we have posed the likelihood, which encapsulates the information about the parameters that is contained in the data. A principled approach would impute the missing $W_i$ values using all of the information at hand. Because we are willing to make assumptions about the likelihood and because the missing data mechanism for $W^{\text{obs}}$ is missing at random, two principled ways to do this are data augmentation methods such as the EM algorithm or the DA algorithm.

One might be suspicious of these approaches – we’ve already seen that EM produces biased estimates in simulation for the kind of missingness that occurs in the Philadelphia Divorce Study (although it works well when the data is missing MCAR). The DA Algorithm works well for estimation of the causal estimands in the simulated settings but perhaps, given the failure of EM, there actually isn’t much extra information in the likelihood and we are simply “cheating” by supplying the result that we wish to achieve by a clever selection of prior. Given the default nature of the prior and the fact that it is designed to be uninformative (it is, indeed, a flat, proper prior), it seems an unlikely tool with which to engineer a highly informative posterior. However, treating this suspicion with seriousness, we explain below why the maximum likelihood approach fails for the Philadelphia Divorce Study and why the DA Algorithm, by contrast, succeeds.
Failure of the Maximum Likelihood Approach

There are several problems with a maximum likelihood approach to this particular problem. The structure of the missingness in the Philadelphia Divorce Study makes it difficult to disentangle always-takers from compliers. As a result, we observe skew and multimodality in the likelihood function for the kind of missing data mechanism that exists in the Philadelphia Divorce Study. The shape of the likelihood therefore departs from the desired asymptotic behavior that we depend on for good MLE estimates. Such multimodality doesn’t appear to occur in the case where the data is MCAR, at least not for these model assumptions.

What is worse, for certain parameters, the MLE will not perform better as the amount of data gets large. Parameters that are weakly identified remained weakly identified because the amount of missing data grows as the amount of data grows and the structure of the missingness remains problematic for these parameters. No matter how much data we have, we will never observe $W_i$ for units, i, such that $Y_{i}^{obs} = 0$ and $Z_i = 0$

Again, this is less of a worry in the MCAR case. To see intuitively that the MLE will perform well asymptotically, we could always get good, if inefficient, maximum likelihood estimates by simply discarding data rows that contain missing data. This subset of the data will satisfy the conditions for asymptotic normality and consistency of the MLE.

Why the DA Algorithm Performs Well

To begin with, note that the crucial information for the DA Algorithm comes from the likelihood, not from the prior, at least in the case of flat or weakly informative priors. Consider the case where the prior is flat, as it is in the example Philadelphia Divorce Study simulated dataset. Recall the marginal posterior plots for the parameters in Figures 3.13 and 3.14. These look nothing like the standard uniform distribution (i.e. they look nothing like the prior). Indeed, the shapes of these posteriors are completely determined by the likelihood.

Let’s denote the parameter vector as capital $\Theta$ and the $i^{th}$ component of this vector
as $\theta_i$. Let’s further denote the vector of remaining components (i.e. all components of the parameter vector besides $\theta_i$) as $\Theta_{-i}$. Then, the marginal posterior for $\theta_i$ is the full posterior with the other parameter components integrated out:

$$
\pi(\theta_i \mid Y) = \int \pi(\Theta \mid Y) d\Theta_{-i}
$$

Factoring, the full posterior into the product of the likelihood and the prior we have:

$$
\pi(\theta_i \mid Y) \propto \int \mathcal{L}(\theta_i, \Theta_{-i} \mid Y) \pi(\Theta) d\Theta_{-i}
$$

But recall that we are using a flat prior for the parameter vector and therefore $\pi(\Theta) = 1$. This suggests that a flat prior will have no effect on the shape of the integral (and that similarly weak priors will have very little effect):

$$
\pi(\theta_i \mid Y) \propto \int \mathcal{L}(\theta_i, \Theta_{-i} \mid Y) d\Theta_{-i}
$$

Indeed, the marginal posterior for $\theta_i$ is just the likelihood function with all of the other parameters integrated out. The shape of the posterior, which diverges significantly from the shape of the marginal prior, is completely determined by the likelihood. The multimodality for some of these parameters reflects the fact that there are mixture components to the likelihood, which are in turn a product of the structured missingness in $W_{obs}$.

The pseudo observations prior is similarly weak. Indeed, the pseudo observations prior that we use is almost identical to the flat prior for the case where the model contains no covariates. We use the flat prior in the argument above simply for ease of explanation.

We have a likelihood function, a proper prior and the assignment mechanism for $Z$ and the missing data mechanism for $W$ are both ignorable. We are guaranteed to get a proper posterior. The biggest weakness is whether the likelihood is well-specified, but, given the difficulty of the problem, we are forced to make some model assumptions.

The DA Algorithm, since we are reporting the full posterior, doesn’t suffer from the problems that arise in the maximum likelihood approach. Those problems arise when we report
a single point estimate whose validity and sampling variance depend heavily on asymptotic properties that are inapplicable for this dataset. Thankfully, the parameters with nasty posteriors are not estimands of interest but rather are nuisance parameters. The estimands, by contrast, have very well-behaved posteriors.

The nastiness of the posteriors for the nuisance parameters raises the question of why the posteriors for the actual estimands are so well-behaved. One way to think about this question is to consider the super population estimand \( \tau_{sp} = p_{co,t} - p_{co,c} \). Recalling Figure 3.13, we see that the marginal posteriors for \( p_{co,t} \) and \( p_{co,c} \) both have high variance and skew. Yet note how similar these posteriors look. Indeed, for this example dataset, \( p_{co,t} \) and \( p_{co,c} \) are quite correlated. In fact, the correlation coefficient between the posterior draws for \( p_{co,t} \) and \( p_{co,c} \) is \( \text{corr}(p_{co,t}, p_{co,c}) = 0.70 \). Recalling a basic fact about the variance of sums, we see that there is substantial variance reduction for this difference due the covariance between \( p_{co,t} \) and \( p_{co,c} \):

\[
\text{Var}(\tau_{sp}) = \text{Var}(p_{co,t} - p_{co,c}) = \text{Var}(p_{co,t}) + \text{Var}(p_{co,c}) - 2\text{Cov}(p_{co,t}, p_{co,c})
\]

Indeed, in the simulated dataset, \( \text{Var}(p_{co,t}) = 0.050 \) and \( \text{Var}(p_{co,c}) = 0.047 \), yet the variance of the estimand, \( \text{Var}(p_{co,c} - p_{co,t}) = \text{Var}(\tau_{sp}) = 0.029 \), a substantial reduction in variance relative to the variance of these two parameters. Similarly, the skews of the two distributions cancel, and we are left with a nice, symmetric bell-shaped curve whose variance is substantially reduced relative to the parameter variances (See Figure 3.12). The finite-population estimand, which is obviously closely related to the super population estimand, has similarly nice properties.

### 3.9 Results for The Philadelphia Divorce Study

We now apply the DA algorithm to estimate the posterior distribution of the complier average causal effect in the Philadelphia Divorce Study for each estimand. Table 3.8 summarizes the
data that was collected for the primary outcome of interest, whether the potential litigant managed to file a divorce lawsuit in Philadelphia county. Note that almost half of the observations are missing their treatment indicators.

<table>
<thead>
<tr>
<th>S(z, w, y)</th>
<th>G_i</th>
<th>Z_i</th>
<th>W_i</th>
<th>Y_i</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>S(0, ?, 0)</td>
<td>{co, nt, at}</td>
<td>0</td>
<td>?</td>
<td>0</td>
<td>152</td>
</tr>
<tr>
<td>S(0, 0, 1)</td>
<td>{co, nt}</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>58</td>
</tr>
<tr>
<td>S(0, 1, 1)</td>
<td>at</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>27</td>
</tr>
<tr>
<td>S(1, 0, 0)</td>
<td>nt</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>S(1, 0, 1)</td>
<td>nt</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>S(1, 1, 0)</td>
<td>{co, at}</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>S(1, 1, 1)</td>
<td>{co, at}</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>41</td>
</tr>
</tbody>
</table>

Table 3.8: Actual Dataset

The estimated CACEs are huge. It is very difficult for litigants to get into court without an attorney. The intervals are wide, but the effects are so large that posterior variance doesn’t prevent us from detecting that there is almost certainly a positive effect and that this effect is likely quite strong. Representation by a lawyer has a huge effect on a complier’s probability of success.

Figure 3.15 shows the full posterior for the finite population causal estimand for each outcome. Similar to the simulation results, the posterior distributions for these estimands are quite well-behaved. While there is substantial variance, the posteriors are nice, symmetric, unimodal bell-shaped curves. One can see that there is higher variance for the divorce outcomes than for the filing outcomes, but the credible intervals are all located far from 0.

Table 3.9 summarizes the posteriors for these four outcomes. We can see that, although the intervals are wide, the effects are very large, especially for the outcomes that only consider Philadelphia divorce court, which is the proper venue for filing all of these cases. For these outcomes, the complier average causal effect is expected to make the difference between almost never succeeding and almost always succeeding.

Restricting, for a moment, to the filing in Philadelphia county outcome, the marginal posterior distributions for the components of the parameter vector \( \mathbf{p} \) are plotted in Figure 3.16. We do not observe the nasty bimodality in any of these distributions (nor do we for the
other outcomes). We nevertheless observe significant skew and posterior variance for some of these parameters.

The marginal posterior distributions for the components of the parameter vector \( q \), in Figure 3.17, are also relatively well-behaved when compared with the example simulated dataset, although there is perhaps some slight bimodality in the marginal posterior for \( q_{at} \), the proportion of alwaystakers.

The large estimates for the CACEs are not exactly a surprise – the intention to treat (ITT) effects were also relatively large, on the order of 28% to 44%. Since the ITT sometimes has the characteristics of a watered-down CACE, it is no surprise that the CACE is so large.

Table 3.9: CACE Summary

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Posterior Mean</th>
<th>Posterior Mode</th>
<th>0.025 Quantile</th>
<th>0.975 Quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>File for Divorce (PA)</td>
<td>0.57</td>
<td>0.55</td>
<td>0.29</td>
<td>0.84</td>
</tr>
<tr>
<td>File for Divorce (PHI)</td>
<td>0.88</td>
<td>0.90</td>
<td>0.71</td>
<td>0.98</td>
</tr>
<tr>
<td>Obtain Divorce (PA)</td>
<td>0.60</td>
<td>0.69</td>
<td>0.13</td>
<td>0.89</td>
</tr>
<tr>
<td>Obtain Divorce (PHI)</td>
<td>0.87</td>
<td>0.91</td>
<td>0.70</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Figure 3.16: Posterior for $p_i$, Philadelphia Divorce Study

Figure 3.17: Posterior for $q_i$, Philadelphia Divorce Study
These large effects cast serious doubt on the constitutionality of Philadelphia’s divorce court procedures. The effects are quite large; it is almost prohibitively difficult for a litigant to obtain a divorce without an attorney and relatively easy to obtain one with an attorney. Since the right to divorce is a fundamental constitutionally protected right that can only be effectuated via the courts, these results suggest that the Philadelphia divorce court is required to provide attorneys to would-be divorcees or else dramatically simplify their procedures.

3.10 Remarks on Model Misspecification

The first half of this dissertation discusses the consequences of model misspecification in model-based Bayesian approaches to causal inference. In the second half, we have proposed a Bayesian methodology for handling two-way noncompliance when there is substantial missingness in the compliance data. We remark here on how the results of the first study informed our modeling choices in the second study.

To begin with, in the simulation study we found that a failure to adequately transform the outcome variable was by far the most consequential form of model misspecification. In the simulation study, the outcome was continuous whereas, by contrast, in the Philadelphia Divorce Study, the outcomes are binary. Since the outcomes are binary, the problem of transforming the outcome appropriately is less of a concern.

In the simulation study, we also found that the choice of prior was also an important factor in the performance of misspecified models. The choice of prior was usually only important when there was a failure to adequately transform the outcome and so the prior is likely less of a concern in the Philadelphia Divorce Study. Nevertheless, we found in the simulation study that fewer problems arose in the misspecified setting when a flat prior was chosen (although there is some risk of overcoverage). We use flat or nearly flat proper priors in the Philadelphia Divorce Study and therefore the reported credible intervals may be conservative. Additionally, we experimented with multiple choices of prior distribution
and showed that the results were quite similar.

If we had included covariates in the Philadelphia Divorce Study model, we would have had to consider the choice of link function, the form of the linear predictor (had we used a GLM) and other issues raised by the modeling of the covariates. With the relatively simple, covariate-free binomial models that we chose, these issues are less of a concern. By omitting covariates in the model, we likely lost some precision in the calculation of our posterior intervals for the causal estimand (another reason for suspecting that the reported intervals are somewhat conservative). Even though the omission of covariates and the use of a flat prior may lead to conservative credible intervals, we believe this is unproblematic for this study because the estimated causal effects are so large. The large magnitude of the treatment effect, which is ascertainable even with the somewhat wide credible intervals that we report, is the central finding of scientific importance in the Philadelphia Divorce Study. We would also rather avoid model misspecification errors that could arise from modeling the relationship between the outcome and the covariates and from posing a strong prior (especially in this study where there is no strong prior information).

A future study could investigate the effects of model misspecification in the setting where the outcome is binary.

### 3.11 Future Extensions

There are a variety of future extensions to this work that could be envisioned. For example:

- A variant of the proposed algorithm that incorporates covariates. There are over 400 covariates and so some dimension reduction would be required, perhaps by using the propensity score as a summary covariate. There is no pre-existing, expert knowledge about which covariates were important, since this study is the first of its kind.

- Sensitivity analysis to the assumptions about the likelihood and prior
• More simulation studies. The simulation studies were very computationally intensive, requiring weeks of computing time. If more computing resources were available, one could conduct a broader range of simulations.

These are all large projects that could be starting points for future studies.

3.12 Conclusion

We began by developing a method for estimating the complier average causal effect for ignorable assignment mechanisms in the presence of substantial missing data in $W$, the vector of indicators for treatment actually received. We demonstrated the efficacy of the proposed maximum likelihood and full Bayesian estimation methods in simulation studies where the treatment received data was missing completely at random. We show that point estimates and intervals, even for nuisance parameters, had desirable bias and coverage properties.

In simulation studies that replicated the missing data mechanism observed in the Philadelphia Divorce Study, we observed that maximum-likelihood methods broke down but that the fully Bayesian approach continued to produce credible intervals for the causal estimand with the correct coverage properties.

We calculated both the ITT, using standard methods, and the CACE, using the methods proposed in this study, for all of the causal estimands. The estimated effect sizes were all very large. The large ITT effects suggest that Philadelphia VIP, the legal services provider with whom we partnered in the study, has a quite efficacious pro bono divorce practice.

The estimated CACE suggests a giant effect of having an attorney on one’s ability to obtain a divorce in Philadelphia, which was the proper venue for all of the study participants. We summarized the divorce law and constitutional law that has bearing on the legality of current Philadelphia divorce court procedure. Based on the CACE effect sizes that we estimate using the new methodology proposed in this study, we argued that the Philadelphia
The divorce court system is in violation of the Due Process Clause of the Fourteenth Amendment to the United States Constitution.
Chapter 4

Conclusion

In this dissertation, we first examined the effects of Bayesian model misspecification on causal inferences via a full factorial simulation experiment. Secondly, we developed new methodology for analyzing randomized experiments that simultaneously have noncompliance and missingness in the indicators of which treatment was received.

The first study highlights which modeling assumptions carry the most risk in model-based Bayesian causal inference. The second study shows that model-based Bayesian methods are sometimes necessary; these methods provide conceptual coherence and analytical power for difficult problems, such as handling noncompliance when compliance data is systematically missing.

In the full factorial simulation experiment we analyzed the risks associated with many forms of model misspecification. We considered the analytical consequences of the choice of the likelihood, the choice of prior and experimental design choices such as the sample size and the assignment mechanism. We relied on Bayesian evaluation criteria to measure the differences between the true posterior distribution and a misspecified posterior distribution. In particular, we considered the coverage of a misspecified credible interval under the true posterior and the Kullback-Leibler divergence from the misspecified posterior distribution to the true posterior distribution.
Of the factors considered in the experiment, we found that the most consequential form of model misspecification is the failure to properly transform the outcome variable. When the outcome is improperly transformed, inferences are additionally very sensitive to the choice of prior. Indeed, even the choice of the true, data generating prior can underperform the choice of a flat prior in such cases. Misspecified posterior intervals were often badly located. Misspecified intervals could be both much too wide and much too narrow, relative to the true posterior interval. Such problems seem to persist even when sample sizes are large and when the required transformation of the outcome appears to be quite mild. Considering these risks, it may be better to avoid complex modeling assumptions unless they are truly necessary for the problem at hand (for example, if Fisher’s exact test will suffice for the problem, one can avoid complex modeling decisions).

We next saw that the Philadelphia Divorce Study data set presents a thorny problem where model-based Bayesian methods greatly outperform other approaches. Inspired by this particular dataset, we developed a methodology for estimating the complier average causal effect for experiments with ignorable assignment mechanisms in the presence of two-way noncompliance and substantial missing data in the indicators for treatment actually received.

We showed the efficacy of the proposed methodology (both the full Bayesian method and the maximum likelihood method) in simulation studies where the treatment indicators were missing completely at random. We then saw that, for simulations with the missing data mechanism observed in the Philadelphia Divorce Study, maximum-likelihood methods failed but the full Bayesian method still performed well; the model-based Bayesian method continued to produce credible intervals for the causal estimand with the correct coverage properties and the posterior mean was quite close to the true causal estimand on average.

We applied these methods to The Philadelphia Divorce Study. The results suggest that it is incredibly difficult to even file for divorce, much less obtain a divorce, without an attorney. With these results in hand, we analyzed the constitutionality of the current divorce court
system in Philadelphia. We argued that Philadelphia is in violation of the Due Process Clause of the Fourteenth Amendment to the United States Constitution because its divorce court procedures significantly impair the ability of pro se litigants to obtain a divorce, which is a constitutionally protected right.

We discuss why the model-based Bayesian method succeeds even though maximum-likelihood methods and the usual method-of-moments instrumental variables approach both fail.

Finally, we argued that the model misspecification problems that we investigated in the simulation study do not affect the results of the Philadelphia Divorce Study. In particular, since the outcomes are binary in the Philadelphia Divorce Study, we cannot so egregiously fail to transform the outcome variable. Additionally, we use flat or nearly flat proper priors, which were observed to mitigate misspecification problems in the simulation study.

In short, while model-based Bayesian methods are often indispensable for causal analysis of randomized experiments, researchers should be cognizant of the risks of model misspecification.
Appendix A

Simulation Study: ANOVA Tables for All Outcomes

A.1 ANOVA for the $PC(Y^{obs}, X, W)$

Table A.1: ANOVA for the mean of $PC(Y^{obs}, X, W)$: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>332.3</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>50.6</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation</td>
<td>4</td>
<td>45.0</td>
</tr>
<tr>
<td>Assignment Mechanism</td>
<td>2</td>
<td>0.6</td>
</tr>
<tr>
<td>Analysis Prior * Assignment Mechanism</td>
<td>2</td>
<td>0.5</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>0.4</td>
</tr>
<tr>
<td>Missing Covariates</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>Analysis Prior * P</td>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>Analysis Likelihood * Missing Transformation</td>
<td>8</td>
<td>0.2</td>
</tr>
<tr>
<td>Analysis Likelihood</td>
<td>2</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Table A.2: ANOVA for the 0.1 quantile of $PC(Y^{obs}, X, W)$: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>216</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>98</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation</td>
<td>4</td>
<td>34</td>
</tr>
<tr>
<td>Assignment Mechanism</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Analysis Prior * Assignment Mechanism</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Analysis Prior * P</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>P * Assignment Mechanism</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation * Assignment Mechanism</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Missing Transformation * Assignment Mechanism</td>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>

Table A.3: ANOVA for the 0.9 quantile of $PC(Y^{obs}, X, W)$: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>371</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation</td>
<td>4</td>
<td>59</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>57</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>0.1</td>
</tr>
<tr>
<td>P * Analysis Prior</td>
<td>2</td>
<td>0.1</td>
</tr>
<tr>
<td>P * Analysis Prior * Missing Transformation</td>
<td>8</td>
<td>0.1</td>
</tr>
<tr>
<td>P * Missing Transformation</td>
<td>8</td>
<td>0.1</td>
</tr>
<tr>
<td>True Likelihood</td>
<td>2</td>
<td>0.1</td>
</tr>
<tr>
<td>Missing Transformation * True Likelihood</td>
<td>8</td>
<td>0.1</td>
</tr>
<tr>
<td>Missing Transformation * Missing Covariates</td>
<td>4</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table A.4: 2 x 5 table for the mean of $PC(Y^{obs}, X, W)$

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^{-1}}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>$\log(y)$</th>
<th>$y$</th>
<th>$\frac{y^{2}-1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>0.917</td>
<td>0.738</td>
<td>0.612</td>
<td>0.966</td>
<td>0.917</td>
<td>0.830</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>0.000</td>
<td>0.102</td>
<td>0.460</td>
<td>0.973</td>
<td>0.000</td>
<td>0.307</td>
</tr>
<tr>
<td>Total</td>
<td>0.458</td>
<td>0.420</td>
<td>0.536</td>
<td>0.969</td>
<td>0.459</td>
<td>0.568</td>
</tr>
</tbody>
</table>
### A.2 ANOVA for the KL Divergence

Table A.5: ANOVA for the mean of the KL divergence: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Prior</td>
<td>4</td>
<td>693881</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation</td>
<td>4</td>
<td>115195</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>2</td>
<td>97422</td>
</tr>
<tr>
<td>N * Analysis Prior</td>
<td>2</td>
<td>2108</td>
</tr>
<tr>
<td>N</td>
<td>1</td>
<td>1769</td>
</tr>
<tr>
<td>Analysis Prior * Missing Covariates</td>
<td>2</td>
<td>1355</td>
</tr>
<tr>
<td>Missing Covariates</td>
<td>2</td>
<td>1271</td>
</tr>
<tr>
<td>Missing Covariates * N</td>
<td>2</td>
<td>1237</td>
</tr>
<tr>
<td>Missing Covariates * Analysis Prior * N</td>
<td>1</td>
<td>1111</td>
</tr>
<tr>
<td>N * Missing Transformation</td>
<td>2</td>
<td>761</td>
</tr>
<tr>
<td>Analysis Prior * P</td>
<td>2</td>
<td>408</td>
</tr>
</tbody>
</table>

Table A.6: ANOVA for the 0.1 quantile of the KL divergence: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>386322</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation</td>
<td>4</td>
<td>128547</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>103294</td>
</tr>
<tr>
<td>N * Analysis Prior</td>
<td>2</td>
<td>3849</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>3541</td>
</tr>
<tr>
<td>Missing Covariates</td>
<td>1</td>
<td>3262</td>
</tr>
<tr>
<td>Analysis Prior * N * Missing Covariates</td>
<td>2</td>
<td>3013</td>
</tr>
<tr>
<td>Missing Covariates * N</td>
<td>2</td>
<td>3009</td>
</tr>
<tr>
<td>Missing Covariates * Analysis Prior</td>
<td>1</td>
<td>2803</td>
</tr>
<tr>
<td>N * Missing Transformation</td>
<td>8</td>
<td>1458</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation * N</td>
<td>8</td>
<td>1357</td>
</tr>
</tbody>
</table>
Table A.7: ANOVA for the 0.9 quantile of the KL divergence: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>772607</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>123656</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation</td>
<td>4</td>
<td>117100</td>
</tr>
<tr>
<td>Assignment Mechanism</td>
<td>2</td>
<td>509</td>
</tr>
<tr>
<td>Missing Transformation * N</td>
<td>8</td>
<td>496</td>
</tr>
<tr>
<td>Analysis Prior * P</td>
<td>2</td>
<td>428</td>
</tr>
<tr>
<td>Analysis Prior * Assignment Mechanism</td>
<td>2</td>
<td>333</td>
</tr>
<tr>
<td>Analysis Prior * Missing Covariates</td>
<td>1</td>
<td>293</td>
</tr>
<tr>
<td>Analysis Prior * N</td>
<td>2</td>
<td>185</td>
</tr>
<tr>
<td>Missing Covariates * N</td>
<td>2</td>
<td>177</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>171</td>
</tr>
</tbody>
</table>

Table A.8: 2 x 5 table for the mean of the KL Divergence

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^2}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>$\log(y)$</th>
<th>$y$</th>
<th>$\frac{y^2-1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>0.8</td>
<td>7.9</td>
<td>10.1</td>
<td>0.7</td>
<td>1.2</td>
<td>4.1</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>46.5</td>
<td>35.4</td>
<td>13.5</td>
<td>0.1</td>
<td>44.8</td>
<td>28.0</td>
</tr>
<tr>
<td>Total</td>
<td>23.7</td>
<td>21.6</td>
<td>11.8</td>
<td>0.4</td>
<td>23.0</td>
<td>16.1</td>
</tr>
</tbody>
</table>

A.3 ANOVA for the Normalized Difference in Posterior Means

Table A.9: ANOVA for the mean of the log absolute standardized bias: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>6759</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior</td>
<td>4</td>
<td>2897</td>
</tr>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>1529</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>57</td>
</tr>
<tr>
<td>Missing Covariates</td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>Analysis Prior * N</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>Missing Transformation * Missing Covariates</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior * Missing Covariates</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Assignment Mechanism</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Analysis Prior * Assignment Mechanism</td>
<td>2</td>
<td>6</td>
</tr>
</tbody>
</table>
Table A.10: ANOVA for the 0.1 quantile of the standardized bias: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>1025732</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior</td>
<td>4</td>
<td>607751</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>579359</td>
</tr>
<tr>
<td>Analysis Prior * N</td>
<td>2</td>
<td>76356</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>49517</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior * N</td>
<td>8</td>
<td>48526</td>
</tr>
<tr>
<td>Missing Transformation * N</td>
<td>8</td>
<td>792</td>
</tr>
<tr>
<td>Analysis Prior * P</td>
<td>2</td>
<td>435</td>
</tr>
<tr>
<td>Missing Covariates</td>
<td>1</td>
<td>393</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>334</td>
</tr>
</tbody>
</table>

Table A.11: ANOVA for the 0.9 quantile of the standardized bias: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>1032415</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>676920</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior</td>
<td>4</td>
<td>645169</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>85760</td>
</tr>
<tr>
<td>Analysis Prior * N</td>
<td>2</td>
<td>82356</td>
</tr>
<tr>
<td>Missing Transformation * N</td>
<td>8</td>
<td>51894</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior * N</td>
<td>8</td>
<td>51076</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>1167</td>
</tr>
<tr>
<td>Analysis Prior * P</td>
<td>2</td>
<td>738</td>
</tr>
<tr>
<td>Analysis Prior * Missing Covariates</td>
<td>1</td>
<td>525</td>
</tr>
</tbody>
</table>

Table A.12: 2 x 5 table for the mean of the Standardized bias

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^2}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>log(y)</th>
<th>$y$</th>
<th>$\frac{y^a-1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>0.00</td>
<td>0.00</td>
<td>$4 \times 10^{10}$</td>
<td>0.00</td>
<td>-0.01</td>
<td>$8 \times 10^9$</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>117</td>
<td>0.11</td>
<td>$10^{15}$</td>
<td>0.00</td>
<td>28</td>
<td>$3 \times 10^{14}$</td>
</tr>
<tr>
<td>Total</td>
<td>58</td>
<td>0.05</td>
<td>$7 \times 10^{14}$</td>
<td>0.00</td>
<td>14</td>
<td>$1.5 \times 10^{14}$</td>
</tr>
</tbody>
</table>

Table A.13: 2 x 5 table for the median of the Standardized bias

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^2}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>log(y)</th>
<th>$y$</th>
<th>$\frac{y^a-1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>-0.02</td>
<td>0.00</td>
<td>-0.05</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.01</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>117</td>
<td>0.09</td>
<td>-0.07</td>
<td>0.00</td>
<td>28</td>
<td>29</td>
</tr>
<tr>
<td>Total</td>
<td>58</td>
<td>0.05</td>
<td>-0.06</td>
<td>0.00</td>
<td>14</td>
<td>14</td>
</tr>
</tbody>
</table>

151
Table A.14: 2 x 5 table for the 0.9 quantile of the Standardized bias

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^2}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>log($y$)</th>
<th>$y$</th>
<th>$\frac{y^2-1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>2.2</td>
<td>0.03</td>
<td>0.15</td>
<td>2.1</td>
<td>2.6</td>
<td>1.4</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>120</td>
<td>0.23</td>
<td>-0.02</td>
<td>0.04</td>
<td>32</td>
<td>30.5</td>
</tr>
<tr>
<td>Total</td>
<td>61</td>
<td>0.13</td>
<td>0.06</td>
<td>1.05</td>
<td>17</td>
<td>16.0</td>
</tr>
</tbody>
</table>

A.4 ANOVA for the Ratio of Posterior Standard Deviations

Table A.15: ANOVA for the mean log of the ratio posterior standard deviations: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>10406</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior</td>
<td>4</td>
<td>221</td>
</tr>
<tr>
<td>Missing Transformation * N</td>
<td>8</td>
<td>59</td>
</tr>
<tr>
<td>Missing Transformation * P</td>
<td>8</td>
<td>58</td>
</tr>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>41</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>Missing Transformation * True Likelihood</td>
<td>8</td>
<td>19</td>
</tr>
<tr>
<td>True Likelihood</td>
<td>2</td>
<td>19</td>
</tr>
<tr>
<td>Analysis Prior * Missing Covariates</td>
<td>1</td>
<td>15</td>
</tr>
</tbody>
</table>

Table A.16: ANOVA for the 0.1 quantile of the ratio posterior standard deviations: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>5590</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior</td>
<td>4</td>
<td>3678</td>
</tr>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>3235</td>
</tr>
<tr>
<td>Analysis Prior * Missing Covariates</td>
<td>1</td>
<td>269</td>
</tr>
<tr>
<td>Missing Transformation * Missing Covariates</td>
<td>4</td>
<td>151</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior * Missing Covariates</td>
<td>4</td>
<td>111</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>51</td>
</tr>
<tr>
<td>Missing Covariates</td>
<td>1</td>
<td>36</td>
</tr>
<tr>
<td>Analysis Prior * Missing Covariates</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>Analysis Prior * P</td>
<td>2</td>
<td>19</td>
</tr>
</tbody>
</table>
Table A.17: ANOVA for the 0.9 quantile of the ratio of posterior standard deviations: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>6869</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior</td>
<td>4</td>
<td>4355</td>
</tr>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>3768</td>
</tr>
<tr>
<td>Analysis Prior * Missing Covariates</td>
<td>1</td>
<td>503</td>
</tr>
<tr>
<td>Missing Transformation * Missing Covariates</td>
<td>4</td>
<td>313</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior * Missing Covariates</td>
<td>4</td>
<td>194</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>55</td>
</tr>
<tr>
<td>Missing Transformation * P</td>
<td>8</td>
<td>50</td>
</tr>
<tr>
<td>Analysis Prior * P</td>
<td>2</td>
<td>38</td>
</tr>
<tr>
<td>Analysis Prior * P * Missing Covariates</td>
<td>2</td>
<td>38</td>
</tr>
</tbody>
</table>

Table A.18: 2 x 5 table for the mean of the ratio of posterior standard deviations

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^2}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>log($y$)</th>
<th>$y$</th>
<th>$\frac{y^2-1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>1.8</td>
<td>0.10</td>
<td>$6 \times 10^{10}$</td>
<td>2.2</td>
<td>2.2</td>
<td>$1.2 \times 10^{10}$</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>10.7</td>
<td>0.06</td>
<td>$7 \times 10^{17}$</td>
<td>1.3</td>
<td>2.8</td>
<td>$1.4 \times 10^{17}$</td>
</tr>
<tr>
<td>Total</td>
<td>6.3</td>
<td>0.08</td>
<td>$3.5 \times 10^{17}$</td>
<td>1.8</td>
<td>2.5</td>
<td>$7 \times 10^{16}$</td>
</tr>
</tbody>
</table>

Table A.19: 2 x 5 table for the median of the ratio of posterior standard deviations

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^2}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>log($y$)</th>
<th>$y$</th>
<th>$\frac{y^2-1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>1.8</td>
<td>0.01</td>
<td>0.01</td>
<td>2.2</td>
<td>2.2</td>
<td>1.2</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>10.7</td>
<td>0.02</td>
<td>0.01</td>
<td>1.3</td>
<td>2.8</td>
<td>3.0</td>
</tr>
<tr>
<td>Total</td>
<td>6.2</td>
<td>0.01</td>
<td>0.01</td>
<td>1.8</td>
<td>2.5</td>
<td>2.1</td>
</tr>
</tbody>
</table>

Table A.20: 2 x 5 table for the 0.9 quantile of the ratio of posterior standard deviations

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^2}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>log($y$)</th>
<th>$y$</th>
<th>$\frac{y^2-1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>2.05</td>
<td>0.08</td>
<td>0.46</td>
<td>2.4</td>
<td>2.4</td>
<td>1.5</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>11.3</td>
<td>0.05</td>
<td>0.26</td>
<td>1.4</td>
<td>3.2</td>
<td>3.2</td>
</tr>
<tr>
<td>Total</td>
<td>6.7</td>
<td>0.07</td>
<td>0.36</td>
<td>1.9</td>
<td>2.8</td>
<td>2.4</td>
</tr>
</tbody>
</table>
A.5 ANOVA for the Ratio of Credible Interval Widths

The ANOVA table for the mean of the log of the credible interval width ratio is presented below. We truncate the table to only display the top 10 factor combinations, ranked by mean squared error. The ANOVA for the mean is done on the log scale to avoid overflow in the computations, which results from a couple of very large outliers. All other tables are based on the unlogged values, as the other outcomes are robust to these outliers.

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Transformation * True Likelihood</td>
<td>8</td>
<td>1541</td>
</tr>
<tr>
<td>True Likelihood</td>
<td>2</td>
<td>1540</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>1351</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior</td>
<td>4</td>
<td>260</td>
</tr>
<tr>
<td>Missing Transformation * N</td>
<td>8</td>
<td>188</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>158</td>
</tr>
<tr>
<td>True Likelihood * N</td>
<td>4</td>
<td>108</td>
</tr>
<tr>
<td>Missing Transformation * True Likelihood * N</td>
<td>16</td>
<td>102</td>
</tr>
<tr>
<td>Missing Transformation * P</td>
<td>8</td>
<td>23</td>
</tr>
<tr>
<td>Missing Transformation * Missing Covariates</td>
<td>4</td>
<td>17</td>
</tr>
</tbody>
</table>

Table A.22: ANOVA for the 0.1 quantile of the CI width ratio: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>5492</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior</td>
<td>4</td>
<td>3678</td>
</tr>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>3321</td>
</tr>
<tr>
<td>Analysis Prior * Missing Covariates</td>
<td>1</td>
<td>277</td>
</tr>
<tr>
<td>Missing Transformation * Missing Covariates</td>
<td>4</td>
<td>152</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior * Missing Covariates</td>
<td>4</td>
<td>111</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>52</td>
</tr>
<tr>
<td>Missing Covariates</td>
<td>1</td>
<td>38</td>
</tr>
<tr>
<td>Analysis Prior * Missing Covariates * P</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>Analysis Prior * P</td>
<td>2</td>
<td>20</td>
</tr>
</tbody>
</table>
Table A.23: ANOVA for the 0.9 quantile of the CI width ratio: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Transformation * Analysis Prior</td>
<td>4</td>
<td>5088</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>4662</td>
</tr>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>2096</td>
</tr>
<tr>
<td>True Likelihood</td>
<td>2</td>
<td>1375</td>
</tr>
<tr>
<td>Missing Transformation * True Likelihood</td>
<td>8</td>
<td>1370</td>
</tr>
<tr>
<td>Missing Transformation * P</td>
<td>8</td>
<td>854</td>
</tr>
<tr>
<td>Missing Transformation * N</td>
<td>8</td>
<td>785</td>
</tr>
<tr>
<td>Missing Transformation * True Likelihood * P</td>
<td>8</td>
<td>617</td>
</tr>
<tr>
<td>True Likelihood * P</td>
<td>4</td>
<td>614</td>
</tr>
<tr>
<td>Missing Transformation * True Likelihood * N</td>
<td>16</td>
<td>607</td>
</tr>
</tbody>
</table>

Table A.24: 2 x 5 table for the mean of the ratio of CI Widths

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^3}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>log($y$)</th>
<th>$y$</th>
<th>$\frac{y^3-1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>1.8</td>
<td>1.5</td>
<td>$4 \times 10^{12}$</td>
<td>2.2</td>
<td>2.2</td>
<td>$8 \times 10^{11}$</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>11</td>
<td>0.87</td>
<td>$10^{19}$</td>
<td>1.3</td>
<td>2.8</td>
<td>$2 \times 10^{18}$</td>
</tr>
<tr>
<td>Total</td>
<td>6.3</td>
<td>1.2</td>
<td>$5 \times 10^{18}$</td>
<td>1.8</td>
<td>2.5</td>
<td>$10^{18}$</td>
</tr>
</tbody>
</table>

Table A.25: 2 x 5 table for the median of the ratio of CI Widths

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^3}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>log($y$)</th>
<th>$y$</th>
<th>$\frac{y^3-1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>1.8</td>
<td>0.2</td>
<td>0.1</td>
<td>2.2</td>
<td>2.2</td>
<td>1.3</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>11</td>
<td>0.2</td>
<td>0.1</td>
<td>1.3</td>
<td>2.8</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>6.2</td>
<td>0.2</td>
<td>0.1</td>
<td>1.8</td>
<td>2.5</td>
<td>2.2</td>
</tr>
</tbody>
</table>

Table A.26: 2 x 5 table for the 0.9 quantile of the ratio of CI Widths

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^3}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>log($y$)</th>
<th>$y$</th>
<th>$\frac{y^3-1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>2.1</td>
<td>1.2</td>
<td>4.5</td>
<td>2.4</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>11</td>
<td>0.7</td>
<td>2.4</td>
<td>1.4</td>
<td>3.2</td>
<td>3.8</td>
</tr>
<tr>
<td>Total</td>
<td>6.7</td>
<td>1.0</td>
<td>3.4</td>
<td>1.9</td>
<td>2.8</td>
<td>3.2</td>
</tr>
</tbody>
</table>
A.6 ANOVA for the KS Statistic

The ANOVA table for the mean of the Kolmogorov-Smirnoff statistic is presented in Table A.27. We truncate the table to only display the top 10 factor combinations, ranked by mean squared error. The conclusions from these tables are very similar to the conclusions that one might reach from the analysis of the KL divergence and the true posterior coverage of the misspecified credible intervals.

Table A.27: ANOVA for the mean of KS Statistic: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>111.9</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>39.2</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation</td>
<td>4</td>
<td>38.9</td>
</tr>
<tr>
<td>Missing Covariates</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>Assignment Mechanism</td>
<td>2</td>
<td>0.5</td>
</tr>
<tr>
<td>Analysis Prior * Assignment Mechanism</td>
<td>2</td>
<td>0.4</td>
</tr>
<tr>
<td>Missing Covariates * Missing Transformation</td>
<td>4</td>
<td>0.3</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>Analysis Prior * P</td>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>Assignment Mechanism * P</td>
<td>2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Table A.28: ANOVA for the 0.1 quantile of the KS Statistic: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>271</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation</td>
<td>4</td>
<td>44</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>40</td>
</tr>
<tr>
<td>Missing Covariates</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Missing Transformation * Missing Covariates</td>
<td>4</td>
<td>0.3</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>0.1</td>
</tr>
<tr>
<td>Analysis Prior * Missing Covariates</td>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>Analysis Prior * P</td>
<td>2</td>
<td>0.1</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation * P</td>
<td>8</td>
<td>0.04</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation * Missing Covariates</td>
<td>4</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Table A.29: ANOVA for the 0.9 quantile of the KS Statistic: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>45</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation</td>
<td>4</td>
<td>35</td>
</tr>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>Assignment Mechanism</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Analysis Prior * Assignment Mechanism</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>0.4</td>
</tr>
<tr>
<td>Analysis Prior * P</td>
<td>2</td>
<td>0.3</td>
</tr>
<tr>
<td>Missing Transformation * Missing Covariates</td>
<td>4</td>
<td>0.3</td>
</tr>
<tr>
<td>Assignment Mechanism * P</td>
<td>4</td>
<td>0.3</td>
</tr>
<tr>
<td>Missing Covariates</td>
<td>1</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Table A.30: 2 x 5 table for the mean of the Kolmogorov-Smirnov Statistic

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^3}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>log($y$)</th>
<th>$y$</th>
<th>$\frac{y^3-1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>0.409</td>
<td>0.362</td>
<td>0.512</td>
<td>0.393</td>
<td>0.437</td>
<td>0.423</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>1.000</td>
<td>0.916</td>
<td>0.650</td>
<td>0.064</td>
<td>1.000</td>
<td>0.726</td>
</tr>
<tr>
<td>Total</td>
<td>0.704</td>
<td>0.639</td>
<td>0.581</td>
<td>0.229</td>
<td>0.718</td>
<td>0.574</td>
</tr>
</tbody>
</table>

A.7 ANOVA for False Positives

Table A.31: ANOVA for the mean of false positives: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Prior</td>
<td>4</td>
<td>41</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation</td>
<td>4</td>
<td>40</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>1</td>
<td>39</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>0.006</td>
</tr>
<tr>
<td>Analysis Prior * P</td>
<td>2</td>
<td>0.005</td>
</tr>
<tr>
<td>Missing Transformation * N</td>
<td>8</td>
<td>0.004</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>0.003</td>
</tr>
<tr>
<td>Analysis Prior * N</td>
<td>2</td>
<td>0.003</td>
</tr>
<tr>
<td>Missing Transformation * P</td>
<td>8</td>
<td>0.003</td>
</tr>
<tr>
<td>Missing Transformation * P * N</td>
<td>16</td>
<td>0.002</td>
</tr>
</tbody>
</table>
Table A.32: ANOVA for the 0.1 quantile of the false positive rate difference: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>31</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation</td>
<td>4</td>
<td>31</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>Analysis Prior * P</td>
<td>8</td>
<td>0.2</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation * P</td>
<td>8</td>
<td>0.2</td>
</tr>
<tr>
<td>Analysis Prior * Analysis Likelihood * Missing Covariates</td>
<td>2</td>
<td>0.1</td>
</tr>
<tr>
<td>Analysis Likelihood * Missing Covariates</td>
<td>2</td>
<td>0.1</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Likelihood * Missing Covariates</td>
<td>8</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table A.33: ANOVA for the 0.9 quantile of false positive rate difference: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>41</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>39</td>
</tr>
<tr>
<td>Analysis Prior * Missing Transformation</td>
<td>4</td>
<td>40</td>
</tr>
<tr>
<td>N * Analysis Prior</td>
<td>2</td>
<td>3.6</td>
</tr>
<tr>
<td>Analysis Likelihood</td>
<td>2</td>
<td>3.6</td>
</tr>
<tr>
<td>N * Missing Transformation</td>
<td>8</td>
<td>3.6</td>
</tr>
<tr>
<td>N * Analysis Likelihood * Analysis Prior</td>
<td>4</td>
<td>3.6</td>
</tr>
<tr>
<td>P * Analysis Likelihood</td>
<td>4</td>
<td>3.6</td>
</tr>
<tr>
<td>N * P * Analysis Prior</td>
<td>4</td>
<td>3.6</td>
</tr>
<tr>
<td>P * Analysis Likelihood * Analysis Prior * Missing Transformation</td>
<td>16</td>
<td>3.6</td>
</tr>
</tbody>
</table>

Table A.34: 2 x 5 table for the mean of the false positives

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$1 - \frac{y^{-x}}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>$log(y)$</th>
<th>$y^\frac{1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>0.001</td>
<td>0.01</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>0.01</td>
<td>0.91</td>
<td>0</td>
<td>0</td>
<td>0.18</td>
</tr>
<tr>
<td>Total</td>
<td>0.004</td>
<td>0.46</td>
<td>0</td>
<td>0</td>
<td>0.13</td>
</tr>
</tbody>
</table>
### A.8 ANOVA for False Negatives

Table A.35: ANOVA for the mean of False negatives: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Prior</td>
<td>4</td>
<td>26</td>
</tr>
<tr>
<td>Missing Transformation</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>N * Missing Transformation</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>P * Missing Transformation</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior * N</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Analysis Prior * P</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>N * P</td>
<td>2</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Table A.36: ANOVA for the 0.9 quantile of false negative rate difference: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>133</td>
</tr>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>40</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior</td>
<td>4</td>
<td>24</td>
</tr>
<tr>
<td>N * Missing Transformation</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>N * Analysis Prior</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>N * Missing Transformation * Analysis Prior</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>P * Missing Transformation</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>N * P * Missing Transformation</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Missing Covariates * Analysis Prior</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Table A.37: 2 x 5 table for the mean of the false negatives

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^{-3}}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>$\log(y)$</th>
<th>$y$</th>
<th>$\frac{y^{4}-1}{3}$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>0.33</td>
<td>0</td>
<td>0.36</td>
<td>0.01</td>
<td>0.03</td>
<td>0.15</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>0</td>
<td>0</td>
<td>0.39</td>
<td>0</td>
<td>0</td>
<td>0.07</td>
</tr>
<tr>
<td>Total</td>
<td>0.17</td>
<td>0</td>
<td>0.38</td>
<td>0.01</td>
<td>0.02</td>
<td>0.11</td>
</tr>
</tbody>
</table>

### A.9 ANOVA for Skew Ratio

159
Table A.38: ANOVA for the 0.1 quantile of the skew ratio: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>6.2</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>Missing Transformation * N</td>
<td>8</td>
<td>0.1</td>
</tr>
<tr>
<td>N * Analysis Prior</td>
<td>2</td>
<td>0.0</td>
</tr>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>0.0</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior</td>
<td>4</td>
<td>0.0</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>0.0</td>
</tr>
<tr>
<td>Missing Transformation * P</td>
<td>8</td>
<td>0.0</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior * Missing Covariates</td>
<td>4</td>
<td>0.0</td>
</tr>
<tr>
<td>Missing Transformation * N * Analysis Prior</td>
<td>8</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table A.39: ANOVA for the 0.9 quantile of the skew ratio: most influential factors

<table>
<thead>
<tr>
<th>Factor</th>
<th>DF</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Transformation</td>
<td>4</td>
<td>9117</td>
</tr>
<tr>
<td>N</td>
<td>2</td>
<td>288</td>
</tr>
<tr>
<td>Missing Transformation * N</td>
<td>8</td>
<td>98</td>
</tr>
<tr>
<td>P</td>
<td>2</td>
<td>27</td>
</tr>
<tr>
<td>Analysis Prior</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>Missing Transformation * Analysis Prior</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>N * Analysis Prior</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>Missing Transformation * P</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Analysis Prior * Missing Covariates</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Missing Transformation * N * Analysis Prior</td>
<td>8</td>
<td>5</td>
</tr>
</tbody>
</table>

Table A.40: 2 x 5 table for the mean of the ratio of absolute skew

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^{-\frac{3}{2}}}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>log(y)</th>
<th>$y$</th>
<th>$y^{\frac{2}{3}}-1$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>6</td>
<td>7.2</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0.97</td>
<td>21</td>
<td>7.3</td>
</tr>
<tr>
<td>Total</td>
<td>5.4</td>
<td>0</td>
<td>0</td>
<td>17.5</td>
<td>13</td>
<td>0.11</td>
</tr>
</tbody>
</table>

Table A.41: 2 x 5 table for the median of the skew ratio

<table>
<thead>
<tr>
<th>Prior / Transformation</th>
<th>$\frac{1-y^{-\frac{3}{2}}}{3}$</th>
<th>$1 - \frac{1}{y}$</th>
<th>log(y)</th>
<th>$y$</th>
<th>$y^{\frac{2}{3}}-1$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat Prior</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>0.97</td>
<td>0.48</td>
</tr>
<tr>
<td>“True” Prior</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>0.89</td>
<td>0.50</td>
</tr>
<tr>
<td>Total</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>0.93</td>
<td>0.49</td>
</tr>
</tbody>
</table>
Appendix B

Divorce Study: Sequential Binomial Modeling for the Compliance Type

B.1 Likelihood

Recall that these components of the likelihood change when modeling the compliance type with sequential binomials.

\[ Pr(G_i = nt | X_i, \theta) = \phi_{nt} \]

\[ Pr(G_i = co | X_i, G_i \in \{co, at\}, \theta) = \phi_{co} \]

The modified complete data log-likelihood is therefore:

\[
l_{comp}(\theta) = \sum_{i=1}^{n} I(G_i = nt) \left[ y_i \log(p_{nt}) + (1 - y_i) \log(1 - p_{nt}) \right] \\
+ I(G_i = at) \left[ y_i \log(p_{at}) + (1 - y_i) \log(1 - p_{at}) \right] \\
+ I(G_i = co)z_i \left[ y_i \log(p_{co,t}) + (1 - y_i) \log(1 - p_{co,t}) \right] \\
+ I(G_i = co)(1 - z_i) \left[ y_i \log(p_{co,c}) + (1 - y_i) \log(1 - p_{co,c}) \right] \\
+ I(G_i = nt) \log(\phi_{nt}) + I(G_i = at) \log((1 - \phi_{nt})(1 - \phi_{co})) + I(G_i = co) \log((1 - \phi_{nt})\phi_{co})
\]
B.2 E-Step / Gibbs Sampler Steps for the Compliance Types

Case 1: \( Z_i = 1, W_i = 1 \) and \( Y_i = y_i \)

\[
Pr(G_i = at \mid Z_i = 1, W_i = 1, Y_i = y_i) = \frac{p_{at}^{y_i}(1 - p_{at})^{1-y_i}(1 - \phi_{co})}{p_{at}^{y_i}(1 - p_{at})^{1-y_i}(1 - \phi_{co}) + p_{co,t}^{y_i}(1 - p_{co,t})^{1-y_i}\phi_{co}}
\]

and

\[
Pr(G_i = co \mid Z_i = 1, W_i = 1, Y_i = y_i) = \frac{p_{co,t}^{y_i}(1 - p_{co,t})^{1-y_i}\phi_{co}}{p_{at}^{y_i}(1 - p_{at})^{1-y_i}(1 - \phi_{co}) + p_{co,t}^{y_i}(1 - p_{co,t})^{1-y_i}\phi_{co}}
\]

Case 2: \( Z_i = 0, W_i = 0, Y_i = y_i \)

\[
Pr(G_i = nt \mid Z_i = 0, W_i = 0, Y_i = y_i) = \frac{p_{nt}^{y_i}(1 - p_{nt})^{1-y_i}\phi_{nt}}{p_{nt}^{y_i}(1 - p_{nt})^{1-y_i}\phi_{nt} + p_{co,c}^{y_i}(1 - p_{co,c})^{1-y_i}(1 - \phi_{nt})\phi_{co}}
\]

and

\[
Pr(G_i = co \mid Z_i = 0, W_i = 0, Y_i = y_i) = \frac{p_{co,c}^{y_i}(1 - p_{co,c})^{1-y_i}(1 - \phi_{nt})\phi_{co}}{p_{nt}^{y_i}(1 - p_{nt})^{1-y_i}\phi_{nt} + p_{co,c}^{y_i}(1 - p_{co,c})^{1-y_i}(1 - \phi_{nt})\phi_{co}}
\]

Case 3: \( Z_i = z_i, W_i = ?, Y_i = y_i \)
\[ Pr(G_i = nt \mid Z_i, Y_i) = \]
\[
\frac{p^Y_{ni}(1 - p_{nt})^{(1-Y_i)}\phi_{nt}}{p^Y_{ni}(1 - p_{nt})^{(1-Y_i)}\phi_{nt} + p^Y_{at}(1 - p_{at})^{(1-Y_i)}(1 - \phi_{nt})(1 - \phi_{co}) + p^Y_{co,c}(1 - p_{co,c})^{(1-Y_i)}(1 - \phi_{nt})\phi_{co}}
\]

and

\[ Pr(G_i = at \mid Z_i, Y_i) = \]
\[
\frac{p^Y_{at}(1 - p_{at})^{(1-Y_i)}(1 - \phi_{nt})(1 - \phi_{co})}{p^Y_{ni}(1 - p_{nt})^{(1-Y_i)}\phi_{nt} + p^Y_{at}(1 - p_{at})^{(1-Y_i)}(1 - \phi_{nt})(1 - \phi_{co}) + p^Y_{co,c}(1 - p_{co,c})^{(1-Y_i)}(1 - \phi_{nt})\phi_{co}}
\]

and

\[ Pr(G_i = co \mid Z_i, Y_i) = \]
\[
\frac{p^Y_{co,c}(1 - p_{co,c})^{(1-Y_i)}\phi_{co}}{p^Y_{ni}(1 - p_{nt})^{(1-Y_i)}\phi_{nt} + p^Y_{at}(1 - p_{at})^{(1-Y_i)}(1 - \phi_{nt})(1 - \phi_{co}) + p^Y_{co,c}(1 - p_{co,c})^{(1-Y_i)}(1 - \phi_{nt})\phi_{co}}
\]

### B.3 Prior

The priors that we can use with the sequential binomial model for compliance type is substantially similar to the priors that we use for the multinomial. If we favor flat priors, we can have uniform priors on \( \phi_{nt} \) and \( \phi_{co} \) (rather than a dirichlet on the multinomial parameter). Similarly, we can use a pseudo observations prior, which is almost identical to the pseudo observations prior used when we model compliance type as multinomial.
The pseudo observations prior for the two-stage binomial is the following:

\[
\begin{align*}
    p_{at} & \sim \text{Beta}\left(\frac{11}{6}, \frac{11}{6}\right) \\
    p_{nt} & \sim \text{Beta}\left(\frac{11}{6}, \frac{11}{6}\right) \\
    p_{co,t} & \sim \text{Beta}\left(\frac{17}{12}, \frac{17}{12}\right) \\
    p_{co,c} & \sim \text{Beta}\left(\frac{17}{12}, \frac{17}{12}\right) \\
    \phi_{nt} & \sim \text{Beta}\left(\frac{8}{3}, \frac{13}{3}\right) \\
    \phi_{co} & \sim \text{Beta}\left(\frac{8}{3}, \frac{8}{3}\right)
\end{align*}
\]
Appendix C

Divorce Study: Pseudo Observations Prior

C.1 General Remarks

Suppose you were to actually add a synthetic observation for every combination of values in the dataset. In the case where no covariates are included in the model, this amounts to $2 \times 2 \times 3 = 12$ synthetic observations, corresponding to all combinations of the levels of \((Y), (Z)\) and \((G)\). However, suppose we added just a single continuous covariate and that this covariate takes different values for every observation. Then, we would add $2 \times 2 \times 3 \times N$ synthetic observations. In the Philadelphia Divorce Study dataset, where we have 312 observations, this amounts to an additional 3744 observations. Obviously, we need a method to weight the influence of these synthetic observations appropriately. Otherwise, the influence of the prior on our inferences will completely swamp the influence of the actual data.

Let’s denote the unweighted pseudo observations prior as $a(\theta)$. In the example above, $a(\theta)$ is equivalent to $2 \times 2 \times 3 \times N$ synthetic observations. In order to limit the weight of these observations, we set our prior accordingly:
In this way, we limit the influence of the prior to be about as influential as an additional \(N_a\) observations, where \(N_a\) is chosen to be small relative to the sample size of the observed data. In the influenza study, for example, the authors chose \(N_a = 30\), which was about 1.5\% the size of the dataset, which contained 1931 observations.

### C.2 Pseudo Observations Prior for The Philadelphia Divorce Study

Returning to the Philadelphia Divorce Study and the models without covariates, the \(2 \times 2 \times 3 = 12\) synthetic observations would have the following contribution to the likelihood:

\[
\begin{align*}
a(\theta) &= Pr(y = 1 \mid z = 1, G = at)Pr(y = 1 \mid z = 0, G = at)Pr(y = 0 \mid z = 1, G = at)Pr(y = 0 \mid z = 0, G = at) \\
&\times Pr(y = 1 \mid z = 1, G = nt)Pr(y = 1 \mid z = 0, G = nt)Pr(y = 0 \mid z = 1, G = nt)Pr(y = 0 \mid z = 0, G = nt) \\
&\times Pr(y = 1 \mid z = 1, G = co)Pr(y = 1 \mid z = 0, G = co)Pr(y = 0 \mid z = 1, G = co)Pr(y = 0 \mid z = 0, G = co) \\
&\times Pr(G = at)^4Pr(G = nt)^4Pr(G = co)^4
\end{align*}
\]

We then set the prior as follows:

\[
p(\theta) = a(\theta)^{N_a}^{12}
\]

If we limited the influence of these synthetic observations similar to the Influenza paper, we might choose \(N_a = 5\), since 5 synthetic observations is similarly small when compared to the 312 observations in the study.

When we have we model the compliance status as multinomial, this gives us the following prior:

\[
p(\theta) = [p_{at}^2(1 - p_{at})^2p_{nt}^2(1 - p_{nt})^2p_{co,c}(1 - p_{co,c})p_{co,t}(1 - p_{co,t})q_{at}^4q_{nt}^4q_{co}^4]^\frac{N_a}{12}
\]
\[ p(\theta) = p_{at}^{\frac{5}{6}}(1 - p_{at})^{\frac{5}{6}} p_{nt}^{\frac{5}{6}}(1 - p_{nt})^{\frac{5}{6}} p_{co,c}^{\frac{5}{12}}(1 - p_{co,c})^{\frac{5}{12}} p_{co,t}^{\frac{5}{12}}(1 - p_{co,t})^{\frac{5}{12}} p_{nt}^{\frac{5}{12}} q_{nt}^{\frac{5}{12}} q_{co}^{\frac{5}{12}} \]

Therefore, in the case without covariates, this means that the pseudo observations prior still factors as conjugate priors for each of the parameter components. Explicitly, we have the following:

\[ p(p_{co,t}, p_{co,c}, p_{at}, p_{nt}, q) = p(p_{co,t})p(p_{co,c})p(p_{at})p(p_{nt})p(q) \]

where

\[ p_{co,t} \sim Beta(17/12, 17/12) \]
\[ p_{co,c} \sim Beta(17/12, 17/12) \]
\[ p_{at} \sim Beta(11/6, 11/6) \]
\[ p_{nt} \sim Beta(11/6, 11/6) \]
\[ q \sim Dirichlet(8/3, 8/3, 8/3) \]

With such a prior, our posterior updates at each iteration of the DA algorithm would be as follows:

\[ p_{i+1}(co, c) \sim Beta\left(\frac{17}{12} + n^{(i)}_{1,co}, \frac{17}{12} + n^{(i)}_{0,co}\right) \]
\[ p_{i+1}(co, t) \sim Beta\left(\frac{17}{12} + n^{(i)}_{1,co}, \frac{17}{12} + n^{(i)}_{0,co}\right) \]
\[ p_{i+1}(at) \sim Beta\left(\frac{11}{6} + n^{(i)}_{1,at}, \frac{11}{6} + n^{(i)}_{0,at}\right) \]
\[ p_{i+1}(nt) \sim Beta\left(\frac{11}{6} + n^{(i)}_{1,nt}, \frac{11}{6} + n^{(i)}_{0,nt}\right) \]
\[ q \sim Dirichlet\left(\frac{8}{3} + n^{(i)}_{1,co} + n^{(i)}_{0,co}, \frac{8}{3} + n^{(i)}_{0,co}, \frac{8}{3} + n^{(i)}_{1,at} + n^{(i)}_{0,at}, \frac{8}{3} + n^{(i)}_{1,nt} + n^{(i)}_{0,nt}\right) \]

The posterior updates are therefore very convenient, even when using a pseudo observations prior, since we still have a piecewise conjugate prior. In the general case, where covariates are included, for example, we will rather be required to use a metropolis step to
update the posterior for each component of the parameter vector. In other words we need to use Metropolis within a block Gibbs sampler.
Bibliography


domized studies,” *Journal of Educational Psychology*, 66.


