

DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY


Fortunes and Misadventures With Parametric Models: They Can Be Confounding, Burdensome and Unstable, Yet Insightful, Powerful and Flexible

Citation

Campos Manzo, Luis Fernando. 2019. Fortunes and Misadventures With Parametric Models: They Can Be Confounding, Burdensome and Unstable, Yet Insightful, Powerful and Flexible. Doctoral dissertation, Harvard University, Graduate School of Arts & Sciences.

Link

http://nrs.harvard.edu/urn-3:HUL.InstRepos:42029487

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

Accessibility

https://accessibility.huit.harvard.edu/digital-accessibility-policy

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

Fortunes and Misadventures with Parametric Models: they can be confounding, burdensome and unstable, yet insightful, powerful and flexible

A dissertation presented
by
Luis Fernando Campos Manzo
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
May 2019

©2019 – Luis Fernando Campos Manzo
All rights reserved.

Fortunes and Misadventures with Parametric Models: they can be confounding, burdensome and unstable, yet insightful, powerful and flexible

Abstract

Statistical models allow us to represent latent structure in data, giving us the ability to wield the power of the unobserved. At the same time, statistical models can confound and trouble us in at least three ways. First, the correspondence between model parameters and physical quantities of interest is not always clear. In the pursuit of causal inference of zero-inflated outcomes, for example, we can employ zero-inflated generalized linear models to harness covariate information and gain precision. However, matching model parameters to causal quantities of interest is not as straight-forward as one would think, especially when involving covariates. I use analysis and simulation to investigate the appropriate use of models here. Secondly, while complex models can capture equally complex structure in data, fitting these models can be a burdensome task. For example, to describe the dependence between time-varying covariates and time-varying outcomes, we might employ latent variable models. The description of these models, however, involve a large number of parameters. I enlist a marginalization strategy that induces a two-stage procedure which greatly simplifies model fitting. Finally, we are not guaranteed to learn about all model parameters when combining models with data. Indeed, it is possible that some parameters are not informed by the data directly at all, but only indirectly through their relationship with other model parameters. I develop strategies for understanding this as the flow of information from data to model parameters using unidentifiable and orthogonal parameters as building blocks. In this thesis, I describe these situations to highlight the difficulties of using parametric models to gain scientific knowledge from data.

Contents

0 Introduction **1**

1 Estimating Causal Effects using Zero-Inflated Outcome Models **5**
- 1.1 Introduction . . . 5
- 1.2 Background . . . 8
- 1.3 Models for Excess Zeros . . . 9
- 1.4 Simple Methods for Estimating Average Treatment Effect (ATE) . . . 13
- 1.5 Methods for Estimating ATE Including Covariate Information . . . 16
- 1.6 Simulation Studies . . . 20
- 1.7 Data Application . . . 31
- 1.8 Concluding Remarks . . . 34

2 Measuring the Effects of Adherence on Time-Varying Health Outcomes **36**
- 2.1 Introduction . . . 36
- 2.2 Data . . . 39
- 2.3 A dynamic model for multivariate health measures . . . 41
- 2.4 Marginal Dynamic Linear Models . . . 44
- 2.5 Application to BP Study . . . 51
- 2.6 Discussion . . . 61

3 A Guided Meditation on the Flow of Information for Model-based Inference **63**
- 3.1 A Normal Example in Double Variation . . . 65
- 3.2 Background Information . . . 67
- 3.3 Likelihood-Based Measures of Information Flow . . . 74
- 3.4 Measures of Information Flow in Bayesian Analysis . . . 82
- 3.5 Discussion . . . 102

Appendix A Technical Details for Chapter 1 **105**
- A.1 Intensive and Extensive Marginal Effects for Two-Part Models . . . 105
- A.2 Mean and Variance of zero-inflated distributions . . . 107
- A.3 Derivations for model-based Confidence Intervals . . . 108
- A.4 Definitions of Quantities used in Simulation Analysis . . . 112

Appendix B Technical Details for Chapter 3 **114**
- B.1 Proofs . . . 114
- B.2 Calculations for Section 3.3 . . . 115

B.3 Calculations for Section 3.4 118
B.4 Measure of Information Conditioning on MAP estimate 126

References **132**

List of Figures

1.1 Covariate adjustments in the analysis of randomized experiments is known to provide precision gains. Here we show two simple methods for covariate adjustment, post-stratification (left) and regression adjustment (right). Here we use three strata of size N_1, N_2 and N_3 respectively. 17

1.2 Data generated from zero-inflated Poisson distributions. (left) data are simulated with the same fraction of structural zeros (i.e. $p_0 = p_1 = 0.25$) and increased sampling means (i.e. $\lambda_0 = 2, \lambda_1 = 6$). (right) data are simulated under a larger fraction of structural zeros and increased sampling means (i.e. $p_0 = p_1 = 0.5$ and $\lambda_0 = 5, \lambda_1 = 9$). 21

1.3 The absolute bias and difference in absolute bias for the Neymanian estimate $\hat{\tau}$ and model-based estimate $\hat{\tau}_{zif}$ for the simulation settings in Table 1.1 (right column) 23

1.4 The confidence-interval coverage for the Neymanian estimate $\hat{\tau}$ and model-based estimate $\hat{\tau}_{zif}$ for the simulation settings in Table 1.1 (right column) 24

1.5 Relative precision of $\hat{\tau}$ compared to $\hat{\tau}_{zif}$, i.e. average $\widehat{SE}$ compared to average $\widehat{SE}_{zif}$ . . . 25

1.6 Relative precision of model-based $\widehat{SE}_{zif}$ compared to Neymanian $\widehat{SE}$ 26

1.7 Data generated from zero-inflated Poisson distributions when a covariate is either not predictive of zero-status (top), or highly predictive of zero status (bottom). (left) Fraction of zeros for $X_i < x$. (right) Relationship between covariate and outcome with a simple linear regression 28

1.8 The absolute bias and coverage for the Neymanian estimate $\hat{\tau}$ and two covariate-adjusted estimates: $\hat{\tau}_{lm}$ and $\hat{\tau}_{zif,x}$. 29

1.9 The relative efficiency of estimates of τ_{sp} and standard error estimates for the simulation settings in Table 1.2 31

1.10 Example Data from an experiment on a MOOC. (top left) Histogram of observed control outcomes, (top right) histogram of observed treatment outcomes, (bottom left) distribution of covariate (age), (bottom right) Relationship between covariate and outcome for treatment and control with a simple linear regression 32

2.1 Example BP and adherence data for a study patient. Blue squares are diastolic BP, red dots are systolic BP, and the black dots near the horizontal axis indicate whether the patient was adherent on a particular day (filled dots indicate adherence, open dots indicate non-adherence). Note that the adherence data for the first nine days of the follow-up period are missing. . . 40

2.2 A 30-day trajectory for a simulated patient from a DLM with a scalar outcome. The bottom of the figure displays the adherence indicator simulated independently with a 90% probability of adherence. The adherence effect is simulated to be large and negative ($\phi = -0.5$). The contribution of the non-dynamic covariates is assumed to be $x_i^T \beta = 130$. 43

2.3 90% central posterior intervals for the overall mean blood pressure as a function of only covariates $(x_i\beta_1, x_i\beta_2)$ for one study subject. The solid lines are the estimated mean DBP and SBP, and envelopes are posterior intervals. 59
2.4 90% central posterior intervals for the mean blood pressure $\{\mu_{it\cdot}\}$ for one study subject. The solid lines are the estimated mean DBP and SBP, and the dashed lines and shading are the pointwise posterior credible intervals. Horizontal dotted lines are drawn at the baseline-estimated mean DBP and SBP. 60

3.1 A representation of the flow of information from the data to the primary parameter θ_1 when the secondary parameter must be estimated (left) and the secondary parameter is assumed (right). 73
3.2 We explore the relationship between the relative redirection penalty $\mathcal{R}_{RP}$ studied with as a function of marginal information and co-information 76
3.3 The relative actual information and redirection penalty for estimating ρ in the correlation variation example for differing fractions of paired observations $\alpha = 0.1, 0.25, 0.5, 0.75, 1$ and for a range of underlying correlation $\rho \in (-1, 1)$. 78
3.4 The flow of information from the data to parameters under the two parameterizations discussed: (left) original parameterization where we're interested in estimating (θ_1, θ_2) and (right) the orthogonal parameterization $(\eta = \theta_1 + \theta_2, \lambda = \theta_2)$ 81
3.5 Posterior distribution for ρ and σ^2 under data $\mathcal{D}_{20}$ generated with with $\rho_0 = 0.5$ and $\sigma_0^2 = 1$. Yellow corresponds to high posterior density. 96
3.6 The two posteriors distribution $p(\sigma^2|\mathcal{D}_{n,\alpha})$ (left) and $p(\sigma^2|\rho_0, \mathcal{D}_{n,\alpha})$ (right) when the true underlying generating values are $\rho_0 = 0.5$, $\sigma_0^2 = 1$, $n = 20$ and $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ . . . 97
3.7 The two posteriors distribution $p(\rho|\mathcal{D}_{n,\alpha})$ (left) and $p(\rho|\sigma_0^2, \mathcal{D}_{n,\alpha})$ (right) when the true underlying generating values are $\rho_0 = 0.5$, $\sigma_0^2 = 1$, $n = 20$ and $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ . . . 98
3.8 Measures of direct and indirect information for σ^2 with replicates shown in colored dots. The mean and two quartiles shown in lines. 99
3.9 Measures of direct and indirect information for ρ with replicates shown in colored dots. The mean and two quartiles shown in lines. 101

B.1 Measures of direct and indirect information for ρ with replicates shown in colored dots. The mean and two quartiles shown in lines. 127
B.2 Measures of direct and indirect information for σ^2 with replicates shown in colored dots. The mean and two quartiles shown in lines. 127

List of Tables

1.1 Settings for Simulation A. Example data sets from the settings highlighted in blue and red are shown in Figure 1.2 left and right, respectively 22
1.2 Settings for Simulation B. Example data sets from the settings highlighted in blue and red are shown in Figures 1.7 top and bottom, respectively 27
1.3 ATE Estimates for Example Intervention in MOOC 34
1.4 Model Output for our model-based covariate-adjusted estimator 34

2.1 Number of BP readings during the study period. 54
2.2 Baseline socio-demographic and health characteristics of the patients in the study cohort. . . 55
2.3 Summaries of covariate effects for the bivariate dynamic linear model. *Effect with 90% posterior interval not containing 0 in DLM, †Significant effect at the 0.1 level in average adherence model, ‡Significant effect at the 0.1 level in dichotomized adherence model. 56
2.4 Measurement error estimates for three models. 57
2.5 Adherence effects for the models considered 58

A.1 Four principal strata defined by potential outcomes from a Hurdle model 107

For Kristina Ng and for my family.
All of my growth in empathy and conscientiousness flows through you.

Acknowledgments

I would like to thank the Graduate School of Arts and Sciences and the Statistics Department at Harvard University. In addition to financial support, they provided a supportive academic environment where ideas can be expressed freely and challenged critically. In particular, I would like to thank my committee members Luke Miratrix, Mark Glickman and Xiao-Li Meng for their support and advice. You all inspire me to continue developing myself into a curious, diligent, and creative researcher. I would like to thank my departmental colleagues for thoughtful and insightful conversations, especially Stephane Shao, Zach Branson, Maxime Rischard, David Jones, Kristen Hunter, Lo-Hua Yuan, and Pierre Jacob. This task would have been an unbearable grind without you. I would be seriously remiss if I did not thank my previous advisers. I would like to thank Danh V. Nguyen and Damla Şentürk for their advice and support during my time at U.C. Irvine and afterward. I am also grateful for Bin Yu, Deborah Nolan, and David Brillinger for all their care and support during my time at U.C. Berkeley. I am thankful to my therapist Rodrigo Barahona who has helped me begin my journey of self-compassion and self-actualization, without which this work would not have been possible. This work would have also been impossible without the steadfast support and constant inspiration from my family. My siblings Andres, Ana, Maria, Alfredo, Juan, Miguel, Flor, and Esteban, and our parents Andres and Lilia motivate me endlessly. You have shown me how to work hard, think big, and be kind. I am driven to inspire my nieces and nephews by showing them all that is possible. Finally, I would like to thank my partner Kristina. You provide me with a wellspring of inspiration and your spirit enriches my life. I take full responsibility for all errors in this dissertation.

A model is a simplification or approximation of reality and hence will not reflect all of reality. ... Box (1976) noted that "all models are wrong, but some are useful." While a model can never be "truth," a model might be ranked from very useful, to useful, to somewhat useful to, finally, essentially useless.

Burnham & Anderson (2002)

0

Introduction

Sure, some models are useful – but even the useful ones can be confusing, burdensome, and unstable. Statistical models allow us to represent the underlying structure in data, giving us the ability to wield the power of the unobserved. This power gives us reason to remain optimistic in the use of statistical models in scientific applications. However, also important are the difficulties that come from the use of models in various contexts. Models can be

1. Insightful, yet confounding: They allow us to incorporate multiple sources of information and explore potential relationships. Yet, the correspondence between model parameters and physical quantities of interest is not always clear.

2. Powerful, yet burdensome: Models can describe complex relationships, yet these models often in-

volve burdensome computation.

3. Flexible, yet unstable: We can build models that describe practically any natural phenomenon. Yet, we are not guaranteed to learn about all model parameters when combining models with data.

In this thesis, we explore these three complications as they arise in the every-day use of statistical models.

Matching parameters with causal estimands

In Chapter 1, we explore how parametric mixture models can help us conduct causal inference analyses. Though the causal estimands of interest are quite simple, they can be complex functions of the model parameters. They can, therefore, lead the analyst into confusing one causal quantity for another – we clarify this. We also explore the potential benefits incorporating covariate information into sophisticated parametric modeling.

Outcomes with excess zeros are a common occurrence in experiments in Massive Open Online Courses (MOOCs) due to significant attrition rates and optional participation. The analysis of these experiments is, therefore, complicated in two ways. Firstly, it is not always clear what estimates derived from models are targeting. We clarify this for models that account for excess zeros. Secondly, using zero-inflated models to estimate causal effects, like the average treatment effect (ATE), should be preferred over simple difference estimates. We show through simulation that model-based and simple difference estimators of ATE exhibit similar performance for the large samples standard in MOOC experiments. Finally, incorporating covariates into ATE estimation is known to improve precision when the covariate is predictive of outcome and worsens estimates only marginally otherwise. Models for excess zeros typically incorporate covariates in non-linear and multifaceted ways. We show that these model-based estimation procedures are only preferable to simple linear regression adjustment when covariate information is strongly predictive of outcomes.

Modeling jointly and computing marginally

In Chapter 2, we use latent variables to model the effect of time-varying covariates on time-varying outcomes. These models allow us to clearly and compactly describe relationships in data. These models are, however, difficult to fit because of the large number of parameters needed in the description of the model. We employ a marginalization strategy that induces a two-stage procedure which greatly simplifies

model fitting.

One of the most significant barriers to medication treatment is patients' non-adherence to a prescribed medication regimen. The extent of the impact of poor adherence on resulting health measures is often unknown, and typical analyses ignore the time-varying nature of adherence. We develop a modeling framework for longitudinally recorded health measures modeled as a function of time-varying medication adherence or other time-varying covariates. Our framework, which relies on Normal Bayesian dynamic linear models (DLMs), accounts for time-varying covariates such as adherence and non-dynamic covariates such as baseline health characteristics. Standard inferential procedures for DLMs associated with sparse and irregularly recorded response data are inefficient. We develop an approach that relies on factoring the posterior density into a product of two terms: a marginal posterior density for the non-dynamic parameters, and a multivariate Normal posterior density of the dynamic parameters conditional on the non-dynamic ones. This factorization leads to a two-stage process for inference in which the non-dynamic parameters can be inferred separately from the time-varying parameters. We demonstrate the application of this model to the time-varying effect of anti-hypertensive medication on blood pressure levels from a cohort of patients diagnosed with hypertension. We compare our model results to ones that incorporate adherence through non-dynamic summaries.

Understanding information flow in Parametric Models

In chapter 3, I explore information flow in parametric models. When we introduce models to data, there is no guarantee that the data can inform our model parameters equally, or at all. We use the parametric structure of models to measure the flow of information from the data to the parameters. We use the cases of unidentifiable and orthogonal parameters as cornerstones on which to build measures of information flow from data to parameters. We discuss some nuances of this situation in both likelihood-based inference and Bayesian inference through two variations of one example.

In the medication adherence problem, we observe adherence daily and outcomes more sparsely, often monthly. Intuitively, the amount of information for the variance parameters in the DLM should diminish with this decrease in observation frequency. We can regard the analysis of data using models as extracting information from the data to inform model parameters. Alternatively, the information from the data flows

to model parameters. Clouding this is the fact that information pass not only *directly* from data to parameters but also *indirectly* among the model parameters. Furthermore, the various parts of the data inform different parts of the parameter space at different rates. In other words, some parameters may drown in information while others remain parched. We explore measures of information flow for decision-making in the model-building process. We focus our efforts on understanding how one might measure this flow information through analysis of example models and provide guidance for model builders.

We must be careful not to confuse data with the abstractions we use to analyze them.

William James

1

Estimating Causal Effects using Zero-Inflated Outcome Models

1.1 Introduction

There are several approaches one can take to analyze randomized experiments that focus on the average treatment effect (ATE) for outcomes exhibiting a large proportion of zeros. For example, one could ignore the structure of the outcomes and use standard simple-difference estimates. Alternatively, one could explicitly model the structure of the outcomes and use model-based estimates. Similarly, one could ignore covariate information or choose to include it in the hopes of increasing precision. A particular challenge

in including covariate information in models for data with excess zeros is that they may be predictive in some ways, suggesting that we include them, but not others, risking precision losses. Deciding to include covariates is, therefore, less straightforward than usual. In this chapter, we investigate the impact of these decisions on estimating the ATE through simulation, analysis and in practice.

Outcomes with a large proportion of zeros commonly occur in randomized experiments across many fields in social science. In education research, for example, we might investigate the causal effect of an intervention on student engagement during a massive online open course (MOOC). If we use the number of posts in online discussions as a measure of engagement, the majority of students who initially sign up for the MOOC will not partake in any online discussions and hence have zero measured engagement. However, the distribution of engagement levels for those who do participate is typically right-skewed with a long tail (Lamb et al., 2015). Various statistical models have been developed to handle count data with excess zeros, for example, zero-inflated Poisson models, hurdle models and other two-part models (like the Tobit model). These models have different data-generating interpretations and are usually selected to reflect reality somewhat.

As in many other applications, pre-treatment covariates are available that are predictive of outcomes and may help increase the precision of estimates. In the previously mentioned example, age correlates with engagement levels among those who participate, but it is not very predictive of participation. Including covariates that are predictive of outcome can help us more precisely estimate the overall treatment effect even if the model is misspecified. See, e.g. Lin (2013) for a discussion on linear regression adjustment. This advice, however, is linked to simple models like linear regression adjustment. Naturally, we question whether or not this holds in models for outcome data with excess zeros where the relationship between covariates and outcomes are typically non-linear and multifaceted.

Despite a plethora of zero-inflated outcome data in real-world applications, there is a scarcity of publications in the causal inference literature that discuss the special considerations we need to address when estimating treatment effects for data with excessive zeros. The majority of articles dealing with zero-inflated outcomes focus either on zero-inflated count models (Bohning et al., 1999; Yau & Lee, 2001) or on two-part models like the Tobit model (Héroux et al., 2014; DeSantis et al., 2014). Recently, Staub (2014); Lee (2017) consider extensive and intensive causal effects via principal stratification and model-

ing, while Keele & Miratrix (2018) consider randomization-based inference for an assortment of hypothesis tests generally tailored to zero-inflated outcome data. Most previous work is fundamentally rooted in the zero-inflated Poisson model. Except for a few works (DeSantis et al., 2014; Lee, 2017; Keele & Miratrix, 2018), they generally do not define causal effects in terms of the Neyman-Rubin potential outcomes framework.

We link the causal inference literature with models for data with excess zeros. We show that we can define these models generically in terms of mixture models that encompass a wide range of models for data with excess zeros. We also derive the causal estimands of the ATE generically for these models and discuss model-free and model-based methods for estimation. We conduct a simulation study to compare the finite sample performance of the confidence intervals resulting from each form of estimation. Finally, we investigate the use of covariates in modeling outcomes with excess zeros. We derive the causal estimand of the ATE for mixture models that include covariates and investigate the impact they have on estimator precision. Throughout, we provide methodological recommendations for how to think about and conduct causal inference under the scenario of a completely randomized experiment.

The structure of the rest of this chapter is as follows: Section 1.2 gives an overview of the Neyman-Rubin causal model; Section 1.3 introduces basic models for handling data with excess zeros, and reviews previous causal work that has been done for zero-inflated data; Section 1.4 defines the ATE estimand and two ways of defining an estimator for it, along with methods to calculate their sampling variance; Section 1.5 discusses methods for estimating the ATE that include covariate information; Section 1.6 describes and shows our simulation results, Simulation A investigate simple methods described in Section 1.4 while simulation B investigates the use of covariate information with methods described in Section 1.5; Section 1.7 investigates an application to an experiment in a MOOC; and finally, in Section 1.8, we end with a discussion.

1.2 Background

1.2.1 Causal Inference Framework

We ground ourselves in the Neyman-Rubin model of potential outcomes (Holland, 1986) as a causal model. The fundamental idea underlying the Neyman-Rubin causal model is the assertion that causality is tied to an *action* (or treatment, intervention, or manipulation), applied to a *unit* (Imbens & Rubin, 2015). Each unit has set of potential outcomes that are *a priori* observable. That is, we can observe any of these outcomes if the unit were to receive the corresponding treatment level. If there are two possible treatments for unit i, $T_i = 0, 1$, we denote these potential outcomes as $(Y_i(0), Y_i(1))$. However, once a treatment is applied, say $T_i = 1$, we observe at most one potential outcome, in this case, $Y_i(1)$. This motivates the view that causal inference is, in essence, a missing data problem (Rubin, 1974).

It is necessary to observe multiple units in order to draw causal inferences. If we assume that treatment assignment for any particular unit does not affect the potential outcomes of any other unit (what Rubin (1980) calls the "stable unit treatment value assignment" (SUTVA) assumption), then observing multiple units allows us to impute missing potential outcomes and thereby estimate a causal treatment effect. Furthermore, the treatment assignment mechanism is crucial for inferring causal effects. From a finite sample population viewpoint the observed outcomes are random, but only because of the treatment randomization used. If we wish to draw inference on the larger "super-population" then our observed outcomes are random due to two sources of variation: the treatment randomization and random sampling from the super-population.

Because in most real-life situations we are interested in saying something about a broader population, not just the finite sample we have at hand, we focus here on assessing super-population treatment causal effects. This view aligns with model-based inferences, which are inherently super-population inferences because they postulate that the potential outcomes are generated from a specified model, representing the super-population. Our simulations reflect this idea.

1.2.2 Average Treatment Effect Causal Estimand

Causal treatment effects come in many different flavors. One of the most common is the average treatment effect (ATE) estimand. In our setting of a completely randomized experiment, ATE analyses are validated by the randomization of the assignment to treatment, without the need for additional assumptions beyond no unit interference (SUTVA). The main critique of ATE analyses is that they only answer questions about the causal effect of the *assignment* to treatment and not the causal effect of *receipt* of treatment. However, from a high-level policy perspective, when one has no control over compliance with the treatment or intervention, ATE gives a reasonable estimate of a causal effect that policymakers can directly observe.

For units $i = 1, \ldots, N$, let T_i denote the treatment indicator for the i^{th} unit, and let $Y_i(T_i = t)$ denote the potential outcome of the i^{th} unit under treatment level $t \in \{0, 1\}$. For simplicity, we consider only two treatment levels. Our estimand of interest is then:

$$\tau_{sp} = \mathbb{E}_{sp}[Y_i(1) - Y_i(0)] \tag{1.1}$$

where the *sp* subscript notation indicates taking an expectation over the distribution generated by random sampling from the super-population as well as the random treatment assignment. Importantly, note that we defined our estimand of interest without making any model assumptions.

1.3 Models for Excess Zeros

Methodological developments for data with excess zeros have generally occurred in two different settings. One is in dealing with count data that is not accurately represented by standard count models because of a large proportion of zeros. For example, Lambert (1992) found it helpful to model the number of defects in manufacturing as a mixture of a standard count model, here a Poisson distribution, and a point-mass at zero. The other setting where we use models with excess zeros is in Econometrics where outcomes of interest, e.g., income, spending on classes of items, are truncated at zero. Tobin (1958) consider modeling the total expenditures on classes of items, e.g., luxury goods, as a mixture of a point mass at zero and a dis-

tribution (possibly continuous) truncated at zero. Indeed, recognizing that a straight line can not possibly describe the relationship between income and expenditure on luxury goods inspired the Tobit model. The two settings described may seem similar but have very different data-generating interpretations that lend themselves to different applications. Below we introduce notation that describes these scenarios before discussing a general framework that ties them together.

1.3.1 Zero-Inflated Models

Given some set of discrete counts, $Y_1, \ldots Y_n$, and a model for those counts, we say the data has *zero-inflation* when the proportion of Y_i s equaling zero is greater than what we would expect from a standard count model. For example, with a $Poisson(\lambda)$ distribution, we expect on average $e^{-\lambda}$ proportion of zeros; that's approximately 13.5% when $\lambda = 2$ and about 5% when $\lambda = 3$. Hence, if we expect $Y_1, \ldots, Y_n$ to come from a $Pois(\lambda = 3)$ model, but find that much more than 5% of the Ys are zero, then we say we have zero-inflated data. A complementary notion is that of zero-deflation in which we observe *fewer* zeros than would be expected under a count model*.

One representation of data with excess zeros is a mixture model. One set of observations may be zero according to a random event (these are called *sampling zeros*) while another set is necessarily zero (termed *structural zeros*). The associated distribution is a mixture of an ordinary count model with a degenerate point mass at zero. For example, the *zero-inflated Poisson (ZIP) model* (Lambert, 1992) assumes that

$$Y \sim \begin{cases} 0 & \text{w.p.} \quad p \\ Pois(\lambda) & \text{w.p.} \quad 1-p \end{cases}$$

*This notion can be handled with some models, but not others as discussed in the following sections

The unconditional probability distribution[†] has

$$P(Y = 0) = p + (1 - p) \exp^{-\lambda}$$
$$P(Y = k) = (1 - p)\frac{\exp^{-\lambda} \lambda^k}{k!}, \quad k = 1, 2, \ldots$$

Since in this model an instance of zero can come either from the zero distribution with probability p or from the Poisson distribution with probability $(1 - p)e^{-\lambda}$, the probability of observing a zero is at least as high as in the Poisson. One complicating feature of this model is the fact that upon observing a zero, we do not know if it is a structural or sampling zero.

This definition is given explicitly for Poisson distributions for educational purposes, but the structure holds for all zero-inflated distributions. Other common zero-inflated distributions include the zero-inflated negative binomial model to address overdispersion, as well as zero-inflated binomial and zero-inflated beta-binomial models to handle bounded counts. We generally denote a zero-inflated distribution as $Y \sim ZIF(p, \theta)$. Here F is a random variable parametrized by θ, hence the parameter set of the distribution describing our outcome Y is (p, θ). The mean parametrizes the Poisson distribution, $\theta = \lambda$, while two parameters would parametrize the Negative binomial, $\theta = (r, q)$. We denote the mean and variance of the sampling distribution, F, as μ and σ^2 respectively. For example, the Poisson sampling distribution has both mean and variance equal to the parameter, $\mu = \sigma^2 = \lambda$, while the Negative binomial has mean $\mu = \frac{qr}{1-q}$ and variance $\sigma^2 = \frac{qr}{(1-q)^2}$.

We can include covariates into zero-inflated models; these models are called zero-inflated generalized linear models (ZI-GLMs). As in all GLMs, we relate the conditional mean of the outcome to covariates via link functions, but here two link functions are needed. One link is needed to relate covariates to the conditional probability of being a structural zero, i.e., relating x to p, and a second link relating covariates to the conditional mean of the sampling distribution, i.e., relating x to μ. In the zero-inflated Poisson distribution, we use a logit link for the zero-part and a log link for the non-zero mean part, i.e. $\text{logit}(p) = x\beta$ and $\log(\mu) = x\gamma$. We can already see the challenge in including covariates for zero-inflated models as we can include them in two places.

[†]IF $Z \sim Bern(p)$ and $Y^* \sim Pois(\lambda)$, then in our notation $Y = (1 - Z)\, Y^* \sim ZIP(p, \lambda)$

1.3.2 Latent-Variable Models

An alternative way to model excess zeros is through models that incorporate latent-variables. This representation is common in Econometrics with models like the Tobit, or two-part models (Heckman, 1979). The observed outcome Y is modeled as a truncated version of a latent variable $Y^\star$ where Y is observed as $Y^\star$ only if it is positive and zero otherwise, i.e

$$Y = \begin{cases} 0 & \text{if} \quad Y^\star < 0 \\ Y^\star & \text{o.w.} \end{cases}$$

The latent variable has a distribution that can be related to covariates. For example, the Tobit model simply relates $Y^\star$ to covariates x linearly so that with $\varepsilon \sim N(0, \sigma^2)$,

$$Y^\star = x\beta + \varepsilon \quad \text{and} \quad Y = Y^\star \, \mathbb{1}\{Y^\star > 0\}.$$

This linear relationship implicitly relates covariates to both the zero-probability p and the non-zero mean μ. It places a linear relationship on the non-zero mean as a function of x, i.e., an identity link, and a Normal-quantile relationship between x and p, i.e., a probit link. Furthermore, the interpretation of β involves both changes in the probability of zero *and* non-zero mean. There are a large number of extensions to Tobit models that allow for complex relationships between covariates and latent variables, more latent variables and more censoring (Schnedler, 2005). While the interpretation of these models makes them useful, the inability to control the links for zero-probability and nonzero-mean separately makes working with them challenging.

1.3.3 General Mixture Models

The two models discussed Sections 1.3.1 and 1.3.2 evolved in different fields of science with different requirements in mind, but can both be written as mixture models. Let $F(\theta)$ represent a random variable

parameterized by θ, then Y distributed as

$$Y \sim \begin{cases} 0 & \text{w.p.} \quad p(X) \\ F(\theta(X)) & \text{w.p.} \quad 1 - p(X) \end{cases}$$

is a mixture of a sampling distribution and a point-mass at zero. We can also relate the parameters in the model to covariates by specifying link functions that relate the covariate to p (signified by $p(X)$) and the mean of the sampling distribution (usually written $\mu(\theta(X))$). If we further restrict F to take strictly non-negative values, then both models in Sections 1.3.1 and 1.3.2 are special cases.

Taking $F(\theta) \equiv \text{Pois}(\lambda)$, and $p(X) = p$ we recover the zero-inflated Poisson model discussed in Section 1.3.1. Again, for this we typically relate the covariates to $p(X)$ through a logit link and to $\lambda(X)$ through a log link. The Tobit model described in Section 1.3.2 uses $F(\theta(X)) \equiv \mathcal{N}(X\beta, \sigma^2)$, implying an identity link for the mean and a probit link to relate X to p. The second link is implicit in the model, but can be seen easily by denoting Φ as the c.d.f. of a standard Normal distribution and seeing that

$$p(x) = \mathbb{P}(Y = 0 | X = x) = \mathbb{P}(x\beta + \varepsilon < 0) = \Phi\left(\frac{x\beta}{\sigma}\right).$$

This general formulation allows us to explore the models discussed under one umbrella. This formulation also pinpoints the general structure of the decisions needed for an analysis of data with excess zeros: (1) specify an outcome model that describes non-zero outcomes, (2) specify a link function that describes the relationship between covariates and average of non-zero outcomes and (3) specify a link function that relates covariates to the probability of zero. For the remainder of the chapter, we will primarily be working with zero-inflated distributions to derive our estimators and analyze simulation results. However, we will keep latent-variable models in mind as these models behave similarly.

1.4 Simple Methods for Estimating Average Treatment Effect (ATE)

In this section, we present simple methods for estimating the ATE estimand. In Section 1.4.1 we begin by presenting the most straightforward method we consider, a simple difference estimator with Neymanian

confidence intervals. This method assumes that the averages of our outcome within treatment groups are approximately normal. This assumption begs the question as to whether incorporating population structure into our estimates improves inference. So, we go on to present our first model-based method for an estimate and confidence interval of the ATE in Section 1.4.2.

1.4.1 Neymanian Estimation and Confidence Intervals

One simple, intuitive way to estimate τ_{sp} is to use a method-of-moments estimator:

$$\hat{\tau} = \frac{1}{N_1} \sum_{i:T_i=1} Y_i(1) - \frac{1}{N_0} \sum_{i:T_i=0} Y_i(0) \tag{1.2}$$

where N_1 is the number of units assigned to active treatment, N_0 is the number of units assigned to control, and $N_1 + N_0 = N$. Because $\hat{\tau}$ is a difference of averages, we can use asymptotic approximations to build confidence intervals by way of the Central Limit Theorem. We can also apply a small-sample correction using a t-distribution quantile instead of a standard Normal multiplier. Denoting the sampling variability with

$$\widehat{SE} = \sqrt{\frac{s_1^2}{N_1} + \frac{s_0^2}{N_0}}, \tag{1.3}$$

our resulting Neymanian-based $(1-\alpha) * 100\%$ confidence interval (CI) is then:

$$\hat{\tau} \pm z_{\alpha/2}\, \widehat{SE} \tag{1.4}$$

where s_1^2 and s_0^2 are the sample variances for the treatment and control outcomes, respectively. This method is applied liberally throughout the literature without regard for the underlying data distribution. A goal of this paper is to justify this naive CI's use in practice. It may seem sloppy to use such a simple method despite observing zero-inflated outcomes; however, we will see that the Normal approximation works quite well, even with noticeable zero-inflation. However, even with this strong and general result, it is reasonable to believe that incorporating information from the data-generating mechanism into the estimation

process could only help us in terms of inference. With this in mind let us introduce estimation procedures that incorporate such structure.

1.4.2 Estimation using Zero-Inflated Models

If we describe our data in terms of structural and sampling zeros and think of our data as coming from a general mixture model (as in Section 1.3.1), we can then write the distribution of our potential outcomes as, say, $Y_i(t) \sim ZIF(p_t, \theta_t)$. Recalling that the mean and variance of $F(\theta_t)$ are, indeed, functions of θ_t, we denote them as $\mu_t \equiv \mu_t(\theta_t)$ and $\sigma_t^2 \equiv \sigma_t^2(\theta_t)$ respectively. Hence, the super-population ATE under the this potential outcome model is

$$\tau_{sp} = \mathbb{E}[Y_i(1) - Y_i(0)] = (1 - p_1)\mu_1 - (1 - p_0)\mu_0.$$

We write this explicitly because it is important to note that parameters of interest and model parameters are not always the same. For example, comparing the sampling means under different treatment regimes, e.g., $\mu_1 - \mu_0$, may be a useful exploratory analysis, but the causal estimand validated by randomization is the ATE. With these results in hand, we can construct estimates and confidence intervals for the population average ATE τ_{sp} using these mixture models.

We fit our model to the observed data[‡] to obtain maximum likelihood (ML) estimates of the model parameters, $(\hat{p}_0, \hat{p}_1, \hat{\theta}_0, \hat{\theta}_1)$. Since our causal estimand of interest is simply a function of these model parameters, we get a ML point estimate for τ_{sp} using

$$\begin{aligned} \hat{\tau}_{zif} &= (1 - \hat{p}_1)\ \mu_1(\hat{\theta}_1) - (1 - \hat{p}_0)\ \mu_0(\hat{\theta}_0) \\ &= (1 - \hat{p}_1)\hat{\mu}_1 - (1 - \hat{p}_0)\hat{\mu}_0 \end{aligned} \tag{1.5}$$

Furthermore, because $\hat{\tau}_{zif}$ is the MLE of τ_{sp}, we use the Delta Method to derive an approximate distribution of $\hat{\tau}_{zif}$ and use this to build a corresponding confidence interval. Let $g(p_0, p_1, \theta_0, \theta_1) = (1 - p_1)\ \mu_1(\theta_1) - (1 - p_0)\ \mu_0(\theta_0)$ and ∇ represent the partial derivatives with respect to the parameters. We can

[‡]The observed data are $Y_i^{obs}(T_i) = T_i\, Y_i(1) + (1 - T_i)\, Y_i(0)$, for $i = 1, ..., N$.

approximate the sampling variance of $\hat{\tau}_{zif}$ as

$$\widehat{SE}_{zif} = \sqrt{\nabla g(\hat{p}_0, \hat{p}_1, \hat{\theta}_0, \hat{\theta}_1)^T \, \hat{\Sigma} \, \nabla g(\hat{p}_0, \hat{p}_1, \hat{\theta}_0, \hat{\theta}_1)} \tag{1.6}$$

where $\hat{\Sigma}$ is the covariance matrix of our parameters. We can then calculate a model-based confidence interval for τ_{sp} as

$$\hat{\tau}_{zif} \pm z_{\alpha/2} \, \widehat{SE}_{zif} \tag{1.7}$$

A complete derivation is given in Appendix A.3.

1.5 Methods for Estimating ATE Including Covariate Information

Introducing covariate information into estimation procedures at the analysis stage can increase the precision of estimates even when they are not used in the design. In essence, covariates that are predictive of the treatment effect can help increase the precision of estimates built using post-stratification (Miratrix et al., 2013), linear regression (Lin, 2013), and even high-dimensional regression (Bloniarz et al., 2016). Consider the following simulated dataset that echoes some of the trends observed in the MOOC example. This example will help us see how post-stratification and regression might increase precision, but will also point out some of the difficulties of working with zero-inflated data.

Figure 1.1 shows simulated data (of total size $N = 100$) where the covariate is certainly predictive of the outcome. Subjects with high values of the covariate seem to have large outcomes when not zero, but the zero outcomes seem to be distributed equally across the covariate. Shown are two possible methods of analysis, a post-stratification analysis (left) and a linear regression adjustment with interactions (right). Both methods capture the overall trend that treatment outcomes are higher than control outcomes but fail to capture relationships for zero-inflated data adequately.

The model implied in a post-stratification analysis with K strata is a piece-wise step function with pre-determined cut-points $(x_0, ..., x_K)$ and unknown heights for treatment and control separately (μ_{k1}, μ_{k0} for $k = 1, ..., K$). Figure 1.1 (left) shows the model fit to this data for three strata of equal length with sam-

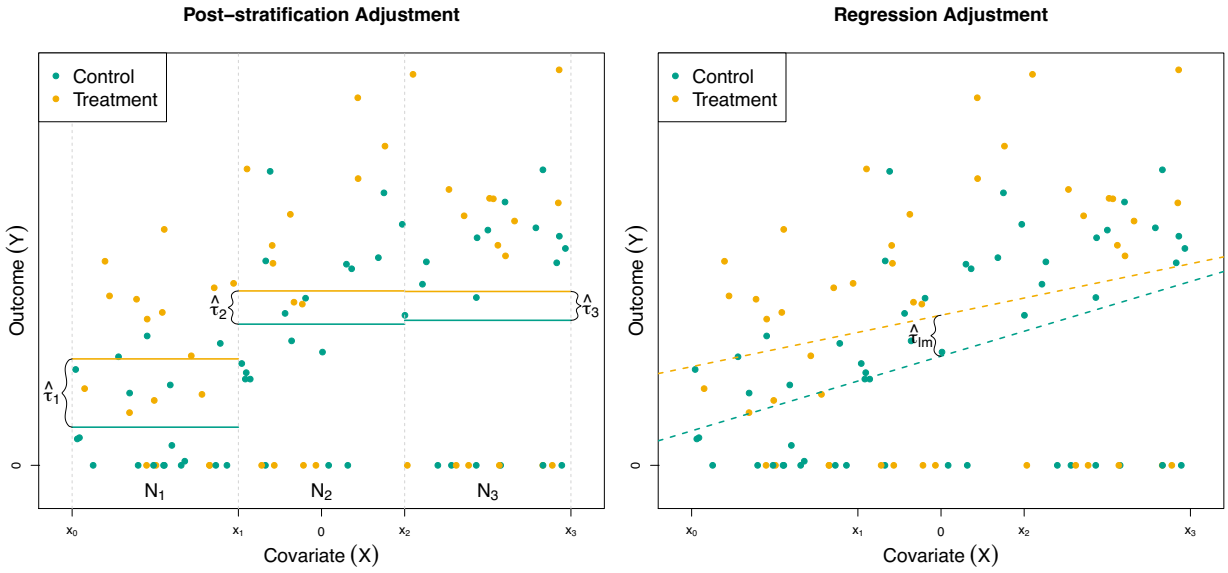


Figure 1.1: Covariate adjustments in the analysis of randomized experiments is known to provide precision gains. Here we show two simple methods for covariate adjustment, post-stratification (left) and regression adjustment (right). Here we use three strata of size N_1, N_2 and N_3 respectively.

ple sizes (N_1, N_2, N_3) and estimated treatment effects $\hat{\tau}_k = \hat{\mu}_{k1} - \hat{\mu}_{k0}$ for $k = 1, ..., 3$. Let $\hat{\mu}_1$ and $\hat{\mu}_0$ be the overall means for treatment and control outcomes. Typically, there are gains in precision when the variability within strata is smaller than the overall variability (Miratrix et al., 2013). In other words, when $\hat{\mu}_{k1}$ and $\hat{\mu}_{k0}$ better describe the data within the group than do $\hat{\mu}_1$ and $\hat{\mu}_0$. For the first strata, $\hat{\mu}_{11}$ and $\hat{\mu}_{10}$ seem to describe the data better than the overall means and, indeed, the variability is smaller within the group than overall. The data in the third strata tell a different story. The large fraction of zeros decrease the overall means, and $\hat{\mu}_{31}$ and $\hat{\mu}_{30}$ no longer seem to describe the structure of the data within the third strata. Furthermore, the variability within this strata is as wide as the overall variability. Nevertheless, because some strata exhibit smaller variability within than across strata, post-stratification may help increase precision.

We can also include covariates with linear regression adjustment, Figure 1.1 (right) shows such an analysis. A linear regression analysis with interactions models the conditional expectation of the potential outcomes given a particular value of the covariate, i.e. we are modeling $\tau_{sp}|x = \mathbb{E}[Y_i(1)|X = x] - \mathbb{E}[Y_i(0)|X = x]$. Similar to post-stratification, we are considering treatment effect estimates for subjects with similar values of the covariate. So, if the covariate is related to the outcome, we will gen-

erally gain precision. If the covariate is not predictive of the outcome, the precision may be hurt in finite samples but is not a problem asymptotically (Lin, 2013). Similar to post-stratification, the linear model represents the data well for negative values of the covariate and poorly for positive values around x_3.

Including covariates that are predictive of outcomes likely increases precision even when the outcomes exhibit zero-inflation. However, as discussed in both post-stratification and regression adjustment, including covariates that are not useful may hurt precision. One the one hand, we must take extra care when considering covariate adjustment in zero-inflated models. While it is possible that covariates are predictive in one part of the model, they may not be predictive in the other. It is, therefore, less apparent when the potential gains are worth the risks. On the other hand, by using models that do not represent the data, we may be leaving precision on the table. In this example, it is clear that a linear model for the non-zero outcomes would provide a much better fit to the data and fit may lead to more precise estimates of the ATE. In the following sections, we consider not just whether we should include covariates in ATE estimation or not, but also the additional question of how.

1.5.1 ATE and Zero-Inflated Generalized Linear Models

In this section, we discuss the use of zero-inflated generalized linear models in the process of estimating the ATE, an extension of Section 1.3.1 to include covariates. When including covariate information, the model-based conditional ATE estimand for a given value of x under zero-inflated distributions remains largely the same, i.e.

$$\tau_{sp}|x = (1 - p_1(x))\mu_1(x) - (1 - p_0(x))\mu_0(x).$$

The covariate can change the probability of a structural zero and the mean of the sampling distribution for treatment and control separately.

If we keep $p_t(x)$ constant with x and take $p = p_1(x) = p_0(x)$, we have that $\tau_{sp}|x = (1-p)\left[\mu_1(x) - \mu_0(x)\right]$. If $\mu_t(x) = x\beta_t$ for $t = 0, 1$, this quantity is a scaled version of what is typically modeled with linear regression. As previously discussed, Lin (2013) studies this situation in detail[§]. We will, therefore, focus

[§]For related discussion on general link functions see Rosenblum & van der Laan (2010)

our efforts on understanding the impact of covariates on the zero-part. Say, then, that the covariate only impacts the probability of a structural zero, i.e. $\mu_t = \mu_t(x)$. In this case, the unconditional ATE estimand is

$$\tau_{sp} = \left(1 - \int p_1(x) dF_X(x)\right) \mu_1 - \left(1 - \int p_0(x) dF_X(x)\right) \mu_0.$$

Here, $dF_X(x)$ is the distribution of our covariate X. The ATE estimand is a more complex function of the parameters and covariates. Nevertheless it is still straightforward to calculate an MLE of τ_{sp} using zero-inflated GLMs.

For example, a zero-inflated Poisson distribution with logit-link for the zero-part, i.e. $\mu_t = \lambda_t$ and $\text{logit}(p_t) = \alpha_t + \beta_t x$ for $t = 0, 1$, the ATE estimand is

$$\begin{aligned}\tau_{sp} &= \left(1 - \int \text{expit}(\alpha_1 + \beta_1 x) dF_X(x)\right) \lambda_1 - \left(1 - \int \text{expit}(\alpha_1 + \beta_1 x) dF_X(x)\right) \lambda_0 \\ &= (1 - E(\alpha_1, \beta_1)) \lambda_1 - (1 - E(\alpha_0, \beta_0)) \lambda_0.\end{aligned}$$

Here, $E(\alpha_t, \beta_t)$ is simply the link function averaged over the distribution of the data. If the covariate is standard Normal (as in our simulation studies), it is the mean of a logit-Normal distribution[¶]. The moments of a logit-Normal distribution are computationally tractable using numerical integration. The GLM fitting procedure returns parameter estimates and their covariances. The maximum-likelihood estimate of τ_{sp} is

$$\hat{\tau}_{zif,x} = \left(1 - E(\hat{\alpha}_1, \hat{\beta}_1)\right) \hat{\lambda}_1 - \left(1 - E(\hat{\alpha}_0, \hat{\beta}_0)\right) \hat{\lambda}_0.$$

The sampling variability $\widehat{SE}_{zif,x}$ can be approximated using the Delta method (details in Appendix A.3). The confidence interval takes the familiar form as

$$\hat{\tau}_{zif,x} \pm z_{\alpha/2} \widehat{SE}_{zif,x}. \tag{1.8}$$

[¶]A random variable is distributed logit-Normal if it is Normally distributed after taking the logit transformation

We show these quantities specifically for the logit-link, but the structure holds generally. We use this example in the simulation studies that follow.

1.6 Simulation Studies

In Section 1.4.1, we introduced the classic Neymanian estimator of the super-population ATE and a corresponding confidence interval. We then went on to derive a general model-based estimator and confidence interval using the zero-inflated mixture models in Section 1.4.2 and discussed a covariate-adjusted model-based estimator in Section 1.5.1. In this section, we investigate the finite-sample properties of these estimators in the hopes of critically analyzing the necessity of model-based estimators.

The Central Limit Theorem gives us confidence that for large enough samples, we will not need models more complicated than the Normal distribution for means. Of course, we would hope that incorporating more information about the nature of the data would improve the efficiency of our inference. Furthermore, models allow us to incorporate potentially relevant covariates that may also help our inferential procedures. We will run several simulations meant to inspect the performance of methods for estimating the ATE. We will first study the impact of zero-inflation on the performance of typical Neymanian estimation procedures compared to the model-based estimation procedures without the use of covariates. This will be Simulation A. We then study the impact of including covariate information into the estimation procedures by comparing linear regression adjustment (an extension of Neymanian procedures) to the zero-inflated generalized linear models. In order to assess the utility of covariates unique to zero-inflated data, we introduce a covariate that helps predict whether or not an observation is a zero. This will be called Simulation B.

1.6.1 Simulation A

The purpose of our first simulation study is to understand how zero-inflated data impacts ATE estimation. The Neymanian procedures rely heavily on the assumption of Normality of means, so we expect poor performance for small samples and highly skewed data. Figure 1.2 illustrates two simulation settings for zero-inflated outcome data. These data are simulated from two zero-inflated Poisson mixtures: one for

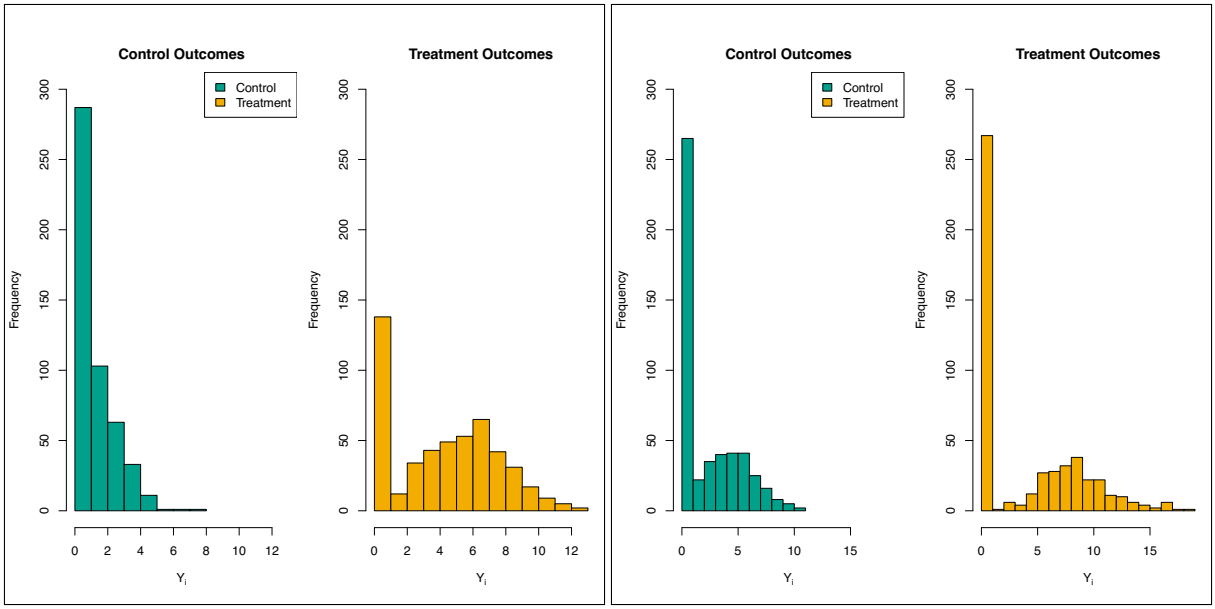


Figure 1.2: Data generated from zero-inflated Poisson distributions. (left) data are simulated with the same fraction of structural zeros (i.e. $p_0 = p_1 = 0.25$) and increased sampling means (i.e. $\lambda_0 = 2, \lambda_1 = 6$). (right) data are simulated under a larger fraction of structural zeros and increased sampling means (i.e. $p_0 = p_1 = 0.5$ and $\lambda_0 = 5, \lambda_1 = 9$).

control, $Y_i|T_i = 0 \sim ZIP(p_0, \lambda_0)$, and another for treatment $Y_i|T_i = 1 \sim ZIP(p_1, \lambda_1)$. Figure 1.2 (left) shows treatment and control groups that look drastically different. The control group seems to exhibit an extreme amount of zero-inflation; in this example, 35% of control outcomes are zeros while only 27% of treatment outcomes are zero. However, once we take into account the fraction of zeros due to the sampling distribution, a $Pois(2)$ in control and $Pois(6)$ in treatment, we find that the fraction of structural zeros to be similar. Specifically, we expect $(1 - 0.25)e^{-2} = 10.15\%$ and $(1 - 0.25)e^{-6} = 0.18\%$ sampling zeros in the control and treatment groups, respectively. So, approximately 25% and 26.8% of the zeros are predicted to be structural in the control and treatment groups, respectively.

Figure 1.2 (right) investigates a scenario where we have a large fraction ($p_1 = p_0 = 0.5$) of zeros and large Poisson means. We end up, in this case, with minimal overlap between the structural zeros and the Poisson distribution. We expect about $(1 - 0.5)e^{-5} = 0.34\%$ and $(1 - 0.5)e^{-9} = 0.006\%$ sampling zeros in the control and treatment groups, respectively. Under this setting, it is readily apparent that the data is coming from some sort of mixture distribution.

Table 1.1 summarizes the simulation settings we implemented. The two sets of simulations highlighted in blue and red correspond to the settings for the example data shown in Figure 1.2 (left) and (right) re-

	Fraction of structural zeros					
	$p = 0.25$			$p = 0.5$		
Simulation	λ_0	λ_1	τ_{sp}	λ_0	λ_1	τ_{sp}
Setting (1)	2	2	0	2	2	0
Setting (2)	5	5	0	5	5	0
Setting (3)	20	20	0	20	20	0
Setting (4)	2	6	3	2	6	2
Setting (5)	5	9	3	5	9	2
Setting (6)	20	24	3	20	24	2

Table 1.1: Settings for Simulation A. Example data sets from the settings highlighted in blue and red are shown in Figure 1.2 left and right, respectively

spectively. For each of the 12 simulation settings described in Table 1.1, we generated 2000 replicated data sets for each level of total sample size ($N = 40, 60, 100, 180, 400, 800, 1000, 2000$). We simulated data independently from zero-inflated Poisson models: control units $Y_i^{obs} \sim ZIPois(p_0, \lambda_0)$, for $i = 1, ..., N_0$, and treated units $Y_j^{obs} \sim ZIPois(p_1, \lambda_1)$, for $j = 1, ..., N_1$, with $N_0 + N_1 = N$. To simplify our simulation design, we only considered a balanced randomized experiment, so we had $N_0 = N_1 = N/2$. Also, we fixed the expected proportion of structural zeros in the control and treatment groups to be the same, i.e. $p_0 = p_1 = p$. The left and right columns of Table 1.1 refer to two simulation sets with differing values of p. We used these data to construct point estimates and standard errors for the population ATE $\tau_{sp} = (1 - p_1)\lambda_1 - (1 - p_0)\lambda_0$ using both the Neymanian $\hat{\tau}$ and model-based approaches $\hat{\tau}_{zif}$.

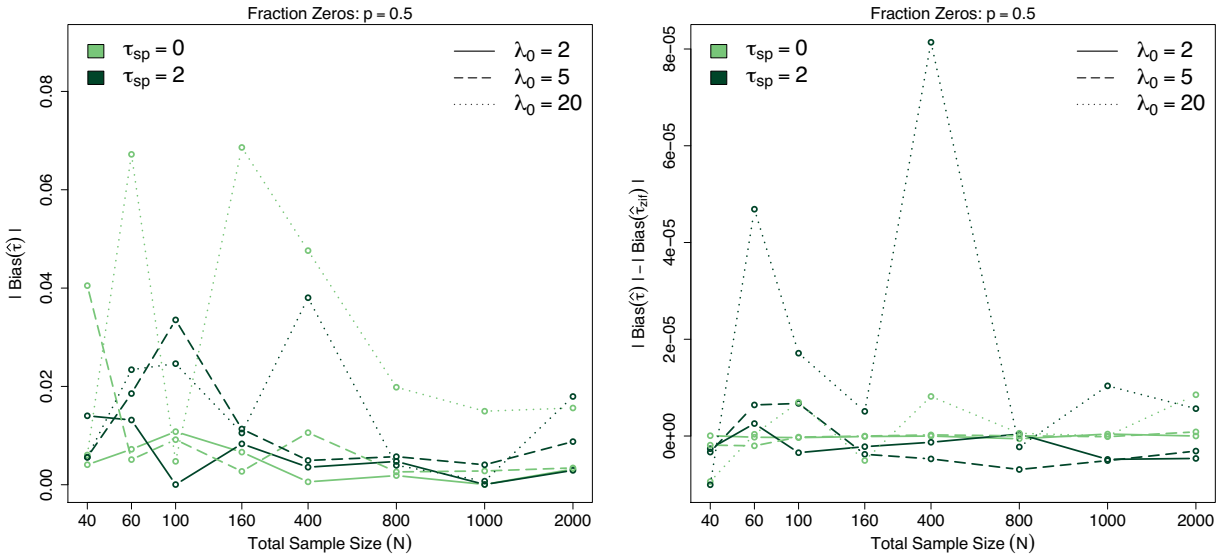
For each simulation setting $(\lambda_0, \lambda_1, p, N)$ we will have 2000 replicated data sets. For each data set $r = 1, ..., 2000$ we will have the following.

1. $\hat{\tau}^{(r)}$ and $\widehat{SE}^{(r)}$ – the Neymanian estimate and standard error described in Section 1.4.1
2. $\hat{\tau}_{zif}^{(r)}$ and $\widehat{SE}_{zif}^{(r)}$ – the model-based estimate and standard error described in Section 1.4.2

Specifying a model can improve estimation and inference by reducing bias or increasing precision in estimation (if correctly specified). These, together, can sometimes lead to better coverage. On the other hand, incorrectly specifying a model can lead to the opposite effects, biased estimates, and worse coverage. As we will show, assuming a zero-inflated model to estimate the ATE does not tend to help in terms of bias

and coverage, but it does impact the precision of estimates, sometimes drastically. See Appendix A.4 for a definition of the terms used in the following analysis.

The Neymanian estimator $\hat{\tau}$ is unbiased, so it is no surprise that the bias is small in relative terms (Figure 1.3 (left)). The model-based estimator is consistent, so it also shows very little bias. Figure 1.3 (right) shows the difference in bias between the two methods. The largest deviations occur in for large sampling means where they still deviate by less than 0.008%. With a Poisson distribution, a large mean implies a variance. Data with high variability gives more opportunity to over or underestimate the mean. The model does not seem to decrease bias in our simulations even in the most extreme cases where the data are severely non-normal (i.e. $p = 0.5, \lambda_0 = 20, \lambda_1 = 24$). This stability stands as a testament to the power of simple averages.

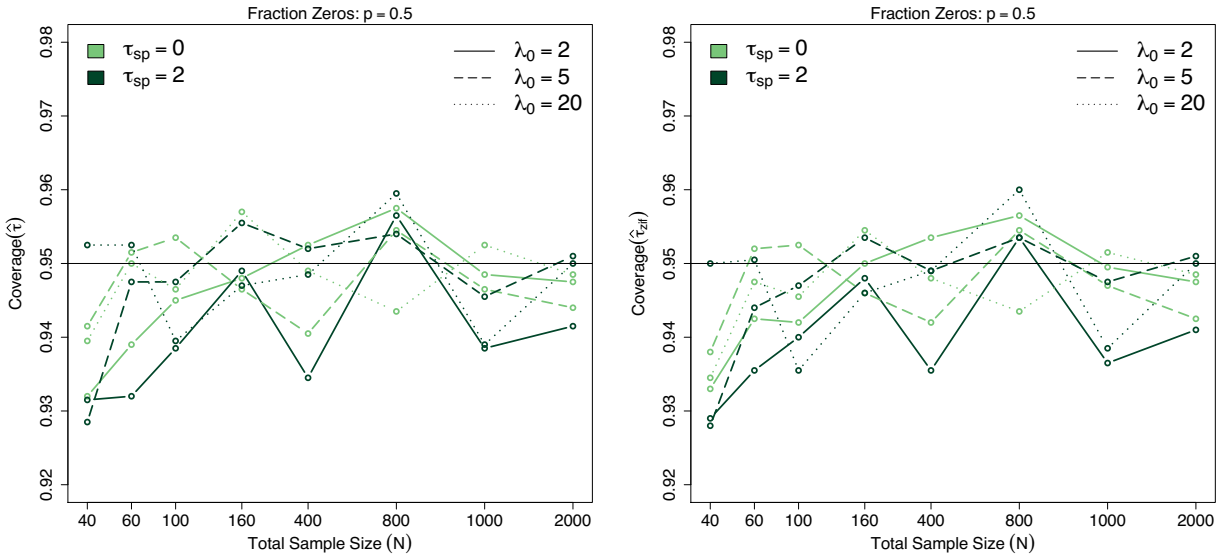


(a) Absolute bias, Neymanian estimate $\hat{\tau}$

(b) Difference in bias, Neymanian estimate $\hat{\tau}$ and model-based estimate $\hat{\tau}_{zif}$

Figure 1.3: The absolute bias and difference in absolute bias for the Neymanian estimate $\hat{\tau}$ and model-based estimate $\hat{\tau}_{zif}$ for the simulation settings in Table 1.1 (right column)

The two methods also have similar coverage properties, Figure 1.4 shows the approximate coverage for the two methods. When samples are small, there is slight under-coverage, but this may be due to noise since we observe similar amounts of under-coverage for large samples. Similar results hols for the simulation with a smaller fraction of zeros $p = 0.25$.



(a) Coverage, Neymanian estimate $\hat{\tau}$

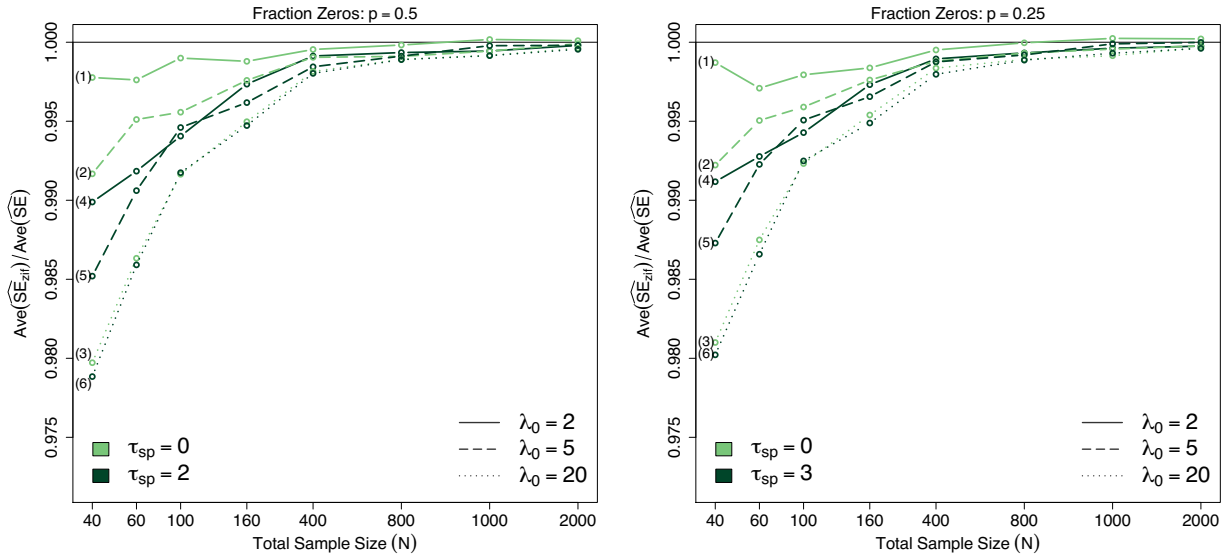
(b) Coverage, model-based estimate $\hat{\tau}_{zif}$

Figure 1.4: The confidence-interval coverage for the Neymanian estimate $\hat{\tau}$ and model-based estimate $\hat{\tau}_{zif}$ for the simulation settings in Table 1.1 (right column)

We expect the model-based standard error estimate, $\widehat{SE}_{zif}$, to be smaller than the Neymanian standard error estimate $\widehat{SE}$ on average. Figure 1.5 contains the precision gains for simulation settings in Table 1.1. Correctly assuming a zero-inflated model gains us the most precision when the sampling means are away from zero (lines (3) and (6)). Indeed, we gain precision regardless of the underlying distribution, but these gains are modest at best. With small samples, we increase precision by about 2%, and these gains are quickly lost. The relative precision gain is below 0.24% starting at $N = 400$. We can also gain precision in estimates of the sampling variability; we investigate this next.

Figure 1.6 shows that the estimates for the sampling variability are more precise if we use a model compared to the model-free Neymanian quantity. The gains in precision are most extreme in the low sampling means cases where the model-based estimates are up to 20% more precise than the Neymanian estimates. Interestingly, the cases that saw the largest gains in precision of τ estimation saw the smallest gains in precision for standard error estimates.

Overall, the model-based estimate for the ATE performed slightly better than the simple Neymanian estimate for small samples in terms of precision, while both showed little bias in simulations. The primary



(a) Simulations in Table 1.1 (left column)

(b) Simulations in Table 1.1 (right column)

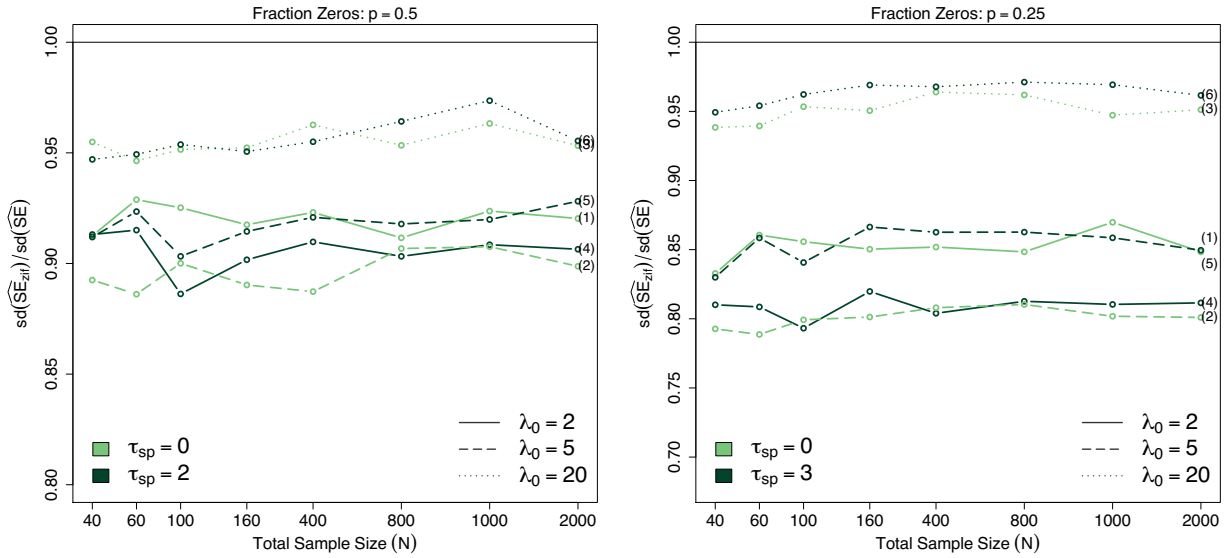
Figure 1.5: Relative precision of $\hat{\tau}$ compared to $\hat{\tau}_{zif}$, i.e. average $\widehat{SE}$ compared to average $\widehat{SE}_{zif}$

way that the model-based estimate outperforms the Neymanian estimate is through the precision of the standard error estimates. Essentially, the Neymanian standard error estimates seem to be sensitive the variability the data while the model-based estimate not. However, the precision of standard error estimates does not translate into tangible gains for inference. We will see next whether incorporating covariates into zero-inflated models changes this.

1.6.2 Simulation B

The purpose of this simulation is to inspect the impact of covariates on our inference procedures. In Section 1.5, we discussed that when covariates are predictive of the outcome, incorporating them into the analysis generally increases precision. The natural next step is to understand whether incorporating covariate information in a specialized way, e.g., with zero-inflated generalized linear models, improves our estimation procedures beyond naively using linear regression adjustment. We focus our efforts on understanding the impact of having a covariate that is predictive of zero-inflation.

To focus our study on the effects of covariates, we consider only the underlying distributions described in Setting (5) in Table 1.1, i.e we fix $\lambda_0 = 5, \lambda_1 = 9, p_0 = p_1 = 0.25$ so that $\tau_{sp} = 3$. For this simulation



(a) Simulations in Table 1.1 (left column)

(b) Simulations in Table 1.1 (right column)

Figure 1.6: Relative precision of model-based $\widehat{SE}_{zif}$ compared to Neymanian $\widehat{SE}$

our covariate will change the conditional probability of zero, $p(x)$, but will keep the overall probability of zero constant at p. For each subject i, covariates and zero-indicators under two potential treatments as

$$\begin{aligned} \text{Covariate:} \quad & X_i \sim \mathcal{N}(0,1) \\ & L_i = \sqrt{\rho}\, X_i + \sqrt{1-\rho}\, \varepsilon_i, \quad \rho \in [0,1], \quad \varepsilon_i \sim \mathcal{N}(0,1) \\ \text{Zero-Indicators:} \quad & Z_i(0) = \mathbb{1}\{L_i < \gamma_0\} \quad \text{and} \quad Z_i(1) = \mathbb{1}\{L_i < \gamma_1\}. \end{aligned}$$

Here, γ_0 and γ_1 control the overall probabilities of being a structural zero the control and treatment groups, respectively. If we set $\gamma_0 < \gamma_1$ a higher fraction of zeros in the treatment group than in the control group on average. The distribution of L_i is standard normal for any value of ρ so that γ_0 and γ_1 control the structural-zero probabilities in a straightforward way. If, say, $\gamma_0 = 0$, then automatically $\mathbb{P}(Z_i(0) = 1) = 0.5$. We select γ_0 and γ_1 so that the overall proportion of zeros is p_0 and p_1, respectively. For example, in the control group $\gamma_0 = F^{-1}_{\mathcal{N}(0,1)}(p_0)$ so that $\mathbb{P}(Z_i(0) = 1) = p_0$.

Here, ρ controls the dependence between the observed covariate X_i and the zero indicators $(Z_i(1), Z_i(0))$. When $\rho = 0$, $Z_i(1)$ and $Z_i(0)$ are independent of X_i and the covariate is not predictive of zeros. On the

	Simulation B					
Simulation	ρ	p_0	p_1	λ_0	λ_1	τ_{sp}
Setting (1)	0	0.25	0.25	5	9	3
Setting (2)	0.2	0.25	0.25	5	9	3
Setting (3)	0.4	0.25	0.25	5	9	3
Setting (4)	0.6	0.25	0.25	5	9	3
Setting (5)	0.8	0.25	0.25	5	9	3

Table 1.2: Settings for Simulation B. Example data sets from the settings highlighted in blue and red are shown in Figures 1.7 top and bottom, respectively

other extreme, when $\rho = 1$, the covariate is perfectly predictive of zeros. Indeed, in this case, we will have complete separation and will not be able to estimate the parameters in our model. By focusing our efforts on understanding the influence of ρ, we will get an understanding of how our estimating procedures might perform under varying degrees of covariate usefulness. We complete the specification of the zero-inflated Poisson by generating the potential outcomes

$$Y_i^*(0) \sim \text{Pois}(\lambda_0) \quad \text{and} \quad Y_i^*(1) \sim \text{Pois}(\lambda_1),$$
$$Y_i(0) = (1 - Z_i(0))\, Y_i^*(0) \quad \text{and} \quad Y_i(1) = (1 - Z_i(1))\, Y_i^*(1).$$

Figure 1.7 shows a simulated dataset where $\rho = 0$ and $\rho = 0.6$, the left panels show the predictive power of the covariate. When $\rho = 0$ (top-left), the proportion of zeros is equal across all x, and hence the covariate will have no predictive power. On the other hand, when $\rho = 0.6$ (bottom-left) the proportion of zeros varies with X_i. We observe more zeros for smaller values of X_i for both treatment and control, and hence X_i will be predictive of zero status. The right panels show the relationship between the covariate and outcome directly. The plotted linear regressions suggest that even a simple regression adjustment might help more precisely estimate the treatment effects when $\rho = 0.6$, but may not be helpful when $\rho = 0$.

In what follows, we will consider comparisons between three estimation methods. For benchmark purposes, we consider the typical Neymanian ATE estimator $\hat{\tau}$. We will also consider two separate methods that use covariate information in estimation. First, we consider the linear regression ATE estimator with

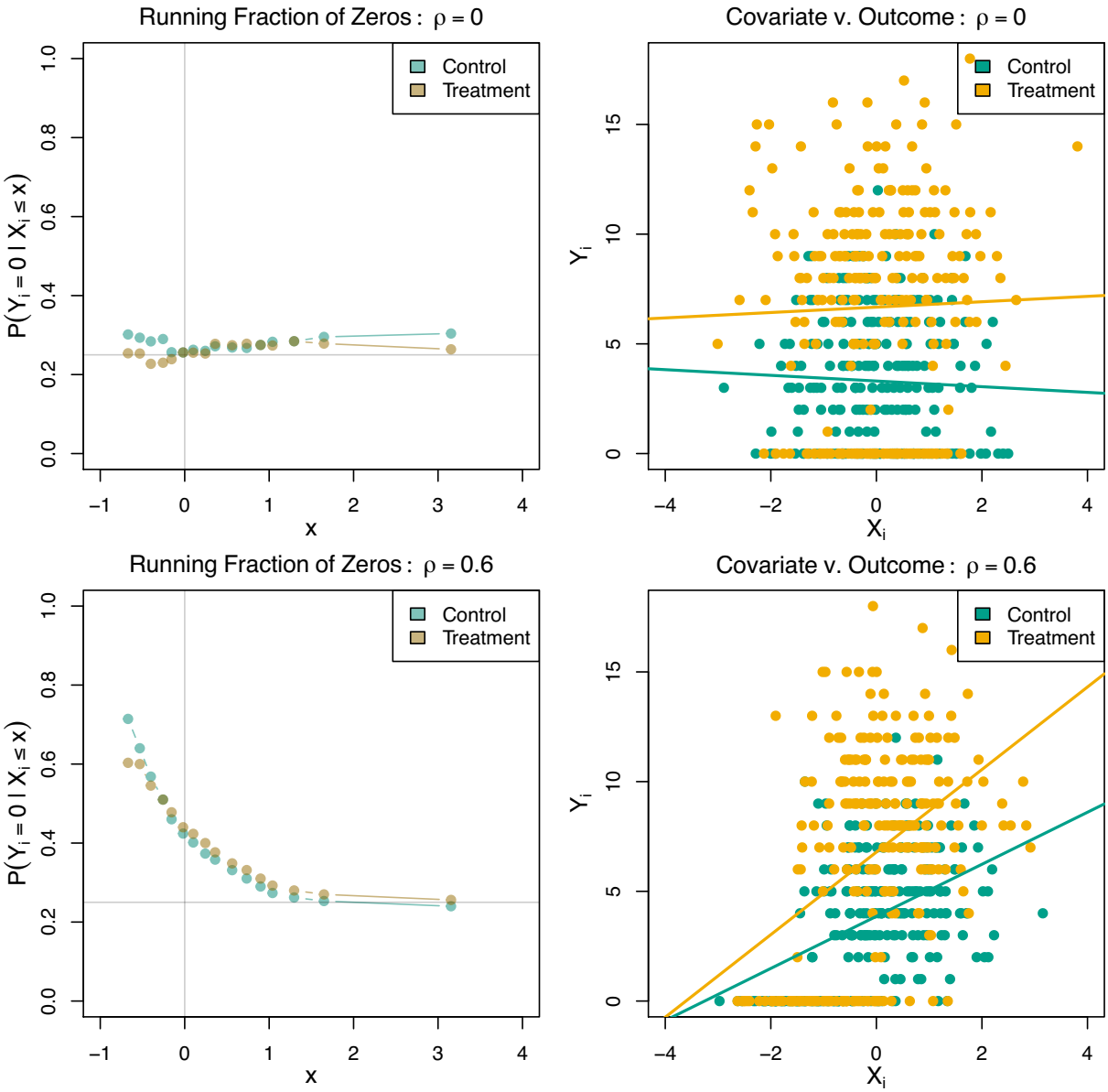
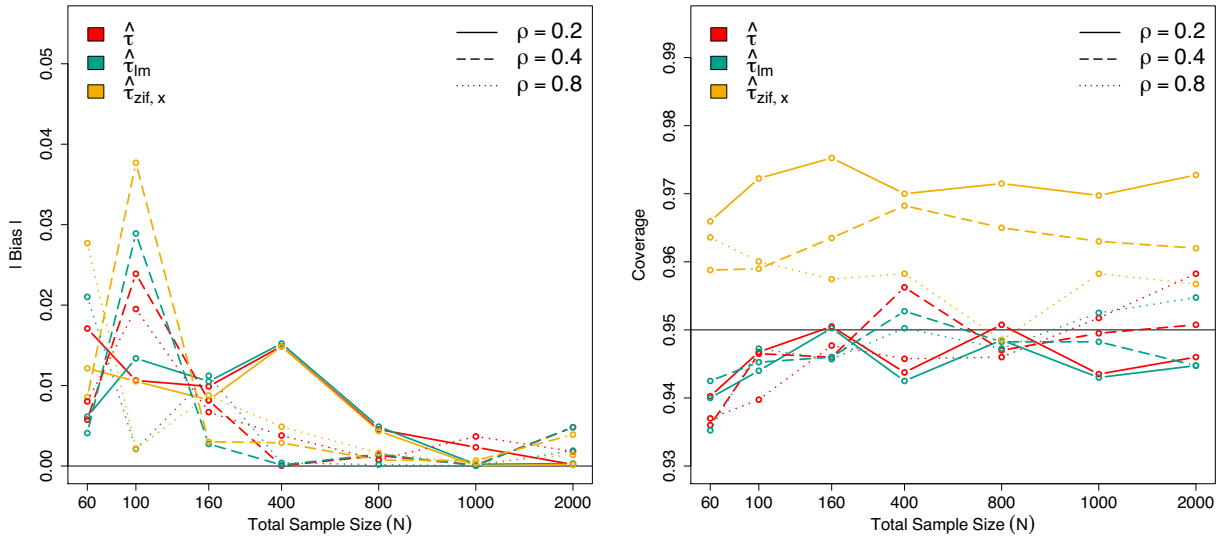


Figure 1.7: Data generated from zero-inflated Poisson distributions when a covariate is either not predictive of zero-status (top), or highly predictive of zero status (bottom). (left) Fraction of zeros for $X_i < x$. (right) Relationship between covariate and outcome with a simple linear regression

treatment indicator interacted with the covariate; we denote that as $\hat{\tau}_{lm}$. We also consider the zero-inflated generalized linear model discussed in Section 1.5.1, $\hat{\tau}_{zif,x}$.

Our first results are that both linear regression adjustment and model-based covariate adjustment are approximately unbiased regardless of the strength of the relationship between the covariate and outcome. Figure 1.8(a) shows the approximate bias for increasing levels of relation between covariate and outcome $\rho = 0.2, 0.4, 0.8$. As expected, the bias decreases with increasing sample size, but even the largest biases are relatively small with relative bias at around 1%. The coverage, on the other hand, starts to paint a more interesting picture.

Figure 1.8(b) shows the coverage properties of the different estimators for increasing sample sizes. Both the standard Neymanian estimate and the linearly adjusted methods tend to have the right coverage. The covariate-adjusted $\hat{\tau}_{zif,x}$ tends to over-cover when the covariate information is not strongly predictive of the outcome, i.e., when ρ is small. When ρ is large $\rho = 0.8$, $\hat{\tau}_{zif,x}$ has similar coverage to the other methods for larger samples. Because the estimators are approximately unbiased, the standard error estimates $\widehat{SE}_{zif,x}$ are the likely culprit when ρ is small.



(a) Absolute bias (b) Coverage

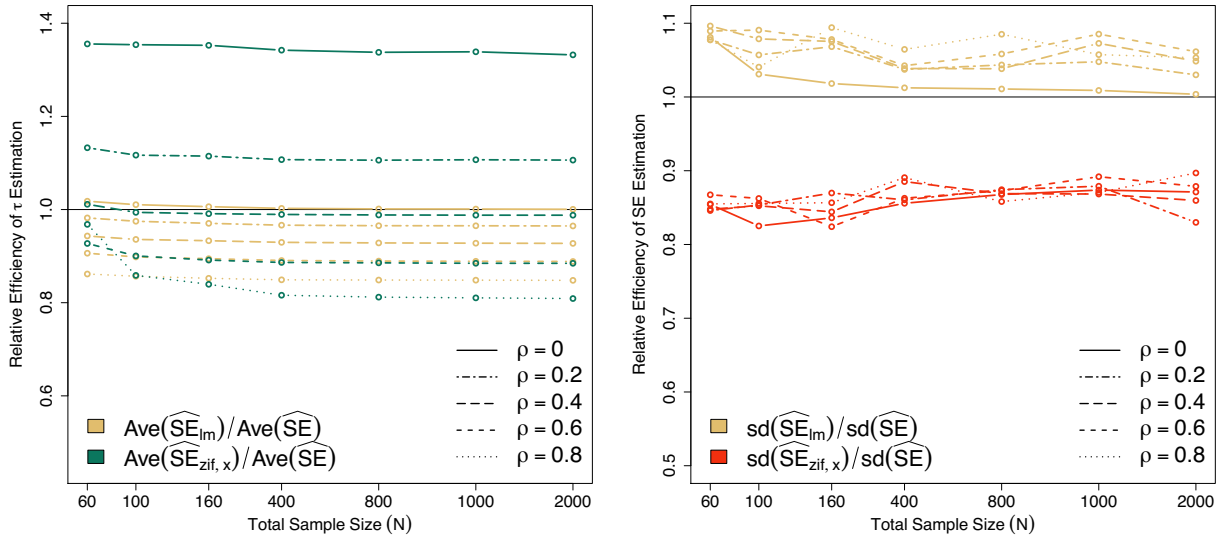
Figure 1.8: The absolute bias and coverage for the Neymanian estimate $\hat{\tau}$ and two covariate-adjusted estimates: $\hat{\tau}_{lm}$ and $\hat{\tau}_{zif,x}$.

Figure 1.9(a) shows the relative efficiency of covariate-adjusted models to the Neymanian estimator for increasing values of ρ. As expected, the linear regression adjustment improves over the typical Neymanian estimator when the covariate information is useful. This is shown in gold in Figure 1.9(a). For example, the linear regression adjustment shows a 10% reduction in standard error when $\rho = 0.6$. The zero-inflated regression, on the other hand, has worse precision when the covariate is not useful in predicting the outcome.

As Lin (2013) discuss, even when covariates do not predict outcomes, as is the case when $\rho = 0$, linear regression estimates are not worse than the Neymanian estimator. However, complex covariate adjustments like $\hat{\tau}_{zif,x}$ do not have this property and are only beneficial when covariates are predictive of the outcome. When $\rho = 0.2$, the standard errors $\widehat{SE}_{zif,x}$ tend to be 11% larger than the Neymanian estimator $\widehat{SE}$. For values of ρ larger than 0.4, the standard errors finally beat out the Neymanian estimator. These results indicate that we should only adjust for covariates in the zero-inflated generalized linear model if the covariates are strongly predictive of zero status. However, it is difficult to say what strongly predictive means for the zero-inflated model. It is not clear in what situations can we expect the equivalent to $\rho = 0.3$, or 0.5, muddying the waters in making this decision.

Simulation A showed that assuming a model increased the precision of standard error estimates by up to 20%. Figure 1.9(b) shows the relative efficiency of the standard error estimates for the different methods. In gold, we see that linear regression adjustment has slightly, but consistently larger relative efficiency (up to 10%) than the Neymanian $\widehat{SE}$. When we use models to incorporate covariate information, the precision of standard error estimates increases by $15 - 20\%$ over, well in line with the previous results.

Many of the results here are in line with the current literature. If a covariate is predictive of the outcome, including it in modeling can help improve precision in estimation. Where the current results diverge most strikingly is in the opposite situation. If the covariate is not predictive, including it in the model can hurt precision. Indeed, we see in our results that precision is lost when linearly adjusting for a non-predictive covariate. However, the losses are minuscule. On the other hand, adjusting for the same covariate with a zero-inflated GLM leads to extreme losses in efficiency. However, it is encouraging to see that when the covariate is significantly predictive of the outcome $\rho = 0.8$, the zero-inflated GLM is indeed the most precise.



(a) Relative efficiency of τ_{sp} estimation

(b) Relative efficiency of SE estimation

Figure 1.9: The relative efficiency of estimates of τ_{sp} and standard error estimates for the simulation settings in Table 1.2

1.7 Data Application

Experiments on Massive Open Online Course (MOOC) platforms perfectly exemplify the scenario explored in this chapter. Typically, researchers are interested in comparing outcomes for engagement like course completion, forum participation, and scores on exams or quizzes. Zero-inflation in these outcome measures is certainly present due to attrition. However, even engaged students may not participate in measurable ways. Many course platforms also collect a set of pre-treatment covariates, like age and previous enrollment in online courses, that may be predictive of participation. In what follows, we explore one such experiment. We analyze the experiment, excluding details because the experimental results remain unpublished.

The purpose of the intervention we analyze here was to increase engagement in MOOCs through social support. There is a growing literature that supports the idea that social support impacts increased engagement and participation in MOOCs. One recent study found that even perceived social support was predictive of engagement in online course (Hsu et al., 2018) and another study found that attrition rates among students who sought help from friends and family than those who did not (Nelimarkka & Hellas, 2018).

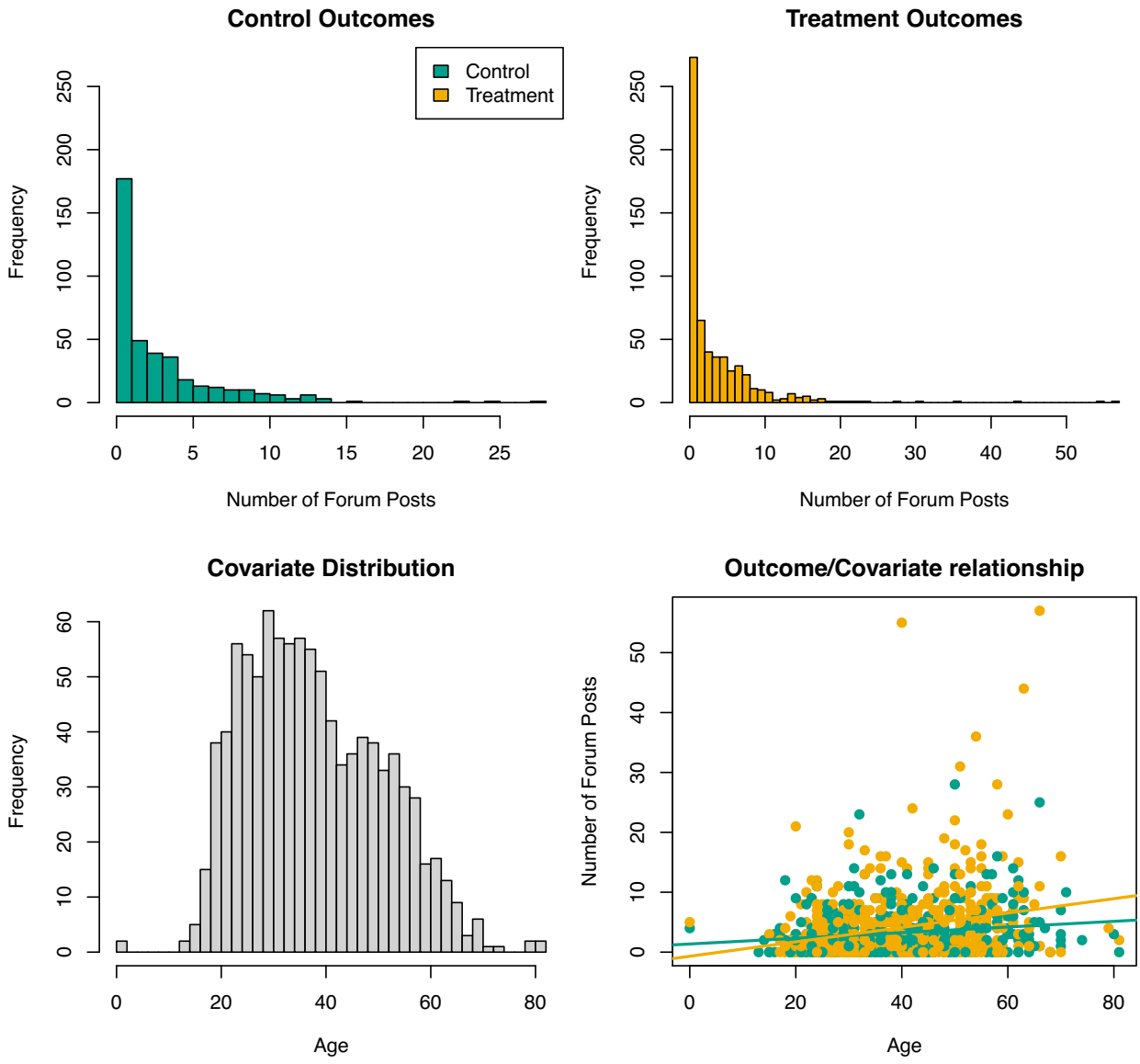


Figure 1.10: Example Data from an experiment on a MOOC. (top left) Histogram of observed control outcomes, (top right) histogram of observed treatment outcomes, (bottom left) distribution of covariate (age), (bottom right) Relationship between covariate and outcome for treatment and control with a simple linear regression

The intervention was conducted on a popular MOOC platform and involved 986 concurrently-enrolled students. The intervention was time-intensive and involved contacting members of the student's social support network to fortify the network. We assigned students to a treatment or control condition at random, 593 were assigned to treatment while 393 were assigned to control. We measure engagement as the number of forum posts the student made during the length of the course.

Figure 1.10 shows the data distribution for the experiment. The number of forum posts (upper panels) shows clear signs of zero-inflation and some potential outliers. We will be using the student's age as a pre-treatment covariate. The bottom panels show the age distribution and the relationship between age and number of forum posts. The linear regressions indicate that the covariate may help predict the outcome, so making covariate adjustment will likely help. It is unclear, however, whether the obvious heteroskedasticity will affect the linear regression adjustment.

The Neymanian estimate of the ATE for our intervention is 0.72, a significant finding ($95\% CI : (0.07, 1.37)$). The model-based estimate $\hat{\tau}_{zif}$ showed the same point-estimate but narrower confidence intervals ($95\% CI : (0.38, 1.05)$). From the simulations, we know that Neymanian standard error estimates tend to vary more than the model-based ones. So we would trust the model-based confidence interval here.

Covariate-adjustment gave mixed results. The linear regression adjusted ATE estimate gave much wider confidence intervals. This loss of precision may be due to sensitivity to heteroskedasticity in the data, but we did not investigate this in our simulations. The covariate-adjusted zero-inflated estimator, on the other hand, still provided a reasonable estimate of the ATE. However, including the covariate inflates the standard error estimates by 14%. The covariate is marginally predictive of the zero-status for the treatment group and not at all for the control group, so the inflated variance falls in line with our simulations (see Table 1.4).

Overall, we recommend the model-based $\hat{\tau}_{zif}$ for estimating the ATE as the data are highly heteroskedastic. Our simulations showed that this estimator typically performs at least as well as the Neymanian estimator in terms of bias and average precision. We do not recommend covariate adjustment with this covariate. However, it may help further increase precision if we adjust for the covariate in for the sampling mean.

Estimator	Estimate	95% Confidence Interval	Interval Length
Neymanian $\hat{\tau}$	0.72	(0.07 ,1.37)	1.3
Model-based $\hat{\tau}_{zif}$	0.72	(0.38, 1.05)	0.67
Regression Adjusted $\hat{\tau}_{lm}$	-2.05	(-3.97, -0.12)	3.85
Adjusted Model-Based $\hat{\tau}_{zif,x}$	0.52	(-0.47, 1.50)	1.97

Table 1.3: ATE Estimates for Example Intervention in MOOC

Sampling Distribution

Parameter	Estimate	SE	p-value
$\log(\lambda_0)$	1.4204	0.0293	0.0000*
$\log(\lambda_1)$	1.5913	0.0212	0.0000*

Zero-part

Parameter	Estimate	SE	p-value
α_0	-0.7190	0.3592	0.0453*
α_1	-0.7161	0.3290	0.0295*
β_0	-0.0123	0.0091	0.1783
β_1	-0.0166	0.0086	0.0525

Table 1.4: Model Output for our model-based covariate-adjusted estimator

1.8 Concluding Remarks

Even in the ideal case of a balanced, completely randomized experiment, there are potentially special issues to consider when estimating a causal treatment effect from zero-inflated count responses. First, as in any proper causal analysis, we need to define our estimand of interest. Domain knowledge and the scientific questions should motivate this choice, rather the modeling procedure. For this chapter, we focused on the super-population average treatment effect and considered several ways of estimating it.

We considered two ways of looking at the data: taking it as it is, without making any model assumptions, or as a two-part mixture distribution. The former perspective readily justifies a simple method-of-moments difference-in-means estimator, $\hat{\tau}$. In contrast, viewing the data as coming from a zero-inflated model, it makes sense to incorporate this model assumption into our estimator. Exploiting the parametric model assumption, our estimator $\hat{\tau}_{zif}$ is then the difference in maximum likelihood estimates for the means of two zero-inflated models. The estimates $\hat{\tau}$ and $\hat{\tau}_{zif}$ are unbiased and consistent, respectively, for the super-population average ATE τ_{sp}. We were surprised, however, to find that while exploiting a

model assumption did increase precision in ATE estimation for small samples, these gains were quickly lost. However, we did find that modeling helped increase the precision of standard error estimates.

Continuing our view of the data as a mixture motivates the use of covariate adjustment via zero-inflated GLMs. Covariate adjustment is useful when the covariate is predictive of the outcome. Lin (2013) show that methods like regression adjustment are a particularly stable way of incorporating covariate information to increase precision. Furthermore, when covariates are not predictive of outcome, it does not make estimation much worse than standard Neymanian estimates. Zero-inflated GLMs, on the other hand, are quite sensitive to non-predictive covariates. Furthermore, the estimators derived from zero-inflated GLMs can be quite complex, even for simple causal estimands like the ATE. This complexity makes it challenging to implement estimators and, consequently, decreases their use in practice.

Our findings shows that a strongly predictive covariate still leads to precision gains of ATE estimates, but there remains much to be explored. We speculated that if the covariate was predictive of the non-zero mean, ATE estimates based on zero-inflated GLMs would have properties similar to those from regression adjustments. However, it may be the case that ATE estimates based on zero-inflated GLMs are simply more unstable when including covariates into the model (especially useless covariates). In which case, linear adjustment may beat out zero-inflated GLM adjustment.

Furthermore, a quality that makes zero-inflated GLMs attractive is that they allow covariates to affect outcomes in two ways. It may be the case that this ability can mitigate some of the precision losses suffered here if the covariate is predictive in one part of the model, even if it is not in the other. This requires further study. Nevertheless, ATE estimates based on zero-inflated GLMs increased the precision of standard error estimates across the board. This fact encourages us to further study the limits of covariate adjustment through complex models.

To photograph truthfully and effectively is to see beneath the surfaces and record the qualities of nature and humanity which live or are latent in all things.

Ansel Adams

2

Measuring the Effects of Adherence on Time-Varying Health Outcomes

2.1 Introduction

Over 85 million American adults, or about one third of the population over 20 years old, suffer from hypertension (Benjamin et al., 2018). Approximately 16% of adults in the United States are unaware that they have hypertension (Benjamin et al., 2018). Left untreated, hypertension can lead to a range of serious and costly health concerns such as cardiovascular disease, stroke, and renal disease (Amery et al., 1985; Probstfield, 1991). Among the many factors associated with uncontrolled BP, poor adherence to

prescribed anti-hypertensive medications is of major concern to clinicians, health care systems, and other stakeholders (Choo et al., 1999; Morisky et al., 1986; Osterberg & Blaschke, 2005). Little doubt exists that patients who adhere poorly to their prescribed medication are at risk for worse BP outcomes (Vitolins et al., 2000; Lam & Fresco, 2015). Given the large variation in adherence, both across patients and temporally within a patient's treatment period, it is an open question how to accurately measure the impact of varying adherence patterns on BP levels. Furthermore, the variation in adherence patterns creates difficulties in accurately measuring the effects of socio-demographic and health risk factors on BP levels.

This paper develops a Bayesian dynamic linear model (West & Harrison, 1997; Durbin & Koopman, 2001; Petris et al., 2009) for health measures recorded over time as a function of time-varying adherence, with a particular application to the effects of anti-hypertensive medication on BP levels. Bayesian dynamic linear models (DLMs) have a long history as a statistical framework for forecasting and measuring trajectories in many domains, including real-time missile tracking and financial securities forecasting, but are rarely used in healthcare applications. We apply DLMs to describe time-varying health measures (like blood pressure levels) as a function of detailed adherence (or other time-varying covariates) and individual demographics and comorbidities measured typically at study baseline. The application of DLMs to the time-varying adherence on health measures is novel, but fits naturally into the DLM framework because these measures can be tracked over time as adherence data accumulate in parallel. Because the DLM framework permits the inclusion of patient-level predictors, the model can be applied to measuring effects of socio-economic or racial characteristics, or the effects of different comorbid conditions. Our model can control for differential medication adherence patterns, resulting in more accurate measurement of the effects of other covariates.

In order to estimate the effects of time-varying adherence on health measures, we need access to three types of data. First, adherence data for each study participant are assumed to be collected through electronic monitoring devices. These devices electronically time-stamp each time the pill container is opened, providing an accurate recording of when a patient took their medication. Second, health and socio-demographic information are typically recorded at the start of medication adherence studies, and are often non-dynamic. Finally, health measures which might be impacted by differential medication adherence are recorded longitudinally at intermittent intervals. Such measures are often recorded at clinic visits, the timing of

which may be determined by the patient. Thus it is quite common for the number of health measures per study patient to be much smaller than the number of days on which medication adherence information is recorded.

Inference using the DLM framework is challenging in this case because there are considerably fewer health measures observed compared to the number of days of adherence measurements. In traditional uses of DLMs, both time-varying covariates and responses are measured at every time point, and inference for the time-varying state parameters can be accomplished using standard Bayesian updating algorithms (West & Harrison, 1997, Chapter 16). With adherence data, when the responses are measured at irregular and infrequent intervals, the usual updating approaches can be demonstrably inefficient. Instead, we develop an inferential approach in a Bayesian setting that takes advantage of factoring the posterior density into a product of two terms. The first term involves the exact marginal likelihood of the DLM, marginalizing out the dynamic state process. The second is the conditional posterior density for the state process parameters. The two-stage procedure allows us to determine the posterior distribution for the non-dynamic parameters using standard Markov chain Monte Carlo (MCMC) techniques and the state process parameters without resorting to needlessly complex computational tools. With a DLM that has normally distributed responses and a stochastic process for the latent states that has normal innovations, the factorization is the product of two multivariate normal densities and a prior. This factored posterior density also easily allows inference for the non-dynamic parameters, which in the setting of medication adherence is likely of interest because the researcher may want to understand the effect of baseline health characteristics on the health measures controlling for differential adherence.

The remainder of this paper is organized as follows. In Section 2.2 we introduce a motivating example and details of the study cohort. We specify the DLM for our framework in Section 2.3. The model, which assumes an autoregressive structure, accounts for possibly multivariate health measures which may or may not be measured simultaneously. We then develop our computational approach for inference in Section 2.4. We apply our methods to modeling BP in Section 2.5, where we compare our methods to typical models used to measure the effects of adherence. We conclude in Section 2.6 with a discussion on the limitations and potential extensions of our methods.

2.2 Data

The data we analyzed were obtained from the baseline pre-randomization period of a trial that studied the effects of a provider-patient communication skill-building intervention on adherence to anti-hypertensive medication and on BP control (clinicaltrials.gov ID: NCT00201149). Patients were recruited from seven outpatient primary care clinics at Boston Medical Center, an inner-city safety-net hospital affiliated with the Boston University School of Medicine. Patients enrolled in the study from August 2004 and June 2006, meeting several eligibility criteria. These included that the patients were of white or black race (African or Caribbean descent), were at least 21 years old, had an outpatient diagnosis of hypertension on at least three different occasions prior to study enrollment, and were currently on anti-hypertensive medication. The cohort size was 869 patients but because the trial was focused on improving adherence, not all patients had BP measured during this period. So, our final study cohort size was 503 patients. The study involved measuring anti-hypertensive medication-taking using Medication Event Monitoring System (MEMS) caps, a particular type of electronic pill-top monitoring device. Patients were given their most frequently taken anti-hypertensive medications in a bottle with a MEMS cap, and were instructed to open the bottle each time they took a dose. Each MEMS cap contained a microprocessor that recorded the date and time whenever the bottle was opened, and the timing information was then downloaded through a wireless receiver after the patient returned the MEMS cap. Our study focused on medication-taking behavior during the entire pre-randomization baseline period of the study. A patient was considered adherent to the prescribed medication on a day if the MEMS cap was opened as many times as the prescribed dosing frequency, and not adherent otherwise. These measurements were recorded on a daily basis, but the duration over which adherence information was measured varied by patient.

Blood pressure measurements were recorded less frequently, as they were obtained as part of routine clinical care. The BP readings were measured using manual or electronic devices, and were obtained by clinical staff including physicians, nurses and medical assistants. In cases where multiple readings were obtained on a single day (typically at the same clinic visit), the individual values were recorded. Diastolic and systolic blood pressure values (DBP and SBP respectively) were recorded separately.

In addition to detailed time-varying adherence and BP readings, other patient-specific baseline infor-

mation was collected. Race (white versus African American), gender, and age at the start of the study were recorded for each patient. From electronic medical records, the following comorbidities (as binary variables) were investigated in our study, given their potential impact on overall BP levels: presence of cerebrovascular disease, congestive heart failure, chronic kidney disease, coronary artery disease, diabetes mellitus, hyperlipidemia, peripheral vascular disease, and obesity (defined as body mass index greater than $30kg/m^2$).

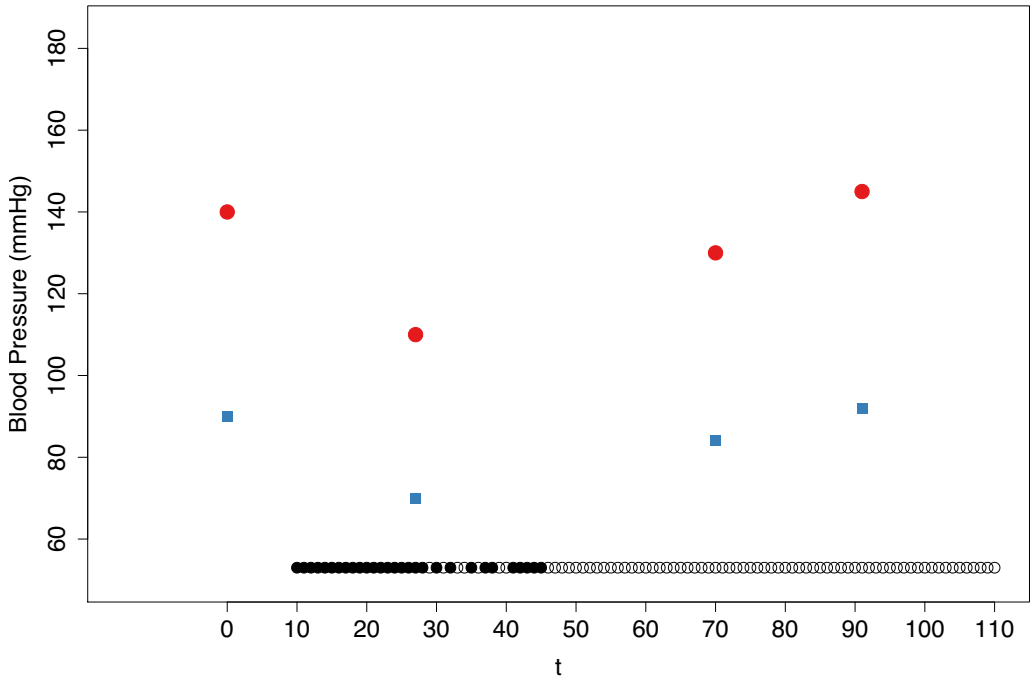


Figure 2.1: Example BP and adherence data for a study patient. Blue squares are diastolic BP, red dots are systolic BP, and the black dots near the horizontal axis indicate whether the patient was adherent on a particular day (filled dots indicate adherence, open dots indicate non-adherence). Note that the adherence data for the first nine days of the follow-up period are missing.

To motivate our modeling framework, consider the data from one of the patients in the study cohort displayed in Figure 2.1. The figure shows the DBP and SBP measures on the four days the patient had their blood pressure measured, and daily indications of whether the patient was adherent. This patient was adherent on 29 of 82 days. However, this simple summary masks the time-varying pattern of the patient's adherence. The patient began the study by being fully adherent to their prescribed medication, during which time their blood pressure remained under control. After 20 days, the patient started becoming less adher-

ent, and starting around day 38 the patient discontinued taking their medication altogether. Over this latter period, the patient's BP increased, and by the end of the study period the patient had DBP and SBP values that were not under control. Besides suggesting a potential link between adherence and BP measures, this example motivates using a dynamic linear model for health measures that accounts for time-varying adherence.

2.3 A dynamic model for multivariate health measures

We propose a statistical framework for multivariate time-varying health measures as a function of medication adherence that is a member of the class of Bayesian dynamic linear models. Let y_{itk} be the value of the k^{th} health measure, with $k = 1, \ldots, K$, for patient i at time t, for $t = 1, \ldots, T_i$. We assume that time is discretized and equally spaced (e.g., days), and that health measures are observed only at times $t_1, t_2, \ldots, t_{m_i}$. Let $\boldsymbol{y}_{it\cdot}$, generally, indicate a vector collecting all outcomes across the dotted index into a column vector. Then the vector $\boldsymbol{y}_{it\cdot}$ collects the outcomes for patient i at time t, i.e., $\boldsymbol{y}_{it\cdot} = (\boldsymbol{y}_{it1}, ..., \boldsymbol{y}_{itK})^T$. Similarly, $\boldsymbol{y}_{i\cdot k}$ collects the k^{th} outcome observed at times $(t_1, ..., t_{m_i})$ for patient i, i.e., $\boldsymbol{y}_{i\cdot k} = (\boldsymbol{y}_{it_1k}, ..., \boldsymbol{y}_{it_mk})^T$. Our framework assumes that $\boldsymbol{y}_{it\cdot}$ is an observed measurement generated from a distribution with mean $\boldsymbol{\mu}_{it\cdot}$ which follows a stochastic process. The framework recognizes that observed health measurements on a given day could vary around a mean level due to various influences including emotional state, activity level, recent alcohol consumption, ambient temperature, and other unobserved factors that might affect the outcomes.

The mean process at time t is modeled as the sum of the contributions of a non-dynamic term and a dynamic term. Specifically, we assume

$$\boldsymbol{y}_{it\cdot} = \boldsymbol{\mu}_{it\cdot} + \boldsymbol{\varepsilon}_{it} = \boldsymbol{\beta}\boldsymbol{x}_i + \boldsymbol{\alpha}_{it\cdot} + \boldsymbol{\varepsilon}_{it} \tag{2.1}$$

where $\boldsymbol{x}_i$ is a vector (including intercepts) of p non-dynamic covariates typically measured at baseline and $\boldsymbol{\beta}$ are the corresponding linear coefficients. Furthermore, $\boldsymbol{\alpha}_{it\cdot}$ is a stochastic process that may depend on dynamic covariates, such as time-varying adherence. Here $\boldsymbol{y}_{it\cdot}, \boldsymbol{\mu}_{it\cdot}, \boldsymbol{\alpha}_{it\cdot}$ and $\boldsymbol{\varepsilon}_{it}$ are all vectors of length K corresponding to the observation, mean outcome process, latent state model and sampling error respec-

tively. The non-dynamic covariate effects $\boldsymbol{\beta}$ are of dimension $K \times p$, and we use $\boldsymbol{\beta}_k$ to denote the k^{th} row of $\boldsymbol{\beta}$, i.e., the non-dynamic covariate effects on the k^{th} outcome. The error term for subject i at time t, $\boldsymbol{\varepsilon}_{it}$, accounts for typical sampling variability, measurement error, and possible correlation of the outcomes within patient and time. We further assume that $\boldsymbol{\varepsilon}_{it} \sim \mathrm{N}(0, \Sigma_\varepsilon)$ with possibly non-diagonal covariance matrix Σ_ε.

If patient i's adherence is tracked for T_i consecutive days, we indicate this measurement set as $\{1, ..., T_i\}$, and denote the collection of time-varying adherence measures as $\{c_{it}\}_{t=1}^{T_i}$. We let $c_{it} = 1$ if patient i was adherent on day t and -1 otherwise. As we describe below, we extend the definition of c_{it} to allow the inclusion of other time-varying covariates. We further let $\mathcal{T}_i = \{t_1, ..., t_{m_i}\} \subset \{1, 2, ..., T_i\}$ denote the set of m_i days on which the outcomes were measured for patient i. The set of outcome measurements for patient i is denoted by $\boldsymbol{y}_{i\cdot\cdot} = \{\boldsymbol{y}_{it\cdot}\}_{t\in\mathcal{T}_i}$. We typically have that $m_i << T_i$ for all i.

The latent process $\boldsymbol{\alpha}_{it\cdot}$ is assumed to be influenced by whether a patient takes the prescribed medication along with any other time-varying covariates, so we assume an $AR(1)$ model on the latent states $\{\boldsymbol{\alpha}_{it\cdot}\}$ that depend on $\boldsymbol{c}_{it\cdot}$ as time-varying covariates. Other forms of stochastic processes are possible, including autoregressive processes of higher order. If $\boldsymbol{c}_{it\cdot}$ is of dimension r, then we assume a process on $\boldsymbol{\alpha}_{it\cdot}$ given by

$$\boldsymbol{\alpha}_{it\cdot} = \boldsymbol{\rho}\boldsymbol{\alpha}_{i,t-1,\cdot} + \boldsymbol{\phi}\boldsymbol{c}_{it\cdot} + \boldsymbol{\nu}_{it} \tag{2.2}$$

where $\boldsymbol{\rho}$ is the K-dimensional diagonal matrix of first-order autoregressive parameters, that is, $\boldsymbol{\rho} = \mathrm{diag}(\rho_1, \rho_2, ..., \rho_K)$. We assume $|\rho_k| < 1$ to ensure stationarity. The time-varying outcome effects $\boldsymbol{\phi}$ is a $K \times r$ matrix. We let $\boldsymbol{\phi}_k$ denote the k^{th} row of $\boldsymbol{\phi}$, i.e., the time-varying covariate effects on the k^{th} outcome. We further assume in (2.2) that the innovations $\boldsymbol{\nu}_{it}$ have zero-mean MVN distributions with diagonal covariance matrix $\Sigma_\nu = \mathrm{diag}(\sigma^2_{\nu 1}, \sigma^2_{\nu 1}, ..., \sigma^2_{\nu K})$. Additionally we assume the initial state $\boldsymbol{\alpha}_{i1} \sim N_K(0, \Sigma_0)$ with $\Sigma_0 = \mathrm{diag}(\sigma^2_{01}, \sigma^2_{01}, ..., \sigma^2_{0K})$. Let θ denote the non-dynamic parameters $\theta = (\boldsymbol{\beta}, \Sigma_\varepsilon, \boldsymbol{\rho}, \boldsymbol{\phi}, \Sigma_\nu, \Sigma_0)$. Equations (2.1) and (2.2) define the distributions $p(\boldsymbol{y}_{it\cdot}|\boldsymbol{\alpha}_{it\cdot}, \boldsymbol{x}_i, \theta)$ and $p(\boldsymbol{\alpha}_{it\cdot}|\boldsymbol{\alpha}_{i,t-1,\cdot}, \boldsymbol{c}_{it\cdot}, \theta)$ respectively.

Figure 2.2 contains a simulated example of a patient with a scalar outcome over a 30-day period corresponding to the model in (2.1) and (2.2). As evidenced in Figure 2.2, the mean process generally decreases

on days when a patient is adherent. However, this is not always the case, and an increase can occur when the corresponding innovation ν_t is large and positive, offsetting the impact of the patient taking their medication. The observed health measure is normally distributed around the mean for the day on which the measure is recorded. On day 1, for example, the health measure is higher than the mean, and on day 15 the health measure is lower than the mean.

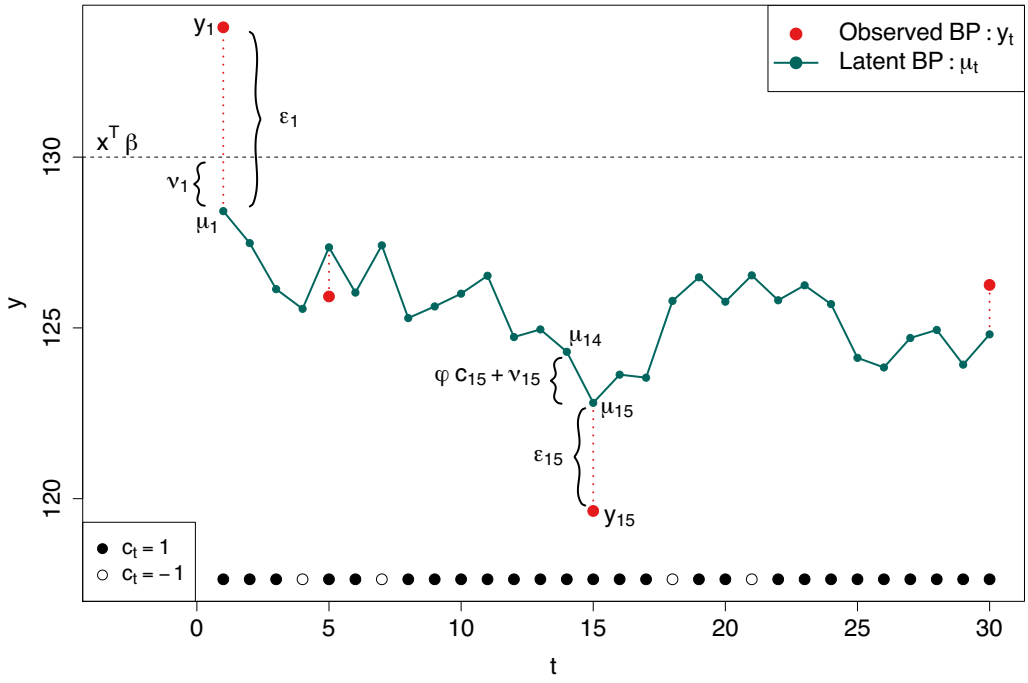


Figure 2.2: A 30-day trajectory for a simulated patient from a DLM with a scalar outcome. The bottom of the figure displays the adherence indicator simulated independently with a 90% probability of adherence. The adherence effect is simulated to be large and negative ($\phi = -0.5$). The contribution of the non-dynamic covariates is assumed to be $x_i^T\beta = 130$.

This model is attractive for relating health measures to time-varying adherence for several reasons. It can account for non-dynamic baseline variables as well as time-varying adherence. This allows us to disentangle the effects of detailed adherence from patient-specific socio-demographic and health covariates. Furthermore, in settings where multiple outcome measures are observed at time t, inference for the sampling variability Σ_ε can be made more precise, and can be separated from the innovation variability Σ_ν.

Additionally, the $AR(1)$ model component has a well-known asymptotic mean under full adherence and full non-adherence. Specifically, the overall effect of repeated days of adherence on the outcome measure

converges for increasing values of t as $\boldsymbol{\alpha}_{it\cdot} \to (I - \boldsymbol{\rho})^{-1}\boldsymbol{\phi}$. We can therefore predict that if patient i were to continue to be fully adherent, their mean outcomes would tend to $\boldsymbol{\mu}_{it\cdot} \to \boldsymbol{\beta}\boldsymbol{x}_i + (I - \boldsymbol{\rho})^{-1}\boldsymbol{\phi}$. An analogous calculation can be performed when the patient is fully non-adherent. This property permits estimating the best-case or worst-case health measure means for perfect adherence or perfect non-adherence, even when patients' adherence level is somewhere in between.

2.4 Marginal Dynamic Linear Models

Inference for the model in Section 2.3 is challenging given the large number of parameters, both the non-dynamic parameters θ as well as the time-varying parameters $\boldsymbol{\alpha}$. In particular, if the number of recorded adherence indicators per patient is large, then the size of $\boldsymbol{\alpha}$ is similarly large. Standard Bayesian inferential methods through posterior simulation approaches for such a highly parameterized model can result in slow convergence and unreliable inferences.

Advances in Bayesian computation have made highly parameterized dynamic models more tractable. With recent developments in sequential Monte-Carlo (SMC), including software packages like Libbi (Murray, 2015), sampling the high-dimensional latent states of the DLM has become computationally feasible and accessible. The majority of these advances are in situations where the likelihoods can not be computed directly and are instead approximated. Marginal sampling schemes, like the particle marginal Metropolis-Hastings (PMMH) algorithm (Andrieu et al., 2010), alternately sample the structural parameters and latent process parameters and accept or reject them with an adjusted Metropolis-Hastings step accounting for the approximation of the likelihood needed in sampling the latent space. Recent work (Bhattacharya & Wilson, 2018) approximates the posterior of the structural parameters on a grid of points, reducing the possible sampling values to a discrete set. SMC has even been adapted to situations where there is known sequential structure of otherwise intractable likelihoods (Chopin et al., 2012). The setting described in Section 2.3, however, contains much more known structure. As we will show in this section, not only can we compute the marginal likelihood of the data given the structural parameters exactly, but we can also compute the exact posterior distribution of the latent process parameters conditional on structural parameters.

Specifically, our approach takes advantage of factoring the joint posterior density of θ and $\boldsymbol{\alpha}$ as follows

$$p(\boldsymbol{\alpha}, \theta|\boldsymbol{y}) = p(\theta|\boldsymbol{y})p(\boldsymbol{\alpha}|\theta, \boldsymbol{y}). \tag{2.3}$$

where we omit the dependence on both $\{\boldsymbol{c}_{it.}\}$ and $\boldsymbol{x}_i$. The first factor in (2.3) is discussed below, while the second factor, the conditional posterior density of the latent process parameters, can be derived exactly and is discussed in Section 2.4.2.

Marginal inference about θ can be accomplished by integrating the left side of (2.3) with respect to $\boldsymbol{\alpha}$, yielding the first factor in the expression. This factor can be expanded using Bayes' Theorem as

$$p(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta)p(\theta)}{p(\boldsymbol{y})}. \tag{2.4}$$

The marginal likelihood, $p(\boldsymbol{y}|\theta)$, in the numerator is determined from

$$p(\boldsymbol{y}|\theta) = \int p(\boldsymbol{y}|\boldsymbol{\alpha}, \theta)p(\boldsymbol{\alpha}|\theta)d\boldsymbol{\alpha}. \tag{2.5}$$

As we derive in Section 2.4.1, the marginal distribution in (2.5) is multivariate normal (MVN) with a mean and variance that depend only on the fixed model parameters θ, the adherence measures $\{\boldsymbol{c}_{it}\}$, and individual covariates $\boldsymbol{x}_i$. This marginalization is made possible by the normality of the latent state innovations $\boldsymbol{\nu}_{it}$ and sampling error $\boldsymbol{\varepsilon}_{it}$. Inference for θ can therefore be obtained directly from this marginal distribution.

2.4.1 Marginal Likelihood of the DLM

The marginalization in Equation (2.5) integrates across both time and outcome dimensions, but for simplicity we explain the marginalization in the case of a DLM with a single outcome variable ($k = 1$) and for one individual, and assume a single baseline covariate ($p = 1$) and single time-varying covariate ($r = 1$) which is assumed to be an adherence indicator. The parameters in θ are therefore $(\beta, \sigma^2_\varepsilon, \rho, \phi, \sigma^2_\nu, \sigma^2_0)$. For now we assume that both adherence indicators and *potential* time-varying outcomes are measured over T consecutive days; the outcomes are $(y_1, y_2, ..., y_T)$ and the adherence indicators are $\boldsymbol{c} = (c_1, c_2, ..., c_T)$. Later, we will subset the outcomes to the *actual* observations times. We can write the vector of outcomes

as the sum of a shared non-time-varying component, a time-varying component and an error term as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix} = \begin{pmatrix} \beta x \\ \beta x \\ \vdots \\ \beta x \end{pmatrix} + \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_T \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{pmatrix} = \beta x \, \mathbb{1}_T + \boldsymbol{\alpha} + \boldsymbol{\varepsilon}. \tag{2.6}$$

where we use the $\mathbb{1}_T$ notation to indicate a column vector of length T consisting of all ones. The following development is conditional on θ and covariates $(x, \boldsymbol{c})$ unless noted otherwise. Thus the first term on the right hand side of Equation (2.6), $\beta x \, \mathbb{1}_T$, is treated as a constant vector. The third term is distributed as $\boldsymbol{\varepsilon} \sim N_T(\mathbf{0}, \sigma_\varepsilon^2 I)$, where $\mathbf{0}$ is a column vector of zeros, and I is a T-dimensional identity matrix.

Because the initial latent variable α_1 is normally distributed, and each α_t conditional on the previous latent variables is a linear combination of normal random variables, then $(\alpha_1, ..., \alpha_T)$ is MVN. The AR(1) structure of the latent states admits a recursive mean and variance calculation

$$\mathbb{E}[\alpha_t] = \rho \mathbb{E}[\alpha_{t-1}] + \phi \, c_t \quad \text{and} \quad \mathbb{V}ar[\alpha_t] = \rho^2 \mathbb{V}ar[\alpha_{t-1}] + \sigma_\nu^2.$$

This recursion and the initial conditions imply a general formula for the mean and variance of α_t

$$\begin{aligned} E_t = \mathbb{E}[\alpha_t] &= \phi \sum_{k=2}^{t} \rho^{t-k} c_k \\ V_t = \mathbb{V}ar[\alpha_t] &= \sigma_\nu^2 \sum_{k=2}^{t} \rho^{2(t-k)} + \rho^{2(t-1)} \sigma_0^2 \end{aligned} \tag{2.7}$$

for $t \in \{2, ..., T\}$ with initial mean $E_1 = \mathbb{E}[\alpha_1] = 0$ and initial variance $V_1 = \mathbb{V}ar[\alpha_1] = \sigma_0^2$. We collect the mean terms into a vector

$$\mathbf{E} = (E_1, ..., E_T). \tag{2.8}$$

Applying a similar recursion to the covariance, we obtain that $\mathbb{C}ov(\alpha_t, \alpha_{t-k}) = \rho^k V_{t-k}$ for $t \in \{2, ..., T\}$

and $k \in \{0, ..., t-1\}$. More compactly, the covariance matrix is

$$\Sigma = \begin{pmatrix} V_1 & \rho\, V_1 & \rho^2\, V_1 & \dots & \rho^{T-1}\, V_1 \\ \rho\, V_1 & V_2 & \rho\, V_2 & \dots & \rho^{T-2}\, V_2 \\ \rho^2\, V_1 & \rho\, V_2 & V_3 & \dots & \rho^{T-3}\, V_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1}\, V_1 & \rho^{T-2}\, V_2 & \rho^{T-3}\, V_3 & \dots & V_T \end{pmatrix}. \tag{2.9}$$

The distribution of the terms in Equation (2.6) can be written explicitly as

$$(y_1, y_2, ..., y_T)^T \sim N_T(\beta x\, \mathbb{1}_T + \mathbf{E}, \Sigma + \sigma_\varepsilon^2 I). \tag{2.10}$$

The distribution in (2.10) is proportional to the marginal likelihood for T sequentially-observed outcomes, marginalizing out the latent parameters.

Let $\mathcal{T} = \{t_1, t_2, ..., t_m\}$ be the *actual* observation times, and let $\boldsymbol{E}_\mathcal{T} = (E_{t_1}, E_{t_2}, ..., E_{t_m})^T$. By properties of the MVN distribution, the outcome vector $y_\cdot = (y_{t_1}, y_{t_2}, ..., y_{t_m})^T$ is also MVN with mean and covariance respectively

$$\mathbb{E}[y_\cdot] = \beta x\, \mathbb{1}_{|\mathcal{T}|} + \boldsymbol{E}_\mathcal{T} \quad \text{and} \quad \mathbb{V}ar[y_\cdot] = \Sigma_{\mathcal{T}\times\mathcal{T}} + \sigma_\varepsilon^2 I_{|\mathcal{T}|}. \tag{2.11}$$

The notation $\Sigma_{\mathcal{T}\times\mathcal{T}}$ indicates subsetting the covariance matrix Σ to the corresponding rows and columns designated by $\mathcal{T}$. If, for example, $\mathcal{T} = \{1, 7, 9\}$ the first, seventh and ninth rows and columns are taken from Σ making $\Sigma_{\mathcal{T}\times\mathcal{T}}$ a 3×3 matrix. One consequence of the marginalization is that if we observe multiple measurements of the same outcome on a single day, they will only vary according to σ_ε^2.

It is worth noting that even though this marginal distribution does not depend on the time-varying parameters, the autoregressive structure remains in both the mean and covariance. For example, the centered marginal mean of y_t can be shown with simple algebraic manipulation to be

$$(\mathbb{E}[y_t] - \beta x) = \rho\, (\mathbb{E}[y_{t-1}] - \beta x) + \phi c_t.$$

This is the same recursive relationship of the dynamic component of our DLM in (2.2). Even though we marginalize out the time-varying parameters, their sequential structure is retained.

The calculations above were derived for one study participant and a single outcome variable. These calculations can also be extended to multivariate outcome measurements. The marginal mean and covariance are both individual- and outcome- dependent, as both $\boldsymbol{\phi}_k$ and α_{itk} vary by outcome and individual. When referring to multiple study participants and multiple outcome variables, we can make the dependence explicit by denoting the quantities in (2.8) and (2.9) for outcome k of individual i as $\mathbf{E}_{ik}$ and Σ_{ik} respectively.

Because the majority of the structure is contained within each outcome (recall $\boldsymbol{\rho}$ is a diagonal matrix), we organize the multivariate outcomes as follows. We collect the outcome k observed through time $t = 1, ..., T_i$ into a $T_i \times 1$ vector $\boldsymbol{y}_{i\cdot k} = (y_{i1k}, y_{i2k}, ..., y_{iT_ik})^T$. The marginal likelihood of a vector of complete outcome measurements can be written as

$$\begin{pmatrix} \boldsymbol{y}_{i\cdot 1} \\ \boldsymbol{y}_{i\cdot 2} \\ \vdots \\ \boldsymbol{y}_{i\cdot k} \end{pmatrix} \mid \theta \sim N_{kT}\left(\boldsymbol{E}_i + \boldsymbol{\beta}\boldsymbol{x}_i \otimes \mathbb{1}_{T_i},\ \Sigma_i + \Sigma_\varepsilon \otimes I_{T_i}\right) \tag{2.12}$$

where $\otimes$ is the Kronecker product, and where

$$\boldsymbol{E}_i = \begin{pmatrix} \mathbf{E}_{i1} \\ \mathbf{E}_{i2} \\ \vdots \\ \mathbf{E}_{ik} \end{pmatrix} \quad \text{and} \quad \Sigma_i = \begin{pmatrix} \Sigma_{i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \Sigma_{i2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \Sigma_{ik} \end{pmatrix}. \tag{2.13}$$

As in (2.11), we can derive the marginal distribution of the outcome measures $\boldsymbol{y}_{i\cdot\cdot}|\theta$ by subsetting the appropriate rows and columns of the quantities in (2.13).

If we assume that outcomes between study participants are independent conditional on θ, the marginal

distribution of outcomes unconditional on the time-varying parameters is given by

$$p\left(\boldsymbol{y}_{1\cdot\cdot}, ..., \boldsymbol{y}_{n\cdot\cdot}|\theta, \boldsymbol{x}, \boldsymbol{c}\right) = \prod_{i=1}^{n} p\left(\boldsymbol{y}_{i\cdot\cdot}|\theta, x_i, \boldsymbol{c}_{i\cdot\cdot}\right). \tag{2.14}$$

This simple marginalization drastically reduces the number of parameters in the model and simplifies computation. We can conduct inference on our non-dynamic parameters in a Bayesian setting by introducing a prior distribution for θ and using the marginal likelihood of θ, which is proportional to (2.14). The posterior density for the non-dynamic parameters is given by

$$p\left(\theta|\left\{\boldsymbol{y}_{i\cdot\cdot}\right\}_{i=1}^{n}, \boldsymbol{x}, \boldsymbol{c}\right) \propto p(\theta) \prod_{i=1}^{n} p\left(\boldsymbol{y}_{i\cdot\cdot}|\theta, x_i, \boldsymbol{c}_i\right). \tag{2.15}$$

Inference can be performed in a straightforward and efficient way via MCMC, or using Hamiltonian Monte Carlo (Homan & Gelman, 2014) sampling to obtain draws from the posterior distribution, as implemented in STAN (Carpenter et al., 2017). Alternatively, posterior samples could be obtained using Sequential Monte-carlo (SMC) in software packages like Libbi (Murray, 2015). SMC works with the sequential likelihoods $p(y_{it_j}|y_{it_1}, ..., y_{it_{j-1}}, \theta)$, which are easily available in our framework because they are conditional distributions of the Multivariate Normal distribution in (2.12).

2.4.2 Inference for the Latent Process

Posterior inference for $\boldsymbol{\alpha}$ can be determined once $p(\theta|\boldsymbol{y})$ has been obtained by exploiting the factorization of the posterior density in (2.3). To do so, we note that the joint distribution of $\boldsymbol{\alpha}$ and $\boldsymbol{y}$ conditional on θ is MVN. For our development, we consider the case in which outcomes are observed for T consecutive days. With a k-dimensional outcome variable, $\boldsymbol{\alpha}$ contains kT elements and $\boldsymbol{y}$ contains kT (scalar) outcomes.

Using the notation in (2.13), the joint distribution of outcomes and time-varying parameters is

$$\left.\begin{pmatrix} \boldsymbol{\alpha}_{i\cdot 1} \\ \vdots \\ \boldsymbol{\alpha}_{i\cdot k} \\ \boldsymbol{y}_{i\cdot 1} \\ \vdots \\ \boldsymbol{y}_{i\cdot k} \end{pmatrix}\right| \theta \sim N_{2kT}\left(\begin{pmatrix} \mathbf{E}_i \\ \\ \mathbf{E}_i + \boldsymbol{\beta}\boldsymbol{x}_i \otimes \mathbb{1}_{T_i} \end{pmatrix}, \begin{pmatrix} \Sigma_i & \Sigma_i \\ \\ \Sigma_i & \Sigma_i + \Sigma_\varepsilon \otimes I_{T_i} \end{pmatrix}\right). \tag{2.16}$$

The quantities on the diagonal of the covariance matrix in (2.16) were determined in Section 2.4.1. The covariance of $\boldsymbol{y}_{i\cdot k}$ and $\boldsymbol{\alpha}_{i\cdot \ell}$ (for $k \neq \ell$), conditional on θ, can be shown to be $\mathbf{0}$ because of independence assumptions made in Section 2.3 (with conditioning on θ suppressed).

$$\begin{aligned} \mathbb{C}ov(\boldsymbol{\alpha}_{i\cdot l}, \boldsymbol{y}_{i\cdot k}) &= \mathbb{E}\left[\mathbb{C}ov\left(\boldsymbol{\alpha}_{i\cdot l}, \boldsymbol{y}_{i\cdot k} | \boldsymbol{\alpha}_{i\cdot k}\right)\right] + \mathbb{C}ov\left(\mathbb{E}[\boldsymbol{\alpha}_{i\cdot l}|\boldsymbol{\alpha}_{i\cdot k}], \mathbb{E}[\boldsymbol{y}_{i\cdot k}|\boldsymbol{\alpha}_{i\cdot k}]\right) \\ &= \mathbb{E}\left[\mathbb{C}ov\left(\boldsymbol{\alpha}_{i\cdot l}, \boldsymbol{\varepsilon}_{i\cdot k}\right)\right] + \mathbb{C}ov\left(\boldsymbol{\alpha}_{i\cdot l}, \boldsymbol{\alpha}_{i\cdot k}\right) = \mathbf{0}. \end{aligned}$$

We should subset the mean vector and covariance matrix of (2.16) as in (2.11) to account for the sporadically observed outcomes, but in this derivation continue to assume that the response values are observed at each time period. The conditional posterior distribution of the latent parameters is determined from the joint distribution in (2.16) by conditioning on $\boldsymbol{y}$ as

$$\left.\begin{pmatrix} \boldsymbol{\alpha}_{i\cdot 1} \\ \vdots \\ \boldsymbol{\alpha}_{i\cdot k} \end{pmatrix}\right| \boldsymbol{y}_{i\cdot 1}, ..., \boldsymbol{y}_{i\cdot K}, \theta \sim N_{kT}\left(\tilde{\alpha}_i, \tilde{\Sigma}_i\right) \tag{2.17}$$

where

$$\tilde{\alpha}_i = \mathbf{E}_i + \Sigma_i\left(\Sigma_i + \Sigma_\varepsilon \otimes I_{T_i}\right)^{-1}\left[\begin{pmatrix} \boldsymbol{y}_{i\cdot 1} \\ \vdots \\ \boldsymbol{y}_{i\cdot k} \end{pmatrix} - \left(\mathbf{E}_i + \boldsymbol{\beta}\boldsymbol{x}_i \otimes \mathbb{1}_{T_i}\right)\right]$$

and

$$\tilde{\Sigma}_i = \Sigma_i - \Sigma_i \left(\Sigma_i + \Sigma_\varepsilon \otimes I_{T_i}\right)^{-1} \Sigma_i.$$

We can use the joint distribution specified in (2.17) along with the posterior distribution on the structural parameters shown in (2.15) to sample from the latent process parameters by first simulating $\theta^* \sim p\left(\theta | \boldsymbol{y}_{1\cdot\cdot}, ..., \boldsymbol{y}_{n\cdot\cdot}, \boldsymbol{x}, \boldsymbol{c}\right)$ and then sampling from $p(\boldsymbol{\alpha}_{i\cdot 1}, ..., \boldsymbol{\alpha}_{i\cdot K} | \boldsymbol{y}_{1\cdot\cdot}, ..., \boldsymbol{y}_{n\cdot\cdot}, \theta^*)$.

2.5 Application to BP Study

We return to modeling time-varying BP as a function of adherence to anti-hypertensive medication and baseline comorbidities and demographics described in Section 2.2. We begin with a discussion of models often used in this task that incorporate adherence as non-dynamic information. We also address missing adherence measures for some patients.

2.5.1 Non-dynamic models incorporating adherence

Adherence to medication is typically incorporated into outcome models in one of two ways. The average adherence for the study period is usually either used directly or is used to dichotomize patients into two groups, those below a certain threshold and those above, indicating "poor" versus "good" adherence (Rose et al., 2011; Lee et al., 2006; Schroeder et al., 2004). Either of these approaches includes adherence as a non-dynamic covariate in the model. Repeated outcome measures are modeled with patient-specific random effects.

We present these alternative models in the context of BP outcomes, a bivariate measure. We label the outcomes 1 for Systolic BP and 2 for Diastolic BP. These models take the form

$$\begin{pmatrix} y_{it1} \\ y_{it2} \end{pmatrix} = \begin{pmatrix} x_i\beta_1 \\ x_i\beta_2 \end{pmatrix} + \begin{pmatrix} \delta_{i1} \\ \delta_{i2} \end{pmatrix} + \begin{pmatrix} \bar{c}_i\gamma_1 \\ \bar{c}_i\gamma_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{it1} \\ \varepsilon_{it2} \end{pmatrix}. \tag{2.18}$$

We consider two different possible adherence measures, $\bar{c}_i$. We first consider average adherence, $\bar{c}_i = T_i^{-1} \sum_{t=1}^{T_i} c_{it}$. In this case we would interpret the adherence effect parameters (γ_1, γ_2) as the differences in

BP of being fully adherent relative to being fully non-adherent, controlling for the baseline covariates. We also consider a dichotomized summary

$$\bar{c}_i = \mathbb{1}\left\{\left(T_i^{-1}\sum_{t=1}^{T_i} c_{it}\right) > 2p - 1\right\},$$

an indicator of overall adherence. In this case (γ_1, γ_2) would be interpreted as the differences in BP for those with "good" ($\bar{c}_i = 1$) versus "poor" ($\bar{c}_i = 0$) adherence. Several values of p were considered to assess the sensitivity to this choice, but we present results for $p = 0.8$ which is a conventional choice (Schroeder et al., 2004). The model in Equation (2.18) includes patient-specific random effects $\delta_{i1} \sim N(0, \sigma^2_{\delta 1})$ and $\delta_{i2} \sim N(0, \sigma^2_{\delta 2})$.

We compare the fit of the above two non-dynamic models to our dynamic model framework. The non-dynamic approach has a potential advantage of being more robust to model misspecification relative to our dynamic model, particularly with the choice of the specific model for the evolution of the health measures as a function of daily adherence. However, incorporating average adherence may mask important time-varying effects of detailed adherence. We explore this tradeoff in Section 2.5.3.

2.5.2 Missing adherence indicators

Not all of the patients in our cohort have completely observed adherence indicators. Among the 503 patients in our analyses, 70 have at least one day of missing adherence. Adherence could be missing due to MEMS cap malfunctions, hospital inpatient stays in which the MEMS containers were not used, or other causes. Fifty of the 70 patients had only one or two missing adherence values. In the most extreme case one patient was missing 48 out of 102 adherence measures.

For patient i, let $\boldsymbol{c}_i^{obs} = \{c_{it}\}_{t \in \mathcal{T}_{obs}}$ be the set of observed adherence values and $\boldsymbol{c}_i^{mis} = \{c_{it}\}_{t \in \mathcal{T}_{mis}}$ be the set of missing adherence values. Letting η_i represent the parameters describing a potential adherence

model for patient i, the posterior density in (2.15) conditional on the observed adherence data only is

$$\begin{aligned} p\big(\theta|\boldsymbol{y}_{i\cdot\cdot}, x_i, \boldsymbol{c}_i^{obs}\big) &= \int p\big(\theta, \boldsymbol{c}_i^{mis}, \eta_i|\boldsymbol{y}_{i\cdot\cdot}, x_i, \boldsymbol{c}_i^{obs}\big)\, d\boldsymbol{c}_i^{mis}\, d\eta_i \\ &= \int p\big(\theta|\boldsymbol{y}_{i\cdot\cdot}, x_i, \boldsymbol{c}_i^{obs}, \boldsymbol{c}_i^{mis}, \eta_i\big) p\big(\boldsymbol{c}_i^{mis}|\boldsymbol{y}_{i\cdot\cdot}, x_i, \boldsymbol{c}_i^{obs}, \eta_i\big) p(\eta_i|\boldsymbol{y}_{i\cdot\cdot}, x_i, \boldsymbol{c}_i^{obs})\, d\boldsymbol{c}_i^{mis}\, d\eta_i \\ &\overset{(a)}{=} \int p\big(\theta|\boldsymbol{y}_{i\cdot\cdot}, x_i, \boldsymbol{c}_i^{obs}, \boldsymbol{c}_i^{mis}, \eta_i\big) p\big(\boldsymbol{c}_i^{mis}|\eta_i\big) p(\eta_i|\boldsymbol{c}_i^{obs})\, d\boldsymbol{c}_i^{mis}\, d\eta_i. \end{aligned} \tag{2.19}$$

The last line in (2.19) labeled (a) involves a set of modeling choices. First, we assume that the adherence model parameters do not depend on other covariates or the outcomes (conditional on the adherence values), and this is reflected by assuming $p(\eta_i|\boldsymbol{y}_{i\cdot\cdot}, x_i, \boldsymbol{c}_i^{obs}) = p(\eta_i|\boldsymbol{c}_i^{obs})$. Other approaches, such as those described by Naranjo et al. (2013), provide a framework for modeling missing time-varying covariates in DLMs with complex models. Our assumption is more conservative since it does not use additional potentially informative data, but it is also likely more robust to model misspecification. Second, $p\big(\boldsymbol{c}_i^{mis}|\boldsymbol{y}_{i\cdot\cdot}, x_i, \boldsymbol{c}_i^{obs}, \eta_i\big) = p\big(\boldsymbol{c}_i^{mis}|\eta_i\big)$ implies that the missing adherence values can be simulated directly from our adherence model, represented by η_i, without regard to individual-level characteristics. Again, this is a conservative choice because adding information to the imputation model could help improve predictions if properly modeled.

We simulate the missing adherence measures from Beta-Bernoulli distributions that depend only on adherence measures for each patient separately. Letting η_i represent patient i's average adherence, the unobserved adherence values are drawn from a Bernoulli distribution

$$\{c_{it}^*\}_{t\in\mathcal{T}}|\eta_i \overset{i.i.d.}{\sim} \text{Bernoulli}(\eta_i)$$

where the $c_{it} = 2c_{it}^* - 1$ take on values $\{-1, 1\}$. Under a uniform prior distribution on patient i's adherence rate, the posterior distribution for η_i is given by

$$\eta_i|\{c_{it}\}_{t\in\mathcal{T}_{obs}} \sim \text{Beta}(n_{i,1} + 1, n_{i,-1} + 1)$$

when they were observed to be adherent $n_{i,1}$ days and non-adherent $n_{i,-1}$ days.

We repeated the simulation process 20 times and combined the posterior samples for a Bayesian multiple-imputation analysis (Little & Rubin, 2002) for our non-dynamic models. For our fully Bayesian model, the simulated adherence values were incorporated into our posterior simulation analyses. We found that in both cases inference of our non-dynamic parameters θ was not sensitive to the missing adherence values.

2.5.3 Analysis of BP measures

Table 2.1 displays the number of BP measurements among our 503 patients during the study period. While a 351 patients had either one or two BP readings, a substantial number of patients had 3 or more. The number of BP readings totalled 1152, averaging 2.29 readings per patient.

Number of BP Readings	1	2	3	4	5	6	7	8+
Number of Patients	226	125	72	28	16	12	15	9
Percent of Patients	44.90	24.90	14.30	5.60	3.20	2.40	3.00	1.79

Table 2.1: Number of BP readings during the study period.

For the period of the study, the proportion of days that patients took their medication varied widely. The proportions ranged from 8.9% to 100%, with a median of 95.1% and a mean of 88.9%. These adherence rates on average were high, with many patients being fully adherent throughout the study period. Only 19.1% of patients had below 80% adherence. The high degree of adherence is consistent with recruiting patients for the study who were continual users of anti-hypertensive medication. The average number of days for which adherence was recorded was 98 days per patient (minimum and maximum of 21 and 395 days, respectively, with an inter-quartile range of 14), with 75% of patients followed between 84 and 112 days.

Baseline summaries of the non-dynamic covariates appear in Table 2.2. A majority of the cohort consisted of women, more patients of black race (African or Caribbean descent) than white, and a large fraction of low-income patients. The cohort also consisted of mostly obese patients, and had a moderately high comorbidity burden (89% had at least one comorbidity). Based on 503 patients with BP readings within 14 days of enrollment, the cohort on average had relatively well-controlled hypertension at baseline, as all patients were prescribed anti-hypertensive medication (though adherent to varying extents). The cohort

	Mean (Std Dev)
Age (y)	$60.3(11.1)$
DBP at Enrollment	$79.7(11.5)$
SBP at Enrollment	$133.5(19.6)$
	Percent
Female	67.5
African-American	54.9
Income below $\$20,000$	44.3
Obese	60.6
Cerebral Vascular Disease	5.6
Congestive Heart Failure	3.8
Renal Insufficiency	6.8
Coronary Artery Disease	14.9
Diabetes	36.2
Hyperlypidemia	57.1
Peripheral Vascular Disease	6.2

Table 2.2: Baseline socio-demographic and health characteristics of the patients in the study cohort.

consisted of 26.2% having DBP>80 mm Hg and SBP>130 mm Hg, and 7.8% having DBP>90 mm Hg and SBP>140 mm Hg.

We assume the same prior distributions for all models when possible. For $k = 1, 2$, indicating systolic and diastolic BP measures separately, we assume the following.

$$\rho_k \sim U(-1,1), \phi_k \sim N(0,25), \gamma_k \sim N(0,25)$$

$$\sigma_{\varepsilon k} \sim U(0,30), \rho_\varepsilon \sim U(-1,1)$$

$$\sigma_{\nu k} \sim U(0,10), \sigma_{0k} \sim U(0,30)$$

$$\beta_{11} \sim N(120,400), \beta_{12} \sim N(80,400)$$

$$\beta_{jk} \sim N(0,400), j = 2, ..., p$$

$$\sigma_{Uk} \sim U(0,30)$$

The prior components were selected to be vague but proper. The intercepts β_{1k} had distributions centered near the typical systolic and diastolic BPs, but had variances that were sufficiently large to acknowledge the uncertainty in the effects. We assumed uniform prior components with compact support for the stan-

dard deviation parameters, as recommended by Gelman et al. (2006). The correlation and autocorrelation parameters were assumed to have uniform priors as in the dynamic model parametrization. Convergence of the MCMC simulated values was inspected with trace plots of multiple chains, and using the Gelman-Rubin convergence statistic (Gelman & Rubin, 1992).

Table 2.3 presents posterior means and 90% central posterior intervals for the non-dynamic covariate effects using the DLM. The point estimates and intervals are reported for the DLM only; the effects in

Variable	Systolic	(90% CI)	Diastolic	(90% CI)
Intercept	132.97	$(129.46, 136.51)^{\star\dagger\ddagger}$	85.68	$(83.65, 87.72)^{\star\dagger\ddagger}$
Sex (male)	-1.1	$(-3.59, 1.38)$	0.53	$(-0.88, 1.94)$
Age (group 1)	-0.01	$(-3.3, 3.27)$	-1.5	$(-3.4, 0.4)$
Age (group 2)	1.87	$(-1.45, 5.19)$	-3.54	$(-5.45, -1.63)^{\star\dagger\ddagger}$
Age (group 3)	5.22	$(1.57, 8.89)^{\star\dagger\ddagger}$	-6.73	$(-8.83, -4.62)^{\star\dagger\ddagger}$
White	-3.31	$(-5.61, -1.01)^{\star\dagger\ddagger}$	-1.58	$(-2.88, -0.26)^{\star}$
Obese	3.07	$(0.76, 5.38)^{\star\dagger\ddagger}$	1.68	$(0.35, 3.01)^{\star\dagger\ddagger}$
Nicotine dependence	-0.9	$(-5.22, 3.42)$	1.3	$(-1.17, 3.76)$
Hyperlipidemia	-1.44	$(-3.71, 0.83)$	-1.34	$(-2.64, -0.04)^{\star}$
Diabetes	1.44	$(-0.96, 3.85)$	-2.98	$(-4.34, -1.61)^{\star\dagger\ddagger}$
Peripheral vascular disease	-1.28	$(-5.75, 3.17)$	-2.6	$(-5.13, -0.06)^{\star\dagger}$
Renal insufficiency	-0.67	$(-4.88, 3.52)$	-2.73	$(-5.14, -0.33)^{\star\dagger\ddagger}$
Benign prostatic hypertrophy	2.82	$(-3.66, 9.3)$	-1.59	$(-5.34, 2.15)$
Coronary artery disease	-1.66	$(-4.77, 1.45)$	-2.76	$(-4.54, -0.98)^{\star\dagger\ddagger}$
Congestive heart failure	-0.78	$(-6.05, 4.46)$	0.59	$(-2.44, 3.59)$
Cerebral vascular disease	2.45	$(-2.19, 7.06)$	-0.17	$(-2.82, 2.48)$

Table 2.3: Summaries of covariate effects for the bivariate dynamic linear model. *Effect with 90% posterior interval not containing 0 in DLM, †Significant effect at the 0.1 level in average adherence model, ‡Significant effect at the 0.1 level in dichotomized adherence model.

the alternative models that were significant at the 0.1 level are indicated with a † (for the average adherence model) or a ‡ (for the dichotomized adherence model). Effects with 90% central posterior intervals not containing 0 are marked with an asterisk ($\star$). Based on the model fits, the estimated covariate effects tend to be similar across all models with the point estimates tending to agree in magnitude and sign. Even though the DLM covariate effects tended to have narrower intervals on average (3% reduction), the significance of the findings tended to agree as well. In particular, the effect of race (white versus non-white) was significantly negative, indicating that whites tended to have lower blood pressure controlling for all other variables and time-varying adherence. Patients who were obese at the beginning of the study tended

to have significantly higher DBP and SBP. These findings are consistent with the results of previous studies (Kressin et al., 2010; Rose et al., 2011) in their significance and direction of the effects. Both of these, except for the effect of being white on mean diastolic BP, agreed with the alternative models in terms of significance and directionality of the effect.

Table 2.4 reports inferences for the standard error and correlations of $(\boldsymbol{\varepsilon}_{it})$ for all three models. Com-

Dynamic linear model				
Variable	Systolic	(90% CI)	Diastolic	(90% CI)
Standard Errors:	13.37	$(12.46, 14.25)$	8.11	$(7.68, 8.54)$
Correlation:	0.62	$(0.57, 0.67)$		
Average Adherence Model				
Variable	Systolic	(90% CI)	Diastolic	(90% CI)
Standard Errors:	14.53	(13.87, 15.21)	8.34	(7.97, 8.72)
Correlation:	0.58	(0.54, 0.62)		
Dichotomized Adherence Model				
Variable	Systolic	(90% CI)	Diastolic	(90% CI)
Standard Errors:	14.5	(13.86, 15.18)	8.34	(7.96, 8.72)
Correlation:	0.58	(0.54, 0.62)		

Table 2.4: Measurement error estimates for three models.

paring the sampling standard deviation estimates across models provides an indication of the gains in modeling the adherence effects as time-varying. The standard error estimates for the alternative adherence models tend to be slightly larger than those given by the DLM, which is consistent with previous work (Rose et al., 2011). The time-varying adherence explicitly captured in the DLM may account for the extra variation in the outcomes of the alternative adherence models through a reduction in the estimated measurement error variance.

Table 2.5 contains the adherence effects estimated from our models. The adherence effects across the three models are not directly comparable, given the different approaches to incorporating adherence. The average adherence model indicates that the difference between those who were fully adherent and those who were fully non-adherent, controlling for other covariates, is about -9.2 and -9.5 for systolic and diastolic, respectively. This implies, for example, that a 10% additive increase in adherence corresponds to a 0.92 and 0.95 reduction in systolic and diastolic blood pressure, respectively. However, this effect size assumes that the relationship between average adherence and blood pressure is linear and holds through-

State Space Model				
Variable	Systolic	(90% CI)	Diastolic	(90% CI)
Adherence effect:	-0.48	(-0.84, -0.2)*	-0.24	(-0.43, -0.09)*
Asymptotic Adherence effect:	-3.87	(-5.98, -1.83)*	-3.15	(-4.38, -1.94)*
Alternative Adherence Model: Average Adherence				
Variable	Systolic	(90% CI)	Diastolic	(90% CI)
Adherence effect:	-9.24	(-16.03, -2.22)*	-9.46	(-13.55, -5.52)*
Alternative Adherence Model: Dichotomized Adherence ($p = 0.8$)				
Variable	Systolic	(90% CI)	Diastolic	(90% CI)
Adherence effect:	-5.49	(-8.26, -2.75)*	-3.78	(-5.27, -2.2)*

Table 2.5: Adherence effects for the models considered

out the entire range of adherence. Given the limited range of average adherence observed in the data (90% of patients have average adherence above 70.5%), interpreting this effect beyond this range is not recommended because it involves extrapolating beyond the data. The dichotomized adherence model shows similar results. In particular this approach involves comparing those with relatively good overall adherence (above 80% adherent) to everyone else. Based on the dichotomized adherence model, the benefit of being in the former group is indicated by a lower blood pressure of -5.49 and -3.78 on average for systolic and diastolic blood pressure.

The DLM gives similar results. The effect of taking medication can be inferred on a daily basis. The results of the model fit suggest a small but significant reduction of blood-pressure from taking the medication on a daily basis, -0.24 mm Hg and -0.12 mm Hg on average for systolic and diastolic BP, respectively. These estimates imply that, accounting for the correlation estimate, a patient who is adherent over consecutive days would experience a long-term reduction in systolic BP by -3.9 mm Hg, and a long-term reduction in diastolic BP by -3.15 mm Hg. The magnitudes of an increase in long-term BP for continued non-adherence is the same but in the opposite direction. Overall, the adherence effects tend to agree in terms of the significance and direction for the different models. However, the DLM provides a clearer interpretation of these parameters that is consistent with the time-varying nature of the data.

Another benefit of the DLM is our ability to infer the latent BP for unobserved days using the procedure discussed in Section 2.4.2. Figure 2.4 contains an example of posterior draws of the mean process $\boldsymbol{\mu}_{it\cdot} = \boldsymbol{\beta}\boldsymbol{x}_i + \boldsymbol{\alpha}_{it\cdot}$ for the patient presented in Figure 2.1.

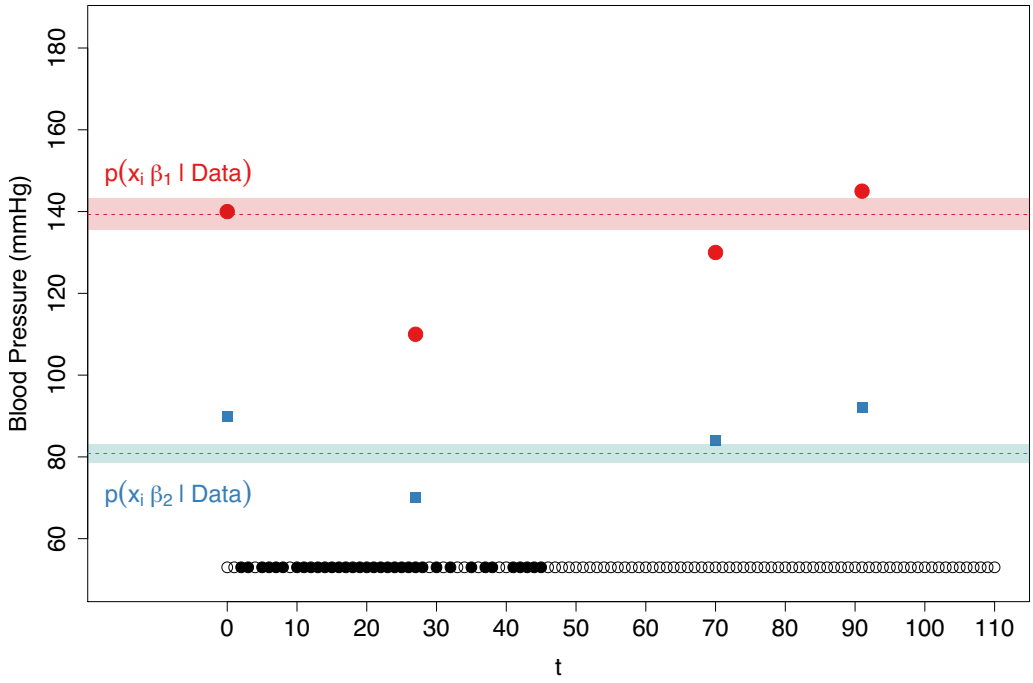


Figure 2.3: 90% central posterior intervals for the overall mean blood pressure as a function of only covariates $(x_i\beta_1, x_i\beta_2)$ for one study subject. The solid lines are the estimated mean DBP and SBP, and envelopes are posterior intervals.

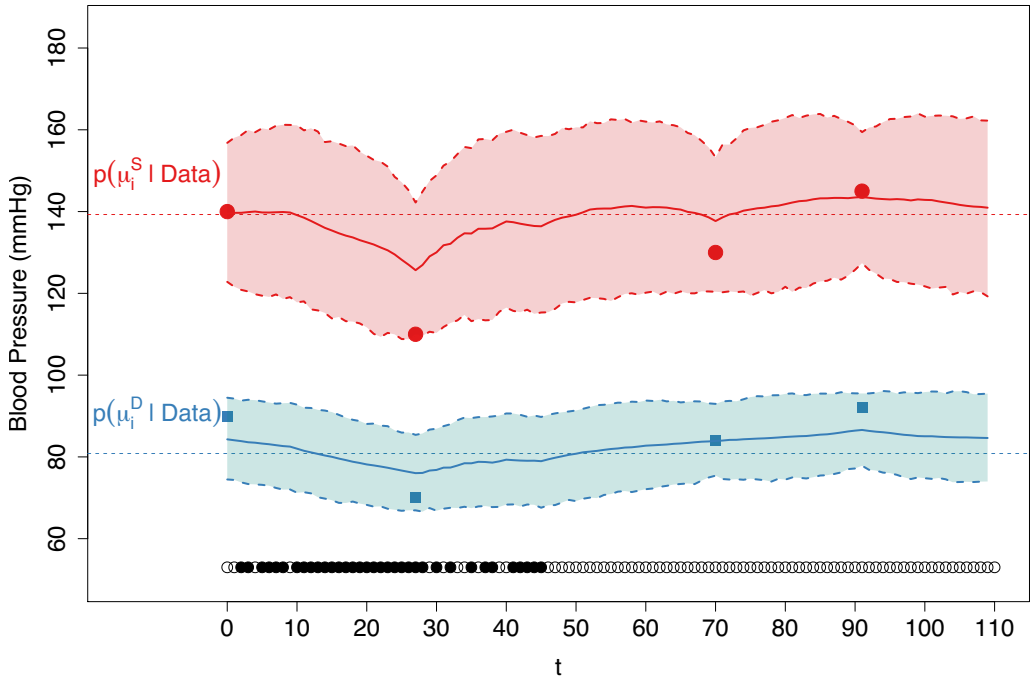


Figure 2.4: 90% central posterior intervals for the mean blood pressure $\{\mu_{it.}\}$ for one study subject. The solid lines are the estimated mean DBP and SBP, and the dashed lines and shading are the point-wise posterior credible intervals. Horizontal dotted lines are drawn at the baseline-estimated mean DBP and SBP.

The solid line indicates the posterior mean of the latent process while the dashed lines indicate a 90% credible envelope across time. That is, for each time t, a 90% credible interval of $\boldsymbol{\alpha}_{it1}, \boldsymbol{\alpha}_{it2} | \{\boldsymbol{y}_{i\cdot\cdot}\}_{i=1}^{n}$ is shown. The estimated mean DBP and SBP processes vary gradually over time, with shifts in direction influenced by medication adherence variation.

2.6 Discussion

In this article we propose a multivariate DLM that can be used for modeling time-varying outcome measures as a function of detailed medication adherence or other time-varying covariates. While DLMs are of common use, our particular setting is unique in that it benefits from a two-stage computational approach made possible by factorizing the posterior density into the dynamic and non-dynamic parameters. Typical analyses in this setting ignore the time-varying structure of medication adherence and instead examine this relationship between adherence and outcomes via correlations with non-dynamic adherence measures (e.g., time-averaged adherence). Our framework explicitly provides a measure of daily impact of medication-taking and meaningful bounds for mean BP while controlling for baseline covariates.

Our modeling approach is sufficiently flexible to permit a wide range of assumptions distinct from those we included in our hypertension application. For example, we assumed an AR(1) modeling structure for the mean outcome process, and this assumption can be justified based on the pharmacokinetic properties of the medications. In other settings, alternative mean processes can be considered, including higher order AR processes, growth curves, and so on. Our baseline covariates were modeled linearly, but our framework permits non-linear inclusion of covariate information for both baseline non-dynamic covariates, as well as the dynamic predictors in the mean process component of our model. A crucial assumption of our framework is that the outcome distribution is multivariate normal, as this assumption allows for the marginalization strategy that leads to an efficient computational procedure. However, with some non-normal outcome distributions, various strategies can be employed that can potentially take advantage of the marginalization idea. For example, non-normal outcome densities can be approximated by normal distributions after which the marginalization can be performed; then a Metropolis-Hastings algorithm would be incorporated into the posterior sampling procedure that would account for the non-normality of the

original sampling distribution. Such a procedure is likely to be far more efficient than sampling the dynamic parameters directly as part of model fitting.

The proposed modeling framework also permits other extensions that are straightforward to incorporate. First, multivariate outcomes recorded at staggered intervals can be accommodated by marginalizing the distribution in (2.12) appropriately. Second, if we believe that the effect of adherence to medication varies from person to person we can use a hierarchical prior on patient-specific adherence effects to share information across individuals. Third, our analyses did not acknowledge differences among anti-hypertensive medications, but our framework easily permits distinguishing differential medication effects. The effects of different medications, perhaps grouped by relevant characteristics such as whether the medication is short-acting or long-lasting, or by medication type (e.g., for hypertension, diuretics, ACE inhibitors, etc.), can be included as separate time-varying effects in the dynamic component of our model. Dosages and dosing frequencies can serve as covariates for the effects of particular medications.

The framework we developed can help establish answers to questions of interest to clinicians and medical researchers that have been difficult to assess through simpler models. In particular, our model can determine the effects of different socio-demographic or health factors on health outcomes that control for detailed time-varying adherence to medication by allowing medication-taking behaviors to change over time. From our framework, we can estimate the daily improvement in being adherent to one's medication, controlling for socio-demographic and health characteristics, but also the likely long-range achievable mean outcomes. We can also use the model to forecast health outcomes as a function of specified patterns of adherence, potentially serving as a tool for medical decision-making by clinicians. Our approach provides a robust framework for understanding the impacts of poor medication adherence as clinicians and patients work together to improve their medication treatment.

Water is so fine that it is impossible to grasp a handful of it. It has no shape of its own but molds itself to the receptacle that contains it. When frozen it crystallizes into a mighty rock. When heated to the state of steam it is invisible but has enough power to split the earth itself.

Bruce Lee

3

A Guided Meditation on the Flow of Information for Model-based Inference

When analyzing data using models, we are extracting information from the data that we hope will inform model parameters. Alternatively, information flows like water to thirsty model parameters. For a fixed model and a set of observed data, the *amount of information* in the data is a limited resource that distributes among the parameters. Statisticians are routinely involved with scientists to help them build models to answer questions of scientific interest. Domain experts have great intuition about what parameters in their models are most (or least) informed by their data, but formalizing this intuition is not always trivial. In this chapter, we build tools that allow us to formalize this intuition. We do this in the hope that it will

sharpen our intuition and allow us to help scientists make more informed model-building decisions. We might remove parameters that are not well-informed by the data, or decide to collect new data that better inform those parameters.

To this end, we seek to understand how the model uses the information in the data to inform model parameters: which parameters are parched and which are sufficiently satiated. The notion that there is limited information in data is not new. Ronald A. Fisher said as much.

> Modern statisticians are familiar with the notion that any finite body of data contains only a limited amount of information on any point under examination; that this limit is set by the nature of the data themselves, and cannot be increased by any amount of ingenuity expended in their statistical examination: that the statistician's task, in fact, is limited to the extraction of the whole of the available information on any particular issue. (Fisher, 1971, D.o.E. pg 40)

This quote was a response to what he saw as an erroneous belief that one could rearrange data in order to gain more information for certain questions. Beyond pointing out this error, Fisher went on to proclaim what a statistician ought to do when tasked to answer a question with data. The statistician ought to extract as much information as possible from the data to answer said question. Of equal importance, the statistician should understand the flow of information from data to model parameters.

This flow, however, is clouded by the fact that information pass not only *directly* from data to parameters but also *indirectly* among the parameters. It is only in special cases that parameters do not have both of these sources of information. For example, unidentifiable parameters (to be made precise later) do not receive *direct information* from data, only *indirect information* from other parameters. Orthogonal parameters, on the other hand, receive only *direct information* from the data and do not pass *indirect information* among themselves.

Beyond different sources of information, varying amounts of information further complicates model-building decisions. The various parts of the data may more or less inform different parts of the parameter space. It is possible that a parameter, though identifiable, receives very little *direct information* relative to other parameters. However, even when extracting very little direct information from the data, this parameter may provide considerable *indirect information* to other model parameters. In this case, the model

builder may rethink including this parameter in their model because it is not being informed well by the data and affects the performance of estimating other parameters.

In this chapter, we explore measures of information flow with an eye on decision-making in the model-building process. We focus our efforts through analysis of examples we introduce in Section 3.1. We provide background information and introduce our two cornerstones of unidentifiable and orthogonal parameters in Sections 3.2.1 and 3.2.2 respectively. In Section 3.3 we discuss measures of information flow based on the Fisher Information matrix. In Section 3.4 we discuss measures based on prior-posterior comparisons for Bayesian analysis. We conclude with a brief discussion in Section 3.5.

3.1 A Normal Example in Double Variation

We begin our exploration of the flow of information by considering two examples that use the Normal distribution. In the first example, we are interested in measuring the correlation coefficient ρ of a Normal bivariate distribution using both bivariate and marginal observations. We refer to this as the *correlation variation*. Specifically, consider observations generated as

$$\begin{aligned}
&\text{Bivariate Observations : } \begin{pmatrix} X_i \\ Y_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \quad i = 1, ..., m; \\
&\text{Marginal Observations : } \quad X_i \sim \mathcal{N}(0, \sigma^2), \quad i = m+1, ..., m+k; \\
&\qquad\qquad\qquad\qquad\quad Y_i \sim \mathcal{N}(0, \sigma^2), \quad i = m+k+1, ..., m+2k.
\end{aligned}$$

We also take the convention that $m = 0$ represents no paired observations and $k = 0$ represents no marginal observations. The paired observations contain direct information about ρ, while the marginal data contain information about σ^2 which in turn indirectly informs the estimation of ρ. As we will see in the coming sections, estimating σ^2 well helps increase the precision in estimating ρ. The observations are collected into one dataset as

$$\mathcal{D}_{m,k} = \left\{ \{(X_i, Y_i)\}_{i=1}^{m}, \{X_i\}_{i=m+1}^{m+k}, , \{Y_i\}_{i=m+k+1}^{m+2k} \right\}. \tag{3.1}$$

In our simulations and analysis we will fix $n = m + k$ and vary m between 0 and n. In our formulation, going from $m = m'$ to $m = m' + 1$ introduces one paired observation and removes one observation from each marginal. We, therefore define the fraction of paired data as $\alpha = m/(m + k)$, we can also denote the dataset as $\mathcal{D}_{m+k,\alpha}$. When $\alpha = 0$ the correlation parameter can not be estimated because ρ is no longer part of the likelihood. As we increase α, i.e. introduce more paired data, estimates of ρ should become more precise. In this sense we can treat α as a proxy for the amount of information for ρ. Increasing α also changes the information for σ^2, in fact it decreases it. Consider the extreme case of $\rho = 1$, the paired data is completely redundant. when $\alpha = 1$ we effectively have $m + k$ observations to estimate σ^2, when $\alpha = 0$, we effectively have $2(m + k)$ observations. This is true to some extent for any $\rho \neq 0$. So, increasing α should also affect our ability to estimate σ^2.

In a second example, we are interested in measuring the means of two Normal distributions. We refer to this as the *mean variation*. Specifically, the observations are generated as

$$\begin{aligned} X_i &\sim N(\theta_1 + \theta_2, 1), \quad i = 1, ..., n_X; \\ Y_j &\sim N(\theta_2, \gamma), \qquad\quad j = 1, ..., n_Y \end{aligned}$$

where (θ_1, θ_2) are unknown model parameters and γ is known. The parameters are related to one another, estimating one helps estimate the other. The quantity γ controls the quality of the data that informs θ_2. If γ is small we can learn θ_2 with fewer samples, but would need more if γ is large. In either case, because our observations $\{X_i\}$ involve both θ_1 and θ_2, γ also affects the estimation of θ_1. At the extreme $\gamma = \infty$, even with an infinite number of samples we can not estimate both θ_1 and θ_2. Here, a case can be made for focusing on the parameterization $(\eta, \lambda) = (\theta_1 + \theta_2, \theta_2)$, an orthogonal parametrization, and dropping λ all together.

Throughout this exposition we consider one parameter to be of primary interest so that there is no ambiguity between direct and indirect information. In the correlation variation, the correlation ρ will be of primary interest so that α balances the amount of direct and indirect information the data provide for ρ. In the mean variation, the parameter θ_1 will be of primary interest, clarifying that γ controls the amount of indirect information for estimating θ_1 and n_X controls the amount of direct information. We will use these

two variations to shed light on both direct and indirect flows of information. In the coming sections, we explore how parameter unidentifiability and orthogonality can be used to understand the direct and indirect information, respectively.

3.2 Background Information

To understand the direct and indirect flow of information, we can use extremes to anchor ourselves. One extreme is the *lack of flow* of direct information. That is, the data provide no information for the parameter of interest and any information it may contain is necessarily through its relation to other parameters. This is often described as *non-identifiability* or unidentifiability and is discussed in Section 3.2.1. Another extreme is the *lack of flow* of indirect information. Here, a parameter is unaffected by the (potentially poor) estimation of other parameters. This is typically seen as a desirable quality of model parameters, one formalization is parameter orthogonality. We explore this in Section 3.2.2.

3.2.1 Direct Information Flow and Identifiability

A statistical model $\{p(\mathcal{D}_n|\theta) : \theta \in \Theta\}$ is said to be *identifiable* if distinct parameter values correspond to distinct probability distributions, i.e.

$$p(\mathcal{D}_n|\theta_1) = p(\mathcal{D}_n|\theta_2) \;\; a.e.\; \mathcal{D}_n \Rightarrow \theta_1 = \theta_2$$

The parameter θ is said to identify this model. Identifiability ensures that the mapping between the parameter-space and model-space is invertible so that learning the model ensures our ability to learn parameters. If a model lacks this property, it is said to be *non-identifiable* or *unidentifiable*.

Even if a model is non-identifiable, it may still be possible to learn about some parameters. Poirier (1998) consider a particular subset of non-identifiable models called Partially Identifiable Models (PIM). Incidentally, the Econometric literature also uses the term PIMs to refer to models whose parameters can be set-identified but not point identified (Christ, 2001; Manski, 2003; Geweke, 2010a; Moon & Schorfheide, 2012), they are also known as Incomplete models (Geweke, 2010b). In other words, parameters can be known up to a range, but can never be precisely learned. This notion is of course related, but we will use

the language of Poirier (1998) and, later, Gustafson (2015) to describe parameter non-identifiability.

Following Gustafson (2015), non-identifiability can be defined on the level of parameters by considering two distinct parameterizations. We first define the original parameterization θ as the *scientific* parameterization to distinguish it from a secondary parameterization $(\eta, \lambda) = g(\theta)$. The parameterization (η, λ) is said to be *transparent* with identifiable parameter η and unidentifiable parameter λ if (1) the likelihood of the data can be written as a function of of η alone,

$$f(\mathcal{D}_n|\theta) = f(\mathcal{D}_n|\eta, \lambda) = f(\mathcal{D}_n|\eta).$$

Moreover, (2) the likelihood $f(\mathcal{D}_n|\eta)$ has the properties that ensure regular parametric asymptotic theory apply. This ensures that the likelihood behaves well enough for conventional inference tools to be used.

The above definition is the product of various lines of research in the use of non-identifiable models in Bayesian analysis (Poirier, 1998; Gustafson, 2009). The primary reason this research occurred in the Bayesian domain is that even when faced with an unidentifiable likelihood, Bayesian parameter inference can proceed. If the analyst uses a proper prior distribution on the scientific parameterization θ, the posterior distribution will also be proper (proof given in Appendix B.1).

Theorem 3.2.1. *Given a proper prior and a likelihood that is integrable over the sample space $\mathcal{D}_n \in \mathscr{D}_n$, the posterior distribution will be proper for almost every $\mathcal{D}_n$, i.e. $\int_\Theta p(\theta)d\theta < \infty$ and $\int_{\mathscr{D}_n} p(\mathcal{D}_n|\theta)d\mathcal{D}_n < \infty$ imply*

$$\int p(\mathcal{D}_n|\theta)p(\theta)d\theta < \infty, \quad a.e. \quad \mathcal{D}_n \in \mathscr{D}_n$$

The role of identifiability in Bayesian modeling has a rich history. Because identifiability is a property of the likelihood, it should be regarded equally whether the inference procedure is classical or Bayesian (Kadane, 1974). This seemingly innocuous thought is common among many statisticians, but a deeper exploration begins to show some cracks. The inferential procedures have radically different consequences. If a model is unidentifiable, we cannot apply classical inference like maximum likelihood without first drastically altering the model to make it identifiable. Bayesian inference with proper priors, on the other hand, can proceed without much trouble as seen in Theorem 3.2.1. In fact, introducing information thorough the prior can be seen as a solution to working with unidentifiable models Gustafson (2005). Regardless, the

Bayesian exploration of identifiability was fruitful in increasing our understanding of identifiability.

In the Bayesian context, both parameters and data are both treated as random quantities. One idea that applied to the likelihood was seamlessly transferred parameters to bring insight on identifiability. In an exploration of the role of conditional independence in statistical theory, Dawid (1979) defined a notion of sufficiency in the context of parameters. If the distribution of $\mathcal{D}_n$ is determined by parameters (η, λ), but we also have that $\mathcal{D}_n$ and λ are independent given η, i.e. $\mathcal{D}_n \perp\!\!\!\perp \lambda \mid \eta$, then λ is not identified and η is said to be a sufficient parameter. In other words, the parameter λ is redundant when η is known. In the same treatise, Dawid (1979) describes the practical consequence of having redundant parameters on posterior inference.

> If η is a sufficient parameter, so that $\mathcal{D}_n \perp\!\!\!\perp \lambda \mid \eta$, and the parameters have a prior distribution, then $\lambda \perp\!\!\!\perp \mathcal{D}_n \mid \eta$ so that $p(\lambda|\mathcal{D}_n, \eta) = p(\lambda|\eta)$. We see that the conditional distribution for the redundant part λ of the parameter, given the sufficient parameter η, is the same in the posterior distribution as in the prior: once we have learned about η from the data, we can learn nothing about λ, over and above what we knew already.* Dawid (1979)

Even in the unidentifiable model case, we will be able to learn about the parameters that matter, but not about others beyond what we believe them to be apriori. The tension here is that the Bayesian analysis, as opposed to classical analysis, seems to retain the ability to make parameter inference, if somewhat hindered, even in situations where we cannot learn about some parameters from the data. We summarize the lack of desire to address this tension with a quote from Lindley (1987)

> In passing it might be noted that underidentifiability causes no real difficulty in the Bayesian approach. If the likelihood does not involve a particular parameter, λ say, when written in the natural form, then the conditional distribution of λ, given the remaining parameters, will be the same before and after the data.†

As Poirier (1998) describe, this seems like a "Bayesian free lunch." The distribution of the unidentifiable

*$\mathcal{D}_n$, η and λ here replace Y, Θ and Φ in the original text for clarity.
†η and λ here replace λ and Ψ in the original text for clarity.

parameter still benefits from data. The marginal posterior distribution of the unidentifiable parameter is

$$p(\lambda|\mathcal{D}_n) = \int p(\lambda, \eta|\mathcal{D}_n)\, d\eta = \int p(\lambda|\eta, \mathcal{D}_n)\, p(\eta|\mathcal{D}_n)\, d\eta \tag{3.2}$$
$$= \int p(\lambda|\eta)\, p(\eta|\mathcal{D}_n)\, d\eta.$$

The marginal posterior distribution here is an average of the prior distribution weighted by the posterior of the identifiable parameter. In other words, the structure of the prior helps provide indirect information from identifiable to unidentifiable parameters. Even when the likelihood is unidentifiable, we can conduct parameter inference, even for the non-identifiable parameters.

Surely then, the Bayesian framework must assume something beyond what the likelihood provides if it is able to retain the ability to conduct parameter inference when classical inference cannot. Poirier (1998) resolve this tension by showing that there are quantities for which the data are uninformative and that the cost of the "free lunch" is the need to specify the prior distribution. The prior distribution wholly determines the inference on the unidentifiable parameters. Modern Bayesian analysts do not place sufficient importance on the prior construction raising the cost of the lunch.

The practical difficulty here, of course, is deciding whether or not a particular parameter is unidentifiable. This is not always clear from typical posterior diagnostics. In fact, marginal prior-posterior comparisons can be misleading as shown in Equation (3.2). When the parameters are unidentifiable, the marginal prior-posterior comparison will show no change when prior distributions are independent, $p(\lambda|\eta) = p(\lambda)$, or the marginal posterior of η shows no change, $p(\eta|y) = p(\eta)$. The latter case occurs when the parameter η is also unidentifiable.

Gustafson developed an asymptotic theory of Bayesian analysis with PIMs in a series of papers (Gustafson, 2005, 2009, 2014, 2015). In PIMs, certain measures of information flow are theoretically justified. Comparing the prior to posterior conditional distributions of non-identifiable parameters (conditional on identifiable parameters) should indeed show *no direct flow* of information. Hence, comparing these distributions leads to a natural measure of direct information. Any measure of difference will yield zero for non-identifiable parameters and non-zero for identifiable parameters (Lindley, 1956; Xie & Carlin, 2006; Raue et al., 2009, 2012). In particular, Xie & Carlin (2006) give measures of identifiability for hierarchical mod-

els that are quite general. We consider measures of this sort for the Bayesian setting, but we also discuss likelihood-based measures in the next section.

3.2.2 Indirect Information Flow and Orthogonality

Orthogonal parameters, like unidentifiable parameters in Section 3.2.1, are defined in terms of the likelihood function. In fact, they are typically defined in terms of the Fisher Information matrix. Following Cox & Reid (1987), if the parameter θ of length p is split into two vectors θ_1, θ_2 of lengths p_1 and p_2. These parameters are said to be orthogonal if

$$i_{st} = \mathbb{E}_{\mathcal{D}_n}\left[\left(\frac{\partial}{\partial\theta_s}\log p(\mathcal{D}_n|\theta)\right)\left(\frac{\partial}{\partial\theta_t}\log p(\mathcal{D}_n|\theta)\right)\right] = 0 \tag{3.3}$$

for $s = 1, ..., p_1$ and $t = p_1 + 1, ..., p_1 + p_2$. This orthogonality is called global if this holds for every value of θ and local at θ^0 if it holds for $\theta = \theta^0$. If we consider the situation when $p_1 = p_2 = 1$ and $\theta = (\eta, \lambda)$ for simplicity, there are several elementary consequences that follow from the orthogonality of these parameters as outlined in Cox & Reid (1987). For example, determining these parameters numerically is greatly simplified and the MLEs, $\widehat{\eta}$ and $\widehat{\lambda}$, are asymptotically independent. There is one particular property that relates to the flow of information, property (iv) in Section 2.2 of Cox & Reid (1987): $\widehat{\eta}_\lambda$, the maximum likelihood estimate for η when λ is given, varies only slowly with λ. If we were to, say, estimate λ poorly, the effect of that on estimating η will be minimal. In other words, the indirect information for η is small, and zero asymptotically. This leads naturally to measures of indirect information where we inspect the relative Fisher information, either given η or not.

Meng & Xie (2013) explore Fisher Information-based measures to study the difference in how procedures incorporate information from data and model assumptions. One of the guiding questions there was to understand how principled estimation procedures, like maximum likelihood, accrue information differently than less principled procedures, like the method of moments. In particular, they show that principled estimation procedures incorporate information from data and modeling assumptions appropriately. They use a variant of our *correlation variation* to study how maximum likelihood estimation incorporates information from data, in the form of new bivariate observations, and from a modeling assumption, in the form

of model restriction (assuming σ^2 as known). Here, the addition of new data contributes to the estimation of the variance σ^2 as well as the correlation ρ. Moreover, making a modeling assumption that σ_2 is known also impacts the estimation of ρ. They separate this by investigating the gains in asymptotic precision for the MLE of ρ under two situations: when σ^2 is known and when it has to be estimated.

Here, a "model assumption" encompasses quite generally the constraining of the model space. For example, adding the assumption of Normality to that of independent and identically distributed (i.i.d.) constrains the model space from all models where the data are i.i.d. to only models where the data are i.i.d and Normal. Of course, much freedom remains in this situation because the model parameters can cover a wide range of data. Similarly, assuming σ^2 to be known constrains the model space.

We can use this framework to study how information flows to different parts of the model. When we assume the *secondary parameter* to be known, we are essentially injecting information into the system that allows us to see the full impact of the data on the *primary parameter*. If the parameters are orthogonal, or nearly orthogonal, we should see that the information for the primary parameter should not change. If, on the other hand, the parameters are far from orthogonal, we should see a radical change in information for the primary parameter if given the secondary parameter.

The asymptotic precision of the MLE, the basis of comparison in Meng & Xie (2013) is defined in terms of the Fisher Information matrix. Consider, for now, a family of distributions whose likelihood is indexed by two parameters (θ_1, θ_2) and whose Fisher information matrix is

$$\mathcal{I}(\theta_1, \theta_2) = \begin{pmatrix} i_{11} & i_{12} \\ i_{21} & i_{22} \end{pmatrix}.$$

A measure of marginal information for θ_1 and information for θ_1 given θ_2 can be defined in terms of these quantities via the asymptotic precision of the MLE. In the case of the 2-dimensional parameter, these are simply

$$I(\theta_1) = i_{11} - \frac{i_{12} i_{21}}{i_{22}} \quad \text{and} \quad I(\theta_1 | \theta_2) = i_{11}. \tag{3.4}$$

These quantities measure the amount of information contained in the data for θ_1 in two situations. First,

$I(\theta_1)$ measures the total amount of information that is passed from the data $\mathcal{D}_n$ to θ_1 both directly, and indirectly. This is captured in Figure 3.1 (left). Second, $I(\theta_1|\theta_2)$ captures the total information for θ_1 *plus the additional* information from assuming a known θ_2. Figure 3.1 (right) indicates assuming θ_2 is known with grey shading. $I(\theta_1|\theta_2)$ can be interpreted as how well we might estimate θ_1 in an ideal setting where we put the full weight of the data to the task without needing to estimate θ_2. $I(\theta_1)$, on the other hand, shows us the information in our actual case. We can also see the loss of precision when needing to estimate θ_2 in addition to θ_1, i.e. $\frac{i_{12}i_{21}}{i_{22}}$.

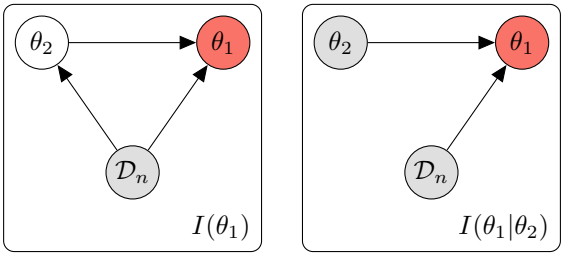


Figure 3.1: A representation of the flow of information from the data to the primary parameter θ_1 when the secondary parameter must be estimated (left) and the secondary parameter is assumed (right).

Following Meng & Xie (2013), the absolute and relative gain in information for θ_1 by conditioning on θ_2 are

$$\mathcal{G}(\theta_1|\theta_2) = I(\theta_1|\theta_2) - I(\theta_1) = \frac{i_{12}i_{21}}{i_{22}} \quad \text{and} \quad \mathcal{R}(\theta_1|\theta_2) = \frac{\mathcal{G}(\theta_1|\theta_2)}{I(\theta_1)} = \frac{I(\theta_1|\theta_2) - I(\theta_1)}{I(\theta_1)}.$$

These measure the gain in information that θ_1 receives from making a modeling assumption that assumes away θ_2 in absolute and relative terms respectively. An alternative interpretation of $\mathcal{G}(\theta_1|\theta_2)$ is that it measures the drain in information incurred by needing to estimate θ_2 in addition to θ_1. If $\mathcal{G}(\theta_1|\theta_2) = 0$, the actual precision for estimating θ_1 is the same as the precision in the ideal case. There are two ways $\mathcal{G}(\theta_1|\theta_2)$ is zero. If the parameters are orthogonal, i.e. $i_{12} = i_{21} = 0$. More precisely estimating one neither helps nor hurts the estimation of the other. Alternatively, if we are able to estimate θ_2 with infinite precision, i.e. $i_{22} = \infty$. In either case, $I(\theta_1)$ and $I(\theta_1|\theta_2)$ measure the same quantity.

We will use these quantities in Section 3.3 to construct likelihood-based measures on information flow. In particular we will factor the information in the ideal setting $I(\theta_1|\theta_2)$ into the sum of information in the

actual information $I(\theta_1)$ and the penalty of diverting some information to estimating θ_2, $\mathcal{G}(\theta_1|\theta_2)$. This breakdown will provide a way of understanding the flow of information from the data to the parameter of interest.

3.3 Likelihood-Based Measures of Information Flow

In this section we introduce measures of information flow by considering the Fisher Information structure. In Section 3.2.2 we introduced measures of information gain, or information drain, as the relative losses in precision of primary parameters when needing to estimate secondary parameters. We build upon those ideas to describe measures of flow of information and use our two examples to build our intuition about these measures.

When considering measures of flow of information from data to a primary parameter θ_1, we can compare the information available in the dataset under two scenarios: if only the primary parameter needs estimation and if both parameters need estimation. If only one parameter, θ_1, needs estimation then the full force of the data is used for estimating θ_1 – the information should be higher. For this we use $I(\theta_1|\theta_2)$ in Equation (3.4). If both parameters need to be estimated, information available is $I(\theta_1)$ in Equation (3.4). The difference between these two, $\mathcal{G}(\theta_1|\theta_2)$, was described in Section 3.2.2 as the relative gain of being in the ideal situation. Putting these together we have that the

$$\text{Ideal Information} = \text{Actual Information} + \text{Redirection Penalty} \tag{3.5}$$

$$\begin{aligned} I(\theta_1|\theta_2) &= I(\theta_1) + \mathcal{G}(\theta_1|\theta_2) \\ i_{11} &= i_{11} - \frac{i_{12}i_{21}}{i_{22}} + \frac{i_{12}i_{21}}{i_{22}}. \end{aligned}$$

The ideal information measures the amount of information available for estimating θ_1 in the ideal setting where no other parameters need be estimated. It makes sense, then, that the actual information is only a fraction of this. The redirection penalty measures the cost that occurs from needing to estimate θ_2 in addition to θ_1, the cost of redirecting some of the information to estimate θ_2. The relationship between the different elements of the Fisher information matrix can be complex. Increasing one will often change the

others in order to retain positive-definiteness of the matrix. These quantities are related to one another via

$$0 \leqslant i_{11} - \frac{i_{12}i_{21}}{i_{22}} \leqslant i_{11} \tag{3.6}$$

The first inequality holds because the Fisher information matrix is semi-definite and the second inequality holds because $i_{12} = i_{21}$. When i_{11} is small, the redirection penalty is also small. This means that the information for θ_1 that gets passed though θ_2 will never exceed the total information available for θ_1.

If the actual information is close to 0, there is a large amount of information being passed through θ_2. This occurs when $\varepsilon > i_{11}i_{22} - i_{12}i_{21} \geqslant 0$ for a small ε. Equality would occur if either $i_{11} = 0$ and $i_{12} = 0$ or if $i_{22} = 0$ and $i_{21} = 0$, that is, if the likelihood was not a function of either θ_1 or θ_2. Equality would also hold if the matrix was degenerate because the parameters were perfectly correlated to one another, i.e. $i_{12} = i_{21} = \sqrt{i_{11}i_{22}}$. That is, either of both parameters are unidentifiable.

We can also study the relative contribution of each form of information by defining the relative actual information and redirection penalty with

$$\mathcal{R}_{AI}(\theta_1) = \frac{I(\theta_1)}{I(\theta_1|\theta_2)} = \frac{i_{11} - i_{12}i_{21}/i_{22}}{i_{11}} = 1 - \frac{i_{12}i_{21}}{i_{11}i_{22}} \quad \text{and} \quad \mathcal{R}_{RP}(\theta_1) = \frac{\mathcal{G}(\theta_1|\theta_2)}{I(\theta_1|\theta_2)} = \frac{i_{12}i_{21}}{i_{11}i_{22}} \tag{3.7}$$

By construction $\mathcal{R}_{AI}(\theta_1) + \mathcal{R}_{RP}(\theta_1) = 1$ and $\mathcal{R}_{AI}(\theta_1) \in [0, 1]$ as long as the Fisher Information matrix is positive definite, i.e. $i_{11}i_{22} - i_{12}i_{21} > 0$. Practically speaking, the relative information measures do not apply the cases explored in Section 3.2.1, when the likelihood is unidentified since the Fisher Information matrix is not invertible in that case. However, the ideal and actual information as well as the redirection penalty can all be investigated even when $i_{11} = 0$. The interpretation of these quantities as asymptotic precision, however, no longer hold.

We can already gain some intuition for the measures of flow in Equation (3.7) by considering the behavior as a function of its component quantities. If we vary the marginal information i_{11}, keeping all else fixed, the relative penalty increases when i_{11} is small, a quality not directly evident in Equation (3.6). In other words, the greatest loss for estimating θ_1 occurs when we already have very little information for it. This also occurs when i_{22} is small, when there is not much information in the data marginally for θ_2, much of the information that would have been used to estimate θ_1 is redirected. The co-information i_{12}

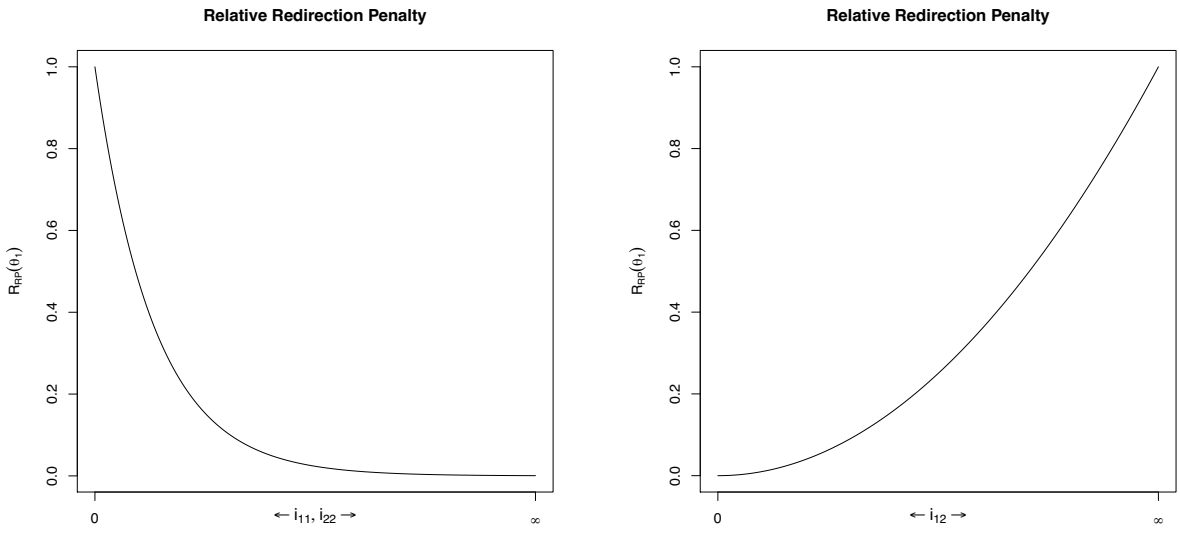


Figure 3.2: We explore the relationship between the relative redirection penalty $\mathcal{R}_{RP}$ studied with as a function of marginal information and co-information

has, of course, the opposite effect. When it is small the penalty incurred is small and when it is large the the penalty is equally large. Essentially when i_{12} is large, there is much more shared information between θ_1 and θ_2, the channel where information can potentially flow is wide.

3.3.1 Exploring a Frequentist Analysis of the Correlation Variation

In Section 3.1 we introduced two variations on a Normal distribution, we first consider the *correlation variation*. The Fisher Information matrix for ρ and σ^2 is

$$\mathcal{I}(\rho, \sigma^2) = \begin{pmatrix} i_{11} & i_{21} \\ i_{12} & i_{22} \end{pmatrix} = \begin{pmatrix} m\frac{1+\rho^2}{(1-\rho^2)^2} & -m\frac{\rho}{\sigma^2(1-\rho^2)} \\ -m\frac{\rho}{\sigma^2(1-\rho^2)} & \frac{m+k}{\sigma^4} \end{pmatrix}. \tag{3.8}$$

The number of marginal observations k is only affects the marginal Fisher information for σ^2 while the number of paired observations affects all of the information measures. From this we can calculate the marginal and conditional information for estimating ρ as described in Equation (3.4) to respectively be

$$I(\rho) = i_{11} - \frac{i_{21}^2}{i_{22}} = m\frac{m + k(1+\rho^2)}{(m+k)\,(1-\rho^2)^2} \quad \text{and} \quad I(\rho|\sigma^2) = i_{11} = m\frac{1+\rho^2}{(1-\rho^2)^2}. \tag{3.9}$$

Immediately, we have that if $m = 0$, we observe no paired observations, the data themselves contain no information for ρ. The conditional information $I(\rho|\sigma^2)$ grows with the number of pairs m observed and is unaffected by the number of marginal observations k. The information when σ^2 is unknown on the other hand is very much impacted by the number of marginal observations. To explore this further, consider the relative actual information and redirection penalty, respectively

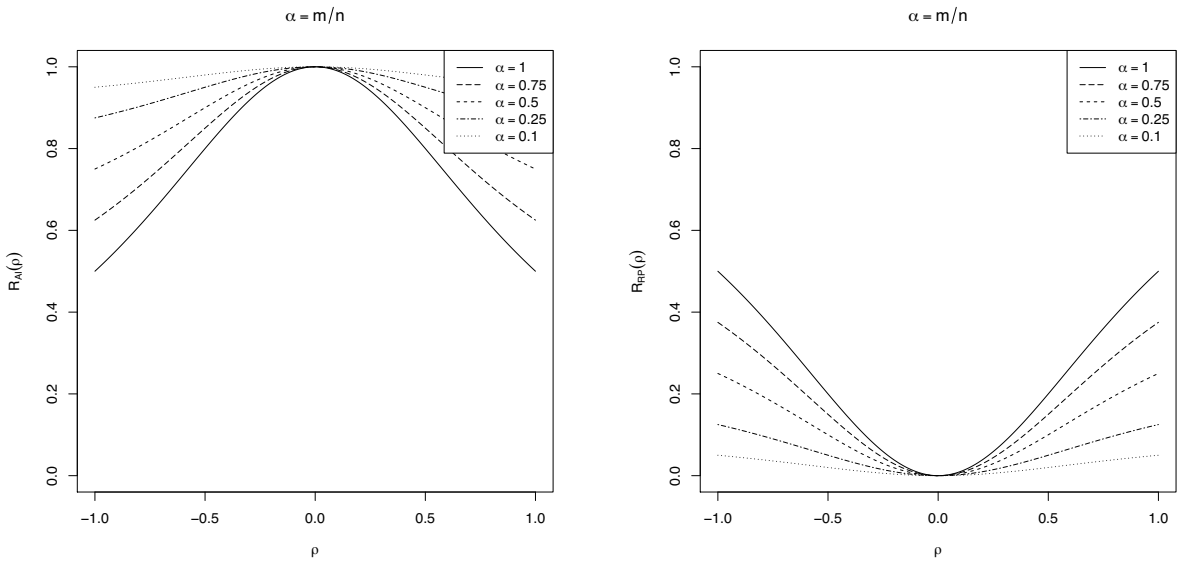
$$\mathcal{R}_{AI}(\rho) = \frac{1+\rho^2-\alpha\rho^2}{1+\rho^2} \quad \text{and} \quad \mathcal{R}_{RP}(\rho) = \frac{m}{m+k}\frac{\rho^2}{1+\rho^2} = \alpha\frac{\rho^2}{1+\rho^2} \tag{3.10}$$

It can be easily verified that in fact $\mathcal{R}_{AI}(\rho) + \mathcal{R}_{RP}(\rho) = 1$. While ρ and σ^2 are not globally orthogonal, they are locally orthogonal at $\rho = 0$ so $\mathcal{R}_{RP}(0) = 0$. The penalty also increases as ρ moves away from zero, indicating a stronger dependence between the parameters and more information lost for estimating ρ.

Marginal observations help estimate ρ. The redirection penalty for ρ in Equation 3.10 is a decreasing function of k. That is, as k increases the precision for estimating ρ increases. When $k = 0$, the redirection penalty is high, especially for larger values of $|\rho|$, this is because the m pairs are used to estimate both ρ and σ^2. Any additional k marginal observations reduce this penalty. Each observation for the marginal helps improve estimation of σ^2, and the paired observations can better estimate ρ. This is a because i_{22} increases with k while all else remains the same.

Because $\mathcal{R}_{RP}(\rho)$ measures information relative to the ideal information, interpreting increasing pairs is difficult. Indeed, with each new observed pair we can better estimate ρ whether or not we know σ^2 – both $I(\rho)$ and $I(\rho|\sigma^2)$ increase with m. However, $\mathcal{R}_{RP}(\rho)$ is also an increasing function of m. Each additional paired observation provides an opportunity for σ^2 to skim some information from what would have been available for ρ alone. However, this relative loss of efficiency does asymptote. The difference in penalty when $m = 50$ and $m = 200$ are negligible.

To understand the trade-off between paired and marginal observations, consider instead a fixed $m + k$ and we vary $\alpha = m/(m+k)$. Figure 3.3 plots Equations 3.10 for several values of α. Recall that the



(a) Relative actual information for ρ

(b) Relative redirection penalty for ρ

Figure 3.3: The relative actual information and redirection penalty for estimating ρ in the correlation variation example for differing fractions of paired observations $\alpha = 0.1, 0.25, 0.5, 0.75, 1$ and for a range of underlying correlation $\rho \in (-1, 1)$.

quantities are undefined when $\alpha = 0$, so we consider $\alpha = 0.1, 0.25, 0.5, 0.75, 1$. When the true underlying correlation $\rho \approx 0$, $\mathcal{R}_{AI}(\rho) \approx 1$ while $\mathcal{R}_{RP}(\rho) \approx 0$ regardless of α. This is because the parameters are locally orthogonal at $\rho = 0$. On the other hand, when $|\rho|$ is near 1, the parameters are far from orthogonal and $\mathcal{R}_{AI}(\rho)$ ranges between 0.5 and 0.95, depending on α. This indicates greater losses in efficiency for larger values of ρ, a fact we saw in Figure 3.2 as well.

Figure 3.3(b) shows the redirection penalty, decreasing the fraction of paired observations leads to smaller penalties. The paired observations are the only observations that can directly inform the estimate of ρ while the marginal observations can only indirectly inform ρ through σ^2. Eventually when we have no bivariate observations, or very few, relative to marginal observations, $\alpha \approx 0$, the penalty decreases. This reduction in penalty occurs because there is not much information to be had by either parameter. The extreme case when $\alpha = 0$ is the case of an unidentifiable likelihood with unidentifiable parameter ρ.

Now that we have built some intuition to how these measures are affected by the fraction of informative data, through α, we move to an example where the impact of the quality of the data might be more easily detectable – the mean variation of our two cases.

3.3.2 Exploring a Frequentist Analysis of the Mean Variation

In the mean variation, we are interested in conducting inference for the related means of two normally distributed random variables. One parameter θ_2 is directly informed by the data $\{Y_j\}_{j=1}^{n_Y}$. This data is of questionable quality controlled by the known variance γ. If γ is large we expect poor estimates of θ_2, especially when n_Y is small. The other parameter θ_1, on the other hand, is informed directly by the data $\{X_i\}_{i=1}^{n_X}$, but also indirectly by $\{Y_j\}_{j=1}^{n_Y}$. Figure 3.4 (left) contains a diagram of the flow of information for this example. The data $\{X_i\}_{i=1}^{n_X}$ informs both parameters while the data $\{Y_j\}_{j=1}^{n_Y}$ informs θ_2 directly and θ_1 indirectly through θ_2. In addition to helping us understand how information of questionable quality flows models, we study this example to explore a topic introduced in Section 3.2.2, parameter orthogonality. Understanding (and finding) orthogonal parameterization is simpler with location parameters than with scale parameters.

The Fisher Information matrix for the $n_X + n_Y$ observations $\{X_i\}_{i=1}^{n_X}, \{Y_j\}_{j=1}^{n_Y}$ is

$$I(\theta_1, \theta_2) = \begin{pmatrix} i_{11} & i_{21} \\ i_{12} & i_{22} \end{pmatrix} = \begin{pmatrix} n_X & n_X \\ n_X & n_X + \frac{n_Y}{\gamma} \end{pmatrix} \tag{3.11}$$

and has determinant $\det(I(\theta_1, \theta_2)) = \frac{n_X n_Y}{\gamma}$. The dependence on γ is clear, if γ is large relative to n_Y or n_X, the matrix will be nearly singular. Furthermore, as is typical, the identifiability of the likelihood is dependent on the number of observations and the quality of the data. If either of n_X, n_Y is zero, or $\gamma = \infty$, we can not possibly learn about both parameters and this matrix is non-invertible. Yet, for any finite γ, even extremely large, and any $n_X, n_y > 0$, even $n_X = n_y = 1$, this matrix will be invertible and the likelihood is identifiable. Because of the Normality of our observations, this information matrix describes the precision of the MLE estimates of θ_1 and θ_2 exactly, and not just asymptotically.

The information measures described in Equation (3.4) are, for this example,

$$I(\theta_1) = i_{11} - \frac{i_{21}^2}{i_{22}} = n_X - \frac{n_X^2}{n_X + n_Y/\gamma} = \frac{n_X n_Y/\gamma}{n_X + n_Y/\gamma} \quad \text{and} \quad I(\theta_1|\theta_2) = i_{11} = n_X. \tag{3.12}$$

When θ_2 is known, the variance of the MLE for estimating θ_1 will be $1/n_X$. In other words, knowing θ_2

implies that $X_i - \theta_2 \sim N(\theta_1, 1)$ and we can use this quantity to estimate θ_1 directly. The effective sample size for estimating θ_1 is n_X if θ_2 is known. However, when examining $I(\theta_1)$, the cost of needing to estimate θ_2 becomes clear – the information accumulates more slowly. The variance of the MLE of θ_1 in this case will be $\frac{n_X+n_Y/\gamma}{n_X n_Y/\gamma} = 1/n_X + \gamma/n_Y$. The effective sample size for estimating θ_1 is less than n_X. The loss of efficiency is γ/n_Y, rather large when γ is large or when n_Y is small.

The relative actual information and redirection penalties for estimating θ_1 are

$$\mathcal{R}_{AI}(\theta_1) = \frac{I(\theta_1)}{I(\theta_1|\theta_2)} = \frac{n_Y/\gamma}{n_X + n_Y/\gamma} \quad \text{and} \quad \mathcal{R}_{RP}(\theta_1) = \frac{n_X}{n_X + n_Y/\gamma} \tag{3.13}$$

When the quality of the data for estimating θ_2 increases, i.e. as γ approaches 0, the redirection penalty also approaches zero. This is because γ influences these quantities only through i_{22} and $i_{22} \to \infty$ as $\gamma \to 0$. The information for estimating θ_1 need not be redirected though θ_2 since θ_2 is being adequately estimated by the high-quality data $\{Y_j\}_{j=1}^{n_Y}$. On the other hand, for low quality data (large γ), the redirection penalty is close to 1, reaching 1 only as $\gamma \to \infty$. Care must be taken with this limit because the Fisher Information matrix in Equation (3.11) will be singular. So, instead of taking limits of γ directly, consider the limit of worsening data quality but more of it, i.e. $n_Y, \gamma \to \infty$ with $n_Y/\gamma \to c$. Here, c represents a re-scaled data quality for $\{Y_j\}$, making c and n_X directly comparable. We can inspect the quantities in Equations (3.11) and (3.13) assuredly since the matrix will be positive definite, even in the limit: $\det(I(\theta_1, \theta_2)) \to n_X c$. If $c = n_X$, the quantity and and quantity of $\{Y_j\}$ is equivalent to $\{X_i\}$ and is worse if $c < n_X$.

The quality of $\{Y_j\}$ plays a major role in estimation of θ_1. The loss of precision in estimating θ_1 that comes from needing to estimate θ_2 is $\frac{n_X^2}{n_X+n_Y/\gamma}$. Taking this loss of precision to the extreme, we have

$$\frac{n_X^2}{n_X + n_Y/\gamma} = \frac{n_X}{1 + n_Y/(n_X\gamma)} \overset{n_Y/\gamma \to c}{\longrightarrow} \frac{n_X}{1 + c/n_X}.$$

If the quality of data is equal for both sets $\{Y_j\}$ and $\{X_i\}$, i.e. $c = n_X$, the loss of precision will still be $n_X/2$. This is also reflected in the redirection penalty: in this case $\mathcal{R}_{RP}(\theta_1) = 1/2$.

Thinking of γ as an index for data quality sheds light on the link between data quality and sample size. The worse our data, the more samples we will need to achieve the same precision. Consider, for simplicity, the situation where $n_X = n_Y = n$. Say we observe $n = n_1$ samples from the model with data quality

$\gamma = \gamma_1$, the variances of the MLEs of (θ_1, θ_2) will be $((1+\gamma_1)/n_1, \gamma_1/n_1)$. Now if we have instead a situation with reduced data quality, $\gamma_2 > \gamma_1$, we would need

$$n_2 = \frac{1+\gamma_2}{1+\gamma_1} n_1$$

samples to achieve the same level of accuracy for estimating θ_1. Under these conditions the multiplier of n_1 is strictly greater than 1. The worse the data quality is for estimating θ_2, the more samples you will need in order to precisely estimate θ_1 because these parameters are dependent.

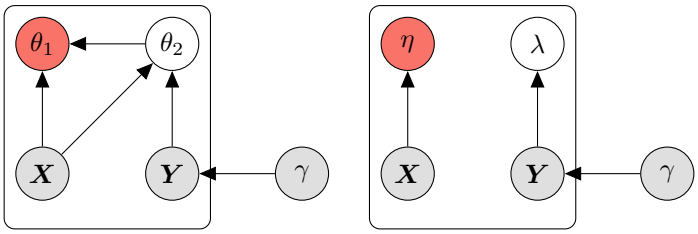


Figure 3.4: The flow of information from the data to parameters under the two parameterizations discussed: (left) original parameterization where we're interested in estimating (θ_1, θ_2) and (right) the orthogonal parameterization $(\eta = \theta_1 + \theta_2, \lambda = \theta_2)$

Orthogonalization and the Mean Variation

We can see that the fate of the quality of estimates for θ_1 and θ_2 are entangled. The information available for estimating θ_2 will directly impact the estimates of θ_1. In recognizing this, it may help to change our perspective on the model. In some sense, if the value of γ is large, we should give up on estimating θ_2. The data relevant for estimating θ_2 is of poor quality. In the extreme case, when $\gamma = \infty$, the likelihood is unidentifiable. The prevailing dogma for dealing with unidentifiable models is that one should not. However, in this example, and others, some parameters can still be estimated with precision. In particular $\eta = \theta_1 + \theta_2$ could be precisely estimated (based on intuition) and estimation this quantity should not be tainted by the quality of the data $\{Y_j\}_{j=1}^{n_Y}$.

Consider the parameter transformation $(\eta, \lambda) = (\theta_1 + \theta_2, \theta_2)$. In this case we can calculate the Fisher

Information matrix as

$$\mathcal{I}(\eta,\lambda)=\begin{pmatrix}1 & -1\\ 0 & 1\end{pmatrix}^T\begin{pmatrix}n_X & n_X\\ n_X & n_X+\frac{n_Y}{\gamma}\end{pmatrix}\begin{pmatrix}1 & -1\\ 0 & 1\end{pmatrix}=\begin{pmatrix}n_X & 0\\ 0 & n_Y/\gamma\end{pmatrix}.$$

Because the parameters are orthogonal,

$$I(\eta)=i_{11}-\frac{i_{21}^2}{i_{22}}=n_X \quad \text{and} \quad I(\eta|\lambda)=i_{11}=n_X. \tag{3.14}$$

So the relative actual information and redirection penalty are

$$\mathcal{R}_{AI}(\eta)=\frac{I(\eta|\lambda)}{I(\eta)}=1 \quad \text{and} \quad \mathcal{R}_{RP}(\eta)=0 \tag{3.15}$$

The relative actual information does not depend on γ nor n_Y as did $\mathcal{R}_{AI}(\theta_1)$ in Equation (3.13). The redirection penalty is zero; the estimation of η is no longer affected by the estimation of λ, or the quality of the data that estimates it. The inference for η is untainted by poor-quality information from the observations $\{Y_i\}$ – essentially, by removing the dependence on the nuisance parameter λ. In other words, because the parameters are orthogonal the flow of poor-quality information Y is blocked from affecting other parameters besides λ. Figure 3.4 (right) shows that this parameterization not only removed the flow of information from γ to the primary parameter η, it also removed the link between the data X and the secondary parameter λ.

3.4 Measures of Information Flow in Bayesian Analysis

In Section 3.3, we use orthogonality as the cornerstone for our measures of information flow. If two parameters were orthogonal, the redirection penalty is zero, and the information in the ideal case and actual case were the same. Likewise, in this section, we use non-identifiability as the cornerstone for our measures. We use Bayes' Theorem as a framework for information transfer from data to model parameters. We show that there is no direct information from the data to non-identifiable parameters. We then use our two examples, where we can move our models toward and away from the non-identifiable case, to study

nearly non-identifiable parameters. We also inspect how prior structure influences non-identifiable (or nearly non-identifiable) parameters.

A recurring theme in the measures of information we consider in Section 3.3 is that they are not always defined. Because we are using the Fisher Information matrix as the basis for understanding the flow of information, we need this matrix to be defined. We also need it to be invertible if we want to interpret the measures, like the relative redirection penalty, in terms of precision loss. However, a non-invertible Fisher Information matrix gives a clear indication that the likelihood is not identifiable, implying no direct information for some parameter. The measures of information considered in Section 3.3 do not reflect this; in this section, we explore measures that do.

To discuss measures of information flow in the Bayesian setting, we first set some notation. Consider a *scientific parametrization* $\theta \in \Theta$ and an invertible reparametrization $(\eta, \lambda) = g(\theta)$. The parameter η is the primary parameter and λ is the secondary parameter. We also consider a proper prior distribution $p(\theta)$ with support Θ. We consider proper prior distributions because we want to study unidentifiable likelihoods and posterior inference is possible in this case, but only with proper priors. The transformation g induces a prior distribution $p(\eta, \lambda)$, which we assume is also proper, with induced support $g(\Theta)$.

Care must be taken when considering the transformed parameter space. The induced support and prior distribution may have a complex structure, even if Θ and $p(\theta)$ do not. For this reason, we denote the support of η and λ respectively

$$N = \{\eta' : (\eta', \lambda) = g(\Theta) \text{ for some } \lambda\} \quad \text{and} \quad \Lambda = \{\lambda' : (\eta, \lambda') = g(\Theta) \text{ for some } \eta\}.$$

We also denote the conditional support for η given λ and λ given η respectively

$$N(\lambda) = \{\eta' : (\eta', \lambda) = g(\Theta)\} \quad \text{and} \quad \Lambda(\eta) = \{\lambda' : (\eta, \lambda') = g(\Theta)\}.$$

We further assume that the induced prior distributions are well-defined and proper on these spaces, e.g. $p(\lambda|\eta) > 0$ for $\lambda \in \Lambda(\eta)$ and $\int_{\Lambda(\eta)} p(\lambda|\eta) d\lambda < \infty$.

Recalling the discussion in Section 3.2.1, when a model is partially identifiable, the likelihood can be written as a function of the identifiable parameter, say η, alone: $p(\mathcal{D}_n|\theta) = p(\mathcal{D}_n|\eta)$. Even in this case,

Theorem 3.2.1 ensures that the posterior distribution for (η, λ) is typically defined. This posterior distribution, however, takes on a particular form when the likelihood is unidentifiable. It can be factores as

$$p(\lambda, \eta \mid \mathcal{D}_n) = p(\lambda \mid \eta, \mathcal{D}_n)p(\eta \mid \mathcal{D}_n) = p(\lambda \mid \eta)p(\eta \mid \mathcal{D}_n) \tag{3.16}$$

In other words, the posterior distribution for η is affected by the data and may even desirable properties like posterior convergence. The posterior conditional for λ, on the other hand is fully determined the prior distribution. It is not updated by the data. If the prior distribution for η and λ are also independent, i.e. $p(\lambda|\eta) = p(\lambda)$, then

$$p(\lambda \mid \eta, \mathcal{D}_n) = p(\lambda \mid \mathcal{D}_n) = p(\lambda \mid \eta) = p(\lambda). \tag{3.17}$$

There is no information transferred to λ. Neither from the data nor the identifiable parameter η.

The parameter transformations will typically create dependencies in prior distributions, so the prior for η and λ can be complex. Nevertheless, we can still understand the marginal posterior of the unidentifiable parameter as a weighted average of the conditional prior.

Theorem 3.4.1. *(Proposition 2 in Poirier (1998)) Let $\{p(\mathcal{D}_n|\theta), \theta \in \Theta\}$ be a family of distributions for the data $\mathcal{D}_n \in \mathscr{D}_n$ indexed by θ. Let $(\eta, \lambda) = g(\theta)$ be an transparent parameterization with identifiable parameter η and unidentifiable parameter λ. Let η_0 denote the parameter that generates $\mathcal{D}_n$. Denote the proper prior distribution $p(\theta)$ and induced prior $p(\eta, \lambda) = p(\lambda|\eta)p(\eta)$. Then*

$$p(\lambda \mid \mathcal{D}_n) = \mathbb{E}_{\eta|\mathcal{D}_n}\left[p(\lambda|\eta)\right] \overset{n\to\infty}{\longrightarrow} p(\lambda|\eta_0).$$

This follows immediately from Equation 3.16 and the assumption of a "well-behaved" posterior $p(\eta|\mathcal{D}_n)$. Even as we gather an infinite amount of information, prior dependencies strongly influence the inference for λ. Intuitively then, in this extreme situation, the information in the data can not inform λ directly, only indirectly through prior relationships. However, what happens when we move away from this extreme.

Consider the general conditional posterior for $\lambda \in \Lambda(\eta)$, so that $p(\lambda \mid \eta) > 0$, we can invert the typical

Bayesian formulation as

$$p(\lambda \mid \eta, \mathcal{D}_n) = \frac{p(\mathcal{D}_n \mid \lambda, \eta) p(\lambda \mid \eta)}{p(\mathcal{D}_n \mid \eta)}.$$
$$\frac{p(\lambda \mid \eta, \mathcal{D}_n)}{p(\lambda \mid \eta)} = \frac{p(\mathcal{D}_n \mid \lambda, \eta)}{p(\mathcal{D}_n \mid \eta)}. \tag{3.18}$$

Clearly, if the likelihood is unidentifiable, then the quantity on the right hand side of Equation (3.18) will be one, implying that the conditional posterior for λ will be the same as the prior. Equally, however, if the quantity on the left hand side of Equation (3.18) is one, this would imply the quantity on the right hand side would be one. If this prior-posterior comparison shows no change then it may be possible to rewrite the likelihood as a function of η alone. This simple observation gives rise to another way of defining PIMs.

A model $p(\mathcal{D}_n|\theta)$ is partially identifiable with identifiable parameter η and unidentifiable parameter λ if

$$KL\left(p(\lambda \mid \eta, \mathcal{D}_n) || p(\lambda \mid \eta)\right) = 0, \text{ for } \eta \in N \text{ and a.e. } \mathcal{D}_n \in \mathscr{D}_n. \tag{3.19}$$

That is, if the relationship in Equation 3.18 holds for every η and almost every dataset $\mathcal{D}_n$ then the model is partially identifiable. Comparing these distributions, either through KL-divergence or visually, will give us a measure direct information flow. If $p(\lambda \mid \eta, \mathcal{D}_n)$ and $p(\lambda \mid \eta)$ are the same, the data are not informing λ, and there is no direct flow of information.

$$\text{Direct Information for } \lambda\text{: } KL\left(p(\lambda \mid \eta, \mathcal{D}_n) || p(\lambda \mid \eta)\right) \tag{3.20}$$

If these distributions are drastically different, then the data is informing the parameter beyond the prior structure, and it is receiving information directly from the data. The difference between marginal posterior and conditional posterior distribution provide a measure of the indirect flow of information among parameters.

$$\text{Indirect Information for } \lambda\text{: } KL\left(p(\lambda \mid \eta, \mathcal{D}_n) || p(\lambda \mid \mathcal{D}_n)\right) \tag{3.21}$$

If these distributions are drastically different, this would indicate a strong dependence between the parameters and hence much shared information. The difference between the marginal and conditional posterior speaks to the indirect information provided by, or the conditional independence of, the conditioning parameter. Indeed, $p(\lambda|\mathcal{D}_n) = p(\lambda|\eta, \mathcal{D}_n)$ is the definition of conditionally independent.

When λ is unidentifiable, the direct information measure is zero for all values of η; otherwise, it is a function of η. Seeing these as a function of η provides some sense of the relative information provided by the parameter, either through the prior in the direct information or through the posterior in the indirect information. It also makes sense to consider these at a point estimate of η. For example, we can evaluate them at the maximum a-posteriori estimate, η_{MAP}, of η. Doing this adds to our interpretation of the KL measures. The direct information measures the additional information the data provide for λ beyond that which other parameters provide. Conversely, the indirect information measures the additional information η provides for λ beyond that which is provided by the data. Similarly, in simulations we can evaluate them at the true data-generating parameters. Compared to evaluating them at MAP estimates, this provides a more concrete definition of information flow. It separates the information provided by conditioning on the parameter from the information provided by the data. These interpretations are useful, but divergence measures are arguably flimsy.

The KL-divergence is a one-number summary. It is necessarily an oversimplification that does not capture all meaningful differences. When comparing prior and posterior densities, however, first and second order differences are relevant, and the KL divergence does capture these. Location shifts indicate poor initial first guesses and decreases in scale indicate precise inference. Take, for example, the KL-divergence between two univariate Normal distributions $\mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$ and $\mathcal{N}(\mu_\star, \sigma_\star^2)$; it is

$$\mathcal{KL}\left(\mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)||\mathcal{N}(\mu_\star, \sigma_\star^2)\right) = -\frac{1}{2}\log\frac{\tilde{\sigma}^2}{\sigma_\star^2} + \frac{\tilde{\sigma}^2 + (\tilde{\mu} - \mu_\star)^2}{2\sigma_\star^2} - \frac{1}{2}. \tag{3.22}$$

Assuming no location shifts and setting $\tilde{\sigma}^2 = \nu\sigma_\star^2$ the KL-divergence simplifies to $0.5(\nu - \log\nu - 1)$; positive for $\nu \neq 1$. Values of $\nu < 1$ indicate posterior contraction, in fact a posterior contraction of half ($\nu = 0.5$) corresponds to a KL-divergence of 0.097. A KL-divergence would be 1 corresponds to a variance reduction of about 95% ($\nu = 0.05$). On the other hand, assuming no posterior contraction ($\tilde{\sigma} = \sigma_\star$)

and denoting $\delta = (\tilde{\mu} - \mu_\star)/\tilde{\sigma}$ the KL-divergence would be $0.5\delta^2$. So, if the posterior mean is 1 standard deviation away from the prior mean, the KL-divergence would be 0.5. Again, a KL-divergence of 1 would correspond to a location shift of about 1.4 standard deviations. Therein lies one difficulty of using a one-dimensional summary to compare distributions. There is no one-to-one map between distributional differences and KL-measures. Still, these measures are useful in giving some clue as to the changes from prior to posterior. We recommend using visual inspection along with these measures to gain a clear understanding of the changes.

This exercise may seem frivolous, but the Normal distribution is as distinctive in Bayesian statistics as it is in classical inference. Indeed, the Bernstein-von Mises Theorem ensures that posterior distributions typically behave like a Normal distribution asymptotically (van der Vaart, 1998). Furthermore, the Normal distribution is a popular choice for approximating posteriors with Variational Inference methods (Blei et al., 2017). Having established the connection between prior-posterior comparisons and information flow, we now explore a Bayesian analysis of the mean variation and correlation variation.

3.4.1 Exploring a Bayesian Analysis of the Mean Variation

One feature of prior-posterior comparisons, like the one seen Equations 3.21 and 3.21, is that the prior structure matters. In particular, prior dependence dictates the appropriate comparison. For the mean variation example, consider the following prior distribution on the transformed parameters space $(\eta, \lambda) = (\theta_1 + \theta_2, \theta_2)$

$$\begin{pmatrix} \eta \\ \lambda \end{pmatrix} \sim \mathcal{N}\left(\mu_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \; \Sigma_0 = \begin{pmatrix} \sigma_1^2 & \phi\sigma_1\sigma_2 \\ \phi\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right).$$

Take (η_0, λ_0) as underlying data generating values. If $\phi = 0$, the parameters are a-priori independent. Otherwise, the parameters will be dependent. We will be primarily interested in how γ changes the amount of information η and λ receive through the Bayesian updating process. We will also be interested in how ϕ changes different prior-posterior comparisons. We will compare the prior and posterior distributions for η,

λ and $\lambda|\eta$. These prior distributions are

$$\eta \sim \mathcal{N}(0, \sigma_1^2), \quad \lambda \sim \mathcal{N}(0, \sigma_2^2) \quad \text{and} \quad \lambda|\eta \sim \mathcal{N}\left(\phi\frac{\sigma_2}{\sigma_2}\eta, (1-\phi^2)\sigma_2^2\right),$$

which we use as the basis of comparison.

If we take equal observations of each data type ($n_Y = n_X = n$), we can further simplify our likelihood to the likelihood of the sufficient statistics, $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ and $\bar{Y}_n = n^{-1}\sum_{j=1}^n Y_j$. The likelihood can be simplified to

$$\begin{pmatrix} \bar{X}_n \\ \bar{Y}_n \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \eta \\ \lambda \end{pmatrix}, \ \Sigma = \frac{1}{n}\begin{pmatrix} 1 & 0 \\ 0 & \gamma \end{pmatrix}\right).$$

Because the covarinace matrix in the likelihood is treated as known, we have that the posterior distribution for (η, λ) is also bivariate Normal. We can calculate the mean vector and variance-covariance matrix of the bivariate normal distribution.

$$\begin{pmatrix} \eta \\ \lambda \end{pmatrix} \Big| \mathcal{D}_n \sim \mathcal{N}\left(\mu^\star, \Sigma^\star\right). \tag{3.23}$$

Denoting the determinant of the Σ_0 to be $|\Sigma_0| = \sigma_1^2\sigma_2^2(1-\phi^2)$ and letting $c = c(\phi, \sigma_1^2, \sigma_2^2, \gamma, n) = \left(n|\Sigma_0| + \sigma_2^2\right)n/\gamma + \left(n\sigma_1^2 + 1\right)$, the posterior mean is

$$\mu^\star = \begin{pmatrix} \mu_\eta^\star \\ \mu_\lambda^\star \end{pmatrix} = \frac{1}{c}\begin{pmatrix} n\bar{X}_n(|\Sigma_0|n/\gamma + \sigma_1^2) + \frac{n}{\gamma}\bar{Y}_n\phi\sigma_1\sigma_2 \\ n\bar{X}_n\phi\sigma_1\sigma_2 + \frac{n}{\gamma}\bar{Y}_n(|\Sigma_0|n + \sigma_2^2) \end{pmatrix}. \tag{3.24}$$

The posterior variance-covariance matrix is similarly

$$\Sigma^\star = \begin{pmatrix} \sigma_\eta^{\star 2} & \sigma_{\eta\lambda}^\star \\ \sigma_{\eta\lambda}^\star & \sigma_\lambda^{\star 2} \end{pmatrix} = \frac{1}{c}\begin{pmatrix} |\Sigma_0|n/\gamma + \sigma_1^2 & \phi\sigma_1\sigma_2 \\ \phi\sigma_1\sigma_2 & |\Sigma_0|n + \sigma_2^2 \end{pmatrix} \tag{3.25}$$

The prior correlation ϕ plays an important role in the posterior distribution. As evident in Equation (3.24), the posterior mean is a function of both $\bar{X}_n$ and $\bar{Y}_n$. However, the contribution of $\bar{Y}_n$ for estimating the posterior mean of η is weighted by $\frac{n\phi\sigma_1\sigma_2}{c\gamma}$. So if ϕ is small, the contribution of $\bar{Y}_n$ to estimating the

mean of η is equally small. Likewise, the contribution of $\bar{X}_n$ to estimating λ is controlled by ϕ. In fact, if $\phi = 0$, the posterior distribution simplifies

$$\begin{pmatrix} \eta \\ \lambda \end{pmatrix} \Big| \mathcal{D}_n \sim \mathcal{N}\left(\begin{pmatrix} \frac{n\sigma_1^2}{n\sigma_1^2+1}\bar{X}_n \\ \frac{n\sigma_2^2/\gamma}{n\sigma_2^2/\gamma+1}\bar{Y}_n \end{pmatrix}, \begin{pmatrix} \frac{\sigma_1^2}{n\sigma_1^2+1} & 0 \\ 0 & \frac{\sigma_2^2}{n\sigma_2^2/\gamma+1} \end{pmatrix} \right). \tag{3.26}$$

In Section 3.3 we discussed how (η, λ) are orthogonalized versions of (θ_1, θ_2). In this example, the posterior correlation is fully driven by prior correlations since the parameters are uncorrelated in the likelihood. Equation (3.26) contains hints at prior posterior comparisons as γ gets large. Recall that γ controls the informativeness of $\bar{Y}_n$ for estimating λ. If $\gamma = \infty$, our model is a PIM with identifiable parameter η and unidentifiable parameter λ. With independent priors, $\phi = 0$, we expect the posterior marginal distribution $p(\lambda|\bar{X}_n, \bar{Y}_n)$ and prior marginal distribution $p(\lambda)$ to be the same. Indeed, setting $\gamma = \infty$ yields that the posterior marginal distribution

$$\lambda|\bar{X}_n, \bar{Y}_n \sim \mathcal{N}(0, \sigma_2^2)$$

is equal to the prior marginal distribution. We can, however, learn more about how the prior structure plays a role in this prior-posterior comparison by inspecting this comparison for varying levels of prior dependence.

If we allow the data quality index to increase so that we get poorer and poorer data quality $\gamma \to \infty$, the posterior mean vector and variance-covariance matrix converge to

$$\mu^\star \xrightarrow{\gamma\to\infty} \begin{pmatrix} \frac{n\sigma_1^2}{n\sigma_1^2+1}\bar{X}_n \\ \frac{n\phi\sigma_1\sigma_2}{n\sigma_1^2+1}\bar{X}_n \end{pmatrix} \quad \text{and} \quad \Sigma^\star \xrightarrow{\gamma\to\infty} \frac{1}{n\sigma_1^2+1} \begin{pmatrix} \sigma_1^2 & \phi\sigma_1\sigma_2 \\ \phi\sigma_1\sigma_2 & |\Sigma_0|n + \sigma_2^2 \end{pmatrix}. \tag{3.27}$$

This is the posterior distribution when the likelihood is unidentifiable, nevertheless, as stated in Theorem 3.2.1, this distribution is defined and proper. While this posterior may be defined, the tension discussed in Section 3.2.1 begins to show itself. It is tempting to boast that the Bayesian framework allows for the estimation of parameters even when it is unidentifiable. It is important to note, however, that the marginal

posterior for λ when $\gamma = \infty$ is

$$\lambda|\mathcal{D}_n \sim \mathcal{N}\left(\phi\frac{n\sigma_1\sigma_2}{n\sigma_1^2+1}\bar{X}_n, \sigma_2^2\frac{(n\sigma_1^2(1-\phi^2)+1)}{n\sigma_1^2+1}\right).$$

The prior correlation, ϕ, influences how much information transfers between the two parameters. For example, if we select $\phi = \sigma_1/\sigma_2$, the posterior means of λ and η will be the same. This quantity also controls the amount of "shrinkage" when comparing the prior variance to the posterior variance for λ. The posterior variance will always be smaller than the prior variance

$$\sigma_2^2\frac{(n\sigma_1^2(1-\phi^2)+1)}{n\sigma_1^2+1} \leqslant \sigma_2^2.$$

Equality occurs when we have independent priors. Indeed, simply comparing marginal prior and posterior variances to detect the amount of learning can mislead us when the priors are dependent.

The accumulation of information for η occurs as typical, while it is much slower for λ. The posterior marginal variance of η is $\frac{\sigma_1^2}{n\sigma_1^2+1}$ and vanishes regardless of ϕ, as n grows to ∞. The accumulation of information for λ, on the other hand, is slower. The variance of the posterior converges but does not vanish, and it instead converges to $\sigma_2^2(1-\phi^2)$. Similarly, the mean converges in expectation to $\eta_0\phi\sigma_2/\sigma_1$, a scaled version of the generating parameters η_0. These quantities look familiar; they are the mean and variance of the conditional distribution of $\lambda|\eta$ evaluated at η_0. As stated in Theorem 3.4.1, the limiting posterior distribution of the unidentifiable parameter λ is the prior distribution evaluated at the true value of the identifiable parameter. To further study the impact of prior-dependence, data quality and sample size on these prior-posterior comparisons we can inspect a global measure of similarity, the KL-divergence.

We can compare the marginal prior and posterior distributions for λ, as

$$\mathcal{KL}\left(p(\lambda|\mathcal{D}_n)||p(\lambda)\right) = -\frac{1}{2}\log A_\lambda + \frac{A_\lambda}{2} + \frac{B_\lambda}{2} - \frac{1}{2}. \tag{3.28}$$

where

$$A_\lambda = \frac{\sigma_\lambda^{\star 2}}{\sigma_2^2} = \frac{|\Sigma_0|n + \sigma_2^2}{c\sigma_2^2} = \frac{n\sigma_1^2(1-\phi^2)+1}{n^2|\Sigma_0|/\gamma + n\sigma_2^2/\gamma + n\sigma_1^2 + 1} \tag{3.29}$$

$$B_\lambda = \frac{\mu_\lambda^{\star 2}}{\sigma_2^2} = \frac{1}{c^2\sigma_2^2}\left(n\bar{X}_n\phi\sigma_1\sigma_2 + \frac{n}{\gamma}\bar{Y}_n(|\Sigma_0|n + \sigma_2^2)\right)^2 \tag{3.30}$$

If this quantity is zero, the two distributions are the same for almost every λ. In this case, that only occurs only if both $A_\lambda = 1$ and $B_\lambda = 0$. We know from Equation 3.19 that when $\gamma = \infty$, this comparison will be zero under independent priors ($\phi = 0$). Surely enough, $A_\lambda = 1$ and $B_\lambda = 0$ only when both $\gamma = \infty$ and $\phi = 0$. There is no direct information from the data for estimating λ, but because $\phi = 0$ there is also no indirect information being passed from η.

The KL-divergence in Equation (3.28) measures the total information the data contains for λ both directly and indirectly for any data quality measure γ and prior correlation ϕ. We can take advantage of that by inspecting these separately. The KL divergence measure in Equation (3.28) in the case where $\gamma \to \infty$ provides a measure of indirect information for λ provided by η. In this case $A_\lambda \to \frac{n\sigma_1^2(1-\phi^2)+1}{n\sigma_1^2+1}$ and $B_\lambda \to \frac{n^2\phi^2\sigma_1^2}{(n\sigma_1^2+1)^2}\bar{X}_n^2$. The KL divergence is then

$$\mathcal{KL}\left(p(\lambda|\mathcal{D}_n)||p(\lambda)\right) = -\frac{1}{2}\log\frac{n\sigma_1^2(1-\phi^2)+1}{n\sigma_1^2+1} + \frac{n\sigma_1^2(1-\phi^2)+1}{2(n\sigma_1^2+1)} + \frac{n^2\phi^2\sigma_1^2}{2(n\sigma_1^2+1)^2}\bar{X}_n^2 - \frac{1}{2}. \tag{3.31}$$

Because λ is unidentifiable, the change in marginal prior to marginal posterior is fully determined by the prior structure. In fact, this quantity is driven by the prior correlation ϕ and the prior variance for η. This quantity is also random and an increasing function of $\bar{X}_n^2$, so samples with extreme means will yield large KL-divergences.

If we take expectations over the data distribution $\mathcal{D}_n|\eta,\lambda$, the KL-divergence above simplifies to

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n|\eta,\lambda}\left[\mathcal{KL}\left(p(\lambda|\mathcal{D}_n)||p(\lambda)\right)\right] &= -\frac{1}{2}\log\frac{n\sigma_1^2(1-\phi^2)+1}{n\sigma_1^2+1} + \frac{n\sigma_1^2(1-\phi^2)+1}{2(n\sigma_1^2+1)} + \frac{n^2\phi^2\sigma_1^2(\frac{1}{n}+\eta^2)}{2(n\sigma_1^2+1)^2} - \frac{1}{2} \\ &= -\frac{1}{2}\log\frac{\sigma_1^2(1-\phi^2)+1/n}{\sigma_1^2+1/n} + \frac{n^2\phi^2\sigma_1^2(\eta^2-\sigma_1^2)}{2(n\sigma_1^2+1)^2}\end{aligned}$$

This quantity is zero when $\phi = 0$ for every η. The posterior inference for λ will not change (from

the prior), even if η is learned perfectly from the data. That is, the posterior marginal distribution for this unidentifiable parameter will be different from the prior, but the prior structure partially determines that difference.

Even when accumulating an infinite number of observations, the marginal posterior for λ will retain uncertainty. Take $\sigma_1^2 = 1$ for simplicity. This quantity converges to $-\frac{1}{2}\log(1-\phi^2) + \phi^2\left(\eta^2-1\right)/2$ as $n \to \infty$. The prior dependence structure remains because the information is useless for estimating λ, even if infinite. It also shows that the prior correlation determines the amount of indirect information that λ receives from η. If the prior is moderately dependent, say $\phi = 0.5$ then the KL-divergence will be $0.14 + 0.25(\eta^2 - 1)$ for large n, an increasing function of η. Changing η will greatly impact the inference for λ since the data is not able to anchor it. Indeed, this is a perfect illustration of the need for checking sensitivity to the prior. Here, changing the prior mean of η will greatly influence the posterior inference of λ.

The above comparison of marginal prior and posterior distributions for λ displays a complex relationship between direct and indirect information. As discussed above, no direct information is passed to unidentifiable parameters if we measure direct information by comparing prior and posterior conditional distributions as in Equation (3.19). The conditional distribution for $\lambda|\eta, \mathcal{D}_n$ are available from standard manipulations

$$\lambda|\mathcal{D}_n, \eta \sim \mathcal{N}\left(\mu^\star_{\lambda|\eta}, {\sigma^\star_{\lambda|\eta}}^2\right)$$

where

$$\mu^\star_{\lambda|\eta} = \mu^\star_\lambda + \frac{\phi\sigma_1\sigma_2}{|\Sigma_0|n/\gamma + \sigma_1^2}\left(\eta - \mu^\star_\eta\right)$$
$${\sigma^\star_{\lambda|\eta}}^2 = \sigma^{\star 2}_\lambda - \frac{\phi^2\sigma_1^2\sigma_2^2}{c\left(|\Sigma_0|n/\gamma + \sigma_1^2\right)}$$

The KL-divergence between this conditional and the prior conditional $p(\lambda|\eta)$ is

$$\mathcal{KL}\left(p(\lambda|\mathcal{D}_n, \eta)||p(\lambda|\eta)\right) = -\frac{1}{2}\log A_{\lambda|\eta} + \frac{A_{\lambda|\eta}}{2} + \frac{B_{\lambda|\eta}}{2} - \frac{1}{2} \tag{3.32}$$

where

$$A_{\lambda|\eta} = \frac{1}{n\sigma_2^2(1-\phi^2)/\gamma + 1} \quad \text{and} \quad B_{\lambda|\eta} = \frac{1}{\sigma_2^2(1-\phi^2)} \left(\frac{|\Sigma_0|n/\gamma}{|\Sigma_0|n/\gamma + \sigma_1^2} \bar{Y}_n - \phi \frac{\sigma_2}{\sigma_1} \frac{|\Sigma_0|n/\gamma}{|\Sigma_0|n/\gamma + \sigma_1^2} \eta \right)^2.$$

The relationship between data quality and this prior-posterior comparison is much simpler than the marginal comparison in Equation (3.28). Taking $\gamma \to \infty$ we have that

$$\mu^\star_{\lambda|\eta} \to \eta \frac{\phi\sigma_2}{\sigma_1}, \quad \text{and} \quad {\sigma^\star_{\lambda|\eta}}^2 \to \sigma_2^2(1-\phi^2),$$

so that

$$\mathcal{KL}\left(p(\lambda|\mathcal{D}_n, \eta)||p(\lambda|\eta)\right) \to 0. \tag{3.33}$$

This measure of direct information λ behaves as expected, it indicates no direct information for it. To understand the behavior for a finite γ, consider the independent prior case $\phi = 0$. The KL-divergence takes a similar form to the marginal comparison in Equation (3.31),

$$\mathcal{KL}\left(p(\lambda|\mathcal{D}_n, \eta)||p(\lambda|\eta)\right) = -\frac{1}{2}\log\frac{1}{n\sigma_2^2/\gamma + 1} + \frac{1}{2(n\sigma_2^2/\gamma + 1)} + \frac{n^2\sigma_2^2/\gamma^2}{2(n\sigma_2^2/\gamma + 1)^2}\bar{Y}_n^2 - \frac{1}{2}$$

Taking expectations over the data distribution $\mathcal{D}_n|\eta, \lambda$, we can simplify this to

$$\mathbb{E}_{\mathcal{D}_n|\eta,\lambda}\left[\mathcal{KL}\left(p(\lambda|\mathcal{D}_n, \eta)||p(\lambda|\eta)\right)\right] = -\frac{1}{2}\log\frac{1}{\frac{n}{\gamma}\sigma_2^2 + 1} + \frac{n^2\sigma_2^2/\gamma^2\left(\lambda^2 - \sigma_2^2\right)}{2(n\sigma_2^2/\gamma + 1)^2} \tag{3.34}$$

Because we are inspecting the conditional distribution of λ given η, we can interpret this as the amount of direct information for λ. A balance between data quality γ, sample size n, and prior marginal variance σ_2^2 determine the KL. Note that this is still a function of λ since we take expectations of the KL over the likelihood, which is indeed a function of parameters. The quantity, n/γ, represents a measure of effective sample size. The higher the effective sample size, the more direct information for λ. Similarly, $1/\sigma_2^2$ represents a prior sample size; a smaller prior variance corresponds to more prior information. The underlying parameter λ plays a similar role here as in the marginal comparisons. Large values of λ lead to greater deviations from the prior. Indeed, the KL grows as $(\lambda - 0)^2$ following the intuition we discussed from

Equation 3.22.

Prior-posterior comparisons help shed light on the information available to parameters from data. This example helped us shed light on how potential measures of information flow in the Bayesian context help disentangle direct from indirect information in a situation where the prior distribution induces both types. We inspected the measures precisely in a few special cases, but we discovered some general trends. One helpful feature of this example was that the indirect information flow was wholly attributable to the prior distribution because the parameters are orthogonal. In the following section, we inspect the correlation variation where the model parameters are not orthogonal, adding interest to our interpretations.

3.4.2 Exploring a Bayesian Analysis of the Correlation Variation

In this section, we analyze the prior-posterior measures of information flow in a situation where the parameters in our model are not orthogonal. We simplify the job by placing independent priors to understand the impact of the dependence of the model parameters in the likelihood. We analyze one dataset in detail and conduct a simulation study to understand how the data change

Recalling the formulation in Section 3.1, we observe m paired observations and k from each marginal and let $\alpha = m/(m+k)$ and $n = m+k$. The likelihood for a dataset $\mathcal{D}_n = \left\{\{(X_i, Y_i)\}_{i=1}^{m}, \{X_i\}_{i=m+1}^{m+k}, \{Y_i\}_{i=m+k+1}^{m+2k}\right\}$ can be written as

$$L(\rho, \sigma^2; \mathcal{D}_n) = \frac{1}{(2\pi)^n \sigma^{2n}(1-\rho^2)^{m/2}} \exp\left(-\frac{1}{2(1-\rho^2)\sigma^2}\left(\sum_{i=1}^{m} X_i^2 + \sum_{i=1}^{m} Y_i^2 - 2\rho\sum_{i=1}^{m} X_i Y_i\right) - \frac{1}{2\sigma^2}\left(\sum_{i=m+1}^{n} X_i^2 + \sum_{i=m+k+1}^{m+2k} Y_i^2\right)\right)$$

In our simulations we vary α, equivalently we hold $m + k$ fixed and vary m. In order to retain a relationship between increasing values of α we keep our paired data and sequentially add new pairs. In other words, going from a dataset with m pairs to $m + 1$ pairs looks like this:

$$\left\{\{(x_i, y_i)\}_{i=1}^{m}, \{x_i\}_{i=m+1}^{m+k}, \{y_i\}_{i=m+k+1}^{m+2k}\right\} \rightarrow \left\{\{(x_i, y_i)\}_{i=1}^{m}, (X_{m+1}, Y_{m+1}), \{x_i\}_{i=m+2}^{m+k}, \{y_i\}_{i=m+k+2}^{m+2k}\right\}.$$

We, essentially, replace one set of independent marginal observations with a new paired observation (shown

in blue) generated from the model. This allows us to study the data sets sequentially, but retain the independence structure of the unpaired data.

In the previous example, we studied how the prior correlation shows through in prior-posterior comparisons in a case when the parameters were orthogonal in the likelihood. This correlation variation certainly does not have orthogonal parameters. The parameters ρ and σ^2 share a complex relationship in the likelihood that we wish to explore directly, so we place priors on ρ and σ^2 independent with marginals

$$\rho \sim \text{Beta}_{(-1,1)}(a, a) \quad \text{and} \quad \sigma^2 \sim \text{Gamma}(3, 1). \tag{3.35}$$

The prior distribution for σ^2 is selected so that σ^2 has a mode of 2 a-priori, far from what the true generating value will be. The prior distribution for ρ is a Beta distribution shifted and scaled to support $\rho \in (-1, 1)$. The hyper-parameter a controls the information for ρ contained in the prior. The priors considered will be symmetric so that the prior mode for ρ is zero with precision increasing with a.

We generate data from our model with true underlying parameters $\rho_0 = 0.5$ and $\sigma_0^2 = 1$. Figure 3.5 shows contour plots of a posterior distribution for (ρ, σ^2) given a random sample of $n = 20$ data points. The prior distributions are given in Equation 3.35 and $a = 2$ is used. The true parameter values are shown in white cross hairs. Figure 3.5 (top-left) shows the posterior distribution when all 20 points are unpaired, i.e. there is no direct information for ρ. Instead, the curvature in the ρ direction is due to the prior. By increasing α we add more pairs, introducing curvature in the ρ direction. Maximum a-posteriori (MAP) estimates for ρ show a compromise between prior and likelihood information. When $\alpha = 0$, no information for ρ is available in the likelihood and the MAP estimate of ρ is appropriately 0. As we increase α to $0.25, 0.75$ and 1, the MAP estimates approach the MLE for ρ (when $\alpha = 1$) of 0.66 and are respectively $0.355, 0.625$ and 0.615.

Changing α also changes our inference for σ^2. When $\alpha = 0$, the MAP estimate of σ^2 is 0.90, close to the true generating value $\sigma_0^2 = 1$. When $\alpha = 1$, however, the MAP estimate is 1.27, closer to the prior mode of 2. The impact of α on the estimates of σ^2 are noticeable. Intuitively, adding paired observations removes information from σ^2, since the paired observation is replacing two independent marginals capable of independently contributing to the inference on σ^2. Recall our thought experiment: when $\rho = 1$, $\alpha = 1$

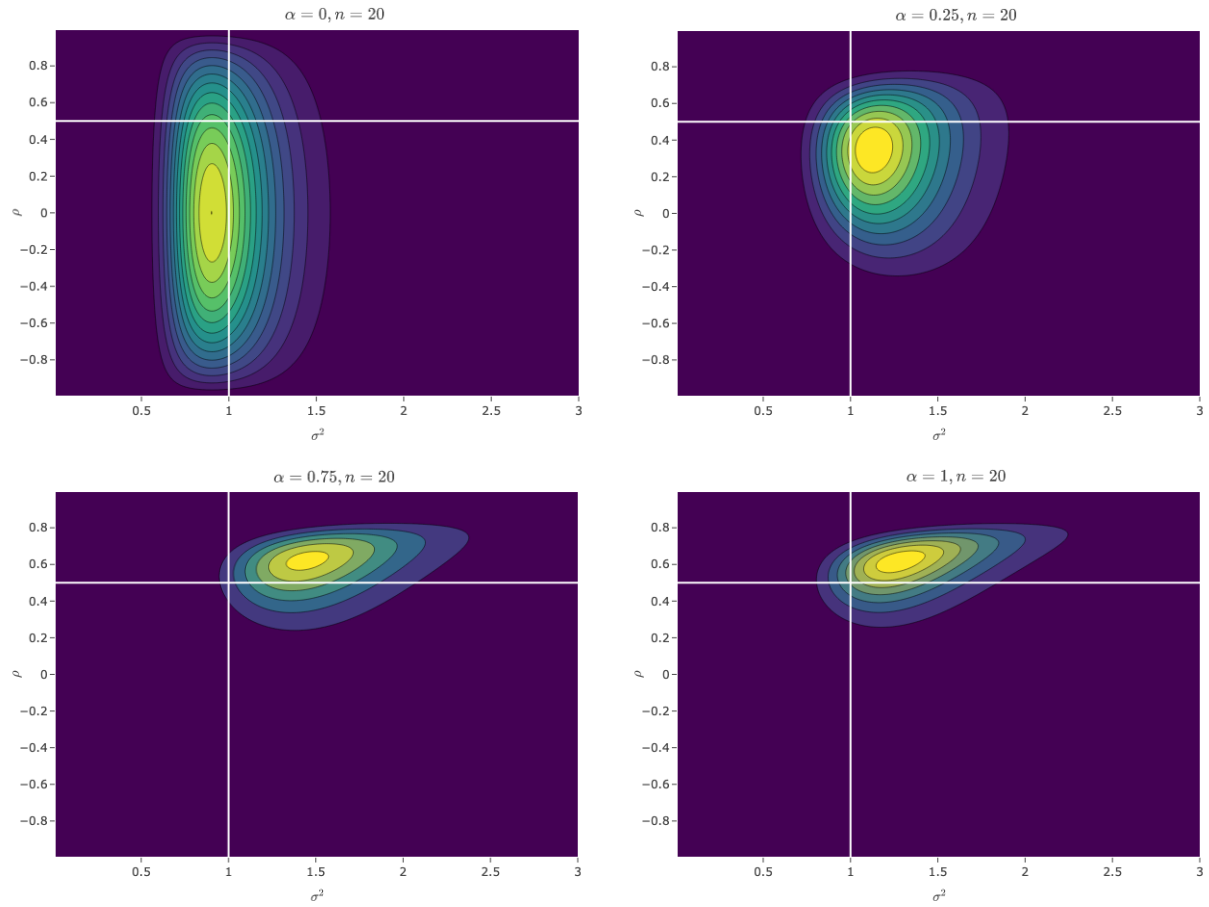


Figure 3.5: Posterior distribution for ρ and σ^2 under data $\mathcal{D}_{20}$ generated with with $\rho_0 = 0.5$ and $\sigma_0^2 = 1$. Yellow corresponds to high posterior density.

gives us effectively n points for inferring σ^2, while $\alpha = 0$ gives us effectively $2n$ such points. In the same experiment, even two paired observations provide the same amount of information for inferring ρ than do any more pairs. Still, it is unclear whether the changes in the inference for ρ and σ^2 are due to (1) noise, (2) the interdependence between ρ and σ^2, or (3) the changing information contained in the data. We continue dissecting this example to analyze (2) and (3) before providing a simulation study that addresses (1).

We can use unidentifiability to focus our conversation on information. When $\alpha = 0$, the likelihood is unidentifiable and is indeed partially identifiable with unidentifiable parameter ρ. Because of this, and

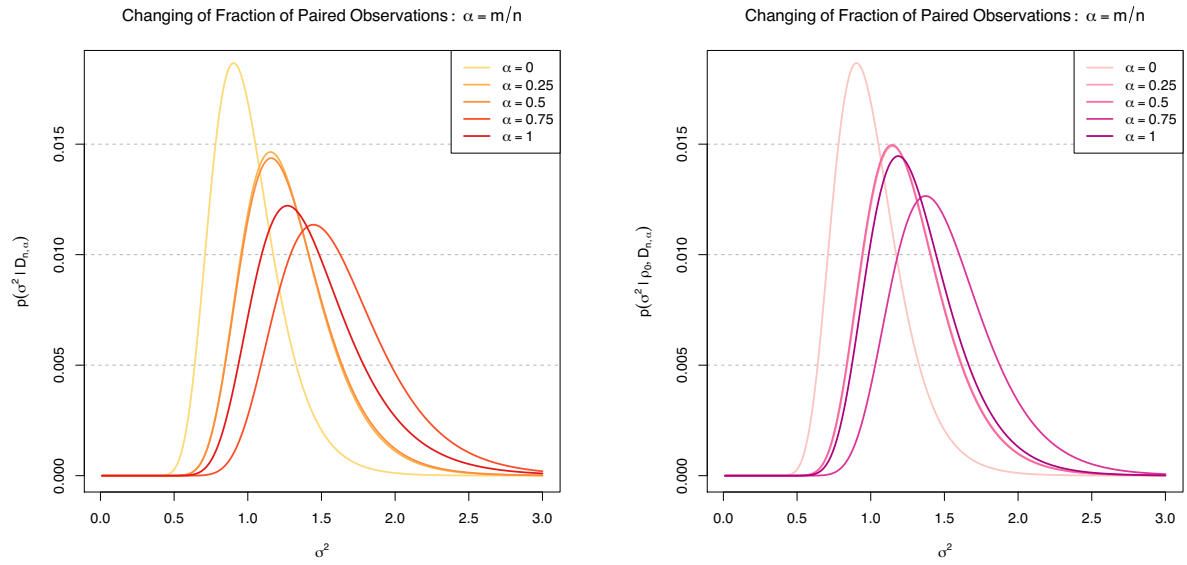


Figure 3.6: The two posteriors distribution $p(\sigma^2|\mathcal{D}_{n,\alpha})$ (left) and $p(\sigma^2|\rho_0, \mathcal{D}_{n,\alpha})$ (right) when the true underlying generating values are $\rho_0 = 0.5, \sigma_0^2 = 1, n = 20$ and $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$

because the priors are independent, the following hold in this situation:

$$p(\sigma^2|\mathcal{D}_{n,\alpha}) = p(\sigma^2|\rho, \mathcal{D}_{n,\alpha}), \tag{3.36}$$

and

$$p(\rho|\mathcal{D}_{n,\alpha}) = p(\rho|\sigma^2, \mathcal{D}_{n,\alpha}) = p(\rho). \tag{3.37}$$

Equation 3.36 indicates that when $\alpha = 0$, the correlation parameter does not change our inference for σ^2. We have a similar conclusion for ρ. But the second equality in Equation 3.37 reminds us that the data contain zero information for ρ. Using Equations 3.36 and 3.37 as anchors, we begin by exploring the inference for σ^2 and then ρ in our example.

Changing α changes both the marginal and conditional inference for σ^2 – see Figure 3.6. When $\alpha = 0$, we effectively have 40 independent marginal observations to infer σ^2 most precisely. Adding more paired data reduces the number of independent marginals and changes the inference, reducing the information. Indeed, comparing $p(\sigma^2|\rho, \mathcal{D}_{n,\alpha})$ to $p(\sigma^2|\rho)$ provides a natural measure of direct information when evaluated at ρ_0. In Figure 3.6 (left), α is sequentially $0, 0.25, 0.5, 0.75$ and 1 the KL-divergence measure of

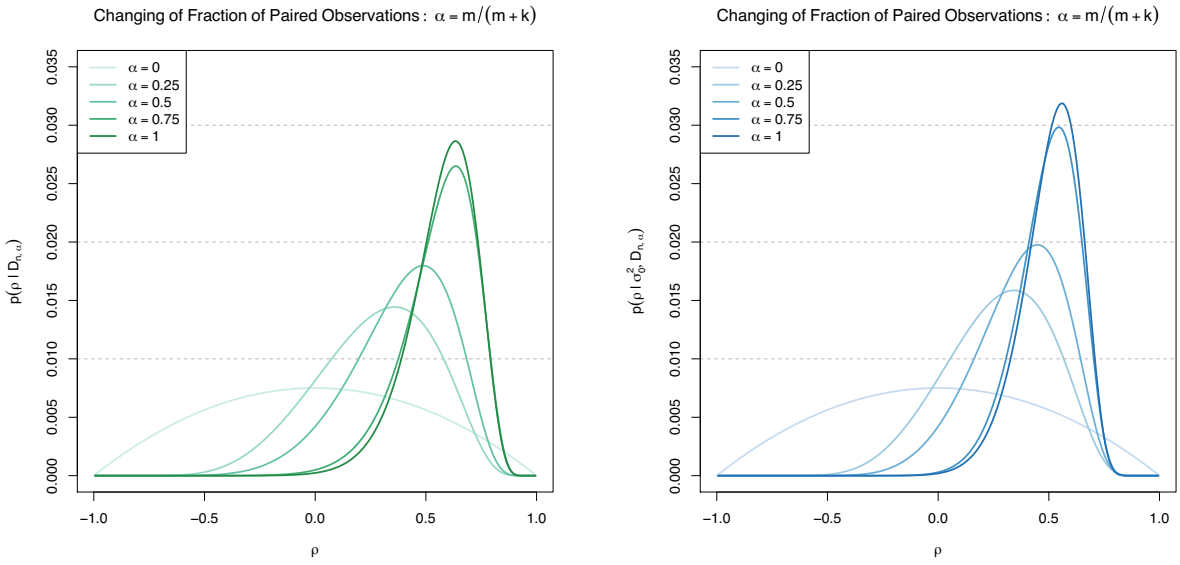


Figure 3.7: The two posteriors distribution $p(\rho|\mathcal{D}_{n,\alpha})$ (left) and $p(\rho|\sigma_0^2, \mathcal{D}_{n,\alpha})$ (right) when the true underlying generating values are $\rho_0 = 0.5$, $\sigma_0^2 = 1$, $n = 20$ and $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$

direct information are $1.82, 1.39, 1.40, 1.13$ and 1.3. These distributions seem to get closer to the prior distribution (whose mode is 2) as we add more pairs. This could, however simply be due to noise.

On the flip side, observing more pairs changes the inference ρ which should, as a consequence, change the inference for σ^2. Indeed, the conditional distributions shown in Figure 3.6 (right) differ from the marginal distributions (left). By conditioning on ρ_0 we are injecting information into the system, similar to our approach in Section 3.3. If the parameters were orthogonal, this additional conditioning would make no difference (because our priors are independent). The bigger the difference, then, the stronger the dependence. In this example the KL-divergence measures of indirect information for σ^2 are $0, 0.0005, 0.0030, 0.0376$ and 0.074 for α equal to $0, 0.25, 0.5, 0.75$ and 1 respectively. The influence of α on ρ is, by construction, more dramatic.

Figure 3.7 shows how marginal and conditional inferences for ρ vary with α. When $\alpha = 0$, we observe no paired observations and recover the prior distribution of ρ. As we increase the fraction of paired observations, our inference becomes sequentially more peaked and centered around the true underlying generating value. That the KL-divergence measure of direct information should increase with alpha from zero is obvious. However, the comparison between the marginal and conditional, the measure of indirect

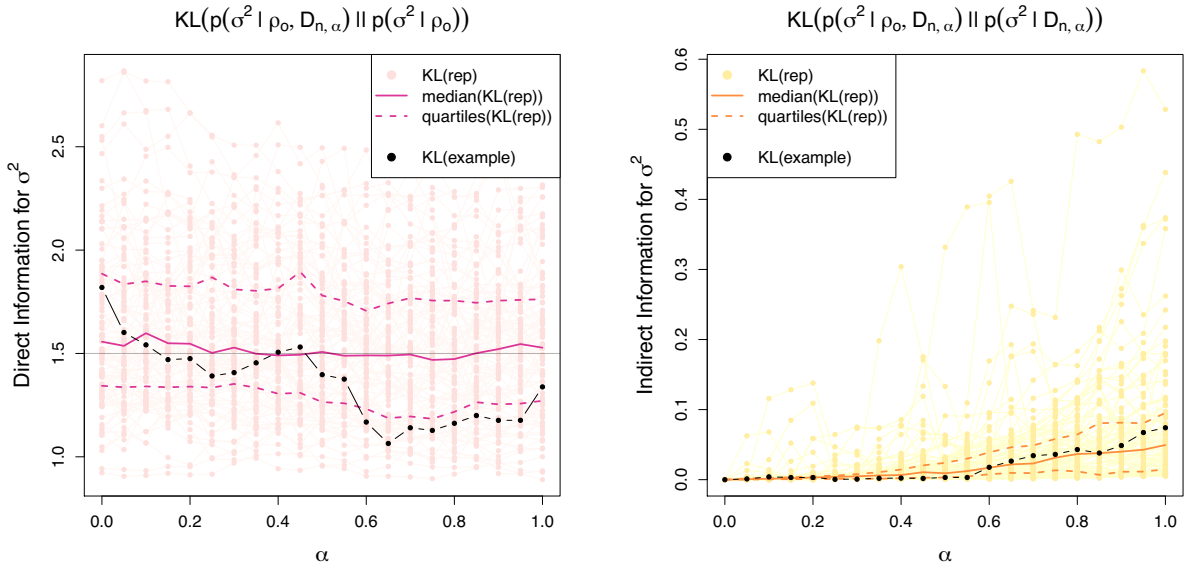


Figure 3.8: Measures of direct and indirect information for σ^2 with replicates shown in colored dots. The mean and two quartiles shown in lines.

information, is more subtle.

If $p(\rho|\sigma_0^2, \mathcal{D}_{n,\alpha})$ is equal to $p(\rho|\mathcal{D}_{n,\alpha})$, then injecting complete information for σ^2 does not add to the inference of ρ beyond what the data already tells us. Indeed, Figure 3.7 shows that $p(\rho|\sigma_0^2, \mathcal{D}_{n,\alpha})$ is more peaked than $p(\rho|\mathcal{D}_{n,\alpha})$ for these data, so the additional information does help. In fact, this comparison is the measure of indirect information information for ρ. In our example dataset, they are $0, 0.02, 0.04, 0.3$ and 0.32 for increasing α. These differences are not large, but they are not zero. This means that the inference on σ^2 indeed changes the inference for ρ. Larger differences occur for the measure of direct information for ρ.

The equality of the distributions in Equation 3.37 anchor the direct information. When $\alpha = 0$, the conditional posterior and conditional prior are equal. A difference would indicate that the data provide information beyond that which is explained by the other parameter. Indeed, for the increasing values of α the KL divergence between $p(\rho|\sigma_0^2, \mathcal{D}_{n,\alpha})$ and $p(\rho|\sigma_0^2)$ are $0, 0.48, 0.72, 1.2$ and 1.3. These differences are much larger than the measures of indirect information for ρ. It is unclear whether these trends of increasing information are systematic of simply due to noise.

Simulation Study

In order to understand whether the trends observed in our example dataset are systematic, we simulate 100 replicated data sets and analyze them for increasing values of α. We keep our sample size at $n = 20$ but vary α to include one new pair at a time. That is, we study our information measures for α at increments of $1/20$.

Figure 3.8 shows our measures of direct and indirect information for σ^2. This variance parameter is always identifiable; the inference for σ^2 should differ from the prior. Indeed, Figure 3.8 (left) shows that the large amount of direct information the data contain for estimating σ^2 is stable across α. The median KL measure is around 1.5, but values above 2 occur frequently. In our first example (shown in black) we observed a decreasing amount of information for σ^2.

This extreme change in information content for σ^2 is atypical. Indeed the median KL divergence among replicates does not noticeably decrease with α. Upon further inspection, we found that both the paired and marginal observations contained several outliers that inflate the inference on σ^2 when α is large. To give a rough idea: when $\alpha = 0.7$, the MLEs are 0.67 for ρ and 1.48 for σ^2, but when $\alpha = 0.3$ the MLEs were instead 0.47 and 1.1 respectively. Because we purposefully selected a prior for σ^2 with a mode of 2, the posterior distributions for higher α were close to the prior. Our measures are sensitive to atypical data, an issue we discuss further below.

The indirect information for σ^2 tells a different study – it is shown in 3.8 (right). Knowing ρ will change our inference for σ^2. The impact is minimal, but for some samples it can be large. This difference is largest when $\alpha = 1$. At this point, the data can most precisely estimate ρ, and that additional information changes the inference for σ^2 most drastically. These observations are interesting, but let us inspect the information content for ρ, our primary parameter.

These measures of information flow are not calculable in practice as we do not know the actual underlying parameter values (ρ_0, σ_0^2). In practice, we can either inspect these quantities as we vary the conditioning parameter or can use a plug-in estimate. One option would be to the MAP estimate of the conditioning parameter, we do this and shown in Appendix B.4. Using a plug-in estimate will, however, underestimate the indirect information. When we condition on the true underlying parameters, we are injecting information from two independent sources: the data and the independent knowledge of the actual parameter. However, conditioning on the MAP estimate injects information from the data into the posterior twice: by

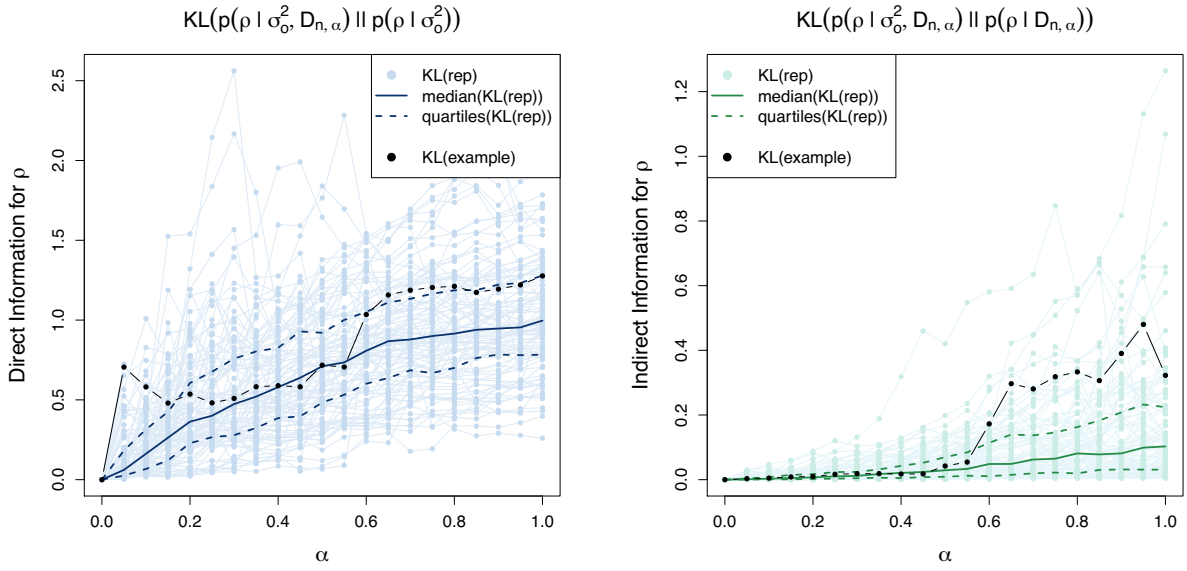


Figure 3.9: Measures of direct and indirect information for ρ with replicates shown in colored dots. The mean and two quartiles shown in lines.

conditioning on the data directly and, again, by conditioning on a data-driven estimate of the parameter. This leads to an underestimation of the indirect information.

The indirect information for ρ shown in Figure 3.9 (right) increases with α, and is relatively small. Our example dataset, shown in black, is typical in the sense that it follows the general trend of other data sets but atypical in that it shows relatively large measures when we introduce more pairs. The measures of indirect information do vary across simulations in our setting, but the majority stay below 0.2 for nearly every value of α. However, because this quantity is positive, the inference for ρ is affected indirectly by σ^2 on average, if slightly. However, as expected, the inference for ρ is more sensitive to the data structure as measured by the direct information.

Figure 3.9 (left) shows our measure of direct information for ρ. The lines and dots in black show the direct information measure for the example analyzed in detail above. Indeed, the information systematically grows as we increase α, but the variability is extreme. Our example showed a KL of 1.2 for $\alpha = 1$. This KL is near the 75^{th} percentile but nowhere near the largest values. Relative to the measures of indirect information, the measures of direct information are large. When $\alpha = 1$, the conditional posterior distribution for ρ tends to be at a KL-divergence of 1 from the prior. To give a rough idea of what this means, we use

KL-divergence formula for two univariate Normal distributions in Equation 3.22. The prior distribution on ρ, if approximated by a Normal, would have mean $\mu_\star = 0$ and standard deviation roughly $\sigma_\star = 0.45$. When $\alpha = 1$ the posterior modes for $\rho|\sigma_0^2$ tend to be roughly around the true generating value of 0.5. If the posterior mode is $\tilde{\mu} = 0.5$, the posterior standard deviations would roughly tend to be $\tilde{\sigma} = 0.18$ (since the KL-divergence tends to be 1); that is, the posterior distributions tend to be 2.5 times more precise than the prior.

Several simulated data sets exhibit spikes in direct information for ρ at α around 0.2 and 0.3 (corresponding to 4 and 6 observed pairs), but then decrease and return to typical values. The first several pairs in these data fall nearly perfectly in line. The conditional posterior distributions for $\rho|\sigma_0^2$, therefore, concentrate near one, creating a spike in the KL-divergence when compared to the prior. After observing a few more pairs, the correlation begins to decrease and so the KL divergence between the prior and posterior return to average levels.

This points to the issue of atypical data discussed earlier. We measure changes in the information as changes in inference. In comparing prior and posterior distributions, we are comparing prior inference for ρ versus posterior inference for ρ. If we observe one new data point that drastically changes our posterior, we say that point contains lots of information for the parameter of interest. This issue is troubling because it gives undue weight to atypical data, like outliers. In the correlation variation observing a high leverage point (one that does not fall into the typical trend of the data), we would say that that point contains a large amount of information for ρ. While this is indeed worrisome, the promise of measures of information may still be worth the trouble. Being able to separate information gained by introducing new data from the information gained from other model parameters is a useful tool in the model building process.

3.5 Discussion

In this chapter, we explored measures of the flow of information from a Bayesian and a classical inference perspective. We used unidentifiable and orthogonal parameters as the cornerstone on which we built our measures of information flow. We discussed the connection between the direct flow of information and unidentifiable parameters. When a parameter is unidentifiable, the data provide no direct information for

that parameter. We also discussed the connection between the indirect flow of information and orthogonal parameters. Parameters are orthogonal if their Fisher co-information is zero; in this case, no information passes between them.

We introduced measures of information flow in the context of both likelihood and Bayesian analysis. In the likelihood analysis framework, we use the Fisher information matrix to measure the impact non-orthogonality through what we called the redirection penalty $\mathcal{R}_{RP}(\theta)$, a quantity between 0 and 1. This quantity measures the impact redirecting part of the information available in the data for estimating θ to aid in the estimation of other model parameters. If the redirection penalty is zero, the inference on θ is as-if it were the only parameter in the model. That is, the precision for estimating θ does not get worse because of the presence of other model parameters. As the redirection penalty increases, the amount of information that flows directly to θ decreases. The precision for estimating θ gets worse precisely because we need to distribute the information available for it to other model parameters.

In the Bayesian analysis framework, we consider prior-posterior comparisons to measure direct and indirect information for model parameters. In the Bayesian setting, model parameters can inform one another through their dependence in the likelihood (as in the likelihood analysis framework), but also their prior relationship. To study the potential flow of information we considered marginal prior-posterior comparisons to measure overall information, conditional prior-posterior comparisons to measure direct information and these two together to measure indirect information. When parameters are orthogonal, and priors are independent, there is no indirect information passed between model parameters. The marginal prior-posterior comparison would show no change. If parameters are orthogonal, but we have dependent priors the correct comparison would, instead, be the difference between the conditional prior and conditional posterior. When model parameters are unidentifiable, there is no direct information available for them in the data. We would observe no change in conditional prior-posterior comparisons, regardless of the prior is structure.

Measures of information flow follow naturally in the Bayesian inference paradigm, the concerns of unidentifiable parameters are of no concern and indeed help anchor our measures. However, the KL-divergence measures we provide are difficult to interpret and do not fully characterize the nature of changing information. The likelihood-based measures, on the other hand, had intuitive and straightforward in-

terpretations. Still, other measures of distributional differences may be more interpretable. de Carvalho et al. (2017) discuss a bounded measure of distributional similarity similar to the Pearson correlation while Xie & Carlin (2006) discuss measuring prior and posterior precision directly, a compromise between our likelihood and Bayesian measures of information.

Our explorations involved the simple case where we have two parameters and consider the flow of information either directly from the data or indirectly through a second parameter. Some explorations can be extended immediately to the case of three or more parameters. For example, with the redirection penalty, we can study the penalty for redirecting the information available for θ_1 through both θ_2 and θ_3, $\mathcal{R}_{RP}(\theta_1|\theta_2, \theta_3)$ or for each θ_2 and θ_3 in turn, $\mathcal{R}_{RP}(\theta_1|\theta_2)$ and $\mathcal{R}_{RP}(\theta_1|\theta_3)$. If θ_1 and θ_2 are informed by the same data, but θ_1 is orthogonal to θ_3 then we would expect $\mathcal{R}_{RP}(\theta_1|\theta_2)$ to be rather large and $\mathcal{R}_{RP}(\theta_1|\theta_3)$ to be 0. This knowledge would allow the model-builder to inspect θ_3 and decide its importance in the model and potentially choose to remove it.

Information, like water, can be elusive. It is difficult to know how it behaves when passed through the intricate channels of a model. However, we can learn about this process by studying how satiated each parameter is and how extensive the channels are between parameters to allow for information to pass. Information is also powerful. It can inform model parameters that validate and cement scientific knowledge like a river leaving the Grand Canyon in its tracks.

A

Technical Details for Chapter 1

A.1 Intensive and Extensive Marginal Effects for Two-Part Models

Another causal analysis relevant when discussing data with excess zero are known as *intensive* and *extensive marginal estimands* – the marginal analysis. This causal analysis can be made precise in zero-truncated models by removing the ambiguity of zero-observations, we call subjects with non-zero outcomes "participants" and subjects with zero outcomes "non-participants". Roughly, the marginal effects measure the impact of a treatment on participation amounts for two separate subgroups of interest, those who would participate regardless of treatment and those who would participate only with treatment. See Heckman (1979) for an overview of Hurdle models and Staub (2014) for a review of intensive and ex-

tensive marginal effects for these models. We give a brief overview here to make the marginal analysis precise. Recall that the data generating mechanism for a Hurdle model is as a mixture of zeros and a zero-truncated distribution, taking this as the data generating mechanism of potential outcomes we have that

$$Y_i(0) \sim ZTPois(p_0, \lambda_0) \quad \text{and} \quad Y_i(1) \sim ZTPois(p_1, \lambda_1)$$

Letting $T_i \in \{0, 1\}$ be the treatment indicator for subject i, we can write the principal strata of interest for the marginal analysis. Following Staub (2014), the principal strata are shown in Table A.1. Because these strata are defined by both potential outcomes, we can never observe these principal strata directly. However, because the hurdle model provides the group labels marginally, we can notice that

$$\begin{aligned} p_0 =& Pr(Y_i = 0|T_i = 0) = \pi^{NP} + \pi^{S1}, & p_1 =& Pr(Y_i = 0|T_i = 1) = \pi^{NP} + \pi^{S2}, \\ (1 - p_0) =& Pr(Y_i > 0|T_i = 0) = \pi^{P} + \pi^{S2} \quad \text{and} & (1 - p_1) =& Pr(Y_i > 0|T_i = 1) = \pi^{P} + \pi^{S1}. \end{aligned}$$

If we are willing to assume there are no Switchers (2), i.e. $\pi^{S2} = 0$, we can identify these proportions because we can estimate both p_0 and p_1 from the data and our model. This is essentially assuming that the intervention doesn't make matters worse and is known as the Monotonicity Assumption (Imbens & Rubin, 2015). Given this assumption, we can break down the ATE causal estimand described above into a weighted average of the intensive marginal effect, τ_I, and the extensive marginal effect, τ_E as

$$\begin{aligned} \mathbb{E}[Y_i(1) - Y_i(0)] &= \tau_I \Pr\left(Y_i(0) = 0, Y_i(1) > 0\right) + \tau_E \Pr\left(Y_i(0) > 0, Y_i(1) > 0\right) \\ &= \pi^{S1}\tau_I + \pi^{P}\tau_E \end{aligned}$$

where $\tau_I = \mathbb{E}\left[Y_i(1)|Y_i(0) = 0, Y_i(1) > 0\right]$ and $\tau_E = \mathbb{E}\left[(Y_i(1) - Y_i(0))|Y_i(0) > 0, Y_i(1) > 0\right]$.

The Extensive Marginal Effect can be interpreted as the overall treatment effect on the subgroup of subjects who would participate regardless of the treatment while the Intensive marginal effect measures the overall treatment effect only among those who are encouraged to participate with treatment. These effects can typically be bounded from observed data, see Staub (2014) for a complete overview on the procedures.

Label	Subgroup	Potential Outcomes	Proportion
NP	Nonparticipants	$(Y_i(0) = 0, Y_i(1) = 0)$	π^{NP}
S_1	Switchers (1)	$(Y_i(0) = 0, Y_i(1) > 0)$	π^{S_1}
S_2	Switchers (2)	$(Y_i(0) > 0, Y_i(1) = 0)$	π^{S_2}
P	Participants	$(Y_i(0) > 0, Y_i(1) > 0)$	π^{P}

Table A.1: Four principal strata defined by potential outcomes from a Hurdle model

A.2 Mean and Variance of zero-inflated distributions

In this section we derive the mean and variance of a general zero-inflated distribution. For the following, let $F(\theta)$ represent a general distribution with finite mean $\mu(\theta)$ and variance $\sigma^2(\theta)$. We can define a zero-inflated random variable as

$$Y_i \sim \begin{cases} 0 & \text{w.p.} \quad p \\ F(\theta) & \text{w.p.} \quad 1-p. \end{cases}$$

For example, for $F = Pois(\lambda)$ we have the parameters $\mu = \sigma^2 = \lambda$. If Y_i is a zero-inflated Poisson we say $Y_i \sim ZIPois(p, \lambda)$, but more generally we could say $Y_i \sim ZIF(p, \theta)$.

We express the ZIF random variable as the product of two random variables. Let

(1) $Z_i \sim Bern(p)$

(2) $X_i \sim F(\theta)$

(3) $X_i \perp\!\!\!\perp Z_i$

Our zero-inflated random variable is then $Y_i \sim (1 - Z_i)X_i$, with mean and variance:

$$\begin{aligned} \mathbb{E}[Y_i] &= \mathbb{E}[(1 - Z_i)X_i] = \mathbb{E}\left[\mathbb{E}[(1 - Z_i)X_i|Z_i]\right] = \mathbb{E}\left[(1 - Z_i)\mathbb{E}[X_i|Z_i]\right] \\ &= \mathbb{E}\left[(1 - Z_i)\mathbb{E}[X_i]\right] = \mathbb{E}\left[\mu(1 - Z_i)\right] \\ &= \mu\left(1 - \mathbb{E}[Z_i]\right) = \mu(1 - p) \end{aligned}$$

$$
\begin{aligned}
var(Y_i) &= \mathbb{E}[var(Y_i|Z_i)] + var(\mathbb{E}[Y_i|Z_i]) \\
&= \mathbb{E}[(1-Z_i)^2 var(X_i|Z_i)] + var(\mu(1-Z_i)) \\
&= \mathbb{E}[(1-Z_i) var(X_i)] + \mu^2 var(1-Z_i) \\
&= \mathbb{E}[(1-Z_i)\sigma^2] + \mu^2 var(Z_i) \\
&= \sigma^2 \mathbb{E}[(1-Z_i)] + \mu^2 p(1-p) \\
&= \sigma^2(1-p) + \mu^2 p(1-p) = (1-p)(\sigma^2 + p\mu^2)
\end{aligned}
$$

where for the variance calculation we made use of the fact that since Z_i is either 0 or 1, $(1-Z_i) = (1-Z_i)^2$.

A.3 Derivations for model-based Confidence Intervals

ATE estimation using Zero-Inflated Models

In Sections 1.4.2 we introduced a procedure for using zero-inflated models for estimating ATE. The approximation of the model parameters $(\hat{p}_0, \hat{p}_1, \hat{\mu}_0, \hat{\mu}_1)$ is rarely done directly, instead it is usually fit with a two part model. In the first part, a binary regression (typically logistic) is fit to a latent mixture indicator with parameters (α_0, α_1). Then, as a second stage, a generalized linear model (with an F-specific link) is fit to the sampling distribution with parameters (θ_0, θ_1). For example, if we fit a zero-inflated Poisson distribution with a logit link for the *zero part*, and a `log` link for the sampling distribution then

$$
p_t = \text{expit}(\alpha_t) \quad \text{and} \quad \mu_t = \exp(\theta_t) \quad \text{for} \quad t = 0, 1.
$$

In order to construct confidence intervals for $\hat{\tau}_{zif}$ we need to consider a Delta Method that accounts for two things. (1) the link functions connecting the parameters $(\hat{\alpha}_t, \hat{\theta}_t)$ to $\hat{p}_t$ and $\hat{\mu}_t$ and (2) the combination of these parameters into the estimate $\hat{\tau}_{zif}$. To be specific, consider the common case of the zero-inflated Poisson. If we fit out outcomes using dummy variables for treatment* we get estimates of $(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\theta}_0, \hat{\theta}_1)$

*Two-group comparisons using models is most cleanly done using dummy variables: `T0=1*(T==0)` and `T1=1*(T==1)`

directly with default log link for the sampling distribution and logit link for the zero part as well as a covariance matrix, $\hat{\Sigma}$, of these estimates.[†]. The first set of transformations invert the link functions for the two parts of the model separately,

$$\begin{pmatrix} \hat{p}_0 \\ \hat{p}_1 \\ \hat{\lambda}_0 \\ \hat{\lambda}_1 \end{pmatrix} = g_1(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\theta}_0, \hat{\theta}_1) = \begin{pmatrix} \text{expit}(\hat{\alpha}_0) \\ \text{expit}(\hat{\alpha}_1) \\ \exp(\hat{\theta}_0) \\ \exp(\hat{\theta}_1), \end{pmatrix}$$

while the second transformation converts these parameter estimates to the effect estimate of interest,

$$\hat{\tau}_{zif} = g_2(\hat{p}_0, \hat{p}_1, \hat{\lambda}_0, \hat{\lambda}_1) = (1 - \hat{p}_1)\hat{\lambda}_1 - (1 - \hat{p}_0)\hat{\lambda}_0.$$

We lay these out separately because it is important to keep in mind the potential sources of variance inflation/compression. If the transformation changes the space drastically, as the `expit` function would for $\hat{\alpha}_t$ far from zero, the variance associated to $\hat{p}_t$ would also be large as a consequence. If, on the other hand, we consider a transformation to the mean involving the reciprocal of a parameter $\hat{\theta}_t$ as we might observe in the mean of a Beta-Binomial distribution, we would observe variance compression for large values of $\hat{\theta}_t$.

Simply joining these transformations will simplify the Delta method calculation since

$$\hat{\tau}_{zif} = g(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\theta}_0, \hat{\theta}_1) = (1 - \text{expit}(\hat{\alpha}_1))exp(\hat{\theta}_1) - (1 - \text{expit}(\hat{\alpha}_0))exp(\hat{\theta}_0).$$

Recall that our estimand of interest can be written as a function of a four-dimensional parameter $(\theta_0, \theta_1, \alpha_0, \alpha_1)$, that is

$$\tau_{sp} = (1 - \text{expit}(\alpha_1)) \exp(\theta_1) - (1 - \text{expit}(\alpha_0)) \exp(\theta_0).$$

[†]We can fit the zero inflated Poisson distribution with `zeroinfl(Y~0+T0+T1|0+T0+T1,dist="poisson")` and covariance matrix can be extracted using the `vcov` function

The standard error of $\hat{\tau}_{zif}$ is obtained using the Delta method. For this, we need the partial derivatives of this function with respect to the original parameter space. This is

$$\nabla\tau_{sp}(\theta_0, \theta_1, \alpha_0, \alpha_1) = \begin{pmatrix} -\left(1 - \text{expit}(\alpha_0)\right)\exp(\theta_0) \\ \left(1 - \text{expit}(\alpha_1)\right)\exp(\theta_1) \\ \text{expit}(\alpha_0)\left(1 - \text{expit}(\alpha_0)\right)\exp(\theta_0) \\ -\text{expit}(\alpha_1)\left(1 - \text{expit}(\alpha_1)\right)\exp(\theta_1) \end{pmatrix}.$$

The standard error estimate of $\hat{\tau}_{zif}$ as

$$\widehat{SE}_{zif} = \sqrt{\nabla g(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\theta}_0, \hat{\theta}_1)^T \, \hat{\Sigma} \, \nabla g(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\theta}_0, \hat{\theta}_1)}.$$

Again, using ML plug-in estimates to approximate our estimator's sampling variance, we obtain a Normal approximation and Delta method-based confidence interval,

$$\hat{\tau}_{zif} \pm z_{\alpha/2} \, \widehat{SE}_{zif}.$$

ATE estimation using Zero-Inflated Models with Covariate Information

In Section 1.5.1 we introduced a method for deriving confidence intervals for the ATE from fitting Zero-Inflated generalized linear models. The unconditional ATE estimand is a function of the 6-dimensional parameter space as well as the covariate distribution $F_X(\cdot)$,

$$\tau_{sp} = \left(1 - E(\alpha_1, \beta_1)\right)\exp(\theta_1) - \left(1 - E(\alpha_0, \beta_0)\right)\exp(\theta_0).$$

When using the typical logistic regression setup, we have $E(\alpha, \beta) = \int \text{expit}\left(\alpha + \beta x\right) dF_X(x)$. This is the expectation of a logit-Normal distribution with parameters α and β which can be calculate through numerical integration when the covariates are (even approximately) Normally distributed.

The maximum likelihood estimate of this quantity is obtained using the plug-in principle. However, the standard error of the estimate shown in Equation (1.8) is obtained using the Delta method. For this, we

need the partial derivatives of τ_{sp} with respect to the original parameter space. These are

$$\nabla\tau_{sp}(\theta_0,\theta_1,\alpha_0,\alpha_1,\beta_0,\beta_1) = \begin{pmatrix} -\left(1-E(\alpha_0,\beta_0)\right)\exp(\theta_0) \\ \left(1-E(\alpha_1,\beta_1)\right)\exp(\theta_1) \\ \frac{\partial}{\partial\alpha_0}E(\alpha_0,\beta_0)\exp(\theta_0) \\ -\frac{\partial}{\partial\alpha_1}E(\alpha_1,\beta_1)\exp(\theta_1) \\ \frac{\partial}{\partial\beta_0}E(\alpha_0,\beta_0)\exp(\theta_0) \\ -\frac{\partial}{\partial\beta_1}E(\alpha_1,\beta_1)\exp(\theta_1) \end{pmatrix}.$$

The partial derivatives have a complex form, we derive these generically as

$$\begin{aligned}
\frac{\partial}{\partial\alpha}E(\alpha,\beta) &= \frac{\partial}{\partial\alpha}\int \text{expit}\left(\alpha+\beta x\right)dF_X(x) \\
&= \int \frac{\partial}{\partial\alpha}\text{expit}\left(\alpha+\beta x\right)dF_X(x) \\
&= \int \text{expit}\left(\alpha+\beta x\right)\left(1-\text{expit}\left(\alpha+\beta x\right)\right)dF_X(x) \\
&= \int \text{expit}\left(\alpha+\beta x\right)dF_X(x) - \int \left(\text{expit}\left(\alpha+\beta x\right)\right)^2 dF_X(x) \\
&= E(\alpha,\beta) - \int \left(\text{expit}\left(\alpha+\beta x\right)\right)^2 dF_X(x).
\end{aligned}$$

The second term is the second moment of a logit-Normal distribution and can be easily calculated via numerical integration. Similarly, the derivative with respect to the covariate effect β is

$$\begin{aligned}
\frac{\partial}{\partial\beta}E(\alpha,\beta) &= \frac{\partial}{\partial\beta}\int \text{expit}\left(\alpha+\beta x\right)dF_X(x) \\
&= \int \frac{\partial}{\partial\beta}\text{expit}\left(\alpha+\beta x\right)dF_X(x) \\
&= \int x\ \text{expit}\left(\alpha+\beta x\right)\left(1-\text{expit}\left(\alpha+\beta x\right)\right)dF_X(x) \\
&= \int x\ \text{expit}\left(\alpha+\beta x\right)dF_X(x) - \int x\ \left(\text{expit}\left(\alpha+\beta x\right)\right)^2 dF_X(x).
\end{aligned}$$

These two terms are also tractable using numerical integration. We can now evaluate the gradient at the

MLE and obtain the confidence intervals for $\hat{\tau}_{zip,x}$ specified in Equation (1.8).

A.4 Definitions of Quantities used in Simulation Analysis

We define some quantities of interest to our different estimates. For each simulation setting we can approximate the the bias and variance of our estimators and standard error estimators with $R = 2000$ replicated data sets. For each dataset we will have four estimators and corresponding standard error estimates, we denote the estimates for the r^{th} with a superscript (r), for the Neymanian, linearly-adjusted, model-based and model-based with covariate adjustment we have

$$(\hat{\tau}^{(r)}, \hat{\tau}^{(r)}_{lm}, \hat{\tau}^{(r)}_{zif}, \hat{\tau}^{(r)}_{zif,x}) \quad \text{and} \quad (\widehat{SE}^{(r)}, \widehat{SE}^{(r)}_{lm}, \widehat{SE}^{(r)}_{zif}, \widehat{SE}^{(r)}_{zif,x}).$$

The bias for the Neymanian estimate is

$$|Bias(\hat{\tau})| = \left| \frac{1}{B} \sum_{r=1}^{R} \hat{\tau}^{(r)} - \tau_{sp} \right|.$$

We can similarly estimate the absolute bias for the other estimators.

We also investigate the the possibility of precision gains, for this we investigate the average standard error estimates for the methods across the replicated data sets. For the Neymanian estimate we have,

$$Ave(\widehat{SE}) = \frac{1}{B} \sum_{r=1}^{R} \widehat{SE}^{(i)}.$$

We can similarly estimate the precision gains for the other estimators.

Finally, we inspect how modeling assumptions might change the precision of standard error estimates.

For this we inspect the precision of SE estimates, it can be approximated with

$$sd(\widehat{SE}) = \frac{1}{B}\sum_{r=1}^{R}\left(\widehat{SE}^{(r)} - Ave(\widehat{SE})\right)^2.$$

B

Technical Details for Chapter 3

B.1 Proofs

Proof of Theorem 3.2.1.

Proof. Given $\int_\Theta p(\theta)d\theta = c_1 < \infty$ and $\int_\mathcal{Y} p(y|\theta)dy = c_2 < \infty$ we can show

$$\int_\Theta p(y|\theta)p(\theta)d\theta < \infty, a.e.\ y \in \mathcal{Y}.$$

Because $p(\theta) > 0, \theta \in \Theta$ and $p(y|\theta) > 0$ for $\theta \in \Theta$ and $y \in \mathcal{Y}$, the Fubini–Tonelli theorem gives us

$$\begin{aligned}\int_{\mathcal{Y}} \int_{\Theta} p(y|\theta)p(\theta)d\theta dy &= \int_{\Theta} \int_{\mathcal{Y}} p(y|\theta)p(\theta)dyd\theta = \int_{\Theta} p(\theta) \int_{\mathcal{Y}} p(y|\theta)dyd\theta \\ &= \int_{\Theta} p(\theta)c_2 d\theta = c_2 \int_{\Theta} p(\theta)d\theta = c_1 c_2 < \infty\end{aligned}$$

Therefore, $\int_{\mathcal{Y}} \int_{\Theta} p(y|\theta)p(\theta)d\theta dy < \infty$, hence $\int_{\Theta} p(y|\theta)p(\theta)d\theta < \infty$ for almost every $y \in \mathcal{Y}$. □

B.2 Calculations for Section 3.3

Here we provide the calculations for the quantities discussed in Section 3.3.

Fisher Information of the Correlation Variation

In this example we observe a total of m observations from a bivariate Normal distribution and $k = n - m$ observations from both corresponding marginals of the bivariate Normal. The likelihood function for $\rho, \sigma_X^2, \sigma_Y^2$ is

$$\begin{aligned}L(\rho, \sigma_X^2, \sigma_Y^2) = L_X L_{XY} L_Y = &\prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp\left(-\frac{x_i^2}{2\sigma_X^2}\right) \\ &\prod_{i=k+1}^{k+m} \frac{1}{2\pi\sqrt{(1-\rho^2)\sigma_X^2\sigma_Y^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{x_i^2}{\sigma_X^2} + \frac{y_i^2}{\sigma_Y^2} - \frac{2\rho x_i y_i}{\sigma_X \sigma_Y}\right)\right) \\ &\prod_{i=k+m+1}^{2k+m} \frac{1}{\sqrt{2\pi\sigma_Y^2}} \exp\left(-\frac{y_i^2}{2\sigma_Y^2}\right)\end{aligned}$$

We later simplify things but in order to reuse some of these keep different marginal variances for X and Y. Consider one observation from each of the three types of data. The second and third type (marginal data) will provide zero Fisher information toward estimating ρ and the second type will only contribute to estimating σ_X and not σ_Y and vise-versa.

X Marginal observation

$$\ell_X = -\frac{x_i^2}{2\sigma_X^2} - \frac{1}{2}\log(\sigma_X^2) + c$$

$$\frac{\partial \ell_X}{\partial(\sigma_X^2)} = \frac{x_i^2}{2\sigma_X^4} - \frac{1}{2\sigma_X^2}$$

$$\frac{\partial^2 \ell_X}{\partial(\sigma_X^2)^2} = -\frac{x_i^2}{\sigma_X^6} + \frac{1}{2\sigma_X^4}$$

So the Fisher Information for σ_X^2 is

$$-\mathbb{E}\left[\frac{\partial^2 \ell_X}{\partial(\sigma_X^2)^2}\right] = \frac{1}{\sigma_X^4} - \frac{1}{2\sigma_X^4} = \frac{1}{2\sigma_X^4}$$

The likelihood here does not involve the correlation or σ_Y.

Y Marginal observation

$$\ell_Y = -\frac{y_i^2}{2\sigma_Y^2} - \frac{1}{2}\log(\sigma_Y^2) + c$$

$$\frac{\partial \ell_Y}{\partial(\sigma_Y^2)} = \frac{y_i^2}{2\sigma_Y^4} - \frac{1}{2\sigma_Y^2}$$

$$\frac{\partial^2 \ell_Y}{\partial(\sigma_Y^2)^2} = -\frac{y_i^2}{\sigma_Y^6} + \frac{1}{2\sigma_Y^4}$$

So the Fisher Information is

$$-\mathbb{E}\left[\frac{\partial^2 \ell_Y}{\partial(\sigma_Y^2)^2}\right] = \frac{1}{\sigma_Y^4} - \frac{1}{2\sigma_Y^4} = \frac{1}{2\sigma_Y^4}$$

Bivariate Observation The likelihood for one bivariate observation (x_i, y_i) is

$$\ell_{XY} = -\frac{1}{2(1-\rho^2)}\left(\frac{x_i^2}{\sigma_X^2} + \frac{y_i^2}{\sigma_Y^2} - \frac{2\rho x_i y_i}{\sigma_X \sigma_Y}\right) - \frac{1}{2}\log(\sigma_X^2) - \frac{1}{2}\log(\sigma_Y^2) - \frac{1}{2}\log(1-\rho^2) + c$$

Consider the special case where $\sigma_X = \sigma_Y = \sigma$, then the likelihood becomes

$$\begin{aligned}\ell_{XY} &= -\frac{1}{2(1-\rho^2)}\left(\frac{x_i^2}{\sigma^2} + \frac{y_i^2}{\sigma^2} - \frac{2\rho x_i y_i}{\sigma^2}\right) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2}\log(1-\rho^2) + c \\ &= -\frac{1}{2(1-\rho^2)\sigma^2}\left(x_i^2 + y_i^2 - 2\rho x_i y_i\right) - \log(\sigma^2) - \frac{1}{2}\log(1-\rho^2) + c \\ &= -\frac{x_i^2 + y_i^2}{2(1-\rho^2)\sigma^2} + \frac{\rho x_i y_i}{(1-\rho^2)\sigma^2} - \log(\sigma^2) - \frac{1}{2}\log(1-\rho^2) + c\end{aligned}$$

The Fisher Information matrix for ρ, σ^2 under the bivariate observations is

$$\mathcal{I}_{XY}(\rho, \sigma^2) = m \begin{pmatrix} \frac{1+\rho^2}{(1-\rho^2)^2} & -\frac{\rho}{\sigma^2(1-\rho^2)} \\ -\frac{\rho}{\sigma^2(1-\rho^2)} & \frac{1}{\sigma^4} \end{pmatrix}.$$

The Marginal observations give us the additional information

$$\mathcal{I}_X(\rho, \sigma^2) = k \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \quad \text{and} \quad \mathcal{I}_Y(\rho, \sigma^2) = k \begin{pmatrix} 0 & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

The total Fisher Information matrix for ρ and σ^2 is therefore

$$\mathcal{I}(\rho, \sigma^2) = \begin{pmatrix} m\frac{1+\rho^2}{(1-\rho^2)^2} & -m\frac{\rho}{\sigma^2(1-\rho^2)} \\ -m\frac{\rho}{\sigma^2(1-\rho^2)} & \frac{m+k}{\sigma^4} \end{pmatrix}.$$

Fisher Information of the Mean Variation

The log likelihood of the mean variation example is

$$\ell(\theta_1, \theta_2) = -\sum_{i=1}^{n_X} \frac{(x_i - (\theta_1 + \theta_2))^2}{2} - \sum_{j=1}^{n_Y} \frac{(y_i - \theta_2)^2}{2\gamma} + c.$$

The Fisher Information matrix is then

$$I(\theta_1, \theta_2) = \begin{pmatrix} n_X & n_X \\ n_X & n_X + \frac{n_Y}{\gamma} \end{pmatrix}$$

B.3 Calculations for Section 3.4

A Bayesian Analysis of the Mean Variation

Here we calculate the posterior distribution for the mean variation with prior distribution

$$\begin{pmatrix}\eta\\ \lambda\end{pmatrix} \sim \mathcal{N}\left(\mu_0 = \begin{pmatrix}0\\ 0\end{pmatrix},\ \Sigma_0 = \begin{pmatrix}\sigma_1^2 & \phi\sigma_1\sigma_2\\ \phi\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right).$$

and likelihood

$$\begin{pmatrix}\bar{X}_n\\ \bar{Y}_n\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\eta\\ \lambda\end{pmatrix},\ \Sigma = \frac{1}{n}\begin{pmatrix}1 & 0\\ 0 & \gamma\end{pmatrix}\right).$$

The marginal and conditional prior distributions are

$$\eta \sim \mathcal{N}(0,\sigma_1^2), \quad \lambda \sim \mathcal{N}(0,\sigma_2^2) \quad \text{and} \quad \lambda|\eta \sim \mathcal{N}\left(\phi\frac{\sigma_2}{\sigma_2}\eta, (1-\phi^2)\sigma_2^2\right)$$

The Posterior Variance-covariance matrix is

$$\begin{aligned}
\Sigma^\star &= \left(\Sigma_0^{-1} + \Sigma^{-1}\right)^{-1}\\
&= \left(\begin{pmatrix}\sigma_1^2 & \phi\sigma_1\sigma_2\\ \phi\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}^{-1} + \begin{pmatrix}1/n & 0\\ 0 & \gamma/n\end{pmatrix}^{-1}\right)^{-1}\\
&= \left(\frac{1}{|\Sigma_0|}\begin{pmatrix}\sigma_2^2 & -\phi\sigma_1\sigma_2\\ -\phi\sigma_1\sigma_2 & \sigma_1^2\end{pmatrix} + \begin{pmatrix}n & 0\\ 0 & n/\gamma\end{pmatrix}\right)^{-1}\\
&= \left(\frac{1}{|\Sigma_0|}\begin{pmatrix}|\Sigma_0|n + \sigma_2^2 & -\phi\sigma_1\sigma_2\\ -\phi\sigma_1\sigma_2 & |\Sigma_0|n/\gamma + \sigma_1^2\end{pmatrix}\right)^{-1} = \left(\frac{1}{|\Sigma_0|}\Sigma_1\right)^{-1}
\end{aligned}$$

Denoting Σ_1 as the matrix on the interior of the parentheses, we continue

$$\Sigma^\star = \frac{|\Sigma_0|}{|\Sigma_1|}\begin{pmatrix}|\Sigma_0|n/\gamma + \sigma_1^2 & \phi\sigma_1\sigma_2\\ \phi\sigma_1\sigma_2 & |\Sigma_0|n + \sigma_2^2\end{pmatrix}$$

The determinant of the matrix Σ_1 simplifies as follows,

$$\begin{aligned}
|\Sigma_1| &= \left(|\Sigma_0|n + \sigma_2^2\right)\left(|\Sigma_0|n/\gamma + \sigma_1^2\right) - \phi^2\sigma_1^2\sigma_2^2 \\
&= \frac{n^2|\Sigma_0|^2}{\gamma} + n|\Sigma_0|\sigma_1^2 + \frac{n|\Sigma_0|\sigma_2^2}{\gamma} + \sigma_1^2\sigma_2^2 - \phi^2\sigma_1^2\sigma_2^2 \\
&= \frac{n|\Sigma_0|}{\gamma}\left(n|\Sigma_0| + \sigma_2^2\right) + |\Sigma_0|n\sigma_1^2 + |\Sigma_0| \\
&= \frac{n|\Sigma_0|}{\gamma}\left(n|\Sigma_0| + \sigma_2^2\right) + |\Sigma_0|\left(n\sigma_1^2 + 1\right)
\end{aligned}$$

Letting $c = c(\phi, \sigma_1^2, \sigma_2^2, \gamma, n) = \left(n|\Sigma_0| + \sigma_2^2\right)n/\gamma + \left(n\sigma_1^2 + 1\right)$, the posterior variance-covariance matrix is

$$\Sigma^\star = \frac{1}{c}\begin{pmatrix} |\Sigma_0|n/\gamma + \sigma_1^2 & \phi\sigma_1\sigma_2 \\ \phi\sigma_1\sigma_2 & |\Sigma_0|n + \sigma_2^2 \end{pmatrix}$$

The posterior mean is similarly

$$\begin{aligned}
\mu^\star &= \Sigma^\star\Sigma^{-1}\begin{pmatrix}\bar{X}_n \\ \bar{Y}_n\end{pmatrix} = \Sigma^\star\begin{pmatrix} n & 0 \\ 0 & n/\gamma\end{pmatrix}\begin{pmatrix}\bar{X}_n \\ \bar{Y}_n\end{pmatrix} = \Sigma^\star\begin{pmatrix} n\bar{X}_n \\ \frac{n}{\gamma}\bar{Y}_n\end{pmatrix} \\
&= \frac{1}{c}\begin{pmatrix} n\bar{X}_n(|\Sigma_0|n/\gamma + \sigma_1^2) + \frac{n}{\gamma}\bar{Y}_n\phi\sigma_1\sigma_2 \\ n\bar{X}_n\phi\sigma_1\sigma_2 + \frac{n}{\gamma}\bar{Y}_n(|\Sigma_0|n + \sigma_2^2)\end{pmatrix}
\end{aligned}$$

The Effect of Independent Priors on the Bivariate Distributions

For the special case of independent priors, i.e. $\phi = 0$, we have that

$$\begin{aligned}
|\Sigma_0| &= \sigma_1^2\sigma_2^2 \\
c_0 = c(\phi = 0, \sigma_1^2, \sigma_2^2, \gamma, n) &= \left(n^2\sigma_1^2\sigma_2^2 + n\sigma_1^2\right)/\gamma + \left(n\sigma_2^2 + 1\right) \\
&= \left(n\sigma_2^2/\gamma + 1\right)\left(n\sigma_1^2 + 1\right),
\end{aligned}$$

so the posterior variance shows that the parameters are also a-posteriori independent:

$$\begin{aligned}\Sigma^{\star} &= \frac{1}{\left(n\sigma_2^2/\gamma+1\right)\left(n\sigma_1^2+1\right)}\begin{pmatrix}\sigma_1^2\sigma_2^2 n/\gamma+\sigma_1^2 & 0\\ 0 & \sigma_1^2\sigma_2^2 n+\sigma_2^2\end{pmatrix}\\ &= \frac{1}{\left(n\sigma_2^2/\gamma+1\right)\left(n\sigma_1^2+1\right)}\begin{pmatrix}\sigma_1^2(n\sigma_2^2/\gamma+1) & 0\\ 0 & \sigma_2^2(n\sigma_1^2+1)\end{pmatrix}\\ &= \begin{pmatrix}\frac{\sigma_1^2}{n\sigma_1^2+1} & 0\\ 0 & \frac{\sigma_2^2}{n\sigma_2^2/\gamma+1}\end{pmatrix}.\end{aligned}$$

The posterior means show that the margins no longer depend on one another, e.g. the posterior mean of η no longer depends on $\bar{Y}_n$,

$$\mu^{\star} = \begin{pmatrix}\frac{n\sigma_1^2}{n\sigma_1^2+1}\bar{X}_n\\ \frac{n\sigma_2^2/\gamma}{n\sigma_2^2/\gamma+1}\bar{Y}_n\end{pmatrix}$$

The Effect of Uninformative Data on the Bivariate Distribution

If we allow the data quality index to increase so that we get poorer and poorer data quality $\gamma \to \infty$ first have that

$$c \to n\sigma_1^2+1$$

So the posterior variance-covariance matrix converges to

$$\Sigma^{\star} \to \frac{1}{n\sigma_1^2+1}\begin{pmatrix}\sigma_1^2 & \phi\sigma_1\sigma_2\\ \phi\sigma_1\sigma_2 & |\Sigma_0|n+\sigma_2^2\end{pmatrix}$$

and the posterior mean converges to

$$\mu^{\star} \to \begin{pmatrix}\frac{n\sigma_1^2}{n\sigma_1^2+1}\bar{X}_n\\ \frac{n\phi\sigma_1\sigma_2}{n\sigma_1^2+1}\bar{X}_n\end{pmatrix}$$

Marginal Distributions of η and λ

The marginal distributions of η and λ are directly available here. In particular,

$$\eta|\mathcal{D}_n \sim \mathcal{N}\left(\mu_\eta^\star, \sigma_\eta^{\star 2}\right) \quad \text{and} \quad \lambda|\mathcal{D}_n \sim \mathcal{N}\left(\mu_\lambda^\star, \sigma_\lambda^{\star 2}\right)$$

with means and variances

$$\begin{aligned}
\mu_\eta^\star &= c^{-1}\left(n\bar{X}_n(|\Sigma_0|n/\gamma + \sigma_1^2) + \frac{n}{\gamma}\bar{Y}_n\phi\sigma_1\sigma_2\right) \\
\sigma_\eta^{\star 2} &= c^{-1}\left(|\Sigma_0|n/\gamma + \sigma_1^2\right) \\
\mu_\lambda^\star &= c^{-1}\left(n\bar{X}_n\phi\sigma_1\sigma_2 + \frac{n}{\gamma}\bar{Y}_n(|\Sigma_0|n + \sigma_2^2)\right) \\
\sigma_\lambda^{\star 2} &= c^{-1}\left(|\Sigma_0|n + \sigma_2^2\right)
\end{aligned}$$

Generally, the Kullback–Leibler divergence (KL-divergence) between two univariate Normal distributions $\mathcal{N}(\mu_\star, \sigma_\star^2)$ and $\mathcal{N}(\mu_0, \sigma_0^2)$ is

$$\mathcal{KL}\left(\mathcal{N}(\mu_\star, \sigma_\star^2)||\mathcal{N}(\mu_0, \sigma_0^2)\right) = -\frac{1}{2}\log\frac{\sigma_\star^2}{\sigma_0^2} + \frac{\sigma_\star^2 + (\mu_\star - \mu_0)^2}{2\sigma_0^2} - \frac{1}{2}.$$

We use this result to compare the marginal prior and posterior distributions for η and λ, but first we calculate some of the quantities needed.

$$\begin{aligned}
A_\eta &= \frac{\sigma_\eta^{\star 2}}{\sigma_1^2} = \frac{|\Sigma_0|n/\gamma + \sigma_1^2}{c\sigma_1^2} = \frac{|\Sigma_0|n}{c\sigma_1^2\gamma} + \frac{1}{c} \\
B_\eta &= \sigma_\eta^{\star 2} + \mu_\eta^{\star 2} = \frac{1}{c}\left(|\Sigma_0|n/\gamma + \sigma_1^2 + \frac{1}{c}\left(n\bar{X}_n(|\Sigma_0|n/\gamma + \sigma_1^2) + \frac{n}{\gamma}\bar{Y}_n\phi\sigma_1\sigma_2\right)^2\right) \\
A_\lambda &= \frac{\sigma_\lambda^{\star 2}}{\sigma_2^2} = \frac{|\Sigma_0|n + \sigma_2^2}{c\sigma_2^2} = \frac{n\sigma_1^2(1-\phi^2) + 1}{n^2|\Sigma_0|/\gamma + n\sigma_2^2/\gamma + n\sigma_1^2 + 1} \\
B_\lambda &= \sigma_\lambda^{\star 2} + \mu_\lambda^{\star 2} = \frac{1}{c}\left(|\Sigma_0|n + \sigma_2^2 + \frac{1}{c}\left(n\bar{X}_n\phi\sigma_1\sigma_2 + \frac{n}{\gamma}\bar{Y}_n(|\Sigma_0|n + \sigma_2^2)\right)^2\right)
\end{aligned}$$

The KL-divergence for the change from marginal prior to posterior for η and λ are therefore

$$\mathcal{KL}\left(p(\eta|\mathcal{D}_n)||p(\eta)\right) = -\frac{1}{2}\log A_\eta + \frac{B_\eta}{2\sigma_1^2} - \frac{1}{2}$$
$$\mathcal{KL}\left(p(\lambda|\mathcal{D}_n)||p(\lambda)\right) = -\frac{1}{2}\log A_\lambda + \frac{B_\lambda}{2\sigma_2^2} - \frac{1}{2}$$

The Effect of Independent Priors on Marginal Distributions

The mean and variance of the marginal distributions are

$$\mu_\eta^\star = \frac{n\sigma_1^2}{n\sigma_1^2+1}\bar{X}_n$$
$$\sigma_\eta^{\star 2} = \frac{\sigma_1^2}{n\sigma_1^2+1}$$
$$\mu_\lambda^\star = \frac{n\sigma_2^2/\gamma}{n\sigma_2^2/\gamma+1}\bar{Y}_n$$
$$\sigma_\lambda^{\star 2} = \frac{\sigma_2^2}{n\sigma_2^2/\gamma+1}.$$

The KL-divergences of the marginal prior-posterior comparisons in this special case is

$$\mathcal{KL}\left(p(\eta|\mathcal{D}_n)||p(\eta)\right) = \frac{1}{2}\log(n\sigma_1^2+1) + \frac{1}{2(n\sigma_1^2+1)} + \frac{n^2\sigma_1^2}{2(n\sigma_1^2+1)^2}\bar{X}_n^2 - \frac{1}{2}$$
$$\mathcal{KL}\left(p(\lambda|\mathcal{D}_n)||p(\lambda)\right) = \frac{1}{2}\log(n\sigma_2^2/\gamma+1) + \frac{1}{2(n\sigma_2^2/\gamma+1)} + \frac{n^2\sigma_1^2/\gamma^2}{2(n\sigma_2^2/\gamma+1)^2}\bar{Y}_n^2 - \frac{1}{2}$$

The Effect of Uninformative Data on Marginal Distributions

If we allow the data quality index to increase so that we get poorer and poorer data quality $\gamma \to \infty$ first have that

$$c \to n\sigma_1^2 + 1$$

so that

$$\mu_\eta^\star \to \frac{n\sigma_1^2}{n\sigma_1^2+1}\bar{X}_n, \qquad \sigma_\eta^{\star 2} \to \frac{\sigma_1^2}{n\sigma_1^2+1}$$
$$\mu_\lambda^\star \to \frac{n\phi\sigma_1\sigma_2}{n\sigma_1^2+1}\bar{X}_n, \quad \text{and} \quad \sigma_\lambda^{\star 2} \to \sigma_2^2\frac{n\sigma_1^2(1-\phi^2)+1}{n\sigma_1^2+1}$$

Here, ϕ controls the amount of marginal shrinkage. The KL-divergence is then

$$\begin{aligned}\mathcal{KL}\left(p(\eta|\mathcal{D}_n)||p(\eta)\right) \to &-\frac{1}{2}\log\frac{1}{n\sigma_1^2+1} + \frac{1}{2\left(n\sigma_1^2+1\right)} \\ &+ \frac{n^2\sigma_1^2}{2(n\sigma_1^2+1)^2}\bar{X}_n^2 - \frac{1}{2} \\ \mathcal{KL}\left(p(\lambda|\mathcal{D}_n)||p(\lambda)\right) \to &-\frac{1}{2}\log\left(\frac{n\sigma_1^2(1-\phi^2)+1}{n\sigma_1^2+1}\right) + \frac{n\sigma_1^2(1-\phi^2)+1}{2\left(n\sigma_1^2+1\right)} \\ &+ \frac{n^2\phi^2\sigma_1^2}{2(n\sigma_1^2+1)^2}\bar{X}_n^2 - \frac{1}{2}\end{aligned}$$

Conditional Distribution of $\eta|\lambda$

We derive the conditional posterior distribution for $\lambda|\eta$ since this is be examined in detail

$$\lambda|\mathcal{D}_n, \eta \sim \mathcal{N}\left(\mu_{\lambda|\eta}^\star, {\sigma_{\lambda|\eta}^\star}^2\right)$$

this has mean and variance

$$\mu_{\lambda|\eta}^\star = \mu_\lambda^\star + \frac{\phi\sigma_1\sigma_2}{|\Sigma_0|n/\gamma + \sigma_1^2}\left(\eta - \mu_\eta^\star\right)$$
$${\sigma_{\lambda|\eta}^\star}^2 = \sigma_\lambda^{\star 2} - \frac{\phi^2\sigma_1^2\sigma_2^2}{c\left(|\Sigma_0|n/\gamma + \sigma_1^2\right)}$$

We examine the KL divergence result to understand the evolution of the posterior when

$$\begin{aligned}
A_{\lambda|\eta} &= \frac{{\sigma^\star_{\lambda|\eta}}^2}{\sigma^2_{\lambda|\eta}} = \frac{|\Sigma_0|n + \sigma_2^2}{c(1-\phi^2)\sigma_2^2} - \frac{\phi^2\sigma_1^2\sigma_2^2}{c\left(|\Sigma_0|n/\gamma + \sigma_1^2\right)(1-\phi^2)\sigma_2^2} \\
&= \frac{|\Sigma_0|}{\left(|\Sigma_0|n/\gamma + \sigma_1^2\right)(1-\phi^2)\sigma_2^2} = \frac{\sigma_1^2\sigma_2^2(1-\phi^2)}{\left(|\Sigma_0|n/\gamma + \sigma_1^2\right)(1-\phi^2)\sigma_2^2} \\
&= \frac{\sigma_1^2}{|\Sigma_0|n/\gamma + \sigma_1^2} = \frac{1}{n\sigma_2^2(1-\phi^2)/\gamma + 1} \\
B_{\lambda|\eta} &= \frac{1}{\sigma^2_{\lambda|\eta}}\left(\mu^\star_{\lambda|\eta} - \mu_{\lambda|\eta}\right)^2 = \left(\mu^\star_\lambda + \frac{\phi\sigma_1\sigma_2}{|\Sigma_0|n/\gamma + \sigma_1^2}\left(\eta - \mu^\star_\eta\right) - \phi\frac{\sigma_2}{\sigma_1}\eta\right)^2 \\
&= \frac{1}{\sigma_2^2(1-\phi^2)}\left(\mu^\star_\lambda - \frac{\phi\sigma_1\sigma_2}{|\Sigma_0|n/\gamma + \sigma_1^2}\mu^\star_\eta + \left(\frac{\phi\sigma_1\sigma_2}{|\Sigma_0|n/\gamma + \sigma_1^2} - \phi\frac{\sigma_2}{\sigma_1}\right)\eta\right)^2 \\
&= \frac{1}{\sigma_2^2(1-\phi^2)}\left(\frac{|\Sigma_0|n/\gamma}{|\Sigma_0|n/\gamma + \sigma_1^2}\bar{Y}_n - \phi\frac{\sigma_2}{\sigma_1}\frac{|\Sigma_0|n/\gamma}{|\Sigma_0|n/\gamma + \sigma_1^2}\eta\right)^2
\end{aligned}$$

The KL-divergence can then be calculated as

$$\mathcal{KL}\left(p(\lambda|\mathcal{D}_n, \eta)||p(\lambda|\eta)\right) = -\frac{1}{2}\log A_{\lambda|\eta} + \frac{A_{\lambda|\eta}}{2} + \frac{B_{\lambda|\eta}}{2} - \frac{1}{2}.$$

The Effect of Independent Priors on the Conditional Distribution

When the priors are independent, $\phi = 0$, the conditional projections no longer have an effect, see formula for $\mu^\star_{\lambda|\eta}$ and ${\sigma^\star_{\lambda|\eta}}^2$ above. Furthermore, the marginal mean and variance formulas were calculated previously, so we have

$$\begin{aligned}
\mu^\star_{\lambda|\eta} &= \mu^\star_\lambda = \frac{n\sigma_2^2/\gamma}{n\sigma_2^2/\gamma + 1}\bar{Y}_n \\
{\sigma^\star_{\lambda|\eta}}^2 &= {\sigma^\star_\lambda}^2 = \frac{\sigma_2^2}{n\sigma_2^2/\gamma + 1}
\end{aligned}$$

The KL Divergence is the same as in the marginal case,

$$\begin{aligned}\mathcal{KL}\left(p(\lambda|\mathcal{D}_n,\eta)||p(\lambda|\eta)\right) &= \mathcal{KL}\left(p(\lambda|\mathcal{D}_n)||p(\lambda)\right)\\ &= \frac{1}{2}\log(n\sigma_2^2/\gamma+1)+\frac{1}{2(n\sigma_2^2/\gamma+1)}+\frac{n^2\sigma_2^2/\gamma^2}{2(n\sigma_2^2/\gamma+1)^2}\bar{Y}_n^2-\frac{1}{2}\end{aligned}$$

The Effect of Uninformative Data on the Conditional Distribution

If we allow the data quality index to increase so that we get poorer and poorer data quality $\gamma \to \infty$ we see that

$$\begin{aligned}\mu^{\star}_{\lambda|\eta} &\to \frac{n\bar{X}_n\phi\sigma_1\sigma_2}{n\sigma_1^2+1}+\frac{\phi\sigma_2}{\sigma_1}\left(\eta-\frac{n\bar{X}_n\sigma_1^2}{n\sigma_1^2+1}\right)\\ &= \eta\frac{\phi\sigma_2}{\sigma_1}+\frac{n\bar{X}_n\phi\sigma_1\sigma_2}{n\sigma_1^2+1}-\frac{n\bar{X}_n\phi\sigma_1\sigma_2}{n\sigma_1^2+1}=\eta\frac{\phi\sigma_2}{\sigma_1}\\ {\sigma^{\star}_{\lambda|\eta}}^2 &\to \frac{|\Sigma_0|n+\sigma_2^2}{n\sigma_1^2+1}-\frac{\phi^2\sigma_1^2\sigma_2^2}{\sigma_1^2(n\sigma_1^2+1)}\\ &= \frac{n\sigma_1^2\sigma_2^2(1-\phi^2)+\sigma_2^2}{n\sigma_1^2+1}-\frac{\phi^2\sigma_2^2}{n\sigma_1^2+1}\\ &= \frac{n\sigma_1^2\sigma_2^2(1-\phi^2)+\sigma_2^2(1-\phi^2)}{n\sigma_1^2+1}\\ &= \sigma_2^2(1-\phi^2)\frac{n\sigma_1^2+1}{n\sigma_1^2+1}=\sigma_2^2(1-\phi^2)\end{aligned}$$

Taking $\gamma \to \infty$, we have that $A_{\lambda|\eta} \to 1$ and $B_{\lambda|\eta} \to 0$ so that

$$\mathcal{KL}\left(p(\lambda|\mathcal{D}_n,\eta)||p(\lambda|\eta)\right) \to -\frac{1}{2}\log 1+\frac{1}{2}+\frac{1}{2\sigma^2_{\lambda|\eta}}0-\frac{1}{2}=0$$

A Bayesian Analysis of the Correlation Variation

Letting $n = m + k$ where we observe m paired observations and k from each marginal, the likelihood for

a dataset $\mathcal{D}_n = \left\{\{(X_i, Y_i)\}_{i=1}^m, \{X_i\}_{i=m+1}^n, \{Y_i\}_{i=m+1}^n\right\}$ can be written simply as

$$\begin{aligned}
L(\rho, \sigma^2; \mathcal{D}_n) &= \prod_{i=1}^{m} \frac{1}{2\pi\sigma^2\sqrt{(1-\rho^2)}} \exp\left(-\frac{\left(X_i^2 + Y_i^2 - 2\rho X_i Y_i\right)}{2(1-\rho^2)\sigma^2}\right) \\
&\quad \prod_{i=m+1}^{n} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{X_i^2}{2\sigma^2} - \frac{Y_i^2}{2\sigma^2}\right) \\
&= \frac{1}{(2\pi)^n \sigma^{2n} (1-\rho^2)^{m/2}} \exp\left(-\frac{1}{2(1-\rho^2)\sigma^2}\left(\sum_{i=1}^{m} X_i^2 + \sum_{i=1}^{m} Y_i^2 - 2\rho \sum_{i=1}^{m} X_i Y_i\right)\right. \\
&\quad \left. - \frac{1}{2\sigma^2}\left(\sum_{i=m+1}^{n} X_i^2 + \sum_{i=m+1}^{n} Y_i^2\right)\right)
\end{aligned}$$

After setting a prior for ρ and σ^2, we can conduct posterior analysis via numerical integration.

B.4 Measure of Information Conditioning on MAP estimate

In Section 3.4.2 we conciser measures of information in the Bayesian context where we condition on the true underlying data-generating parameter, something we cannot do in practice. Here, we show the same simulation results when conditioning on the MAP estimates. Figures B.1 and B.2 show the measures of information for ρ and σ^2 respectively when conditioning on the MAP estimates. The measures of direct information are the same, however the measures of indirect information are very different. Indeed, when conditioning on MAP point-estimates we severely underestimate the amount of indirect information. That is because we're injecting information information from the data twice, once for by conditioning on the data directly and again by conditioning on a data-driven estimate of the parameter.

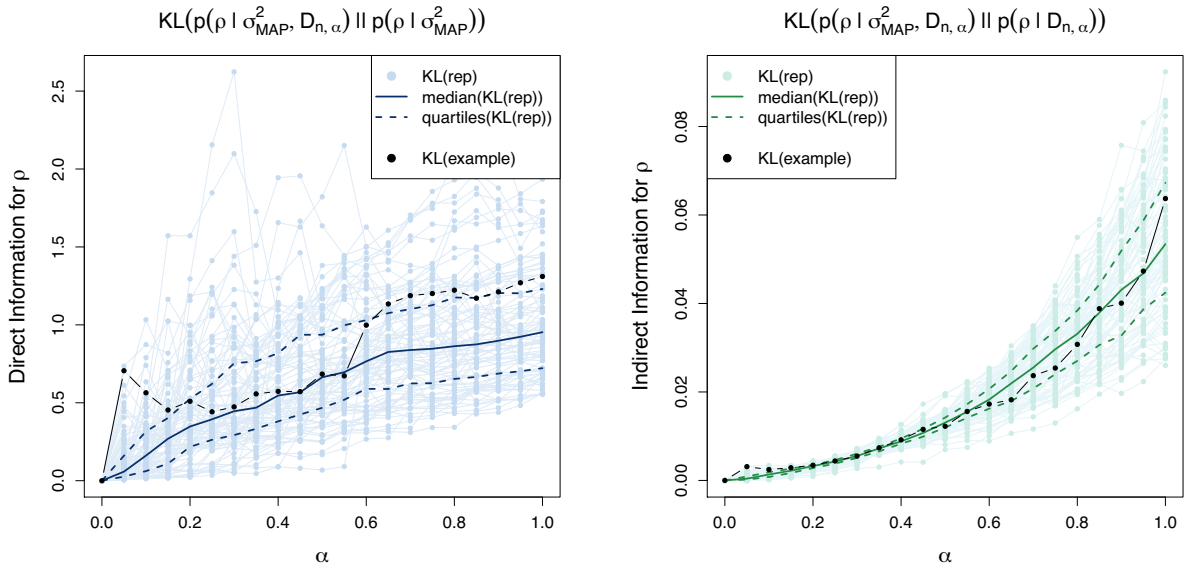


Figure B.1: Measures of direct and indirect information for ρ with replicates shown in colored dots. The mean and two quartiles shown in lines.

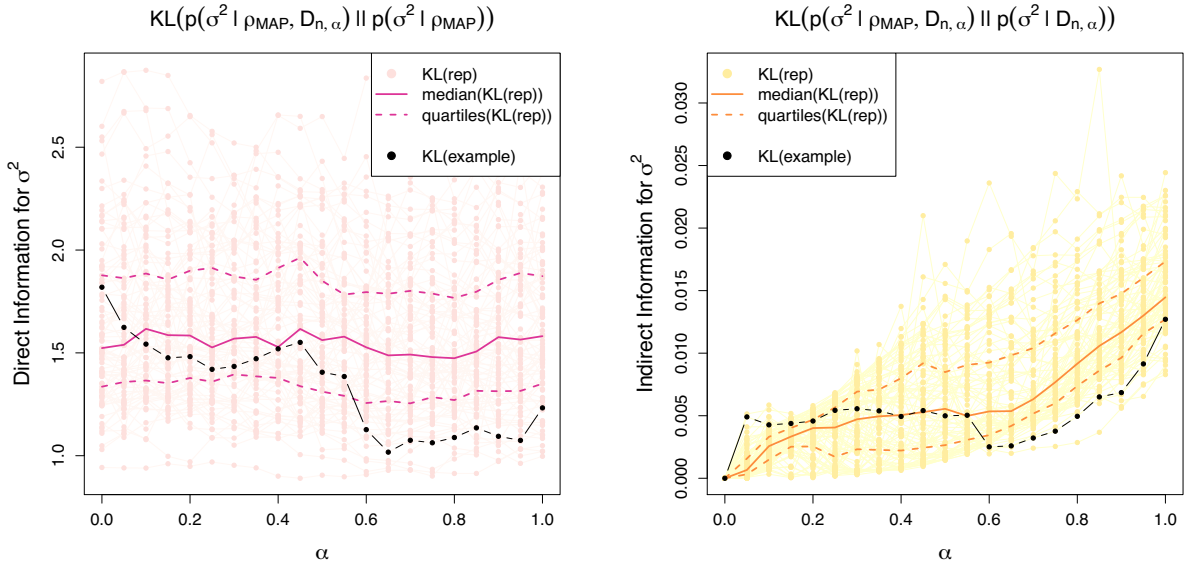


Figure B.2: Measures of direct and indirect information for σ^2 with replicates shown in colored dots. The mean and two quartiles shown in lines.

Bibliography

Amery, A., Brixko, P., Clement, D., De Schaepdryver, A., Fagard, R., Forte, J., Henry, J. F., Leonetti, G., O'Malley, K., & Strasser, T. (1985). Mortality and morbidity results from the European Working Party on High Blood Pressure in the Elderly trial. *The Lancet*, 325(8442), 1349–1354.

Andrieu, C., Doucet, A., & Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3), 269–342.

Benjamin, E. J., Virani, S. S., Callaway, C. W., Chamberlain, A. M., Chang, A. R., Cheng, S., Chiuve, S. E., Cushman, M., Delling, F. N., Deo, R., et al. (2018). Heart disease and stroke statistics—2018 update: a report from the american heart association. *Circulation*, 137(12), e67–e492.

Bhattacharya, A. & Wilson, S. P. (2018). Sequential bayesian inference for static parameters in dynamic state space models. *Computational Statistics & Data Analysis*, 127, 187 – 203.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518), 859–877.

Bloniarz, A., Liu, H., Zhang, C.-H., Sekhon, J. S., & Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27), 7383–7390.

Bohning, D., Dietz, E., Schlattmann, P., Mendonca, L., & Kirchner, U. (1999). The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 162(2), 195–209.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791–799.

Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multimodel Inference : A Practical Information-theoretic Approach.*, volume 2nd ed. Springer.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1), 1–32.

Choo, P. W., Rand, C. S., Inui, T. S., Lee, M.-L. T., Cain, E., Cordeiro-Breault, M., Canning, C., & Platt, R. (1999). Validation of patient reports, automated pharmacy records, and pill counts with electronic monitoring of adherence to antihypertensive therapy. *Medical care*, 37(9), 846–857.

Chopin, N., Jacob, P. E., & Papaspiliopoulos, O. (2012). SMC 2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3), 397–426.

Christ, C. (2001). Simultaneous equation estimation: Overview. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 14106 – 14110). Oxford: Pergamon.

Cox, D. R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(1), 1–39.

Dawid, P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 1–31).

de Carvalho, M., Page, G. L., & Barney, B. J. (2017). On the geometry of Bayesian inference. *arXiv.org*.

DeSantis, S. M., Lazaridis, C., Ji, S., & Spinale, F. G. (2014). Analyzing Propensity Matched Zero-Inflated Count Outcomes in Observational Studies. *Journal of Applied Statistics*, 41(1).

Durbin, J. & Koopman, S. J. (2001). *Time series analysis by state space methods*. Oxford; New York: Oxford University Press.

Fisher, Ronald Aylmer, S. .-. (1971). *The Design of Experiments*. New York :: Hafner Publishing Company, 8th edition.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3), 515–534.

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472.

Geweke, J. (2010a). *Complete and Incomplete Econometric Models*. Princeton University Press.

Geweke, J. (2010b). *Complete and Incomplete Econometric Models*. The Econometric and Tinbergen Institutes lectures. Princeton: Princeton University Press. OCLC: 659549748.

Gustafson, P. (2005). On Model Expansion, Model Contraction, Identifiability and Prior Information: Two Illustrative Scenarios Involving Mismeasured Variables. *Statistical Science*, 20(2), 111–140.

Gustafson, P. (2009). What Are the Limits of Posterior Distributions Arising From Nonidentified Models, and Why Should We Care? *Journal of the American Statistical Association*, 104(488), 1682–1695.

Gustafson, P. (2014). Bayesian inference in partially identified models: Is the shape of the posterior distribution useful? *Electronic Journal of Statistics*, 8(1), 476–496.

Gustafson, P. (2015). *Bayesian Inference for Partially Identified Models*. New York: Chapman and Hall/CRC.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1), 153–161.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945–960.

Homan, M. D. & Gelman, A. (2014). The No-U-turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1), 1593–1623.

Hsu, J.-Y., Chen, C.-C., & Ting, P.-F. (2018). Understanding mooc continuance: An empirical examination of social support theory. *Interactive Learning Environments*, 26(8), 1100–1118.

Héroux, J., Moodie, E. E., Strumpf, E., Coyle, N., Tousignant, P., & Diop, M. (2014). Marginal structural models for skewed outcomes: identifying causal relationships in health care utilization. *Statistics in Medicine*, 33(7), 1205–1221.

Imbens, G. W. & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Kadane, J. B. (1974). The Role of Identification in Bayesian Theory. *Studies in Bayesian Econometrics and Statistics*, (pp. 175–191).

Keele, L. & Miratrix, L. (2018). Randomization inference for outcomes with clumping at zero. *The American Statistician*, 0(0), 1–10.

Kressin, N. R., Orner, M. B., Manze, M., Glickman, M. E., & Berlowitz, D. (2010). Understanding contributors to racial disparities in blood pressure control. *Circulation: Cardiovascular Quality and Outcomes*, 3(2), 173–180.

Lam, W. Y. & Fresco, P. (2015). Medication adherence measures: an overview. *BioMed Research International*, 1, 1–12.

Lamb, A., Smilack, J., Ho, A., & Reich, J. (2015). Addressing common analytic challenges to randomized experiments in moocs: Attrition and zero-inflation. In *Proceedings of the Second (2015) ACM Conference on Learning @ Scale*, L@S '15 (pp. 21–30). New York, NY, USA: ACM.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1), 1–14.

Lee, J. K., Grace, K. A., & Taylor, A. J. (2006). Effect of a pharmacy care program on medication adherence and persistence, blood pressure, and low-density lipoprotein cholesterol: A randomized controlled trial. *Journal of the American Medical Association*, 296(21), 2563–2571.

Lee, M.-j. (2017). Extensive and intensive margin effects in sample selection models: racial effects on wages. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3), 817–839.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1), 295–318.

Lindley, D. V. (1956). On a Measure of the Information Provided by an Experiment. *The annals of mathematical statistics*, 27(4), 986–1005.

Lindley, D. V. (1987). *Bayesian Statistics a Review*. Society for Industrial and Applied Mathematics.

Little, R. J. & Rubin, D. B. (2002). *Statistical analysis with missing data. John Wiley & Sons*. New York.

Manski, C. F. (2003). *Partial Identification of Probability Distributions*. Springer-Verlag New York.

Meng, X.-L. & Xie, X. (2013). I Got More Data, My Model is More Refined, but My Estimator is Getting Worse! Am I Just Dumb? *Econometric Reviews*, 33(1-4), 218–250.

Miratrix, L. W., Sekhon, J. S., & Yu, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society. Series B*.

Moon, H. R. & Schorfheide, F. (2012). Bayesian and frequentist inference in partially identified models. *Econometrica*, 80(2), 755–782.

Morisky, D. E., Green, L. W., & Levine, D. M. (1986). Concurrent and predictive validity of a self-reported measure of medication adherence. *Medical care*, 24(1), 67–74.

Murray, L. M. (2015). Bayesian state-space modelling on high-performance hardware using libbi. *Journal of Statistical Software*, 67(10), 1–26.

Naranjo, A., Trindade, A. A., & Casella, G. (2013). Extending the State-Space Model to Accommodate Missing Values in Responses and Covariates. *Journal of the American Statistical Association*, 108(501), 202–216.

Nelimarkka, M. & Hellas, A. (2018). Social help-seeking strategies in a programming mooc. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, SIGCSE '18 (pp. 116–121). New York, NY, USA: ACM.

Osterberg, L. & Blaschke, T. (2005). Adherence to medication. *New England Journal of Medicine*, 353(5), 487–497.

Petris, G., Petrone, S., & Campagnoli, P. (2009). Dynamic linear models. In *Dynamic Linear Models with R* (pp. 31–84). New York, NY: Springer, New York, NY.

Poirier, D. J. (1998). Revising Beliefs in Nonidentified Models. *Econometric Theory*, 14(04), 483–509.

Probstfield, J. L. (1991). Prevention of stroke by antihypertensive drug-treatment in older persons with isolated systolic hypertension-final results of the Systolic Hypertension in the Elderly Program (SHEP). *Journal of the American Medical Association*, 265(24), 3255–3264.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., & Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15), 1923–1929.

Raue, A., Kreutz, C., Theis, F. J., & Timmer, J. (2012). Joining forces of Bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984), 20110544–20110544.

Rose, A. J., Glickman, M. E., D'Amore, M. M., Orner, M. B., Berlowitz, D., & Kressin, N. R. (2011). Effects of daily adherence to antihypertensive medication on blood pressure control. *The Journal of Clinical Hypertension*, 13(6), 416–421.

Rosenblum, M. & van der Laan, M. J. (2010). Simple, Efficient Estimators of Treatment Effects in Randomized Trials Using Generalized Linear Models to Leverage Baseline Variables. *The International Journal of Biostatistics*, 6.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701.

Rubin, D. B. (1980). Comment on 'Randomization analysis of experimental data in the Fisher randomization test' by Basu. *Journal of the American Statistical Association*, 75(371), 591–593.

Schnedler, W. (2005). Likelihood estimation for censored random vectors. *Econometric Reviews*, 24(2), 195–217.

Schroeder, K., Fahey, T., & Ebrahim, S. (2004). How can we improve adherence to blood pressure–lowering medication in ambulatory care? Systematic review of randomized controlled trials. *Archives of Internal Medicine*, 164(7), 722–732.

Staub, K. E. (2014). A causal interpretation of extensive and intensive margin effects in generalized tobit models. *The Review of Economics and Statistics*, 96(2), 371–375.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26(1), 24–36.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Vitolins, M. Z., Rand, C. S., Rapp, S. R., Ribisl, P. M., & Sevick, M. A. (2000). Measuring adherence to behavioral and medical interventions. *Controlled Clinical Trials*, 21, 188S–194S.

West, M. & Harrison, J. (1997). *Bayesian forecasting and dynamic models (2nd ed.)*. New York, NY, USA: Springer-Verlag New York, Inc.

Xie, Y. & Carlin, B. P. (2006). Measures of Bayesian learning and identifiability in hierarchical models. *Journal of Statistical Planning and Inference*, 136(10), 3458–3477.

Yau, K. K. W. & Lee, A. H. (2001). Zero-inflated poisson regression with random effects to evaluate an occupational injury prevention programme. *Statistics in Medicine*, 20(19), 2907–2920.