UNIVERSITY OF CALIFORNIA,
IRVINE

Causal Inference and Model Selection in Complex Settings

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Shandong Zhao

Dissertation Committee:
Professor David A. van Dyk, Co-Chair
Associate Professor Yaming Yu, Co-Chair
Associate Professor Scott Bartell

2014

© 2014 Shandong Zhao

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

ACKNOWLEDGMENTS

First of all, I would like to express my sincere appreciation to my PhD advisor Prof. David van Dyk. David has not only provided me guidance, continous support, and encouragement over my entire Ph.D study and research, but also set me a perfect example of an independent, diligent, and respectable researcher. His passion, professionalism, and humor makes him a great mentor and friend. I believe these six-year-long research experience under his supervison will be my priceless treasure in my life and set a solid foundation for my future career.

I would also like to thank Aneta Siemiginowska and Vinay Kashyap. Both of you have provided me enormous help to understand the Astrophysics from the very basic background of the universe to the most advanced theory with regard to supernova detection. Without your assistance, I cannot imagine I would have a chance to introduce a methodology for astronomers helping them select the proper statistical tool for model selection. My daughter will learn your names when she hear me telling the stories behind the sky. She might feel bored with those terminologies but I am sure I will feel my enthusiasm again.

My thanks also goes to Kosuke Imai, James King, and Yaming Yu. I was so lucky to have the chance collaborating with all of you. We might not have a lot of chance working together but you sure all have left marks on my research career. I would also thank my fellow friends, Jin Xu, Jian Liang, Jing Liu, and David Stenning. I have spent most of my Ph.D career with you. You all witness my success, frustration, and endeavor. Thanks so much for your support and I wish you great success on your future career and life.

Mostly, to my wife Chunting Gu and my little angel Allison. You two are the best gift I ever have. You give me a warm and wonderful home; You show me hope, passion, and power to conquer any obstacle; You make me believe I will never walk alone.

Last but not least, this work was funded in part by NSF grants SES-05-50980, DMS-09-07522, and DMS-12-09232. University of California, Irvine also provided substantial funding support to my Ph.D research.

CURRICULUM VITAE

Shandong Zhao

EDUCATION

Doctor of Philosophy in Statistics **2014**
University of California, Irvine *Irvine, CA*

Bachelor of Science in Mathematics **2008**
Shanghai Jiaotong University *Shanghai, China*

RESEARCH EXPERIENCE

Graduate Research Assistant **2008–2014**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **2009–2014**
University of California, Irvine *Irvine, California*

ABSTRACT OF THE DISSERTATION

Causal Inference and Model Selection in Complex Settings

By

Shandong Zhao

Doctor of Philosophy in Statistics

University of California, Irvine, 2014

Professor David A. van Dyk, Co-Chair

Associate Professor Yaming Yu, Co-Chair

Propensity score methods have become a part of the standard toolkit for applied researchers who wish to ascertain causal effects from observational data. While they were originally developed for binary treatments, several researchers have proposed generalizations of the propensity score methodology for non-binary treatment regimes. In this article, we firstly review three main methods that generalize propensity scores in this direction, namely, inverse propensity weighting (IPW), the propensity function (P-FUNCTION), and the generalized propensity score (GPS), along with recent extensions of the GPS that aim to improve its robustness. We compare the assumptions, theoretical properties, and empirical performance of these methods. We propose three new methods that provide robust causal estimation based on the P-FUNCTION and GPS. While our proposed P-FUNCTION-based estimator preforms well, we generally advise caution in that all available methods can be biased by model misspecification and extrapolation. In a related line of research, we consider adjustment for posttreatment covariates in causal inference. Even in a randomized experiment, observations might have different compliance performance under treatment and control assignment. This posttreatment covariate cannot be adjusted using standard statistical methods. We review the principal stratification framework which allows for modeling this effect as part

of its Bayesian hierarchical models. We generalize the current model to add the possibility of adjusting for pretreatment covariates. We also propose a new estimator of the average treatment effect over the entire population.

In a third line of research, we discuss the spectral line detection problem in high energy astrophysics. We carefully review how this problem can be statistically formulated as a precise hypothesis test with point null hypothesis, why a usual likelihood ratio test does not apply for problem of this nature, and a doable fix to correctly quantify the p-value using the likelihood ratio test statistic via posterior predictive p-values. However, as p-values (including posterior predictive p-values) tend to overstate the evidence for the alternative hypothesis for precise hypothesis testing, we review a Bayesian alternative method to do the line detection problem using the Bayes factor. Although Bayes factors are generally criticized to be sensitive to the choice of prior distributions, we show that such prior dependence can reflect different scientific questions and thus be sensible. In fact, p-values have similar "subjective influence" in that testing for the existance of a line at a fixed location or in an area with broad range can lead to very different conclusions. This is usually known as the look elsewhere effect in astrophysics.

Chapter 1

Background

In this Chapter, we discuss the background of the three scientific problems studied in this thesis; proposing two robust propensity score based methods to do causal inference for fixed non-binary treatment in observational studies; generalizing the principal stratification framework to adjust for the influence of different compliance performance under treatment and control group in a randomized experiment; and a careful study of the sensitivity of the Bayes Factors to the choice of prior distributions when testing for spectral lines in high energy astrophysics of Chapter 2, 3, and 4. A more complete review of the relevant literature for each of the three topics is included in the early sections of each chapter.

1.1 Causal Inference with Propensity Score

Accessing the causal effect from observational studies is a challenging task. The problem is that without randomization, you cannot expect balance for any covariate so that you have to adjust for all of them. Doing so using a parametric model like regression is prone to misspecification while doing so non-parametrically (e.g, matching) is difficult because of

the curse of dimensionality. When the treatment is binary, the propensity score methods of Rosenbaum and Rubin (1983) solved this fundamental problem by reducing the dimension, so you can match or subclassify. They show that under the assumption of unmeasured confounding, adjusting for the one-dimensional propensity score, rather than potentially high-dimensional covariates, is sufficient for unbiased estimation of causal effect.

Despite their popularity, the original propensity score method is only applicable to a binary treatment. Over the past decade, several methods have been proposed to allow the application of the propensity score for non-binary treatment regimes. Among these, we review the three primary generalizations, namely, inverse probability weighting (Robins *et al.*, 2000a) (RHB), propensity function (Imai and van Dyk, 2004) (IvD), and the generalized propensity score (Hirano and Imbens, 2004) (HI) (including two of its extensions (Flores *et al.*, 2012) (FFGN)). While HI's method is designed to estimate the dose response function (DRF), IvD estimates the average treatment effect (RHB can be used with a variety of estimands). We introduce an extended method based on IvD that is capable of making robust estimation for the full DRF.

We review the assumptions and theoretical properties of these methodologies when used to estimate a DRF and compare their empirical performance via a series of simulation studies, including one that is based on a real dataset. We demonstrate that the response model used by HI is less flexible than those typically used with propensity score methods and the methods proposed by FFGN to address this probelm can exhibit undesirable properties. We then propose a robust variant of HI's method. In summary, we find estimating the full DRF with a continuous treatment in an observational study is challenging, as all available methods can be biased by model misspecification and extrapolation. Researchers should be cautious when making such causal estimate.

1.2 Generalized Principal Stratification Framework

While pretreatment covariates adjustment is a challenging task for observational studies (see section 1.1), coping with posttreatment covariate such as the compliance rate can be more demanding even in a double blind randomized experiment. For example, participants might not always take designated dose of either drug or placebo due to either side-effect or preference. This covariate cannot be adjusted using standard statistical methods such as regression, as the treatment might have an effect on the posttreatment variables (Frangakis and Rubin, 2002).

We encountered problem of this type when evaluating the effectiveness of the READ 180 program based on a recent large implementation study conducted in high-poverty school district located in southeastern Massachusetts (Hartry *et al.*, 2008). The READ 180 program is a mixed-method approach designed to help struggling readers in grades 4 to 12. In the implementation study, participating students are randomly assigned to either the READ 180 program or a district after-school program which serves as a control group. Kim *et al.* (2010) conducted ANCOVA on the vocabulary measure using relevant pretreatment information as covariates and no significant difference were found between the two groups. However, they found the attendance rate (defined as the percentage of attended days) were significantly higher for students in the READ 180 group than students in the district after-school group.

To properly adjust for the influence of the attendance rate, we regard it as the compliance variable and apply the principal stratification method of Jin and Rubin (2008) for causal effect estimation. Under the principal stratification framework, each observation has two potential compliance rate: one for the treatment and one for control. These two compliances are considered to be fixed and uninfluenced by the treatment in the same way as such pretreatment covariates as age and gender. Hence, causal effect can be estimated for observations with same compliance (defined as principal strata). However, in practice, part of the

compliance for any observation is always missing and we can only observe his compliance to the group he is assigned to. To solve this problem, Jin and Rubin model compliance as well as the response variable conditioned on the compliance using Bayesian hierarchical models. They use hypothetical data points with complete compliance information (computed under certain assumptions) as the prior for the compliance parameters.

We review and select valid assumptions of the principal stratification method for the READ 180 analysis, and propose two generalizations to its current framework. First, we allow adjusting for pretreatment covariates in the conditional model of the response variable given the compliance information. Accordingly, we design two estimators to access the treatment effect within each principal stratum after the pretreatment adjustment. Secondly, we propose a method to compute the average treatment effect over the entire population, which can be regarded as the weighted average of the within principal strata treatment effect where the weights are equal to the density of each principal stratum. Lastly, we introduced a way to do sensitivity analysis for the potentionally influential hypothetical complete-data prior distribution. After adjusting for the attendance rate, we find that although the treatment effect is still not significant in all of the principal strata, the average treatment effect over the entire population does appear significant.

1.3 Bayes Factor for Model Selection in Astrophysics

Distinguishing a faint spectral line from a chance fluctuation in data with low photon counts is a challenging problem in high energy astrophysics. Statistically, it can be thought of as a test for the presence of a component in a finite-mixture distribution. In particular, it falls into the category of precise hypothesis testing as the null hypothesis specifies a point value of zero for the intensity of the spectral line (Berger and Delampady, 1987). Unforturnately, the common routine to quantify the evidence via computing a likelihood ratio test (LRT)

statistic and calibrating it according to its nominal asymptotic distribution does not apply for this problem as the standard regularity conditions required for the asymptotic theory are not satisfied. Protassov *et al.* (2002) provide a detailed discussion of this problem and propose a doable fix using posterior predictive p-values (ppp-value), which bypass the asymptotic theory of the LRT, find its posterior predictive distribution empirically, and then calculate the correctly-calibrated p-value. However, since the ppp-value shares a similar definition and interpretation of the classical p-value, compared to the posterior probability of the null hypothesis, it aslo tends to overstate the evidence of the more complicated model when used for testing precise hypotheses (Berger and Delampady, 1987; Berger and Sellke, 1987).

Bayes factors, on the other hand, provide a principled alternative summary statistic for doing model selection. However, they are criticized for being "subjective" in that decisions based on Bayes factors can be sensitive to the choices of the prior distribution. We carefully study such prior influence for each of the model parameters in a simple yet popular class of spectral line detection problems via both simulation and real data analysis. We find that the prior influence can actually be interpreted in a non-subjective manner. Different priors can reflect different scientific questions such as where and how strong of a spectral line astronomers are looking for. Moreover, p-values are also prone to similar subjective influence in that testing for the existance of a spectral line at a known location versus at an unknown location within a certain energy range will lead to significantly different p-values. This is typically known as the look elsewhere effect in astronomy and pyhsics (Gross and Vitells, 2010). Overall the Bayes Factor is usually more conservative for detecting the spectral line compared to the (overstated) p-value (as well as ppp-values).

We find that for the spectral line detection problem, the prior distributions for the emission line parameters are expected to be much more important. We suggest plotting the decision boundary based on the Bayes factor as a function of the hyper-parameters for these prior distributions. We then compare the set of priors that vote for (i) model with a spectral line;

(ii) model without a spectral line; (iii) indifference between the two models. If all reasonable priors correspond to one of these three sets, the model selection is completed. Otherwise, we cannot clearly enunciate the outcome of the comparison and need to state how the outcome depends on the choice of the prior distributions.

Chapter 2

Using Propensity Score based Methods for Fixed Non Binary Treatments

2.1 Introduction

Adjusting for observed confounding variables is one of the most common strategies used across numerous scientific disciplines when making causal infererence in observational studies. Researchers find that the results based on regression adjustments can be sensitive to model specification when applied to the data where the treatment and control groups differ substantially in terms of their pre-treatment covariates. The propensity score methods of Rosenbaum and Rubin (1983), hereafter RR, aim to address this fundamental problem by reducing the covariate imbalance between the two groups. RR showed that under the assumption of no unmeasured confounding, adjusting for the propensity score, rather than potentially high-dimensional covariates, is sufficient for unbiased estimation of causal effects

and this can be done by simple nonparametric methods such as matching and subclassification.

Despite their popularity, one limitation of the original propensity score methods is that they are only applicable to a binary treatment. About a decade ago, several researchers proposed generalization of the propensity score methodology for non-binary treatment regimes (Robins, Hernán, and Brumback, 2000a; Imbens, 2000; Hirano and Imbens, 2004; Imai and van Dyk, 2004). Such extensions have widened the applicability of propensity score methods and are indeed becoming increasingly popular themselves, Google Scholar citation counts of the aforementioned papers are 1174, 800, 240, and 318, respectively, as of Oct 29, 2013).

All of these methods, however, require users to overcome the challenges of first correctly modeling a treatment variable as a function of a possibly large number of pre-treatment covariates and second modeling the response variable. These represent significant difficulties in practice. Standard diagnostics based on the comparison of the covariate distributions between the treatment and control groups are not directly applicable to non-binary treatment regimes and the final inference can be quite sensitive to the choice of response model. Flores *et al.* (2012), hereafter FFGN, propose two extensions to the method of Hirano and Imbens (2004) that aim to provide more robust estimation through a move flexible response model.

In this article, we closely examine the three primary propensity score-based methods for causal inference with non-binary treatements, namely, inverse probability weighting (IPW) of Robins, Hernán, and Brumback (2000a), hereafter RHB; the propensity function (P-FUNCTION) of Imai and van Dyk (2004), hereafter IvD; and the generalized propensity score (GPS) of Hirano and Imbens (2004), hereafter HI, along with the FFGN extensions. We compare the assumptions and theoretical properties of these alternative methodologies when used to estimate a dose response function (DRF) and examine their empirical performance in practice. Our primary message is cautionary: estimating the full DRF with a continuous treatment in an observational study is challenging. Researchers should be cautious when

attempting to do so.

G-estimation (Robins *et al.*, 1992) is also based on propensity scores, in that it involves the distribution of the treatment given observed covariates. Although it can be used to estimate the effect of time-varying treatments, it is not commonly used with other non-binary treatments. Both because of this and because g-estimation is methodologically quite distinct from the other methods we consider, we do not include it in our detailed comparisons.

The GPS and particularly the IPW methods enjoy wide application beyond the estimation of a DRF. Like g-estimates, they can be used, for example, to estimate the effect of time-varying treatments and with longitudinal data (e.g., Hernán *et al.*, 2002; Hogan and Lancaster, 2004; Moodie and Stephens, 2012); Ertefaie and Stephens (2010) provides a comparison of the two approaches in this setting. IPW is a foundational statistical method that dates back at least to Horvitz and Thompson (1952); its application is endemic to survey sampling, missing-data methods, and causal inference; reviews include Schafer and Kang (2008), Seaman and White (2011), and Stuart (2010). Likewise, the methods of IvD and HI each formalize a framework for inference, but are based on ideas that appear earlier in the literature (Joffe and Rosenbaum, 1999; Lu *et al.*, 2001; Rosenbaum, 1987; Imbens, 2000). In this article we focus on the estimation of the DRF in an observational study in which both the covariates and treatment are fixed, as outlined in HI. This is the simplest setting beyond estimating an average treatment effect and is a well-defined testing ground for the clear comparison of available methods. As we shall see, even in this simple setting standard methods may exhibit unacceptable statistical properties.

Section 2.2 reviews the theoretical properties of the original propensity score methodology and its interrelated generalizations. The GPS, for example, is closely related to the probability weight of IPW, but without stabilized weights (Robins, 1998, 1999); whereas the P-FUNCTION uses a completely different quantification of the propensity for treatment. While the GPS is designed to estimate the DRF, the P-FUNCTION estimates the average

treatment effect, and IPW can be used with a variety of estimands. In Section 2.3, we compare the methods of RHB, IvD, HI, and FFGN both theoretically and empirically. We demonstrate that the response model used by HI is less flexible than those typically used with propensity score methods and that the methods proposed by FFGN to address this problem can exhibit undesirable properties. We also show that one of FFGN's methods can be improved by using the stabilized weights of RHB, effectively implementing IPW to estimate the full DRF. In Section 2.4, we compare these methods with a new proposal and show how the method of IvD can also be extended for robust estimation of the full DRF. The efficacy of the proposed methodology is illustrated through simulation studies in Section 2.4 and an empirically-based study in Section 2.6. Section 2.7 offers concluding remarks and Appendix 2.5 introduces a robust variant of HI's method.

2.2 Methods for Estimating the DRF

Suppose we have a simple random sample of size n with each unit consisting of a p-dimensional column vector of pretreatment covariates, $\boldsymbol{X}_i$, the observed univariate treatment, T_i, and the outcome variable, Y_i. Although IvD's method can be applied to multivariate treatments, here we assume the treatment is univariate to facilitate comparison with the methods of RHB and HI. We omit the subscript when referring to generic values of $\boldsymbol{X}_i$, T_i, and Y_i.

We denote the potential outcomes by $\mathcal{Y} = \{Y_i(t), t \in \mathcal{T} \text{ for } i = 1, \ldots, n\}$, where $\mathcal{T}$ is a set of possible treatment values and $Y_i(t)$ is a function that maps a particular treatment level of unit i, to its outcome. This setup implies the *stable unit treatment value assumption* (Rubin, 1990) that the potential outcome of each unit is not a function of treatment level of other units and that the same version of treatment is applied to all units. In addition, we assume *strong ignorability of treamtent assignment*, i.e., $Y(t) \perp\!\!\!\perp T \mid \boldsymbol{X}$ and $p(T = t \mid \boldsymbol{X}) > 0$ for all

$t \in \mathcal{T}$, which implies no unmeasured confounding (RR).

2.2.1 The propensity score with a binary treatment

RR considered the case of treatment variables that take on only two values, $\mathcal{T} = \{0, 1\}$, where $T_i = 1$ ($T_i = 0$) implies that unit i receives (does not receive) the treatment and defined the *propensity score* to be the conditional probability of assignment to treatment given the observed covariates, i.e., $e(\boldsymbol{X}) = p(T = 1 \mid \boldsymbol{X})$. In practice, $e(\boldsymbol{X})$ is typically estimated using a parametric treatment assignment model $p_\psi(T = 1 \mid \boldsymbol{X})$ where ψ is a vector of unknown parameters. The appropriateness of the fitted model can be assessed via the celebrated balancing property of $e(\boldsymbol{X})$, namely, that covariates should be independent of the treatment conditional on the propensity score, $\boldsymbol{X} \perp\!\!\!\perp T \mid e(\boldsymbol{X})$. In particular, the fitted model, $\hat{e}(\boldsymbol{X}) = p_{\hat{\psi}}(T = 1 \mid \boldsymbol{X})$ should not be accepted unless adjusting for $\hat{e}(\boldsymbol{X})$ results in adequate balance.

In order to estimate causal quantities, we must properly adjust for $\hat{e}(\boldsymbol{X})$. RR propose three techniques: matching, subclassification, and covariance adjustment. We focus on subclassification and covariance adjustment because they are more closely related to the generalizations for non-binary treatments. The key advantage of propensity scores when applying these methods, and the inverse weighting method discussed below, is dimension reduction. They only require adjustment for a scalar variable $\hat{e}(\boldsymbol{X})$ rather than for the entire covariate vector.

With subclassification (RR), we adjust for $\hat{e}(\boldsymbol{X})$ by dividing the observations into several subclasses based on $\hat{e}(\boldsymbol{X})$. Individual response models are then fitted within each subclass, adjusting for $\hat{e}(\boldsymbol{X})$ and sometimes $\boldsymbol{X}$ along with T. The overall causal effect is then computed as the weighted average of the within-class coefficients of T, with weights proportional to the size of subclass. The standard error of the causal effect is typically computed by treating

the within-subclass estimates as independent of one another.

With covariance adjustment (RR), we regress the response variable on $\hat{e}(\boldsymbol{X})$ separately for the treatment and control groups. Specifically, we divide the data into the treatment and control groups and fit the regression model, $\mathrm{E}(Y \mid \boldsymbol{X}, T = t) = \alpha_t + \beta_t \cdot \hat{e}(\boldsymbol{X})$, seperately for $t = 0, 1$. The average causal effect is then estimated as

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{\alpha}_1 + \hat{\beta}_1 \cdot \hat{e}(\boldsymbol{X}_i) - \hat{\alpha}_0 - \hat{\beta}_0 \cdot \hat{e}(\boldsymbol{X}_i)\Big) = (\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\beta}_1 - \hat{\beta}_0) \cdot \overline{\hat{e}(\boldsymbol{X})} \tag{2.1}$$

where $\overline{\hat{e}(\boldsymbol{X})}$ is the sample mean of the estimated propensity score.

In addition to the techniques in RR, inverse propensity score weighting can be used to estimate causal quantities (e.g., Rosenbaum, 1987; Robins, 1998; Robins *et al.*, 2000a; Imbens, 2000). Because the equalities

$$\mathrm{E}\left\{\frac{TY}{e(\boldsymbol{X})}\right\} = \mathrm{E}\{Y(1)\} \quad \text{and} \quad \mathrm{E}\left\{\frac{(1-T)Y}{1-e(\boldsymbol{X})}\right\} = \mathrm{E}\{Y(0)\},$$

hold, the inverse weighting estimate,

$$\sum_{i=1}^{N}\left(\frac{T_i Y_i}{\hat{e}(\boldsymbol{X}_i)} - \frac{(1-T_i)Y_i}{1-\hat{e}(\boldsymbol{X}_i)}\right)$$

is an unbiased estimate of the average causal effect. Because $\hat{e}(\boldsymbol{X}_i)$ may vary greatly with $\boldsymbol{X}_i$, its recipoical can be very unstable (e.g., RHB, Kang and Schafer, 2007, Sections 2.3.1–2.3.2). RHB show that replacing the weight $1/\hat{e}(\boldsymbol{X}_i)$ by the stabilized weight, $p_{\hat{\psi}_0}(T_i)/\hat{e}(\boldsymbol{X}_i)$, can substantially mitigate this instability, where $p_{\hat{\psi}_0}(T_i)$ is a parameterized model for the marginal distribution of T. A similar stabilization uses $\frac{1}{\hat{e}(\boldsymbol{X}_i)} / \sum_{i=1}^{n} \frac{T_i}{\hat{e}(\boldsymbol{X}_i)}$ for the treatment group (Hirano *et al.*, 2003). Robins (1999) proposes a doubly robust augmented IPW estimator which can be more efficient than IPW and can protect against misspecification of the treatment model, but according to RHB can be more involved to implement, see also Wang

et al. (2007).

Matching, subclassification, covariance adjustment, and (stabilized) inverse weighting all aim to provide robust flexible adjustment for $\hat{e}(\boldsymbol{X})$ in the response model. As we shall see below, the flexibility of the response model is important especially for non-binary treatment regimes. This is because unlike the treatment assignment model which has an effective diagnostic tool based on the balancing property of propensity scores, the response model lacks such diagnostics.

2.2.2 Propensity score: methods for non-binary treatments

Suppose now that $\mathcal{T}$ is a more general set of treatment values, perhaps categorical or continuous. It is in this setting that RHB extended IPW, that IvD introduced the P-FUNCTION and that HI introduced the GPS. (IvD also allow for multi-variate treatments, which we do not discuss in this paper.) In what follows, we review and compare these generalizations of propensity score methods. In particular, we consider the following aspects of propensity score adjustment with binary treatments that RHB, IvD, and HI all generalize:

1. Treatment assignment model: Model the distribution of the treatment assignment given covariates to estimate the propensity score, i.e., $\hat{e}(\boldsymbol{X})$
2. Diagnostics: Validate $\hat{e}(\boldsymbol{X})$, by checking for covariate balance, i.e., $T \perp\!\!\!\perp \boldsymbol{X} \mid \hat{e}(\boldsymbol{X})$
3. Response model: Model the distribution of the response given the treatment, adjusting for $\hat{e}(\boldsymbol{X})$ via matching, subclassification, covariance adjustment, or IPW
4. Causal quantities of interest: Estimate the causal quantities of interest and their standard error based on the fitted response model

2.2.1 Treatment assignment model. As in the case of the binary treatment, we begin by modeling the distribution of the observed treatment assignment given the covariates using a parametric model, $p_\psi(T \mid \boldsymbol{X})$, where ψ is a set of parameters. Common choices of $p_\psi(T|\boldsymbol{X})$ include the Gaussian or multinomial regression models when the treatment variable is continuous or categorical, respectively. HI define the GPS as $R = r(T, \boldsymbol{X}) = p_\psi(T \mid \boldsymbol{X})$. That is, the GPS is equal to the treatment assignment model density *evaluated* at the observed treatment variable and covariate for a particular individual. This is analogous to the propensity score for the binary treatment, which can be written as $e(\boldsymbol{X}) = r(1, \boldsymbol{X}) = p_\psi(T = 1 \mid \boldsymbol{X})$.

Before HI coined the term GPS, the same quantity was used by RHB in IPW. In particular, RHB considered using weights equal to $1/r(T, \boldsymbol{X})$ and then, noting the instability of these weights, suggested using the stabilized weights $W = W(T, \boldsymbol{X}) = p_{\psi_0}(T)/p_\psi(T \mid \boldsymbol{X})$ instead. For example, if we use a normal linear model for $p_\psi(T \mid \boldsymbol{X})$, i.e., $T_i \mid \boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{X}_i^T\beta, \sigma^2)$, we might use a normal model for $p_{\psi_0}(T)$, i.e., $T_i \sim \mathcal{N}(\mu, \tau^2)$. Although RHB first proposed the quantity, we use HI's now standard term, GPS, for $r(T, \boldsymbol{X})$.

IvD summarize $p_\psi(T \mid \boldsymbol{X})$ in a manner that is qualitatively different from RHB and HI. First, they define the P-FUNCTION to be the entire conditional density (or mass) function of the treatment, namely $e_\psi(\cdot \mid \boldsymbol{X}) = p_\psi(\cdot \mid \boldsymbol{X})$. This is also analogous to the propensity score for the binary treatment case because $e_\psi(\cdot \mid \boldsymbol{X})$ is completely determined by $e(\boldsymbol{X}) = p_\psi(T = 1 \mid \boldsymbol{X})$. In order to summarize the P-FUNCTION, IvD introduce the *uniquely parameterized propensity function* assumption which states that for every value of $\boldsymbol{X}$, there exists a unique finite-dimensional parameter, $\boldsymbol{\theta} \in \Theta$, such that $e_\psi(\cdot \mid \boldsymbol{X})$ depends on $\boldsymbol{X}$ only through $\boldsymbol{\theta}_\psi(\boldsymbol{X})$. In other words, $\boldsymbol{\theta}$ uniquely represents $e\{\cdot \mid \boldsymbol{\theta}_\psi(\boldsymbol{X})\}$, which we may therefore write as $e(\cdot \mid \boldsymbol{\theta})$ or simply $\boldsymbol{\theta} = \boldsymbol{\theta}_\psi(\boldsymbol{X}_i)$. For example, if we model the treatment, $T_i \sim \mathcal{N}(\boldsymbol{X}_i^\top\boldsymbol{\beta}, \sigma^2)$ with $\psi = (\boldsymbol{\beta}, \sigma^2)$, the scalar $\boldsymbol{\theta}_i = \boldsymbol{X}_i^\top\boldsymbol{\beta}$ uniquely represents $e_\psi(\cdot \mid \boldsymbol{X}_i)$. In practice, ψ, ψ_0, $\boldsymbol{\theta}_i$, W_i, R_i, and r_i are estimated from data; we denote their estimates $\hat{\psi}$, $\hat{\psi}_0$,

$\hat{\boldsymbol{\theta}}_i$, $\hat{W}_i$, $\hat{R}_i$, and $\hat{r}_i$.

2.2.2 Diagnostics. Diagnostics for the treatment assignment model rely on balancing properties of the IPW, P-FUNCTION and GPS. For example, IvD shows that the P-FUNCTION is a balancing score, i.e., $T \perp\!\!\!\perp \boldsymbol{X} \mid e(\cdot \mid \boldsymbol{\theta})$. IvD suggest checking balance by regressing each covariate on T and $\hat{\boldsymbol{\theta}}$, e.g., using Gaussian and/or logistic regression and comparing the distribution of the t-statistics for each of the resulting regression coefficients of T with the standard normal distribution via a normal quantile plot. Improvement in balance can be assessed by constructing the plot again in the same manner except that $\hat{\boldsymbol{\theta}}$ is left out of each regression. Although not typical used, this diagnostic is equally applicable in the binary treatment case.

HI, on the other hand, show that $\mathbf{1}\{T = t\}$ is independent of $\boldsymbol{X}$ given $r(t, \boldsymbol{X})$, where $\mathbf{1}\{\cdot\}$ is an indicator function and the GPS is evaluated at $t \in \mathcal{T}$. Following the covariate balancing property for the binary propensity score, HI construct a series of binary treatments by coarsening the original treatment T in the form $\{t_j < T \leq t_{j+1}\}$ for some $t_1, t_2, \ldots, t_J$. Covariate balance is then checked for these binary treatment variables by first subclassifying units on $\hat{r}(\widetilde{T}_j, \boldsymbol{X})$, where $\widetilde{T}_j$ is the median of the treatment variable among units with $\mathbf{1}\{t_j < T \leq t_{j+1}\} = 1$. Then, two-sample t-tests are performed within each subclass to compare the mean of each covariate among units with $\mathbf{1}\{t_j < T \leq t_{j+1}\} = 0$ against that among units with $\mathbf{1}\{t_j < T \leq t_{j+1}\} = 1$. Finally the within-subclass differences in means and the variances of these differences are combined to compute a single t-statistic for each covariate. HI suggest repeating this diagnostics for several choices of $\{t_1, \ldots, t_J\}$ that cover the range of observed T.

With IPW T and $\boldsymbol{X}$ constitute a weighted sample from a population where T and $\boldsymbol{X}$ are independent. Thus, if the weights are correctly specified, the weight-adjusted T and $\boldsymbol{X}$ should be consistent with independence. This can be checked, for example, by computing weighted

correlations or multiplying counts of discrete or discretized variables by their weights, see RHB.

Because the same treatment models are used with all three methods the diagnostics for each may be used for all. In any case, failure to reject the null hypothesis of perfect balance does not imply balance and hence the diagnostics must be interpreted carefully. In fact, a small (within subclass) sample size, may limit the ability to detect a lack of balance (Imai *et al.*, 2008).

2.2.3 Response model. The response models proposed by RHB, IvD, and HI are quite different, with HI relying heavily on parametric assumptions. IvD propose two response models. The first is completely analogous to the subclassification technique proposed by RR. Individual response models are fitted within each subclass, adjusting for $\hat{\boldsymbol{\theta}}$ and typically $\boldsymbol{X}$ along with T. The second is a smooth coefficient model (SCM), which allows the intercept and slope to vary smoothly as a function of the P-FUNCTION

$$E(Y \mid T, \hat{\boldsymbol{\theta}}) \;=\; f(\hat{\boldsymbol{\theta}}) + g(\hat{\boldsymbol{\theta}}) \cdot T, \tag{2.2}$$

where $f(\cdot)$ and $g(\cdot)$ are unknown smooth continuous functions. In our numerical illustrations, we fit this model using the R package `mgcv` developed by Simon Wood, in which smooth functions are represented as a weighted sum of known basis functions; and the likelihood is maximized with an added smoothness penalization term. We use penalized cubic regression splines as the basis functions, with dimension equal to five.

In contrast, HI propose to estimate the conditional expectation of the response as a function of the observed treatment, T, and the GPS, $\hat{R}$. They recommend using a flexible parametric function of the two arguments and give the following Gaussian quadratic regression model,

$$E(Y \mid T, \hat{R}) \;=\; \alpha_0 + \alpha_1 \cdot T + \alpha_2 \cdot T^2 + \alpha_3 \cdot \hat{R} + \alpha_4 \cdot \hat{R}^2 + \alpha_5 \cdot T \cdot \hat{R}. \tag{2.3}$$

This can be viewed as a generalization of RR's covariance adjustment technique, which in the binary treatment case involves regressing Y on $\hat{e}(\boldsymbol{X})$ separately for the treatment and control groups. HI, on the other hand, parametrically estimate the average outcome for all possible treatment levels simultaneously via the quadratic regression on T given in (2.3). Non-parametric response models are suggested by FFGN, see Section 2.2.5.

Finally, RHB, illustrate IPW by linearly adjusting for the treatment,

$$E(Y \mid T) \;=\; \beta_0 + \beta_1 \cdot T, \tag{2.4}$$

where (β_0, β_1) are fit via linear regression with stabilized weights, $\hat{W}_i$. The model in (2.4) is just an example, IPW can be used with a variety of flexible response models for robust estimation of the DRF (e.g., Wang *et al.*, 2007; Ertefaie and Stephens, 2010; Bodnar *et al.*, 2004; Hernán *et al.*, 2002). In our numerical studies we use a kernel-based regression to facilitate comparison with a GPS-based proposal of FFGN, see Section 2.2.5.

2.2.4 Estimating causal quantities. Some methods focus on estimating the DRF and others on the average causal effect. We illustrate how the DRF can be estimated with IPW in Section 2.2.5. Under (2.4), the estimated average treatment effect is the weighted least squares estimate of β_1. The P-FUNCTION was designed to estimate the average causal effect. Under (2.2) this involves averaging $g(\hat{\boldsymbol{\theta}}_i)$ across all units. Bootstrap standard errors are computed by resampling the data and refitting both the treatment assignment and response models. With subclassification, computing the estimated average causal effect proceeds exactly as in the binary case. Because a response model is fit conditional on T within each subclass, we can also in principle average these fitted models and estimate the DRF. While we illustrate this possibility in our simulations, we advocate a flexible non-parametric approach in Section 2.4.1.

In contrast, to estimate the DRF, HI computes the average potential outcome on a grid of treatment values. In particular, at treatment level t, they compute

$$\hat{E}\{Y(t)\} \;=\; \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{\alpha}_0+\hat{\alpha}_1\cdot t+\hat{\alpha}_2\cdot t^2+\hat{\alpha}_3\cdot \hat{r}(t,\boldsymbol{X}_i)+\hat{\alpha}_4\cdot \hat{r}(t,\boldsymbol{X}_i)^2+\hat{\alpha}_5\cdot t\cdot \hat{r}(t,\boldsymbol{X}_i)\Big). \quad (2.5)$$

In (2.5) the mean response had all individuals received dose t is estimated by computing a *dose-specific* score: r is evaluated at t, not T. While not unprecedented (Rosenbaum, 1987; Imbens, 2000), this distinguishes the method from both IPW and the P-FUNCTION, for which there is a single score for each individual. This difference has ramifications. When the inverse GPS is used as a weight, it cannot be stabilized in the standard manner because $p_{\hat{\psi}_0}(t)$ does not vary among individuals so that $p_{\hat{\psi}_0}(t)/r(t,\boldsymbol{X}_i) \propto 1/r(t,\boldsymbol{X}_i)$. Standard errors can be calculated using the bootstrap, taking into account the estimation of both the GPS and model parameters.

In practice, we are often interested in the *relative* DRF, $E\{Y(t)-Y(0)\}$, which compares the average outcome under each treatment level with that under the control, i.e., $t=0$. Of course, in some studies there is no control *per se* and we revert to $E\{Y(t)\}$. In our simulation studies we report the relative DRF while in our applied example we report the DRF which is more appropriate in its particular context.

2.2.5 Extensions of the method of HI and RHB. Unfortunately, the quadratic regression in (2.3) is not sufficiently flexible for robust estimation of the DRF, see Section 2.3. Bia *et al.* (2011) and FFGN point out that misspecification of (2.3) can result in biased causal quantites and FFGN proposes two alternatives. The first generalizes (2.3) with,

$$E(Y \mid T,\hat{R}) \;=\; \beta(T,\hat{R}) \quad (2.6)$$

where $\beta(T, \hat{R})$ is a flexible nonparametric model; in our numerical studies we use a SCM.[1] The DRF, $\hat{E}\{Y(t)\}$, and its standard errors are computed as in(2.5), but with

$$\hat{E}\{Y(t)\} \;=\; \frac{1}{n}\sum_{i=1}^{n}\hat{\beta}[t, \hat{r}(t, \boldsymbol{X}_i)] \tag{2.7}$$

Because the SCM is a function of the GPS, we refer to this method as SCM(GPS).

FFGN's second method involves a GPS version of inverse probability weighting. We refer to this method as IPW_0 because it uses naive rather than stabilized weights; recall that stabilization is not possible when using $1/\hat{r}(t, \boldsymbol{X}_i)$ as weights, see Section 2.2.4. The IPW_0 estimate of the DRF is

$$\hat{E}\{Y(t)\} \;=\; \frac{\sum_{i=1}^{N}\tilde{K}_{h,X}(T_i - t)\cdot Y_i}{\sum_{i=1}^{N}\tilde{K}_{h,X}(T_i - t)}, \tag{2.8}$$

where $\tilde{K}_{h,X}(T_i - t) = K_h(T_i - t)/\hat{r}(t, \boldsymbol{X}_i)$, $K(\cdot)$ is a kernel function with the usual properties, h is a bandwidth satisfying $h \to 0$ and $Nh \to \infty$ as $N \to \infty$, and $K_h(\cdot) = h^{-1}K(\cdot/h)$. This is the local constant regression (Nadaraya-Watson) estimator but now with each individual's kernel weight being divided by its GPS at t. To avoid boundary bias and to simplify derivative estimation, the IPW_0 estimates $E\{Y(t)\}$ using a local linear regression of Y on T with a weighted kernel function $\tilde{K}_{h,X}(T_i - t)$, i.e.,

$$\hat{E}\{Y(t)\} \;=\; \frac{D_0(t)S_2(t) - D_1(t)S_1(t)}{S_0(t)S_2(t) - S_1^2(t)}, \tag{2.9}$$

where $S_j(t) = \sum_{i=1}^{N}\tilde{K}_{h,X}(T_i - t)(T_i - t)^j$ and $D_j(t) = \sum_{i=1}^{N}\tilde{K}_{h,X}(T_i - t)(T_i - t)^j Y_i$. The global bandwidth can be chosen following the procedure of Fan and Gijbels (1996). We use (2.9)

[1] FFGN propose a nonparametric kernel estimator with polynomial regression of order 1 (Fan and Gijbels, 1996), but we use the SCM to facilitate comparisons of the methods. As with (2.2), we use the `mgcv` package with penalized cubic regression splines as the basis functions with dimension equal to five for both T and $\hat{R}$ along with a tensor product.

Table 2.1: Estimates of the DRF. Methods differ only in their response models, not their treatment models.

Estimate	Description
IvD	Using $\hat{\boldsymbol{\theta}}$ to form S subclasses, fit a linear model within each subclass and average the S fitted models to estimate DRF.
HI	Estimate the DRF with (2.3) and (2.5).[a]
SCM(GPS)	Same as HI, but with the quadratic regressions in (2.3) and (2.5) replaced by a SCM as in (2.6) and (2.7).
IPW_0	Same as HI, but with the quadratic regressions in (2.3) and (2.5) replaced by the kernel smoothing regression in (2.8) and (2.9).
IPW_{SW}	As in (2.4), but with the linear regression replaced by the kernel smoothing regression in (2.8) and (2.9), but with $\tilde{K}_{h,X}(T_i - t) = K_h(T_i - t)\hat{W}(T_i, \boldsymbol{X}_i)$.
SCM(P-FUNCTION)	Estimate the DRF with (2.12) and (2.13), see Section 2.4.

[a] In the linear fit of Simulation II, the quadratic models in (2.3) and (2.5) are replaced by linear models.

as the IPW_0 estimator in our numerical studies.

Unfortunately, IPW_0 can be very unstable, owing to the infinite variance of $1/r(t, \boldsymbol{X}_i)$, at least when the treatment is continuous (RHB). To improve IPW_0 and to implement the method of RHB for robust estimation of the DRF, we replace $1/\hat{r}(t, \boldsymbol{X}_i)$ with Robins' stabilized weight, $\hat{W}(T, \boldsymbol{X})$. We denote this method IPW_{SW}, where the subscript indicates its stabilized weights. Notice that these weights are evaluated at T rather than t. This is a fundamental difference from the methods layed out by HI. In this regard IPW_{SW} should be viewed as a flexible implementation of the method of RHB with (2.4) replaced by kernel smoothing regression of the form in (2.8), but with $\tilde{K}_{h,X}(T_i - t) = K_h(T_i - t)\hat{W}(T_i, \boldsymbol{X}_i)$. Table 2.1 summarizes the specific estimates of the DRF that we review in Sections 2.3-2.6.

2.3 Comparing IPW, the GPS, and the P-Function

In this section, we examine the differences between the methods of RHB, IvD, HI, and FFGN using both simulation studies and theoretical comparisons. The key differences lie in how the method summarizes $p(T \mid \boldsymbol{X})$: the GPS and the weights used in IPW both evaluate

this density at the observed covariate, whereas the P-FUNCTION uniquely parameterizes it. As we show below, this difference leads to alternative response models and markedly divergent results.

2.3.1 Simulation study I

In our first simulation study, we generate $2,000$ observations, each of which includes a single continuous covariate, X, a continuous univariate treatment, T, and a response variable, Y. We simulate $X_i \overset{\text{ind}}{\sim} \mathcal{N}(0.5, 0.25)$ and $T_i \mid X_i \overset{\text{ind}}{\sim} \mathcal{N}(X_i, 0.25)$ and assume that the potential outcome is distributed as $Y_i(t) \mid T_i, X_i \overset{\text{ind}}{\sim} \mathcal{N}(10X_i, 1)$ for all $t \in \mathcal{T}$. In this simulation study the true treatment effect is zero and the true DRF is five for all t. We deliberately choose this simple setting where any reasonable method should perform well. Fitting a simple linear regression of Y on T yields a statistically significant treatment effect estimate of roughly five. However, adjusting for X in the regression model is sufficient to yield an estimate that is much closer to and is not statistically different from the true effect of zero.

Using the correctly specified treatment assignment model, $T_i \mid X_i \overset{\text{ind}}{\sim} \mathcal{N}(X_i, 0.25)$, the marginal distribution of the the treatment, $T_i \sim \mathcal{N}(0.5, 0.5)$, and the response models given in Table 2.1, we implement the HI, IvD, SCM(GPS), IPW_0, and IPW_{SW} methods. For the purposes of illustration, we do not adjust for $\hat{\boldsymbol{\theta}}$ within each subclass when using IvD's method. Owing to the linear structure of the generative model, doing so would dramatically reduce bias even with a small number of subclasses. Here we illustrate, instead, how bias can be reduced by increasing the number of subclass; we implement IvD with $S = 5, 10,$ and 50 subclasses. For the methods other than IvD, we use a grid of ten equally spaced points between -0.5 and 1.5, $t_1, \ldots, t_D$ with $D = 10$, to compute the relative DRF and its derivative. Standard errors are computed using 1,000 bootstrap replications.

Figure 2.1 presents the results. In the first row, we plot the estimated relative DRF while

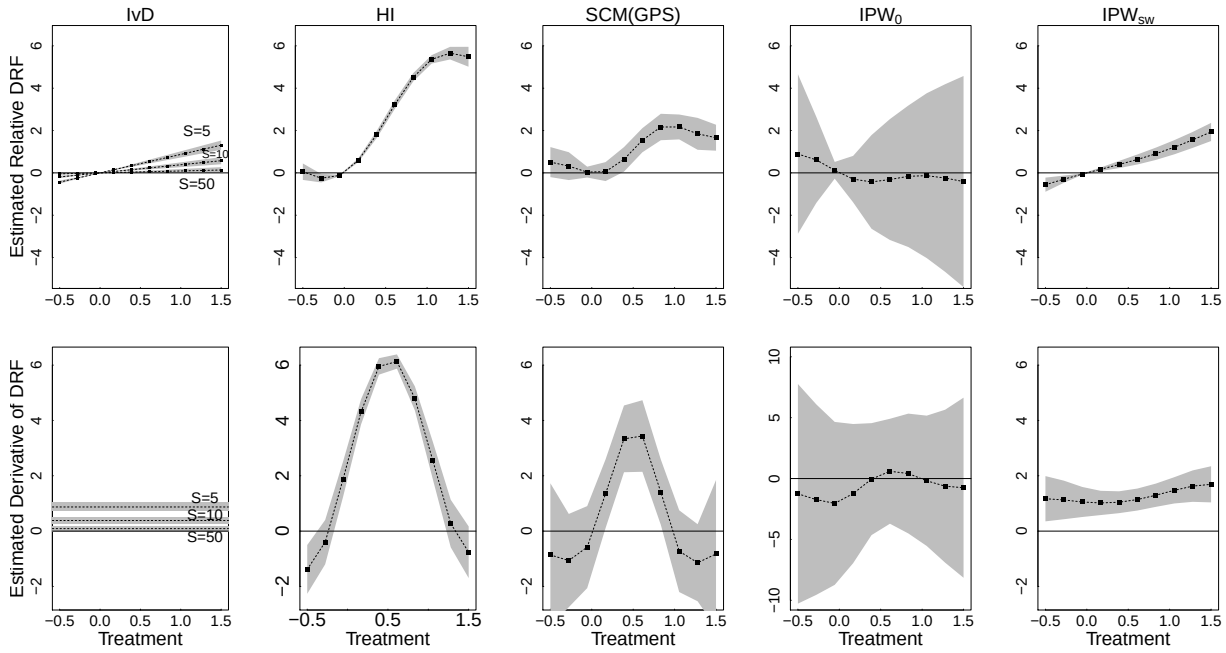

Figure 2.1: Results of Simulation Study I. The first row plots the estimated relative DRF with the horizontal solid line representing the true relative DRF. For IvD, we use $S = 5, 10, 50$ subclasses. The solid diagonal line for the method of HI is the unadjusted regression of Y on T. The second row plots the estimated derivatives of the DRF with the solid line representing the truth. In both rows, the grey shaded areas represent 95% confidence intervals. The estimated derivative for IPW_0 is plotted on a different scale as its standard error is significantly larger than that of the other methods.

the second row plots the estimated derivative of the DRF. For HI, SCM(GPS), IPW_0, and IPW_{SW} the derivative is computed as

$$\frac{1}{2}\left[\frac{\hat{E}\{Y(t_{d+1})\} - \hat{E}\{Y(t_d)\}}{t_{d+1} - t_d} + \frac{\hat{E}\{Y(t_d)\} - \hat{E}\{Y(t_{d-1})\}}{t_d - t_{d-1}}\right] \tag{2.10}$$

for $d = 2, \ldots D-1$. For $d = 1$, we simply use the first term in (2.10) and for $d = 10$ we use the second term in (2.10). For IvD, the derivative is the weighted average of the within subclass linear regression coefficient; 95% point-wise confidence intervals are shaded gray.

Figure 2.1 shows that even in this simple simulation, all methods except IPW_0 miss the true relative DRF and its derivative, albeit to differing degrees. The behavior of IvD's estimate

improves with more subclasses, a luxury we can afford here because of the large sample size. IvD makes the general recommendation that the within subclass models be adjusted for $\boldsymbol{X}$ or at least for $\hat{\boldsymbol{\theta}}$. Because of the simple structure of this simulation, doing so would result in a correctly specified model even with a single subclass, eliminating bias in the estimated average treatment effect.[2] We do not recommend estimating the DRF by averaging the unadjusted within subclass models, but do so here to facilitate comparisons between the methods. We propose a new estimate of the DRF using the P-FUNCTION in Section 2.4.1.

The performance of HI's method is particularly poor; it differs only slightly from the unadjusted regression. Although SCM(GPS) offers limited improvement, it also introduces a cyclic artifact into the fit. We will see this pattern again and discuss it in Section 2.3.6. IPW_0, on the other hand, results in an unstable fit that is characterized by large standard errors. The performance of these methods are especially troubling both because the GPS was expressly designed to estimate the DRF and because the current simulation setup is so simple. Given their performance here, it is difficult to expect these methods to succeed in more realistic settings. While IPW_{SW} also misses the mark in Simulation I, it is important to emphasize that RHB did not propose to use IPW to estimate the full DRF. The primary goal of this paper is to explain why the GPS-based methods can fail and to provide a more robust estimate of the DRF.

One reason that GPS-based methods can perform poorly is that their response model are based on overly strong parametric assumptions, especially (2.3). This is illustrated in Figure 2.2 which compares the fitted mean potential outcome as a function of the GPS and T under the HI model (left panel) and under SCM(GPS) (middle panel). The fitted potential

[2]It would also complicate estimation of the DRF. Because the treatment assignment mechanism is strongly ignorable given the propensity function (IvD), we aim to adjust for the propensity function in a robust manner in the response model. Thus, adjusting for $\hat{\theta}$ within the subclasses poses no conceptional problem. In practice, however, $\hat{\theta}$ tends to be fairly constant within subclasses and its coefficient tends to be correlated with the intercept. A solution is to recenter $\hat{\theta}$ within each subclass. Because, we propose a more robust strategy for estimating the DRF using the P-FUNCTION in Section 2.4.1, however, we do not purse such adjustment strategies here.

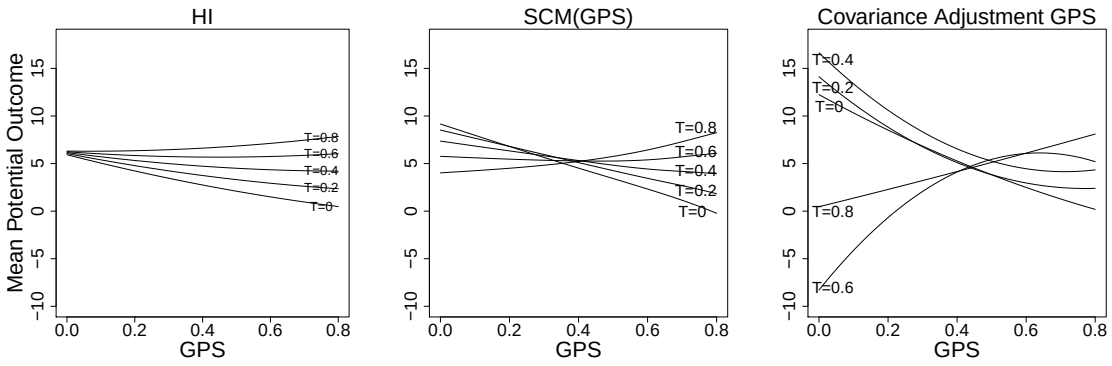

Figure 2.2: The Varying Flexibility of the Response Models. The plots show the mean potential outcome as a function of the GPS and T under HI's quadratic response model (left panel), fitted SCM(GPS) (middle panel), and covariance adjustment GPS (right panel). Covariance adjustment GPS fits a quadratic regression ($Y \sim R + R^2$) in each of several subclasses based on T, see Appendix 2.5 for details. We use 10 subclasses but only plot five. Subclassification is by far the most flexible of the three response models.

outcomes differs substantially and are considerably more constrained under the quadratic model of HI. To fit an even more flexible response model, we subclassified the data into 10 subclasses based on T, and fit a quadratic regression for Y as a function of the GPS seperately within each of the subclasses. Five out of the 10 within subclass fit are plotted in the right most plot in Figure 2.2. The results differs substantially from HI and reveals the considerable constraint of the quadratic response model. Subclassifying on T in this way leads to a new response model and a corresponding new GPS-based estimate of the DRF; this method is discussed in Appendix 2.5.

2.3.2 Simulation study II

Although IvD's response model in Simulation I is misspecified in terms of its adjustment for X, the method may benefit from its assumption that the DRF is linear in T. We address this in the second simulation that compares the performance of the methods under several alternative generative models. We also explore the frequency properties of the methods.

Suppose we have a simple random sample of $2,000$ observations that includes a trivariate

normal covariate, (X_1, X_2, X_3), with mean vector $(1, 1, 4)$, component variances all equal to one, $\text{Corr}(X_1, X_2) = 0.3$, $\text{Corr}(X_1, X_3) = -0.4$, and $\text{Corr}(X_2, X_3) = 0.6$. Suppose further that the treatment is generated according to $T \mid \boldsymbol{X} \overset{\text{ind}}{\sim} \mathcal{N}(X_1 - X_2^2 + 0.5X_3, 1)$ and the response according to one of three response models:

Gaussian Linear DRF: $Y(t) \mid t, \boldsymbol{X} \overset{\text{ind}}{\sim} \mathcal{N}(X_1 + 2X_2 + t, 9)$,

Gaussian Quadratic DRF: $Y(t) \mid t, \boldsymbol{X} \overset{\text{ind}}{\sim} \mathcal{N}((X_1 + 2X_2 + t)^2, 9)$, or

Lognormal DRF: $Y(t) \mid t, \boldsymbol{X} \overset{\text{ind}}{\sim} \exp\left\{\frac{\mathcal{N}(X_1 + 2X_2 + t, 9)}{5}\right\}$.

Unlike the Gaussian response models the lognormal model exhibits significant homoscedasticity; the central 95% of the conditional variances range from 0.3 to 17.8.

To isolate the difference between the methods, we use the correctly specified treatment model in our analyses. (We use the R function `density` to estimate the marginal distribution of T for use in stabilized weights.) For the fitted response model under HI and IvD's methods, we consider Gaussian regression models that are linear and quadratic in T. In particular for the method of IvD we fit (i) $Y \sim T$ and (ii) $Y \sim T + T^2$ within each of $S = 10$ equally sized subclasses, and for the method of HI we fit (i) $Y \sim T + R + R^2 + R \cdot T$ and (ii) $Y \sim T + T^2 + R + R^2 + R \cdot T$. With IvD, the relative DRF is computed by averaging the coefficients of the within subclass models. The response models given in Table 2.1 are used for SCM(GPS), IPW_0, and IPW_{SW}. For the GPS-based methods and IPW, the relative DRF is evaluated at ten equally spaced values of t between -1.9 to 3.4. The entire procedure was repeated using all methods on each of 1,000 data sets generated with the same covariate and treatment models and with each of the three response models. All of the fitted response models are misspecified in their adjustment for X and/or T, as we expect in practice. Thus, this simulation study investigates the robustness of the methods to typical misspecification of the response model.

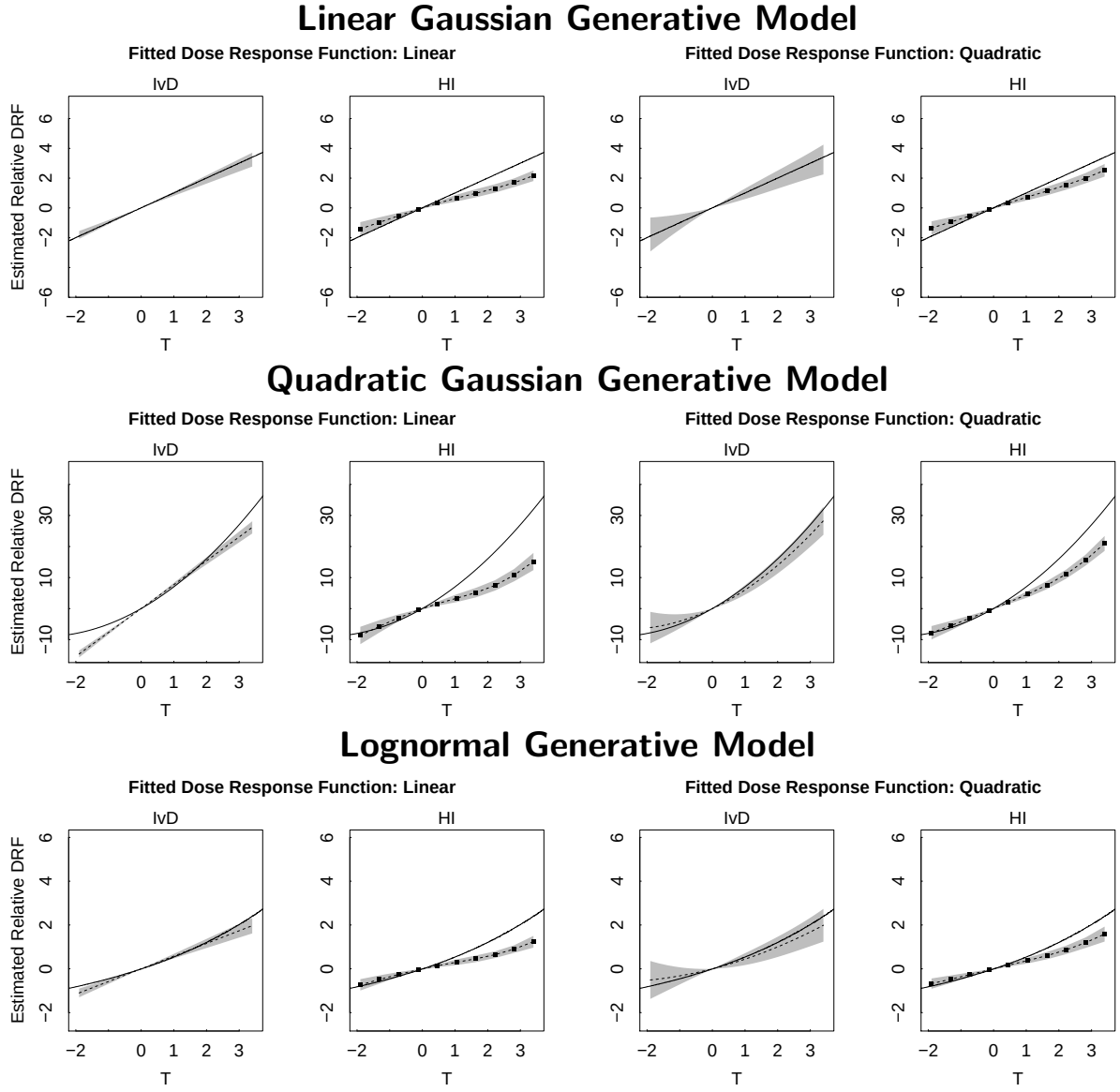

Figure 2.3: Estimated Relative DRFs in Simulation Study II Using the Methods of IvD and HI. The solid lines plot the true relative DRFs, the dashed lines plot the means of the fitted relative DRFs across 1000 simulations, and the gray shaded regions plot two standard deviation pointwise intervals across the 1000 fits. The evenly-spaced grid of evaluation points used with HI are also plotted as solid circles. The method of HI shows appreciable bias with all six combinations of generative and fitted response models. The method of IvD, on the other hand, is biased only when the fitted model is linear and the generative is not.

Figures 2.3 and 2.4 report the average of the estimated relative DRFs across the simulations (dashed lines) along with their two standard deviation intervals (shaded regions). The true relative DRF functions are plotted as solid lines; rows correspond to the three response

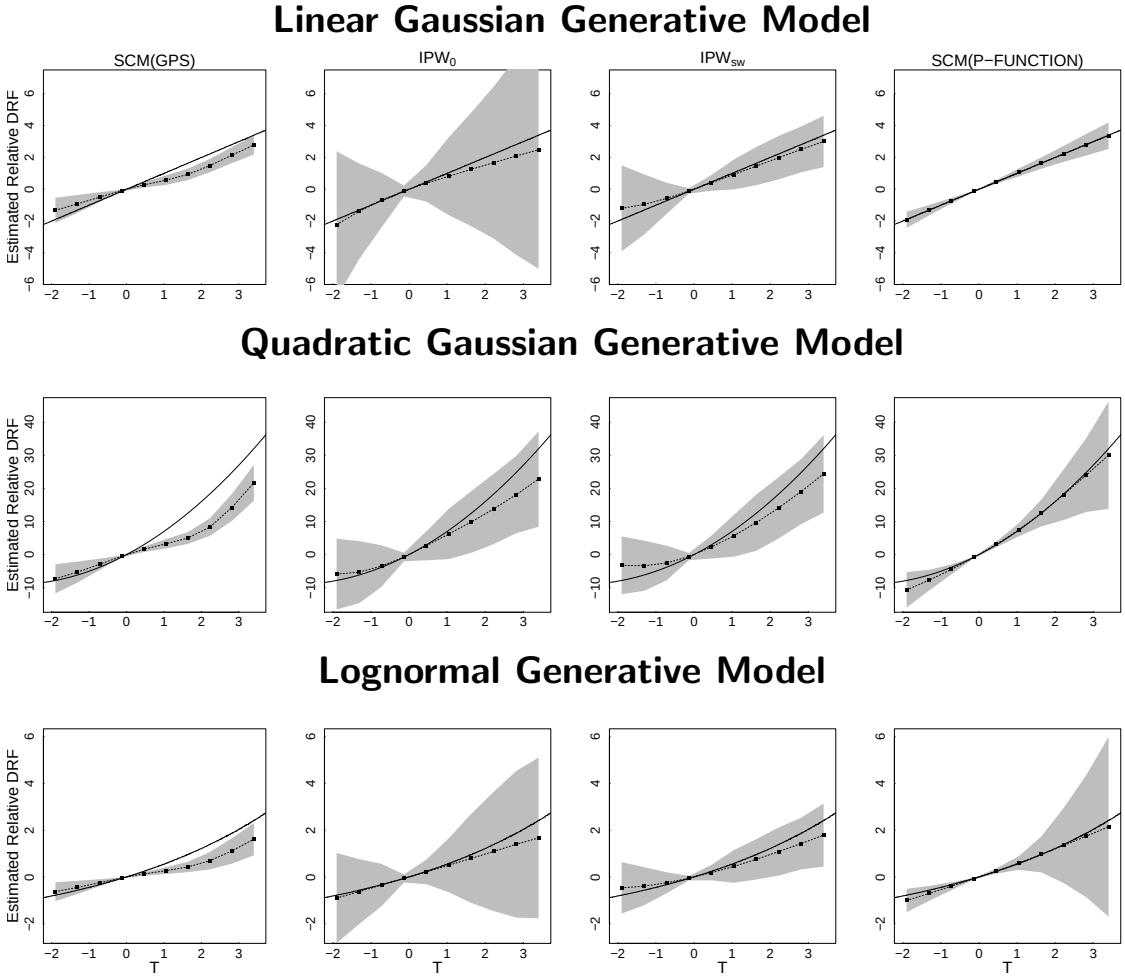

Figure 2.4: Estimated Relative DRFs in Simulation Study II for the SCM(GPS), IPW_0, IPW_{SW}, and SCM(P-FUNCTION) Methods. Solid lines, dashed lines and gray regions represent the true relative DRFs the means of the 1000 fitted relative DRFs and 95% pointwise intervals. Points represent the evenly-spaced grid points. The SCM(P-FUNCTION) method is discussed in Section 2.4.1.

models. The left pair of columns in Figure 2.3 give results when the fitted model is linear under the IvD method (column 1) and the HI method (column 2). The right pair of columns give results when the fitted model is quadratic under the methods of IvD (column 3) and HI (column 4).

The IvD method performs reasonably well when the generative model is linear (row 1). When IvD is fit with a quadratic model (column 3), the DRF is estimated with little bias though the estimate has higher variability. IvD exhibits significant bias only when the fitted

model is linear and the true DRF is non-linear (column 1, rows 2 and 3). The HI method, on the other hand, exhibits appreciable bias even when the fitted response model matches the true model in its functional dependence on t. Like the IvD method, the bias is most acute when the fitted model is linear but the true DRF is not. Unlike with IvD, however, the 95% frequency intervals of HI miss the true value completely across a wide range of treatment values.

The first three columns of Figure 2.4 give result for SCM(GPS), IPW_0, and IPW_{SW}. All three improve on HI when the true DRF is quadratic, although the variances are larger. For the linear DRF, SCM(GPS) is comparable to HI while IPW_0 exhibits enormous variance, and IPW_{SW} exhibits moderate variance. Except for the variance of IPW_0, the heteroscedasticity of the lognormal response model does not significantly effect the three methods.

2.3.3 Simulation study III

The simulation study aims to investigate the effect of a heteroscedastic treatment and the robustness of the methods to misspecification of the treatment model. The simulation setup is exactly as in Simulation study II (i.e., the same sample size, covariate distribution, and replication) but we consider a heteroscedastic treatment, $T \mid \boldsymbol{X} \overset{\text{ind}}{\sim} \mathcal{N}(X_1 - X_2^2,\ 0.25X_3^2)$. The response is generated according to the Gaussian quadratic DRF of Simulation study II. This simulation is repeated using two fitted treatment models:

Correctly Specified Treatment Model: $T \mid \boldsymbol{X} \overset{\text{ind}}{\sim} \mathcal{N}(\beta_0 + \beta_1 X_1 + \beta_2 X_2^2,\ \sigma^2 X_3^2)$

Misspecified Treatment Model: $T \mid \boldsymbol{X} \overset{\text{ind}}{\sim} \mathcal{N}(\beta_0 + \beta_1 X_1 + \beta_2 X_2^2,\ \sigma^2)$.

(We consider misspecification of both the mean and variance of the treatment model in Section 2.6.2.) The parameters of both treatment models are fit via maximum likelihood and the marginal treatment model is fit using the R function `density`. The estimated DRFs

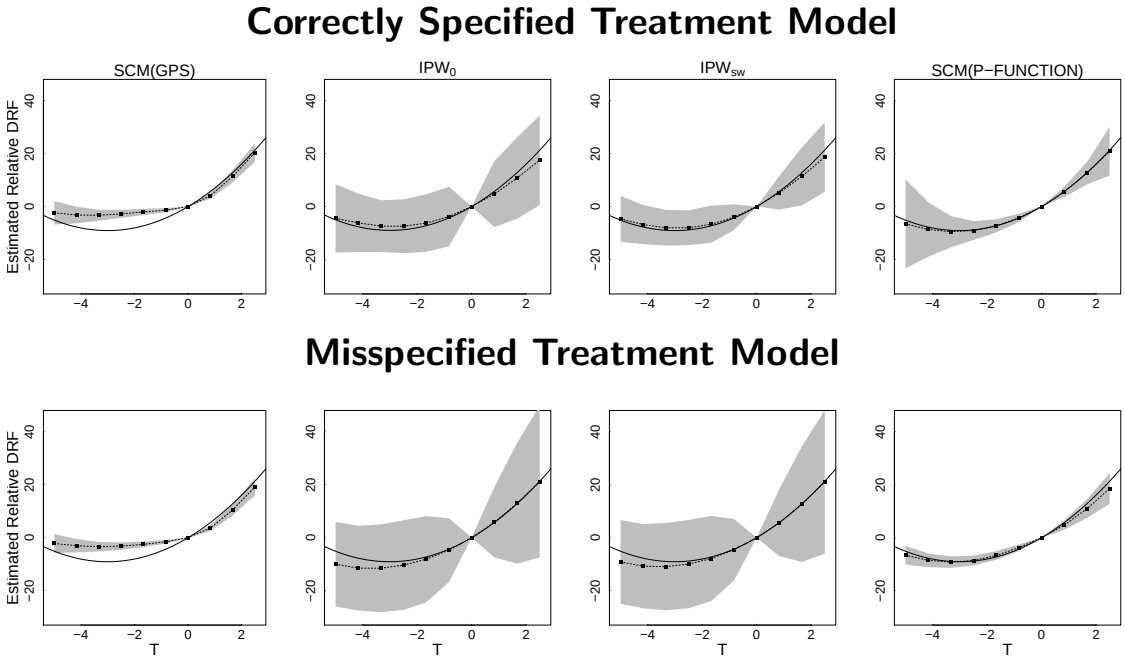

Figure 2.5: Estimated Relative DRFs under the Heteroscedastic Treatment of Simulation Study III. Solid lines, dashed lines and gray regions represent the true relative DRFs the means of the 1,000 fitted relative DRFs and 95% pointwise intervals.

fit with SCM(GPS), IPW_0, and IPW_{SW} using the response models in Table 2.1 appear in the first three columns of Figure 2.5. Using the bandwidth of Fan and Gijbels (1996) as suggested by FFGN leads to numerical instability in 708 of the 1000 datasets fit with IPW_0 under the correctly specified treatment model. The bandwidth can be tuned to improve stability; results in Figure 2.5 are based on a single bandwidth that gave stable results in 793 of the 1,000 datasets. Although this procedure would be difficult to implement with a single dataset, it gives the best possible representation to IPW_0. The standard bandwidth of Fan and Gijbels (1996) was used with the misspecified treatment model. The specification of the treatment model has little effect on SCM(GPS). For both of the IPW methods, the choice of treatment model effects the variance more than the bias of the fitted DRF, and, IPW_0 is much more stable with the *misspecified model*. Of the three methods IPW_{SW} is clearly best in terms of coverage and stability.

Taking the results of Simulations I–III together, IPW_{SW} preforms better than the other methods. The GPS methods (HI, SCM(GPS), and IPW_0) are explicitly designed to estimate the DRF but seem ill-suited to the task.

2.3.4 Simulation study III.1

Although the SCM(GPS) estimate of the DRF shows some bias in Simulation studies II and III, the cyclic bias that it exhibits in Simulation I is much less pronounced in Figures 2.4 and 2.5. To see if the cyclic bias exists in more complex settings, we extend the well-known simulation study of Kang and Schafer (2007) to a continuous treatment. In particular, we independently simulate $Z_{ij} \sim N(0,1)$ for $i = 1, \ldots, 2000$ and $j = 1, \ldots, 4$ and generate

$$T_i = -Z_{i1} + 0.5Z_{i2} - 0.25Z_{i3} - 0.1Z_{i4} + \sigma_i,$$

where $\sigma_i \sim N(0,1)$, and

$$Y_i = 210 + 27.4Z_{i1} + 13.7Z_{i2} + 13.7Z_{i3} + 13.7Z_{i4} + 10T_i + \epsilon_i$$

where $\epsilon_i \sim N(0,1)$. We estimate the relative DRF by applying the methods of HI, SCM(GPS), IPW_0, IPW_{SW}, and SCM(P-FUNCTION) to each of 1000 replicated data sets. We use the correctly specified treatment model and the fitted response models given in Table 2.1. Figure 2.6 shows that the cyclic bias remains a problem for SCM(GPS) and that large variances continue to plague IPW_0. The stabilized weights of IPW_{SW} clearly improve its performance, relative to IPW_0. Nonetheless, SCM(P-FUNCTION) dominated the other methods.

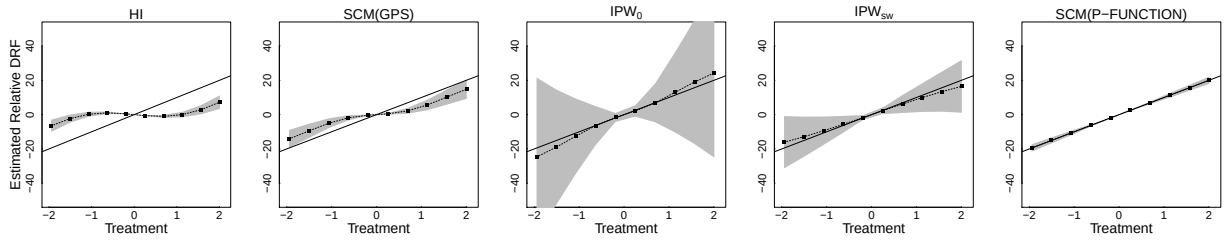

Figure 2.6: Estimated Relative DRF in Simulation V. Solid lines, dashed lines and gray regions represent the true relative DRFs the means of the 1000 fitted relative DRFs and 95% pointwise intervals. The evaluation points are identical for all plots. SCM(GPS) exhibits a cyclic artifact and IPW_0 is quite unstable. The SCM(P-FUNCTION) method proposed in Section 2.4.1 again ourperforms the other methods.

2.3.5 Theoretical considerations and methodological implications

To better understand the simulation results, we consider the tradeoff in assumptions required by IPW, the GPS, and the P-FUNCTION. All three require a form of conditional independence between T and $\boldsymbol{X}$—after weighting, conditioning on the GPS, and conditioning of the P-FUNCTION, respectively. While IvD make the strong theoretical assumption of a uniquely parameterized propensity function, IPW and the method of HI effectively make the same assumption in that they typically use the same parametric treatment models.

To flesh this out, recall that the three methods generalize RR's propensity score. In the binary case, $p(T|\boldsymbol{X})$ is uniquely determined by $e(\boldsymbol{X}) = p(T = 1|\boldsymbol{X})$. IvD focus on *uniquely* determining the full conditional distribution of T given $\boldsymbol{X}$, and assume this conditional distribution is parameterized and can be uniquely represented by $\boldsymbol{\theta}$. RHB and HI, on the other hand, do not constrain the treatment assignment model in this way and instead, following the binary propensity score, evaluate $p(T|\boldsymbol{X})$ to compute propensity weights or GPS. In this way the GPS does *not* uniquely determine $p(T|\boldsymbol{X})$. There may be multiple distributions that are equal when evaluated at a particular t. The assumption of a uniquely parameterized propensity function constrains the choice of treatment model that can be used for a P-FUNCTION. In practice, however, the same treatment models are typically used by all three methods.

Comparing IvD and HI, the assumption of IvD allows a stronger form of *strong ignorability of the treatment assignment given the propensity function.* In particular, Result 2 of IvD states

Ignorability of IvD: $p\{Y(t) \mid T, e(\cdot \mid \boldsymbol{\theta})\} = p\{Y(t) \mid e(\cdot \mid \boldsymbol{\theta})\}$ for every t,

Whereas, in their Theorem 2.1., HI show

Ignorability of HI: $p_T\{t \mid r(t, \boldsymbol{X}), Y(t)\} = p_T\{t \mid r(t, \boldsymbol{X})\}$ for every t.

In the case where T is categorical, HI's ignorablity implies that $\mathbf{1}\{T = t\}$ and $Y(t)$ are independent given $r(t, \boldsymbol{X})$, where the GPS is evaluated at the particular value of t in the indicator function. Although achieving conditional independence of $Y(t)$ and T would require conditioning on a family of GPS, HI provide an insightful moment calculation to show how the response model described in Section 2.2.2 can be used to compute the DRF. Nonetheless conditioning on either R or $r(t, \boldsymbol{X})$, for any particular value of t does not guarantee that T will be uncorrelated with the potential outcomes. The fact that the GPS does not constitute a single score for each individual restricts the response models that can be used. Subclassification, for example, is not feasible unless the classifying variable is low dimensional. The advantage of IvD over HI is that $Y(t)$ and T are conditionally independence given the low-dimensional score, $\boldsymbol{\theta}$, enabling the use of a wide-range of response models. Similarly, IPW only requires that $Y(t)$ and T be independent in the weighted sample: any model that allows for weighting can be used.

2.3.6 Simulation study IV: The potential cyclic bias of SCM(GPS)

The DRF fitted with SCM(GPS) in Simulation I exhibits a cyclic artifact that does not exist in the underlying DRF; Section 2.3.4 provides another simulation study in which this

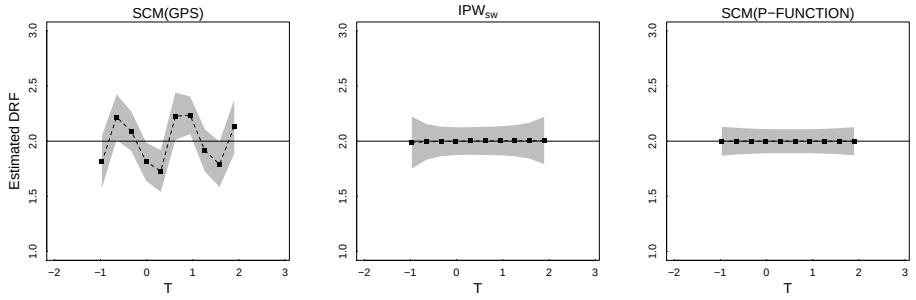

Figure 2.7: Estimated DRF for Simulation IV. Solid lines, dashed lines and gray regions represent the true DRFs the means of the 1,000 fitted DRFs and 95% pointwise intervals. Only SCM(GPS) exhibits the cyclic bias. The SCM(P-FUNCTION) is introduced in Section 2.4.

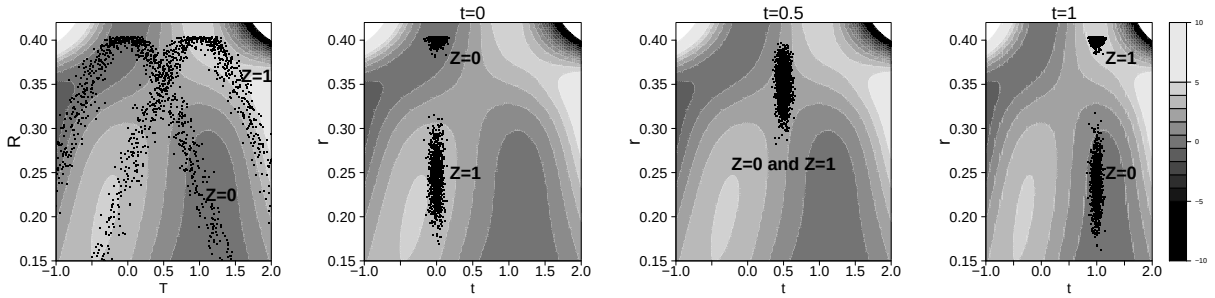

Figure 2.8: How the SCM(GPS) fit can lead to a cyclic artifact in the the fitted DRF. The leftmost panel overlays a scatter plot of T and the GPS, $\hat{R}$, on a heat map of the fitted SCM(GPS) response model in Simulation IV. The other three panels overlay scatterplots of $(t, \hat{r}(t, \boldsymbol{X}_i))$, with t equal to 0, 0.5, and 1. (We jitter in the T direction to improve visualization.) The panels show that as t increases the $(t, \hat{r}(t, \boldsymbol{X}_i))$ clusters move from local minima to local maxima and back, resulting in a cyclic pattern in the fitted DRF.

effect is even more pronounced. Here we present a simulation study that investigates the origin of this cyclic bias. In particular, we independently generate $Z_i \sim$ Bernoulli(0.5), $X_i \sim \mathcal{N}(Z_i, 0.01)$, $T_i \sim \mathcal{N}(X_i, 1)$, and $Y_i \sim \mathcal{N}(4Z_i, 1)$, for $i = 1, \ldots, 2000$. Using the correct treatment model, we estimate the DRF using SCM(GPS) at ten evenly spaced theoretical percentiles of T. We repeat the entire fitting procedure on each of 1,000 replicated data sets and plot the average of the estimated DRF and their pointwise two standard deviation intervals in Figure 2.7. For comparison, we also present results for IPW_{SW}. The cyclic bias of the SCM(GPS) fit is evident.

To see the source of the cyclic bias, we plot the fitted response model, the SCM given in (2.6), as a heat map along with a scatter plot of the observed $(T_i, \hat{R}_i)$ in the leftmost panel

of Figure 2.8. The two bell-shaped curves that appear in the plotted values of $(T_i, \hat{R}_i)$ stem from the definition of the GPS; $\hat{R}_i$ is the value of the fitted density function of T. By design, X clusters around the two values of Z; these two clusters correspond to the two bell-shaped curves.

The overlapping bell-shaped curves in the observed $(T_i, \hat{R}_i)$ induce a cyclic pattern in the fitted SCM-response model. To estimate the DRF at t, the fitted response model is evaluated at and averaged over each $\hat{r}(t, \boldsymbol{X}_i)$. This is illustrated in the second panel of Figure 2.8 which plots $(t, \hat{r}(t, \boldsymbol{X}_i))$ with $t = 0$ on top of the fitted SCM. The cluster of points at the top of the panel land in a local minimum resulting in the dip of the fitted DRF in Figure 2.7. The third and fourth panels show that as t increases to 0.5 and 1.0, the values of $(t, \hat{r}(t, \boldsymbol{X}_i))$ continue to cluster, but the clusters shift from minima to maxima of the fitted SCM, leading to the cyclic pattern in the fitted DRF.

The patterned behavior of the GPS (illustrated in the first panel of Figure 2.7) means that the response model is particularly difficult to accurately represent, even with a flexible non-parametric model, and that extrapolation is especially likely. Unfortunately, this is inevitable: when estimating the DRF we must evaluate the fitted response model at each value of $\hat{r}(t, \boldsymbol{X}_i)$, including at unobserved combinations of t and $\hat{r}(t, \boldsymbol{X}_i)$, see (2.6) and (2.7). This is a difficulty with the underlying response model, regardless of the choice of fitted response model. Although Simulation IV uses a simple setting to clearly explain the cyclic bias of SCM(GPS), the bias persist in more complex settings (see Figures 2.1, 2.4, 2.14, and 2.6).

2.4 New Methods for Estimating the DRF

In this section, we propose three new methods for robust estimation of the DRF using the

P-FUNCTION and GPS based on SCM and covariance adjustment.

2.4.1 The SCM(P-FUNCTION) Estimate

IvD developed the P-FUNCTION to estimate the average treatment effect, rather than the full DRF. Nonetheless we use the framework of IvD to compute the DRF in Simulation studies I and II (see Figures 2.1 and 2.3). The method we employ, however, is constrained by its dependence on the parametric form of the within subclass model. Practitioners would generally prefer a robust and flexible DRF, and here we propose a procedure that allows such estimation. We view this estimate as the best available for a DRF in an observational study.

We begin by writing the DRF as

$$E[Y(t)] = E\Big[\ E[Y(t) \mid \boldsymbol{\theta}]\ \Big] = E\Big[\ E[Y(T) \mid \boldsymbol{\theta}, T = t]\ \Big] \tag{2.11}$$

where the first equality follows from the law of iterated expectation and the second from the strong ignorability of the treatment assignment given the P-FUNCTION. We estimate the DRF using the right-most expression in (2.11) which we flexibly model using a SCM,

$$E[Y(T) \mid \boldsymbol{\theta}, T = t] \ = \ f(\boldsymbol{\theta}, T), \tag{2.12}$$

where $f(\cdot)$ is a smooth function of $\boldsymbol{\theta}$ and T. In practice we replace $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$ from the fitted treatment model. We approximate the outer expectation in (2.11) by averaging over the empirical distribution of $\hat{\boldsymbol{\theta}}$, to obtain an estimate of the DRF using a SCM of the P-FUNCTION,

$$\hat{E}[Y(t)] \ = \ \frac{1}{n}\sum_{i=1}^{n} \hat{f}(\hat{\boldsymbol{\theta}}_i, t), \tag{2.13}$$

where $\hat{f}(\cdot)$ is the fitted SCM. We refer to this method of estimating the DRF as the SCM(P-FUNCTION) method and typically evaluate (2.13) on a grid of values of $t_1, \ldots, t_D$ evenly spaced in range of the observed treatments. Bootstrap standard errors are computed on the same grid.

Comparing (2.6)–(2.7) with (2.12)–(2.13), SCM(GPS) and SCM(P-FUNCTION) are algorithmically very similar. The primary difference is the choice between the P-FUNCTION and GPS in the response model. As we shall see, this change has a siginificant effect on the statistical properties of the estimates. Simply put, $\boldsymbol{\theta}$ is a much better behaved predictor variable than is R. When using Gaussian linear regression for the treatment model, for example, $\boldsymbol{\theta} = \boldsymbol{X}_i^\top \beta$, whereas R is the Gaussian density evaluated at T. As illustrated in Section 2.3.6, the dependence of the GPS on t and the non-monotonicity of this dependence both complicate the response model and pose challenges to robust estimation.

Computing $\hat{E}[Y(t_0)]$ with (2.13) for some particular t_0 involves evaluating $\hat{f}(\cdot, t_0)$ at every observed value of $\hat{\theta}_i$. Invariably, the range of $\hat{\theta}$ observed among units with T near t_0 is smaller than the total range of $\hat{\theta}$, at least for some values of t_0. Thus, evaluating (2.13) involves some degree of extrapolation, at least for some values of t. Luckily, this problem is relatively easy to diagnose with a scatter plot of the observed values of $(T_i, \hat{\theta}_i)$. The estimate in (2.13) may be biased for values of the treatment where the range of observed $\hat{\theta}_i$ is relatively small. As we illustrate in our simulation studies, however, (2.13) appears quite robust and this bias is small relative to the biases of other available methods.

2.4.2 The numerical performance of SCM(P-FUNCTION): Simulation studies I–IV revisited

We now revisit the simulation studies from Section 2.3, which illustrate the potentially misleading results and/or high variance of existing estimates of the DRF. Here, we compare

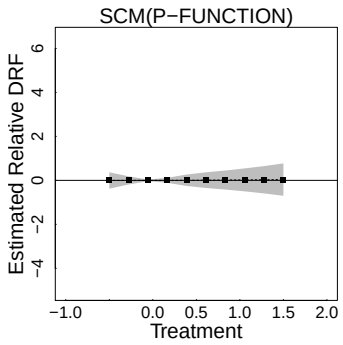

Figure 2.9: Estimated Relative DRF Using SCM(P-FUNCTION) in Simulation Study I. The solid (dashed) lines represent the true (fitted) relative DRF, the 95% confidence bands are plotted in grey, and the grid points are identical to those in Figure 2.1. The fitted relative DRF is much improved compared with those of HI, SCM(GPS), IPW0, and IPW_{SW} but without the linear assumptions of IvD (see Figure 2.1).

these results with those of SCM(P-FUNCTION). In all cases, SCM(P-FUNCTION) was fitted using the same (correctly specified or misspecifed) treatment assignment model and with the same equally-spaced grid points. When fitting the SCM, we continue to use the penalized cubic regression spline basis for both parameters (R and T) and a tensor product to construct a smooth fit of the continuous function $f(\theta, T)$ (see `mgcv` R-package documentation). Figure 2.9 and the rightmost panel of Figure 2.7 show the fitted (relative) DRF for SCM(P-FUNCTION) in Simulations I and IV, respectively. The performance of SCM(P-FUNCTION) is a dramatic improvement over all other methods in Simulation I and over SCM(GPS) in Simulation IV (see Figures 2.1, 2.7, and 2.9).

The rightmost column of Figure 2.4 presents the results of the SCM(P-FUNCTION) method in Simulation study II. Comparing Figures 2.3 and 2.4 again illustrates the advantages of the proposed method. The fits in Figure 2.3 are quite dependent on the parametric choice of the response model, whereas the non-parametric fits illustrated in Figure 2.4 do not require a parametric form. Among the non-parametric methods, the advantage of SCM(P-FUNCTION) over SCM(GPS) and IPW_0 is clear. It essentially eliminates bias with only a small increase in variance. Only IPW_{SW} has comparable statistical properties in this simulation.

In Simulation III, the correctly specified heteroscedastic treatment model is parameterized by a bivariate function; $\boldsymbol{\theta}$ has two components which are equal to the mean and log-variance of the fitted treatment model. Both components of $\boldsymbol{\theta}$ are used as predictor variables in the SCM used to model the response. The rightmost column of Figure 2.5 shows that SCM(P-FUNCTION) performs very well in Simulation III, especially with the simpler misspecified treatment model.

Overall, SCM(P-FUNCTION) consistently provides better estimates than HI, SCM(GPS), and IPW_0, both in terms of bias and variance. The IPW_{SW} also preforms better than these method and is comparable to SCM(P-FUNCTION) in Simulations II and IV. In the bulk of our numerical studies (Simulations I and III, as well as in Simulation V in Appendix 2.3.4 and the applied example in Section 2.6), however, SCM(P-FUNCTION) outperforms even IPW_{SW}.

2.5 Covariance Adjustment GPS and Covariance Adjustment P-Function

2.5.1 Covariance adjustment for catagorical treatments

One of the response models suggested by RR for a binary treatment in an observational study involves *covariance adjustment*. With this method, the response variable is regressed on the fitted propensity score separately for the treatment and control groups. Suppose we use the GPS in place of the propensity score in the context of a binary treatment. Specifically, for units in the treatment group, we use the ordinary propensity score, $R_i = r(1, \mathbf{X}_i) = p_\psi(T = 1 \mid \mathbf{X}_i)$, but for units assigned to the control group, we use the probability of control rather than the probability of treatment, $R_i = r(0, \mathbf{X}_i) = p_\psi(T = 0 \mid \mathbf{X}_i)$. Because the GPS is

equal to the propensity score for treatment units and is equal to one minus the propensity score for control units (Imbens, 2000), it is easy to see that the usual covariance adjustment is equivalent to fitting the following regression model,

$$Y_i \ \sim \ \alpha_t + \beta_t \hat{R}_i, \tag{2.14}$$

separately for the treatment and control units, i.e., $t = 0$ and 1. The linear transformation of the predictor variable does not effect the predicted value of the response for the control group.

After fitting the model given in equation (2.14), the average of the two potential outcomes can be estimated by averaging the fitted values over all units in the sample. That is, we compute

$$\hat{E}\{Y(t)\} \ = \ \frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{\alpha}_t + \hat{\beta}_t \ \hat{r}(t, \mathbf{X}_i) \right\}, \tag{2.15}$$

for $t = 0, 1$. The estimated average causal effect is simply the difference $\hat{E}\{Y(1)\} - \hat{E}\{Y(0)\}$, which is equivalent to the estimate reported in equation (2.1). Thus, with a binary treatment, the method of HI is equivalent to RR's covariate adjustment, except that HI propose a quadratic rather than a linear response model.

Suppose now that the treatment variable is categorical with more than two levels. In principle, exactly the same procedure can be applied. Namely, the regression model given in equation (2.14) can be fitted separately for units in each treatment group and the average potential outcome can be computed using the formula of equation (2.15) for each level of the treatment. We refer to this procedure as *covariate adjustment* GPS *for categorical treatments*. The relative DRF can be estimated as $\hat{E}\{Y(t)\} - \hat{E}\{Y(0)\}$ for each t. This procedure's validity follows directly from the theory of RR because it only considers two treatments at a time.

If the categorical treatment variable is ordinal with a meaningful numerical scale, we can use the quadratic regression model of equation (2.3) suggested by HI. However, such a model is restrictive because the slope for GPS in the model changes in a particular way across the treatment levels. Figure 2.2 shows that this assumption may be too strong to justify in practice.

The usefulness of the covariance adjustment GPS for categorical variables is limited by our ability to fit multiple regression models with limited data. When the treatment takes a large number of values, the method may be infeasible. This problem is even more acute for continuous treatments where it is simply impossible to fit a separate regression model for each observed treatment level. We now discuss the covariate adjustment for continuous treatments.

2.5.2 Covariance adjustment GPS for continuous treatments

To use covariance adjustment with a continuous treatment variable, we propose to subclassify the data on the treatment variable rather than on the GPS or the P-FUNCTION. To facilitate the computation of standard errors via bootstrap (see below), we form the subclasses using the theoretical quantiles of the fitted treatment assignment model. This is typically easy to accomplish via Monte Carlo. We draw a large sample from the fitted treatment assignment model with parameters fixed at their fitted values and covariates sampled from their observed values and estimate the theoretical quantiles based on this sample. We also compute the theoretical median, or its Monte Carlo approximation, within each subclass and denote it as t_s for $s = 1, \ldots, S$ with S the number of subclasses.

With the subclassifed data in hand, we fit the model defined in equation (2.14) separately for each subclass. Alternatively, we can use a more flexible model. Here, we consider both quadratic regression, i.e., $Y \sim \alpha_t + \beta_t \hat{R} + \gamma_t \hat{R}^2$, and the SCM of $E[Y|\hat{R}] = f_s(\hat{R})$. We then

compute the GPS for each unit at the median treatment value within each subclass, i.e., $\hat{r}(t_s, \mathbf{X}_i)$ for $i = 1, \ldots, n$ and $s = 1, \ldots, S$. Finally, we estimate the DRF by computing $\hat{E}\{Y(t_s)\}$ for each t_s using equation (2.15) or an appropriate generalization of it if a different response model is used. The derivative of the DRF at t_s can be estimated as in (2.10). Notice that the grid values at which we compute the DRF are different than those advocated by HI. Ours are based on percentiles of the fitted treatment assignment model, whereas theirs are equally spaced in the range of observed treatments.

The standard bias-variance tradeoff arises when selecting the number of subclasses, S. We generally defer to Cochran's advice and use about five (Cochran, 1968). Sensitivity to the choice of S can be quantified by repeating the entire procedure with S equal to approximately three and ten. One source of bias in this procedure results from using units with a range of treatment values to fit the model given in equation (2.14) (or a more flexible version of it). This bias will be especially acute in subclasses with a relatively wide range of the treatment value. If the distribution of the treatment has tails in either direction this correspond to extreme evaluation points of the DRF, t_1 and t_S. Thus, in some cases, we might want to increase the number of subclasses, especially when the extremities of the DRF are of interest. This point is illustrated in Sections 2.5.4.

We approximate the standard errors of the estimated DRF and its estimated derivative via bootstrap resampling. We resample the data, fit the treatment model, subclassify, and compute the DRF and its derivative for each resampled data as described above. We use the same evaluation points, $t_1, \ldots, t_S$ for each resampled data set. Because both the treatment assignment model and the response model are fitted to each bootstrap sample, this procedure accounts for both sources of uncertainty.

2.5.3 Covariance adjustment P-FUNCTION for continuous treatments

As shown in Section 2.3.5, the P-FUNCTION allows a stronger form of *strong ignorability of the treatment assignment given the propensity function* than the GPS. Thus, we could replace the GPS used in the covariance adjustment GPS method by P-FUNCTION and still get a legitimate method, which we define to be the covariance adjustment P-FUNCTION. In particular, firstly of all we subclassify the data on the treatment variable. Secondly, within each subclass, we fit a seperate response model, for example SCM of,

$$E[Y|\hat{\boldsymbol{\theta}}] = f_s(\hat{\boldsymbol{\theta}}) \tag{2.16}$$

Lastly, we estimate the DRF at the median treatment value of each subclass using

$$\hat{E}\{Y(t_s)\} \;=\; \frac{1}{n}\sum_{i=1}^{n} \hat{f}_s(\hat{\boldsymbol{\theta}}_i) \tag{2.17}$$

The standard errors of the estimated DRF and its estimated derivative are also approximated via bootstrap resampling, which is completely analogous to the covariance adjustment GPS method.

2.5.4 The numerical performance of Covariance Adjustment GPS and P-FUNCTION Method

We now examine the performance of covariance adjustment GPS and P-FUNCTION in Simulations I, II, and III. In Simulation I, we again use the the correct treatment assignment model and $S = 10$ subclasses with grid points at the $5\%, 15\%, \ldots,$ and 95% quantiles of T. (Using $S = 5$ or 15 gives similar results.) We compare three within subclass response models, which are (i) $Y \sim R$, (ii) $Y \sim R + R^2$, and (iii) $Y \sim f(R)$, where $f(\cdot)$ is a SCM for the covariance adjustment GPS method. For the covariance adjustment P-FUNCTION

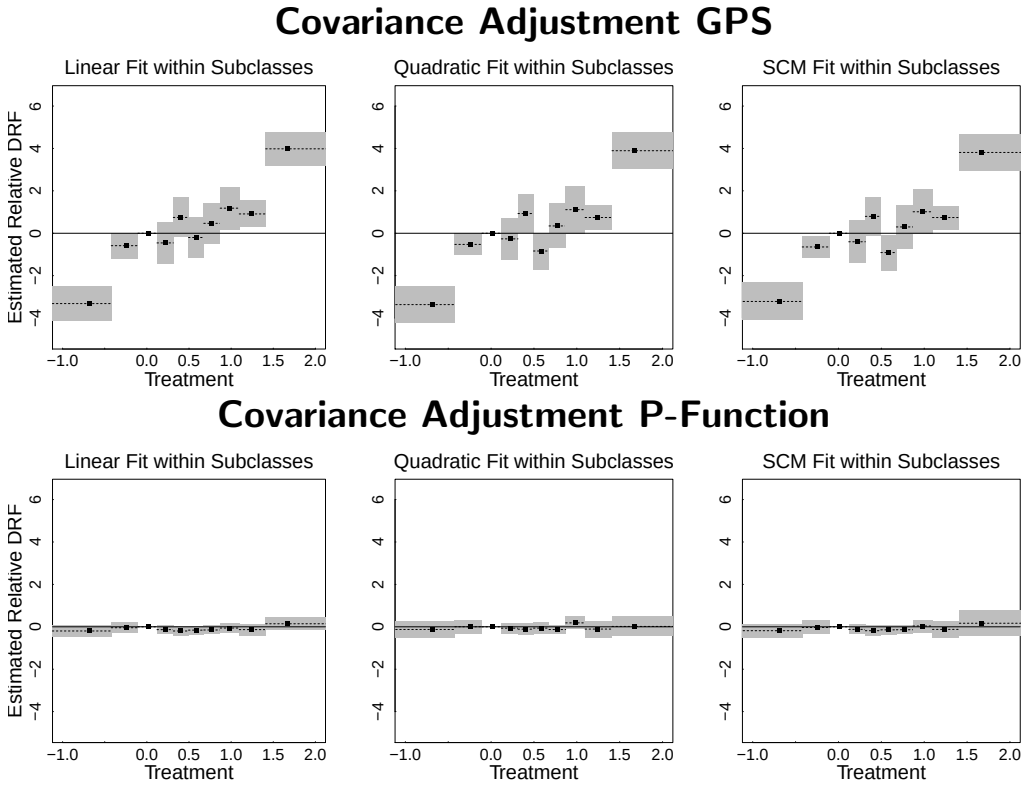

Figure 2.10: Estimated Relative DRFs in Simulation Study I for the Covariance Adjustment GPS and P-FUNCTION Method. The three columns correspond to the three within subclass models. In all plots the solid (dashed) lines represent the true (fitted) relative DRF and 95% confidence bands based on 1000 bootstrap replications are plotted in grey.

method, we simply replace the GPS R by the P-FUNCTION $\boldsymbol{\theta}$ in the response models. The results are shown in Figure 2.10. The three response models are labelled linear, quadratic, and SCM fit within subclasses, respectively. For the covariance adjustment GPS method, the response models are conditional on R, rather than on T as in Section 2.3.1 because covariance adjustment GPS subclassifies on T. As mentioned in Section 2.5.2, the fitted relative DRF exhibit bias in extreme subclasses owing to the relatively large range of treatment levels in these classes. On the other hand, the covariance adjustment P-FUNCTION method works very well in all subclasses. Because the three within subclass models used with both methods lead to very similar fits, we only present results for the quadratic model in the rest of this article.

Figure 2.11 shows the estimated relative DRFs in Simualtion II. We still use the correct

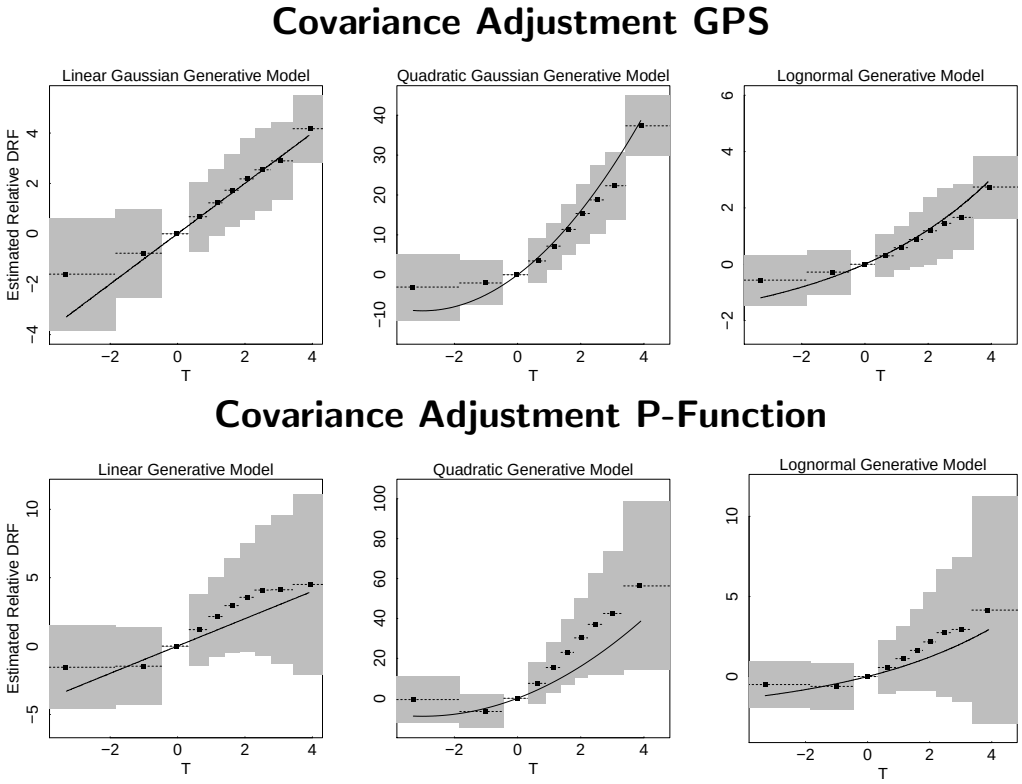

Figure 2.11: Estimated Relative DRFs in Simulation Study II for the Covariance Adjustment GPS and P-FUNCTION Method. Solid lines represent the true relative DRF and dashed lines the average of the fitted relative DRFs across 1000 simulations. Points represent the theoretical quantiles of T used to construct the subclasses. The grey shaded regions represent pointwise intervals containing 95% of the 1000 fitted relative DRFs. Note that the scale of the y-axis for the first row is the same Figure 2.3 and 2.4 while the second row is plotted in a different scale as the covariance adjustment P-FUNCTION method shows significantly larger standard deviation than all other methods.

treatment assignment model and a quadratic response model for both covariance adjustment GPS and P-FUNCTION method within each of $S = 10$ subclasses. (Using $S = 7$ or 13 and/or the other two within subclass models yields similar results.) For the covariance adjustment GPS method, the result is similar to Simualtion I: except in the two most extreme subclasses, the estimated DRF appears to be essentially unbiased. The fitted relative DRF deteriorates in the extreme, more heterogeneous treatment subclasses. Because the distribution of treatment is left skewed, this is less of a problem for the right-most than for the left-most subclass. However, unlike in Simuation study I, the covariance adjustment P-FUNCTION method is now outperformed by the covariance adjustment GPS method.

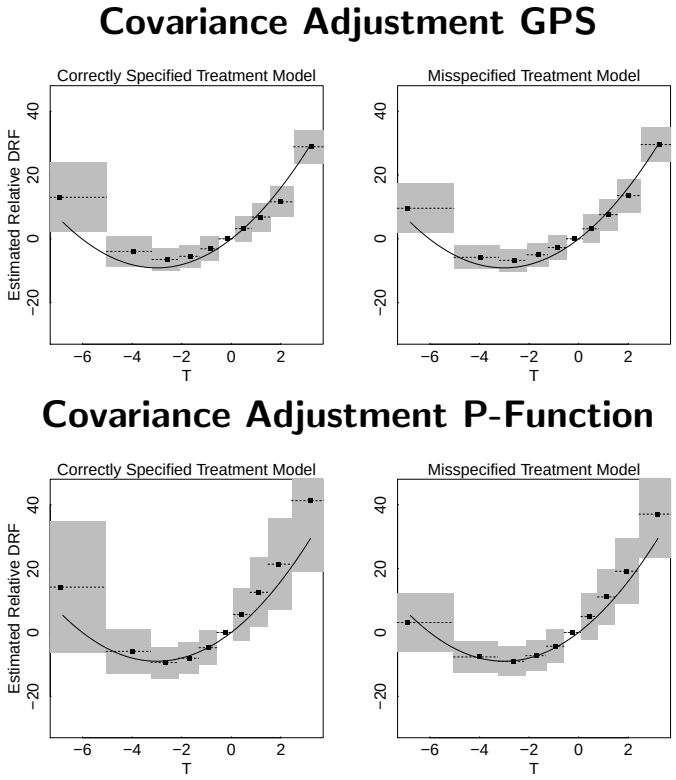

Figure 2.12: Estimated Relative DRFs in Simulation Study III for the Covariance Adjustment GPS and P-FUNCTION Method. Solid lines represent the true relative DRF and dashed lines the average of the fitted relative DRFs across 1000 simulations. Points represent the theoretical quantiles of T used to construct the subclasses. The grey shaded regions represent pointwise intervals containing 95% of the 1000 fitted relative DRFs.

Its estimated DRF is not only biased in most subclasses but also shows significant larger variance than all available methods (Note that we use a different scale on the y-axis in Figure 2.11 to account for such large variance).

Interestingly, the covariance adjustment GPS and P-FUNCTION methods perform about the same in Simulation III, which is shown in Figure 2.12. In this simulation study, we use both correctly and misspecified treatment models. Under the correctly specified treatment model, we use a quadratic response model $Y \sim R + R^2$ for the covariance adjustment GPS method and $Y \sim \theta_1 + \theta_1^2 + \theta_2 + \theta_2^2 + \theta_1 \cdot \theta_2$ for the covariance adjustment P-FUNCTION method within each of the $S = 10$ subclasses, where (θ_1, θ_2) consists the P-FUNCTION in this case. Under the misspecified treatment model, the response model stays the same for

the covariance adjustment GPS method (although the estimated GPS, $\hat{R}$, differs) while for the covariance adjustment P-FUNCTION method it turns into $Y \sim \theta_1 + \theta_1^2$. We simply discard θ_2 in this case as it is no longer a component of the P-FUNCTION.

2.6 Example: The effect of smoking on medical expenditures

2.6.1 Background

We now illustrate the available methods by estimating the DRF of smoking on annual medical expenditures. The data we use were extracted from the 1987 National Medical Expenditure Survey (NMES) by Johnson *et al.* (2003). Its details information about frequency and duration of smoking allows us to continuously distinguish among smokers and estimate the effects of smoking as a function of how much they smoke. The response variable, medical costs, is verified by multiple interviews and additional data from clinicians and hospitals. IvD used the propensity function to estimate the average effect of smoking on medical expenditures. We extend their analysis and study estimation of the full DRF. Like IvD, we adjust for the following subject-level covariates: age at the times of the survey, age when the individual started smoking, gender, race (white, black, other), marriage status (married, widowed, divorced, separated, never married), educational level (college graduate, some college, high school graduate, other), census region (Northeast, Midwest, South, West), poverty status (poor, near poor, low income, middle income, high income), and seat belt usage (rarely, sometimes, always/almost always).

To measure the cumulative exposure to smoking based on the self-reported smoking frequency

and duration, Johnson *et al.* (2003) proposed the variable of `packyear`, defined as

$$\texttt{packyear} = \frac{\text{number of cigarettes per day}}{20} \times \text{number of years smoked.} \tag{2.18}$$

We use log(`packyear`) as our treatment variable. We follow Johnson *et al.* (2003) and IvD and discard all individuals with missing values and conduct a complete-case analysis, yielding a sample of 9,073 smokers. Although in general complete-case propensity-score-based analyses produce biased causal inference unless the data are missing completely at random (D'Agostino and Rubin, 2000), Johnson *et al.* (2003) showed that accounting for the missing data using multiple imputation did not significantly affect their results.

Because the observed response variable, self-reported medical expenditure, denoted Y, is semicontinuous, we use the two-part model of Duan *et al.* (1983). This involves first modeling the probability of spending some money on medical care, $\Pr(Y > 0 \mid T, \boldsymbol{X})$, where $T = \log(\texttt{packyear})$, and $\boldsymbol{X}$ represents the covariates; and then modeling the conditional distribution of Y given T and $\boldsymbol{X}$ for those who reportedf positive medical expenditure. To illustrate and compare methods for computing the DRF, we concentrate on the second part of this model. Because the distribution of Y is skewed, we consider the model $p(\log(Y) \mid Y > 0, T, \boldsymbol{X})$.

For our treatment assignment model, we use a Gaussian linear regression adjusted for all available covariates and the second order terms of two age covariates. The model was fitted using sampling weights provided with the original data. This is the same treatment assignment model used by IvD who demonstrate that it achieves adequate balance.

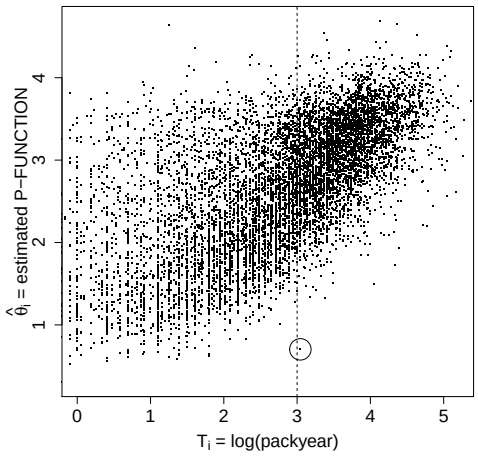

Figure 2.13: A Diagnostic for SCM(P-FUNCTION). Because the range of the $\hat{\theta}_i$ when $T_i > 3$ is less than the overall range of $\hat{\theta}_i$ estimating the DRF for $t > 3$ involves extrapolation under the SCM and thus possible bias. The single individual with T_i slightly larger than three and $\hat{\theta}_i$ less than one is circled. Although this datapoint may mitigate bias for t near three, the fitted DRF for $t > 3$ may be seriously biased.

2.6.2 Simulation study based on the smoking data

This simulation study aims to mimic the characteristics of the actual data with the goal of comparing the statistical properties of the proposed methods in as realistic a setting as possible. In particular, we do not alter the observed covariates or treatment and use the same fitted treatment model used by IvD. Figure 2.13 presents a scatter plot of the observed treatment variable, $T_i = \log(\texttt{packyear})$, and the values of the P-FUNCTION from the fitted treatment assignment model, $\hat{\theta}_i$. As discussed in Section 2.4.1, this plot can be used as a diagnostic for SCM(P-FUNCTION). Recall that this estimate requires that we fit a SCM to predict the response variable as a function of T_i and $\hat{\theta}_i$. To estimate the DRF at t we must evaluate the fitted SCM at $(t, \hat{\theta}_i)$ using each observed $\hat{\theta}_i$ in the data set. This involves extrapolation and thus possible bias if the range of θ_i at a particular value of t is less than the overall range of θ_i. Judging from Figure 2.13, this is a concern for t greater than about three. There is a solitary individual with T_i slightly above three and $\hat{\theta}_1 < 1$ that is circled in Figure 2.13. Even this single point can guard against significant extrapolation bias for t

less than three, but the concern remains for larger t. We emphasize that this diagnostic is preformed before the response model is fit.

To explore the robustness of the methods to different DRFs, we simulate the response variable under three known DRFs and attempt to reconstruct them using HI, SCM(GPS), IPW_0, IPW_{SW}, SCM(P-FUNCTION), covariance adjustment GPS and P-FUNCTION. In particular, we assume $\log(Y_i(t)) \sim \mathcal{N}(E[\log(Y_i(t))], 0.5^2)$ with $t = \log(\texttt{packyear})$ and consider three functional forms for $E[\log(Y_i(t))]$:

$$\begin{aligned}
\text{Quadratic DRF}: \quad E[\log(Y_i(t))] &= \frac{4}{25} \cdot t^2 + [\log(\texttt{age}_i)]^2 \\
\text{Piecewise} - \text{Linear DRF}: \quad E[\log(Y_i(t))] &= \begin{cases} -4 - 0.5 \cdot t + [\log(\texttt{age}_i)]^2, & t \leq 2 \\ -5 - 2.3 \cdot (t-2) + [\log(\texttt{age}_i)]^2, & t > 2, \end{cases} \\
\text{Hockey} - \text{Stick DRF}: \quad E[\log(Y_i(t))] &= \begin{cases} -8.1 + [\log(\texttt{age}_i)]^2, & t \leq 3 \\ -8.1 + 1.5 \cdot (t-3)^2 + [\log(\texttt{age}_i)]^2, & t > 3, \end{cases}
\end{aligned}$$

where `age` is the age at the time of the survey. We include `age` because it is the covariate most correlated with $\log(\texttt{packyear})$ and thus most able to bias a naive analysis. Each of the response models was fitted using the sampling weights.[3]

Each of the seven methods was fitted to one data set generated under each of the three DRFs. We evaluate each DRF at ten points equally spaced between the 5% and 95% quantiles of $\log(\texttt{packyear})$. The results fitted by all existing and the SCM(P-FUNCTION) methods

[3]When using IPW_0 or IPW_{SW}, we construct new weights by multipling the weights required by IPW and the sampling weights. We also take the sampling weights into account when estimating the marginal distribution of the treatment with IPW_{SW} (using the `density` function in R). Ignoring the sampling weights leads to similar results.

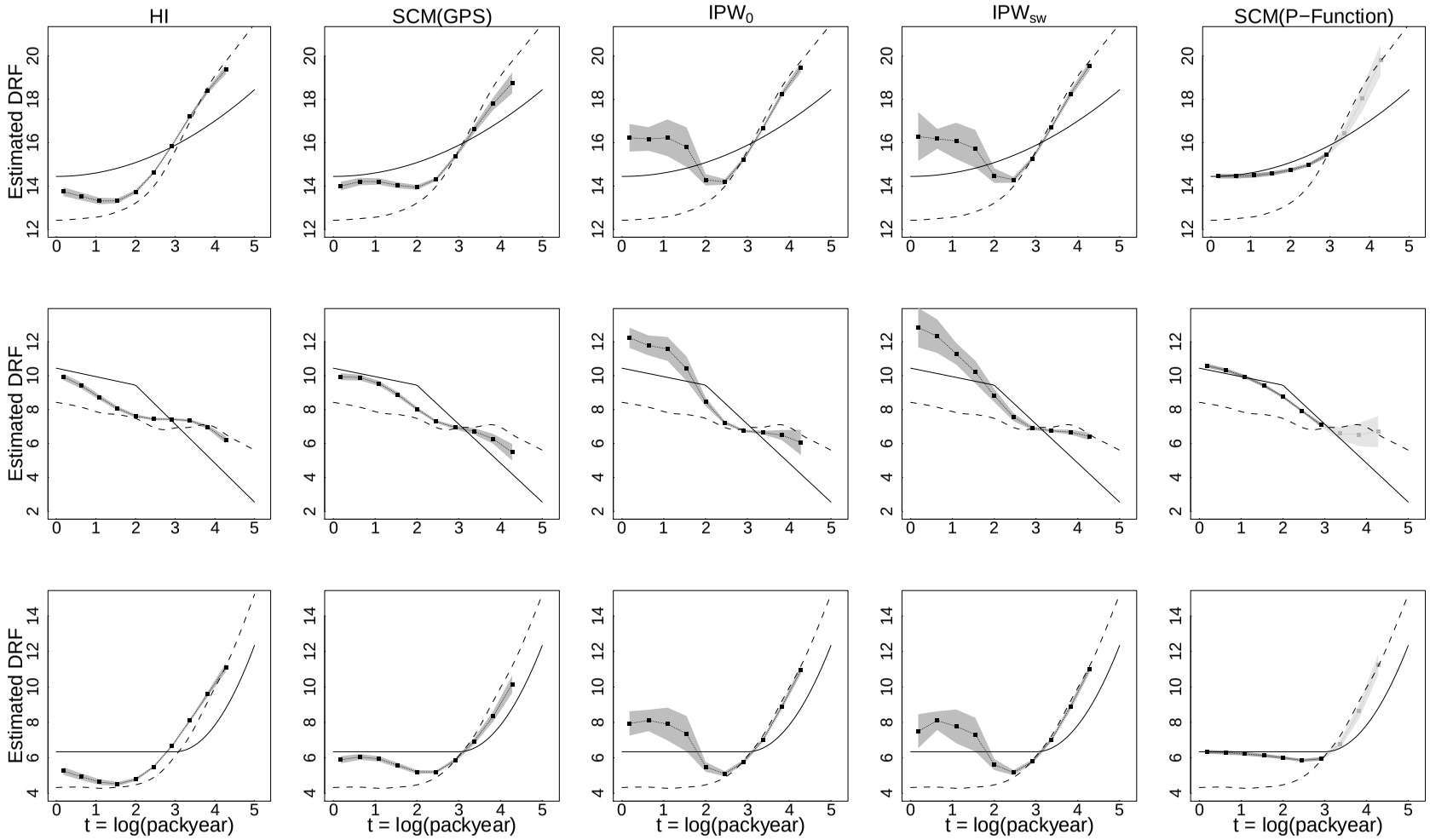

Figure 2.14: Estimated DRFs for the Simulation Based on Smoking Data. The five columns correspond to the method of HI, SCM(GPS), IPW_{SW}, IPW_0 and SCM(P-FUNCTION) respectively. In all plots the dotted lines with bullets correspond to the fitted DRF. The true DRF is plotted as solid lines while dashed lines represent the fitted SCM of $\log(Y)$ on T, unadjusted for the covariates. The evaluation points are evenly-spaced in t. The 95% asymptotic confidence bands plotted in grey are based on 1000 bootstrap replications. A lighter shade of grey is used in the right-most column for $t > 3$ because the estimate is less reliable in this region. The performance of the SCM(P-FUNCTION) clearly dominates the other methods, especially for $t < 3$.

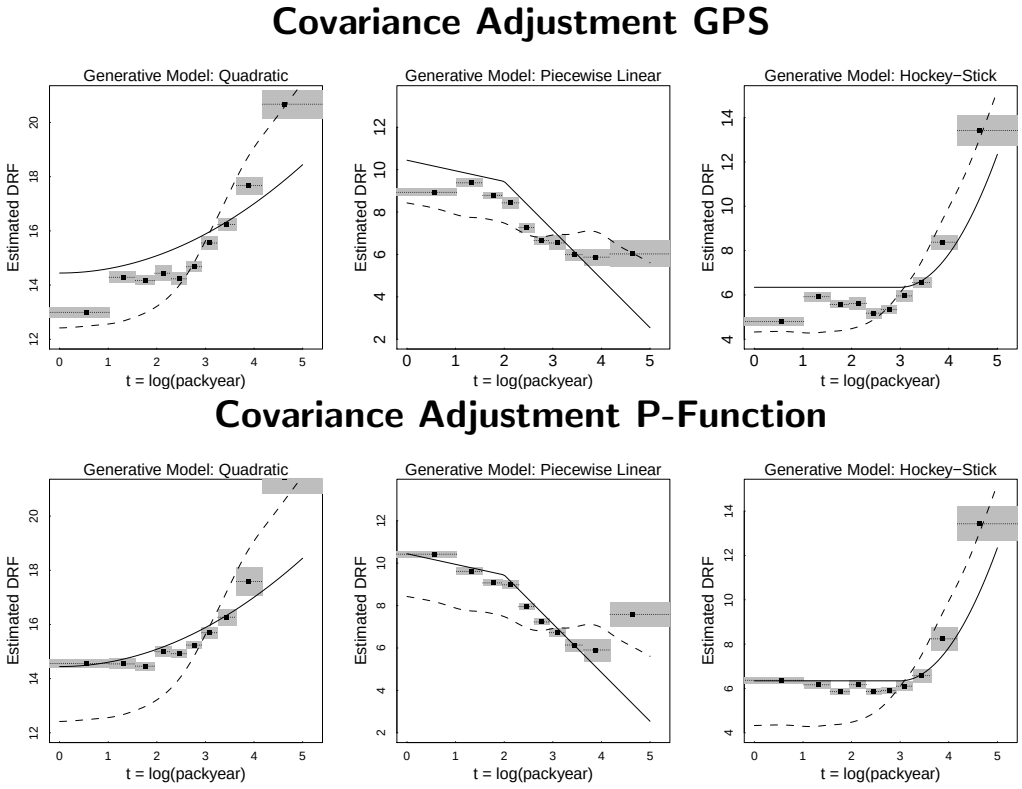

Figure 2.15: Estimated DRF for the Simulation Based on Smoking Data Using the Covariance Adjustment GPS and P-FUNCTION Method. In all plots the dotted lines with bullets correspond to the fitted DRF. The true DRF is plotted as solid lines while dashed lines represent the fitted SCM of $\log(Y)$ on T, unadjusted for the covariates. Evaluation points are based on the theoretical quantiles of $\log(\texttt{packyear})$. The 95% asymptotic confidence bands plotted in grey are based on 1000 bootstrap replications.

appear in Figure 2.14 [4] in which rows correspond to the three generative models and columns represent the method used to fit the DRF. In Figure 2.15, where we show the results under covariance adjustment GPS and P-FUNCTION methods, such order is reversed. In all plots, the true DRF is plotted as a solid line and a directly fitted SCM of $\log(Y)$ on T as a dashed line. This SCM fit is a simple bench mark; it does not account for covariates in any way, in particular it does not adjust for any summary of the treatment assignment model. The fitted DRFs are plotted with dotted lines and bullets indicate the grid where the estimates are evaluated. The shaded regions represent 95% point-wise bootstrap confidence intervals.

[4]The general shape of the estimated DRF fitted by the covariance adjustment GPS and P-FUNCTION are very similar to that fitted by SCM(GPS) and SCM(P-FUNCTION). But the blocky nature of the covariance adjustment method may lead many users to prefer the smooth fitted DRF obtained with SCM method.

The diagnostic described in Figure 2.13 indicates possible bias in the SCM(P-FUNCTION) method for $t > 3$. Thus, we plot the fit in this region in light grey to emphasize its potential unreliability.

The HI fit misses the true DRF under all three generative models, even the quadratic DRF which coincides with the parametric dependence of $\log(Y)$ on T under HI's response model. Instead, HI's fitted DRF tends to follow the unadjusted SCM fit of $\log(Y)$ on T. Although SCM(GPS) improve somewhat on HI, it still exhibits a cyclic pattern; notice its cubic-like fits in the first and third rows. Unfortunately, IPW_0 again exhibits instability, although in this case it takes the form of bias rather than variance. Although IPW_{SW} performs significantly better than IPW_0 in our simulation studies, here the methods are essentially indistinguishable. Finally, SCM(P-FUNCTION) closely matches the true DRF under all three generative models, at least for $t < 3$. As discussed above, we suspect bias for $t > 3$ and see that the fitted DRF reverts to the unadjusted SCM in this range. The quality of the fit can be improved still further by increasing the dimension of the basis used in the SCM. We do not purse this strategy, however, for fear of over fitting. Overall, SCM(P-FUNCTION) appears to be the most reliable, especially considering the diagnostic that alerts us the ranges of t where there is the potential for bias.

2.7 Concluding Remarks

Propensity score methods have gained wide popularity among applied researchers in a number of disciplines. Although they were originally designed exclusively for binary treatment regimes, the fact that treatment variables of interest are not binary in many research settings has led to proposals for generalized propensity score methods. These methods are applicable to a variety of non-binary treatment regimes, and their applications are becoming increasingly common.

In this article, we compare the three most frquently used generalized propensity score methods, IPW of Robins *et al.* (2000a), the GPS of Hirano and Imbens (2004) and the P-FUNCTION of Imai and van Dyk (2004), as well as the two GPS-based methods of Flores *et al.* (2012). First, we show that the suggested implementation of the HI method is sensitive to misspecification of the response model. Second, we show that while SCM(GPS) exhibits substantial improvement over HI's method, it remains biased and/or can exhibit a cyclic artifact in some situations. Third, we demonstrate that while Flores *et al.*'s IPW_0 can be highly unstable, using RHB's stabilized weights can significantly improve its performance. Finally, while IvD provides a relatively robust estimate of the average causal effect, its main limitation is its inability to estimate the DRF. We show how to obtain an estimate of the DRF based on the P-FUNCTION and empirically compare its performance to that of the other estimates. We also give an explanation as to why the SCM(P-FUNCTION) method outperforms the SCM(GPS) method. While SCM(P-FUNCTION) preforms well in comparison to other methods, it remains biased in realistic settings. *We emphasize that researchers should be cautious when using any method to estimate the full* DRF *with a continuous treatment in an observational study.*

There are several important challenges that must still be addressed. We have largely assumed that the propensity weights, GPS, and P-FUNCTION can be correctly estimated. This is an optimistic assumption given that modeling a multi-valued or continuous treatment in a high-dimensional covariate space is much more difficult than doing so for a binary treatment (see e.g., Imai and Ratkovic, 2013, for one recent method that addresses this issue). Diagnostic tools developed for the binary treatment case are also not directly applicable to general treatment regimes. Even more challenging is diagnosing misspecification in the response model. As we have illustrated, this can lead to significant bias in the estimated DRF. Our proposals rely on implementing more flexible response models in more natural spaces, but principled diagnostics for the response model remain elusive. Diagnosing and correcting for inbalance in either the P-FUNCTION or the GPS is another difficulty. Since

the subpopulation that has propensity for treatment varies with the dose, the estimated dose response function is in effect the treatment effect on a varying subpopulation. Future research must develop methods for estimating the propensity weights, GPS, and P-FUNCTION in the presence of possible misspecification of the treatment assignment model and the DRF in the presence of possible misspecification of the response model, as well as diagnostics for both models.

Chapter 3

Using Principal Stratification to Adjust for Imperfect Attendance When Estimating the Effect of the Read 180 Program

3.1 Introduction

Causal inference can be a challenge in more than observation studies. Many double-blind placebo-controlled randomized experiments with active drugs suffer from complications caused by treatment noncompliance, i.e., patients might take only part or none of the assigned dose, whether active or placebo. For example, Efron and Feldman (1991) analyzed a subset of the Lipid Research Clinics Coronary Primary Prevention Trial (LRC-CPPT) data, which was a placebo-controlled double-blind randomized clinical trial designed to study the effectiveness of cholestyramine for lowering cholesterol levels. The complication in this analysis is that

most of the patients in the experiment took only part of their assigned dose; this is typical partial-compliance behavior. The only data available are the treatment assignment, the proportion of the assigned drug or placebo taken, and the observed cholesterol reduction. Efron and Feldman found that better compliance is associated with larger reductions in cholesterol levels not only in the Treatment group but also in the Control group.

To properly adjust for this posttreatment variable of compliance, Frangakis and Rubin (2002) propose a general framework based on Rubin's causal model (Holland, 1986) called principal stratification. Jin and Rubin (2008) use this framework to re-analyze this data set. They regard both placebo compliance and chelestyramine compliance as characteristics of patients, just as any pre-treatment covariates; and there exists a so-called "true dose-response relationship", which is the dose response function that would have been observed is dose had been randomly assigned and 100% compliance had been enforced. They then model the compliance and dose response function via Bayesian hierarchical modelling.

The framework of Jin and Rubin (2008) can be applied to various circumstances in both natural science and social sciences and has received increasing attention. A couple of its application includes: Wolfson and Gilbert (2010) study the statistical identifiability of estimands for the surrogate endpoint problem; Long *et al.* (2010) estimate causal effects in trials involving multitreatment arms subject to non-compliance under a Bayesian framework; Mattei and Mealli (2011) investigate new augmented experimental designs to disentangle direct and indirect treatment effects; and Schwartz *et al.* (2011) proprose a Bayesian semiparametric generalization which replaces part of its parametric models by a Dirichlet process mixture model.

In this chapter, we build on the framework of Jin and Rubin (2008) and apply it to assess the effectiveness of the READ 180 program (a mixed-method approach designed to help struggling students). In particular, on the basis of their framework, we introduce a way to adjust pre-treatment covariates into the Bayesian hierarchical models; we then propose two different

estimands to evaluate the average treatment effect accordingly; we also discuss a method to check the influence of the prior distribution suggested by Jin and Rubin. The rest of the chapter is organized in the following manner. Section 3.2 provides a detailed introduction of the background and scientific objective of the READ 180 program. The statistical details of the principal stratification framework and our extension appear in Section 3.3 and 3.4. We perform a simulation study as well as a real data analysis in Section 3.5. A discussion appears in Section 3.6.

3.2 Scientific Objective of the READ 180 Program

READ 180 is a mixed-method approach (Slavin *et al.*, 2008) to literacy instruction that is designed to help struggling readers in grades 4 to 12 improve their word reading efficiency, reading comprehension and vocabulary, and oral reading fluency. In the full 90-min version of READ 180, teachers begin with a 20-30 min whole group lesson and then create small groups of children who participate in three 20-min activities, in which reading practice is scaffolded by computer activities, leveled books, and teacher lessons tailored to the reading level of each small group (Hasselbring and Goin, 2004). To evaluate its effectiveness, a large implementation study was recently conducted in a high-poverty school district located in southeastern Massachusetts (Hartry *et al.*, 2008). Children were recruited from three elementary schools with a large percentage of struggling readers, who were identified as children in grades four to six who scored below proficiency on the Massachusetts Comprehensive Assessment System (MCAS), a standard based assessment of the state English language arts curriculum. Children were eligible for this study if they scored below proficiency on their most recent MCAS English language arts assessment (Grades 3-5). In this study, 294 children received active consent to participate in the study. Of these, black and Latino children comprised over 70% of the sample and the children received free- or reduced-price lunch

comprised 81%.

During the study, participating students are randomized to the READ 180 program and the district after-school program, which are both administered 4 days per week for around 23 weeks from October 2005 to April 2006. There were a total of 20 certified teachers in this study; 10 teachers used the READ 180 materials and 10 teachers used the district after-school program. For both groups of children, the after-school program began with a one-hour session that involved snack and homework assistance. In the second one-hour session, teachers followed the curriculum of either the READ 180 program or the district after-school program. In the present study, the 90-minute READ 180 model was adapted to fit a 60-minute timeframe in order to accommodate the districts after-school program schedule. As a result, READ 180 teachers only implemented three 20-minute rotations of the whole group lesson. The district after-school program, on the other hand, did not involve structured time in the daily schedule to practice reading leveled text. Pretest and posttest measures on reading, vocabulary, comprehension measures for the sample of children in READ 180 and the district after-school program are recorded. Other available pretest covariates include grade, gender, free or reduced-price lunch indicator, and ethinicity.

To estimate the treatment effect of the READ 180 program, Kim *et al.* (2010) conducted ANCOVA on each posttest score using the relevant pretest score as the covariate. Children in the READ 180 program scored significantly higher on the measure of oral reading fluency than the control group. But no significant differences were found between the two groups on the vocabulary and comprehension measure. However, attendance rates (defined as the percentage of attended days) were significantly higher for children in the READ 180 group than for children in the district after-school program. Hence, despite the randomization experimental design, the results from the ANCOVA analysis cannot be treated as causal as we cannot remove the confounding factor of attendance rate. In another word, we now have to answer the question of whether the improvement of the oral reading fluency measure is

due to the effectiveness of the READ 180 program itself or those extra minutes the children in the READ 180 group stayed in the program. Note that this problem is not trivial as the attendance rate is not a pretest measure. Thus, we cannot simply adjust for it as a covariate into the ANCOVA model. In this Chapter, we propose to remove this confounding factor by regarding the attendance rate as the compliance variable and adjust for it using the principal stratification framework.

3.3 Review of the Principal Stratification Method

In this Section, we review the principal stratification framwork for partial compliance (Jin and Rubin, 2008) in the context of the READ 180 analysis. We carefully compare and select relevant assumptions that are reasonbale for the READ 180 dataset. We then propose a generalization of the existing Bayesian hierarchical models that allows adjusting for pretest covariates. We also introduce two estimands based on our generalized hierarchical models to assess the average treatment effect over the entire population.

We follow the notation of the Rubin Causal Model (Holland, 1986) (as in Chapter **??**) and the Rubin principal stratification framework (Frangakis and Rubin, 2002). Formally, we use N to denote the number of students in the sample. Let Z_i represent the treatment assignment for student i ($i = 1, 2, \ldots, N$), where $Z_i = T$ if student i is assigned to the READ 180 group (hereafter treatment), and $Z_i = C$ if student i is assigned to the district after-school group (hereafter control). Then $Y_i(Z_i = T)$ is the potential outcome of student i if assigned treatment and $Y_i(Z_i = C)$ if assigned control. The causal effect of treatment assignment for student i is $E_i = Y_i(Z_i = T) - Y_i(Z_i = C)$, and the average causal effect across all n students is $\bar{E} = (\sum_{i=1}^{N} E_i)/N$. Let $D_i(Z_i = T) \equiv D_i(T)$ denote the proportion of READ 180 classes attended by student i if assigned treatment, and $D_i(C)$ denote the proportion of of READ 180 classes attended by student i if assigned control. Likewise, let $d_i(T)$ denote the

Table 3.1: Principal Stratification Structure of Extended Partial Compliance. The "$\star$" represents observed data while "?" represents missing data.

i	X_i	Z_i	$D_i(T)$	$D_i(C)$	$d_i(T)$	$d_i(C)$	$Y_i(T)$	$Y_i(C)$
1	$\star$	T	$\star$	?	$\star$	?	$\star$	?
2	$\star$	T	$\star$	?	$\star$	?	$\star$	?
3	$\star$	T	$\star$	?	$\star$	?	$\star$	?
4	$\star$	T	$\star$	?	$\star$	?	$\star$	?
5	$\star$	C	?	$\star$	?	$\star$	?	$\star$
6	$\star$	C	?	$\star$	?	$\star$	?	$\star$
7	$\star$	C	?	$\star$	?	$\star$	?	$\star$
8	$\star$	C	?	$\star$	?	$\star$	?	$\star$

proportion of district after-school classes attended by student i if assigned treatment, and $d_i(T)$ denote the proportion of district after-school classes attended by student i if assigned control. Pretreatment covariates such age and gender are denoted as X_i.

The principal strata are defined as the combination of potential compliance pairs $S = [D(T), D(C), d(T), d(C)]$. In general, each principal stratum is composed of units with the same value of $D_i(T), D_i(C), d_i(T)$, and $d_i(C)$. The principal causal effect in stratum S is defined as $\bar{E}_S = \text{AV}E_{i \in S}[Y_i(T) - Y_i(C)]$, i.e., the average causal effect for all observations who fall into this particular principal stratum. Table 3.1 illustrates the general structure of this framework using eight patients evenly distributed into treatment and control. The "$\star$" sign denotes observed data, and the "?" sign denotes unobserved or missing data. For units 1-4 assigned to the treatment group, the potential compliances and outcomes in the treatment arm, $D_i(T)$, $d_i(T)$, and $Y_i(T)$ are observed, whereas the corresponding values in the control arm, $D_i(C)$, $d_i(C)$, and $Y_i(C)$ are missing. The opposite holds for units 5-8 who are assigned to the contrl group.

3.3.1 Assumptions

In addition to the two standard assumptions of the Rubin Causal Model, namely *Stable unit treatment value assumption* (SUTVA; Rubin (1980)) and *Ignorable treatment assignment assumption* (Rubin, 1978) that we have already reviewed in Chapter **??**, the following are the relevant assumptions with regard to extended partial compliance.

1. *Access monotonicity.* We differentiate two levels of access monotonicity.

   (a) *General access monotonicity.* This assumption has two parts: treatment access monotonicity, $D_i(T) \geq D_i(C)$; and placebo access monotonicity, $d_i(T) \leq d_i(C)$. Treatment access monotonicity implies that for every student, the amount of READ 180 classes he attends if assigned treatment will be greater than or equal to the amount of district after-school classes he would take if assigned control. The rationale is that students have more convenient access to the READ 180 classes if they are assigned treatment other than control. Analogously for placebo access monotonicity: the amount of district after-school classes each student attends under control will be greater than or equal to those he would take under treatment, again because control student has more convenient access to district after-school classes.

   (b) *Strong access monotonicty.* This assumption also has two parts: strong treatment access monotonicity, $D_i(C) = 0$; and strong placebo access monotonicity, $d_i(T) = 0$. That is, first, no student of the treatment group has access to the district after-school program, and, second, no student of the control group has access to READ 180 program.

   In the READ 180 analysis, we make the *Strong access monotonicty* assumption as by the design of the study, once a student is assigned to one of the program, he is unable

to switch to the other program or to access the other program for the duration of the study.

2. *Side-effect monotonicity.* There are two versions of this assumption: negative side-effect monotonicity, $D_i(T) < d_i(C)$; and positive side-effect monotonicity, $D_i(T) > d_i(C)$. If the READ 180 program has more negative side-effects than the district after-school program, the amount of classes student i takes under treatment will be no more than the amount of classes he would take under control, and vice versa for positive side-effect monotonicity. Often, one or the other of these assumptions will be reasonable, but certainly not always. In fact, in some cases the sample could be realistically viewed as a mixture of two subgroups, one susceptible to positive side effects and the other susceptible to negative side effects.

   In our analysis, because the READ 180 program involves activities that are tailored to the reading level of small groups of participating students, we expect that it will do a better job at encouraging students stay longer in the class when compared to the un-tailored district after-school program. Thus, we make the positive side-effect monotonicity assumption.

3. *Perfect blind.* This assumption asserts $D_i(T) = d_i(C)$. That is, the amount of READ 180 classes attended by a student under treatment is exactly equal to the amount of district after-school classes he would take if assigned to the control group. This assumption typically requires the treatment to be perceived is identically to the control with absolute no side effects. This assumption can be overly strong for the READ 180 dataset. We do not make this assumption in the following analysis.

4. *Equipercentile equating of compliances.* This assumption states that $D_i(T) = f[d_i(C)] = F_D^{-1}\{F_d[d_i(C)]\}$, where $F_D(\cdot)$ and $F_d(\cdot)$ are the cumulative distribution functions (CDFs) of $D(T)$ and $d(C)$, respectively. In practice, under this assumption, the "equating function" $f(\cdot)$ is estimated by the relationship between the empirical CDFs of observed

$D(T)$ and observed $d(C)$. Note that this one-to-one mapping function $f(\cdot)$ precludes the possibility that two students who attend the same amount of district after-school classes under control have a different attendence rate under treatment, possibly because of different tolerances to the READ 180 program. Thus, although this assumption is a weakening of the *perfect-blind* assumption, it is more restrictive than *side-effect monotonicity* assumption which allows for this possibility. We do not make this assumption for the READ 180 analysis.

In summary, we make the *Stable unit treatment value assumption*, *Ignorable treatment assignment assumption*, *strong access monotonicity* assumption, and the *positive side-effect monotonicity* assumption. Our assumptions are almost identical to those Jin and Rubin (2008) made except that they made the *negative side-effect monotonicity* instead of the *positive side-effect monotonicity* due to the expected negative side-effects of the treatment under their consideration.

3.3.2 Bayesian Hierarchical Model

As we make essentially identical assumptions as Jin and Rubin (2008), we first review their proposed models to estimate the causal effect. In the end of this section, we propose a generalization to improve their current model.

Under these four assumptions, the principal stratum is simplified to $S = [D(T), 0, 0, d(C)]$, or, more simply, $S = [D(T), d(C)] = [D, d]$, i.e., the amount of READ 180 classes a student would attend if he is assigned to the treatment group and the amount of district after-school classes he would attend if assigned to the control group. Recall that for each student i, there is always one missing value for the pair of (D_i, d_i). If the student is assigned to the treatment group, D_i is observed while d_i is missing; if he is assigned to the control group, d_i is observed while D_i is missing. Thus, we rewrite $S \equiv (S_{\text{obs}}, S_{\text{mis}})$. Similarly, we rewrite the potential

outcome $(Y(T), Y(C)) \equiv (Y_{\text{obs}}, Y_{\text{mis}})$. In a Bayesian statistical analysis, the quantity of the interest, the posterior distribution of the parameter $\boldsymbol{\theta}$, can thus be writed as

$$
\begin{aligned}
p(\boldsymbol{\theta}|Z, S_{\text{obs}}, Y_{\text{obs}}) &\propto p(\boldsymbol{\theta}) \cdot p(Z, S_{\text{obs}}, Y_{\text{obs}}|\boldsymbol{\theta}) && (3.1)\\
&\propto p(\boldsymbol{\theta}) \cdot \int\int p(Z, S_{\text{obs}}, S_{\text{mis}}, Y_{\text{obs}}, Y_{\text{mis}}|\boldsymbol{\theta})\ dY_{\text{mis}} \cdot dS_{\text{mis}} && (3.2)\\
&\propto p(\boldsymbol{\theta}) \cdot \int\int p(S_{\text{obs}}, S_{\text{mis}}, Y_{\text{obs}}, Y_{\text{mis}}|\boldsymbol{\theta}, Z) \cdot p(Z)\ dY_{\text{mis}} \cdot dS_{\text{mis}} && (3.3)\\
&\propto p(\boldsymbol{\theta}) \cdot \int\int p(S_{\text{obs}}, S_{\text{mis}}, Y_{\text{obs}}, Y_{\text{mis}}|\boldsymbol{\theta})\ dY_{\text{mis}} \cdot dS_{\text{mis}}, && (3.4)
\end{aligned}
$$

where (3.4) follows from the randomization of Z. Analytical inference from (3.4) can be difficult as it involves the integral over S_{mis} and Y_{mis}. Jin and Rubin (2008) instead propose to base inference on the joint posterior distribution of $(\theta, S_{\text{mis}}, Y_{\text{mis}})$ via standard data augmentation technique (Tanner and Wong, 1987). Computationally, they use a Beta distribution for the compliance to treatment, D_i,

$$D_i|\theta \sim \text{Beta}(\alpha_1, \alpha_2). \tag{3.5}$$

When specifying $p(d|D, \theta)$, they inforce the *positive side-effect monotonicity* assumption via a Beta distribution for relative compliance d_i/D_i,

$$\frac{d_i}{D_i}\bigg| D_i, \theta \sim \text{Beta}(\alpha_3, \alpha_4). \tag{3.6}$$

Given the principal stratum (D_i, d_i), they assume a Normal distribution for $Y_i(T)$ with mean linear in D_i and d_i,

$$Y_i(T)|D_i, d_i, \theta \sim \text{N}(\beta_0 + \beta_1 D_i + \beta_2 d_i, \sigma_T^2), \tag{3.7}$$

and a Normal distribution for $Y_i(C)$ with also linear regression on D_i and a linear regression

on d_i,

$$Y_i(C)|D_i, d_i, \theta \sim \mathrm{N}(\gamma_0 + \gamma_1 D_i + \gamma_2 d_i, \sigma_C^2). \tag{3.8}$$

Assuming $Y(T)$ and $Y(C)$ are conditional independent given principal stratum (D, d), the complete data likelihood for student i is,

$$\begin{aligned} p(S_{\mathrm{obs},i}, S_{\mathrm{mis},i}, Y_{\mathrm{obs},i}, Y_{\mathrm{mis},i}|\boldsymbol{\theta}) =& p(Y_i(T)|D_i, d_i, \boldsymbol{\theta}) \cdot p(Y_i(C)|D_i, d_i, \boldsymbol{\theta}) \cdot \\ & p(d_i|D_i, \boldsymbol{\theta}) \cdot p(D_i|\boldsymbol{\theta}) \end{aligned} \tag{3.9}$$

Combining (3.6) $-$ (3.8), the complete-data likelihoood for $\boldsymbol{\theta}$ is

$$\begin{aligned} \prod_i p(S_{\mathrm{obs},i}, S_{\mathrm{mis},i}, Y_{\mathrm{obs},i}, Y_{\mathrm{mis},i}|\boldsymbol{\theta}) = & \prod_i \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\ D_i^{\alpha_1-1}(1-D_i)^{\alpha_2-1}\frac{\Gamma(\alpha_3+\alpha_4)}{\Gamma(\alpha_3)\Gamma(\alpha_4)} \\ & \times \Big(\frac{d_i}{D_i}\Big)^{\alpha_3-1}\Big(1-\frac{d_i}{D_i}\Big)^{\alpha_4-1}\frac{1}{D_i} \\ & \times \prod_{i\in\{Z_i=C\}} \frac{1}{\sqrt{2\pi\sigma_T^2}}\exp\Big[-\frac{(Y_i(T)-\beta_0-\beta_1 D_i-\beta_2 d_i)^2}{2\sigma_C^2}\Big] \\ & \times \prod_{i\in\{Z_i=T\}} \frac{1}{\sqrt{2\pi\sigma_C^2}}\exp\Big[-\frac{(Y_i(C)-\gamma_0-\gamma_1 D_i-\gamma_2 d_i)^2}{2\sigma_T^2}\Big]. \end{aligned} \tag{3.10}$$

For the prior distributions, they assume

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma}) = p(\boldsymbol{\alpha}) \cdot p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma}).$$

Of this, $p(\boldsymbol{\alpha})$ are specified through hypothetical data points. In particular, this corresponds to adding to the observed-data likelihood function six extra observations of (D, d) selected in the following way: First, we construct a dataset of (D, d) values for all the students in the sample, where missing parts are computed according to the equipercentile equating assumption; Second, we select the 1st, 21st, 41st, 61st, 81st, and 100th percentiles in this

dataset and add these six prior data points (with complete (D, d) values but missing all Y values) to the actual data. On the other hand, we use a improper prior for the other parameters, i.e., $p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma}) \propto (\sigma_C \sigma_T)^{-2}$. This is a standard choice for Bayesian regressions (Gelman *et al.*, 2004). In another word, we are equivalently using non-informative priors for all parameters except for $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$. Jin and Rubin (2008) believe this prior is reasonable because it weakly pulls the posterior distribution of $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ toward the equipercentile equating assumption. We will go back to this point in section 3.4.3.

Jin and Rubin (2008) use the average principal causal effect $E[Y(T)-Y(C)|S] = \mathrm{AV}E_{i\in S}[Y_i(T)-Y_i(C)]$ as their estimand. For a given principal strata (D, d), this is

$$\beta_0 + \beta_1 D + \beta_2 d - \gamma_0 - \gamma_1 D - \gamma_2 d \tag{3.11}$$

according to (3.7) and (3.8). If we can get a sample $(\boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\gamma}^{(k)}, \boldsymbol{\sigma}^{(k)})$, $k = 1, \ldots, K$ from their joint posterior distribution, (3.11) can then be estimated via

$$\frac{1}{K}\sum_{k=1}^{K}\{\beta_0^{(k)} + \beta_1^{(k)} D + \beta_2^{(k)} d - \gamma_0^{(k)} - \gamma_1^{(k)} D - \gamma_2^{(k)} d\}. \tag{3.12}$$

3.3.3 Computation

To explore the joint posterior distribution of the parameters, Jin and Rubin (2008) apply Markov chain Monte Carlo (MCMC) using the basic idea of the Gibbs sampler (Geman and Geman, 1984), which involves the following steps in each iteration:

1. Given the parameter $\boldsymbol{\theta}$ and observed data, draw the missing data d_i or D_i for $i =$

$1, \ldots, N$. For the treatment group members, we draw $d_i^{(t)}$ from the distribution

$$\begin{aligned} d_i|\boldsymbol{\theta}, D_i, Y_i(T) \;\propto\; & d_i^{\alpha_3-1} \cdot (D_i - d_i)^{\alpha_4-1} \\ & \cdot \exp\left[-\frac{(Y_i(T) - \beta_0 - \beta_1 D_i - \beta_2 d_i - \beta_3 X_{1i} - \beta_4 X_{2i})^2}{2\sigma_T^2}\right]. \end{aligned} \tag{3.13}$$

via the Metropolis-Hastings method. Similarly, we draw $D_i^{(t)}$ for the control group members from the following distribution

$$\begin{aligned} D_i|\boldsymbol{\theta}, D_i, Y_i(C) \;\propto\; & D_i^{\alpha_1-\alpha_3-\alpha_4} \cdot (1 - D_i)^{\alpha_2-1} \cdot (D_i - d_i)^{\alpha 4-1} \\ & \cdot \exp\left[-\frac{(Y_i(C) - \gamma_0 - \gamma_1 D_i - \gamma_2 d_i - \gamma_3 X_{1i} - \gamma_4 X_{2i})^2}{2\sigma_C^2}\right]. \end{aligned} \tag{3.14}$$

2. Given the $D_i, d_i, Y_{\text{obs},i}$, and other parameters, draw the parameters $\alpha_1, \alpha_2, \alpha_3$, and α_4. Since the methods for $\alpha_1, \alpha_2, \alpha_3$, and α_4 are similar, we use α_1 as an example which involves drawing α_1 from the following distribution:

$$\alpha_1|D_i, d_i, Y_{\text{obs},i}, \alpha_2 \;\propto\; \prod_i \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)} d_i^{\alpha_1-1} \tag{3.15}$$

Because this is also not a standard distribution, we still use the Metropolis-Hasting. In particular, we first draw $\alpha_1^\star$ from a truncated normal distribution, $\alpha_1^\star \sim \text{N}(\alpha_1^{(k-1)})$, $\alpha_1^\star > 0$ (k is the MCMC iteration index). We then calculate its normalizing constant $c_1 = \int_0^\infty \phi(x - \alpha_1^{(k-1)})dx$, where $\phi(\cdot)$ represents the pdf function for a standard Normal distribution. Next, we calculate the normalizing constant for the left-truncated normal distribution $\text{N}(\alpha_1^\star, 1)$ of "jumping back": $c_1^\star = \int_0^\infty \phi(x - \alpha_1^\star)dx$. Lastly, we accept $\alpha_1^\star$ with probability

$$p_1 = \frac{c_1}{c_1^\star} \prod \frac{\Gamma(\alpha_1^\star + \alpha_2^{(k-1)})\Gamma(\alpha_1^{(k-1)})}{\Gamma(\alpha_1^{(k-1)} + \alpha_2^{(k-1)})\Gamma(\alpha_1^\star)} d_i^{\alpha_1^\star - \alpha_1^{(k-1)}}.$$

3. Given the $D_i, d_i, Y_{\text{obs},i}$, and other parameters, draw the parameters $\boldsymbol{\beta}, \boldsymbol{\gamma}$, and $\boldsymbol{\sigma}$, which

simply comprises two standard Bayesian regressions (Gelman *et al.*, 2004).

3.4 Generalization of the Principal Stratification Method

As you might have noticed, none of the hierarchical models (3.5) − (3.8) involves the adjustment of pretreatment covariates. In this section, we introduce a generalization that allows for utilizing those extra pretreatment information in (3.7) and (3.8). In addition, we propose another causal effect estimand based on (3.11) which is able to estimate the average treatment effect over the entire population. We also suggest another way to check the influence of the hypothetical complete data prior for $\boldsymbol{\alpha}$.

3.4.1 Covariate Adjustment

First, without loss of generality, we update the regression model of (3.7) and (3.8) using

$$Y_i(T)|D_i, d_i, \theta \sim \mathrm{N}(\beta_0 + \beta_1 D_i + \beta_2 d_i + \beta_3 X_{1i} + \beta_4 X_{2i}, \sigma_T^2) \tag{3.16}$$

$$Y_i(C)|D_i, d_i, \theta \sim \mathrm{N}(\gamma_0 + \gamma_1 D_i + \gamma_2 d_i + \gamma_3 X_{1i} + \gamma_4 X_{2i}, \sigma_C^2), \tag{3.17}$$

where X_1 and X_2 are two pretreatment covariates. It is straightforward to generalize (3.16) and (3.17) to the cases with more than two covariates. Thus, we stick with the simplest case

with just two covariates for illustration. Computation wise, we replace (3.13) and (3.14) by

$$\begin{aligned} d_i|\boldsymbol{\theta}, D_i, Y_i(T) &\propto d_i^{\alpha_3-1}\cdot(D_i-d_i)^{\alpha_4-1} \\ &\quad\cdot\exp\left[-\frac{(Y_i(T)-\beta_0-\beta_1 D_i-\beta_2 d_i-\beta_3 X_{1i}-\beta_4 X_{2i})^2}{2\sigma_T^2}\right] \qquad (3.18) \\ D_i|\boldsymbol{\theta}, D_i, Y_i(C) &\propto D_i^{\alpha_1-\alpha_3-\alpha_4}\cdot(1-D_i)^{\alpha_2-1}\cdot(D_i-d_i)^{\alpha 4-1} \\ &\quad\cdot\exp\left[-\frac{(Y_i(C)-\gamma_0-\gamma_1 D_i-\gamma_2 d_i-\gamma_3 X_{1i}-\gamma_4 X_{2i})^2}{2\sigma_C^2}\right]. \qquad (3.19) \end{aligned}$$

The other steps of the Gibbs sampler are the same.

Under (3.5),(3.6),(3.16) and (3.17), the principal causal effect within the stratum of (D,d), $E[Y(T)-Y(C)|S]$, is now

$$(\beta_0-\gamma_0)+(\beta_1-\gamma_1)D+(\beta_2-\gamma_2)d+(\beta_3-\gamma_3)E(X_1|D,d)+(\beta_4-\gamma_4)E(X_2|D,d). \quad (3.20)$$

Unfortunately, estimating $E(X_1|D,d)$ and $E(X_2|D,d)$ in practice is nontrivial, especially when D and d are continuous[1]. Here we provide two Monte Carlo estimators for (3.20). The naive estimator approximates $E(X_1|D,d)$ and $E(X_2|D,d)$ by their observed sample means, $\bar{X}_1$ and $\bar{X}_2$, i.e., we assume X_1 and X_2 are independent of (D,d). In this case, (3.20) can be estimated via

$$\frac{1}{K}\sum_{k=1}^{K}(\beta_0^{(k)}-\gamma_0^{(k)})+(\beta_1^{(k)}-\gamma_1^{(k)})D+(\beta_2^{(k)}-\gamma_2^{(k)})d+(\beta_3^{(k)}-\gamma_3^{(k)})\bar{X}_1+(\beta_4^{(k)}-\gamma_4^{(k)})\bar{X}_2, \quad (3.21)$$

where $\boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\gamma}^{(k)}$ $(k=1,\ldots,K)$ is a posterior sample of the relevant parameters. We call it the naive within strata treatment effect estimator.

To avoid the independence assumption between $\boldsymbol{X}$ and (D,d), we model their relationship using linear regression. In particular, if we knew the complete compliance (D,d) for each

[1]Jin and Rubin (2009) discuss how to estimate such quantities when D and d are discrete.

observation, we could fit the following linear regressions in order[2]

$$\begin{aligned} X_1 &\sim \mathrm{N}(\eta_{10} + \eta_{11} D + \eta_{12} d, \sigma_1^2) \\ X_2 | X_1 &\sim \mathrm{N}(\eta_{20} + \eta_{21} D + \eta_{22} d + \eta_{23} X_1, \sigma_2^2). \end{aligned} \tag{3.22}$$

Use (3.22), we can estimate $E(X_1|D,d)$ and $E(X_2|D,d)$ according to the fitted regression,

$$\begin{aligned} \hat{E}(X_1|D,d) &= \hat{\eta}_{10} + \hat{\eta}_{11} D + \hat{\eta}_{12} d \\ \hat{E}(X_2|D,d) &= \hat{\eta}_{20} + \hat{\eta}_{21} D + \hat{\eta}_{22} d + \hat{\eta}_{23} \hat{E}(X_1|D,d) \end{aligned}$$

Of course, we do not know the complete compliance (D,d) in practice. Fortunately, at each iteration of the MCMC run, we sample the missing $D_i^{(k)}$ and $d_i^{(k)}$ for each relevant observation, where $k = 1, \ldots, K$ index the MCMC iterations. Thus, at each MCMC iteration, we can then fit (3.22), record the corresponding regression coefficients as $\boldsymbol{\eta}_1^{(k)}$ and $\boldsymbol{\eta}_2^{(k)}$, and estimate $E(X_1|D,d)$ and $E(X_2|D,d)$ as

$$\begin{aligned} \hat{E}^{(k)}(X_1|D,d) &= \hat{\eta}_{10}^{(k)} + \hat{\eta}_{11}^{(k)} D + \hat{\eta}_{12}^{(k)} d \\ \hat{E}^{(k)}(X_2|D,d) &= \hat{\eta}_{20}^{(k)} + \hat{\eta}_{21}^{(k)} D + \hat{\eta}_{22}^{(k)} d + \hat{\eta}_{23}^{(k)} \hat{E}^{(k)}(X_1|D,d). \end{aligned} \tag{3.23}$$

We can then estimate (3.21) via,

$$\frac{1}{K} \sum_{k=1}^{K} (\beta_0^{(k)} - \gamma_0^{(k)}) + (\beta_1^{(k)} - \gamma_1^{(k)}) D + (\beta_2^{(k)} - \gamma_2^{(k)}) d + (\beta_3^{(k)} - \gamma_3^{(k)}) \hat{E}^{(k)}(X_1|D,d) + (\beta_4^{(k)} - \gamma_4^{(k)}) \hat{E}^{(k)}(X_2|D,d) \tag{3.24}$$

We call (3.24) the regression within strata treatment effect estimator.

[2]For discrete pre-treatment covariates, we replace the linear regression by a logistic regression.

3.4.2 Global Average Treatment Effect

In addition to the within strata treatment effect, researchers might also be interested in the average treatment effect over the entire population, $E[Y(T) - Y(C)]$. We call it the global treatment effect. According to the law of iterated expectation,

$$E[Y(T) - Y(C)] = E_S\big[E[Y(T) - Y(C)|S]\big], \tag{3.25}$$

where $E_S(\cdot)$ represents the expectation with regard to the distribution of principal strata. In another word, the global treatment effect can be regarded as the integral of the within strata treatment effect over the strata distribution. Mathematically, under our four assumptions, we can rewrite (3.25) as

$$\begin{aligned} E[Y(T) - Y(C)] &= E_{(D,d)}\big\{E[Y(T) - Y(C)|D, d]\big\} \\ &= \int E[Y(T) - Y(C)|D, d] \cdot p(D, d)\ dD\ dd \end{aligned} \tag{3.26}$$

If $E[Y(T) - Y(C)|D, d]$ and $p(D, d)$ are known for all possible strata (D, d), (3.26) can be numerically approximated by

$$\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N} E[Y(T) - Y(C)|D_m, d_n] \cdot p(D_m, d_n) \tag{3.27}$$

where $D_m, m = 1, \ldots, M$ and $d_n, n = 1, \ldots, N$ represent fine grids of the unit interval. Of course in practice, neither $E[Y(T) - Y(C)|D, d]$ nor $p(D, d)$ is known. As discussed earlier, $E[Y(T) - Y(C)|D, d]$ can be estimated via either the naive or regression within strata treatment effect estimator. For $p(D, d)$, if we know $\boldsymbol{\alpha}$,

$$\begin{aligned} p(D, d) &= p(d|D, \alpha_3, \alpha_4) \cdot p(D|\alpha_1, \alpha_2) \\ &= p_{\text{beta}}(\frac{d}{D}, \alpha_3, \alpha_4) \cdot p_{\text{beta}}(D, \alpha_1, \alpha_2) \end{aligned} \tag{3.28}$$

where $p_{\text{beta}}(\cdot)$ represents the density of a Beta distribution. It is then straightforward to estimate (3.28) via

$$\hat{p}(D,d) = \frac{1}{K}\sum_{k=1}^{K} p_{\text{beta}}(\frac{d}{D}, \alpha_3^{(k)}, \alpha_4^{(k)}) \cdot p_{\text{beta}}(D, \alpha_1^{(k)}, \alpha_2^{(k)}). \tag{3.29}$$

As a result, the average treatment effect over the entire population (3.26) can then be estimated via

$$\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N} \hat{E}[Y(T) - Y(C)|D_m, d_n] \cdot \hat{p}(D_m, d_n) \tag{3.30}$$

where $\hat{E}[Y(T) - Y(C)|D_m, d_n]$ are given by either (3.21) or (3.24) and $\hat{p}(D_m, d_n)$ is given by (3.29).

3.4.3 Sensitivity Analysis for the Prior Influence

It is worth noting that in practice non of the samples will have complete observation of both D and d. Thus, our current prior for $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ using hypothetical data points with complete observation of (D, d) might have strong influence on the final result, i.e., even if we only add six extra observations, their relative influence might be huge compared to the total sample size of a couple of hundreds. To access the prior influence of those added extra complete observations, we perform a sensitivity analysis via calculating the global treatment effect under the three, six, and ten extra complete observation priors. Ideally, the results under these three different priors should not differ significantly.

3.5 Simulation And Real Data Analysis

In this section, we firstly perform a simulation study to illustrate the methodology of (3.4.1) and (3.4.2). Then, we will use this method to adjust for the influence of partial attendance on the causal effect estimate of the READ 180 program.

3.5.1 Simulation Study

Before we use the principal stratification method to account for the attendance rate effect on the READ 180 program. We perform a simulation study mimicing the READ 180 data set. In particular, we make the *SUTVA, ignorable treatment assignment, strong access monotonicity*, and *positive side-effect monotonicity* assumptions, use (3.5),(3.6),(3.16) and (3.17) as the generative model, and sample 294 observations using the following parametric model

$$\begin{aligned} D_i &\sim \text{Beta}(\alpha_1, \alpha_2) \\ \left.\frac{d_i}{D_i}\right| D_i &\sim \text{Beta}(\alpha_3, \alpha_4) \\ X_{1i}|D_i, d_i &\sim \text{N}(50, 10^2) \\ X_{2i}|D_i, d_i &\sim \text{N}(30D_i + 70d_i^2, 10^2) \\ Y_i(T)|D_i, d_i &\sim \text{N}(\beta_0 + \beta_1 D_i + \beta_2 d_i + \beta_3 X_{1i} + \beta_4 X_{2i}, 36^2) \\ Y_i(C)|D_i, d_i &\sim \text{N}(\gamma_0 + \gamma_1 D_i + \gamma_2 d_i + \gamma_3 X_{1i} + \gamma_4 X_{2i}, 36^2), \end{aligned} \tag{3.31}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_C, \sigma_T)$, $\boldsymbol{\alpha} = (3.9164, 1.324, 7.21, 1)$, $\boldsymbol{\beta} = (621, 130, 80, 1, 2)$, $\boldsymbol{\gamma} = (621, 100, 80, 1, 1)$, and $(\sigma_T, \sigma_C) = (36, 36)$[3]. We assign 146 observations to the control group and the other 148 units to the treatment group.

The time series plot for 5000 draws of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and σ_T are shown in Figure 3.1[4], from which

[3]The parameter values of the generative models are chosen to mimic the real data set.

[4]The time series plot for $\boldsymbol{\gamma}$ and σ_C are similar to that of $\boldsymbol{\beta}$ and σ_T.

Table 3.2: The 95% Posterior Credible Intervals for All Parameters in Simulation Study. *True value* represents the values of parameters that are used to generate the data; *C.I.* the 95% credible interval; *Median* the median of posterior draws; and *S.D.* the standard deviation of the posterior draws for each parameter.

	True Value	*C.I.*	*Median*	*S.D.*
α_1	3.92	[3.14, 4.63]	3.83	0.38
α_2	1.32	[1.12, 1.63]	1.36	0.13
α_3	7.21	[4.07, 12.47]	6.74	2.18
α_4	1.00	[0.66, 1.74]	1.01	0.28
β_0	621	[574, 654]	614	21
β_3	1	[0.51, 1.81]	1.17	0.33
β_4	2	[1.36, 2.42]	1.89	0.27
γ_0	621	[570, 720]	640	40
γ_3	1	[0.00, 1.27]	0.64	0.32
γ_4	1	[0.10, 1.25]	0.67	0.29
σ_T	36	[27, 40]	35	3.24
σ_C	36	[28, 40]	34	3.03
β_1	130	[-37, 519]	233	147
β_2	80	[-344, 254]	-30	160
γ_1	100	[-267, 279]	-34	169
γ_2	80	[-51, 469]	263	153

we can tell that the autocorrelation for all of the parameters are relatively low. Table 3.2 lists the 95% credible intervals for all parameters. According to Table 3.2, all of the true parameter values used to generate the data are included in their corresponding 95% credible intervals. However, the standard deviations of $\beta_1, \beta_2, \gamma_1, \gamma_2$ are very large, which suggests identifiability problems in models (3.16) and (3.17). We believe this is not a severe problem as we only use these four parameters for prediction.

We then use both (3.21) and (3.24) to evaluate the within strata treatment effect for the simulation study. The strata are defined using the 25%, 50%, 75%, and 90% quantile of D and d respectively. The results are shown in Table 3.3 where panel 1 and 2 correspond to the estimators of (3.21) and (3.24), respectively. For both panels, the three numbers in each cell are the true treatment effect according to (3.20), the estimated treatment effect ((3.21) or (3.24)), and the posterior quantile of true treatment effect. For panel 1, the posterior quantile

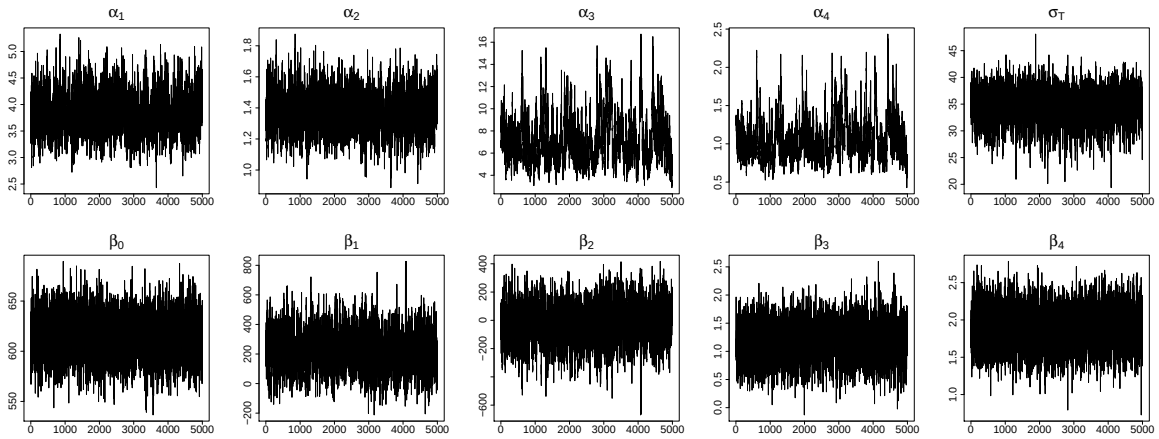

Figure 3.1: Timeserie plot for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and σ_T of the Simulation Study.

of true treatment effect is calculated as the quantile of (3.20) against the empirical posterior distribution of $(\beta_0^{(k)} - \gamma_0^{(k)}) + (\beta_1^{(k)} - \gamma_1^{(k)})D + (\beta_2^{(k)} - \gamma_2^{(k)})d + (\beta_3^{(k)} - \gamma_3^{(k)})\bar{X}_1 + (\beta_4^{(k)} - \gamma_4^{(k)})\bar{X}_2$, i.e., the Monte Carlo sample used to compute (3.21).

According to the posterior quantiles of Table 3.3, although the regression model of $X_2|D, d$ in (3.22) for the regression within strata treatment effect estimator is misspecified, all posterior quantile in panel 2 are within the range of $(2.5\%, 97.5\%)$ which implies that all of the 95% credible interval given by (3.24) contain the true treatment effects. However, this is not the case for the naive within strata treatment effect estimator (3.21). For three pairs of (D, d) strata, the true treatment effect fall outside of its 95% credible interval. Using regression models to adjust for the relationship between $\boldsymbol{X}$ and (D, d) does appear to help even if the parametric form of the regression is not correctly specified.

To estimate the global average treatment effect, we let D_m and d_n be 50 equally spaced points between 0.05 and 0.95 and $K = 500$ in our numerical calculation. Using (3.21) for (3.30), we estimate the average treatment effect as 80.34 with standard deviation 6.52; On the other hand, using (3.24) for (3.30), the mean treatment effect is equal to 82.87 with standard deviation 6.25. The true average treatment effect is equal to 77.1.

To access the prior influence of six added extra complete observations, we calculate the aver-

Table 3.3: Treatment Effect within Strata. Strata are defined as The 25%, 50%, 75%, And 90% Quantile of D and d respectively. For each cell of the table, the three numbers represent the true treatment effect, estimated treatment effect, and the posterior quantile of true treatment effect. The true treatment effects are displayed in bold.

Naive within Strata Treatment Effect Estimator

		d			
		.53	.66	.79	.86
D	.61	**56**, 76, 1.8%			
	.77	**66**, 112, 6.8%	**77**, 80, 29.2%		
	.87	**72**, 134, 11.2%	**83**, 102, 22.0%	**96**, 70, 99.6%	
	.95	**77**, 152, 12.4%	**87**, 120, 22.4%	**101**, 88, 75.4%	**109**, 71, 99.8%

Regression within Strata Treatment Effect Estimator

		d			
		.53	.66	.79	.86
D	.61	**56**, 60, 30.8%			
	.77	**66**, 106, 9.6%	**77**, 83, 14%		
	.87	**72**, 134, 11.2%	**83**, 112, 11.8%	**96**, 89, 82.8%	
	.95	**77**, 157, 11.8%	**87**, 134, 12.0%	**101**, 112, 27.2%	**109**, 100, 83.6%

age treatment effect over the entire population under the three, six, and ten extra complete observation priors. If we add three extra complete observations, the average treatment effect using (3.24) for (3.30) has a mean estimate equal to 84.27 and standard deviation of 6.74; With ten extra complete observation prior, the mean and standard deviation estimate becomes 81.67 and 5.91 respectively. In general, both the bias and the standard deviation of the estimator decrease if we increase the strength of the prior, but in an reasonable manner.

3.5.2 Real Data Analysis

We now present our analysis of the the real READ 180 data. Our main goal is to access the effect of READ 180 program adjusting for the influence of attendance rate. Recall that attendance rate is a posttreatment covariate. The students in the READ 180 program have significantly better attendance compared to those in the district after-school program. A two-sample t-test gives a p-value of 0.04. The average attendance rate are 77.5% and 69.2% for the students in the READ 180 program and the district after-school program,

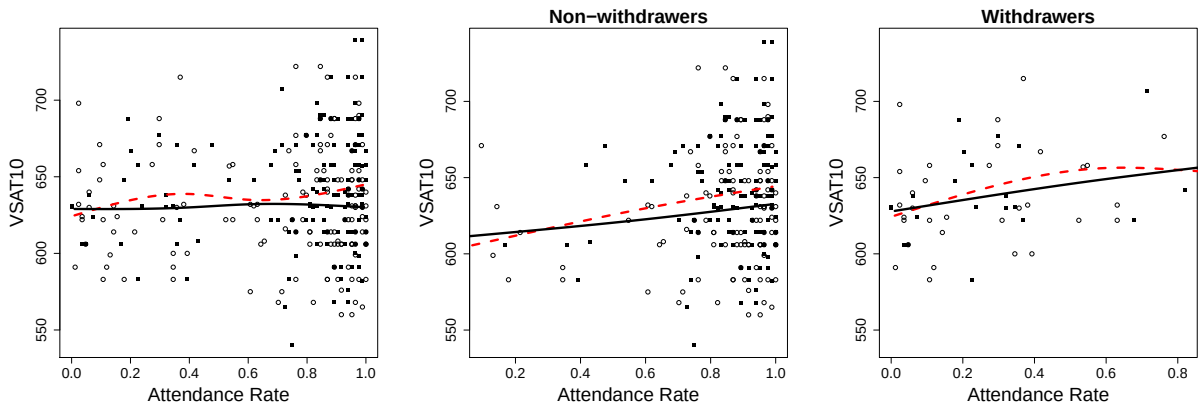

Figure 3.2: The Scatter Plot of VSAT10 And Attendance. The first panel contains all students in the sample, while the second and third contain only non-withdrawer and withdrawer students. For all plots, the black dots represent students who are assigned to the READ 180 group and the white circles represents students who are assigned to the district after-school group. The red dashed line shows the Natural spline fit of VSAT10 on attendance for students in the treatment group. The black solid lines shows the similar fit for students in the control group.

respectively. We use the SAT10 vocabulary scale score (hereafter VSAT10) as the outcome variable in this section.

As discussed in section 3.3.1, we make the *SUTVA, ignorable treatment assignment, strong access monotonicity*, and *positive side-effect monotonicity* assumptions. We use (3.5) and (3.6) to model compliance (attendance rate) under treatment and control group, where D_i represents the attendance rate for student i if he were in the READ 180 program and d_i the attendance rate if he were in the district after-school program[5]. As for the regression model (3.16) and (3.17), in order to specify its parametric form on D and d, we fit two independent natural spline regression models for VSAT10 against attendance rate seperately for students in the treatment and control group. The two fits are shown as the red dashed (treatment group) and black solid (control group) lines in the first panel of Figure 3.2. The fits appears linear for the district after-school program. For the READ 180 program, on the other hand, the relationship is more complicated. For students with low attendance rate (below 40%)

[5]To avoid numerical instability, if a student has perfect attendance (D_i or d_i equal to 1), we re-code D_i or d_i to 0.999; If he has zero attendance (D_i or d_i equal to 0), we re-code D_i or d_i to 0.001.

and high attendance rate (above 60%), the relationships between the two variables appear linear. In addition, we find that most of the students with low attendance rate actually withdrew from the program before the end of the study. To investigate this, we reproduce the natural spline regression fit of VSAT10 on attendance rate for students who did and did not withdrew seperately, see middle and right panel of Figure 3.2. The red dashed and black solid lines again represent the estimated fits for students in the treatment and control group. For students who did not withdraw, the relationships between VSAT10 and attendance rate are linear for both the treatment and control group. Since there are only a few students who withdrew, we confine our attention to the students who did not withdraw and decide the first order of D and d suffice for (3.16) and (3.17). It is worth noting that we cannot simply add a binary variable indicating the withdraw status into (3.16) and (3.17) to adjust for its effect as it is also a posttreatment variable. We provide a discussion about this topic in section 3.6.

To decide which pretreatment covariates we need adjust in (3.16) and (3.17), we first regress VSAT10 linearly on all available pretreatment covariates, of which we then pick the three covariates which are significant. They are adjusted title recognition score, Baseline DIBELS score, and special education status. The special education status is a discrete variable with five levels. Apart from the baseline level (do not have special education), there are four levels of special education. However, three out of the four levels have less than or equal to seven observations (the other level has 47). Thus, we convert the special education status into a binary variable corresponding to either have or do not have special education. In particular, (3.16) and (3.17) now take the form of

$$Y_i(T)|D_i, d_i, \theta \sim \mathrm{N}(\beta_0 + \beta_1 D_i + \beta_2 d_i + \beta_3 X_{1i} + \beta_4 X_{2i} + \beta_5 X_{3i}, \sigma_T^2)$$
$$Y_i(C)|D_i, d_i, \theta \sim \mathrm{N}(\gamma_0 + \gamma_1 D_i + \gamma_2 d_i + \gamma_3 X_{1i} + \gamma_4 X_{2i} + \gamma_5 X_{3i}, \sigma_C^2),$$

where X_1, X_2, and X_3 stand for adjusted title recognition score, Baseline DIBELS score, and

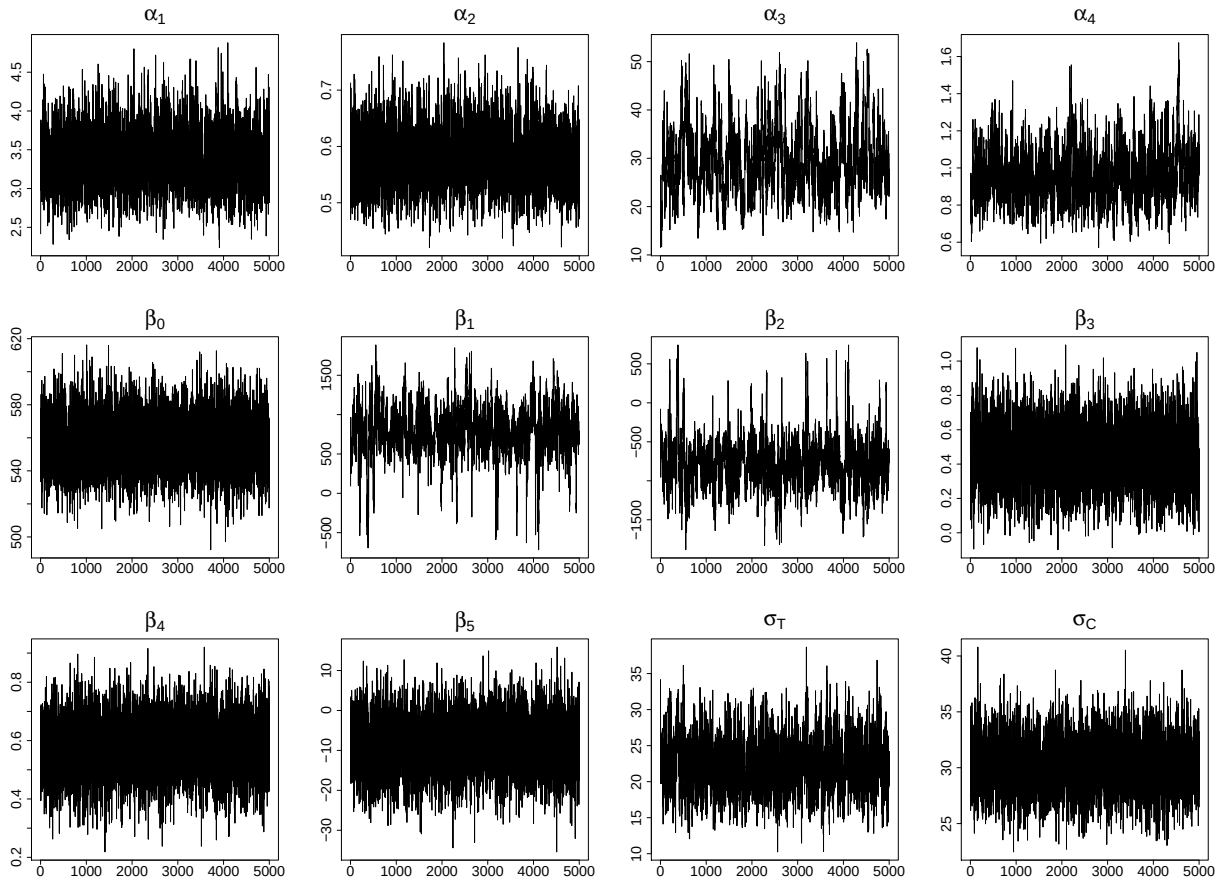

Figure 3.3: Time Series Plot for the Posterior Draws.

special education status, respectively.

Under the priors we discussed in section 3.3.2, we draw 50,000 samples from the joint posterior distribution using the Gibbs sampler in section 3.3.3. Because the autocorrelation for $\boldsymbol{\alpha}$ are relatively high. We thinning the MCMC chain by 10 (we discard everything except every tenth draw of the chain). This gives us a total of 5000 samples. The time series plot of the final 5000 draws for $\boldsymbol{\alpha}, \boldsymbol{\beta}$, and $\boldsymbol{\sigma}$ are shown in Figure 3.3 (the time series plot for $\boldsymbol{\gamma}$ appear similar to that for $\boldsymbol{\beta}$). Overall the auto-correlation is low and the chains perform similar to that of the simulation study. Table 3.4 lists the 95% posterior credible intervals for all parameters.

We then use the regression within strata treatment effect estimator (3.24) to evaluate the

Table 3.4: The 95% Posterior Credible Intervals for All Parameters in the Real Data Analysis. *C.I.* represents the 95% credible interval; *Median* the median of posterior draws; and *S.D.* the standard deviation of the posterior draws for each parameter.

	C.I.	*Median*	*S.D.*
α_1	[2.71, 4.16]	3.34	0.37
α_2	[0.48, 0.69]	0.58	0.05
α_3	[18.22, 44.96]	28.84	6.73
α_4	[0.71, 1.27]	0.94	0.14
β_0	[526, 591]	558	17
β_3	[0.11, 0.80]	0.46	0.18
β_4	[0.38, 0.77]	0.58	0.10
β_4	[-22.29, 4.95]	-8.55	6.91
γ_0	[508, 575]	541	17
γ_3	[-0.15, 0.76]	0.31	0.23
γ_4	[0.42, 0.91]	0.67	0.12
γ_5	[-34.29, -4.29]	-19.17	7.56
σ_T	[15.17, 30.46]	21.89	3.81
σ_C	[25.66, 34.95]	29.98	2.34
β_1	[94, 1392]	800	303
β_2	[-1397, -67]	-795	310
γ_1	[-436, 693]	212	283
γ_2	[-669, 472]	-180	286

effect of the READ 180 program for several values of D and d. In particular, we set $K = 500$ while the values of D and d were chosen to be their 10%, 25%, 50%, 75%, 85%, and 95% quantiles. Table 3.5 reports the estimated principal causal effects, including its standard deviation estimate. Out of the 15 strata, none of the within strata treatment effect is significant (All of the $mean \pm 2S.D$ interval contain zero). To estimate the global average treatment effect, we use (3.24) for (3.30) and let D_m and d_n be 50 equally spaced points between 0.05 and 0.95. We estimate the average treatment effect as 15.37 with standard deviation 6.45. If instead, we add three and ten extra complete observation as the prior for $\boldsymbol{\alpha}$, the average treatment effect estimate appear to be 14.29 and 15.54 with standard deviation 6.05 and 6.59, respectively. The average treatment effect over the entire population does seem to be siginficant.

Table 3.5: Posterior Median And Standard Deviation of Representative Principal Causal Effects. The posterior standard deviation are the figures in parentheses.

		d				
		.10	.81	.899	.959	.989
D	.17	52.33 (28.19)				
	.83	108.59 (62.65)	4.93 (4.89)			
	.90	152.21 (89.65)	48.67 (27.35)	-8.98 (9.52)		
	.96	189.81 (112.83)	86.17 (50.37)	28.52 (16.25)	-10.35 (10.12)	
	.99	208.55 (124.42)	104.91 (61.93)	47.27 (27.48)	8.4 (6.99)	-11.03 (10.47)

3.6 Conclusions And Discussions

In this chapter, we estimate the treatment effect of the READ 180 program on word reading efficiency, reading comprehension, vocabulary, and oral reading fluency for struggling reader in grade 4-12. Our estimate need to account for differences in attendance pattern for the treatment and control group. While standard statistical methods like the intention-to-treat analysis are much simpler, they do not address the different compliance rate between the treatment and control group. Thus, we implement the principal stratification method proposed by Jin and Rubin (2008), which treat both compliance to treatment, D, and compliance to control, d, as psychological characteristics of students and then estimated the expected effect of assignment to treatment versus assignment to control for each type of student jointly defined by D and d. These groups of students are called principal strata.

We carefully review their framework and select the necessary assumptions in the context of the READ 180 study design. While Jin and Rubin (2008) do not make use of any pretreatment covariate in their Bayesian hierarchical models, we propose a generalization that allows for adjusting pretreatment covariate into the regression model of the outcome variable conditional on the principal strata. Accordingly, we introduce two ways to compute the within strata treatment effect for our generalized regression model, which is the causal effect estimator reported by Jin and Rubin (2008). In addition, we propose another causal effect estimator, which averages the within strata treatment effect over the principal strata

distribution and can be regarded as the average treatment effect over the entire population. Based on this global average treatment effect estimator, we discuss a way to do the sensitivity analysis for the potentionally influential hypothetical complete data prior distribution suggested by Jin and Rubin (2008). We apply our generalized principal stratification method to both the READ 180 dataset and a simulation study designed to mimic the real dataset. We find that although none of the within strata treatment effect is significant, the average treatment effect does appear to be significant.

Although the principal stratification method seems to work properly for the READ 180 data, we find there are still a few aspects of the model that require further improvement. First, the parametric models with regard to the compliance rate do not make use of the pretreatment covariates. Jin and Rubin (2008) briefly mentions the possibility to further rewrite the current Beta parameters as a exponential transformation of the linear combinations of pretreatment covariates, i.e., $\alpha_i = \exp(\boldsymbol{X}\boldsymbol{\eta}^i)$ for $i = 1, 2, 3, 4$. However, no further progress has been reported on the success of this application. Second, as we have seen from the READ 180 dataset, in those observational studies where compliance rate differs between the treatment and control group, dropping out rate might probably also be different between the two groups. However, the dropping out variable (withraw variable for the READ 180 dataset) is also a posttreatment variable like compliance rate, which cannot be simply adjusted using standard statistical method like regression while ignoring this factor might cause estimation bias. To adjust for its effect, we can treat it like the compliance rate, i.e., every observation in the sample has two potential probabilities to drop out of the study under treatment and control group, which can be modeled via a logistic regression model. Conditioning on the dropping out status and compliance rate, the outcome variable can thus be modeled using a regression adjusting for pretreatment variables. In another word, we can add another level of models for the dropping out variable into our current Bayesian hierarchical models.

Chapter 4

Quantifying The Sensitivity of The Bayes Factor on The Choice of Prior Distribution in High-Energy Astrophysical Analysis

4.1 Introduction

Distinguishing a faint spectral line or a new source from a chance fluctuation in data with low photon counts is a challenging statistical task in high-energy astrophysical analysis. It is common practice to characterize the problem in statistical terms as a test for the presence of a component in a finite-mixture distribution and address it by computing a likelihood ratio test and calibrating it according to its nominal asympototic distributions (Murakami *et al.*, 1988; Fenimore *et al.*, 1988; Yoshida *et al.*, 1992; Palmer *et al.*, 1994; Freeman *et al.*, 1999; Piro *et al.*, 1999; Fu *et al.*, 2008; Lee *et al.*, 2012). Unfortunately,

as shown in Protassov *et al.* (2002), the standard regularity conditions required for the aysmptotic theory do not always apply to goodness-of-fit tests of this nature, even with a large sample size. As a result, the likelihood ratio statistic does not follow its *known* nominal distribution even asymptotically and the p-values computed in the common routine are misleading and uninterpretable. Protassov *et al.* (2002) provide a solution for the problem by bypassing the asymptotic theory, finding its posterior predictive distribution empirically, and then calculating the corresponding p-value, called the posterior predictive p-value (hereafter, ppp-value). Their proposal is versatile and can be applied to any spectral line detection problem as long as we are able to draw a sample of the unknown parameters from their joint posterior distribution. However, the summary statistic for the decision making criteria, the ppp-value, shares a similar definition and interpretation of the classical p-value. As pointed by Berger and Delampady (1987) and Berger and Sellke (1987), the p-value tends to overstate the evidence of the more complicated (alternative) model when used for testing a precise hypotheses, which is exactly the case for the spectral line detection problem where we want to compare models with and without the spectral line. This problem is fundamental and arises from the definition of the p-value itself. Despite this, there is a lack of model selection metrics other than the p-value for spectral line detection in the astrophysics. Our goal is to introduce another Bayesian method for model selction in this setting, namely the Bayes Factor.

Bayes Factors, e.g., Kass and Raftery (1995), are a powerful summary statistic for model selection which can be applied to a broad range of scientific problems, including comparing non-nested models. However, they are criticized for being "subjective" in that two analyses based on the same data set but with different prior distribution can lead to different conclusions. One of our goals is to carefully study the influence of the choice of prior distribution on the Bayes factor for each of the model parameters in a simple yet popular class of spectral line detection problems via both simulation and real data analysis. We find that although the Bayes factor does depends on the prior assignment, the prior influence can be interpreted in a

non-subjective manner since different priors correspond to different scientific questions such as the particular range of energies where an astronomer might look for the spectral line and how strong of a spectral line the astronomer is looking for. In fact, both p-values and ppp-values are subjective to similar "subjective influence", which is known as the *look elsewhere effect* in astronomy and physics (Gross and Vitells, 2010; Ranucci, 2012). When computing p-values, this "subjective influence" is not typically the effect of a prior distribution or even viewed as subjective. On the other hand, as pointed out by Berger and Delampady (1987) and Berger and Sellke (1987), we find the Bayes Factor is more conservative for detecting the spectral line when compared to p-values and ppp-values.

The rest of the Chapter is organized as follows. Section 4.2 introduces the spectral line detection problem from a statistical point of view. We explain why the usual p-value based on the standard asympototic distribution of the likelihood ratio test is not appropriate for this type of problem. We also detail how the ppp-value and Bayes Factor can be used instead. We carefully analyze the disadvantages of both the ppp-value and Bayes Factor being as a model selection criteria in the context of spectral line detection. In Section 4.3 we review several common methods for computing the Bayes Factor and compare their relative effectiveness and efficiency for the spectral line problem. We then introduce our methodology to study the influence of the choice of prior distributions on the Bayes factor as well as its comparison to the look elsewhere effect of the ppp-values in Section 4.4. In Section 4.5, we perform two Simulation studies and a real data analyses where one of the simulation studies focus on the statistical properties ignoring all instrumental errors and the other mimics real data sets. Conclusion are stated in Section 4.6.

4.2 Model Selection Techniques

4.2.1 Statistical Setup

A model for an energy spectrum can be seperated into two basic set of model components: a set of continuum and a number of emission lines. The continuum decides the general shape of the spectrum. It describes the distribution over the entire energy range of interest. Emission lines, on the other hand, are local positive aberrations from the continuum. We follow Park *et al.* (2008) and assume a standard spectral model that includes a single continuum term along with several spectral lines, with $\boldsymbol{\theta}^C$ and $\boldsymbol{\theta}^L$ representing the parameters for the continuum and emission line respectively. Because X-ray emission is measured by counting photons in each of a number of energy bins, we use the Poisson distribution to independently model the observed photon numbers, represented by $\boldsymbol{D} = (D_1, \ldots, D_J)$ where J represents the number of energy bins. Theoretically, the expected photon counts in a particular energy bin, denoted $\Lambda_j(\boldsymbol{\theta})$, can thus be written as

$$\Lambda_j(\boldsymbol{\theta}) = a(\boldsymbol{\theta}^A, E_j) \cdot \left\{ \Delta_j f(\boldsymbol{\theta}^C, E_j) + \sum_{k=1}^{K} \lambda_k \pi_j(\mu_k, \nu_k) \right\},\ j = 1, \ldots, J \tag{4.1}$$

where $a(\boldsymbol{\theta}^A, E_j)$ is an absorption model with parameter $\boldsymbol{\theta}^{\boldsymbol{A}}$; Δ_j and E_j are the width and mean energy of bin j, $f(\boldsymbol{\theta}^C, E_j)$ is the expected photon counts per unit energy due to the continuum term at energy E_j, K is the number of emission lines, λ_k is the expected photon counts due to the emission line k, and $\pi_j(\mu, \nu)$ is the proportion of an emission line centered at energy μ and with width ν that falls into bin j. Although the model in equation (4.1) is of primary scientific interest, a more complex statistical model is needed to address the errors involved in the data collection processes. There are three main sources of such instrumental errors. First, a photon that arrives with energy corresponding to energy bin j might be mistakenly recorded in energy bin i. The probabilistics of such errors are tabulated in the

redistribution matrix. Secondly, the effective geometric area associated with the energy bin might be much less than its polished surface area due to reflectivity, vignetting, and other effects. We define this effect as "effective area" of the energy bin. Lastly, observation for any source of interest is subject to background contamination. Mathematically, we modify equation (4.1) via

$$\begin{aligned} &\Xi_l(\boldsymbol{\theta}) = \sum_{j=1}^{J} M_{lj}\Lambda_j(\boldsymbol{\theta})d_j + \boldsymbol{\theta}_l^B,\ l = 1, \ldots, L \\ &Y_l \overset{\text{ind}}{\sim} \text{Pois}(\Xi_l(\boldsymbol{\theta})) \end{aligned} \tag{4.2}$$

where $\Xi_l(\boldsymbol{\theta})$ is the expected observed Poisson counts in energy bin l; M is the redistribution matrix with M_{ij} representing the probability of a photon that should have arrived in energy bin i but mistakenly recorded by energy bin j; d_j is the effective area of bin j; $\boldsymbol{\theta}_l^B$ is the expected photon counts in energy bin l originating from the background other than the source of interest.

For ease of discussion, we suppose that all bins are all of equal energy width and a Photo-electric absorption (Hall, 1936). Likewise, we suppose the continuum term in (4.1) is parameterized as a powerlaw model, i.e., $f(\boldsymbol{\theta}^C, E_j) = \alpha E_j^{-\beta}$ where α and β represent the normalization and photon index, respectively; there is only one emission line; the emission line in (4.1) is modeled as a delta function (so that the width of the delta function is effectively the width of the energy bin in which it resides). Under these assumptions, equation (4.1) simplies to

$$\Lambda_j(\boldsymbol{\theta}) = e^{-\text{nH}\cdot\sigma(E_j)} \cdot (\alpha E_j^{-\beta} + \lambda \cdot \delta_j(\mu)) \tag{4.3}$$

where the first factor represents the absorption model with $\boldsymbol{\theta}^A = \text{nH}$ and $\sigma(E_j)$ are the known photo-electric cross-section; $\boldsymbol{\theta} = (\boldsymbol{\theta}^C, \boldsymbol{\theta}^L)$, $\boldsymbol{\theta}^C = (\alpha, \beta)$, and $\boldsymbol{\theta}^L = (\lambda, \mu)$. $\delta_j(\mu)$ is

defined as

$$\delta_j(\mu) = \begin{cases} 1 & \mu = j \\ 0 & \mu \neq j \end{cases}$$

As discussed in Kashyap *et al.* (2010), from a statistical point of view, the emission line detection problem can be regarded as a model selection problem between the two candidate models:

$$\begin{aligned} \text{Continuum Model} &: \lambda = 0 \\ \text{Continuum} + \text{Spectral Line Model} &: \lambda > 0 \end{aligned} \tag{4.4}$$

We introduce a running example based on (4.2), (4.3) and (4.4) which we use extensively to illustrate ideas in Section 4.2.2 - 4.2.6.

Example 1: Suppose we have no instrumental errors with $L = 1000$ equally spaced energy bins between 0.3 to 7 keV. To generate a spectrum, we follow (4.3), set nH $= 0$, $\alpha = 50$, $\beta = 1.69$, $\mu = 1.3$ keV, $\lambda = 20$, and compute the expected photon counts in each energy bin $\Lambda_l(\boldsymbol{\theta})$. We then independently sample a simulated spectrum from the Poisson distribution with intensity equal to the expected photon counts for each energy bin, i.e.,

$$Y_l \overset{\text{ind}}{\sim} \text{Pois}(\Lambda_l(\boldsymbol{\theta}))$$

One simulated spectrum is shown in Figure 4.1. Our general goal is to quantify evidence for the comparison of the models with and without the spectral line.

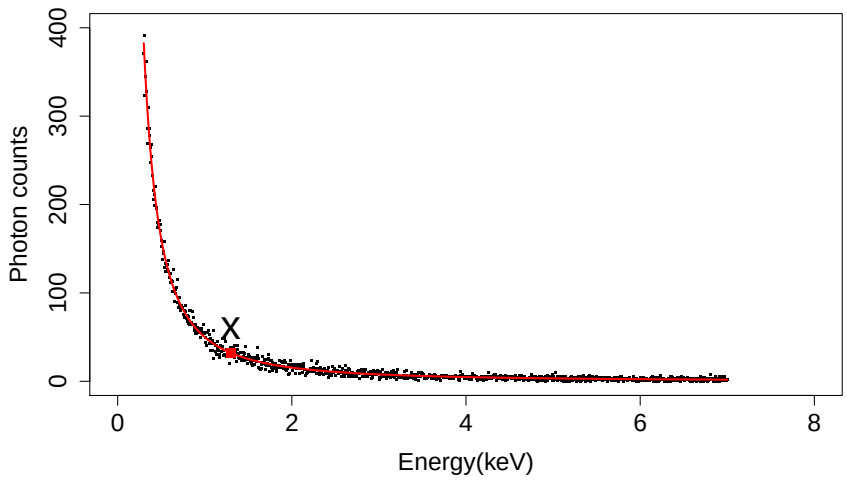

Figure 4.1: A Simulated Spectrum for the Running Example. The red line plots the theoretical functional relationship of the power law model. The red dot shows the energy level where the delta function emission line is placed while the black cross represents the simulated photon counts at this particular energy level. All other photon counts are plotted as black dots.

4.2.2 Hypothesis-testing Using P-values

Between the two candidate models in (4.4), the Continuum model is simpler. According to Occam's Razor Parsimony (Ariew, 1978), if two models predict nature equally well, we prefer the simpler model. It is then natural to give the simple model priority so that the more complicated Continuum + Spectral Line model should not be adopted unless it is by some measure better than the simple model (e.g., more able to predict the observed data). Classically, statisticians formalize this reasoning through hypothesis testing. The simpler or more parsimonious theory is called the *null hypothesis*, H_0, while the more complicated theory is called the *alternative hypothesis*, H_1. Thus, in terms of Neyman-Pearson Framework statistical hypothesis testing, (4.4) can be rewritten as

$$\begin{aligned} &H_0 : \lambda = 0 \\ &H_1 : \lambda > 0. \end{aligned} \tag{4.5}$$

In term of the emission line detection problem, H_0 and H_1 represent the Continuum model and the Continuum+Spectral Line model respectively, and we use $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ to represent the relevant parameters under each model. In particular, (4.5) is a precise hypothesis (Berger and Delampady, 1987) as H_0 involves testing a point value of the parameter. The degree of evidence is typically quantified through the specification of a *test statistic* that is a function of the observed data, denoted $T(\boldsymbol{D})$, with a distribution that is known at least approximately under the assumption that H_0 is correct. The distribution of $T(\boldsymbol{D})$ under H_0 is called the *null distribution*. We then compute the the test statistic for our data, denoted $T(\boldsymbol{D}^{\mathrm{obs}})$, and compare the result to the null distribution. One common way to do this is by computing the probability of obtaining a test statistic at least as extreme as the one that was actually observed. This probability is called *p-value* and we reject the null hypothesis if the p-value is less than a pre-defined threshold called the *significance level*. Mathematically,

$$\text{p-value}(\boldsymbol{\theta}_0) = \begin{cases} P(T(\boldsymbol{D}) \geq T(\boldsymbol{D}^{\mathrm{obs}}) \mid \boldsymbol{\theta}_0, \mathrm{H}_0), & \text{if large values of } T(\boldsymbol{D}) \text{ support } \mathrm{H}_1 \\ P(T(\boldsymbol{D}) \leq T(\boldsymbol{D}^{\mathrm{obs}}) \mid \boldsymbol{\theta}_0, \mathrm{H}_0), & \text{if small values of } T(\boldsymbol{D}) \text{ support } \mathrm{H}_1 \end{cases} \tag{4.6}$$

The implied logic behind a small p-value is that either H_0 is true and a rare event happened, or H_0 is false.

Example 1(a): To illustrate this in the context of the running example, we assume nH, α, β, and μ are known and simulate a spectrum with $\text{nH} = 0, \alpha = 10, \beta = 1.69$, $\lambda = 10$, and $\mu = 1$ keV (the 105^{th} bin). Suppose we use the observed counts in the bin where the line resides, $\boldsymbol{D}_{105}$, as the test statistic; The observed counts in this bin is equal to 14. To compute the p-value, we firstly derive the null distribution assuming the null hypothesis is correct. Since there are no unknown parameters under H_0, $\boldsymbol{D}_{105} \sim \text{Pois}(\alpha E^{-\beta})$ with $\alpha = 10, \beta = 1.69$, and $E = 1$, i.e., the null distribution of $\boldsymbol{D}_{105}$ is Pois(10). Assume the test statistic for our observed data, $T(\boldsymbol{D}^{\mathrm{obs}})$, is equal to 14. Under H_0, obtaining a test statistic "at least as extreme as the one that was actually observed" is equivalent to observing a test statistic which is larger

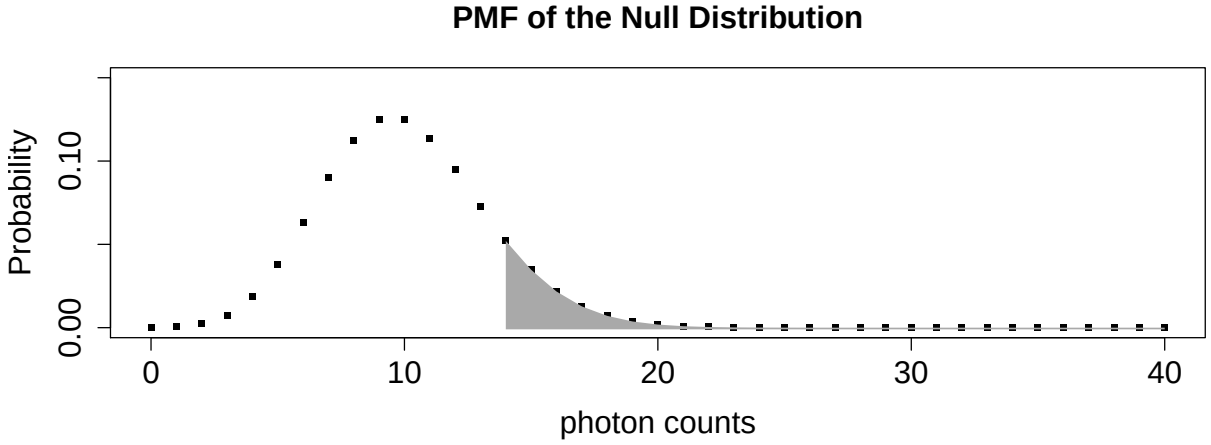

Figure 4.2: The Probability Mass Function (PMF) of Null Distribution for The Test Statistic $T(\boldsymbol{D})$ in The Running Example. The grey area, 13.55%, corresponds to the actual p-value in this problem.

than 14, because H_1 corresponds to a theory that expects more photon counts in this bin than H_0. According to (4.6), the p-value is then $P(T(\boldsymbol{D}) \geq 14)$, where $T(\boldsymbol{D}) \sim \text{Pois}(10)$. In Figure 4.2, it is represented by the grey area under the probability mass function of the null distribution of $T(\boldsymbol{D})$.

By definition, p-values require that we derive the null distribution of $T(\boldsymbol{D})$. Otherwise, the observed value of $T(\boldsymbol{D})$ cannot be calibrated. Even in simple problems this can be challenging, for example, if in practice α and β are unknown, we can no longer explicitely derive the null distribution for the test statistic of $\boldsymbol{D}_{105}$ as it depends on $\boldsymbol{\theta}_0$. The p-value in this case also depends on $\boldsymbol{\theta}_0$. If possible, we want to find a test statistic whose null distribution does not depend on any unknown parameters. Statistically, an ideal option would be to use a statistic that is *ancillary* under H_0, i.e., a statistic whose distribution does not depend on $\boldsymbol{\theta}_0$. In practice, a simple yet popular ancillary statistic that is widely applicable in many scientific fields is the likelihood ratio test (hereafter LRT) statistic. To formally introduce the LRT, we need firstly to define the likelihood. Assume $D_1, \ldots, D_n$ is an independent sample from a probability density function or probability mass function

$f(D|\boldsymbol{\theta})$, the likelihood function is defined as

$$L(\boldsymbol{\theta}|D_1,\ldots,D_n) = L(\boldsymbol{\theta}|\boldsymbol{D}) = f(\boldsymbol{D}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(D_i|\boldsymbol{\theta}) \tag{4.7}$$

Let Θ denote the parameter space. The likelihood ratio test statistic for testing H_0: $\boldsymbol{\theta} \in \Theta_0$ versus H_1: $\boldsymbol{\theta} \in \Theta_0^c$, denoted $T_{\mathrm{LRT}}(\boldsymbol{D})$, is defined as,

$$T_{\mathrm{LRT}}(\boldsymbol{D}) = -2\log(R(\boldsymbol{D})), \text{ where } R(\boldsymbol{D}) = \frac{\sup\limits_{\Theta_0} L(\boldsymbol{\theta}|\boldsymbol{D})}{\sup\limits_{\Theta} L(\boldsymbol{\theta}|\boldsymbol{D})} \ . \tag{4.8}$$

For large samples ($n \to \infty$) under suitable regularity conditions (Wilks, 1938; Chernoff, 1954), the null distribution of the LRT statistic converges in distribution to a χ^2 distribution,

$$T_{\mathrm{LRT}}(\boldsymbol{D}) \xrightarrow{D} \chi^2_d \ , \text{ as } n \to \infty \tag{4.9}$$

where $\xrightarrow{D}$ denotes convergence in distribution and d is the difference between the number of free parameters specified by Θ and Θ_0. For our running example, the numerator of the LRT statistic is likelihood of the observed data being computed under H_0 and the denominator of $T_{\mathrm{LRT}}(\boldsymbol{D})$ is the maximum likelihood under H_0. Because we assume the line location is known, λ is the only unknown parameter in H_1. Hence, the degree of freedom of the putative asympototic χ^2 distribution, d, is equal to 1.

As rule of thumb, there are two important conditions that must be satisfied for the proper calibration of the LRT statistic using the approximated χ^2 distribution. First, *the two models that are being compared must be nested.* This means that null hypothesis must be a special case of the alternative hypothesis with a subset of its parameters being restricted, i.e., $\Theta_0 \subset \Theta$. Second, *the null values of the parameters fixed under* H_0 *may not be on the boundary of the set of possible parameter values, i.e.,* Θ_0 *must not be on the boundary of*

Θ. For our running example 1(a), the first condition implies for example the LRT cannot be used to test a continuum only model against an emission line only model. The second condition is violated when we test (4.5) because λ must be nonnegative while its null value, zero, is exactly on the boundary of its parameter space. As a result, the LRT statistic is not χ^2 distributed even approximately. Using a χ^2 table to compute p-values for this LRT can lead to unpredicatable results (Protassov *et al.*, 2002).

One possible solution when the null distribution of a test statistic cannot be derived analytically is to calibrate it using non-parametric techniques such as the bootstrap (Efron, 1979). It can be formulated as,

Algorithm for computing bootstrap p-value:

`Initialize:` Under H_0, find estimates of the nuisance parameters, $\hat{\boldsymbol{\theta}}_0$, for example, via maximum likelihood.

`Step 1:` For $s = 1 \ldots S$,

1. Simulate $\tilde{\boldsymbol{D}}^{(s)} \sim P(D|\hat{\boldsymbol{\theta}}_0, \mathrm{H}_0)$, i.e., simulate a replicate data set according to the Continuum model using estimated parameters.
2. Compute the test statistic $T(\tilde{\boldsymbol{D}}^{(s)})$ for each of the simulated data set.

`Step 2:` Calibrate the observed test statistic using the empirical null distribution, i.e.,

$$\begin{aligned} \text{bootstrap p-value} &= P(T(\boldsymbol{D}) \geq T(\boldsymbol{D}^{\mathrm{obs}})|\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}_0) \\ &\approx \frac{1}{S}\sum_{s=1}^{S} \mathcal{I}[T(\tilde{\boldsymbol{D}}^{(s)}) > T(\boldsymbol{D}^{\mathrm{obs}})], \end{aligned} \tag{4.10}$$

where $\mathcal{I}$[statement] is an indicator function that is equation to 1 if the statement is true and 0 otherwise.

4.2.3 Posterior Predictive P-values

Instead of fixing the model parameters at their estimated value under H_0, Meng (1994) and Gelman *et al.* (1996) proposed a Bayesian revision of the bootstrap p-value called posterior predictive p-value. As with the bootstrap p-values, we use a Monte Carlo simulation to access the null distribution of a test statistic, but rather than simulating replicate data with parameter values that are fitted to the data under H_0, Monte Carlo simulation is run accounting for posterior uncertainty in the parameters.

To formalize this, we review some basic points of Bayesian statistical inference. Bayesian statistics involves exploring the so-called posterior distribution of the parameters conditioned on data $\boldsymbol{D}$ and model H_k, $p(\boldsymbol{\theta}_k|\boldsymbol{D}, H_k)$. One can derive the posterior distribution using Bayes theorem,

$$P(\boldsymbol{\theta}_k|\boldsymbol{D}, H_k) = \frac{L(\boldsymbol{\theta}_k|\boldsymbol{D}, H_k) \cdot P(\boldsymbol{\theta}_k|H_k)}{P(\boldsymbol{D}|H_k)} \text{ for } k = 1, 2 \tag{4.11}$$

where $L(\boldsymbol{\theta}_k|\boldsymbol{D}, H_k)$ is the likelihood function; $P(\boldsymbol{\theta}_k|H_k)$ is the prior distribution for the parameters and represents information about the parameters known prior to observing the data $\boldsymbol{D}$; $P(\boldsymbol{D}|H_k)$ is the marginal distribution of data, which can be derived via,

$$P(\boldsymbol{D}|H_k) = \int L(\boldsymbol{\theta}_k|\boldsymbol{D}, H_k) P(\boldsymbol{\theta}_k|H_k) d\boldsymbol{\theta}_k. \tag{4.12}$$

Since its value clearly does not depend on $\boldsymbol{\theta}_k$, $P(\boldsymbol{D}|H_k)$ is a normalizing constant for the posterior distribution, $P(\boldsymbol{\theta}_k|\boldsymbol{D}, H_k)$. Using (4.11) allows us to combine information from expert knowledge or previous studies through the prior distribution with information contained in the current data through the likelihood. To illustrate the procedure of computing the ppp-value for an arbitrary test statistic $T(\boldsymbol{D})$, this involves:

Algorithm for computing ppp-values:

`Step 1:` For $s = 1 \ldots S$,

1. Simulate parameter values $\boldsymbol{\theta}_0^{(s)}$ from $P(\theta_0|\boldsymbol{D}, \mathrm{H}_0)$.
2. Simulate $\tilde{\boldsymbol{D}}^{(s)} \sim P(D|\boldsymbol{\theta}_0^{(s)}, \mathrm{H}_0)$, i.e., simulate a replicate of data set under the Continuum Model for each simulated value of the parameters obtained in step 1.
3. Compute the test statistic $T(\tilde{\boldsymbol{D}}^{(s)})$.

`Step 2:` Compute the posterior predictive p-value,

$$\begin{aligned} \text{ppp-value} &= P\big(T(\boldsymbol{D}) \geq T(\boldsymbol{D}^{\text{obs}})|\boldsymbol{D}\big) = \mathrm{E}\Big[P\big(T(\boldsymbol{D}) \geq T(\boldsymbol{D}^{\text{obs}})|\boldsymbol{D}, \boldsymbol{\theta}_0\big)\Big|\boldsymbol{D}\Big] \\ &= \mathrm{E}\big[\text{p-value}(\boldsymbol{\theta}_0)\big|\boldsymbol{D}\big] \\ &\approx \frac{1}{S}\sum_{s=1}^{S} \mathcal{I}[T(\tilde{\boldsymbol{D}}^{(s)}) > T(\boldsymbol{D}^{\text{obs}})]. \end{aligned} \tag{4.13}$$

The ppp-value is the proportion of the Monte Carlo simulations that results in a value of $T(\tilde{\boldsymbol{D}}^{(s)})$ more extreme than the test statistic obtained for the observed data, $T(\boldsymbol{D}^{\text{obs}})$. It is treated as a p-value, with small values indicating evidence for the more complex model, but addresses the problem associated with the bootstrap p-values of pretending to known the nuisance parameter values $\boldsymbol{\theta}_0$ by fixing them at their sample estimates $\hat{\boldsymbol{\theta}}_0$. It can be used in virtually any scenario as long as we are able to sample the posterior distribution.

However, ppp-values have been criticized for being conservative relative to p-values (Sinharay and Stern, 2003; Bayarri and Castellanos, 2007; Dey *et al.*, 1998; Robins *et al.*, 2000b). Intuitively, since the posterior distribution combines both prior information and information contained in the data, sampling parameters from their posterior distribution inevitably picks those parameters values that are more likely under H_0. The data sets generated with these parameter values are therefore more consistent with H_0, making it difficult to criticize H_0. In practice, small values of ppp-values denote surprise or incompatibilty between the null

hypothesis and the data. This is based on intuitive stemming from the uniform distribution of p-values under H_0. Robins *et al.* (2000b), however, showed that ppp-values are not uniformly distributed even asymptotically under the null hypothesis. Rather, their asymptotic distributions tend to be more concentrated around 0.5. This problem is mitigated if the test statistic is (asympototically) ancillary, which fortunately is the case for the LRT.

4.2.4 A Principled Bayesian Method for Model Selection

A formal Bayesian analysis allows us to compare two hypotheses in a more direct manner. Let $P(H_0)$ and $P(H_1) = 1 - P(H_0)$ be the prior probabilities for the two hypotheses, which represent the researchers' initial knowledge of the relative likelihood of the two hypotheses before the data is observed. For the moment, we leave aside the question of how these quantities are determined. Using Bayes' theorem, we can then derive the posterior probabilities for the two hypotheses $P(H_0|\boldsymbol{D})$ and $P(H_1|\boldsymbol{D}) = 1 - P(H_0|\boldsymbol{D})$ via,

$$P(H_k|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|H_k)P(H_k)}{P(\boldsymbol{D}|H_0)P(H_0) + P(\boldsymbol{D}|H_1)P(H_1)}, \text{ for } k = 0, 1. \tag{4.14}$$

This probability only depends on the marginal distribution of the data under each model, $P(\boldsymbol{D}|H_k)$, and the prior probabilities of the models, $P(H_k)$, and does not involve any model parameters. The odds ratio, $P(H_0|\boldsymbol{D})/P(H_1|\boldsymbol{D})$, can be formatted as,

$$\frac{P(H_0|\boldsymbol{D})}{P(H_1|\boldsymbol{D})} = \frac{P(\boldsymbol{D}|H_0)}{P(\boldsymbol{D}|H_1)} \frac{P(H_0)}{P(H_1)}.$$

Similar to the posterior distribution of parameters, the posterior odds is also a combination of information contained in the data and in the prior distribution, represented by $P(\boldsymbol{D}|H_0)/P(\boldsymbol{D}|H_1)$ and $P(H_0)/P(H_1)$ respectively. The ratio of the posterior odds and

Table 4.1: The Interpretation of The Bayes factor

$\log_{10}(B_{01})$	B_{01}	Evidence against H_1
0 to 0.5	1 to 3.2	Not worth more than a bare mention
0.5 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

the prior odds of the two models,

$$B_{01} = \frac{P(\mathrm{H}_0|\boldsymbol{D})}{P(\mathrm{H}_1|\boldsymbol{D})} \Big/ \frac{P(\mathrm{H}_0)}{P(\mathrm{H}_1)} = \frac{P(\boldsymbol{D}|\mathrm{H}_0)}{P(\boldsymbol{D}|\mathrm{H}_1)}, \tag{4.15}$$

is called a Bayes factor (in favor of the null hypothesis) and does not depend on the prior probability of the two models. As we shall see, however, the Bayes Factor does depend on $P(\boldsymbol{\theta}_0|\mathrm{H}_0)$ and $P(\boldsymbol{\theta}_1|\mathrm{H}_1)$, the prior distributions of the model parameters under each hypothesis. In the simplest case when both hypotheses are simple distributions with no free parameters, for example, testing $\lambda = 0$ versus $\lambda = 1$ with known (α, β, μ) in the running example, $B_{01} = P(\boldsymbol{D}|\lambda = 0)/P(\boldsymbol{D}|\lambda = 1)$ is simply the likelihood ratio. In the general cases, when there are unknown parameters under either or both of the hypotheses, the Bayes factor still has a form similar to that of a likelihood ratio, but with the likelihood $L(\boldsymbol{D}|\hat{\boldsymbol{\theta}}_k, \mathrm{H}_k)$ replaced by the marginal densities of the data $P(\boldsymbol{D}|\mathrm{H}_k)$ for $k = 1, 2$. In words, the LRT involves finding the MLE under both models, computing their ratio, and deriving the p-value by referring to the (perhaps known asympototic) null distribution of the test. Bayes factors, on the other hand, average the likelihood over the prior distribution of relevant parameters and then compute the ratio of these marginal data densities. Interestingly the computation of ppp-values could also be thought as taking average. But instead of averaging the likelihood like the Bayes factor, they average the p-values under different nuisance parameter values over their posterior distribution, as can be seen from (4.13).To interpret the Bayes factor, (Jeffreys, 1961) suggested using half units on the $\log_{10}$ scale. A simplified version of his scale obtained by pooling two of his categories appears in Table 4.1.

4.2.5 The Fallible P-value

There is an extensive literature discussing whether a p-value provides adequate "evidence" for the comparison of two models. A fundamental problem is that compared to the Bayes factor (or posterior odds), p-value (including the posterior predictive p-values) is simply much harder to interpret. It is critical to remember that the p-value is not $p(\mathrm{H}_0|\boldsymbol{D})$ as it is often interpreted. If we only knew that $\boldsymbol{D}$ were as extreme or more extreme than $\boldsymbol{D}^{\mathrm{obs}}$, the p-value would be similar to the posterior probability of the null hypothesis. Formally,

$$E = \{\text{any data } \boldsymbol{D} \text{ such that } |T(\boldsymbol{D})| \geq |T(\boldsymbol{D}^{\mathrm{obs}})| \ \}, \tag{4.16}$$

Berger and Sellke (1987) quantitively show that the p-value is frequently very close to $p(\mathrm{H}_0|E)$. In terms of the definition for the p-value, the event E represents the "tail event" whose probability is being calculated under H_0. It is obvious that there may be a vast difference between being told that $\boldsymbol{D} = \boldsymbol{D}^{\mathrm{obs}}$ and $\boldsymbol{D} \in E$. The latter provides substantially stronger evidence against H_0 in that $p(\mathrm{H}_0|E)$ is typically much smaller than $p(\mathrm{H}_0|\boldsymbol{D}^{\mathrm{obs}})$. To illustrate this, we use the running example 1(a) of Section 4.2.2. Recall we assume nH, α, β, and μ are all fixed; The only unknown parameter is λ. We simulate a spectrum with $\mathrm{nH} = 0, \alpha = 10, \beta = 1.69$, $\lambda = 10$, and $\mu = 1$ keV (the 105^{th} bin) and use the observed counts in the bin where the line resides, $\boldsymbol{D}_{105}$, as the test statistic; The observed counts in this bin is equal to 14. For this example, $E = \{\boldsymbol{D}_{105} \geq 14\}$. Hence,

$$p(\mathrm{H}_0|E) = p(\mathrm{H}_0) \cdot \frac{p(E|\mathrm{H}_0)}{p(E)} = p(\mathrm{H}_0) \cdot \frac{\sum_{k=14}^{\infty} p(\boldsymbol{D}_{105} = k|\mathrm{H}_0)}{\sum_{k=14}^{\infty} p(\boldsymbol{D}_{105} = k)}.$$

Note that

$$\frac{p(\boldsymbol{D}_{105} = k|\mathrm{H}_0)}{p(\boldsymbol{D}_{105} = k)} < \frac{p(\boldsymbol{D}_{105} = 14|\mathrm{H}_0)}{p(\boldsymbol{D}_{105} = 14)}, \text{ for each } k \geq 15.$$

Because larger counts in the 105^{th} energy bin is less and less possible under the null hypothesis. Therefore

$$p(\mathrm{H}_0|E) = p(\mathrm{H}_0) \cdot \frac{\sum_{k=14}^{\infty} p(\boldsymbol{D}_{105} = k|\mathrm{H}_0)}{\sum_{k=14}^{\infty} p(\boldsymbol{D}_{105} = k)} < p(\mathrm{H}_0) \cdot \frac{p(\boldsymbol{D}_{105} = 14|\mathrm{H}_0)}{p(\boldsymbol{D}_{105} = 14)} = p(\mathrm{H}_0|\boldsymbol{D}^{\mathrm{obs}}).$$

In addition, as shown by Berger and Delampady (1987), there are dramatic conflicts between the classical p-value (or observed significance level) and the Bayes factor (or posterior odds) when testing precise hypotheses, which is the case for the emission line detection problem. In particular, "p-values are typically at least an order of magnitude smaller than Bayes factor or posterior probability for H_0". For example, a p-value of 0.05, which is usually considered as significant evidence against H_0, may reflect the actual posterior probability of H_0 near 1/2 and Bayes factor near 1. (Although one might argue such difference might have something to do with the choice of prior distribution for Bayes factors and posterior probabilities, (Berger and Delampady, 1987) show that the smallest posterior probability under a series of prior distributions of a point null hypothesis is still much larger than the corresponding p-value.) To illustrate the discrepancy between the p-value and the Bayes factor for testing precise hyothesis, we present a simple example inspired by the Lindley-Jeffrey paradox (Lindley, 1957), where the posterior probabilities could approach 1 while the p-value stays tiny.

Example 1(b) Suppose now nH, α, β, μ are all fixed. The only unknown parameter is λ. The alternative hypothesis takes the form of $\mathrm{H}_1 : \lambda = \lambda_0$. Suppose nH = 0, $\alpha = 500$, $\beta = 0.1$, and $\mu = 1$. We still use the observed counts in the bin where the line resides, $\boldsymbol{D}_{105}$, as the test statistic. The null distribution in this case is then Poisson(500). Assume that based on the observed data $\boldsymbol{D}^{\mathrm{obs}}$, we find $0.04 <$ p-value < 0.05. One might naively expect that the chances that H_0 holds conditioned on this p-value, $P(\mathrm{H}_0 \mid 0.04 \leq \text{p-value} \leq 0.05)$,

is close to the interval of $(0.04, 0.05)$. However, using Bayes theorem,

$$P_{\lambda_0}(\mathrm{H}_0 \mid 0.04 \leq \text{p-value} \leq 0.05) = P_{\lambda_0}(\mathrm{H}_0 \mid 537 \leq \boldsymbol{D}^{\text{obs}} \leq 539) = \frac{P(537 \leq \boldsymbol{D}^{\text{obs}} \leq 539 \mid \mathrm{H}_0) \cdot P(\mathrm{H}_0)}{P(537 \leq \boldsymbol{D}^{\text{obs}} \leq 539 \mid \mathrm{H}_0) \cdot P(\mathrm{H}_0) + P(537 \leq \boldsymbol{D}^{\text{obs}} \leq 539 \mid \mathrm{H}_1) \cdot (1 - P(\mathrm{H}_0))} \tag{4.17}$$

This probability can depend heavily on the alternative hypotheses, in particular, the value of λ_0. To illustrate this, we plot $P(537 \leq \boldsymbol{D}^{\text{obs}} \leq 539 \mid \mathrm{H}_0)$ and $P(537 \leq \boldsymbol{D}^{\text{obs}} \leq 539 \mid \mathrm{H}_1)$ in the left panel of Figure 4.3 when $\lambda_0 = 30$, where the black shaded area corresponds the former probability and the black and grey shaded area correspond to the latter probability. As the center of the right PMF curve is equal to $500 + \lambda_0$, the black and grey shaded area will depend on the value of λ_0. If λ_0 is extremely large, it can be very tiny. On the other hand, since the left PMF curve stays the same regardless of the value of λ_0. The black area keeps unchanged. As an example, (4.17) is around $p(\mathrm{H}_0)$ when $\lambda_0 = 77.9$ while it is almost 1 for $\lambda_0 = 200$. The complete functional relationship between (4.17) and λ_0 is shown in the right panel of Figure 4.3 assuming $P(\mathrm{H}_0) = P(\mathrm{H}_1) = 0.5$.

4.2.6 The Fallible Bayes Factor

As we can see from (4.12), the Bayes factor depends on the choice of prior distribution. Sometimes, different prior distributions can lead to significantly different conclusions. To illustrate this, we return to our running example.

Example 1(c): Suppose $\mathrm{nH} = 0$, $\alpha = 500$, $\beta = 0.1$, and $\lambda = 50$ are all known, but μ is unknown. We simulate a spectrum assuming there is an emission line located at $\mu = 1$ keV, i.e., energy bin 105. Suppose the observed photon counts in this particular energy bin, $\boldsymbol{D}_{105}$, is equal to 550. We use a uniform prior distribution for μ centered at the true emission line location, i.e., $p(\mu) \sim \mathrm{U}(1 - \eta/2, 1 + \eta/2)$ where η controls the width of the uniform

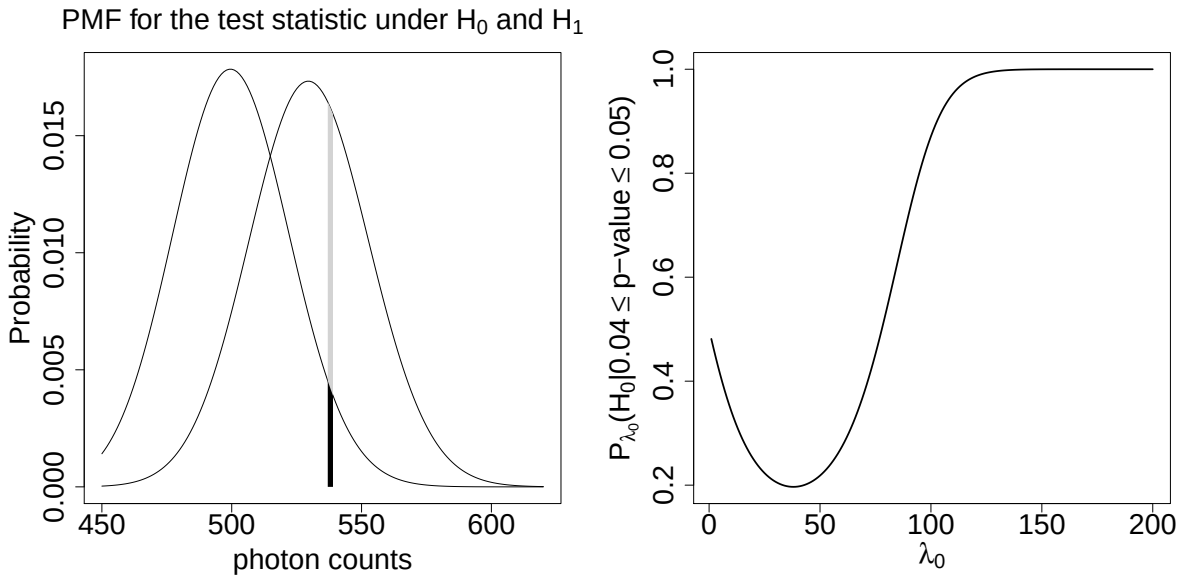

Figure 4.3: The left panel plots the PMF function of Pois(500) and Pois(530), i.e., the PMFs for the test statistic under H_0 and H_1. The black shaded area represent the probability of $P(537 \leq S_0 \leq 539 \mid H_0)$ while the black and grey area combined corresponds to $P(537 \leq S_0 \leq 539 \mid H_1)$. The right panel plots $P_{\lambda_0}(H_0 \mid 0.04 \leq \text{p-value} \leq 0.05)$ as a function of λ_0 assuming $P(H_0) = P(H_1) = 0.5$.

distribution. We then plot the Bayes factor as a function of η in Figure 4.4. The Bayes factor shows substantial to strong evidence for the emission line when η is small, but the evidence diminishes as η increases. Eventually, the Bayes factor fails to distinguish the two models. The choice of prior distribution is often viewed as reflecting subjectiveness. Thus, the very different conclusions due purely to the choice of prior distribution is why the Bayes factor is often viewed as "subjective". It is worth noting, however, that different values of η could represent different sensible scientific questions under different circumstances. To see this, think about two different cases where we use the likelihood ratio test for line detection,

1. Test for a emission line at a given location.

2. Test for a emission line with an unknown location.

The second type of test can be viewd as multiple tests of the first type and thus results in a different p-value. (This is known as the look elsewhere effect in astronomy and physics.)

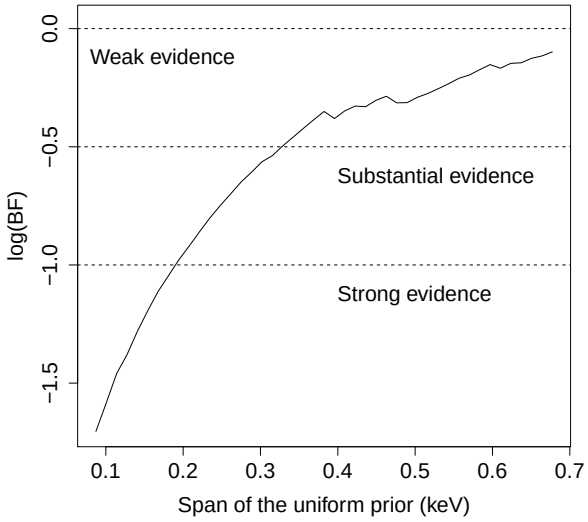

Figure 4.4: Logarithm of the Bayes Factor as A Function of the Emission Line Location Prior for Example 1(b).

In the same way different prior distributions on μ may correspond to situations in which researchers have different objectives about where they wish to look for an emission line.

Unfortunately, noninformative (e.g., "objective") priors (Berger, 1985) do not exist for testing the precise hypotheses, including the line detection problem (Berger and Delampady, 1987; Berger and Sellke, 1987). Furthermore, improper priors should be used with extreme caution when computing the Bayes factors. They can lead to particularly nonsensical answers when assigned to those parameters that are only defined in one of the two (null or alternative) models. We can illustrate this in the running example.

Example 1(d): Assume all parameters are unknown. We use independent uniform distributions over a large region of the parameter space as the prior for $(\boldsymbol{\theta}^C, \boldsymbol{\theta}^L)$, i.e.,

$$p(\boldsymbol{\theta}^C, \boldsymbol{\theta}^L) = \begin{cases} c^C \cdot c^L, & \text{if } \boldsymbol{\theta}^C \in \Theta^C \text{ and } \boldsymbol{\theta}^L \in \Theta^L \\ 0, & \text{otherwise} \end{cases}$$

where Θ^C and Θ^L define the region of possible values of the parameters; c^C and c^L are normalizing constants equal to the areas of Θ^C and Θ^L, respectively. This is appealing as it represents a common case when researchers only have vague prior information for the relevant parameters. Using this prior distribution in (4.12),

$$B_{01} = \frac{P(\boldsymbol{D}|\mathrm{H}_0)}{P(\boldsymbol{D}|\mathrm{H}_1)} = \frac{\int_{\Theta^C} P(\boldsymbol{D}|\boldsymbol{\theta}^C, \mathrm{H}_0) \cdot c^C \; d\boldsymbol{\theta}^C}{\int_{\Theta^C} \int_{\Theta^L} P(\boldsymbol{D}|\boldsymbol{\theta}^C, \boldsymbol{\theta}^L, \mathrm{H}_1) \cdot c^C \cdot c^L \; d\boldsymbol{\theta}^C \; d\boldsymbol{\theta}^L}. \tag{4.18}$$

Since $\boldsymbol{\theta}^C$ is defined in both hypotheses, its normalizing constant c^C appears in both $P(\boldsymbol{D}|\mathrm{H}_0)$ and $P(\boldsymbol{D}|\mathrm{H}_1)$, which appears in the numerator and denominator of (4.18) so that c^C cancels in the Bayes factor. We can use non-informative or reference priors for $\boldsymbol{\theta}^C$, e.g., $p(\boldsymbol{\theta}^C) \propto 1$, which can be regarded as the limit of a Uniform distribution whose range increases to infinity (Pérez and Berger, 2002). To illustrate this, an example based on the running example can be found in Appendix A.1. On the other hand, since $\boldsymbol{\theta}^L$ is only defined in the alternative hypothesis, c^L does not cancels when computing the Bayes factor. Changing the area of Θ^L, i.e., c^L, changes the Bayes factor. Thus, the value of the Bayes factor is completely determined by the choice of Θ^L. In the limit as $c^L \to \infty$, we omit c^L in the expression for the improper prior. The resulting expression for the Bayes factor is uninterpretable. This is not an indictment on the Bayes factor, but does mean that we need to be careful when choosing prior distributions.

4.2.7 Methodological Aim

In this article, we compare the use of both the classical p-value and the ppp-value with the use of the Bayes factor to quantify the evidence for an emission line in a high energy spectrum. We illustrate our comparison via both simulation studies and real data analyses in Section 4.5. We show that there are cases where the Bayes factor can give consistent conclusions under a range of prior distributions and that using the Bayes factor in these

cases is uncontroversial. In other case, although the Bayes factor may be sensitive to the choice of prior distribution, such "subjectiveness" can be interpreted in that different prior distributions represent different scientific questions such as where to look for an emission line; in the full energy range or only in a restricted area. As we shall see, such choice influences not only the Bayes factor but also the p-values via the look-elsewere effect. We give general advice as to how to specify the prior distributions for the emission line detection problem.

4.3 The Computation of Bayes Factor

Computing a Bayes factor requires the evaluation of the integrals in (4.12), for $k = 1, 2$; This typically requires numerical methods. This is a surprisingly difficult computation task. Both marginal data densities of the Bayes factor may involve integrating a non-Gaussian, possibly multimodal probability density function over a high dimensional space where its value may be close to zero over a large area. In this section, we review and compare three common strategies to do this calculation and give a recommended procedure for the emission line detection problem. See Kass and Raftery (1995) for further discussion on other methods.

4.3.1 Laplace's Approximation

As discussed in Kass and Raftery (1995), a useful approximation in order to compute (4.12) is to assume that the posterior density is highly peaked about its maximum $\tilde{\theta}$, the posterior mode. This is usually the case if the likelihood function is highed peaked near its maximum $\hat{\theta}$, which is typically the case for large samples. Mathematically, let $l(\boldsymbol{\theta}_k) = \log\{L(\boldsymbol{\theta}_k|\boldsymbol{D}_k, \mathrm{H}_k)P(\boldsymbol{\theta}_k|\mathrm{H}_k\}$ denote the log-posterior distribution. We can rewrite

(4.12) as

$$
\begin{aligned}
P(\boldsymbol{D}_k|\mathrm{H}_k) &= \int \exp\{l(\boldsymbol{\theta}_k)\} d\boldsymbol{\theta}_k \\
&\approx \int \exp\{l(\tilde{\boldsymbol{\theta}}_k) - \frac{1}{2}(\boldsymbol{\theta}_k - \tilde{\boldsymbol{\theta}}_k)' \mathcal{H}(\tilde{\boldsymbol{\theta}}_k)(\boldsymbol{\theta}_k - \tilde{\boldsymbol{\theta}}_k)\} d\boldsymbol{\theta}_k \qquad (4.19) \\
&= \exp\{l(\tilde{\boldsymbol{\theta}}_k)\}(2\pi)^{d/2} |\mathcal{H}(\tilde{\boldsymbol{\theta}}_k)|^{-1/2} \cdot \\
&\qquad \int (2\pi)^{-d/2} |\mathcal{H}(\tilde{\boldsymbol{\theta}}_k)|^{1/2} \exp\{-\frac{1}{2}(\boldsymbol{\theta}_k - \tilde{\boldsymbol{\theta}}_k)' \mathcal{H}(\tilde{\boldsymbol{\theta}}_k)(\boldsymbol{\theta}_k - \tilde{\boldsymbol{\theta}}_k)\} d\boldsymbol{\theta}_k \\
&= (2\pi)^{d/2} \cdot |\mathcal{H}(\tilde{\boldsymbol{\theta}}_k)|^{-1/2} \cdot L(\tilde{\boldsymbol{\theta}}_k|\boldsymbol{D}, \mathrm{H}_k) \cdot P(\tilde{\boldsymbol{\theta}}_k|\mathrm{H}_k) \qquad (4.20)
\end{aligned}
$$

where $\mathcal{H}$ is the Hession matrix, $\partial^2 l(\boldsymbol{\theta}_k)/\partial\boldsymbol{\theta}_k \cdot \partial\boldsymbol{\theta}_k'$. The approximate equality of (4.19) holds true because we replace $l(\boldsymbol{\theta}_k)$ with its second order Taylor approximation expanded at $\tilde{\boldsymbol{\theta}}_k$. The equality in (4.20) makes use of the fact that integral of a (multivariate Guassian) density is equal to 1.

The Laplace's approximation works well when the the second approximation equality of (4.19) holds, i.e., the log-posterior is quadratic and peaked around its mode. In terms of the emission line detection problem, however, this is usually not the case. As least not if line is weak or moderate in strength, which is typical when a model selection criteria is required to assist with decision making. In this case, the joint posterior distribution for (λ, μ) typically contains several disjoint local modes. To illustrate this, we consider again the running example.

Example 1(e): Suppose now only nH is known; α, β, λ, and μ are all unknown. We simulate a spectrum with nH $= 0$, $\alpha = 50$, $\beta = 1.69$, $\mu = 1.3$ keV, and $\lambda = 10$. We use Uniform priors for α, β, and λ and a discretized Normal prior [1] centered at the true generative value for μ,

[1]Because μ is a discretized random variable, we evaluate its density at all possible 1000 bin locations according to a Normal density of $N(\mu_0, \sigma^2)$ and then re-weight it into a probability mass function. We denote such prior distribution as $DN(\mu_0, \sigma^2)$.

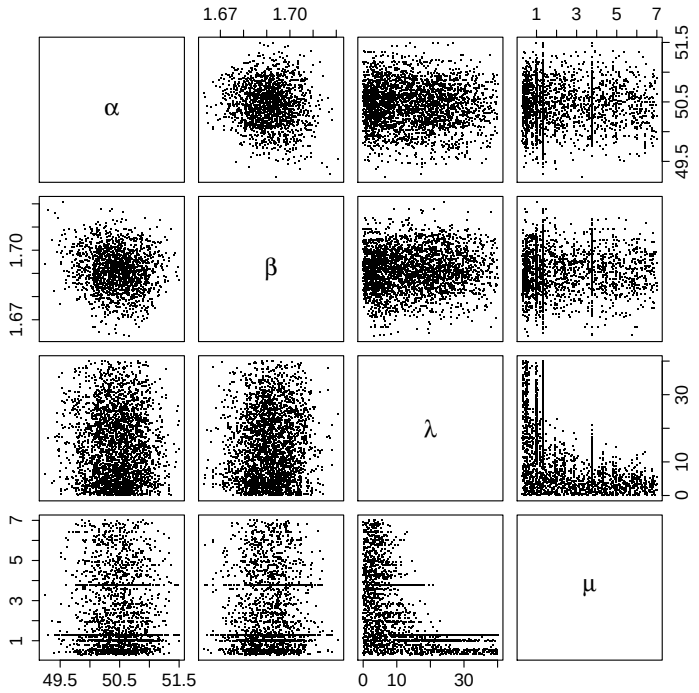

Figure 4.5: The Scatterplot of 3000 Draws from the Joint Posterior Distribution for the Running Example 1(d). The posterior distribution is highly multimodal and non-Gaussian.

i.e.,

$$\alpha \sim U(0, 100),\ \beta \sim U(0, 10),\ \lambda \sim U(0, 40),\ \text{and}\ \mu \sim DN(1.3, 3^2).$$

We then plot the the scatterplot of 3000 draws from the joint posterior distribution for this example in Figure 4.5, which is clearly multimodal.

4.3.2 Monte Carlo Integration

Monte Carlo integration can also be used to compute Bayes factors. The simplest Monte Carlo integration estimate of (4.12) is

$$\hat{P}_1(\boldsymbol{D}|\mathrm{H}_k) = \frac{1}{N}\sum_{i=1}^{N} L(\boldsymbol{\theta}_k^{(i)}|\boldsymbol{D}, \mathrm{H}_k), \tag{4.21}$$

where $(\boldsymbol{\theta}_k^{(1)}, \ldots, \boldsymbol{\theta}_k^{(N)})$ is a random sample from the prior distribution $P(\boldsymbol{\theta}_k|\mathrm{H}_k)$ (Raftery and Banfield, 1990; McCulloch and Rossi, 1991). However, as discussed in Kass and Raftery (1995), a major difficulty with this estimator is that because the prior distribution is generally much more diffuse than the likelihood, most of the $\boldsymbol{\theta}_k^{(i)}$ in a prior sample have small likelihood values. In another word, $\hat{P}_1(\boldsymbol{D}|\mathrm{H}_k)$ tends to be dominated by a few samples which have large likelihood values. As a result, $\hat{P}_1(\boldsymbol{D}|\mathrm{H}_k)$ is inefficient and exhibits large Monte Carlo variance. Consider Example 1(d) except that we now assume nH, α, β, and λ are all known, and only μ is unknown. Under this setting, the numerator of the Bayes factor, $P(\boldsymbol{D}|\mathrm{H}_0)$, becomes a constant. Thus, the Monte Carlo integration estimate of B_{01} is

$$\hat{B}_{01} = \frac{P(\boldsymbol{D}|\mathrm{H}_0)}{\frac{1}{N}\sum_{i=1}^{N} L(\boldsymbol{\theta}_1^{(i)}|\boldsymbol{D}, \mathrm{H}_1)} = \frac{1}{\frac{1}{N}\sum_{i=1}^{N} \frac{L(\boldsymbol{\theta}_1^{(i)}|\boldsymbol{D}, \mathrm{H}_1)}{P(\boldsymbol{D}|\mathrm{H}_0)}}, \tag{4.22}$$

where $\boldsymbol{\theta}_1^{(i)}$ $(i = 1, \ldots, N)$ are the samples from the prior distribution of μ. If we let $N = 10000$, however, more than 50% of the total sum,

$$\sum_{i=1}^{N} L(\boldsymbol{\theta}_1^{(i)}|\boldsymbol{D}, \mathrm{H}_1)/P(\boldsymbol{D}|\mathrm{H}_0),$$

is contributed by the 1% of the $\{\boldsymbol{\theta}_1^{(1)}, \ldots, \boldsymbol{\theta}_1^{(10000)}\}$ that have the highest likelihood. Note that for this example we only have one parameter involved for the integral estimation. The simple Monte Carlo integration estimate can exhibit far worse efficiency as the dimension of the parameter space increases.

The efficiency of the Monte Carlo integration can be improved using importance sampling (Geweke, 1989), adaptive Gaussian quadrature (Genz and Kass, 1993), and bridge (path) sampling (Meng and Wong, 1996; Gelman and Meng, 1998). The first technique requires a good approximation to the likelihood function while the second one assumes the likelihood function is peaked around a dominant mode. Unfortunately, both techniques do not work

well for the emission line detection problem. The likelihood function under the alternative model is usually bumpy and highly multimodal, especially with regard to the parameter of μ. This is evident judging from last row and column of Figure 4.5. The third technique, bridge sampling, is designed to compute the ratio of two normalizing constant, which is exactly what is needed when computing a Bayes factor. However, this method works best for problems in which the common parameter space between the two candidate models, $\Theta_0 \cap \Theta_1$, has non-zero measure, which is not the case for the line detection problem. Note that since the parameter space for λ always has zero measure under H_0 and nonzero measure under H_1, the common parameter space for the line detection problem always has zero measure.

4.3.3 Nested Sampling

A third way to compute the Bayes Factor is via Nested sampling (Skilling, 2006; Feroz and Hobson, 2008). This method is designed primely to find the marginal density of the data,

$$P(\boldsymbol{D}|\mathrm{H}) = \int L(\boldsymbol{\theta}|\boldsymbol{D})P(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{4.23}$$

Unlike Monte Carlo integration which uses a sample from the prior distribution and then averages over the likelihood of the sample, Nested Sampling approximates the numerical integral from a different perspective. We start with the following lemma. For a positive-valued random variable X, assume it has pdf f and cdf F, then

$$\int_0^\infty (1 - F(x))dx = \int_0^\infty xf(x)dx = \mathrm{E}(X). \tag{4.24}$$

To prove (4.24), note that

$$\begin{aligned}
\int_0^\infty (1 - F(x))dx &= \int_0^\infty (1 - P(X < x))dx \\
&= \int_0^\infty P(X > x)dx \\
&= \int_0^\infty \int_x^\infty f(y) \cdot dy \cdot dx \\
&= \int_0^\infty \int_0^y f(y) \cdot dx \cdot dy \qquad (4.25) \\
&= \int_0^\infty f(y) \int_0^y dx \cdot dy \\
&= \int_0^\infty yf(y)dy = \int_0^\infty xf(x)dx = \mathrm{E}(X)
\end{aligned}$$

where equation (4.25) is simply changing the order the double integral. Since the likelihood function $L(\boldsymbol{\theta}|\boldsymbol{D})$ is a non-negative real function defined on Θ, it has its own distribution. In fact, its cumulative distribution function of $\lambda = L(\boldsymbol{\theta}|\boldsymbol{D})$ can be defined by

$$F(\lambda) \equiv \int_{L(\boldsymbol{\theta}|\boldsymbol{D})<\lambda} P(\boldsymbol{\theta})d\boldsymbol{\theta}. \qquad (4.26)$$

On the other hand, (4.23) can be regarded as the expected value of the likelihood function, $\mathrm{E}(L(\boldsymbol{\theta}|\boldsymbol{D}))$. If we define

$$X(\lambda) \equiv 1 - F(\lambda) = \int_{L(\boldsymbol{\theta}|\boldsymbol{D})>\lambda} P(\boldsymbol{\theta})d\boldsymbol{\theta},$$

by (4.24), the desired integral of (4.23) is then equal to $\int_0^\infty X(\lambda)d\lambda$. Note that we may invert the function $X(\lambda)$ and rewrite the integral as

$$\begin{aligned}
P(\boldsymbol{D}|\mathrm{H}) &= \int_0^\infty X(\lambda)d\lambda \qquad (4.27) \\
&= \int_0^1 X^{-1}(p)dp, \qquad (4.28)
\end{aligned}$$

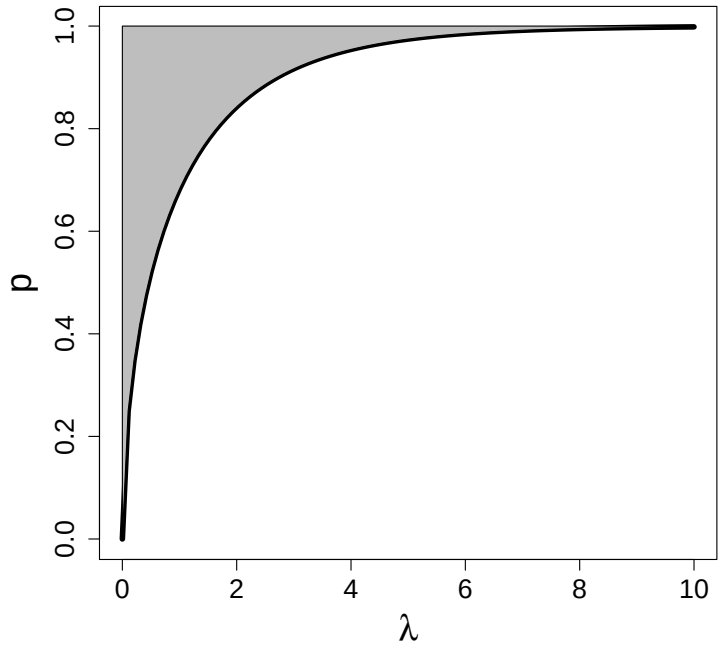

Figure 4.6: Relationshp between $L(\boldsymbol{\theta}|\boldsymbol{D})$, $X(\lambda)$, And the Marginal Density of the Data. The x-axis plots the value of the likelihood function; The solid black line represents its cdf, $1 - X(\lambda)$; The area of the grey region reflects the marginal density of the data.

where $X^{-1}(p) = \lambda$ is that likelihood λ such that $P(L(\boldsymbol{\theta}|\boldsymbol{D}) > \lambda) = p$, e.g., $X^{-1}(0.9)$ is the 90% quantile of the likelihood function. The equivalence of (4.27) and (4.28) can be illustrated via Figure 4.6, where the x-axis represents the likelihood value λ; the solid black line (and also the y-axis) is its CDF defined by (4.26). Our desired integral, $\int_0^\infty X(\lambda)d\lambda$, is thus the area of the grey region. (4.27) solves this integral with regard to λ while (4.28) with regard to p. Rewrite the inverse function of $X(\lambda)$ as $\mathcal{L}(X)$, i.e., $\mathcal{L}(X(\lambda)) = \lambda$, our target integral is then

$$P(\boldsymbol{D}|\mathrm{H}) = \int_0^1 \mathcal{L}(X)dX. \tag{4.29}$$

As an example, the functional relationship between $\mathcal{L}$ and X for a standard Normal distribution is shown in the left panel of Figure 4.7, where the grey area corresponds to the target

integral $P(\boldsymbol{D}|\mathrm{H})$. Given a sequence of X of decreasing values,

$$0 = X_{m+1} < X_m < \ldots < X_2 < X_1 < X_0 = 1.$$

If we can calculate their corresponding $\mathcal{L}_i = \mathcal{L}(X_i)$, the integral in (4.29) is bounded by

$$\sum_{i=1}^{m} (X_i - X_{i+1})\mathcal{L}_i \leq P(\boldsymbol{D}|\mathrm{H}) \leq \sum_{i=1}^{m} (X_{i-1} - X_i)\mathcal{L}_i + X_m \mathcal{L}_{\max}. \tag{4.30}$$

This is most straightforward to judge from the the right panel of Figure 4.7. The left-hand side of (4.30) is the sum of the dark grey shaded rectangles, which strictly stay under the curve of $\mathcal{L}(X)$. The right-hand side of (4.30), on the other hand, is the sum of lightgrey shaded rectangles, which are always on top of $\mathcal{L}(X)$.

A good estimate of $P(\boldsymbol{D}|\mathrm{H})$ is obtained if the upper bound and lower bound are similar, which is the case as m grows. In summary, nested sampling is performed as follows. The iteration counter is first set to $i = 0$ and the initial prior volume X_0 is set to 1. A predefined N "active" samples are drawn from the prior distribution $P(\boldsymbol{\theta})$. At $i = 1$ step, the samples are sorted in order of their likelihood. The smallest (denoted as L_1) is removed from the active set (becoming "inactive") and replaced by a sample drawn from the prior distribution subject to the constraint that the likelihood function of the new sample is larger than L_1. The prior volume contained within the new active sample is a random variable given by $X_1 = t_1 X_0$, where t_1 is a random variable with probability density function $P(t) = Nt^{N-1}$, i.e., t_1 has the same distribution as the largest of N samples drawn uniformly from the interval $[0, 1]$. At each subsequent iteration i, the removal of the lowest likelihood point (denoted as L_i) in the active set, the drawing of a replacement sample with likelihood greater than L_i, and the reduction of the corresponding prior volume $X_i = t_i X_{i-1}$ are repeated. The likelihood for the first removed sample L_1 can be treated as an estimate of $\mathcal{L}_1$. This is true because L_1 is the smallest likelihood of uniform sample of size N. Thus, it can be used to estimate the

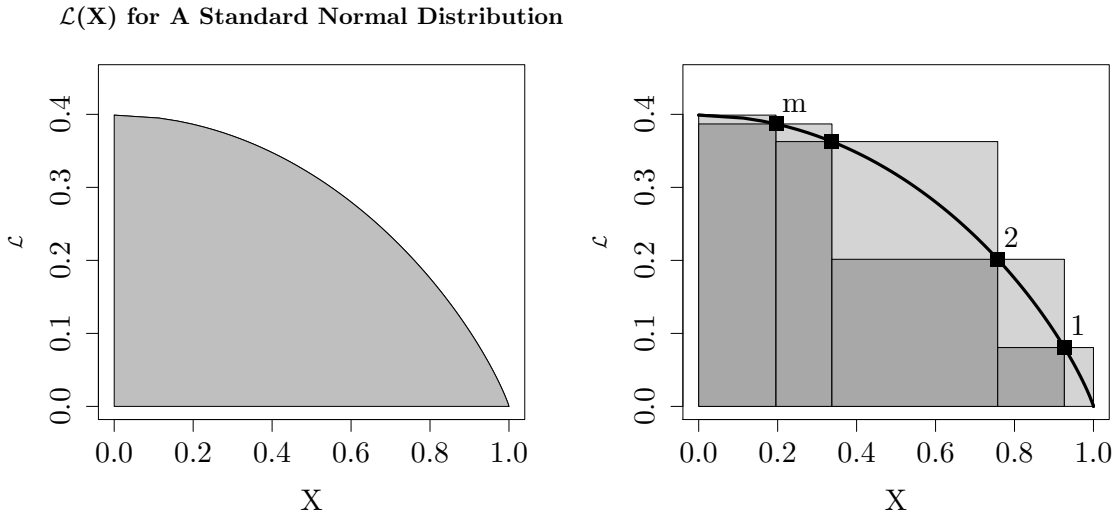

Figure 4.7: Nested Sampling Illustration. The left plot shows the functional relationship between $\mathcal{L}$ and X for a standard Normal distribution where area of the gray region equals the marginal density of the data. The right plot illustrates how we can numerically compute the marginal density of the data if can evaluate $\mathcal{L}(X_i)$ at a right-to-left sequence of m points.

$(\frac{N-1}{N})^{\text{th}}$ quantile of the whole likelihood function, i.e., $X^{-1}(\frac{N-1}{N})$ or $\mathcal{L}(\frac{N-1}{N})$. On the other hand, $\mathrm{E}(X_1) = \frac{N-1}{N}$. Hence, $L_1 \approx \mathcal{L}(X_1) = \mathcal{L}_1$. Similarly, at the i^{th} iteration, the smallest likelihood among the N live points, L_i, is an estimate of $X^{-1}((\frac{N-1}{N})^i)$ or simply $\mathcal{L}_i$. With such approximation, it is then straightforward to estimate the integral using numerical technique like (4.30). We stop the iteration by monitoring the size of the contribution $(X_{i+1} - X_i)\mathcal{L}_{i+1}$. Up till now, the only challenge left for the nested sampling algorithm is how to draw samples under the contraint of increasing likelihood at each iteration. A detailed discussion and an efficient algorithm can be found in Feroz and Hobson (2008) and Feroz *et al.* (2009).

In our experience with the emission line detection problem, Nested Sampling using the MultiNest algorithm works the best for our simulation studies. We apply it using the `PyMultiNest`, a Python wrapper for MultiNest (Feroz and Hobson, 2008) written by Johannes Buchner. A detailed tutorial that explains how to configure `PyMultiNest` so that it can work together with `CIAO` and `Sherpa` for X-ray analysis appears in Appendix A.2.

4.4 Methodology

In this section, we describe how we study the influence of prior distribution on the Bayes factor. We also introduce a set of candidate models to faciliate the comparison of the prior influence on the Bayes factor and on the ppp-value.

4.4.1 Graphical Representation Method

Here we propose a method to quantify the effect of the choice of prior distributions on the Bayes factor. Consider again the Example 1(b). In it α, β, and λ are fixed. The only unknown parameter is μ and we use a discretized Normal distribution centered at its value under the generative model as its prior distribution. We introduce a prototype analysis in a comparison of the Continuum model and Continuum+Spectral line model, by plotting the decision boundary based on the Bayes factor as a function of the prior standard deviation of μ. The details are shown in Figure 4.8, where we simulate three data sets with a strong, moderate, and weak line and plot their corresponding functional relationships between the log Bayes factor and the prior standard deviation of μ. We compare the set of priors that results in (i) evidence for the Continuum model; (ii) evidence for the Continuum+Spectral line model; and (iii) indifference between the two models. If all reasonable priors correspond to one of these three sets, the model comparison has a clear outcome. E.g., the red line in Figure 4.8 shows evidence for an emission line for all priors, and the blue line shows evidence for no line for all priors. If, on the other hand, the range of reasonable priors extends into two (or three) of these sets (e.g., black line), we cannot clearly enunciate the outcome of the comparison, but we can state how that outcome depends on the choice of the prior distributions.

In practice, the Bayes factor depends on more than one hyper-parameter. But the influence

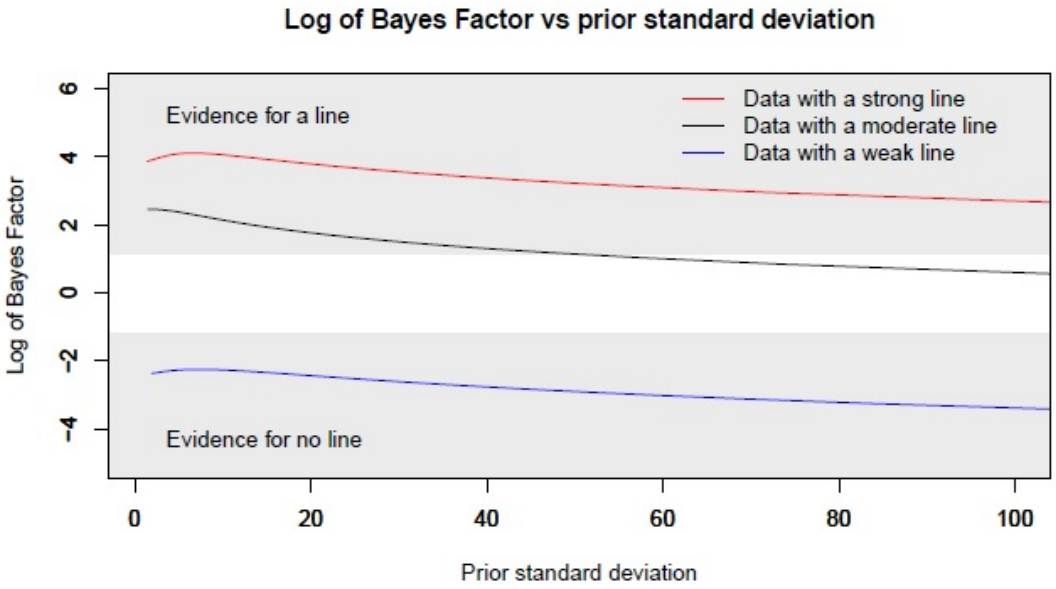

Figure 4.8: Log Bayes Factor as a Function of Prior Standard Deviation of μ for Three Simulated Spectra in the Running Example. Bayes factors indicating evidence for and against the Continuum+Spectral line model are shaded. For some data sets (red and blue lines) the better model is clear, regardless of the prior. For others (black line) the plot indicates what priors correspond to evidence for or against the line or simply no conclusion.

of prior distributions for those parameters appearing in only one of the models under comparison are expected to be most important. We verify this for the line detection problem via the Simulation studies describe in Section 4.5.1. As a result, we focus our attention on the hyper-parameters for λ and μ. In another word, we extend the graphical representation prototype to include all of the hyper-paramters for λ and μ. When the number of the hyper-parameters are equal to 2, the Figure 4.8 becomes a three-dimensional heatmap; if the number of hyper-parameters is larger than or equal to 3, we use a series of heatmaps.

4.4.2 Quantitative Model Comparison

First, we describe the settings under which we study the influence of the choice of prior distributions on the Bayes factor. The two candidate models under comparison are the

Continuum model and the Continuum+Spectral line model. In particular, we assume a powerlaw Continuum and a delta function spectral line. Following (4.3), we can write the model selection problem as testing the following two hypotheses

$$\begin{aligned} \mathrm{H}_0 :&\Lambda_j(\boldsymbol{\theta}) = e^{-\mathrm{nH}\cdot\sigma(E_j)} \cdot \alpha E_j^{-\beta} \\ \mathrm{H}_1 :&\Lambda_j(\boldsymbol{\theta}) = e^{-\mathrm{nH}\cdot\sigma(E_j)} \cdot (\alpha E_j^{-\beta} + \lambda\delta_j(\mu)) \end{aligned} \tag{4.31}$$

where α, β are the parameters for the Continuum and λ, μ are the parameters for the spectral line and nH is the absorption parameter with $\sigma(E_j)$ is the known photo-electric cross-section. We assume the spectrum is observed at 1000 energy bins equally spaced between 0.3 to 7 keV so that $\boldsymbol{E} = \{0.3, 0.3067, \ldots, 7\}$.

We perform two set of simulation studies based on (4.31). The first simulation ignores absorption (i.e., nH = 0) and all instrumental errors including photon redistribution, varying effective area, and background contamination. It provides insight into the sensitivity of Bayes factors to the prior distribution for the line detection problem. The second simulation, on the other hand, mimics a real data set that is analyzed in Section 4.5.3. In both simulation studies, we assign Uniform prior distribution to α, β, and nH with no hyper-parameters; for λ, we use a Uniform distribution with hyper-parameter η while for μ, we use a discretized Normal distribution with hyper-parameter μ_0 and σ. In particular,

$$\lambda \sim U(0, \eta),\ \mu \sim DN(\mu_0, \sigma^2). \tag{4.32}$$

To implement the graphical method of Section 4.4.1, we assign a grid of values for η and σ, i.e., (η_i, σ_j) where $i = 1, \ldots, I$ and $j = i, \ldots, J$. Then, we plot the Bayes factor as a function of $(\eta_i, \sigma_j, \mu_0)$, denoted as $B_{01}(\eta, \sigma, \mu_0)$, using a heatmap with regard to (η, σ) with μ_0 being fixed to different values in each of several heatmaps.

Recall that we are interested in not only the influence of the prior distribution on the Bayes

factor, but also the comparison of this prior influence with that on ppp-values, i.e., the effect of the range of μ considered and correction for the look elsewhere effect. To do this, we consider three different alternative hypotheses, H_1, H_2, and H_3

$$\begin{aligned} &\mathrm{H}_0 : \text{no line,} \\ &\mathrm{H}_1 : \text{line at a known location,} \\ &\mathrm{H}_2 : \text{line in a known energy range of (a, b), and} \\ &\mathrm{H}_3 : \text{line at an unspecified location.} \end{aligned} \tag{4.33}$$

H_1 and H_3 can be regarded as special cases of H_2. When $a = b$, H_2 is equivalent to H_1 while when $a = E_0$ and $b = E_{1000}$, H_2 is equivalent to H_3. When computing p-values, the correction for the look elsewhere effect depends on the choice of alternative hypotheses. To compute the ppp-value using the LRT under these three alternative hypotheses, we restrict the region where the LRT searches for line to the interval of (a, b), i.e., for the third step of (4.13), we use

$$T_{\mathrm{LRT}}(\tilde{\boldsymbol{D}}^{(s)}) = -2\log(R(\tilde{\boldsymbol{D}}^{(s)})), \text{ where } R(\tilde{\boldsymbol{D}}^{(s)}) = \frac{\sup\limits_{\Theta_0} L(\boldsymbol{\theta}|\tilde{\boldsymbol{D}}^{(s)})}{\sup\limits_{\Theta, \mu \in [a,b]} L(\boldsymbol{\theta}|\tilde{\boldsymbol{D}}^{(s)})} . \tag{4.34}$$

Such restriction is equivalent to imposing a Uniform prior distribution, $\mu \sim U(a, b)$ on μ for the Bayes factor, which we use to replace the discrete Normal prior of (4.32) when comparing the prior influence between Bayes factors and ppp-values. For each pair of (a, b), we denote the corresponding ppp-value as $ppp(a, b)$. To compare the prior dependency of the Bayes factor to the look elsewhere effect on the ppp-value, we use a grid of (a, b) pairs, (a_k, b_k) where $k = 1, \ldots, K$. We then plot both the posterior probability of H_0, $P(\mathrm{H}_0|\boldsymbol{D})$, and the ppp-value as a function of $(b_k - a_k)$. (We assume H_0 and H_1 are equally likely as a priori when computing the posterior probability of H_0, i.e., $P(\mathrm{H}_0) = P(\mathrm{H}_1) = 0.5$.) We compare these two quantities in this way because the ppp-value tends to be interpreted as $P(\mathrm{H}_0|\boldsymbol{D})$

in practice while $(b_k - a_k)$ represents the area of the region where we search for the emission line. Apart from $(b_k - a_k)$, the Bayes factor (and $P(\mathrm{H}_0|\boldsymbol{D})$) also depends on the value of the other hyper-parameter η (The effect of the prior for α and β are neglible). Our comparison plot will include Bayes factors computed under several values of η.

4.5 Numerial Studies

In this section, we perform two simulation studies to understand the influence of the choice of prior distribution on the Bayes factor. We also compare the prior dependency of the Bayes factor to that of the ppp-value. Finally, we use the Bayes Factor to detect an emission line in each of six real *Chandra* observations of source PG 1634+706 and compare our findings to both a simulation designed to mimic the real data and existing published results based on the posterior predictive distribution (Park *et al.*, 2008).

4.5.1 Simulation I

We start with a simpler simulation without considering the instrumental errors and absorption (we set nH = 0). In particular, we follow (4.31) to choose between a Continuum model and a Continuum+Spectral line model. For the generative model, we set $\alpha = 50$, $\beta = 1.69$, and $\mu = 1.3$. To vary the evidence in the data for the model with added emission line, we directly change the observed photon count in the bin corresponding to the emission line, $\boldsymbol{D}_{150}$. In particular, we set $\boldsymbol{D}_{150}$ to 49, 54, or 60, which correspond to about 3, 4, and 5 standard deviation above the intensity that is expected under the powerlaw continuum at this bin, 32.

Because the influence of the prior distributions for α and β are negligible[2] compared to those

[2]Using different priors for α and β has very little effect on the Bayes factor in this simulation settings.

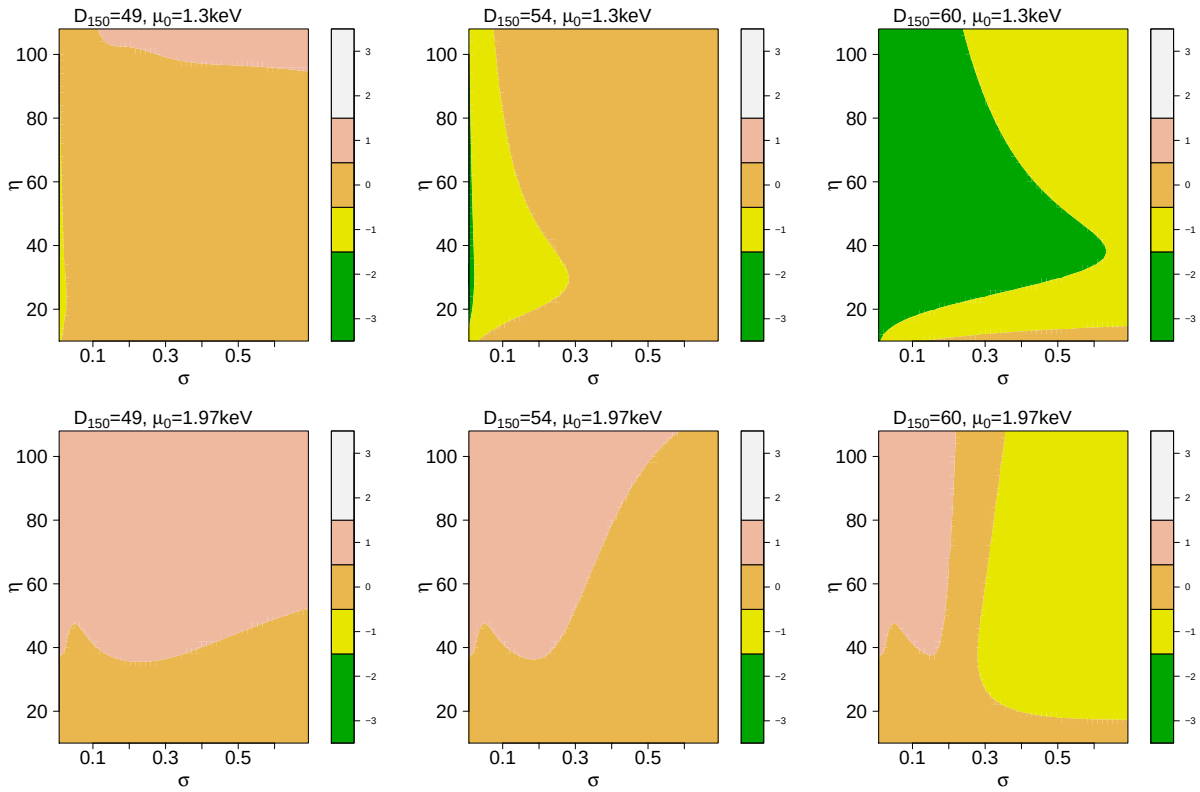

Figure 4.9: Heat Maps of The $\log_{10}(B_{01})$ in Simulation I. The first row correspond to $\mu_0 = 1.3$ while the second row $\mu_0 = 1.97$. Three columns represent the three different values of $\boldsymbol{D}_{150}$. From left to right, we have $\boldsymbol{D}_{150}$ equal to 49, 54, and 60 respectively. The different colors of the heat map correspond to the Jeffrey scale for the Bayes factor interpretation.

of λ and μ, we treat α and β as known in this simulation and we fix both at their values under the generative model. On the other hand, we follow (4.32) to set the prior distributions for λ and μ. To implement the graphical representation method, we let $I = J = 50$; $\{\eta_i\} = 10, 12, \ldots, 108$; $\{\sigma_j\} = 0.0067, 0.0201, \ldots, 0.6633$. We use two different values, 1.3 and 1.97, for μ_0 of which the first value corresponds to the true line location. For this simple simulation, we use the `integrate()` function of R to compute the Bayes factor. To present our findings, we plot the $\log_{10}[B_{01}(\eta_i, \sigma_j, \mu_0)]$ as a function of η and σ using two heat maps; one with $\mu_0 = 1.3$ and the other with $\mu_0 = 1.97$. The heatmaps are repeated for each of the three different values of $\boldsymbol{D}_{150}$. The results are shown in Figure 4.9.

Judging from the first row of Figure 4.9, when the peak of the prior for the emission line

Table 4.2: The Influence of Line Location on the ppp-value.

H_a \ $\boldsymbol{D}_{150}$	49	54	60
H_1: line at a known location	0.004	0.001	0
H_3: line at an unspecified location	0.539	0.184	0.006

location parameter is near the value used to generate the data, for each η, the Bayes factor increases as σ decreases. On the other hand, for each σ, the Bayes factor first increases and then decreases as η increases. Although the Bayes factor is indeed prior dependent, its dependence is quite predictable. For example, μ_0 and σ control the energy region where we look for the line. Intuitively, searching for the line in a small neighborhood of the true line location increases the chance of discovery (correct μ_0 and small σ). In fact, this is also the case for the ppp-value. Table 4.2 lists the two different ppp-values[3] comparing H_0 *vs* H_1 and H_0 *vs* H_3 of (4.33), respectively. The ppp-values are significantly larger under the alternative hypothesis with an unspecified line location. It is much harder to find evidence for an emission line without knowning its location using ppp-values.

To give a more detailed look at the comparison between the prior dependency of the Bayes factor to the look elsewhere effect on the ppp-value, we follow Section 4.4.2 and plot $P(H_0|\boldsymbol{D})$ as well as ppp-values in Figure 4.10 for $\boldsymbol{D}_{150} = 54$, where the horizon axis represents the span of the Uniform prior distribution, $(b_k - a_k)$. In particular,

$$P(\mu) \sim \begin{cases} U(1.3 - \frac{b_k - a_k}{2}, 1.3 + \frac{b_k - a_k}{2}), & \text{if } b_k - a_k \leq 2 \\ U(0.3, b_k - a_k + 0.3), & \text{if } b_k - a_k > 2 \end{cases}$$

As expected, both $P(H_0|Y)$ and the ppp-value increase with more diffuse prior distribution for μ, which is equivalent of searching the emission line in a larger energy range. However, the posterior probability of H_0 (and also the Bayes factor) is consistently more conservative

[3]We use 1000 replicates of data sets for the computation of the ppp-value. It is also equivalent to the bootstrap p-value in this case since there is no free parameter under H_0.

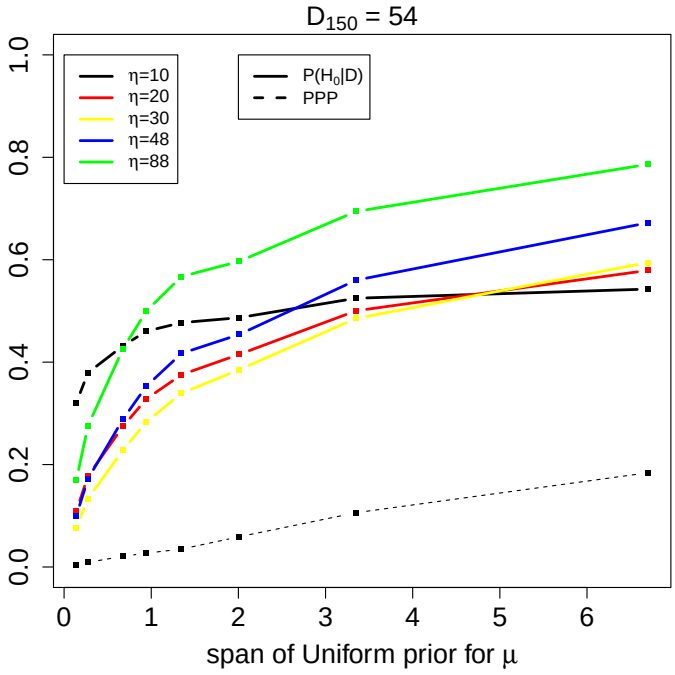

Figure 4.10: The Comparison Between The Posterior Probability of H_0 And The PPP-Value. The solid lines represent $P(H_0|\boldsymbol{D})$ where different colors correspond to different values of η. The ppp-value is represented by the black dashed line.

compared to the ppp-value. All of the solid lines are well above the dashed line by a large margin in Figure 4.10. This agrees with the criticism that p-values (including ppp-values) tend to overstate the evidence for the alternative hypothesis for testing a precise hypotheses Berger and Delampady (1987).

4.5.2 Simulation II

In this section, we perform another simulation study intended to mimic one of the real data sets, ObsID 47 of PG 1634+706. The details of the data set is discussed in Section 4.5.3. The problem of interest is still to choose between a Continuum model and a Continuum+Spectral line model using the model described in (4.31). For the generative model, we set nH = 0.064, $\alpha = 0.00043$, $\beta = 1.99$, and $\mu = 2.88$. The first three parameter values are equal to the

maximum likelihood estimates fitted by `Sherpa`[4] to ObsID 47, while the value of μ is based on the published results of Park *et al.* (2008). According to the generative model, the powerlaw intensity, $\alpha E^{-\beta}$, at $E = 2.88$ is equal to 0.00005. We use three different values for the intensity parameter of the emission line, $\lambda = 0.000005,\ 0.00001$, and 0.000025, which corresponds to a line of 10%, 20%, and 50% of the continuum intensity. To generate simulated data, we use the `fake_pha` routine in `Sherpa` to obtain one spectrum under each value of λ. For the instrumental errors, we use the same photon redistribution, varying effective area, and background contamination information associated with ObsID 47. We call these three simulated spectra the weak line, moderate line, and strong line simulation respectively. All three simulated spectra as well as the observed ObsID 47 are plotted in Figure 4.11.

To study the influence of the prior distributions on the Bayes factor, we assign the following Uniform distributions as the prior for α, β, and nH,

$$\alpha \sim U(0, 0.001),\ \beta \sim U(0, 10),\ \text{and nH} \sim U(0, 0.1)$$

These priors are non-informative in that they contain the range of plausible values for the three parameters. For the other two parameters, we use the prior distribution in (4.32) except that we fix $\mu_0 = 2.74$ and replace the discretized Normal distribution into a discretized truncated Normal distribution with a lower bound of 1, i.e.,

$$\lambda \sim U(0, \eta),\ \mu \sim TDN_1(2.74, \sigma^2). \tag{4.35}$$

We fix $\mu_0 = 2.74$ because this is the energy level where the Fe K_α emission line is expected to be for ObsID 47. We use the same μ_0 to analyze the real data sets. (Note that this is somewhat different from the generative value of $\mu = 2.88$.) We use the discretized truncated Normal prior distribution in order to avoid regions with potential calibration issues and

[4] `Sherpa` is the modeling and fitting application for `CIAO`, which is capable of fitting complex statistical models with instrumental errors explained in 4.2.1.

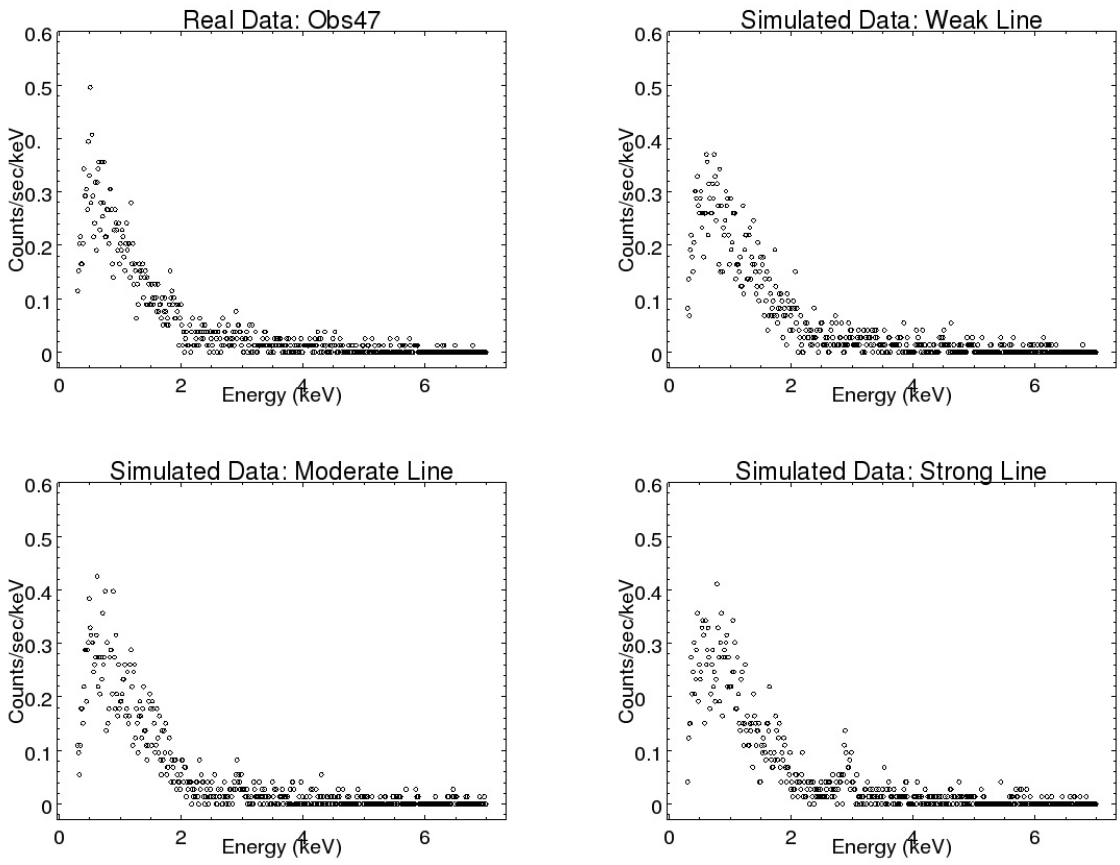

Figure 4.11: The Spectra of the Simulated Data Sets with Instrumental Errors based on ObsID 47.

effects related to absorption, i.e., we confine our attention to energies above 1 keV when searching for emission lines in ObsID 47 as Park *et al.* (2008). Because our simulated spectra share the same instrumental error information as ObsID 47, we make the same restriction in this simulation.

For the grid points of (η_i, σ_j), we set $I = J = 5$; $\{\eta_i\}$ are five equally spaced values between 0.000005 to 0.00007; $\{\sigma_j\}$ are five equally spaced values between 0.1 to 1.1. To compute the Bayes factor, instead of using the `integrate()` function of `R` as in Simualtion I, we use Nested Sampling via the `PyMultiNest` and `Sherpa` Python modules. A tutorial that describes how to configure these modules so that they can work together in Python is provided in Appendix A.2. The heat maps of the $\log_{10}(\mathrm{B}_{01}(\eta, \sigma, \mu_0 = 2.74))$ appears in the first row of

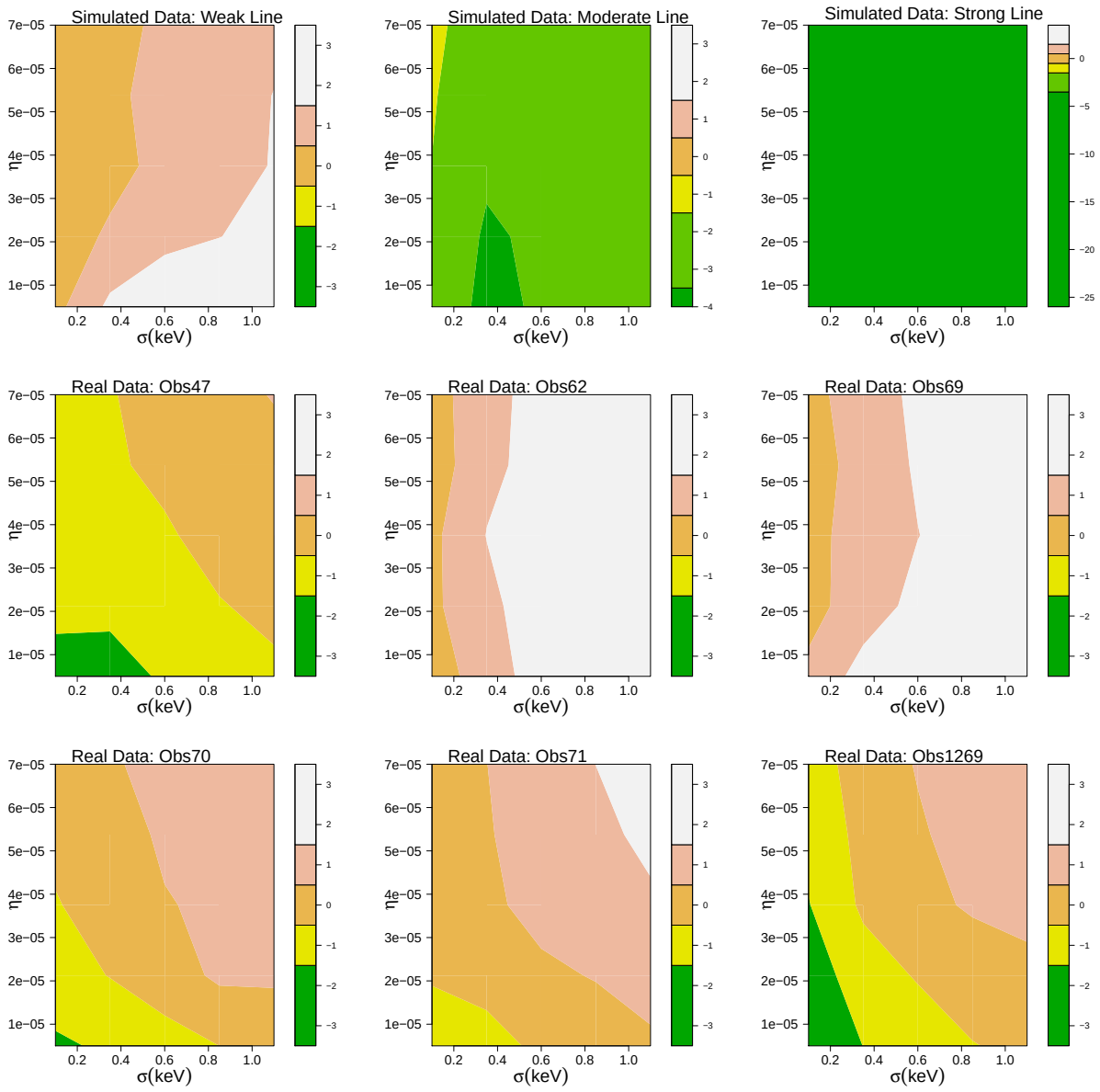

Figure 4.12: Heat Maps of the $\log_{10}(B_{01})$ for the Three Simulated Spectra And Six Real *Chandra* Observations. Note that for the moderate line and strong line cases of the simulated data, their heat maps have different scales compared to the others as these two simulated spectra have much smaller Bayes Factor than all other data sets.

Figure 4.12. Judging from the heat maps, under all of the prior distributions we choose, the Bayes factor favors the model with the emission line for both the moderate and strong line simulation. For the weak line simulation, on the other hand, the result depends on the choice of prior distribution on μ and λ.

As in Simulation I, to compare the prior dependency of Bayes factor to the look elsewhere effect on ppp-values, we use a Uniform prior distribution for μ to that is symmetric of 2.74 keV, i.e., $\mu \sim U(a, b)$ where $(a+b)/2 = 2.74$. To account for the instrumental errors, we use `pyBLoCXS` to draw a sample from the joint posterior distribution of (α, β). (`pyBLoCXS` can be run either within `Sherpa` or as a standalone Python module.) The ppp-value computed is based on 500 replications of the data sets. The results with $b_k - a_k = 0.1, 0.2, 0.5, 0.7, 1, 1.5$ are shown in the first row of Figure 4.13[5].

Both $P(\mathrm{H}_0|\boldsymbol{D})$ and the ppp-value strongly prefer the model with an emission line for the strong line case. When the simulated line is of moderate intensity, the ppp-value indicate strong evidence for the emission line, while $P(\mathrm{H}_0|\boldsymbol{D})$ does so only if $P(\mu)$ includes the true location of the line[6]. However, the stength of evidence from the ppp-value is always larger than that from the Bayes factor. The case of weak line is most interesting. For it, $P(\mathrm{H}_0|\boldsymbol{D})$ consistently finds little evidence for the Continuum+Spectral line model. The ppp-value, on the other hand, gives different answers depending on the value of $(b_k - a_k)$. It supports the model with a line if the search region is large, i.e., $b - a \geq 0.7$. A possible explanation is that in a detector bin outside the energy interval of $(2.74 - 0.35, 2.74 + 0.35)$, a large photon count is recorded due to chance variation. If the LRT is allowed to search for the emission line in a large enough interval containing this particular bin, it will be detected as an potential emission line and boost the evidence for the Continuum+Spectral line model; otherwise, we

[5]Because the likelihood function under H_1 is bumpy and multi-modal, finding the LRT statistic itself is challenging for the ppp-value computation. We use `Sherpa` with 30 equally spaced starting values between 1.24 and 4.24 keV for μ to find the LRT statistic for each of the replicated data set, which is the estimate with the largest LRT statistic among the 30 fits.

[6]Note that the first grid pair of (a, b) for the prior of μ is equal to $(2.64, 2.84)$. Since the value of μ for the generative model is equal to 2.88, it does not include the true line location.

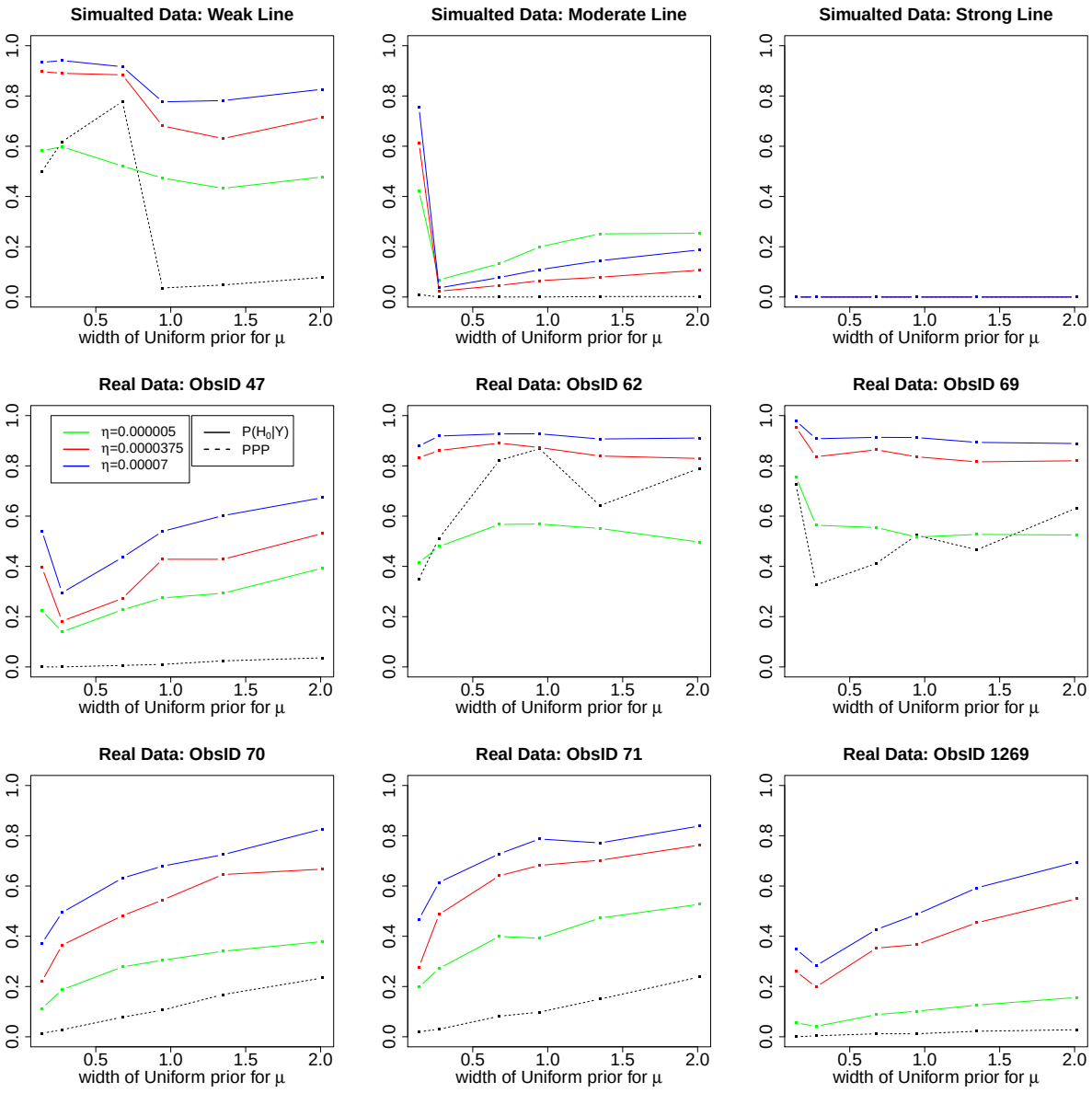

Figure 4.13: The Comparison Between The Posterior Probability of H_0 And The PPP-Value. For all plots, the solid lines represent $P(H_0|\boldsymbol{D})$ with different colors representing different values of η (Green, red, and blue lines correspond to $\eta = 0.000005, 0.0000375$ and 0.00007 respectively). The ppp-value is represented by the black dashed line. For the case of strong simulated line, all four lines are exactly equal.

cannot find evidence for the emission line using LRT.

4.5.3 Real Data Analysis

In this section, we repeat the analysis performed on simulation II to six real *Chandra* observations, PG 1634+706. PG 1634+706 (redshift $z = 1.334$) is a radio-quiet and optically bright quasar (Steidel and Sargent, 1991). It is very luminous in X-rays with the 2-10 keV band luminosity exceeding 10^{46} erg s^{-1} (Jiménez-Bailón *et al.*, 2005). The iron emission line in such luminous sources is expected to be weaker than in lower luminosity active galactic nucleus (AGN) (Nandra *et al.*, 1997). The quasar as observed with *The Advanced Satellite for Cosmology and Astrophysics* (ASCA) (George *et al.*, 2000) and *X-ray Multi-Mirror Mission* (XMM-Newton) (Page *et al.*, 2005) and no line was detected at the energy of the 6.4 keV Fe K_α line (observed at $E_{obs} = 2.738$ keV). However, the narrow line was detected in Haro-Corzo *et al.* (2007) analysis of one *Chandra* data set, ObsID 1269, at $E_{obs} = 2.84$ keV. Park *et al.* (2008) analyzed the complete six data sets using a new efficient MCMC algorithms to explore the complex posterior distribution of the location of narrow emission lines and to test for the presence of narrow emission lines using the ppp-values. They fit the model given in (4.31), i.e., they assume a powerlaw model for the continuum and one delta function emission line. Part of their results are summarized in Table 4.3, which presents the 95% HPD regions for the delta function line locations. As such posterior distributions are usually multimodal, each of the 95% HPD regions is composed of a number of disjoint intervals. Table 4.3 also shows the local modes associated with each interval as well as its posterior probability[7]. Five out of the six data sets are shown to have a local mode close to 2.74 keV, where the Fe K_α emission line is identified. The only exception is ObsID 69.

Here we follow Park *et al.* (2008), but we use Bayes factors rather than ppp-values to search for a delta function emission line in the spectra of PG 1634+706 with energy near 2.74

[7]Only those intervals that have posterior probabilities greater than 5% are presented in the table

Table 4.3: 95% HPD Regions of the Delta Function Line Location from Park *et al.* (2008).

Observered Data Set	Posterior Mode (keV)	95% HPD Region (keV)	Posterior Probability (%)
ObsID 47	2.885	(2.44, 3.14)	72.48
	5.915	(5.44, 5.92)	8.48
ObsID 62	1.885	(1.00, 1.97)	25.88
	2.785	(2.05, 3.02)	21.65
	3.925	(3.30, 4.06)	28.27
	5.395	(4.99, 5.41)	8.17
ObsID 69	1.955	(1.51, 2.65)	55.75
	3.535	(2.84, 3.62)	12.41
	3.935	(3.70, 4.01)	11.79
ObsID 70	2.795	(2.37, 3.17)	63.22
	5.945	(5.34, 6.00)	15.96
ObsID 71	2.325	(1.75, 2.45)	8.81
	2.815	(2.50, 3.01)	42.11
	5.625	(5.38, 5.72)	30.25
ObsID 1269	2.995	(2.69, 3.08)	84.96

keV. When we look for emission lines, we confine our attention ot energies above 1 keV to avoid regions with potential calibration issues and effects related to absorption. We use the same priors, the same number of data replicates for the ppp-value, and the same method of computing the LRT statistic via `Sherpa`. The second and third rows of Figure 4.12 show the heat maps for the $\log_{10}(B_{01}(\eta, \sigma, \mu_0 = 2.74))$. Among the six observations, we do not find any evidence for the Spectral Line in Obs 62 and Obs 69 using the Bayes factor. This is not surprising for Obs 69 because its posterior distribution of μ does not have a local mode near 2.74 keV as shown in Tabel 4.3. Obs 62, on the other hand, does have a local mode at 2.785 keV. However, the posterior probability associated with this mode, 21.65%, is relatively small compared to the other four observations. For the two observations with highest posterior probability for local modes near 2.74 keV, Obs 47 and Obs 1269, the Bayes factor also show the strongest evidence for the emission line under most prior settings. As we can see from the comparison between our simulated spectra and the Obs 47, even if there exists an emission line near 2.74 keV, its intensity might not be very large. Thus, the Bayes factor shows the strongest evidence for the model with the line under prior distributions that

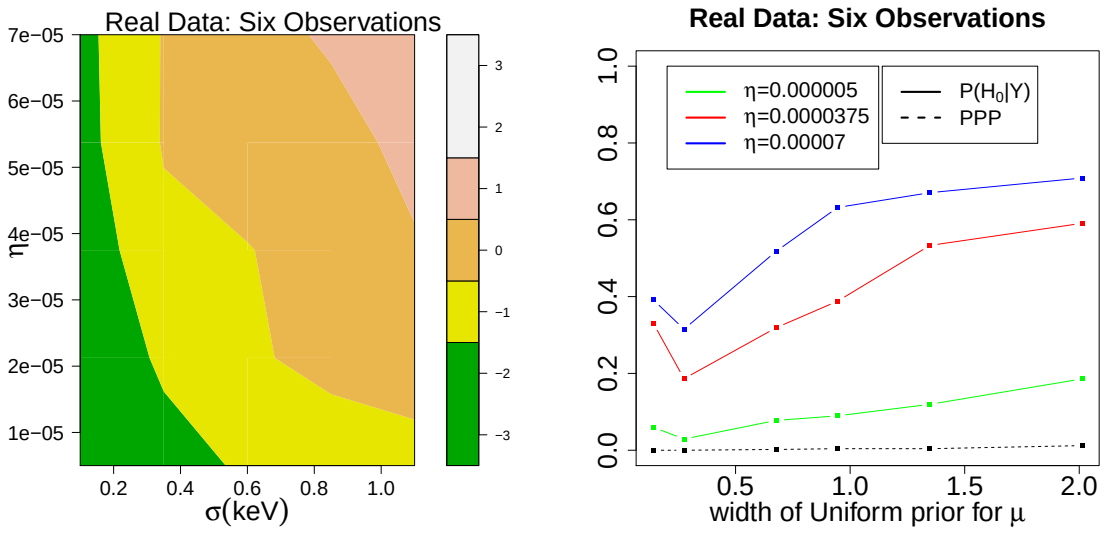

Figure 4.14: Heat Maps of the $\log_{10}(B_{01})$ When Fitting Six Real *Chandra* Observations Altogether.

favor weak lines, i.e., those with small values of η. Our findings are consistent with Park *et al.* (2008). We also tried combining all six observations. The corresponding heat map appears in the left panel of Figure 4.14. It looks most similar to the case of Obs 1269 but provides larger area with evidence for the emission line than all six observations.

The second and third rows of Figure 4.13 and the right panel of Figure 4.14 show similar results for the comparison between the $P(\mathrm{H}_0|\boldsymbol{D})$ and the ppp-value. When both the Bayes factor and the ppp-value find evidence for the model with line, the Bayes factor is more conservative compared with the ppp-value. This can be interpreted in two ways. Firstly, when the search region is the same, $P(\mathrm{H}_0|\boldsymbol{D})$ is significantly higher than the ppp-value. Secondly, if the search region is increased, the effect on on the Bayes factor is also larger than that on the ppp-value. In summary, the ppp-value always overstate the evidence for the spectral line relative to the Bayes factor.

4.6 Concluding Remarks

The spectral line detection problem in high-energy astrophysics is a typical example involving the test of a precise hypothesis. The asysmpototic distribution of the popular likelihood ratio test cannot be used for this problem as the regularity conditions for its asymptotic sampling distribution do not apply. Generally speaking, the null distribution of the likelihood ratio test statistic is unknown. While the ppp-value gives us a Monte Carlo mechanism for evaluating the null distribution, like the p-value, it tends to overstate the evidence for the more complicated model. The Bayes factor, on the other hand, provides an Bayesian alternative for the model selection and is applicable to a even larger class of spectral line / source detection problems. However, it is criticized to be sensitive to the choice of prior distributions.

In this chapter, we carefully study the prior influence of the Bayes factor in the context of spectral line detection problem. Although such prior dependency is usually thought of as a disadvantage for the Bayes factor, we find it could be quite useful from an objective point of view in practice. We discuss the specification of the prior distributions for a simple but typical class of spectral line detection problems, which involve a power law continuum with one extra delta function emission line. We find that we can use non-informative and even improper prior distribution for the continuum parameters. This prior has only a limited influence on the Bayes factor. The prior distribution for the location and intensity of the delta function emission line do have large influence on the Bayes factor. Different prior distribution for these two parameters can lead to different conclusions with the same data set. This effect, however, is not unlike the sensitivity of the ppp-value to the choice of search region. More importantly, we find that the different prior distributions represent different types of scientific questions. The prior for the emission line location parameter specifies the region where we search for the line. We are penalized if we decide to search in a large area ("look elsewhere effect"). This effect influences the ppp-values in the same manner. The

prior for the intensity parameter of the emission line on the other hand, represent the relative strength of the line that we are looking for. Such information could be collected from either historical observations or experts knowledge. Intuitively, the Bayes factor performs best if we determine in advance what type of line to look for. When compared to the ppp-value, the Bayes factor tends to be more conversative towards the null hypothesis, which agrees with the well known tendency of p-values to overstate evidence for the (more complicated) alternative hypothesis if the null hypothesis is precise.

Bibliography

Ariew, R. (1978). *Ockham's Razor: A Historical and Philosophical Analysis of Ockham's Principle of Parsimony*. University Microfilms International.

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

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, New York.

Berger, J. O. and Delampady, M. (1987). Testing precise hypotheses. *Statistical Science* **2**, 3, 317–352.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p-values and evidence (with discussion). *Journal of the American Statistical Association* **82**, 112–139.

Bia, M., Flores, A. C., and Mattei, A. (2011). Nonparametric estimators of dose-response functions. *CEPS/INSTEAD* .

Bodnar, L. M., Davidian, M., Siega-Riz, A. M., and Tsiatis, A. A. (2004). Marginal structural models for analyzing causal effects of time-dependent treatments: An application in perinatal epidemiology. *American Journal of Epidemiology* **159**, 10, 926–934.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics* **25**, 3, 573–578.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **24**, 295–313.

D'Agostino, Ralph B., J. and Rubin, D. B. (2000). Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association* **95**, 451, 749–759.

Dey, D. K., Gelfand, A. E., Vlachos, P. K., and Swartz, T. B. (1998). A simulation-intensive approach for checking hierarchical models. *Sociedad de Estadística e Investigación Operativa. Test* **7**, 2, 325–346.

Duan, N., Manning, W. G., Morris, C. N., and Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. *Journal of Business & Economics Statistics* **1**, 2, 115–126.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* **7**, 1, 1–26.

Efron, B. and Feldman, D. (1991). Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association* **86**, 9–17.

Ertefaie, A. and Stephens, D. A. (2010). Comparing approaches to causal inference for longitudinal data: Inverse probability weighting versus propensity scores. *The International Journal of Biostatistics* **6**.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman and Hall, London.

Fenimore, E. E., Conner, J. P., Epstein, R. I., Klebesadel, R. W., Laros, J. G., Yoshida, A., Fujii, M., Hayashida, K., Itoh, M., Murakami, T., Nishimura, J., Yamagami, Y., Kondo, I., and Kawai, N. (1988). Interpretations of multiple absorption features in a gamma-ray burst spectrum. *Astrophysical Journal* **335**, L71–L74.

Feroz, F. and Hobson, M. (2008). Multimodal nested sampling: An efficient and robust alternative to markov chain monte carlo methods for astronomical data analyses. *Monthly Notices of the Royal Astronomical Society* **384**, 449–463.

Feroz, F., Hobson, M., and Bridges, M. (2009). Multinest: An efficient and robust bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society* **398**, 1601–1614.

Flores, C. A., Flores-Lagunes, A., Gonzalez, A., and Neumann, T. C. (2012). Estimating the effects of length of exposure to instruction in a training program: The case of job corps. *The Review of Economics and Statistics* **94**, 1, 153–171.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21–29.

Freeman, P. E., Graziani, C., Lamb, D. Q., Loredo, T. J., Fenimore, E. E., Murakami, T., and Yoshida, A. (1999). Statistical analysis of spectral line candidates in gamma-ray burst grb 870303. *The Astrophysics Journal* **524**, 753–771.

Fu, Y., Chen, J., and Li, P. (2008). Modified likelihood ratio test for homogeneity in a mixture of von mises distributions. *Journal of Statistical Planning and Inference* **138**, 667–681.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, 2nd edn.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistica Science* **13**, 2, 163–185.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6**, 733–807.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Genz, A. and Kass, R. E. (1993). Subregion adaptive integration of functions having a dominant peak. *technical report, Carnegie Mellon University, Dept. of Statistics* .

George, I. M., Turner, T. J., Yaqoob, T., Netzer, H., Laor, A., Mushotzky, R. F., Nandra, K., and Takahashi, T. (2000). X-ray observations of optically selected, radio-quiet quasars. i. the asca results. *The Astrophysics Journal* **531**, 52–80.

Geweke, J. (1989). Bayesian inference in econometric models using monte carlo integration. *Econometrica* **57**, 1317–1340.

Gross, E. and Vitells, O. (2010). Trial factors for the look elsewhere effect in high energy physics. *The European Physical Journal C* **70**, 525–430.

Hall, H. (1936). The theory of photoelectric absorption for x-rays and γ-rays. *Reviews of Modern Physics* **8**, 358.

Haro-Corzo, S. A. R., Binette, L., Krongold, Y., Benitez, E., Humphrey, A., Nicastro, F., and Rodríguez-Martínez, M. (2007). Energy distribution of individual quasars from far-ultraviolet to x-rays. i. intrinsic ultraviolet hardness and dust opacities. *The Astrophysics Journal* **662**, 145.

Hartry, A., Fitzgerald, R., and Porter, K. (2008). Implementing a structured reading program in an afterschool setting: Problems and potential solutions. *Harvard Educational Review* **78**, 1, 181–210.

Hasselbring, T. S. and Goin, L. I. (2004). Literacy instruction for older struggling readers: What is the role of technology? *Reading and Writing Quarterly* **20**, 123–144.

Hernán, M. A., Brumback, B., and Robins, J. M. (2002). Estimating the causal effect of zidovudine on cd4 count with a marginal structural model for repeated measures. *Statistics in Medicine* **21**, 1689–1709.

Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, chap. 7. Wiley.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161–1189.

Hogan, J. W. and Lancaster, T. (2004). Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical Methods in Medical Research* **13**, 17–48.

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

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* **171**, 2, 481–502.

Imai, K. and Ratkovic, M. (2013). Covariate balancing propensity score. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* Forthcoming.

Imai, K. and van Dyk, D. A. (2004). Casual inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* **99**, 854–866.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87**, 3, 706–710.

Jeffreys, H. (1961). *Theory of Probability (3rd ed.)*. Oxford University Press.

Jiménez-Bailón, E., Piconcelli, E., Guainazzi, M., Schartel, N., Rodriguez-Pascual, P. M., and Santos-Lleo, M. (2005). The xmm-newton view of pg quasars: ii. properties of the fe k-alpha line. *Astronomy Astrophysics* **435**, 449.

Jin, H. and Rubin, D. B. (2008). Principal stratification for causal inference with extended partical compliance. *Journal of the American Statisticial Association* **103**, 481, 101–111.

Jin, H. and Rubin, D. B. (2009). Public schools versus private schools: Causal inference with partial compliance. *Journal of Educational and Behavioral Statistics* **34**, 1, 24–25.

Joffe, M. M. and Rosenbaum, P. R. (1999). Propensity scores. *American Journal of Epidemiology* **150**, 4, 327–333.

Johnson, E., Dominici, F., Griswold, M., and Zeger, S. L. (2003). Disease cases and their medical costs attributable to smoking: An analysis of the national medical expenditure survey. *Journal of Econometrics* **112**, 135–151.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* **22**, 4, 523–539.

Kashyap, V. L., van Dyk, D. A., Connors, A., Freeman, P. E., Siemiginowska, A., Xu, J., and Zezas, A. (2010). On computing upper limits to source intensities. *The Astrophysical Journal* , 719, 900–914.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of The American Statistical Association* **90**, 430, 773–794.

Kim, J. S., Samson, J. F., Fitzgerald, R., and Hartry, A. (2010). A randomized experiment of a mixed-methods literacy intervention for struggling readers in grades 4 - 6: effects on word reading efficiency, reading comprehension and vocabulary, and oral reading fluency. *Reading and Writing* **23**, 1109–1129.

Lee, K. J., Guillemot, L., Yue, Y. L., Kramer, M., and Champion, D. J. (2012). Application of the gaussian mixture model in pulsar astronomy - pulsar classification and candidates ranking for the fermi 2fgl catalogue. *Monthly Notices of the Royal Astronomical Society* **424**, 2832–2840.

Lindley, D. V. (1957). A statistical paradox. *Biometrika* **44**, 187–192.

Long, Q., Little, R. J. A., and Lin, X. (2010). Estimating causal effects in trials involving multitreatment arms subject to non-compliance: A bayesian framework. *Journal of the Royal Statistical Society: Series C* **59**, 513–531.

Lu, B., Zanutto, E., Hornik, R., and Rosenbaum, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association* **96**, 1245–1253.

Mattei, A. and Mealli, F. (2011). Augmented designs to assess principal strata direct effects. *Journal of the Royal Statistical Society: Series B* **73**, 729–752.

McCulloch, R. E. and Rossi, P. E. (1991). A bayesian approach to testing the arbitrage pricing theory. *Journal of Econometrics* **49**, 141–168.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics* **22**, 3, 1142–1160.

Meng, X.-L. and Wong, W. h. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica* **6**, 831–860.

Moodie, E. E. and Stephens, D. A. (2012). Estimation of dose-response functions for longitudinal data using the generalised propensity score. *Statistical Methods in Medical Research* **21**, 2, 149–166.

Murakami, T., Fujii, M., Hayashida, K., Itoh, M., Nishimura, J., Yamagami, T., Conner, J. P., Evans, W. D., Fenimore, E. E., Klebesadel, R. W., Yoshida, A., Kondo, I., and Kawai, N. (1988). Evidence for cyclotron absorption from spectral features in gamma-ray bursts seen with ginga. *Nature* **335**, 234–235.

Nandra, K., George, I. M., Mushotzky, R. F., Turner, T. J., and Yaqoob, T. (1997). Asca observations of seyfert 1 galaxies. ii. relativistic iron kα emission. *The Astrophysical Journal* **477**, 2, 602.

Page, K. L., Reeves, J. N., O'Brien, P. T., and Turner, M. J. L. (2005). Xmm-newton spectroscopy of high-redshift quasars. *Monthly Notics of the Royal Astronomical Society* 195–207.

Palmer, D. M., Teegarden, B. J., Schaefer, B. E., Cline, T. L., Band, D. L., and Ford, L. A. (1994). Batse gamma-ray burst line search. 1: Search for narrow lines in spectroscopy detector data. *The Astrophysical Journal* , 2, L77–L80.

Park, T., van Dyk, D. A., and Siemiginowska, A. (2008). Searching for narrow emission lines in x-ray spectra: Computation and methods. *The Astrophysical Journal* , 688, 807–825.

Pérez, J. M. and Berger, J. O. (2002). Expected-posterior prior distribution for model selection. *Biometrika* , 3, 491–511.

Piro, L., Costa, E., Feroci, M., Frontera, F., Amati, L., Dal Fiume, D., Antonelli, L. A., Heise, J., in 't Zand, J., Owens, A., Parmar, A. N., Cusumano, G., Vietri, M., and Perola, G. C. (1999). The x-ray afterglow of the gamma-ray burst of 1997 may 9: Spectral variability and possible evidence of an iron line. *The Astrophysical Journal* , 2, L73.

Protassov, R., van Dyk, D. A., Connors, A., Kashyap, V. L., and Siemiginowska, A. (2002). Statistics, handle with care: Detecting multiple model components with the likelihood ratio test. *The Astrophysical Journal* , 1, 545–549.

Raftery, A. E. and Banfield, J. D. (1990). Stopping the gibbs sampler, the use of morphology and other issues in spatial statistics. *Annals of the Institute of Statistical Mathematics* 32–43.

Ranucci, G. (2012). The profile likelihood ratio and the look elsewhere effect in high energy physics. *Nuclear Instruments and Methods in Physics Research A* 77–85.

Robins, J. M. (1998). Marginal structural models. *1997 Proceedings of the American Statistical Association, Section on Bayesian Statistical Science* 1–10.

Robins, J. M. (1999). Marginal structural models *versue* structural nested models as tools for causal inference. In *M.E. Halloran and D. Berry, Editors. Statistical Models in Epidemiology: The Environment and Clinical Trials*, 95–134. NY: Springer-Verlag.

Robins, J. M., Blevins, D., Ritter, G., and Wulfsohn, M. (1992). G-estimation of the effect of prophylaxis therapy for pneumocystis cariniii pneumonia on the survial of aids patients. *Epidemiology* 319–336.

Robins, J. M., Hernán, M. A., and Brumback, B. (2000a). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**, 550–560.

Robins, J. M., van der Vaart, A., and Ventura, V. (2000b). The asymptotic distribution of p-values in composite null models. *Journal of American Statistical Association* **95**, 1143–1172.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association* **82**, 398, 387–394.

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

Rubin, D. B. (1978). Bayesian-inference for causal effects–role of randomization. *The Annals of Statistics* **6**, 34–58.

Rubin, D. B. (1980). Comments on "randomization analysis of experimental data: The fisher randomization test". *Journal of the American Statistical Association* **75**, 591–593.

Rubin, D. B. (1990). [on the application of probability theory to agricultural experiments. essay on principles. section 9.] comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science* **5**, 472–480.

Schafer, J. L. and Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods* **13**, 4, 279–313.

Schwartz, S. L., Li, F., and Mealli, F. (2011). A bayesian semiparametric approach to intermediate variables in causal inference. *Journal of the American Statistical Association* **106**, 1331–1344.

Seaman, S. R. and White, I. R. (2011). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research* **22**, 278–295.

Sinharay, S. and Stern, H. S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference* **111**, 209–221.

Skilling, J. (2006). Nested sampling for general bayesian computation. *Bayesian Analysis* **1**, 4, 833–860.

Slavin, R. E., Cheung, A., Groff, C., and Lake, C. (2008). Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly* **43**, 290–322.

Steidel, C. C. and Sargent, W. L. W. (1991). Emission-line and continuum properties of 92 bright qsos - luminosity dependence and differences between radio-selected and optically selected samples. *The Astrophysical Journal* **382**, 433–465.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science* **25**, 1, 1–21.

Tanner, M. A. and Wong, W. H. (1987). An application of imputation to an estimation problem in grouped lifetime analysis. *Technometrics* **29**, 23–32.

Wang, Y., Bembom, O., and van der Laan, M. J. (2007). Data-adaptive estimation of the treatment-specific mean. *Journal of Statistical Planning And Inference* **137**, 1871–1887.

Wilks, S. S. (1938). The large sample distribution of the likelihood ratio for testing composite hypothesis. *The Annals of Mathematical Statistics* **9**, 60.

Wolfson, J. and Gilbert, P. (2010). Statistical identifiability and the surrogate endpoint problem, with application to vaccine trials. *Biometrics* **66**, 1153–1161.

Yoshida, A., Murakami, T., Nishimura, J., Kondo, I., and Fenimore, E. E. (1992). in gamma-ray burst: Observations, analyses and theories. 399. Cambridge: Cambridge Univ. Press, c edn.

Appendix A

Appendix

A.1 Using Improper Prior for Common Parameters

In this section, we use the running example in Section 4.2.1 to show that it is possible to assign improper prior to parameters that appear in both H_0 and H_1. For the running example, assume now β and μ are known so that α and λ are the only unknown parameter we need to integrate against for the computation of the Bayes Factor. For their priors, instead of assigning a prior directly for λ, we assume

$$P\left(\frac{\lambda}{\alpha}\right) \sim U(0, \eta).$$

In another word, we specify the prior distribution for the line intensity under the scale of its relative intensity compared to the continuum model. To complete the prior specification for (α, λ), we use two different priors for $P(\alpha)$. Firstly, suppose $P(\alpha) \sim U(0, N)$ where N is an

integer, then the Bayes Factor could be re-writed as

$$B_{01}^{N} = \frac{\eta}{\displaystyle\int_0^{\eta} \frac{(1+\tilde{\lambda}/E_{\mu}^{-\beta})^{Y_\mu}}{(1+\tilde{\lambda}/\sum_i^L E_i^{-\beta})^{\sum_i^L Y_i+1}} \cdot \frac{P(\tilde{z}\le N)}{P(z\le N)}\, d\tilde{\lambda}}. \tag{A.1}$$

where $z \sim \text{Gamma}(\sum_i^L Y_i + 1,\ 1/\sum_i^L E_i^{-\beta})$, $\tilde{z} \sim \text{Gamma}(\sum_i^L Y_i + 1,\ 1/(\sum_i^L E_i^{-\beta} + \tilde{\lambda}))$, and $L = 1000$ represents the total number of detector bins.

If instead we use an improper prior on the range of $(0, +\infty)$, i.e., $P(\alpha) \propto 1$ for $\alpha > 0$, the Bayes Factor in this case is equal to

$$B_{01}^{\text{improper}} = \frac{\eta}{\displaystyle\int_0^{\eta} \frac{(1+\tilde{\lambda}/E_{\mu}^{-\beta})^{Y_\mu}}{(1+\tilde{\lambda}/\sum_i^L E_i^{-\beta})^{\sum_i^L Y_i+1}}\, d\tilde{\lambda}}. \tag{A.2}$$

Simply note

$$\frac{(1+\tilde{\lambda}/E_{\mu}^{-\beta})^{Y_\mu}}{(1+\tilde{\lambda}/\sum_i^L E_i^{-\beta})^{\sum_i^L Y_i+1}} \cdot P(\tilde{z}\le N) \quad \le \quad \frac{(1+\tilde{\lambda}/E_{\mu}^{-\beta})^{Y_\mu}}{(1+\tilde{\lambda}/\sum_i^L E_i^{-\beta})^{\sum_i^L Y_i+1}}$$

since $P(\tilde{z} \le N) \le 1$. By Lebegue dominated convergnence theorem,

$$\begin{aligned}
\lim_{N\to\infty} B_{01}^{N} &= \eta \cdot \frac{\lim_{N\to\infty} P(z\le N)}{\lim_{N\to\infty} \displaystyle\int_0^{\eta} \frac{(1+\tilde{\lambda}/E_{\mu}^{-\beta})^{Y_\mu}}{(1+\tilde{\lambda}/\sum_i^L E_i^{-\beta})^{\sum_i^L Y_i+1}} \cdot P(\tilde{z}\le N)\, d\tilde{\lambda}} \\
&= \eta \cdot \frac{1}{\lim_{N\to\infty} \displaystyle\int_0^{\eta} \frac{(1+\tilde{\lambda}/E_{\mu}^{-\beta})^{Y_\mu}}{(1+\tilde{\lambda}/\sum_i^L E_i^{-\beta})^{\sum_i^L Y_i+1}} \cdot P(\tilde{z}\le N)\, d\tilde{\lambda}} \\
&= \eta \cdot \frac{1}{\displaystyle\int_0^{\eta} \lim_{N\to\infty} \frac{(1+\tilde{\lambda}/E_{\mu}^{-\beta})^{Y_\mu}}{(1+\tilde{\lambda}/\sum_i^L E_i^{-\beta})^{\sum_i^L Y_i+1}} \cdot P(\tilde{z}\le N)\, d\tilde{\lambda}} \\
&= \eta \cdot \frac{1}{\displaystyle\int_0^{\eta} \frac{(1+\tilde{\lambda}/E_{\mu}^{-\beta})^{Y_\mu}}{(1+\tilde{\lambda}/\sum_i^L E_i^{-\beta})^{\sum_i^L Y_i+1}}\, d\tilde{\lambda}} \ = \ B_{01}^{\text{improper}}
\end{aligned} \tag{A.3}$$

Thus, we have successfully shown that the Bayes Factor under an improper prior for α is equal to the limit of those Bayes Factors under a series of proper Uniform priors. Hence, using improper prior for α is legitimate in this example.

A.2 Tutorial About Software Configuration For Bayes Factor Computation

The following tutorial will introduce how to configure a PC with a fresh installed Ubuntu-12.04-i386 to compute the Bayes factor using Nested Sampling inside `CIAO`.

A.2.1 Intall `CIAO`

The first step is to download and install `CIAO`. Relevant information for this step could be found from `http://cxc.harvard.edu/ciao/download/`. In this tutorial, I choose to do a custom installation. In particular, my installation includes all "Binary Packages" but excludes the "Calibration Database" ("CALDB"). During the installation, I use default options for each step except that I specify the installation directory to be `/opt/ciao-4.6`. There is a series of smoke test after the installation is completed. Do not panic if you see some of the tests failed. `CIAO` is a large analysis package for X-ray analysis while the only part we need for the Bayes factor computation is its modeling and fitting package `Sherpa`. In fact, my installation has 3 out of 42 smoke tests failed but it turns out my remaining steps are completely uninfluenced. However, if you want to diagnose your installation, a current bug list could be found at `http://cxc.harvard.edu/ciao/bugs/`.

A.2.2 Install the Python Module for Nested Sampling

MultiNest is an efficient way to do Nested Sampling. In order to perform MultiNest in Python, I use the `PyMultiNest` module developed by Johannes Buchner. For the installation, I follow the *pymultinest 0.4 documentation*, which could be found at `http://johannesbuchner.github.io/PyMultiNest/install.html`. Following the instruction, I do came across an error of "`A required library with BLAS API not found`" when trying to build the MultiNest. I solved the problem by searching `liblapack` and `libblas` in the Software center of Ubuntu and install all relevant packages.

A.2.3 Calling PyMultiNest And Sherpa as Python Modules

With `PyMultiNest` and `Sherpa` successfully installed, we could now call both as Python modules and calculate the Bayes factor using Nested Sampling taking into account of all instrumental errors. However, since the module of `PyMultiNest` is built against the local binary of Python while the module of `Sherpa` is built against the Python binary inside `CIAO`, it is very difficult to correctly specify the Python library pathes. The way I tried is to do the following:

1. Edit the `.ciaorc` file. In particular, firstly uncomment `PYTHON_PATH postpend`. Then uncomment `PYTHON CIAO` and change it into the location of your local Python binary. For my pc, this is `/usr/bin`. This step tells `CIAO` to use your local binary of Python even after `CIAO` is started.

2. Start `CIAO` using "`source /your_path_of_ciao_installation/bin/ciao.bash`". Doing so will automatically add those `Sherpa` relevant libraries pathes to the `PYTHONPATH`.

3. Start Python by simply typing "`python`" in your terminal. Because of step 1, this should now call the local Python binary of which we build the `PyMultiNest` module

for. To import `PyMultiNest`, we could simply type "`import pymultinest`". On the other hand, because of step 2, we should be available to import the `Sherpa` module as well via "`from sherpa.astro.ui import *`"

4. Sometimes you might see errors like "`RuntimeError: module complied against API version 7 but this version of numpy is 6`" when trying to import `Sherpa`. This is because the module of `PyMultiNest` and `Sherpa` are built against different Python binary, which obviously might involves using different version of extension libraries like `Numpy`. To avoid such errors, make sure the relevant modules for your local Python binary is up to date. And because the current `PYTHONPATH` record pathes to both `Numpy` modules for the two Python binaries, you could go to the location of the `Numpy` library for either Python binary and then import the `Sherpa` module so that you know exactly which `Numpy` would be imported.

A.2.4 Running The Test Code

With each step properly accomplished, you should now be able to run the following test code to calculate $\log(P(D|M))$, i.e., the logarithm of the marginal probability of data or half of the Bayes factor. The model we use here, M, is a power law subject to Photo-electric absorption, which corresponds to `xsphabs.abs1*powlaw1d.pl` in `Sherpa`. Our energy range of interest is from 0.3 to 7 keV. We use Uniform distributions as prior for all three parameters. In particular, we assume

$$\alpha \sim U(0, 0.001),\ \gamma \sim U(0, 10),\ nH \sim U(0, 0.1)$$

where α, γ are power law parameters while nH is the parameter for the Photo-electric absorption.

```
from sherpa.astro.ui import *
import pymultinest, math, numpy
import os, threading, subprocess
if not os.path.exists("chains"): os.mkdir("chains")

load_pha('/path to your data')
energy_range = {"lo":.3,"hi": 7}
notice_id(1,**energy_range)
set_stat("cash")
set_source("xsphabs.abs1*powlaw1d.pl")

def lfactorial(y):
    tmp = math.log( math.factorial(y) )
    return tmp

Y = get_data().counts
lf_Y = map(lfactorial,Y)

### Transform from the Uniform U(0,1) distribution into
### corresponding prior distributions
def myprior(cube, ndim, nparams):
    cube[0] = cube[0]*0.001
    cube[1] = cube[1]*10
    cube[2] = cube[2]*0.1

### Use the Sherpa calc_stat() function to calculate
### the cash statistic for each set of simulated parameters
```

```
def myloglike(cube, ndim, nparams):
    set_par(pl.ampl,  val=cube[0], frozen=True)
    set_par(pl.gamma, val=cube[1], frozen=True)
    set_par(abs1.nH, val=cube[2], frozen=True)
    lkhd1 = -0.5*calc_stat()- sum( lf_Y )
    return lkhd1

parameters = ["x", "y", "m"]
n_params = len(parameters)

### Run the Nested Sampling using MultiNest algorithm
pymultinest.run(myloglike, myprior, n_params, resume = False, verbose = False, sampling_
a = pymultinest.Analyzer(n_params = n_params)
s = a.get_stats()
log_evidence = s["global evidence"]
```

If you are able to run the entire sample code without errors, congratulations! You now would be able to compute the Bayes factor for astronomical data subject to intrumental errors. For any two candidate models of your interest, they just need to calculate their logarithm of the marginal probabilities. Then their difference will tell you the corresponding logarithm of the Bayes factor.