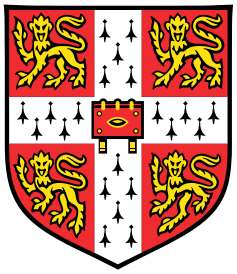
Weighting and moment conditions in Bayesian inference



Andrew Ho Man Yiu

MRC Biostatistics Unit
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Christ's College

March 2021

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Acknowledgements and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Acknowledgements and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Andrew Ho Man Yiu
March 2021

Abstract

Weighting and moment conditions in Bayesian inference

Andrew Ho Man Yiu

The work presented in this thesis was motivated by the goal of developing Bayesian methods for "weighted" biomedical data. To be more specific, we are referring to *probability weights*, which are used to adjust for distributional differences between the sample and the population. Sometimes, these differences occur by design; data collectors can choose to implement an unequal probability sampling frame to optimize efficiency subject to constraints. If so, the probability weights are known and are traditionally equal to the inverse of the unit sampling probabilities. It is often the case, however, that the sampling mechanism is unknown. Methods that use estimated weights include so-called *doubly robust* estimators, which have become popular in causal inference.

There is a lack of consensus regarding the role of probability weights in Bayesian inference. In some settings, it is reasonable to believe that conditioning on certain observed variables is sufficient to adjust for selection; the sampling mechanism is then deemed *ignorable* in a Bayesian analysis. In Chapter 2, we develop a Bayesian approach for case-cohort data that ignores the sampling mechanism and outperforms existing methods, including those that involve inverse probability weighting. Our approach showcases some key strengths of the Bayesian paradigm—namely, the marginalization of nuisance parameters, and the availability of sophisticated computational techniques from the MCMC literature. We analyse data from the EPIC-Norfolk cohort study to investigate the associations between saturated fatty acids and incident type-2 diabetes.

However, ignoring the sampling is not always beneficial. For a variety of popular problems, weighting offers the potential for increased robustness, efficiency and bias-correction. It is also of interest to consider settings where sampling is nonignorable, but weights are available (only) for the selected units. This is tricky to handle in a conventional Bayesian

framework; one must either make ad-hoc adjustments, or attempt to model the distribution of the weights. The latter is infeasible without additional untestable assumptions if the weights are not exact probability weights—e.g. due to trimming or calibration. By contrast, weighting methods are usually simple to implement in this context and are virtually model-free.

Chapters 3 and 4 develop approaches that are capable of combining weighting with Bayesian modelling. A key ingredient is to define target quantities as the solutions to moment conditions, as opposed to "true" components of parametric models. By doing so, the quantities coincide with the usual definitions if working model assumptions hold, but retain the interpretation of being projections if the assumptions are violated. This allows us to nonparametrically model the data-generating distribution and obtain the posterior of the target quantity implicitly. Crucially, our approaches still enable the user to directly specify their prior for the target quantity, in contrast to common nonparametric Bayesian models like Dirichlet processes.

The scope of our methodology extends beyond our original motivations. In particular, we can tackle a whole class of problems that would ordinarily be handled using estimating equations and robust variance estimation. Such problems are often called *semiparametric* because we are interested in estimating a finite-dimensional parameter in the presence of an infinite-dimensional nuisance parameter. Chapter 4 studies examples such as linear regression with heteroscedastic errors, and quantile regression.

Acknowledgements

This thesis would not have been possible without funding from the Medical Research Council. An earlier version of Chapter 3 was published as "Inference under unequal probability sampling with the Bayesian exponentially tilted empirical likelihood" in *Biometrika* (Volume 107, Issue 4, December 2020, p. 857-873). The work in Chapters 1, 2 and 4 are being prepared for submission.

First, I would like to thank my supervisors Dr. Robert Goudie and Dr. Brian Tom, who collaborated with me on all of the work presented in this thesis. Their patience, insight, generosity and guidance have made the last three and a half years a thoroughly enjoyable and productive experience. In particular, I would like to believe that some of their commitment to clear and accessible writing has rubbed off on me!

I am also thankful for helpful discussions with Dr. Stephen Sharp and Dr. Paul Newcombe, who are co-authors on the work in Chapter 2. Dr. Sharp also provided access to the EPIC-Norfolk dataset that was analysed in the chapter. My adviser Dr. Shaun Seaman gave valuable suggestions for the work in Chapter 3, which led to substantial improvements.

I would like to give special thanks to all of the support and technical staff members at the MRC Biostatistics Unit; their exceptional effort and planning have allowed work to continue as smoothly as one could ask for in such difficult circumstances.

Finally, I would like to thank my family for their love and unwavering support.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

The purpose of statistical inference is to generalize from data to an underlying process or population. Standard methods in the statistical toolbox—e.g. logistic and Cox regression—assume that the data are drawn directly from the target. In practice, however, datasets are often afflicted with missing observations and selection bias, and the failure to account for these issues can lead to drastically misleading results. Recent high-profile examples include the early forecasting of COVID-19 (Zhao et al., 2021) and the polling for the 2020 US general election (Panagopoulos, 2021). Collecting more data, which increases the sampling proportion relative to the population (if assumed to be finite), will not necessarily suffice. In fact, it can make matters worse; as Meng (2018) states, "The bigger the data, the surer we fool ourselves."

The work in this thesis was motivated by the objective of developing Bayesian methods for resolving this problem. The standard Bayesian approach (e.g. Section 8, Gelman et al., 2013) assumes that the missingness/selection mechanism is *ignorable* and drops out from the likelihood function if we condition on a sufficiently rich set of observed variables. However, adjusting for these variables can be difficult if they are high-dimensional, particularly if the sample size is relatively small. And the interpretations of our target quantities—such as regression coefficients—may become obscured if we condition on variables that are not of substantive interest.

Probability weighting provides a simple alternative; the units with observed data are weighted by the inverse of their sampling probabilities. This idea originated in the survey literature (Horvitz and Thompson, 1952) but has gained widespread interest due to the rising popularity of causal inference, particularly in the use of so-called *doubly robust* estimators. An overview of some of these developments is provided in §1.1.

It is often argued that probability weights should not be ignored, even if they are ignorable (Robins and Ritov, 1997, Hahn et al., 2020). But it is unclear how they should be incorporated into a Bayesian analysis. These difficulties have led some authors to suggest that Bayesian inference is inappropriate for handling this problem (Robins and Wasserman, 2012a,b, Robins et al., 2015). We discuss these issues in §1.2 and argue why we believe that a Bayesian approach can be desirable. Furthermore, we establish our philosophy of projection-based Bayesian estimation and describe how this can handle not only weighting but a wide-ranging class of problems involving moment conditions.

1.1 A selective overview

The existing literature on unequal probability sampling is vast, covering a wide range of estimands, conditions and applications. Our intention is not to provide a comprehensive review. Instead, we will examine a particular strand of work that will highlight the development of several core concepts and techniques. For illustrative purposes, we will focus throughout this section on the canonical example of estimating the mean of a one-dimensional outcome. Suggestions on how to generalize to other quantities will be provided later.

Our starting point is design-based survey inference (§1.1.1). In contrast to mainstream statistics, the population of interest is assumed to be finite, and the inferential uncertainty is attributed solely to the stochastic sampling mechanism. Remarkably, this framework enables consistent estimation under virtually no assumptions; it is perhaps one of the few exceptions to the famous motto "all models are wrong"[1]. But this does not restrict us from incorporating more traditional models; we will discuss how model-assisted estimation can leverage auxiliary variables and outcome regression models to increase efficiency without sacrificing robustness.

Semiparametric estimation forms the core of our overview (§1.1.2). In the 1990s, James Robins and his collaborators established a theoretical framework for coarsened data problems, including the *missing at random* mean estimation problem that we focus on. From a historical perspective, it is interesting to see how some of their ideas were developed in parallel with design-based inference. Motivated by finding semiparametric efficient estimators, they re-discovered model-assisted estimation as a special case.

The literature on semiparametric estimation is often very abstract. Our intention is to provide a relatively accessible introduction to certain concepts with an emphasis on the statistical motivations. Where possible, we have tried to avoid technical details, but some

[1]The only model assumption is that the sampling mechanism is actually implemented as assumed.

are important to properly understand the advantages and limitations of the theory. We also collect in Appendix A the proofs of some key facts that are frequently cited in the literature, but for which the proofs are either difficult to find in one place, or are contained as special cases within much more general results (accompanied by intimidating sets of notation and terminology!). While the results we prove are certainly not novel, we believe that we have contributed to making the reasoning more intuitive and self-contained. Our main references were van der Vaart (1998) and van der Vaart (2002).

A subsection (§1.1.2.2) is dedicated to the phenomenon of increased asymptotic efficiency from using estimated probability weights. Partly, this is because the material helps to bridge the gap between the (design-based inference motivated) inverse probability estimators and the more sophisticated estimators presented later on in the section. But we also believe that this topic is fascinating in its own right. Existing accounts of the phenomenon tend to be either almost completely mathematical (i.e. in terms of orthogonal projections in Hilbert spaces) or almost completely heuristic (along the lines of the circus elephants example given here). We focus on a middle ground and try to provide statistical intuition.

Much of the work in the literature assumes that the sampling/missingness mechanism is unknown. In this setting, the user is required to construct estimates of both the propensity score—which determines selection—and the outcome regression function. Doubly robust estimators (§1.1.2.3) have enjoyed popularity due to their ability to protect against the potential inconsistency of one of the two estimators. In recent years, researchers have discovered that doubly robust estimators have the benefit of being able to incorporate flexible machine learning methods while retaining attractive statistical properties. We will close this section by providing a brief introduction to these state-of-the-art methods (§1.1.3).

1.1.1 Design-based survey inference

Let $y_1, \ldots, y_N$ be constant values of an outcome variable for a population of known size $N < \infty$. The target quantity is the population outcome mean

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i.$$

The selection indicator variables $R_1, \ldots, R_N$ take the value 1 if the corresponding population outcome is observed, and 0 otherwise. The first-order sampling probabilities $\pi_i = \mathbb{P}(R_i = 1) = \mathbb{E}(R_i)$ are assumed to be known by design.

We can visualize this set-up by imagining N cards laid face down hiding the values of the outcome underneath. The data collector designs a sampling frame to determine how the cards are assumed to be randomly selected and flipped over. This differs from the *superpopulation* framework of conventional statistics, where a model for the outcome variables is specified, and the cards are assumed to have been drawn from an infinite deck. We will return to the superpopulation approach in the next subsection.

Why might a data collector be interested in a sampling frame that assigns unequal sampling probabilities? The answer lies in increased efficiency. It is often reasonable to believe that we can find strata within which the outcome variation is lower than in the whole population. Units in smaller strata can be assigned a relatively high sampling probability to help ensure each stratum is well-represented in the sample. Additionally, some strata could be of more interest than others. In Chapter 2, we study a sampling design where individuals who become cases are over-sampled because they provide more information in a survival analysis than controls.

Since we are likely to believe that the sampling probabilities are correlated with the outcome for the above reasons, a simple average of the observed outcome values will not suffice. Horvitz and Thompson (1952) introduced the following estimator

$$\hat{\mu}_{HT} = \frac{1}{N}\sum_{i=1}^{N}\frac{R_i y_i}{\pi_i.}$$

The inverse of the sampling probabilities are used to form a weighted average of the observed outcome values, such that units with a relatively small sampling probability are given more weight and vice-versa. Since $\pi_i = \mathbb{E}(R_i)$, it is clear that the Horvitz-Thompson estimator is unbiased (some authors use the term *design-unbiased* to emphasize the design-based set-up).

Despite the simplicity and unbiasedness of the Horvitz-Thompson estimator, it is rarely used in practice. The following example, paraphrased from Basu (1971), illustrates why. A circus owner wishes to estimate the average weight of his 50 elephants. Due to the cumbersome nature of weighing an elephant, he decides to form his estimate based on a single measurement. Upon reviewing the results from 3 years ago, he proposes to select Sambo, an elephant of average weight previously. The circus statistician is horrified and insists that the owner uses the Horvitz-Thompson estimator because it is design-unbiased. They devise a sampling frame where Sambo is selected with probability 99/100 and the remaining probability is shared equally among the other 49 elephants. Sambo is selected, and the statistician produces the absurd estimate of 2/99 multiplied by Sambo's weight. The incredulous owner asks what the estimate would have been if Jumbo, the big elephant, was

selected. The statistician replies "98 multiplied by Jumbo's weight", and subsequently loses his job.

In response to this example, Hájek (1971) proposed the alternative estimator

$$\hat{\mu}_{HJ} = \left(\sum_{j=1}^{N} \frac{R_j}{\pi_j} \right)^{-1} \sum_{i=1}^{N} \frac{R_i y_i}{\pi_i}.$$

Regardless of which elephant is selected, the Hájek estimator would be equal to the raw single measurement taken, and the circus statistician's job would have been saved. This superior performance may seem surprising at first; we appear to have gained from replacing the known population size N with an estimate $(\sum_{j=1}^{N} R_j/\pi_j)$. One possible explanation for this is that the Hájek weights are guaranteed to sum to 1; thus, the estimator will lie in the convex hull of the observed outcomes—a property known as *sample-boundedness*. The Horvitz-Thompson estimator does not have this guarantee and is prone to poor behaviour when the weights are variable, as is the case in Basu's elephants example.

Working with a finite population, the notion of consistency in a design setting is slightly different than usual. Let $\{\mathscr{F}_N\}$ be an increasing sequence of (possibly random) finite populations with associated sequences of sampling frames and finite population outcome means $\{\bar{y}_N\}$. An estimator $\hat{\mu}$ is *design-consistent* if for any $\varepsilon > 0$

$$\lim_{N \to \infty} \mathbb{P}_{\mathscr{F}_N}(|\hat{\mu} - \bar{y}_N| > \varepsilon) = 0 \quad \text{a.s.},$$

where $\mathbb{P}_{\mathscr{F}_N}$ denotes the probability with respect to $\mathscr{F}_N$ and its associated sampling frame. Both the Horvitz-Thompson and Hájek estimators are design-consistent under very mild conditions; see Fuller (2009) for further details.

Treating the outcome values as constants does not preclude the use of models and auxiliary variables. Given now that we also observe (possibly vector) auxiliary variable values $x_1, \ldots, x_N$ for each unit in the population, the *difference estimator* takes the general form

$$\hat{\mu}_{DIFF} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{R_i y_i}{\pi_i} + \left\{ 1 - \frac{R_i}{\pi_i} \right\} \hat{m}(x_i) \right), \tag{1.1}$$

where $\hat{m}$ is—loosely speaking—an estimate of the regression function "$\mathbb{E}(y \mid x)$" fitted using the data. Setting $\hat{m} = 0$ recovers the Horvitz-Thompson estimator.

The design-consistency of the difference estimator does not depend on correct specification of the regression function as long as $\hat{m}$ converges appropriately to some fixed function m

as the sample size increases; for this reason, use of the difference estimator is often referred to as *model-assisted estimation*, as opposed to model-based estimation. To see this, we can rewrite (1.1) as

$$\hat{\mu}_{DIFF} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{R_i y_i}{\pi_i} + \left\{1 - \frac{R_i}{\pi_i}\right\} m(x_i) + \left\{1 - \frac{R_i}{\pi_i}\right\}\{\hat{m}(x_i) - m(x_i)\}\right).$$

The first term is the Horvitz-Thompson estimator, the second term will tend to 0 due to the Horvitz-Thompson estimator for $m(x)$, and the remaining term will tend to 0 by the convergence of $\hat{m}$ to m. A more thorough argument will be given later, albeit in a slightly different setting.

If we fit our regression model as usual, the difference estimator can suffer the same kind of poor behaviour as the Horvitz-Thompson estimator. A possible strategy for fixing this can be seen by rewriting (1.1) as

$$\hat{\mu}_{DIFF} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{m}(x_i) + \frac{R_i}{\pi_i}\{y_i - \hat{m}(x_i)\}\right).$$

If we are able to fit $\hat{m}$ in such a way that

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{R_i}{\pi_i}\{y_i - \hat{m}(x_i)\}\right) = 0, \tag{1.2}$$

the difference estimator will take the form of a regression estimator[2], leading to more stable estimates. For example, if the outcomes are binary and we use a logistic regression model, the difference estimator will be guaranteed to lie between 0 and 1.

When $\hat{m}$ is estimated using a generalized linear model with the canonical link function, there are two simple ways (Firth and Bennett, 1998) to attain (1.2). The first is to implement weighted maximum likelihood with weights equal to R_i/π_i, so that (1.2) is equal to the intercept component of the weighted score function. The second is to perform standard maximum likelihood with an extra covariate $1/\pi_i$ added to the regression model; the component of the score function corresponding to this new covariate will be equal to (1.2). This approach is linked to the method of *targeted learning* that will be discussed later.

It is common that the auxiliary variables are also only observed for the sampled units. In this case, model-assisted estimation is still possible if the population mean of the auxiliary variables is known (or can be reasonably approximated). Using a linear regression model

[2]The sample mean of a regression function.

fitted with ordinary least squares, (1.1) specializes to

$$\begin{aligned}\hat{\mu}_{DIFF} &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{R_i y_i}{\pi_i} + \left\{1 - \frac{R_i}{\pi_i}\right\} x_i^{\mathrm{T}}\hat{\beta}_{OLS}\right) \\ &= \hat{\mu}_{HT} + \left(\frac{1}{N}\sum_{i=1}^{N} x_i^{\mathrm{T}}\right)\hat{\beta}_{OLS} - \frac{1}{N}\sum_{i=1}^{N}\left(\frac{R_i x_i^{\mathrm{T}}}{\pi_i}\right)\hat{\beta}_{OLS},\end{aligned}$$

where $\hat{\beta}_{OLS}$ is the ordinary least squares coefficient estimated from the sampled units. This is known as the linear generalized regression estimator (Cassel et al., 1976, Särndal et al., 1992).

1.1.2 Semiparametric estimation

Let us look at the problem from a slightly different perspective. Suppose that we observe independent and identically distributed data $D_1, \ldots, D_n$ from a distribution P known to belong to a set $\mathscr{P}$ of probability measures on a measurable space $(\mathscr{D}, \mathscr{A})$, where for each i, $D_i = (X_i, R_i, R_i Y_i)$, X_i is a vector of covariates, Y_i is a one-dimensional real outcome, and R_i is binary. We refer to $\mathscr{P}$ as our *model*.

As before, the target quantity is the outcome mean. In this setting, we denote the outcome mean by $\mu(P)$, where μ is the mapping $\mu : \mathscr{P} \to \mathbb{R}$ with $P \mapsto \mathbb{E}_P[Y]$. We assume that R and Y are independent given X—this assumption is sometimes referred to as *strong ignorability* (Rosenbaum and Rubin, 1983). For the time being, we will also assume that the function

$$\pi(X) = P(R = 1 \mid X),$$

named the *propensity score* (Rosenbaum and Rubin, 1983), is known, and that π is bounded away from 0 with probability 1—this is known as the *positivity* assumption. Aside from the above, we make no restrictions on our model. We will use the shorthand notation $Pf := \int f(d)\, dP(d)$. In particular, $\mathbb{P}_n f = n^{-1}\sum_{i=1}^{n} f(D_i)$, where $\mathbb{P}_n$ is the empirical measure.

There are a few differences to the set-up in the previous subsection. First, the outcomes and covariates are now treated as random variables, as if they were drawn from a hypothetical infinite superpopulation. Moreover, the covariates are assumed to be sufficient to adjust for the selection bias, and the distribution of R given X is known. Also, we have restricted our attention to a particular type of sampling design known as *Poisson sampling* (Fuller, 2009), where each unit is sampled independent of the others. A consequence is that the number of sampled units is random.

Nevertheless, we can attempt to estimate the outcome mean in much the same way as before. For example, the Horvitz-Thompson estimator

$$\hat{\mu}_{HT} = \mathbb{P}_n \left[\frac{RY}{\pi(X)} \right]$$

is again unbiased—this follows from using iterated expectations $\mathbb{E}\{\mathbb{E}(\cdot \mid X)\}$ and the strong ignorability assumption. In fact, assuming that $\text{var}[RY/\pi(X)] < \infty$ we can use the central limit theorem to deduce that the Horvitz-Thompson estimator is $\sqrt{n}$-consistent and asymptotically normally distributed (CAN) since it takes the form of a sample average.

The Hájek estimator

$$\hat{\mu}_{HJ} = \mathbb{P}_n \left[\frac{R}{\pi(X)} \right]^{-1} \mathbb{P}_n \left[\frac{RY}{\pi(X)} \right]$$

is also CAN. We can see this by defining the Hájek estimator to be the solution to the unbiased[3] estimating equation

$$\mathbb{P}_n[S_{HJ}(D, \mu)] := \mathbb{P}_n \left[\frac{R(Y - \mu)}{\pi(X)} \right] = 0.$$

Thus, the Hájek estimator belongs to the class of *Z-estimators* ("Z" stands for zero) and admits the following expansion under regularity conditions (van der Vaart, 1998):

$$\sqrt{n}(\hat{\mu}_{HJ} - \mu(P)) = -\sqrt{n}\mathbb{P}_n \left[P \left\{ \frac{\partial S_{HJ}}{\partial \mu}(D, \mu(P)) \right\}^{-1} S_{HJ}(D, \mu(P)) \right] + o_P(1).$$

We deduce that $\sqrt{n}(\hat{\mu}_{HJ} - \mu(P))$ converges in distribution to a mean-zero normal distribution with the "sandwich" covariance matrix

$$P \left\{ \frac{\partial S_{HJ}}{\partial \mu}(D, \mu(P)) \right\}^{-1} P \left\{ S_{HJ}(D, \mu(P))^2 \right\} P \left\{ \frac{\partial S_{HJ}}{\partial \mu}(D, \mu(P)) \right\}^{-1}.$$

The above can be estimated by replacing P with $\mathbb{P}_n$ and $\mu(P)$ with $\hat{\mu}_{HJ}$.

Can we do better than these two estimators? For parametric problems with standard regularity conditions, it is well-known that maximum likelihood estimation is *asymptotically efficient*. This means that the maximum likelihood estimator is CAN with the smallest possible asymptotic variance among all *regular* estimators. Roughly speaking, the asymptotic behaviour of a regular estimator is continuous with respect to the data-generating distribution—we will provide a more precise discussion later.

[3] This refers to the fact that S_{HJ} has mean zero under P when evaluated at $\mu = \mu(P)$.

Our model, however, is semiparametric; $\pi(X)$ is known, but the joint distribution of (Y,X) is unspecified, leaving us with an infinite-dimensional nuisance parameter. Given our interest in CAN estimators, a reasonable first step is to restrict our attention to estimators that admit an expansion

$$\sqrt{n}(\hat{\mu} - \mu(P)) = \sqrt{n}\mathbb{P}_n\{\psi(D)\}) + o_P(1), \tag{1.3}$$

where ψ is a measurable function with $P\{\psi(D)\} = 0$ and $P\{\psi(D)^2\} < \infty$. Such estimators are called *asymptotically linear* and ψ is called the *influence function*[4] of $\hat{\mu}$. We have already seen that the Horvitz-Thompson and Hájek estimators belong to this class.

The asymptotic variance of an asymptotically linear estimator is equal to the variance of its influence function. If we could characterize the set of possible influence functions, then we might be able to construct more precise estimators, e.g. by using the influence function to form a set of unbiased estimating equations, as was the case for the Hájek estimator. It turns out that such a characterization is indeed possible. The influence function of any regular and asymptotically linear (RAL) estimator must be a *gradient* of the target quantity with respect to the model. In order to elaborate on this statement, we will introduce some background in the following subsection.

1.1.2.1 Background

What does it mean to say that one estimator is more efficient than another for estimating a quantity at a distribution P? Clearly, any definition must depend on more distributions than just P itself. Otherwise, the constant estimator evaluated at P would always be considered efficient. Somehow, the complexity of the model must be taken into account. A parametric estimator might be viewed as inappropriate for a semiparametric model because there are distributions surrounding P that are not contained in the smaller, parametric model. Thus, there will generally exist certain directions along which we could deviate slightly from P such that the parametric estimator exhibits undesirable behaviour. Efficiency theory focuses on estimators that are insensitive to small local changes in the data-generating distribution in any direction.

The notion of "direction" is formalized by considering smooth, one-dimensional paths contained in the model $\mathscr{P}$ that pass through P. The set of permitted directions is the *tangent space* $\dot{\mathscr{P}}_P$ of the model $\mathscr{P}$ at P, containing measurable functions $g : \mathscr{D} \to \mathbb{R}$ such that $Pg = 0$ and $Pg^2 < \infty$. The tangent space is a subset of the Hilbert space $L_2(P)$ consisting

[4]This refers to the fact that $\psi(D_i)$ quantifies the (asymptotic) influence of a single observation D_i on the value of the estimator.

of all measurable functions $h : \mathscr{D} \to \mathbb{R}$ with $Ph^2 < \infty$, equipped with the inner product $\langle h_1, h_2 \rangle = P[h_1 h_2]$ and norm $\|h\| = \sqrt{Ph^2}$. We say that two elements h_1 and h_2 are *orthogonal* if $P[h_1 h_2] = 0$.

For each $g \in \dot{\mathscr{P}}_P$, we can exhibit a *parametric submodel* $\{P_{t,g} : t \in (-\varepsilon, \varepsilon) \subset \mathbb{R}\} \subset \mathscr{P}$ satisfying $P_{t,g}|_{t=0} = P$ and for every value d, we have

$$g(d) = \frac{\partial}{\partial t} \log dP_{t,g}(d)|_{t=0}. \tag{1.4}$$

In words, a parametric submodel is a one-dimensional model parameterized by t that is contained in $\mathscr{P}$ and passes through P at $t = 0$ with score function g.

The exact form of the parametric submodel is not particularly important[5]. We are only concerned with its score function g as it passes through P at $t = 0$. For bounded g, a common construction is

$$dP_{t,g} = (1 + tg)dP,$$

with ε chosen small enough such that the submodel stays within $\mathscr{P}$. We emphasize that parametric submodels are not meant to be substantively meaningful, despite "model" appearing in the name. As suggested above, we can simply view parametric submodels as paths contained in $\mathscr{P}$ that cross P in directions identified by their score functions.

We can now give a precise definition of regularity. An estimator $\hat{\mu}$ is called *regular* for estimating $\mu(P)$ relative to $\dot{\mathscr{P}}_P$ if there exists a probability measure L such that

$$\sqrt{n}(\hat{\mu} - \mu(P_{1/\sqrt{n},g})) \overset{P_{1/\sqrt{n},g}}{\rightsquigarrow} L \tag{1.5}$$

for every $g \in \dot{\mathscr{P}}_P$ and any parametric submodel $\{P_{t,g}\}$ with score function g. For each N, the underlying data-generating distribution is $P_{1/\sqrt{n},g}$ and the arrow $\rightsquigarrow$ denotes convergence in distribution. In less formal terms, the limiting distribution of the estimator at P is the same, no matter which direction we approach from.

We also require that the target quantity possesses the following property: we say that μ is *pathwise differentiable* at P with respect to $\dot{\mathscr{P}}_P$ if

- the mapping $t \mapsto \mu(P_{t,g})$ is differentiable, and

[5] A technical requirement is *differentiability in quadratic mean* (see Chapter 25 of van der Vaart (1998)). This is used to establish a property known as *local asymptotic normality*, which allows a local change of measure that is central to the proof of Theorem 1.1.

- there exists[6] is real, a necessary and sufficient condition for existence is that the derivative map is continuous and linear in g, by the Riesz representation theorem (see p. 363 of van der Vaart (1998)). a fixed, measurable function $\psi : \mathscr{D} \to \mathbb{R}$ such that

$$\frac{\partial \mu(P_{t,g})}{\partial t}|_{t=0} = P[\psi g] \tag{1.6}$$

for every $g \in \dot{\mathscr{P}}_P$ and any parametric submodel $\{P_{t,g}\}$ with score function g. We call ψ a *gradient*[7] of μ at P. Gradients are not unique; for any measurable $h : \mathscr{D} \to \mathbb{R}$ such that $P[hg] = 0$ for all $g \in \dot{\mathscr{P}}_P$ (we say that h is *orthogonal* to $\dot{\mathscr{P}}_P$), $\psi + h$ is also a gradient.

The definition of pathwise differentiability above can be motivated as follows. Suppose we try to form a distributional Taylor expansion

$$\mu(P_{t,g}) - \mu(P) = (P_{t,g} - P)[\psi] + R_2(P_{t,g}, P), \tag{1.7}$$

where $R_2(P_{t,g}, P)$ is simply the left-hand side minus the first term on the right—nothing is assumed about it yet! Now suppose that we divided both sides by t and took the limit as $t \to 0$. Clearly, the left-hand side converges to $\partial \mu(P_{t,g})/\partial t|_{t=0}$. Assuming that the order of differentiation and integration can be exchanged, the first term on the right converges to

$$\frac{\partial}{\partial t} P_{t,g}(\psi)|_{t=0} = \int \psi \frac{\partial}{\partial t} dP_{t,g}|_{t=0} = \int \psi \left(\frac{\partial}{\partial t} \log dP_{t,g} \right) dP_{t,g}|_{t=0} = P[\psi g]. \tag{1.8}$$

By the definition (1.6), we deduce that $\lim_{t \to 0} R_2(P_{t,g}, P)/t = 0$.

This suggests that a gradient can be viewed as a type of first-order distributional derivative. Furthermore, the remainder term $R_2(P_{t,g}, P)$ must depend on the difference between $P_{t,g}$ and P in some higher-order way that allows it to vanish at a faster than linear rate. This will be very important when we discuss double robustness later on.

From the preceding definitions, we can deduce that the constant estimator $\hat{\mu} = \mu(P)$ is not regular. Indeed,

$$\sqrt{n}(\mu(P) - \mu(P_{1/\sqrt{n},g})) \to -\frac{\partial \mu(P_{t,g})}{\partial t}|_{t=0} = -P[\psi g],$$

[6] Since the target quantity

[7] The literature often uses the name "influence function", due to the connection with RAL estimators discussed later. However, since this definition is not quite equivalent to that of asymptotically linear estimators, we find the terminology potentially misleading and prefer to use "gradient" instead (used in, for example, van der Laan and Robins (2003)).

where the right-hand side is a constant that depends on g and is non-zero for at least some directions[8]. Thus, the constant estimator exhibits a type of local asymptotic bias. This will be the case for any estimator that relies on more information than is assumed in the model.

As stated earlier, we wish to characterize the influence functions of asymptotically linear estimators. This is achieved with the following theorem:

Theorem 1.1. *Suppose that $\hat{\mu}$ is an asymptotically linear estimator with influence function ψ, and μ is pathwise differentiable at P with respect to $\dot{\mathscr{P}}_P$. Then $\hat{\mu}$ is regular if and only if ψ is a gradient of μ at P.*

A formal proof is given in §A.1, but the interpretation is obscured by the technical details. Considering the importance of this result, we believe that it is instructive to provide a more informal account that aids intuition.

Let $\{P_{t,g}\}$ be a parametric submodel with score function g. We can write

$$\sqrt{n}(\hat{\mu} - \mu(P_{1/\sqrt{n},g})) = \underbrace{\sqrt{n}(\hat{\mu} - \mu(P))}_{①} - \underbrace{\sqrt{n}(\mu(P_{1/\sqrt{n},g}) - \mu(P))}_{②}.$$

It is immediate that

$$② \to \frac{\partial \mu(P_{t,g})}{\partial t}|_{t=0}$$

as $n \to \infty$. Under P, we know that term ① converges in distribution to $\mathscr{N}(0, P\psi^2)$. If we switch to the sequence $\{P_{1/\sqrt{n},g}\}$, it can be shown that

$$① \overset{P_{1/\sqrt{n},g}}{\rightsquigarrow} \mathscr{N}(0, P\psi^2) + \frac{\partial}{\partial t} P_{t,g}(\psi)|_{t=0},$$

i.e. we get an extra derivative term. The expectation of ψ under P is assumed to be exactly 0, but this is generally not the case under $P_{1/\sqrt{n},g}$. Thus, we might expect there to be an extra "drift" factor $\sqrt{n}P_{1/\sqrt{n},g}(\psi)$, which converges to the derivative term above. But we already know from (1.8) that

$$\frac{\partial}{\partial t} P_{t,g}(\psi)|_{t=0} = P[\psi g].$$

The estimator $\hat{\mu}$ is regular if and only if the extra terms vanish; that is,

$$\frac{\partial \mu(P_{t,g})}{\partial t}|_{t=0} = P[\psi g].$$

This is precisely the definition of a gradient introduced earlier.

[8] Unless μ is constant in all directions, i.e. $\mu(P)$ is already known to be the truth!

The above sketch yields the interpretation that if the data-generating distribution is perturbed slightly from P, the influence function of a RAL estimator will always follow the target quantity by drifting in the same direction.

1.1.2.2 Improving estimators by using estimated weights

Let us return to our motivating problem. Recall that we observe N i.i.d. copies of $D = (X, R, RY)$ from a distribution P known to belong to a set $\mathscr{P}$ of probability measures on a measurable space $(\mathscr{D}, \mathscr{A})$, where R and Y are independent given X. The density of a single observation takes the form

$$p_{Y|X}(y \mid x)^r p_{R|X}(r \mid x) p_X(x),$$

where $p_{Y|X}(y \mid x)$ and $p_X(x)$ are completely unspecified, but $p_{R|X}(r \mid x)$ is known to take the form $\pi(x)^r(1-\pi(x))^{1-r}$. The target quantity is

$$\mu(P) = \int y p_{Y|X}(y \mid x) p_X(x)\, dydx. \tag{1.9}$$

In §A.2, we show that the set of all mean-zero gradients is

$$\left\{ \psi_c(D) = \frac{R(Y - c(X))}{\pi(X)} + c(X) - \mu(P) \right\}, \tag{1.10}$$

where $c(x)$ ranges over all one-dimensional measurable functions of x. Theorem 1.1 implies that the influence function of any RAL estimator must belong to this set. In particular, $c(x) \equiv 0$ and $c(x) \equiv \mu(P)$ correspond to the influence functions of the Horvitz-Thompson and Hájek estimators respectively.

Remarkably, both estimators—and indeed, any estimator that solves an estimating equation of the form

$$\mathbb{P}_n \left(\frac{R(Y - c(X))}{\pi(X)} + c(X) - \mu \right) = 0 \tag{1.11}$$

for arbitrary $c(x)$–can be improved by replacing the known function π with a maximum likelihood estimate. To be more specific, suppose that we have specified a smooth parametric model $\{\pi(X;\alpha)\}$ such that $\pi(X) \equiv \pi(X;\alpha_0)$ for some parameter value α_0. The likelihood contribution of α is

$$\prod_{i=1}^{n} \pi(X_i;\alpha)^{R_i} (1 - \pi(X_i;\alpha))^{1-R_i}$$

and the maximum likelihood estimator $\hat{\alpha}$ can be found by solving the score equation

$$\frac{1}{n}\sum_{i=1}^{n} S_\alpha(R_i, X_i; \alpha) = \frac{1}{n}\sum_{i=1}^{n} \frac{R_i - \pi(X_i;\alpha)}{\pi(X_i;\alpha)(1-\pi(X_i;\alpha))}\dot{\pi}(X_i;\alpha) = 0,$$

where $\dot{\pi}$ is the derivative of $\pi(X_i;\alpha)$ with respect to α. After replacing $\pi(X)$ with $\pi(X;\hat{\alpha})$, the Horvitz-Thompson and Hájek estimators remain consistent and asymptotically normal but have asymptotic variances that are less than or equal to before. This phenomenon has been referred to as a paradox (Henmi and Eguchi, 2004) since the additional randomness from estimating α has increased efficiency rather than the other way round.

Before we provide a mathematical justification, let us illustrate this "paradox" with an example. We revisit the circus owner from the previous section, who once again wishes to estimate the average weight of his 50 elephants. This time, he is able to weigh more than one! Ten of his elephants are African elephants, and the remainder are all Asian elephants. Knowing that African elephants are generally much heavier, the owner decides to stratify by species—the covariate X—to increase efficiency. As a result, he implements a Poisson sampling design where the African elephants each have selection probability 1/2 and the Asian elephants each have selection probability 1/5. Fifteen elephants are selected with 8 African and 7 Asian.

In this case, the Horvitz-Thompson estimator is equal to

$$\frac{1}{50}(2 * T_{Afr} + 5 * T_{Asn}),$$

where T_{Afr} and T_{Asn} are the total weights of the selected African and Asian elephants respectively. Effectively, we have created a *pseudo-population* of 16 African elephants and 35 Asian elephants by replicating the ones who were selected. Given that the proportion of African elephants is higher in the pseudo-population than in the actual population, it seems likely that the Horvitz-Thompson estimate will be larger than the truth. The Hájek estimator replaces the factor of 1/50 above with 1/51; that is, we divide by the size of the pseudo-population. This might lead to a slight improvement, but the aforementioned problem remains.

Suppose that we estimate the selection probabilities instead. There were 10 African elephants in the population and 8 were selected; thus, we estimate that each African elephant had a probability of 8/10 of being selected—this is the maximum likelihood estimate for the saturated regression model. The corresponding estimate for the Asian elephants is 7/40.

Both the Horvitz-Thompson and Hájek estimators with the estimated probabilities equal

$$\frac{1}{50}\left(\frac{10}{8} * T_{Afr} + \frac{40}{7} * T_{Asn}\right).$$

We have created a pseudo-population with 10 African elephants and 40 Asian elephants: exactly the same as the actual population. We would therefore expect this estimate to be better than the ones using the true selection probabilities. This type of adjustment, where the covariates in the sample are balanced to match the target population, is an example of *poststratification*. We will discuss this further later on.

Let us take a closer look at (1.10). Any influence function ψ_c contained in this set can be written[9] as

$$\psi_c(D) = \underbrace{\frac{R(Y - m(X))}{\pi(X)}}_{①} + \underbrace{\left(\frac{R}{\pi(X)} - 1\right)(m(X) - c(X))}_{②} + \underbrace{\{m(X) - \mu(P)\}}_{③}, \tag{1.12}$$

where $m(X)$ is the outcome regression function $\mathbb{E}_P[Y \mid X]$. It is straightforward to verify that the covariance between any pair of the three terms is 0. Thus,

$$\text{var}(\psi_c) = \text{var}(①) + \text{var}(②) + \text{var}(③). \tag{1.13}$$

The variances of terms ① and ③ are exactly the "unexplained" and "explained" variances of Y given X respectively. By the law of total variance, they sum to $\text{var}(Y)$:

$$\text{var}(Y) = \mathbb{E}_P[\text{var}(Y \mid X)] + \text{var}(\mathbb{E}_P[Y \mid X]) = \text{var}(①) + \text{var}(③).$$

Term ② is the most interesting one in our context. It is precisely the sampling error discussed in the example. If $c \equiv 0$ (corresponding to the Horvitz-Thompson estimator) and $x = \{\text{African elephant}\}$, the expression

$$\sum_{i:X_i = x}\left(\frac{R_i}{\pi(x)} - 1\right) m(x)$$

is equal to the difference in the number of African elephants in the pseudo-population and the actual population, multiplied by the average weight of an African elephant. The variance of ② is the component of $\text{var}(\psi_c)$ that we can reduce by estimating π.

[9] We have just added and subtracted $(R/\pi(X) - 1)m(X)$.

In general, replacing the true value of a nuisance parameter with a maximum likelihood estimate has the (asymptotic) effect of removing the variance explained by the score function at the cost of adding variance from enlarging the model. If the initial estimator is already asymptotically efficient in the model where the nuisance parameter is known, then no nuisance score can cut the variance further—we discuss asymptotically efficient estimators in the next subsection. The cost of enlarging the model is perhaps what makes this paradox surprising. The crucial ingredient in our problem is the fact that μ of (1.9) does not depend on π—the circus owner's choice of sampling mechanism has no bearing on the weight of his elephants! As a result, μ does not vary in the directions that we have expanded the model, and no asymptotic variance is added.

In §A.3, we verify that estimating π by maximum likelihood turns an estimator that solves (1.11) with influence function ψ_c into an estimator with influence function

$$\psi_c^*(D) = \psi_c(D) - \mathbb{E}_P[\psi_c(D) \mid S_\alpha(R, X, \alpha_0)],$$

and therefore,

$$\operatorname{var}(\psi_c^*) = \mathbb{E}_P[\operatorname{var}\{\psi_c(D) \mid S_\alpha(R, X, \alpha_0)\}] \leqslant \operatorname{var}(\psi_c)$$

by the law of total variance. As suggested above, the variance of ψ_c^* is the variance of ψ_c that is unexplained by the score function $S_\alpha(R, X, \alpha_0)$. Hence, the larger the model for π, the more we reduce the variance. This is also clear from the geometrical perspective, viewing the conditional expectation as the orthogonal projection of ψ_c onto the linear space spanned by the components of $S_\alpha(R, X, \alpha_0)$.

As a final note, we point out that the Horvitz-Thompson and Hájek estimators can still be computed when X is completely unobserved and $\pi(X)$ is only observed for the individuals with $R = 1$, but this is not the case for any estimator for which π must be estimated. Such a scenario is common in survey settings, and we give particular attention to it in Chapter 3. This might lead one to believe that the gain in efficiency is simply due to leveraging the information in X. However, this phenomenon occurs for any initial estimator that solves an estimating equation of the form (1.11), including ones that rely on X for the whole sample.

1.1.2.3 The efficient influence function and double robustness

Our representation of a potential influence function in (1.12) and corresponding variance decomposition in (1.13) suggest that we should try to find a RAL estimator with $c(x) \equiv m(x)$.

This is because terms ① and ③ are independent of c, and the variance of ② vanishes, so

$$\psi_{\text{eff}}(D) = \frac{R(Y - m(X))}{\pi(X)} + \{m(X) - \mu(P)\} \tag{1.14}$$

must have the smallest variance[10] within the set (1.10). We call ψ_{eff} the *efficient influence function*. Theorem 1.1 implies that any estimator that is asymptotically linear with influence function ψ_{eff} must have the smallest asymptotic variance at P within the class of RAL estimators. Before we can conclude that this is something worth caring about, we should revisit the restrictions on this class.

First, we have asymptotic linearity. We introduced this restriction because estimators with this property are asymptotically normal. But it turns out that $\mathcal{N}(0, P[\psi_{\text{eff}}^2])$ is not only the best possible limiting distribution for RAL estimators, but for regular estimators in general. The convolution theorem (Theorem 25.20, pg. 366 of van der Vaart (1998)) states that, provided the tangent space $\dot{\mathcal{P}}_P$ is a convex cone[11], any limiting distribution L of a regular estimator is the convolution of $\mathcal{N}(0, P[\psi_{\text{eff}}^2])$ with some probability measure M. In other words, L can be represented by the sum of a $\mathcal{N}(0, P[\psi_{\text{eff}}^2])$ variable and some independent noise. Thus, any regular estimator that attains $\mathcal{N}(0, P[\psi_{\text{eff}}^2])$ exactly is necessarily optimal. Furthermore, Lemma 25.23 of van der Vaart (1998) implies that a regular estimator has limiting distribution $\mathcal{N}(0, P[\psi_{\text{eff}}^2])$ if and only if it is asymptotically linear with influence function ψ_{eff}. Hence, there is no loss in generality by imposing asymptotic linearity in the context of best regular estimators. Following convention (p.367, van der Vaart, 1998), we will say that an estimator is *asymptotically efficient*, if it is regular and attains $\mathcal{N}(0, P[\psi_{\text{eff}}^2])$ asymptotically.

This leaves the restriction of regularity. Recall that the asymptotic behaviour of regular estimators is robust to small changes in the data-generating distribution. As mentioned before, this rules out estimators that rely on more information than is contained in the model. But it also excludes certain types of shrinkage estimators. Famously, the James-Stein estimator (Stein, 1956), which is not regular (pg. 119, van der Vaart (1998)), uniformly outperforms the sample mean with respect to mean squared error when estimating the mean of a multivariate normal distribution in 3 dimensions or higher. It achieves this by shrinking the sample mean

[10] Interestingly, its variance is equal to $\text{var}(Y)$, so an estimator that is asymptotically linear with influence function ψ_{eff} has the same asymptotic variance as the sample mean of Y with complete data.

[11] Recall that the tangent space is the set of score functions, or "directions", permitted by the model for paths that pass through P. The tangent space is a cone if it satisfies: $g \in \dot{\mathcal{P}}_P, a \geqslant 0 \implies ag \in \dot{\mathcal{P}}_P$. As stated in pg. 363 of van der Vaart (1998), it is rarely a loss of generality to make this assumption.

towards zero, inducing a favourable bias-variance trade-off. Some authors (e.g. Efron and Morris, 1973) have motivated this approach using empirical Bayes.

Nevertheless, estimators that attain the $\mathcal{N}(0, P[\psi_{\text{eff}}^2])$ limiting distribution do enjoy a certain type of optimality that does not depend on regularity.

Theorem 1.2. *Let $l : \mathbb{R} \to [0, \infty)$ be a subconvex*[12] *loss function. If $\dot{\mathscr{P}}_P$ is a convex cone, then any estimator $\hat{\mu}$ satisfies*

$$\sup_I \liminf_{n\to\infty} \sup_{g\in I} \mathbb{E}_{P_{1/\sqrt{n},g}} l(\sqrt{n}\{\hat{\mu} - \mu(P_{1/\sqrt{n},g})\}) \geqslant \int l \, d\mathcal{N}(0, P[\psi_{\mathit{eff}}^2]). \tag{1.15}$$

The first supremum is taken over all finite subsets of $\dot{\mathscr{P}}_P$.

This is the *local asymptotic minimax theorem* (van der Vaart, 1992). Much like the definition of regular estimators (1.5), the preceding display (1.15) concerns neighbourhoods of P that contract at rate $1/\sqrt{n}$. Roughly speaking, the worst-case risk over any such sequence of neighbourhoods is asymptotically lower-bounded by the risk of a $\mathcal{N}(0, P[\psi_{\text{eff}}^2])$ variable.

Historically, the concepts of "regular" and "local asymptotic minimax" were developed in the context of parametric models to salvage the theory of maximum likelihood after so-called "superefficient" estimators were discovered—the James-Stein estimator mentioned earlier is an example. With this in mind, the limitations of semiparametric efficiency theory may appear to be analogous to those of maximum likelihood. Shrinkage estimation is ubiquitous in modern applications due to the phenomenon of "big data" and the excellent empirical performance of data-adaptive methods, whereas maximum likelihood estimates are prone to overfitting. What we will see later, however, is that semiparametric theory has another utility that actually makes it particularly useful for modern statistical problems.

Let us return to the efficient influence function (1.14). In order to use this to construct an estimator, we must replace the unknown outcome regression function $m(x)$ with an estimate $\hat{m}(x)$. Using the estimating equation approach, we obtain the estimator

$$\hat{\mu}_{\text{eff}} = \mathbb{P}_n \left(\frac{R(Y - \hat{m}(X))}{\pi(X)} + \hat{m}(X) \right).$$

Recall the difference estimator (1.1) introduced earlier in the context of design-based estimation; both estimators take the same form! What was originally an ad-hoc method for incorporating auxiliary variable regression modelling into a design-consistent estimator can

[12] A function is *bowl-shaped* if the sublevel sets $\{x : l(x) \leqslant c\}$ are convex and symmetric about the origin. It is called *subconvex* if these sets are closed.

in fact be motivated by semiparametric efficiency theory. We will defer discussing conditions on $\hat{m}$ to the next section.

So far, we have assumed that the propensity score $\pi(x)$ is known. But outside of survey estimation and randomized experiments, this is an unreasonable assumption. Suppose now that we work with the nonparametric model where π is unknown and completely unspecified aside from positivity. This nonparametric model contains the previous semiparametric model where π was assumed to be known. Thus, the new tangent space at P must contain the old one, and any gradient for the nonparametric model is also a gradient for the semiparametric model. In particular, any mean-zero gradient must be contained in (1.10).

But as we observed earlier, the target quantity μ does not depend on π, i.e. if we perturb P in a direction where π changes but all else is kept fixed, then μ does not vary. So by definition (1.6), any gradient for the nonparametric model is orthogonal to all possible scores for π. In §A.4, we show that ψ_{eff} is the only element of (1.10) that satisfies this, and that it is indeed a gradient. Therefore, ψ_{eff} is the unique mean-zero gradient in the nonparametric model and remains the efficient influence function.

In the previous subsection, we discussed using a maximum likelihood estimate of π to increase efficiency. If we have an estimator that is asymptotically linear with influence function ψ_{eff}, then no further efficiency can be gained this way because ψ_{eff} is already orthogonal to all scores for π. Heuristically, we could interpret ψ_{eff} as the limiting influence function from estimating π using larger and larger models. In simple cases with finite discrete X, like the elephants example, we could attain ψ_{eff} by using the (fully nonparametric) saturated regression model.

The fact that ψ_{eff} remains the efficient influence function implies that the problem of estimating μ has not become harder in the sense of the efficiency bound theorems described earlier. Of course, the problem is certainly harder in practical terms because π must now be estimated without a guarantee of consistency.

Let

$$\hat{\psi}_{\text{eff}}(D) = \frac{R(Y - \hat{m}(X))}{\hat{\pi}(X)} + \hat{m}(X) - \mu(P),$$

where we have replaced the true π and m in (1.14) with estimates $\hat{\pi}$ and $\hat{m}$ respectively. This yields the estimator

$$\hat{\mu}_{\text{DR}} = \mathbb{P}_n \left(\frac{R(Y - \hat{m}(X))}{\hat{\pi}(X)} + \hat{m}(X) \right) = \mathbb{P}_n[\hat{\psi}_{\text{eff}}] + \mu(P). \tag{1.16}$$

Until recently, it has been routine to estimate both π and m with estimating equations (especially maximum likelihood) derived from working models, yielding $\sqrt{n}$-consistent and asymptotically normal estimators under correct specification. Robins et al. (2000) pointed out that $\hat{\mu}_{\text{DR}}$ is *doubly robust*; that is, $\hat{\mu}_{\text{DR}}$ is consistent as long as at least one of the two working models is correctly specified. If both are correctly specified, then $\mathbb{P}_n[\hat{\psi}_{\text{eff}}] - \mathbb{P}_n[\psi_{\text{eff}}] = o_P(n^{-1/2})$, such that

$$\sqrt{n}(\hat{\mu}_{\text{DR}} - \mu(P)) = \sqrt{n}\mathbb{P}_n[\psi_{\text{eff}}] + o_P(1).$$

We say that $\hat{\mu}_{\text{DR}}$ is *locally efficient* at any distribution P that is contained in the intersection of both working models.

A recent review of the estimating equation approach can be found in Rotnitzky and Vansteelandt (2014), and comprehensive treatments can be found in van der Laan and Robins (2003) and Tsiatis (2006). We give a more general overview of double-robustness in the next section.

1.1.3 Towards data-adaptive estimation

Let us take a closer look at $\hat{\mu}_{\text{DR}}$ of (1.16) with the following decomposition:

$$\begin{aligned}
\sqrt{n}(\hat{\mu}_{\text{DR}} - \mu(P)) &= \sqrt{n}\mathbb{P}_n[\hat{\psi}_{\text{eff}}] && (1.17)\\
&= \underbrace{\sqrt{n}\mathbb{P}_n[\psi_{\text{eff}}]}_{①} + \underbrace{\sqrt{n}P[\hat{\psi}_{\text{eff}}]}_{②} + \underbrace{\sqrt{n}(\mathbb{P}_n - P)[\hat{\psi}_{\text{eff}} - \psi_{\text{eff}}]}_{③}. && (1.18)
\end{aligned}$$

The first equality is true by definition, and the second equality follows simply from adding and subtracting terms ① and ②. Our motivation for constructing $\hat{\psi}_{\text{eff}}$ was to target term ①, which converges to the optimal limiting distribution $\mathcal{N}(0, P[\psi^2_{\text{eff}}])$. If terms ② and ③ converge to 0 in probability as $n \to \infty$, then $\hat{\mu}_{\text{DR}}$ is asymptotically efficient.

We can give an explicit expression for ② and upper-bound it using the Cauchy-Schwarz inequality:

$$\sqrt{n}P[\hat{\psi}_{\text{eff}}] = \sqrt{n}P\left[\pi\left(\frac{1}{\hat{\pi}} - \frac{1}{\pi}\right)(m - \hat{m})\right] \leqslant \sqrt{n}\left\|\frac{1}{\hat{\pi}} - \frac{1}{\pi}\right\| \|m - \hat{m}\|. \tag{1.19}$$

The cross-term structure provides a form of double robustness[13]. As long as the combined rate of convergence of $\hat{\pi}$ and $\hat{m}$ exceeds $\sqrt{n}$, term ② will converge to 0 in probability. This can be achieved, for example, if $\|\hat{\pi}^{-1} - \pi^{-1}\| = O_P(n^{-1/2})$ (e.g. using logistic regression) and $\|\hat{m} - m\| = o_P(1)$.

Alternatively, $\hat{\pi}$ and $\hat{m}$ could both be estimated using flexible data-adaptive techniques. This is particularly attractive in high-dimensional settings where the smoothness conditions imposed by estimating equation approaches become difficult to justify. Convergence rates of $o_P(n^{-1/4})$ are attainable by many popular machine learning methods under relatively mild assumptions on, for example, sparsity or number of derivatives of π and m (see Chernozhukov et al. (2018), particularly the discussion in §3.2).

Term ③ is often controlled using Donsker conditions derived from empirical process theory (van der Vaart, 1998, van der Laan and Robins, 2003). But this may be overly restrictive for high-dimensional settings and machine learning methods (Chernozhukov et al., 2018). A simple way to avoid complexity conditions is through *sample splitting*. We separate our dataset into a "training" sample—used to estimate $\hat{\pi}$ and $\hat{m}$—and a "validation" sample—used to construct $\hat{\mu}_{DR}$. The efficiency lost by reducing our sample size is recovered by swapping the roles of the samples to construct another estimator and taking an average; this step is called *cross-fitting*. Intuitively, this helps us to avoid potential overfitting from plugging in nuisance parameter estimates derived from the same dataset (cf. the Bayesian "sin" of using the data twice).

For simplicity, assume that n is even. Let $\hat{m}^{(1)}$ and $\hat{\pi}^{(1)}$ be estimated from $D_{(n/2)+1}, \ldots, D_n$. Using the remaining data points, we construct

$$\hat{\mu}_{\mathrm{DR}}^{(1)} = \frac{2}{n} \sum_{i=1}^{n/2} \left(\frac{R_i(Y_i - \hat{m}^{(1)}(X_i))}{\hat{\pi}^{(1)}(X_i)} + \hat{m}^{(1)}(X_i) \right).$$

The estimator $\hat{\mu}_{\mathrm{DR}}^{(2)}$ is similarly constructed by swapping the two halves of the dataset. Finally, we set $\check{\mu}_{\mathrm{DR}} = (\hat{\mu}_{\mathrm{DR}}^{(1)} + \hat{\mu}_{\mathrm{DR}}^{(2)})/2$. In §A.5, we show that the conditions

$$\sqrt{n} \left\| \frac{1}{\hat{\pi}} - \frac{1}{\pi} \right\| \|m - \hat{m}\| = o_P(1)$$
$$\|\hat{\psi}_{\mathrm{eff}} - \psi_{\mathrm{eff}}\| = o_P(1)$$

[13] This doesn't quite capture the original definition where one of the estimators is allowed to be inconsistent, so that the distance to the truth is $O_P(1)$. There are some subtleties here depending on the estimation method. We again refer to the articles cited at the end of the previous subsection.

are sufficient for $\check{\mu}_{\mathrm{DR}}$ to be asymptotically efficient. The variance of $\check{\mu}_{\mathrm{DR}}$ can be estimated using sandwich estimation, with each "layer" of the sandwich estimated by averaging across the splits. In this particular case, the estimator takes the simple form $\check{\sigma}^2 = (\hat{\sigma}^2_{(1)} + \hat{\sigma}^2_{(2)})/2$, where

$$\hat{\sigma}^2_{(1)} = \frac{2}{n}\sum_{i=1}^{n/2}\left(\frac{R_i(Y_i - \hat{m}^{(1)}(X_i))}{\hat{\pi}^{(1)}(X_i)} + \hat{m}^{(1)}(X_i) - \hat{\mu}^{(1)}_{\mathrm{DR}}\right)^2,$$

and $\hat{\sigma}^2_{(2)}$ is defined correspondingly. Wald intervals constructed with $\check{\sigma}^2$ are uniformly valid asymptotically (Chernozhukov et al., 2018).

The above can be generalized to more than two folds of the dataset. In particular, we may believe that estimating the functions m and π is far more challenging than estimating the one-dimensional (or in general, finite-dimensional) μ, in which case we could use proportion $(1 - K^{-1})$ of the dataset as the training sample for a whole number $K > 2$, and the remaining proportion K^{-1} for validation. We cycle through K folds of the dataset and average the resulting estimators. There is no difference in the asymptotic sense, but we may obtain finite-sample performance gains.

It is not a coincidence that estimators based on ψ_{eff} enable flexible estimation of the nuisance parameters. Recall that in §1.1.2.1, we discussed how gradients could be viewed as first-order derivatives of the target quantity in a distributional Taylor expansion. For given estimators $\hat{m}$ and $\hat{\pi}$, let $\hat{P}$ be a distribution in $\mathscr{P}$ with outcome regression function $\hat{m}$ and propensity score $\hat{\pi}$. Then the efficient influence function for μ at $\hat{P}$ must be

$$\begin{aligned}\psi_{\mathrm{eff},\hat{P}}(D) &= \frac{R(Y - \hat{m}(X))}{\hat{\pi}(X)} + \hat{m}(X) - \mu(\hat{P}) \\ &= \hat{\psi}_{\mathrm{eff}}(Z) + \mu(P) - \mu(\hat{P}),\end{aligned}$$

which satisfies $\hat{P}[\psi_{\mathrm{eff},\hat{P}}] = 0$. By expanding $\mu(P)$ around $\mu(\hat{P})$ in the sense of (1.7), we obtain

$$\begin{aligned}\mu(P) - \mu(\hat{P}) &= (P - \hat{P})[\psi_{\mathrm{eff},\hat{P}}] + R_2(P,\hat{P}) \\ &= P[\hat{\psi}_{\mathrm{eff}}] + \mu(P) - \mu(\hat{P}) + R_2(P,\hat{P}).\end{aligned}$$

Hence, term ② in the decomposition (1.18) is in fact equal to $-\sqrt{n}R_2(P,\hat{P})$. Since this is true for any such $\hat{P}$, we can see why (1.19) only depends on $\hat{m}$, $\hat{\pi}$ and P. As discussed earlier, we would a priori expect $R_2(P,\hat{P})$ to vanish with respect to the difference between P and $\hat{P}$ in a higher-order way. Indeed, we saw in (1.19) that it takes the form of a second-order cross-term product. It is special, however, that terms involving "$\|\hat{m} - m\|^2$" and "$\|\hat{\pi} - \pi\|^2$"

do not appear, which leads to double robustness. When π is known and $\hat{\pi} \equiv \pi$, the remainder term vanishes and $\hat{m}$ is permitted to converge at an arbitrarily slow rate.

On the other hand, a naïve regression estimator $\mathbb{P}_n[\hat{m}]$ will generally have *first-order bias* involving $\|\hat{m} - m\|$—as opposed to the *second-order bias* $R_2(P, \hat{P})$—and will fail to be $\sqrt{n}$-consistent when $\hat{m}$ is flexibly estimated. Worse yet, the limiting behaviour of such an estimator may be complex and/or poorly understood, which makes it difficult to perform inference. The analogous situation of Horvitz-Thompson or Hájek estimators with flexibly estimated $\hat{\pi}$ is similarly problematic.

This suggests that the use of $\hat{\psi}_{\text{eff}}$ is a way of debiasing the initial estimates $\hat{m}$ and $\hat{\pi}$. The method of *targeted learning* (van der Laan and Rose, 2011, 2018) makes this more explicit. Suppose, for example, that Y is binary, such that μ is known to lie in $[0, 1]$. Let $\hat{m}_{\text{init}}$ and $\hat{\pi}$ be machine learning estimates of m and π respectively. We construct a new estimate of m

$$\hat{m}_{\text{TL}}(X) = \text{expit}\left(\text{logit}\{\hat{m}_{\text{init}}(X)\} + \frac{\hat{\lambda}}{\hat{\pi}(X)}\right),$$

where $\hat{\lambda}$ is the maximum likelihood estimate of the regression coefficient for $1/\hat{\pi}(X)$ with offset $\text{logit}\{\hat{m}_{\text{init}}(X)\}$ using the units with complete data; that is, $\hat{\lambda}$ solves

$$\frac{1}{n}\sum_{i=1}^{n} \frac{R}{\hat{\pi}(X)}\left[Y - \text{expit}\left(\text{logit}\{\hat{m}_{\text{init}}(X)\} + \frac{\lambda}{\hat{\pi}(X)}\right)\right] = 0.$$

As a result,

$$\hat{\mu}_{\text{TL}} \equiv \mathbb{P}_n[\hat{m}_{\text{TL}}(X)] = \mathbb{P}_n\left[\frac{R(Y - \hat{m}_{\text{TL}}(X))}{\hat{\pi}(X)} + \hat{m}_{\text{TL}}(X)\right],$$

which takes the form (1.16). The parameter λ acts to debias the initial estimate $\hat{m}_{\text{init}}$ in the so-called *least favourable direction* $1/\pi$. The estimator $\hat{\mu}_{\text{TL}}$ has the same asymptotic behaviour as other double robust estimators, but has the advantage of being a regression estimator. This guarantees that the estimate will lie in the permitted interval $[0, 1]$, which could lead to improved finite-sample inference. We had previously discussed this issue in §1.1.1 in the context of design-based difference estimators. The solution proposed by Firth and Bennett (1998) that involved adding the covariate $1/\pi$ was similar to the above but differed in that λ and the (initial) outcome regression model were fitted simultaneously.

1.2 The Bayesian paradigm

1.2.1 The Robins-Ritov example

Recall the problem set-up from §1.1.2. We observe independent and identically distributed data $D_1, \ldots, D_n$ from a distribution P on a measurable space $(\mathscr{D}, \mathscr{A})$, where for each i, $D_i = (X_i, R_i, R_i Y_i)$, X_i is a vector of covariates, Y_i is an outcome, and R_i is binary. Strong ignorability and positivity are assumed as before, and we will additionally assume that the outcome is binary and the covariates are known to be uniformly distributed on the space $[0,1]^k$ for large k (i.e. X is very high-dimensional). The likelihood function specializes to the form

$$\mathscr{L}(m, \pi) = \mathscr{L}_1(m)\mathscr{L}_2(\pi), \tag{1.20}$$

where

$$\begin{aligned} m(x) &= \mathbb{E}_P[Y \mid X = x] = P(Y = 1 \mid X = x) \\ \pi(x) &= P(R = 1 \mid X = x) \\ \mathscr{L}_1(m) &= \prod_{i=1}^{n} \left\{ m(X_i)^{Y_i} [1 - m(X_i)]^{1-Y_i} \right\}^{R_i} \\ \mathscr{L}_2(\pi) &= \prod_{j=1}^{n} \pi(X_j)^{R_j} [1 - \pi(X_j)]^{1-R_j}. \end{aligned}$$

The target quantity is $\mu(P) = \mathbb{E}_P[Y] = \int_{[0,1]^k} m(x)\, dx$. For now, we will assume that π is known.

Previously, we discussed how estimators that employ inverse probability weighting could be used to estimate μ. Within this class, we could try to attain efficiency with

$$\hat{\mu}_{\text{eff}} = \mathbb{P}_n \left(\frac{R(Y - \hat{m}(X))}{\pi(X)} + \hat{m}(X) \right),$$

where $\hat{m}$ is an estimator of m. Sample-splitting and cross-fitting techniques can ensure that $\hat{m}$ need only be $L_2(P)$-consistent in order to achieve asymptotic efficiency, enabling the application of flexible machine learning methods. Furthermore, we could guarantee that $\hat{\mu}_{\text{eff}}$ lies between 0 and 1 with a carefully constructed $\hat{m}$ (by the targeted learning approach, for example). From here on, we will use $\check{\mu}_{\text{eff}}$ to refer to an estimator that uses all of the above techniques, i.e. sample-splitting, cross-fitting, flexibly estimated $\hat{m}$, and bounding.

For any $\delta > 0$, $\check{\mu}_{\text{eff}}$ satisfies

$$\sup_{P \in \mathscr{P}} P(|\check{\mu}_{\text{eff}} - \mu(P)| > \delta) \to 0$$

as $n \to \infty$; that is, $\check{\mu}_{\text{eff}}$ is *uniformly consistent*[14]. Moreover, if $C_n(\alpha)$ is the $(1-\alpha)$-Wald interval constructed with $\check{\mu}_{\text{eff}}$ and the cross-fitted sandwich variance estimator $\check{\sigma}^2$ from §1.1.3, then

$$\sup_{P \in \mathscr{P}} |P\{\mu(P) \in C_n(\alpha)\} - (1-\alpha)| \to 0 \tag{1.21}$$

as $n \to \infty$, provided that $\hat{m}$ converges in $L_2(P)$ to some limit—not necessarily the true m—for all $P \in \mathscr{P}$. Robins and Ritov (1997) emphasized the importance of such uniform properties because they guarantee that for any given tolerance level, there exists a minimal sample size at which μ will be sufficiently well-estimated regardless of the true P.

However, Robins and Ritov (1997) proved that any estimator that ignores the known π will not be uniformly consistent. A consequence is that no interval that ignores π can both satisfy (1.21) and also shrink to 0 in expectation as n grows; otherwise, the midpoint of the interval would be a uniformly consistent estimator. A heuristic explanation for this lack of uniformity is as follows. We have not placed any restrictions[15] on m, which permits m to be very badly behaved ("wiggly"). For any given sample size, we can always find an m that is so wiggly that it cannot possibly be adequately estimated from the data. Although $\check{\mu}_{\text{eff}}$ uses estimates of m, it does not rely on these estimates being accurate, much like the design-based model-assisted estimators that preceded it. Any biases resulting from estimating m are corrected by using the known π.

The result only depends on X being continuous—it also holds if X is univariate. The problem can be avoided by making smoothness restrictions on m. But the point of emphasizing that X is high-dimensional is that as the dimension increases, the smoothness requirements to pool the information in the data become increasingly stringent and difficult to justify. As Coombs (1964) states, "we buy information with assumptions"; the sparser the data, the more we have to buy.

Robins and Ritov (1997) argued that the above has dire implications for "strict likelihood" methods, including any Bayesian model that excludes π from the prior for m. This is because the likelihood function $\mathscr{L}$ of (1.20) factors into $\mathscr{L}_1(m)$ and $\mathscr{L}_2(\pi)$, where the latter is constant because π is known; thus, a procedure that obeys the strict likelihood principle

[14] In fact, we could replace $|\check{\mu}_{\text{eff}} - \mu(P)|$ with $n^{0.5-\varepsilon}|\check{\mu}_{\text{eff}} - \mu(P)|$ for any $\varepsilon > 0$.

[15] Aside from the minimal requirements of measurability and being bounded between 0 and 1.

will use $\mathscr{L}_1$ only to estimate μ. Generally, this will lead to inference that ignores π, unless π is "artificially" added to $\mathscr{L}_1$ (Robins et al., 2000).

An example of such an "artificial" construction was already discussed earlier in §1.1.1 in the context of design-based model-assisted estimators. For a logistic regression model, Firth and Bennett (1998) (see also: Scharfstein et al. (1999)) suggested augmenting the regression function with an additional covariate $1/\pi(x)$:

$$m(x;\beta,\lambda) \equiv \text{expit}\left(\sum_{h=1}^{H}\beta_h b_h(x) + \frac{\lambda}{\pi(x)}\right), \tag{1.22}$$

where $\{b_h\}_{h=1}^{H}$ is a prespecified set of basis functions, $\beta = (\beta_1 \ldots, \beta_H)$ is the vector of basis coefficients, and λ is the one-dimensional coefficient of the covariate $1/\pi(x)$. If β and λ are estimated with maximum likelihood, then the resulting estimator of μ will be asymptotically equivalent to $\hat{\mu}_{\text{eff}}$ with $\hat{m}(x) \equiv m(x;\hat{\beta}_{MLE},\hat{\lambda}_{MLE})$ and will therefore be uniformly consistent. Under some additional conditions, this can also be true for estimators derived from a Bayesian model using (1.22) with priors on β and λ (due to the Bernstein-von Mises theorem, possibly under misspecification[16]: Kleijn and van der Vaart (2012)).

Robins and Ritov (1997) and Robins and Wasserman (2012a,b) discuss reasons for why they believe a committed subjective Bayesian would not specify a prior for m that depends on π. Some of these arguments are critiqued by Sims (2012a,b,c,d). While this is an interesting topic philosophically, we believe that the practical implications are limited (see §1.2.2). Therefore, we will omit a discussion of these details.

Regardless, Robins and Wasserman (2012b) point out that specifying a prior for m that depends on π is necessary but not sufficient to achieve desirable frequentist properties like uniform consistency. The model and the prior must be carefully constructed in order to mimic a frequentist estimator, e.g. the construction in (1.22). They conclude that such approaches are examples of *frequentist pursuit* and have no benefits over the original procedures they are based on.

Robins et al. (2015) discuss further issues that arise if π is unknown. If m and π are jointly estimated with a likelihood approach, the resulting estimate of π will depend on the outcome regression model. As a consequence, a misspecified outcome regression model will generally lead to an inconsistent estimate of π, even if the propensity score model is correct. This phenomenon of model "feedback" in a Bayesian setting has been discussed by McCandless et al. (2010), Zigler et al. (2013) and Saarela et al. (2016).

[16] In this case, the credible sets will generally fail to attain nominal coverage asymptotically.

We can be more explicit by adapting the previous example described by (1.22). Suppose we have specified a parametric model $\{\pi(X;\alpha)\}$ for the propensity score, e.g. another logistic model. The outcome regression model becomes

$$m(x;\beta,\lambda,\alpha) \equiv \text{expit}\left(\sum_{h=1}^{H}\beta_h b_h(x) + \frac{\lambda}{\pi(x;\alpha)}\right), \tag{1.23}$$

and the likelihood function factors as follows:

$$\mathscr{L}(m,\pi) \equiv \mathscr{L}_1(\beta,\lambda,\alpha)\mathscr{L}_2(\alpha).$$

Note that α is included in both factors, so the maximum likelihood estimate $\hat{\alpha}_{MLE}$ will converge to the value that minimizes the KL-divergence from the truth to the joint model, rather than just the propensity score model. Thus, if the original, non-augmented outcome regression model[17] is misspecified, then $\pi(x;\hat{\alpha}_{MLE})$ will generally be inconsistent; the maximum likelihood estimate of λ will tend to a non-zero limit, and $\hat{\alpha}_{MLE}$ will be pulled towards a compromise between the two models. This is likely to lead to an inconsistent estimator of μ.

In contrast, we can construct a doubly robust estimator by estimating π and m sequentially. First, $\hat{\alpha}_{MLE}$ is obtained by maximizing $\mathscr{L}_2(\alpha)$ only. Then, we obtain $\hat{\beta}_{MLE}$ and $\hat{\lambda}_{MLE}$ by maximizing $\mathscr{L}_1(\beta,\lambda,\hat{\alpha}_{MLE})$. The resulting estimator of μ will be consistent if the propensity score model is correct, even if the outcome regression model is not. Moreover, we would attain asymptotic efficiency if both models were correct.

As before, the analogous Bayesian approach exhibits similar behaviour to maximum likelihood and is therefore afflicted with the same problem. Zigler et al. (2013) empirically investigated the negative effects of model feedback in a Bayesian setting. There have been several proposals for how a Bayesian could "cut" the feedback and prevent a potentially incorrect outcome regression model from contaminating the propensity score model. McCandless et al. (2010) suggested updating the prior for π with $\mathscr{L}_2(\pi)$ only; this could be implemented with a Gibbs sampler, where π and m are updated in turn. But since there is no proper underlying joint model, the sampler may not converge. The convergence issue has been addressed by Plummer (2015) (and also by Jacob et al. (2017) and Liu and Goudie (2020)). Graham et al. (2016), like the sequential doubly robust method described above, proposed using a preliminary estimate of π derived from the propensity score model alone.

[17] That is, the model defined by (1.23) with $\lambda \equiv 0$.

This is plugged into the likelihood function for m to be used for the updating procedure. The uncertainty in the estimate must be accounted for by making a variance correction.

1.2.2 Discussion

1.2.2.1 Inverse probability weighting vs. poststratification

From a less technical perspective, the Robins-Ritov example can be interpreted as a comparison between inverse probability weighting and poststratification.

To illustrate this, let us return once more to the circus owner from the previous section, who again wishes to estimate the average weight of his 50 elephants. Buoyed by his previous success, he decides to implement another stratified sampling design. This time, he stratifies not only on species, but also age, gender, skin colour and length of tusk. Fourteen elephants are selected and weighed. As he begins to carry out his analysis, the owner realises with horror that his previous estimation strategy will not generalize. Last time, he calculated the totals for the two different species and computed a weighted average. The weights were chosen to calibrate the balance of the sample covariates to the population. However, two of his covariates—age and length of tusk—are continuous; it is impossible to compute the totals within different levels. And even if he were to discretize those two covariates, he notices that there are many combinations of covariates present in the population that are not represented in the sample. It seems that his only choice is to enforce substantial dimension reductions and smoothing in order to pool the limited information contained in his sample. The owner despairs at the fact that his estimate will likely be heavily biased, and his interval, which will be too narrow, will likely fail to cover the truth.

At this point, the former circus statistician—still unemployed—strides in and calls out "The Horvitz-Thompson estimator is still unbiased!". Upon realising that his experiment can be saved, the circus owner jumps with joy and the pair share a tearful embrace. The statistician gets his job back and forges a long and successful career working with the circus.

The moral of the story is that while poststratification is straightforward and efficient in simple settings, it can become unwieldy when more covariates are involved. If heavy smoothing is required to handle continuous covariates and to make up for insufficient data, then the inference might be unreliable. This is the essence of the Robins-Ritov-Wasserman argument. Meanwhile, inverse probability weighted estimators with known weights remain unbiased (or approximately/asymptotically unbiased).

Bayesian inference—in the standard set-up considered by Robins and Ritov (1997)—is inherently a poststratification-type method because it conditions on all of the data[18]. Once the covariates are observed, they must be adjusted for in the analysis.

1.2.2.2 Why Bayes?

Before we discuss and develop potential Bayesian approaches to this problem, it is important to ascertain reasons for why a Bayesian approach might be desirable here. The arguments in Robins and Ritov (1997) and Robins and Wasserman (2012a) concern a holistic subjective Bayes viewpoint, from which the hypothetical user specifies a prior that perfectly encapsulates their subjective beliefs. But this is unreasonable in practice, especially in the high-dimensional setting in which the example is based. In any case, this philosophy excludes any practicing Bayesian who would consider methods such as nonparametric Bayes, default priors for nuisance parameters, and sample-size-dependent model choices. We subscribe to the notion presented by Gelman and Shalizi (2013), who state: "In practice, the various parts of the model have functional forms picked by a mix of substantive knowledge, scientific conjectures, statistical properties, analytical convenience, disciplinary tradition and computational tractability."

If we acknowledge that our prior probabilities are not purely epistemic, then we must address how they should be interpreted. In this thesis, we choose to view the prior as a regularization device, and we will evaluate the performance of procedures from a frequentist standpoint. In this respect, our philosophy aligns with the "Calibrated Bayes" perspective of Rubin (1984) and Little (2011). A calibrated Bayesian procedure offers the potential for improved small-sample inference via the prior but allows the data to dominate as the sample size increases, yielding valid frequentist inference asymptotically. This does not undermine the importance of subject-matter knowledge and expertise, which are still crucial for deciding how the priors are specified.

The reduced dependence on asymptotics is important because it can be difficult to determine whether asymptotic approximations are reasonable for a given dataset. This is especially pertinent when inverse probability weighting is involved due to its well-documented risks of poor finite-sample performance. In a setting where π is unknown, Kang and Schafer (2007) demonstrated empirically that doubly robust estimators can perform far worse than outcome regression estimators when both working models are misspecified, despite the fact that the data show scant evidence of these misspecifications.

[18] In §1.2.2.4 (and also in Chapters 3-4), we will explore a different perspective for Bayesian inference that is distinct from poststratification.

Problems surrounding inverse probability weighting are not necessarily avoided by using machine learning methods. Compared to estimating equation approaches, the increased flexibility demands larger sample sizes for asymptotics to become relevant. If assumptions of sparsity and linear/logistic models are required to handle high-dimensional covariates, then inverse probability weighting with estimated weights is particularly dangerous and arguably fails to justify the risks involved.

Moreover, the fact that the second-order bias converges to 0 faster than $n^{-1/2}$ provides no guarantees that it is small/negligible in finite samples. Analogously, a computer scientist would be cautious about calling a polynomial-time algorithm "fast". The presence of $1/\hat{\pi}$ in (1.19) suggests that the second-order bias is particularly unforgiving of estimation error in regions of X with few selected individuals, which is disturbing because that is also where estimating m is most difficult.

When the outcomes are bounded, we discussed in previous sections how one could use link functions to force doubly robust estimators within the bounds of the parameter space. But arguably, the benefits are only cosmetic; if a set of inverse probability weights are so extreme that a naïve doubly robust estimator lies outside the parameter space, then a corresponding bounded doubly robust estimator using the same weights cannot be viewed as being reliable. If anything, this makes it more likely that a user inadvertently reports an unreliable estimate without realizing the presence of this issue.

One might ask why a Bayesian approach to regularization should be preferred over others. We again emphasize the importance of subject matter knowledge and expertise, and point out that a prior provides a natural and intuitive way for users to incorporate these beliefs, even if the specification is not entirely subjective. There is also growing empirical evidence to suggest that Bayesian regularization can offer significant performance gains over competitors. For example, Bayesian Additive Regression Trees (BART) (Chipman et al., 2007, 2010) has been shown to often dominate more classically-minded counterparts like gradient boosting (Friedman, 2001) and random forests (Breiman, 2001) in extensive simulations (e.g. Chipman et al., 2010, Dorie et al., 2019). The exact reasons for this are not yet clear, but one could speculate that averaging over the parameter space is better-suited in practice to complex prediction problems than optimizing loss functions. A discussion in the context of Bayesian deep learning can be found in Wilson (2020).

Averaging can also be beneficial computationally. Marginalization of nuisance parameters can often be far more computationally efficient than profiling/joint optimization. In Chapter 2, we present an example where a profile maximum likelihood procedure using

the EM-algorithm (Dempster et al., 1977) fails to converge appropriately, but our Bayesian procedure works smoothly.

Finally, it is widely accepted that Bayesian inference is effective at solving complex problems involving multiple data sources. For example, the propagation of uncertainty by integration provides a natural framework for evidence synthesis (Ades and Sutton, 2006); repeatedly integrating over the posterior distributions of parameters allows one to propagate uncertainty across multiple data sources within a single analysis. Prior shrinkage also automatically accommodates multiple comparisons (van Zwet and Cator, 2020). This motivates the development of approaches that can be embedded into a larger, encompassing analysis.

1.2.2.3 Pragmatic compromises

Now that we have established our motivations for pursuing a Bayesian analysis, we can investigate whether it is possible to find pragmatic approaches that deviate slightly from the holistic Bayesian framework while still retaining some of the benefits.

Recall that a fully Bayesian analysis can struggle in the Robins-Ritov example because the high-dimensional, continuously distributed covariate vectors must be conditioned upon. We can try to circumvent this issue by only conditioning on a low dimensional summary of the covariates that is sufficient to adjust for the selection bias; effectively, we pretend that X was unobserved aside from such a summary. The propensity score satisfies this condition, i.e.

$$R \perp\!\!\!\perp Y \mid \pi(X), \tag{1.24}$$

which can be shown as follows:

$$P(R=1 \mid Y, \pi(X)) = \mathbb{E}[\underbrace{\mathbb{E}\{R \mid Y, X, \pi(X)\}}_{\pi(X)} \mid Y, \pi(X)] = P(R=1 \mid \pi(X)).$$

The first equality uses the tower property of expectations and strong ignorability, and the second equality is simply due to $\pi(X) = P(R=1 \mid \pi(X))$.

Thus, the target quantity $\mu(P)$ can be expressed as

$$\mu(P) = \mathbb{E}[P(Y=1 \mid \pi(X))] = \mathbb{E}[P(Y=1 \mid \pi(X), R=1)].$$

The problem now takes the same form as before; the only difference is that the high-dimensional X has been replaced by the one-dimensional $\pi(X)$. We can proceed to specify

a model and prior for the conditional probability $P(Y = 1 \mid \pi(X))$, and then compute the posterior using the data from the individuals with $R = 1$. This determines the posterior for μ since the distribution of $\pi(X)$ is known by assumption.

In general, $\pi(X)$ will be a continuous variable, so the Robins-Ritov theorem still applies; that is, a Bayesian estimate of μ derived from conditioning only on $\pi(X)$ will still fail to be uniformly consistent in the fully nonparametric model. But in practice, we can expect a binary regression function with a one-dimensional covariate bounded between 0 and 1 to be well-estimated for moderate sample sizes. By avoiding inverse probability weighting, we should in fact obtain relatively stable estimates.

This approach is free of "frequentist pursuit" since there is no attempt to imitate a frequentist estimator. The user is completely unrestricted with regards to the specification of the model and prior. The only Bayesian "sin" committed—aside from using π, which we have already addressed—is throwing away/ignoring data. Usually, we might be concerned about the loss of information. But this concern is perhaps based on intuition from low-dimensional parametric models. In this situation, we actually stand to learn more from the data—in the sense of deriving a more precise estimate of the target quantity—due to the substantial dimension reductions of the model.

In survey settings where X is unobserved and has unknown distribution, and $\pi(X)$ is only observed for the selected units, the approach outlined above can be adapted with some additional steps (Zanganeh and Little, 2015, Si et al., 2015). But this involves modelling the distribution of $\pi(X)$ for the unselected units, which has no substantive value and is potentially difficult to specify. Moreover, it is unappealing to treat the data provider as an adversary and try to model information that has been withheld. For this setting, weighting methods are far more attractive. In Chapter 3, we develop a method that combines the simplicity of weighting with the benefits of Bayesian modelling.

In the observation setting, where π is unknown, some authors have suggested incorporating an estimate $\hat{\pi}$ into the outcome regression model to improve performance. Ray and van der Vaart (2018) proposed augmenting an outcome regression model for Y given X with the covariate $1/\hat{\pi}(X)$—similar to the approaches of Firth and Bennett (1998), Scharfstein et al. (1999) and targeted learning—with the intention of correcting the first-order bias of the posterior for μ. Hahn et al. (2020) were instead motivated by the intuition that the selection mechanism should provide useful information about the outcome-covariate relationship. For instance, a practitioner may be more likely to assign treatment to patients that are deemed to be relatively vulnerable. Thus, the estimated propensity score has the potential to be a

useful transformation of the covariates that makes it easier to capture complex dependencies between Y and X.

The troubling aspect from a Bayesian perspective is the data-dependent prior. The asymptotic results of Ray and van der Vaart (2018) avoid this problem by assuming that π is estimated from a separate sample. But in practice, this may involve splitting the dataset, which would reduce the sample size for the prior update. Hahn et al. (2020), however, insist that the data-dependent prior is justified because the outcome regression model is conditional on the covariates and selection variables used to estimate π. They compare this with the Zellner g-prior (Zellner, 1986) for linear regression, where the prior covariance of the regression coefficients is estimated from the observed covariates.

This argument is valid for showing that the data is not used twice, but some undesirable practical aspects remain. If we wish to endow our results with a frequentist interpretation, then we must ascertain the appropriate form of hypothetical repetitions of the experiment. By using a superpopulation approach, it is implicit that we are interested in providing results that generalize to individuals outside of the dataset. This is particularly important in causal inference. Thus, it is inappropriate to consider hypothetical repetitions that condition on a particular realization of the covariates and selection[19]. Furthermore, a data-dependent prior violates *coherence*, in the sense of Bissiri et al. (2016): the form of the posterior will depend on the order that we receive the data. For example, if we receive data sequentially, as is common in clinical trials, then we must either change the prior with each new batch, or accept that our inferences would have been different had we received all the data at once. This can complicate sequential decision-making, which is often cited as a strength of Bayesian inference. Given that the estimated propensity score is used as a black box proxy for subject-matter knowledge, it seems preferable to circumvent these problems through prior specification of meaningful transformations of the covariates based on expert elicitation. Some of the issues outlined above partially motivate the developments in Chapter 4.

The idea of ignoring part of the data can be taken to the extreme of replacing the entire dataset by a set of summary statistics. Several authors (Monahan and Boos, 1992, Robins, 2004, Hoff and Wakefield, 2013, Wang et al., 2017) have suggested leveraging the asymptotic normality of estimators to construct approximate likelihoods. If an estimator $\hat{\mu}$ satisfies

$$\sqrt{n}(\hat{\mu} - \mu) \rightsquigarrow \mathcal{N}(0, \sigma^2)$$

[19]Hahn et al. (2020) do indeed use frequentist metrics to evaluate performance in their simulations, and they perform the repetitions unconditionally; that is, the covariates and selection are randomized across iterations.

for unknown asymptotic variance σ^2, then the following "likelihood function" is approximately valid given a sufficiently large sample size:

$$\mathscr{L}(\mu) \propto \exp\left\{-\frac{n(\hat{\mu}-\mu)^2}{2\hat{\sigma}^2}\right\},$$

where $\hat{\sigma}^2$ is a robust variance estimator, e.g. the sandwich estimator. By specifying a prior $p(\mu)$, we can directly construct a "posterior" for μ:

$$p(\mu \mid \hat{\mu}) \propto \mathscr{L}(\mu)p(\mu).$$

The estimator $\hat{\mu}$ plays the role of the data, while $\hat{\sigma}^2$ is treated as known.

The limitations of this approach are clear. We are stuck in a "Catch-22" situation where we must appeal to asymptotic theory to justify the variance and normality approximations, but asymptotic theory also suggests that the prior will only have a higher-order effect on the posterior (Bernstein-von Mises theorems are proved in Robins (2004) and Wang et al. (2017)). In particular, we fail to attain the previously described goal of "Calibrated Bayes" that allows the user to be agnostic about the applicability of asymptotics.

Other approximate likelihoods have been proposed in the literature: approximate pivot likelihood (Boos and Monahan, 1986), empirical likelihood (Owen, 2001), bootstrap likelihoods (Davison et al., 1992) and implied likelihood (Efron, 1993). A review can be found in Efron and Tibshirani (1994). Chapter 3 studies the *exponentially tilted empirical likelihood* (Jing and Wood, 1996, Corcoran, 1998, Lee and Young, 1999), which was given a Bayesian justification by Schennach (2005). We will give further arguments that support its use in Bayesian inference and investigate its asymptotic properties.

1.2.2.4 A projection-based framework for Bayesian inference

In §1.2.1, we described the so-called "feedback" problem, in which joint modelling of the outcome regression function and propensity score can lead to model contamination arising from misspecification of the outcome regression model. This was compared unfavourably to doubly robust methods, which fit the models separately/sequentially. But we now argue why this phenomenon is misleading.

Recall from §1.1.2.3 that doubly robust estimators operate in the fully nonparametric model, where the conditional distributions of Y given X and R given X are minimally specified. The "working models" for m and π—defined explicitly, or by estimating equations/loss functions—are not assumed to be correct because they only cover proper subsets of the

nonparametric model. Of course, this is also implicit from the definition of double robustness. The estimators $\hat{m}$ and $\hat{\pi}$ can therefore be interpreted as estimating the projections of P onto the working models, rather than the "true" parameters. The type of projection depends on the loss function. For example, maximum likelihood corresponds to negative log-likelihood loss and minimizing KL-divergence.

Furthermore, since the doubly robust estimators fail to be consistent for the functional

$$\mu(P) = \mathbb{E}_P[Y] \tag{1.25}$$

across the whole nonparametric model, we can infer that the actual estimand is in fact

$$\mu_{proj}(P) = \mathbb{E}_P\left[\frac{R(Y - m_{proj}(X;P))}{\pi_{proj}(X;P)} + m_{proj}(X;P)\right], \tag{1.26}$$

where $m_{proj}(\cdot;P)$ and $\pi_{proj}(\cdot;P)$ are the projections of P onto the working models for m and π respectively. This new estimand coincides with the original in (1.25) if either $m_{proj} \equiv m$ or $\pi_{proj} \equiv \pi$.

In contrast, the Bayesian joint estimation approach that suffers from feedback conditions on both working models being correct; that is, the intersection of both working models is assigned prior probability 1. Thus, it is no surprise that inference goes awry under misspecification! A more appropriate Bayesian approach for the set-up above is to specify a fully nonparametric Bayesian model, and the nonparametric posterior for P will then induce a posterior for $\mu_{proj}(P)$. This roughly corresponds to our approach in Chapter 3.

The idea of defining target quantities as model-free projections, as opposed to true components of parametric models, can be generalized to handle several long-standing issues in Bayesian inference. One important example is the question of how to deal with model misspecification. If Bayesian updating is carried out using a misspecified model, the posterior will—under some regularity conditions—concentrate on the element in the model that minimises the KL-divergence to the truth (also called the "pseudo-true" value). But the coverage of credible intervals from this posterior will generally fail to converge to nominal levels (Kleijn and van der Vaart, 2012). Another related example is the problem of heteroscedastic errors and non-linearity in linear regression. This is studied in detail from a frequentist viewpoint by Buja et al. (2019a,b), who also advocate the projection-based perspective. Both cases can be resolved to certain extents by redefining the target quantities and using nonparametric modelling. We discuss this further in Chapter 4.

The crucial ingredient in this framework is the choice of nonparametric prior. A natural candidate is the Dirichlet process (Ferguson, 1973), which is weakly supported on the set of all probability measures on the sample space as long as the sample space is equal to the support of the base measure (Ghosh and Ramamoorthi, 2003, Ghosal and van der Vaart, 2017). The Dirichlet process can be made even more convenient by letting the base measure tend to 0. This leads to the so-called *Bayesian bootstrap* posterior (Rubin, 1981) that assigns probability 1 to the set of distributions supported only on the observed data. Posterior samples of P can be efficiently and exactly drawn without MCMC; it suffices to repeatedly sample vectors of uniform Dirichlet weights of length n, which are assigned to the data points as probabilities.

The first proposal to use the Bayesian bootstrap for constructing posteriors of projected estimands was perhaps the *weighted likelihood bootstrap* (Newton and Raftery, 1994). The motivation was simply to provide a computationally efficient approximation to a parametric Bayes posterior. Posterior computation consisted of repeated (Dirichlet) weighted maximum likelihood, which corresponds to finding the value in the parametric model that is closest to the posterior draw of P in terms of KL-divergence. Although the developments of MCMC have rendered these computational benefits redundant, Lyddon et al. (2018) realised the potential of the weighted likelihood bootstrap for handling model misspecification, and extended the approach. The Bayesian bootstrap has also been applied to instrumental variables estimation and quantile regression (Chamberlain and Imbens, 2003), and doubly robust estimation (Saarela et al., 2016).

The glaring deficiency of the Dirichlet process—and by extension, the Bayesian bootstrap—is the inability to directly incorporate a prior on the target quantity. Kessler et al. (2015) proposed a general heuristic for combining nonparametric priors with informative, marginally specified priors on finite-dimensional parameters. But this involves deriving or estimating the original marginal prior induced by the nonparametric prior, such as a Dirichlet process, which appears to be infeasible unless the parameter is very low-dimensional. This is a problem that we address in Chapters 3 and 4.

1.3 Conclusions

1.3.1 Generalizations to other estimands

Our running example—estimating an outcome mean from incomplete data—is closely related to estimating an average treatment effect in causal inference (e.g. Morgan and Winship, 2007).

Suppose we wish to estimate $\mathbb{E}Y^1 - \mathbb{E}Y^0$, where Y^1 and Y^0 denote the *counterfactual* outcome variables for treated and not treated respectively. Furthermore, suppose that $R = 1$ if an individual is assigned to the treatment group, and $R = 0$ if the individual is assigned to the control group. We observe independent and identically distributed data $D_1, \ldots, D_n$, where for each i, $D_i = (X_i, R_i, R_i Y_i^1, (1-R_i)Y_i^0)$, and X_i is a vector of covariates as before. We assume that X is sufficient to adjust for confounding[20]; that is,

$$Y^r \perp\!\!\!\perp R \mid X \quad \text{for } r \in \{0,1\},$$

and the propensity score $\pi(X) = P(R = 1 \mid X)$ is bounded away not only from 0 but also from 1 with probability 1. Under these assumptions, the average treatment effect is identifiable, and we can apply the previously described methods for incomplete data to first estimate $\mathbb{E}Y^1$ from (X, R, RY^1) and again to estimate $\mathbb{E}Y^0$ from $(X, R, (1-R)Y^0)$.

Returning to the incomplete data set-up, we can also generalize the methodology to other estimands besides outcome means. Recall that the Hájek estimator solves

$$\mathbb{P}_n \left[\frac{R(Y-\mu)}{\pi(X)} \right] = 0,$$

where we have weighted the complete data influence function $Y - \mu$ by $R/\pi(X)$. Suppose now that we observe independent and identically distributed data $D_1, \ldots, D_n$, where for each i, $D_i = (X_i, R_i, R_i Z_i)$. The assumptions are similar to before; we have only replaced Y with a more general variable Z, and μ is replaced with a more general estimand γ. For example, $Z = (Y, W)$ and γ is the linear regression coefficient of Y on W. Suppose also that γ can be estimated by solving

$$\mathbb{P}_n[u(Z;\gamma)] = 0$$

given complete data, where $P[u(Z;\gamma(P))] = 0$, e.g. $u(Z;\gamma) = W^{\mathrm{T}}(Y - W\gamma)$ for ordinary least squares estimation. It is then clear that the Hájek-style estimating equation

$$\mathbb{P}_n \left[\frac{Ru(Z;\gamma)}{\pi(X)} \right] = 0.$$

is unbiased and can be used to estimate γ for incomplete data if π is known.

The estimating equation above belongs to the class described by

$$\mathbb{P}_n \left[\frac{Ru(Z;\gamma)}{\pi(X)} - \phi(X,\gamma) \left\{ \frac{R}{\pi(X)} - 1 \right\} \right] = 0,$$

[20] Also referred to as "no unmeasured confounding" and "conditional exchangeability".

where $\phi(X, \gamma)$ is an arbitrary measurable function of X and γ. The most efficient member of this class is obtained by setting $\phi \equiv \phi_{\text{eff}}(Z, \gamma) \equiv \mathbb{E}_P[u(Z;\gamma) \mid X]$ (Tsiatis, 2006). It is straightforward to verify that the efficient influence function (1.14) for estimating the outcome mean can be recovered by replacing $u(Z;\gamma)$ with $Y - \mu$. However, while $Y - \mu$ is the unique complete-data influence function for μ (see §A.2), there are generally multiple complete-data influence functions for an estimand γ. For instance, the influence function for the linear regression coefficient is not unique; we can use weighted least squares or robust alternatives. Thus, finding the efficient influence function for γ involves optimizing over u as well as ϕ. It is not necessarily the case that the optimal choice of u is the efficient influence function for complete data (van der Laan and Robins, 2003, Tsiatis, 2006). Moreover, the efficient influence function can be difficult to compute. One could therefore opt for a compromise estimator that is not fully efficient but is relatively easy to implement.

Similar to before, it is likely that π and ϕ_{eff}—for a particular choice of u—are both unknown. If estimating equations are used for both, then the resulting estimator for γ is doubly robust and locally efficient in its class. We will revisit this in Chapter 3.

1.3.2 Thesis outline

Chapters 2-4 form the core of this thesis. In Chapter 2, we develop a Bayesian framework for analyzing case-cohort study data under the Cox model. The case-cohort study design employs an unequal probability sampling frame wherein certain covariates are measured for all cases and a random subset of the controls. Our method is applied to the EPIC-Norfolk cohort study to investigate the associations between saturated fatty acids and incident type-2 diabetes. Chapter 3 studies the use of the Bayesian exponentially tilted empirical likelihood to resolve the issues described in this first chapter. This approach builds on the projection-based perspective described in §1.2.2.4, and we prove asymptotic results to justify its use. Some of the shortcomings of this method are addressed in Chapter 4, where we introduce a new nonparametric Bayesian model called the *exponentially tilted Bayesian bootstrap*. We develop algorithms to sample from the posterior and explore its behaviour across a variety of examples. Finally, Chapter 5 discusses some of the limitations of our work and identifies potential avenues for future research.

The notation defined in this chapter will also be used in Chapters 3 and 4, roughly following the conventions of the missing data and moment condition inference literature. In Chapter 2, we will instead use notation that is more standard in survival analysis in order to facilitate comparisons with papers proposing competing methodology. For example, we have

used R to denote the binary selection variable in this chapter, but R is more commonly used to denote the at-risk indicator in survival analysis. Further details of these differences are discussed in §2.2.1.

Chapter 2

A Bayesian framework for case-cohort Cox regression

2.1 Introduction

This chapter develops methodology for the case-cohort study design (Prentice, 1986), which is an increasingly common approach for studying prospective epidemiological associations. Time and cost constraints, as well as concerns over the wastage of valuable biological material (Borgan and Samuelson, 2017), can render it infeasible to obtain certain covariates on a full cohort. The case-cohort design circumvents this issue by requiring complete covariate measurements on only a randomly sampled subcohort along with all remaining incident cases, allowing one to efficiently target the quantities of interest while retaining identifiability. An advantage over the similarly motivated nested case-control design (Thomas, 1977) is the ability to reuse the subcohort for multiple endpoints (Kulathinal and Arjas, 2006).

Existing proposals for analysing case-cohort data are mostly based on the Cox proportional hazards model (Cox, 1972), although other models have been considered (e.g. Lu and Tsiatis, 2006, Zeng and Lin, 2014, Steingrimsson and Strawderman, 2017). The most widely used approach is weighted Cox regression, motivated by the intuition that the oversampling of cases can be balanced by an appropriate overweighting of the subcohort controls. The methods of Prentice (1986) and Barlow (1994) are the most commonly applied (Sharp et al., 2014). In both proposals, cases sampled outside of the subcohort enter into the analysis only at their respective failure times, allowing for the partially collected covariates—referred to as *expensive* covariates hereafter—to be time-dependent.

Assuming time-independence permits more efficient weighting approaches. Kalbfleisch and Lawless (1988) and Chen and Lo (1999) proposed weighting schemes based on inverse

probability weighting and post-stratification respectively. However, neither approach can make use of potentially available information on the unsampled controls, such as auxiliary variables and censoring times. Borgan et al. (2000) suggested several methods to address this issue, one of which was later augmented by Kulich and Lin (2004) to increase efficiency. Yet, weighted Cox estimators cannot be fully efficient; Nan et al. (2004) studied the semiparametric efficiency bound for the problem and quantified the amount of efficiency lost. It is unclear whether estimators that achieve the bound can be constructed in general.

Alternatives to weighted Cox regression have been proposed that use the full cohort data more efficiently and avoid the potential instability of inverse probability weights. Keogh and White (2013) described how multiple imputation can be applied to the problem, treating the expensive covariates for unsampled individuals as missing data. This requires a conditional imputation model for the expensive covariates given all observed variables, including the event time and case indicator. Care is required to avoid incompatibility issues (Morris et al., 2013) with the proportional hazards model: Keogh and White (2013) implemented the imputation with either a simple generalized linear model, or with rejection sampling using a preliminary marginal model. Full likelihood methods have assumed that the censoring mechanism is ignorable given the observed data. Nonparametric maximum likelihood estimation with the EM-algorithm was proposed by Scheike and Martinussen (2004), later extended by Zeng and Lin (2014) to include auxiliary variables and shown to be semiparametric efficient. However, computation is numerically unstable for more than three continuous auxiliary variables. Kulathinal and Arjas (2006) considered Bayesian analysis with data augmentation (Tanner and Wong, 1987), specifying a fully parametric form for the baseline cumulative hazard function.

We introduce a novel Bayesian framework for case-cohort Cox regression under the ignorable censoring assumption stated earlier; time-independence will be also be assumed since it is sufficient for our application and simplifies the descriptions, but we will discuss how this can be relaxed. The basic procedure is carried out in two stages. First, we obtain the posterior of the conditional distribution of the expensive covariates given the fully observed covariates using only the data from individuals with complete measurements—we refer to this as the restricted posterior. Samples from this restricted posterior serve as inputs to a pseudo-marginal Metropolis-Hastings algorithm (Lin et al., 2000, Beaumont, 2003, Andrieu and Roberts, 2009). This procedure yields the interpretation of using a likelihood function equal to the average of a set of Cox partial likelihoods, each computed from a dataset formed from the original with a different instance of imputed values for the missing expensive covariates. In this regard, our method shares a conceptual similarity with multiple imputation,

but is fully Bayesian and is automatically free of incompatibility issues with the Cox model. For large and moderate-dimensional datasets, we also propose extensions to the method based on modified versions of the correlated pseudo-marginal algorithm (Deligiannidis et al., 2018) that facilitate faster mixing.

Unlike Kulathinal and Arjas (2006), who require a fully specified joint model for the expensive and fully observed covariates, we allow for the (nuisance) marginal distribution of the fully observed covariates to be ignored. Moreover, our model for the baseline cumulative hazard function is nonparametrically specified and integrated out; this obviates sampling a potentially high-dimensional (or even infinite-dimensional) parameter, and leads to more robust inference for the log-hazard ratio than using a parametric model specification. With no auxiliary variables, and a discrete model for the expensive covariates, the likelihood reduces to the nonparametric likelihood used by Scheike and Martinussen (2004). When auxiliary variables are available, the conditional model for the expensive covariates can be arbitrarily specified, without the three dimensional covariate ceiling of the Zeng and Lin (2014) kernel estimation approach.

In §2.2, we introduce our method in a general setting, and propose modifications to the basic algorithm that facilitate improved mixing. Simulations comparing the performance of our approach to previous proposals are presented in §2.3. In §2.4, we apply our method to the EPIC-Norfolk study with the objective of investigating the associations between individual plasma phospholipid saturated fatty acids and incident type 2 diabetes. A challenging aspect is incorporating the compositional fatty acid data into the Cox model. Previous studies treated the proportions as absolute measurements, and used them directly. On the other hand, we first apply the additive logratio transformation (Aitchison, 1982) to the data. We discuss how this produces more reliable and interpretable results. To assess the effectiveness of our method and model for studying this application, we carried out a novel synthetic data experiment using a generating mechanism that exploits the case-cohort design and resamples from the real dataset.

2.2 Bayesian case-cohort Cox regression

2.2.1 Notation and background

First, consider the Cox proportional hazards model (Cox, 1972) for complete data. Let $D^0 = (Y, \Delta, Z, W)$, where $Y = \min(T, C)$, T and C denote the failure time and right-censoring time respectively, $\Delta = I(T \leqslant C)$ and $(Z, W) \in \mathbb{R}^m$ is a vector of time-independent covariates—

later on, there is a probability that Z is unobserved. The conditional hazard function of T given (Z,W) is $\lambda(t) = \exp(\beta_1^{\mathrm{T}} Z + \beta_2^{\mathrm{T}} W)\lambda_0(t)$, where $\beta = (\beta_1, \beta_2)$ is the log-hazard ratio describing the effects of the covariates, and $\lambda_0(t)$ is the baseline hazard function. Let $\Lambda_0(t) = \int_{s=0}^{t} \lambda_0(s)\, ds$ be the baseline cumulative hazard function. Suppose we observe an independent and identically distributed sample $D_i^0 = (Y_i, \Delta_i, Z_i, W_i)$ $(i = 1, \ldots, n)$ and let $R_i(t) = I(t \leqslant Y_i)$ be the at-risk indicator at time t for individual i. Assuming that T and C are conditionally independent given (Z,W), the parameter β can be estimated by maximizing the Cox partial likelihood[1] (Cox, 1972)

$$\prod_{i=1}^{n} \left\{ \frac{\exp(\beta_1^{\mathrm{T}} Z_i + \beta_2^{\mathrm{T}} W_i)}{\sum_{j=1}^{n} R_j(T_i) \exp(\beta_1^{\mathrm{T}} Z_j + \beta_2^{\mathrm{T}} W_j)} \right\}^{\Delta_i}, \tag{2.1}$$

which is equivalent to solving the partial score equations

$$\sum_{i=1}^{n} \Delta_i \left\{ (Z_i, W_i)^{\mathrm{T}} - \frac{\sum_{j=1}^{n} R_j(T_i) \exp(\beta_1^{\mathrm{T}} Z_j + \beta_2^{\mathrm{T}} W_j)(Z_j, W_j)^{\mathrm{T}}}{\sum_{j=1}^{n} R_j(T_i) \exp(\beta_1^{\mathrm{T}} Z_j + \beta_2^{\mathrm{T}} W_j)} \right\} = 0 \tag{2.2}$$

Suppose now that the covariates Z_i—which we will refer to as the expensive covariates—are measured for only a random subset of the cohort. Suppose also that we observe an independent and identically distributed sample X_i $(i = 1, \ldots, n)$ of auxiliary covariates that can be used to predict the unmeasured values of Z. More explicitly, we observe $D_i = (Y_i, \Delta_i, A_i Z_i, A_i, W_i, X_i)$ $(i = 1, \ldots, n)$, where A_i is a binary variable indicating whether the expensive covariates for individual i have been measured, and the other variables are defined as before. In a standard case-cohort design, $A_i = 1$ if individual i is a case, or a control sampled into the subcohort. Let $\mathscr{S} = \{i : A_i = 1\} \subset \{1, \ldots, n\}$ denote the set of individuals with measured Z_i, and let $\bar{\mathscr{S}} = \{1, \ldots, n\} \backslash \mathscr{S}$. We will make use of the shorthand notation of indexing by sets, e.g. $X_{\mathscr{S}} = \{X_i : i \in \mathscr{S}\}$.

We point out that the notation defined above differs slightly from Chapters 1, 3 and 4. In particular, we have used A to denote the binary selection variable, rather than R, and Z denotes the expensive covariates instead of the set of fully observed variables, as previously used in §1.3.1 (and also later in Chapter 3). We have chosen to do this to adhere to the conventions of the survival analysis literature, making it easier for the reader to switch between our work and the papers cited in this chapter.

We make the following assumptions:

[1] In the presence of ties, (2.1) takes the Breslow form of the partial likelihood (Breslow, 1972), which is the form we will use for the whole of this chapter.

Assumption 2.1. *For each $i = 1,\ldots,n$, C_i is independent of (T_i, Z_i) given (W_i, X_i).*

Assumption 2.2. *The vector $(A_1,\ldots,A_n)$ is independent of $(Z_1,\ldots,Z_n)$ given $\{(Y_j,\Delta_j,W_j,X_j) : j = 1,\ldots,n\}$.*

Assumption 2.1 strengthens the conditional independence assumption for full-data Cox regression, requiring further that C_i be independent of Z_i given (W_i, X_i) for each $i = 1,\ldots,n$. This will hold, for example, if the censoring is administrative. Assumption 2.2 is guaranteed to hold for standard case-cohort studies since the subcohort selection mechanism is known by design, and is either fully randomized or stratified on the baseline covariates X_i.

Of the 32 case-cohort analyses reviewed by Sharp et al. (2014), all but one used the Cox model, and 20 papers employed weighted Cox regression; the remaining papers carried out a standard unweighted Cox analysis with the sampled units. The motivation for weighted Cox regression is similar to that of the Horvitz-Thompson estimator and the other weighted estimators described in Chapter 1. Since the complete-data partial score equations (2.2) are unusable due to the unmeasured expensive covariates, they are replaced by a weighted version that involves only the sampled units.

We will describe the weighted Cox regression methods under the standard unstratified case-cohort design, where

$$\mathbb{P}(A_i = 1 \mid Y_i,\Delta_i,W_i,X_i) = \begin{cases} 1, & \text{for } \Delta_i = 1 \\ p, & \text{for } \Delta_i = 0, \end{cases}$$

and p is a known proportion that is strictly greater than 0. Let n_0 and m_0 be the number of controls in the full cohort and subcohort respectively. Define n_1 and m_1 for the cases similarly.

Weighted Cox estimators solve

$$\sum_{i=1}^{n} \Delta_i \left\{ (Z_i, W_i)^{\mathrm{T}} - \frac{\sum_{j=1}^{n} \omega_{ij} R_j(T_i) \exp(\beta_1^{\mathrm{T}} Z_j + \beta_2^{\mathrm{T}} W_j)(Z_j, W_j)^{\mathrm{T}}}{\sum_{j=1}^{n} \omega_{ij} R_j(T_i) \exp(\beta_1^{\mathrm{T}} Z_j + \beta_2^{\mathrm{T}} W_j)} \right\} = 0 \tag{2.3}$$

for a chosen set of weights $\{\omega_{ij}\}$. By design, ω_{ij} must be equal to zero if the j-th individual is a control who was not selected into the subcohort. The original approach by Prentice (1986) weighted all elements of the subcohort equally, but only included a case outside the subcohort at its failure time. This was due to the fact that Prentice considered time-dependent covariates and did not assume that the full covariate histories for cases outside the subcohort would be available. Asymptotic justification for the resulting estimator was only established

Table 2.1 Weight for individual j at failure time T_i. P, Prentice (1986); SP, Self and Prentice (1988); KL, Kalbfleisch and Lawless (1988); CL, Chen and Lo (1999).

	P	SP	KL	CL(I)	CL(II)
Case in $\mathscr{S}$	1	1	1	1	1
Case not in $\mathscr{S}$	$\mathbb{1}(i=j)$	0	1	1	1
Control in $\mathscr{S}$	1	1	$1/p$	n_1/m_1	n_0/m_0

later in Self and Prentice (1988). Here, the authors considered a slightly different estimator for which the cases outside the subcohort are left out altogether. Thus, the unknown full cohort quantities are estimated using only the randomly sampled subcohort. They proved consistency and asymptotic normality of their estimator, and argued that the earlier estimator by Prentice (1986) would generally behave similarly if the contribution of the appended cases is negligible for large samples. The Self and Prentice (1988) estimator has not been used in practice due to its low efficiency (e.g. Borgan et al., 2000).

Kalbfleisch and Lawless (1988) suggested an inverse probability weighted estimator; controls in the subcohort are up-weighted by the reciprocal of the subcohort sampling proportion, similar to the Horvitz-Thompson estimator. Chen and Lo (1999) proposed two weighting schemes, both of which can be viewed as variations of the Kalbfleisch and Lawless method with estimated weights. If the size of the full cohort is unknown, the controls are up-weighted by the number of cases in the full cohort divided by the number of cases in the subcohort. Otherwise, the number of controls outside the subcohort is known, and efficiency can be improved by instead estimating the weights by the number of controls in the full cohort divided by the number of controls in the subcohort. This relates to our discussion of estimated weights and post-stratification in Chapter 1. As one would expect, it can be shown (e.g. Borgan and Samuelson, 2017) that the Chen and Lo (1999) approaches improve on the other methods described in terms of efficiency. The different sets of weights are summarized in Table 2.1.

2.2.2 The pseudo-marginal algorithm

In this section, we provide a brief overview of an MCMC approach that is crucial to the methodology we develop later. Suppose we wish to draw samples from an analytically intractable probability density $p(\beta)$, where β is real-valued[2]. Since we are unable to evaluate

[2]Naturally, $p(\beta)$ will be the marginal posterior of the log-hazard ratio.

$p(\beta)$ point-wise, standard MCMC methods like the Metropolis-Hastings algorithm are infeasible. Suppose, however, that we are able to find a non-negative approximation $\hat{p}(\beta \mid U)$ satisfying $\mathbb{E}_U[\hat{p}(\beta \mid U)] = p(\beta)$, where U is a random variable with marginal density $p(u)$, i.e. $\hat{p}(\beta \mid U)$ is an unbiased estimator of the true $p(\beta)$. We can then construct an MCMC sampler that targets the "joint density" of (β, U) that is proportional to $\hat{p}(\beta \mid U)p(U)$; at stationarity, the samples of β are drawn from the exact marginal distribution $p(\beta)$. Thus, this type of algorithm is called *pseudo-marginal* (Andrieu and Roberts, 2009).

In particular, we are interested in the setting where $p(\beta) = \int p(\beta, \theta)\, d\theta$, where θ is a latent variable[3] and $p(\beta, \theta)$ is a tractable joint density. If $p(\beta, \theta) = h(\beta, \theta)g(\theta)$ for non-negative integrable functions h and g (they need not be $p(\beta \mid \theta)$ and $p(\theta)$ respectively), then a natural choice of unbiased estimator is

$$\hat{p}(\beta \mid \theta_1^\dagger, \ldots, \theta_B^\dagger) \propto \frac{1}{B} \sum_{b=1}^{B} h(\beta, \theta_b^\dagger),$$

where B is a positive integer, and $\theta_1^\dagger, \ldots, \theta_B^\dagger$ are independent and identically distributed according to the density proportional to g. In the context of the previous set-up, the auxiliary variable U is equal to $(\theta_1^\dagger, \ldots, \theta_B^\dagger)$.

Classically, this latent variable problem is handled using a Metropolis-Hastings algorithm that targets $p(\beta, \theta)$. A Gibbs sampler is a typical choice if we are able to sample from the conditionals $p(\beta \mid \theta)$ and $p(\theta \mid \beta)$; this is the *data augmentation* method of Tanner and Wong (1987). However, the mixing for such a sampler is likely to be prohibitively slow if θ is high-dimensional and strongly correlated with β (Andrieu and Roberts, 2009). The EM-algorithm (Dempster et al., 1977)—the analogous approach for maximizing a likelihood with latent variables—suffers the same deficiencies. The motivation of the pseudo-marginal algorithm was to create a far more computationally efficient alternative. The caveat is that we do not obtain samples of θ from its true marginal distribution, but that is not a concern for us since we are only interested in β.

The seemingly obvious choice of proposal distribution for U is its marginal density $p(u)$; that is, each proposal for U is independent of the current value. But this can lead to poor mixing if U is very high-dimensional. Deligiannidis et al. (2018) proposed the *correlated pseudo-marginal algorithm*, which uses a *pre-conditioned Crank-Nicolson proposal* (Cotter et al., 2013) for U. This choice of proposal is particularly well-suited for high-dimensional parameters, and the authors showed that there was a substantial gain in efficiency over the

[3]Later on, θ will be the set of all unmeasured expensive covariates.

non-correlated algorithm. Assuming that $p(u)$ is analytically tractable, there is no loss of generality to assume that U is standard multivariate normal by using inversion techniques. Given the current value U, a new proposal U' is drawn by setting $U' = \rho U + \varepsilon\sqrt{1-\rho^2}$, where $\rho \in [0,1]$ and $\varepsilon \sim \mathcal{N}(0,I)$. Higher values of ρ lead to higher acceptance probabilities at the expense of slower exploration of the parameter space; the value can be tuned accordingly.

We will see, however, that the auxiliary variable U in our method does not have a tractable marginal density. This is addressed in §2.2.4, where we extend the correlated pseudo-marginal algorithm for our purposes.

2.2.3 Model and inference

Under the general set-up described in §2.2.1, the likelihood function for the data $D_1, \ldots, D_n$ is equal to

$$\left[\prod_{i\in\mathscr{S}} \{\exp(\beta_1^{\mathrm{T}} Z_i + \beta_2^{\mathrm{T}} W_i)\lambda_0(Y_i)\}^{\Delta_i} \exp\left\{-e^{\beta_1^{\mathrm{T}} Z_i + \beta_2^{\mathrm{T}} W_i}\Lambda_0(Y_i)\right\} p(Z_i \mid W_i, X_i)\right] \left[\prod_{j\in\bar{\mathscr{S}}} \int \exp\left\{-e^{\beta_1^{\mathrm{T}} z_j + \beta_2^{\mathrm{T}} W_j}\Lambda_0(Y_j)\right\} p(z_j \mid W_j, X_j) dz_j\right] \tag{2.4}$$

multiplied by

$$\left\{\prod_{k=1}^{n} p(C_k \mid W_k, X_k)^{1-\Delta_k} \mathbb{P}(C_k \geqslant Y_k \mid W_k, X_k)^{\Delta_k} p(W_k, X_k)\right\} p(A_1, \ldots, A_n \mid \{(Y_j, \Delta_j, W_j, X_j) : j = 1, \ldots, n\}). \tag{2.5}$$

This is derived by taking the full likelihood for the Cox model with complete data (van der Vaart, 1998, p.425) and integrating out the missing expensive covariates $\{Z_j : j \in \bar{\mathscr{S}}\}$. In this section, we will describe our model restrictions for the different terms in the likelihood, and explain how to carry out inference on the hazard ratio.

The baseline cumulative hazard function Λ_0 is set to be a step function with jumps only at the failure times. Let $\Delta\Lambda_0(Y_i)$ denote the jump size of Λ_0 at Y_i for $\Delta_i = 1$. Then, the baseline hazard $\lambda_0(Y_i)$ equals $\Delta\Lambda_0(Y_i)$ if $\Delta_i = 1$, and 0 otherwise, and $\Lambda_0(t) = \sum_{i:\Delta_i=1, Y_i\leqslant t} \Delta\Lambda_0(Y_i)$. This idea was introduced by Breslow (1972) to motivate both the Cox partial likelihood estimator—from a nonparametric maximum likelihood perspective—and the Breslow estimator of the baseline cumulative hazard function. Scheike and Martinussen (2004) and Zeng and Lin (2014) extended this approach for case-cohort data.

We specify a Bayesian bootstrap prior for Λ_0

$$p(\Lambda_0) \propto \prod_{i:\Delta_i=1} \Delta\Lambda_0(Y_i)^{-1}.$$

For complete data, Kim and Lee (2003) referred to this as the "Poisson form Bayesian bootstrap" and showed that the resulting inference for β is equivalent to Bayesian analysis with the Cox partial likelihood. We will see that a similar phenomenon arises with case-cohort data. Kalbfleisch (1978) and Sinha et al. (2003) motivated this prior by considering the limit of a sequence of gamma process priors that become progressively more noninformative. This is similar to how the original Bayesian bootstrap (Rubin, 1981) can be motivated by considering the noninformative limit of a sequence of Dirichlet process priors.

For the terms of the form $p(Z \mid W,X)$ in (2.4), we require a regression model for the expensive covariates Z given the fully observed covariates (W,X). This will be used to predict the missing expensive covariate values and its specification is left to the user. We denote the parameter of this model by γ, which can be infinite-dimensional. The priors for β and γ are also left to the user, aside from the requirement of joint prior independence of Λ_0, β, and γ.

In (2.5), we set the models for the censoring $p(C_k \mid W_k,X_k)$, the fully observed covariates $p(W_k,X_k)$ and the selection $p(A_1,\ldots,A_n \mid \{(Y_j,\Delta_j,W_j,X_j) : j = 1,\ldots,n\})$ to be a priori independent of (Λ_0,β,γ). Thus, (2.5) will drop out of the subsequent analysis and no further specification of these models is needed.

It follows that the posterior for (Λ_0,β,γ) given $D_1,\ldots,D_n$ is proportional to

$$\begin{aligned}&\left[\prod_{i\in\mathscr{S}} \exp(\beta_1^{\mathrm{T}}Z_i+\beta_2^{\mathrm{T}}W_i)^{\Delta_i} \exp\left\{-e^{\beta_1^{\mathrm{T}}Z_i+\beta_2^{\mathrm{T}}W_i}\Lambda_0(Y_i)\right\} p(Z_i \mid W_i,X_i,\gamma)\right] \\ &\left[\prod_{j\in\bar{\mathscr{S}}} \int \exp\left\{-e^{\beta_1^{\mathrm{T}}z_j+\beta_2^{\mathrm{T}}W_j}\Lambda_0(Y_j)\right\} p(z_j \mid W_j,X_j,\gamma)dz_j\right] p(\gamma)p(\beta).\end{aligned} \tag{2.6}$$

Let

$$p(\gamma \mid D_{\mathscr{S}}) \propto \left[\prod_{i\in\mathscr{S}} p(Z_i \mid W_i,X_i,\gamma)\right] p(\gamma) \tag{2.7}$$

be the posterior for γ given only the data for individuals in $\mathscr{S}$—the set of individuals with measured Z_i. We refer to this as the *restricted posterior* of γ. By integrating (2.6) with respect to Λ_0, applying Fubini's theorem to exchange the order of integration with the missing

covariates, and then integrating with respect to γ, we find that

$$p(\beta \mid D_1,\ldots,D_n) \propto \int \prod_{k=1}^{n} \left\{ \frac{\exp(\beta_1^{\mathrm{T}} z_k + \beta_2^{\mathrm{T}} W_k)}{\sum_{l=1}^{n} R_l(T_k) \exp(\beta_1^{\mathrm{T}} z_l + \beta_2^{\mathrm{T}} W_l)} \right\}^{\Delta_k} \left[\prod_{i \in \mathscr{S}} \delta\{z_i = Z_i\} dz_i \right] \left[\prod_{j \in \bar{\mathscr{S}}} p(z_j \mid W_j, X_j, \gamma) dz_j \right] p(\gamma \mid D_{\mathscr{S}}) d\gamma \, p(\beta) \tag{2.8}$$

where $\delta\{\cdot\}$ is the Dirac delta function. (A more detailed derivation can be found in §B.1.) Thus, the posterior of β is proportional to the prior of β multiplied by the Cox partial likelihood averaged across the restricted posterior predictive distribution of the missing covariates.

Although this averaged Cox partial likelihood is probably intractable, it is generally possible to draw values of the missing covariates from the restricted posterior predictive distribution, either exactly or by MCMC methods. This provides us with a computational strategy to sample from the marginal posterior of β using the pseudo-marginal algorithm. Let B be a positive integer (the choice of which is suggested below). Define the distribution of a $B \times |\bar{\mathscr{S}}|$ random variable Z^{mis} by

$$p(z^{\mathrm{mis}} \mid W_{\bar{\mathscr{S}}}, X_{\bar{\mathscr{S}}}, D_{\mathscr{S}}) = \prod_{b=1}^{B} \int \prod_{j \in \bar{\mathscr{S}}} p(z_j^{(b)} \mid W_j, X_j, \gamma_b) p(\gamma_b \mid D_{\mathscr{S}}) d\gamma_b, \tag{2.9}$$

where $\{z_j^{(b)} : j \in \bar{\mathscr{S}}, b = 1,\ldots,B\}$ are the components of z^{mis}. We can sample Z^{mis} as follows: draw B independent values $\gamma_1,\ldots,\gamma_B$ from the restricted posterior (2.7), and for each $b = 1,\ldots,B$ and each $j \in \bar{\mathscr{S}}$, draw from $p(z \mid W_j, X_j, \gamma_b)$; Z^{mis} takes the value of the set of imputed covariates. By combining Z^{mis} with the measured values of Z, this procedure yields B datasets with complete covariate measurements. Define the function h by the mean of the partial likelihood functions across all datasets:

$$h(\beta, Z^{\mathrm{mis}}) = B^{-1} \sum_{b=1}^{B} \left[\prod_{k:\Delta_k=1} \frac{\exp(\beta_1^{\mathrm{T}} Z_k + \beta_2 W_k)}{\sum_{l=1}^{n} R_l(Y_k) \exp(\beta_1^{\mathrm{T}} Z_l^{(b)} + \beta_2 W_l)} \right]$$

where $Z_l^{(b)}$ is the expensive covariate for individual l in the b-th imputed dataset.

Let $q(\tilde{\beta} \mid \beta)$ be a user-specified proposal distribution for β. Algorithm 2.1 describes the basic template for sampling from the marginal posterior of β. The algorithm can be viewed as a Metropolis-Hastings algorithm for the augmented parameter $(\beta, z^{\mathrm{mis}})$ with

Algorithm 2.1: Sampling from the marginal posterior of β

Input initial parameter value $\beta^{(0)}$.
Draw $Z^{\text{mis}}_{(0)}$ from $p(z^{\text{mis}} \mid W_{\bar{\mathscr{S}}}, X_{\bar{\mathscr{S}}}, D_{\mathscr{S}})$ (2.9).
For $r = 1$ to $r = N$
(a) Propose $\tilde{\beta}$ from $q(\beta \mid \beta^{(r-1)})$.
(b) Draw $\tilde{Z}^{\text{mis}}$ from $p(z^{\text{mis}} \mid W_{\bar{\mathscr{S}}}, X_{\bar{\mathscr{S}}}, D_{\mathscr{S}})$.
(c) With probability $\min\left\{1, \frac{q(\beta^{(r-1)}|\tilde{\beta})p(\tilde{\beta})h(\tilde{\beta},\tilde{Z}^{\text{mis}})}{q(\tilde{\beta}|\beta^{(r-1)})p(\beta^{(r-1)})h(\beta^{(r-1)},Z^{\text{mis}}_{(r-1)})}\right\}$,
set $\beta^{(r)} = \tilde{\beta}$ and $Z^{\text{mis}}_{(r)} = \tilde{Z}^{\text{mis}}$.
Otherwise, set $\beta^{(r)} = \beta^{(r-1)}$ and $Z^{\text{mis}}_{(r)} = Z^{\text{mis}}_{(r-1)}$.
Output $(\beta^{(1)}, \ldots, \beta^{(N)})$.

proposal distribution $q^*(\tilde{\beta}, \tilde{Z}^{\text{mis}} \mid \beta, z^{\text{mis}}) = q(\tilde{\beta} \mid \beta) p(\tilde{z}^{\text{mis}} \mid W_{\bar{\mathscr{S}}}, X_{\bar{\mathscr{S}}}, D_{\mathscr{S}})$. The acceptance probability for the r-th iteration with proposal $(\tilde{\beta}, \tilde{Z}^{\text{mis}})$ and current value $(\beta^{(r-1)}, Z^{\text{mis}}_{(r-1)})$ can now be written as

$$\min\left\{1, \frac{q^*(\beta^{(r-1)}, Z^{\text{mis}}_{(r-1)} \mid \tilde{\beta}, \tilde{Z}^{\text{mis}}) p(\tilde{\beta}) p(\tilde{Z}^{\text{mis}} \mid W_{\bar{\mathscr{S}}}, X_{\bar{\mathscr{S}}}, D_{\mathscr{S}}) h(\tilde{\beta}, \tilde{Z}^{\text{mis}})}{q^*(\tilde{\beta}, \tilde{Z}^{\text{mis}} \mid \beta^{(r-1)}, Z^{\text{mis}}_{(r-1)}) p(\beta^{(r-1)}) p(Z^{\text{mis}}_{(r-1)} \mid W_{\bar{\mathscr{S}}}, X_{\bar{\mathscr{S}}}, D_{\mathscr{S}}) h(\beta^{(r-1)}, Z^{\text{mis}}_{(r-1)})}\right\}.$$

Thus, Algorithm 2.1 converges to stationarity with an invariant distribution function proportional to $p(\beta) p(z^{\text{mis}} \mid W_{\bar{\mathscr{S}}}, X_{\bar{\mathscr{S}}}, D_{\mathscr{S}}) h(\beta, z^{\text{mis}})$. By construction, the expectation of $h(\beta, Z^{\text{mis}})$ with respect to $p(z^{\text{mis}} \mid W_{\bar{\mathscr{S}}}, X_{\bar{\mathscr{S}}}, D_{\mathscr{S}})$ is proportional to $p(D_1, \ldots, D_n \mid \beta)$ in β; the marginal invariant distribution of β is therefore equal to the true marginal posterior. If MCMC is required to draw restricted posterior values of γ, it is straightforward to modify Algorithm 2.1 to sample the further augmented parameter $(\beta, z^{\text{mis}}, \gamma_1, \ldots, \gamma_B)$.

Since Algorithm 2.1 is a pseudo-marginal algorithm that uses an average of unbiased estimators (as opposed to a particle filter), and computation time scales roughly linearly in B, the results of Sherlock et al. (2017) suggest that the optimal computational tradeoff between number of iterations N and number of estimators B is achieved by setting $B = 1$. If parallel computing is available with negligible overheads, B should be set equal to the number of available cores, so that the B partial likelihood functions are computed in parallel.

2.2.4 Modifications to improve mixing

For large datasets with moderate to high dimensional covariates, such as our application in §2.4, Algorithm 2.1 may not be sufficient to ensure good mixing. In this section, we describe how improved mixing can be attained.

As discussed in §2.2.2, the correlated pseudo-marginal algorithm (Deligiannidis et al., 2018) improves on the efficiency of the standard pseudo-marginal algorithm by correlating the current and proposed values of the variables that are used to obtain the estimate of the likelihood factor (Z^{mis} in our set-up). However, this method requires the distribution of these variables to be inverted into a standard multivariate normal distribution; for the restricted posterior predictive distribution of Z^{mis} given by (2.9), this will generally be impossible in practice due to intractability.

We solve this by instead considering the restricted posterior predictive distribution of Z^{mis} *conditional* on $\gamma_1, \ldots, \gamma_B$. In equation (2.9), the factors of the form $p(z_j^{(b)} \mid W_j, X_j, \gamma^{(b)})$ are user-specified probability density/mass functions. Generally, this means that we can analytically or numerically evaluate a deterministic function φ such that $\varphi(U, W_{\bar{\mathscr{F}}}, X_{\bar{\mathscr{F}}}, \gamma_1, \ldots, \gamma_B)$ has the distribution of Z^{mis}, where $U \sim \mathscr{N}(0_M, I_M)$ for $M = B \times \bar{\mathscr{F}}$, independent of $\gamma_1, \ldots, \gamma_B$. This motivates Algorithm 2.2, a modified version of the correlated pseudo-marginal algorithm in which the set of parameters is augmented by $\gamma_1, \ldots, \gamma_B$, and the values of U are correlated to the level determined by $\rho \in (-1, 1)$. When $\rho = 0$, Algorithm 2.2 is equivalent to Algorithm 2.1. Recall that increasing ρ leads to higher acceptance probabilities but slower exploration of the parameter space, and the value can be tuned accordingly. We justify the algorithm in §B.2.

If this is insufficient to ensure adequate mixing, we can correlate $\gamma_1, \ldots, \gamma_B$ as well. In the case where the restricted posterior $p(\gamma \mid D_{\mathscr{S}})$ admits an analytic expression, it is straightforward to extend Algorithm 2.2 by replacing step (b) with a correlated proposal using the normal inversion strategy employed for Z^{mis}. We take this approach in §2.4, albeit only for a subparameter of γ. Otherwise, we can sample $\gamma_1, \ldots, \gamma_B$ using a Metropolis-Hastings algorithm with a proposal distribution chosen to induce a suitable level of correlation.

2.3 Simulation study

In this section, we provide an initial assessment of our proposal by comparing its performance with the weighted Cox regression methods of Prentice (1986), Kalbfleisch and Lawless (1988) and Chen and Lo (1999) (the more efficient version labelled CL(II) in Table 2.1). All three

Algorithm 2.2: Correlated sampling algorithm

Input initial parameter value $\beta^{(0)}$
Draw $U^{(0)} \sim \mathcal{N}(0_M, I_M)$.
Draw i.i.d. $\gamma_1^{(0)}, \ldots, \gamma_B^{(0)} \sim p(\gamma \mid D_{\mathscr{S}})$.
Compute $Z_{(0)}^{\text{mis}} = \varphi(U^{(0)}, W_{\bar{\mathscr{S}}}, X_{\bar{\mathscr{S}}}, \gamma_1^{(0)}, \ldots, \gamma_B^{(0)})$.
For $r = 1$ to $r = N$
(a) Draw a proposal $\tilde{\beta}$ from $q(\beta \mid \beta^{(r-1)})$.
(b) Draw i.i.d. $\tilde{\gamma}_1, \ldots, \tilde{\gamma}_B \sim p(\gamma \mid D_{\mathscr{S}})$.
(c) Draw $\varepsilon \sim \mathcal{N}(0_M, I_M)$ and set $\tilde{U} = \rho U^{(r-1)} + \sqrt{(1-\rho^2)}\varepsilon$.
(d) Compute $\tilde{Z}^{\text{mis}} = \varphi(\tilde{U}, W_{\bar{\mathscr{S}}}, X_{\bar{\mathscr{S}}}, \tilde{\gamma}_1, \ldots, \tilde{\gamma}_B)$.
(e) With probability $\min\left\{1, \frac{q(\beta^{(r-1)}|\tilde{\beta})p(\tilde{\beta})h(\tilde{\beta}, \tilde{Z}^{\text{mis}})}{q(\tilde{\beta}|\beta^{(r-1)})p(\beta^{(r-1)})h(\beta^{(r-1)}, Z_{(r-1)}^{\text{mis}})}\right\}$, set $(\beta^{(r)}, U^{(r)}) = (\tilde{\beta}, \tilde{U})$.
Otherwise, set $(\beta^{(r)}, U^{(r)}) = (\beta^{(r-1)}, U^{(r-1)})$
Output $(\beta^{(1)}, \ldots, \beta^{(N)})$.

existing methods can be implemented using the R package `survival`. Since these methods are unable to incorporate auxiliary covariates to improve the prediction of the missing expensive covariates, we considered the special case where there are no auxiliary covariates to enable direct comparisons. In §2.4.3, we perform further experiments that are closely based on the application to the EPIC-Norfolk study.

Failure times were independently and identically generated for a full cohort size of $n = 2000$ using a Weibull baseline hazard function

$$\lambda(t) = \exp(\beta_0 Z)\eta\nu t^{\nu-1},$$

where β_0 is the target quantity. The expensive covariate Z was generated from $\mathcal{N}(0,1)$. The censoring times took the value 3 with probability 0.2, and were otherwise uniformly distributed between 0 and 3. The sets of values of (β_0, η, ν), with $\beta \in \{-0.3, 0, 0.3\}$, were chosen such that the average proportion of cases (approximately 4%) roughly corresponded to that of the application. The subcohort sampling proportion $p = 0.04$ was chosen similarly.

For our Bayesian method, computation was carried out using Algorithm 2.1. We specified a Bayesian bootstrap model (Rubin, 1981) for the distribution of Z. A new value of Z^{mis} is proposed as follows: sample a set of probability weights from Dirichlet$(1, \ldots, 1)$, each corresponding to an observed value of Z in $\mathscr{S}$; conditional on the weights, independently draw each missing covariate from the observed set of Z values. For β, we specified an improper uniform prior on $\mathbb{R}$ and used a normal random walk proposal: $q(\tilde{\beta} \mid \beta^{(r-1)}) =$

$\mathcal{N}(\beta^{(r-1)}, \sigma^2)$. For the proposal variance, we used four times the estimated variance of the Chen and Lo (1999) estimator (using a weighted Cox analysis). With parallel computing, the communication overhead dominated the computation time of the likelihood estimator; thus, the number of estimators B was set to 1. The first 1000 Metropolis-Hastings iterations were discarded, and the subsequent 20000 iterations were used for analysis. We chose the posterior mean as the Bayes point estimator.

Table 2.2 summarizes the performance of the four methods across 2000 Monte Carlo trials. The relative efficiencies were computed by taking the ratio of the mean squared errors relative to the complete data analysis, where information on all variables is available for the full cohort. The coverage properties of the Bayesian method were assessed by examining the proportion of trials where β_0 was contained in the central 95% posterior credible region. For the remaining procedures, we have reported the coverage from 95% Wald intervals with robust variance estimates.

Our proposal substantially outperformed the three weighted Cox approaches in all settings: the Bayes estimator was approximately unbiased with smaller standard deviations, leading to a significant reduction in efficiency loss relative to the complete data analysis. The central posterior credible regions also exhibited frequentist coverage close to nominal levels, improving on the Prentice method in particular. We draw attention to the fact that we have specified a noninformative prior for β and a nonparametric model for Z which makes virtually no modeling assumptions. Thus, there is ample scope to make further performance gains if prior substantive knowledge is available.

We mention also that we implemented the nonparametric maximum likelihood estimator (Scheike and Martinussen, 2004, Zeng and Lin, 2014), which is computed using an EM-algorithm. However, we were unable to obtain numerical convergence for any of the sets of parameter values, so we excluded this estimator from the comparisons.

Additionally, we investigate how the computation time of our method scales with dataset size. The set-up is the same as before, with parameter values $\beta_0 = 0.3$, $\eta = 0.01$, $\nu = 2.0$. In Figure 2.1, we have plotted the computation times across 100 trials for n ranging from 1000 to 10000 in steps of 1000. Additionally, we have plotted the fitted curve from a quadratic model, which suggests that our method and implementation are $\mathcal{O}(n^2)$.

Table 2.2 Comparison of log-hazard ratio estimates for 2000 replicates. CL, Chen and Lo (1999); KL, Kalbfleisch and Lawless (1988); ESD, empirical standard deviation; RMSE, root mean squared error; RE, relative efficiency; Cov, coverage.

	$\beta_0 = -0.3, \eta = 0.01, \nu = 2.0$					$\beta_0 = -0.3, \eta = 0.02, \nu = 1.2$				
Estimator	Bias	ESD	RMSE	RE	Cov (%)	Bias	ESD	RMSE	RE	Cov (%)
Complete	0.000	0.109	0.109	1.000	95.00	0.000	0.107	0.107	1.000	95.70
Bayes	0.013	0.145	0.146	0.561	94.80	0.012	0.141	0.142	0.563	95.85
CL	-0.021	0.206	0.207	0.278	94.25	-0.017	0.190	0.191	0.312	94.50
KL	-0.021	0.206	0.207	0.278	94.25	-0.017	0.190	0.191	0.312	94.50
Prentice	-0.012	0.202	0.202	0.293	90.15	-0.010	0.186	0.187	0.325	91.20

	$\beta_0 = 0, \eta = 0.01, \nu = 2.0$					$\beta_0 = 0, \eta = 0.02, \nu = 1.2$				
Estimator	Bias	ESD	RMSE	RE	Cov (%)	Bias	ESD	RMSE	RE	Cov (%)
Complete	0.002	0.115	0.115	1.000	94.40	0.002	0.113	0.113	1.000	95.00
Bayes	0.004	0.161	0.161	0.508	94.80	0.005	0.163	0.163	0.485	95.10
CL	0.003	0.194	0.194	0.352	95.25	0.003	0.181	0.181	0.394	95.75
KL	0.003	0.194	0.194	0.352	95.25	0.003	0.181	0.181	0.394	95.75
Prentice	0.002	0.191	0.191	0.363	90.20	0.003	0.178	0.178	0.405	91.20

	$\beta_0 = 0.3, \eta = 0.01, \nu = 2.0$					$\beta_0 = 0.3, \eta = 0.02, \nu = 1.2$				
Estimator	Bias	ESD	RMSE	RE	Cov (%)	Bias	ESD	RMSE	RE	Cov (%)
Complete	0.001	0.114	0.114	1.000	94.10	0.001	0.112	0.112	1.000	93.65
Bayes	-0.008	0.151	0.151	0.571	94.80	-0.008	0.149	0.150	0.564	94.90
CL	0.023	0.204	0.206	0.307	93.70	0.019	0.192	0.193	0.340	94.30
KL	0.023	0.204	0.206	0.307	93.70	0.019	0.192	0.193	0.340	94.30
Prentice	0.014	0.201	0.202	0.319	89.45	0.012	0.188	0.189	0.355	90.65

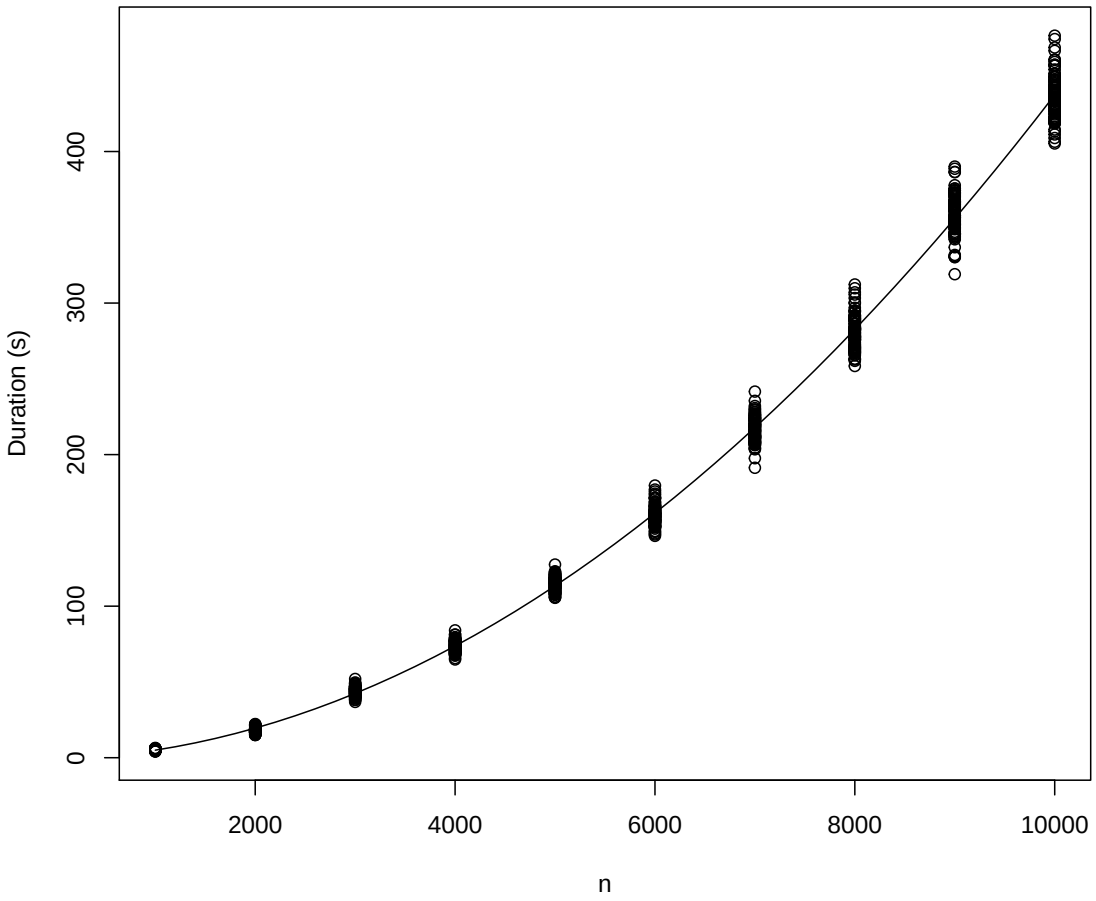


Fig. 2.1 Computation times by dataset size.

2.4 Application to the EPIC-Norfolk study

2.4.1 Study design and data preparation

We apply our methodology to investigate the associations between individual saturated fatty acids and incident type 2 diabetes, using data from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Norfolk study (Day et al., 1999). The original cohort study included 25,639 men and women aged 40 to 79. Between 1993 and 1997, all participants were invited to undergo a baseline health check, during which anthropometric measurements and blood samples were taken by trained nurses. Participants were also required to complete a health and lifestyle questionnaire. Follow-up concluded on 31st December 2007; the follow-up time for each participant was taken to be the total number of days from the recruitment date to diabetes diagnosis or the censoring date. This form of administrative censoring implies that Assumption 2.1 is satisfied.

As one of twenty-six centres contributing to the EPIC-InterAct case-cohort study (Langenberg et al., 2011), a random subcohort of size 1025, along with the remaining 863 incident cases, were selected to have their blood samples analyzed for fatty acid composition. The quantities of the fatty acids were expressed as a percentage of total plasma phospholipid fatty acids (mol%). Among the 27 fatty acids with relative concentrations greater than 0.05%, 9 were identified as saturated fatty acids (SFAs), belonging to 3 different groups: 2 odd-chain SFAs (pentadecanoic acid, C15:0; heptadecanoic acid, C17:0), 3 even-chain SFAs (myristic acid, C14:0; palmitic acid, C16:0; stearic acid, C18:0) and 4 very-long-chain SFAs (arachidic acid, C20:0; behenic acid, C22:0; tricosanoic acid, C23:0; lignoceric acid, C24:0).

We identified age at recruitment, sex, waist circumference, body mass index and physical activity index as potential confounders of the effects of the saturated fatty acids on incident type 2 diabetes. Additionally, we have chosen to incorporate 5 dietary variables from the questionnaires to help predict the missing values of the fatty acids. These are daily intakes (grams per day) of: potatoes and other tubers, fruit, fish and shellfish, meat and meat products, and dairy products.

Individuals with prevalent type 2 diabetes (855 individuals) or unknown diabetes status (5 individuals), as well as those with missing confounder (1832 individuals) or dietary data (310 individuals), were excluded from analysis. Following Forouhi et al. (2014), we also excluded individuals with a ratio of energy intake to energy requirement in the bottom or top 1% as probable dietary misreporters (432 individuals). There remain 22219 individuals in the dataset, with a subcohort of size 886 (860 controls and 26 incident cases) and 771 non-subcohort incident cases. From this, 14 subcohort individuals and 95 non-subcohort incident

cases have missing fatty acid measurements. Instead of excluding these individuals and losing valuable data on cases, we have chosen to assume that this missingness is independent of the values of the missing fatty acid data given the available information, so that Assumption 2.2 is still satisfied.

2.4.2 Model specification

We set W to be the potential confounders described in §2.4.1. Sex was represented by a binary variable. The physical activity index data were categorical with four levels: "Inactive", "Moderately inactive", "Moderately active" and "Active". This information was decomposed into three binary dummy variables with "Active" as the reference category. The remaining confounders—age, waist circumference, and body mass index—were scaled by their full cohort standard deviations. The auxiliary variable X was set to be the 5 dietary variables after undergoing the log-transformation $x \mapsto \log(1+x)$.

The fatty acid data are compositional—the relative concentrations of the individual fatty acids sum to 100%. To address this, we applied the additive logratio transformation (Aitchison, 1982). Denote a fatty acid measurement value by $z' = (z'_1, \ldots, z'_9, z'_O)$, where $z'_1, \ldots, z'_9$ are the relative concentrations of the 9 SFAs, and z'_O is the total relative concentration of all remaining fatty acids. If all entries of z' are non-zero, its additive logratio image in $\mathbb{R}^9$ is

$$\left(\log \frac{z'_1}{z'_O}, \ldots, \log \frac{z'_9}{z'_O}\right). \tag{2.10}$$

Otherwise, we first take the zero replacement strategy described in Greenacre (2019). Any zero entries of z' are replaced by half of the smallest possible positive measurement. In this case, since measurements are given to two decimal places of a percentage, all zeros are replaced by 0.005%. Set Z to be the transformed fatty acid vector as described after scaling each component by its standard deviation within the subcohort. In §2.4.4, we discuss interpretations and the advantages over direct use of the compositional data.

Let $V = (1, W^{\mathrm{T}}, X^{\mathrm{T}})^{\mathrm{T}}$. We specify a multivariate normal linear regression model

$$Z \mid W, X, \xi, \Sigma \sim \mathcal{N}(\xi^{\mathrm{T}} V, \Sigma) \tag{2.11}$$

where $\xi \in \mathbb{R}^{13\times 9}$ and $\Sigma \in \mathbb{R}^{9\times 9}$. Let $n_{\mathscr{S}} = |\mathscr{S}| = 1548$, the total number of individuals with fatty acid measurements. We use the Jeffreys prior

$$p(\xi, \Sigma) \propto |\Sigma|^{-(9+1)/2} = |\Sigma|^{-5},$$

which can be interpreted as the noninformative limit of a matrix normal-inverse Wishart prior (Gelman et al., 2013). By conjugacy, the restricted posterior distributions are

$$\xi \mid \Sigma, Z_{\mathscr{S}}, W_{\mathscr{S}}, X_{\mathscr{S}} \sim \mathscr{MN}(\hat{\xi}, (V_{\mathscr{S}}^{\mathrm{T}} V_{\mathscr{S}})^{-1}, \Sigma) \tag{2.12}$$

$$\Sigma | Z_{\mathscr{S}}, W_{\mathscr{S}}, X_{\mathscr{S}} \sim \mathscr{IW}(\Psi, n_{\mathscr{S}}), \tag{2.13}$$

where $\mathscr{MN}$ and $\mathscr{IW}$ denote the matrix normal and inverse Wishart distributions respectively and

$$\hat{\xi} = (V_{\mathscr{S}}^{\mathrm{T}} V_{\mathscr{S}})^{-1} V_{\mathscr{S}}^{\mathrm{T}} Z_{\mathscr{S}} \quad \text{(least squares estimator)}$$
$$\Psi = (Z_{\mathscr{S}} - V_{\mathscr{S}} \hat{\xi})^{\mathrm{T}} (Z_{\mathscr{S}} - V_{\mathscr{S}} \hat{\xi}) \quad \text{(residual sum of squares).}$$

The remaining notation follows §2.2. For the log-hazard ratio β, we specified independent, weakly informative Student-t priors for each of the components, all centered at 0 with 3 degrees of freedom.

2.4.3 Synthetic data experiment

To assess our method and model specification, we analyzed synthetic datasets designed to resemble the real data. Our design takes advantage of the fact that the subcohort data are a random sample from the full cohort; thus, the empirical distribution of the subcohort data should provide a reasonable approximation of the target population distribution. Synthetic datasets were generated as follows: 1. repeatedly sample with replacement from the subcohort to generate a synthetic full cohort of half the size of the original; 2. generate a new subcohort of half the size as the original subcohort by sampling without replacement from the synthetic full cohort. Step 1 generates a new full cohort dataset using the empirical distribution of the subcohort, and step 2 implements the case-cohort design. We proceed to analyze the dataset without using the fatty acid data for the unsampled controls, as in a standard case-cohort analysis.

The factor of a half introduces a cross-validation element to the experiment, guaranteeing that a substantial proportion of the original subcohort controls will not be sampled into the synthetic subcohort; this way, the predictive performance of the regression model (2.11) is evaluated.

For the purpose of these experiments, we removed all individuals in the original subcohort with missing fatty acid measurements before generating the synthetic data. As a result, we are able to compare our results to a truth: the Cox estimator consistently estimates the hazard

ratio, so we can find the true hazard ratio of the generating distribution to an arbitrary level of accuracy by computing the Cox estimate for a large dataset generated by resampling from the original subcohort—we used $n = 2000000$.

The size and complexity of the datasets necessitated a correlated sampling algorithm to achieve good mixing; we took the approach described at the end of §2.2.4, correlating both the missing fatty acid variables Z^{mis} and the regression coefficients ξ. The full details are provided in §B.3. For each dataset, we discarded the first 100000 iterations of the sampler and used the following 300000. We determined that this was sufficient for chain convergence by examining the trace plots for several trials.

The results for 200 synthetic datasets are summarized in Figure 2.2 and Table 2.3; we compared the performance of the posterior mean estimator with the Prentice estimator. Figure 2.2 contains violin plots of the estimates for the log-hazard ratios with reference to the true values. Table 2.3 compares the numerical performance results of both estimators. The efficiency gain results were found by dividing the mean squared error of the Prentice estimator by the mean squared error of the Bayes estimator and computing the percentage difference. We can roughly interpret the efficiency gain values as the increase in sample size required for the Prentice estimator to match the performance of the Bayes estimator at the current sample size.

The results demonstrate that the method and model are effective for analyzing the data and produce substantial efficiency gains over the Prentice estimator. We also mention the fact that the experimental design favors the Prentice estimator; in each trial, the set of unsampled controls contains exact replicates of the subcohort controls, which matches the implicit modelling assumptions of the Prentice estimator. This will not be the case in the real application, so we expect the actual performance gains to be even greater.

2.4.4 Results for the EPIC-Norfolk data

For the application, we used the same sampling algorithm as the one in §2.4.3 (described fully in §B.3). We discarded the first 200000 iterations of the sampler, and used the following 800000 for analysis. The convergence diagnostics can be found in §B.4.

To interpret the results, we recall that the fatty acid data—originally compositional—were additive logratio transformed using (2.10), and then scaled by their respective estimated standard deviations. For concreteness, let us specifically consider the saturated fatty acid C14:0. The posterior mean estimate of the hazard ratio is 1.18 (Table 2.4), implying that an increase of 1 standard deviation in the logratio corresponding to C14:0, keeping all

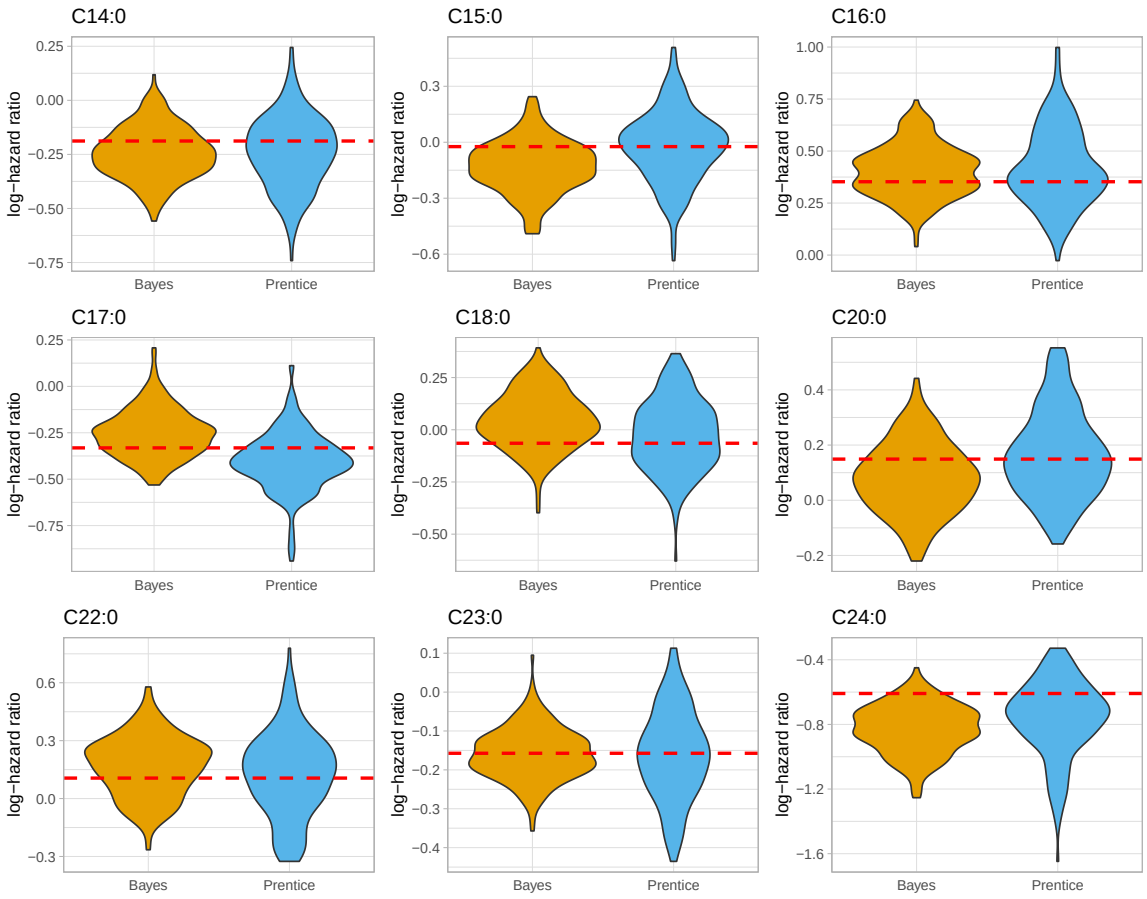


Fig. 2.2 Posterior mean and Prentice estimates of the saturated fatty acid log-hazard ratios. The red dashed lines represent the true values.

Table 2.3 Comparison of log-hazard ratio estimates in the synthetic data experiment. ESD, empirical standard deviation; RMSE, root mean squared error; EG, efficiency gain.

	C14:0				C15:0				C16:0			
Estimator	Bias	ESD	RMSE	EG	Bias	ESD	RMSE	EG	Bias	ESD	RMSE	EG
Prentice	-0.051	0.164	0.172		0.000	0.189	0.189		0.069	0.186	0.199	
Bayes	-0.055	0.119	0.131	+72%	-0.106	0.142	0.177	+14%	0.051	0.124	0.134	+120%

	C17:0				C18:0				C20:0			
Estimator	Bias	ESD	RMSE	EG	Bias	ESD	RMSE	EG	Bias	ESD	RMSE	EG
Prentice	-0.072	0.156	0.172		0.040	0.168	0.173		0.017	0.158	0.159	
Bayes	0.078	0.130	0.152	+29%	0.108	0.134	0.172	+1%	-0.068	0.127	0.144	+21%

	C22:0				C23:0				C24:0			
Estimator	Bias	ESD	RMSE	EG	Bias	ESD	RMSE	EG	Bias	ESD	RMSE	EG
Prentice	0.019	0.217	0.218		-0.004	0.115	0.115		-0.154	0.234	0.280	
Bayes	0.061	0.155	0.167	+71%	-0.002	0.066	0.066	+200%	-0.220	0.151	0.266	+11%

other logratios and confounders fixed, increases the hazard of type 2 diabetes onset by 18%. Framing this with respect to a particular individual, the change occurs if their *absolute* quantity of C14:0 increases, with all else kept equal. This way, the only logratio that changes is the one corresponding to C14:0; the ratios of the other saturated fatty acids to the reference category (the total of all remaining fatty acids) remain the same as before. Cox regression with isometric logratio transformed compositional data has previously been proposed (McGregor et al., 2020), but this produces much less interpretable results than what is described above.

A review and meta-analysis of previous studies can be found in Huang et al. (2019). To the best of our knowledge, our work is the first to use transformed fatty acid data to investigate this problem. There are several reasons why we believe that this is preferable over direct use of the raw data. First, as noted by Pearson (1897), treating proportions as absolute measurements runs the risk of introducing "spurious correlation" into the analysis. In Figure 2.3, we observe that the moderate negative correlation on the original scale between C16:0 and C18:0—by far the two most abundant saturated fatty acids—is removed after transformation. Also, additive changes in percentages ignore the inherently relative nature of the data. For example, an increase from 0% to 1% of a fatty acid is viewed as equivalent to an increase from 4% to 5%. One could further argue that increasing the proportion of a single fatty acid while keeping some others fixed does not correspond to any type of meaningful hypothetical intervention. Moreover, the total proportion of all omitted fatty acids (e.g. all

non-saturated fatty acids, or everything apart from the even-chain SFAs) is forced to decrease in order for the proportions to sum to 100%, making the analysis strongly dependent on the choice of included fatty acids. This could partly explain the disparity in results across studies. In contrast, our use of the transformation gives us the previously described interpretation of increasing the absolute quantity of a fatty acid. This corresponds to a more intuitive intervention, and only depends on the particular fatty acid that is being changed.

The meta-analysis by Huang et al. (2019)—with 10 studies included—suggested that there was conclusive evidence for the effects of only three saturated fatty acids: C15:0 and C17:0 (inverse association with type 2 diabetes), and C14:0 (positive association). In this regard, our results for C17:0 and C14:0 are consistent with the existing literature. It is less clear-cut for C15:0, although there is a weak indication that an inverse association is present.

Even-chain SFAs account for the bulk of the total amount of saturated fatty acids, and they have been linked to an increased risk of type 2 diabetes in several studies (e.g. Forouhi et al., 2014, Lu et al., 2018). Our results for C14:0 and C16:0 support this link, but no evidence of association was found for C18:0. We conjecture that the disparity for C18:0 can be explained by our use of transformed fatty acid data. On the raw data scale, increasing the proportion of C18:0 while keeping the proportions of the other SFAs fixed forces the total proportion of non-saturated fatty acids to decrease. On the transformed scale, this corresponds to an increase in *all* of the logratios. In §B.5, we provide an informal calculation that shows how the effects from the other logratios could indicate a positive association for C18:0, even when such an association does not exist. Particularly, the relatively small standard deviation of C16:0 on the transformed scale (Table 2.4) allows its strong positive association to dominate. This suggests that the effects from C18:0 found by previous studies may in fact be mostly due to C16:0 instead.

Comparatively few studies have investigated the association between very-long-chain SFAs and type 2 diabetes. Forouhi et al. (2014) analyzed data from the EPIC-InterAct Project, which incorporates data from 26 studies from 8 different countries in Europe, including the EPIC-Norfolk dataset. This analysis suggested that all four of the very-long-chain SFAs examined here are inversely associated with type 2 diabetes. Our findings for C22:0 differ, instead supporting a positive association, matching the conclusions of Lin (2018) using data from a Chinese population. On the other hand, our results indicate inverse associations for C20:0 and C24:0; this heterogeneity within an SFA group supports the argument that the effect of each SFA should be studied separately.

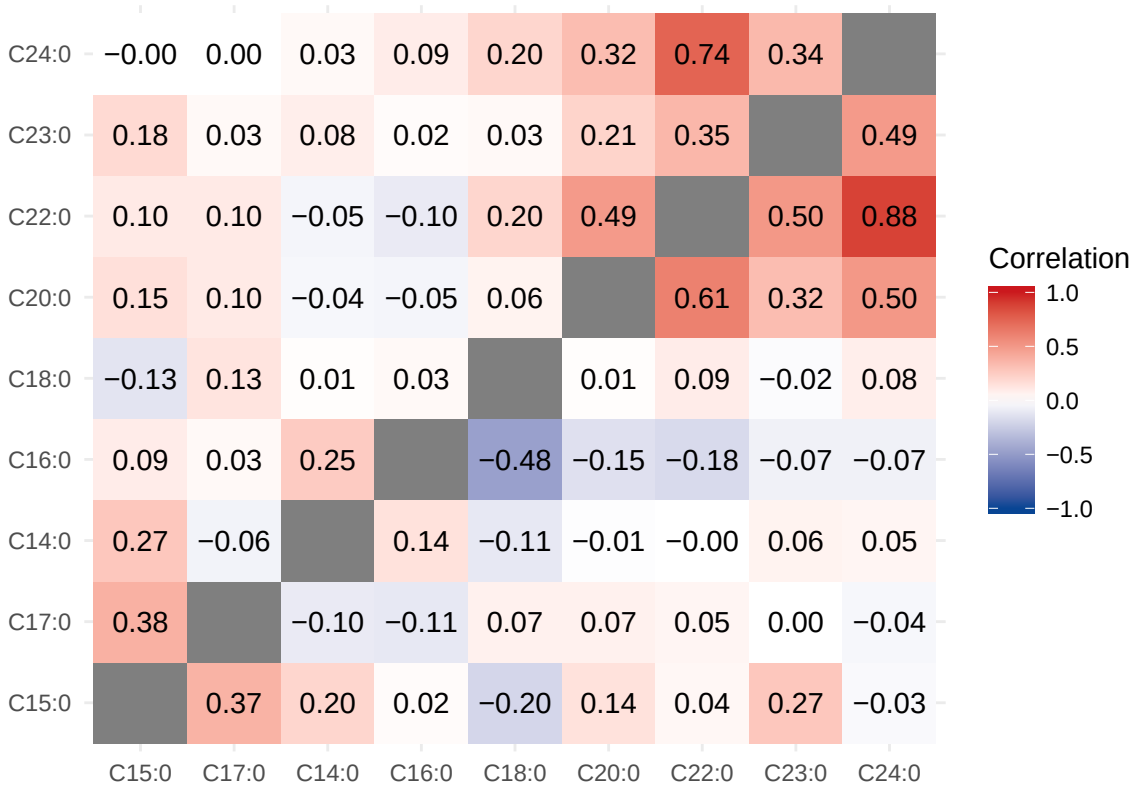


Fig. 2.3 Estimated correlations between the saturated fatty acids using the subcohort data. Values below the diagonal were computed from the raw data; values above the diagonal were computed from the additive logratio transformed data.

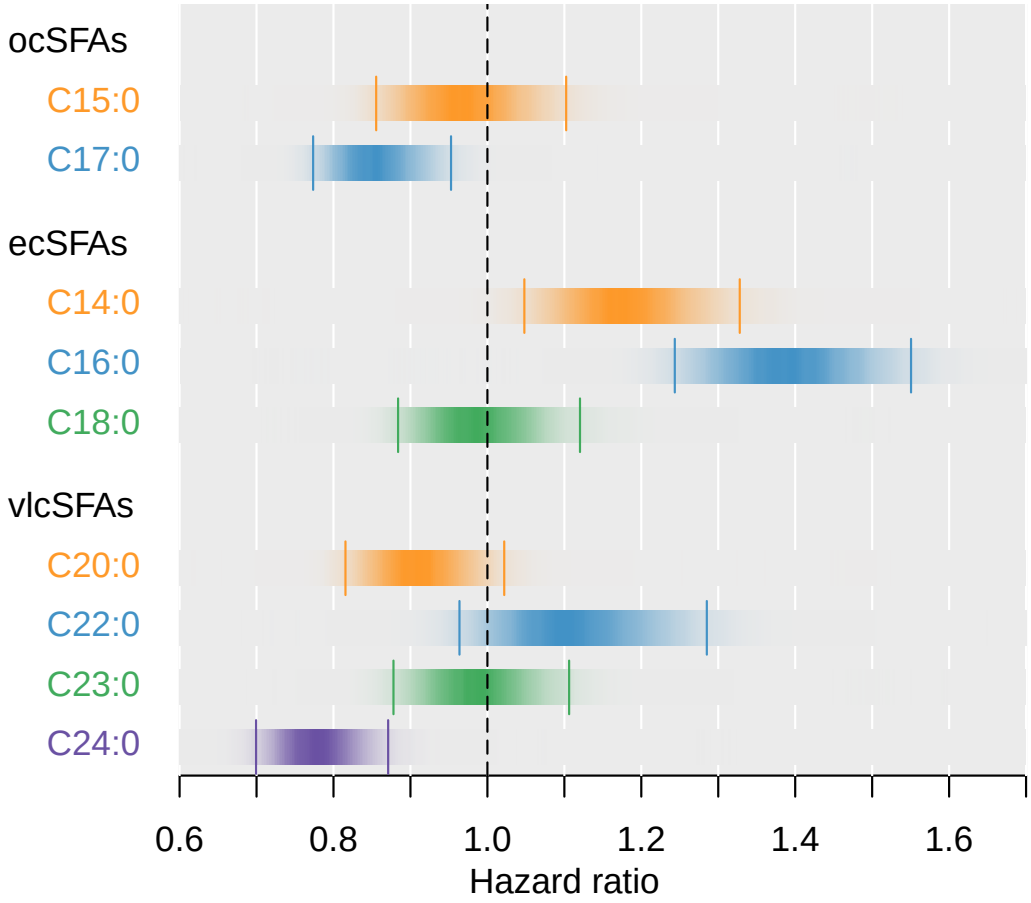


Fig. 2.4 Posterior distributions of the saturated fatty acid hazard ratios. The darkness of the strips is proportional to the posterior density, with the central 95% credible regions indicated. ocSFAs, odd-chain saturated fatty acids; evSFAs, even-chain saturated fatty acids; vlcSFAs, very-long-chain saturated fatty acids.

Table 2.4 Data summaries for the subcohort individuals with complete data, and analysis results. The raw data are expressed as percentages of the total phospholipid fatty acids. SFA, saturated fatty acid; ocSFAs, odd-chain saturated fatty acids; evSFAs, even-chain saturated fatty acids; vlcSFAs, very-long-chain saturated fatty acids; ALR, additive logratio transformed; SD, standard deviation; HR, hazard ratio. "HR 95%" refers to the central 95% credible interval; "$\mathbb{P}(\text{HR} \leqslant 1)$" refers to the posterior probability that the hazard ratio does not exceed 1.

		Raw data	ALR data	Analysis results		
Group	SFA	Mean (SD)	Mean (SD)	HR Mean	HR 95%	$\mathbb{P}(\text{HR} \leqslant 1)$
ocSFAs	C15:0	0.25% (0.07%)	-5.42 (0.27)	0.97	(0.86, 1.10)	0.684
	C17:0	0.43% (0.09%)	-4.86 (0.26)	0.86	(0.77, 0.95)	0.998
ecSFAs	C14:0	0.39% (0.10%)	-4.95 (0.26)	1.18	(1.05, 1.33)	0.003
	C16:0	30.12% (1.54%)	-0.59 (0.07)	1.39	(1.24, 1.55)	0.000
	C18:0	13.97% (1.32%)	-1.36 (0.11)	0.99	(0.88, 1.12)	0.585
vlcSFAs	C20:0	0.16% (0.05%)	-5.89 (0.31)	0.91	(0.82, 1.02)	0.947
	C22:0	0.29% (0.10%)	-5.27 (0.24)	1.11	(0.96, 1.29)	0.074
	C23:0	0.14% (0.07%)	-6.13 (0.70)	0.99	(0.88, 1.11)	0.601
	C24:0	0.24% (0.08%)	-5.44 (0.26)	0.78	(0.70, 0.87)	1.000

2.5 Discussion

This chapter introduced a novel methodology for case-cohort Cox regression. We are able to incorporate auxiliary variables to help predict the missing covariate values and are unrestricted in our choice of prediction model; this differs from multiple imputation (Keogh and White, 2013), which requires careful specification of prediction models to avoid incompatability with the Cox model. The models for the nuisance parameters, including the baseline cumulative hazard function, are nonparametrically specified and then integrated out, facilitating robust and convenient inference. By modifying the basic sampling algorithm, the method scales effectively to datasets with a large sample size and a moderate number of covariates, in contrast to nonparametric maximum likelihood estimation (Zeng and Lin, 2014). We demonstrated this scalability in our analysis of the EPIC-Norfolk study, where we used 19 covariates —5 confounders, 5 auxiliary covariates, and 9 expensive covariates—with a sample size of 22219. Simulations suggest that we obtain substantial efficiency gains over weighted Cox regression approaches (e.g. Prentice, 1986), which are the status quo in practice. As part of our analysis of the EPIC-Norfolk study data, we also developed a new

approach for handling compositional data in the Cox model that provides more reliable and interpretable results compared to previous studies.

There is ample scope to extend our framework. We have assumed that the covariates are time-independent since this was sufficient for our application, where only baseline measurements were available. This assumption can be relaxed by building on the results of Sinha et al. (2003), which provided a Bayesian justification of the Cox partial likelihood in various settings.

The nested case-control design (Thomas, 1977) is similar to the case-cohort design in the sense that full covariate measurements are obtained for all cases, but only for a sample of controls. Like the nonparametric maximum likelihood approach of Scheike and Juul (2004), Scheike and Martinussen (2004) and Zeng and Lin (2014), it is straightforward to adapt our method to the nested case-control design under similar assumptions. Generalizing our method to other survival models like Zeng and Lin (2014) is an area for future research.

Another important direction for further work is variable selection. Existing proposals are few in number and revolve around weighted Cox regression (Ni et al., 2016, Newcombe et al., 2018). Extending our framework to perform variable selection will not only allow more efficient use of data, but also has the advantage of adopting the principled Bayesian approach to variable selection (Clyde and George, 2004).

Chapter 3

Inference under unequal probability sampling with the Bayesian exponentially tilted empirical likelihood

3.1 Introduction

In this chapter, we develop an inferential framework for estimation in the presence of unequal probability sampling. We work under two settings. The first, which we refer to as the design setting, assumes a selection mechanism determined by the data collector, but only partial design information in the form of sampling probabilities for the selected individuals is provided to the analyst. This is frequently encountered when analyzing public-use survey datasets (Zanganeh and Little, 2015, Si et al., 2015, Wang et al., 2017). The second is an observational setting where the selection mechanism is unknown but is assumed to be ignorable conditional on a set of fully observed covariates.

We introduced these two settings in §1.1 and §1.2. Recall that in both cases, it is common in practice to use semiparametric estimators that incorporate inverse probability weighting. If the selection probabilities are known, weighting methods are simple to implement and require few modelling assumptions to attain consistency and asymptotic normality. In the observational setting, inverse probability weights estimated from a selection model can be combined with an imputation model to produce doubly robust estimators. If the models are fitted using estimating equations, the resulting doubly robust estimator is consistent as long as one of the models is correctly specified, and attains the semiparametric efficiency bound if both models are correct. Doubly robust estimators also facilitate bias correction, enabling

valid frequentist inference when using (highly regularized) machine learning methods, provided that they converge sufficiently quickly. However, while the large sample properties of these estimators are attractive, their reliability for small or moderately-sized datasets is less justified theoretically. Particularly, the use of inverse probability weighting can lead to disastrous performance in the presence of model misspecificiation and a practical violation of positivity (Kang and Schafer, 2007). For small sample inference, a careful choice of prior distribution in a Bayesian approach can offer both regularization and a systematic way of incorporating informative knowledge into the analysis, with this influence gradually relaxed as the sample size increases. Further motivations for pursuing a Bayesian approach were discussed in detail in §1.2.2.2.

A significant drawback of a standard Bayesian approach is the requirement of stronger structural assumptions on the data distribution and the sampling mechanism. In the design setting, one option is to specify a flexible regression imputation model with the sampling probability included as a covariate to adjust for the selection bias. To obtain estimates for the target population, the sampling probability can be integrated out using a sampling probability model conditional on selection (Zanganeh and Little, 2015, Si et al., 2015). Alternatively, one can use a sample likelihood approach (Pfeffermann et al., 2006) which truncates the dataset to just the sampled individuals and requires the specification of a conditional selection model. Both approaches involve directly modelling the dependence structure between the incomplete data and the sampling probabilities, rather than using the unavailable design variables specified by the data collector. This is potentially difficult to specify correctly. Moreover, by including the sampling probabilities in a conditional approach, the interpretation of the target quantities—e.g. regression coefficients—can become obscured.

The difficulties in the observational setting are exemplified by the Robins-Ritov example discussed in §1.2.1. Conventional Bayesian estimators will generally fail to be doubly robust; either the selection mechanism is ignored, or the model parameters are a priori dependent, such that misspecification of just one model can feed back into the other, precluding consistency (Zigler et al., 2013, Robins et al., 2015). We have seen in Chapter 2 that ignoring the selection can be beneficial for efficiency, but one may wish to protect against model misspecification by leveraging the selection data to make weaker modelling assumptions.

Our framework offers the practical benefits of Bayesian statistics, along with the attractive asymptotic guarantees of frequentist semiparametric estimators. Central to our approach is a novel application of Bayesian exponentially tilted empirical likelihood (Schennach, 2005), a methodology that forms a posterior by combining a prior with a likelihood function defined by moment conditions. We specialize to the domain of Z-estimation since many

proposed semiparametric estimators (e.g. Hájek, 1971, Robins et al., 1994, Scharfstein et al., 1999, Cao et al., 2009, Rotnitzky et al., 2012) are Z-estimators, and the unbiased estimating equations they solve are used to define a set of corresponding moment constraints. We prove Bernstein–von Mises theorems showing that the resulting Bayesian exponentially tilted empirical likelihood posterior becomes approximately normal, centred at the chosen estimator with matching asymptotic variance; the choice of prior is unrestricted, outside of continuity and non-zero mass in a neighbourhood of the probability limit of the estimator. Thus, the posterior shares analogous properties of the estimator, such as double robustness and local efficiency, and the frequentist coverage of any credible set will be approximately equal to its credibility. In particular, the latter implication extends the large-sample posterior properties proved by Chib et al. (2018) and provides an interpretation of the credible sets as regularized or shrinkage estimators of confidence sets, filling a conceptual gap otherwise left empty due to the procedure not being fully Bayesian.

Additionally, we prove that a separation condition, similar to what is required by Theorem 1 of Chib et al. (2018), is implied under standard assumptions for the consistency of Z-estimators. This allows the user to avoid a potentially difficult verification. Schennach (2005) provided an interpretation of Bayesian exponentially tilted empirical likelihood which justifies its use as a Bayesian procedure. However, the conditions of this result are not satisfied in our design setting in §2.3. We establish an alternative interpretation, connecting the likelihood function with a proper likelihood arising from an exponential family of maximum entropy distributions and suggest that this paves the way for future work. Proofs of all results are found in §C.3.

Our approach offers the ability to obtain modified versions of existing estimators with improved properties, even in the absence of informative priors. For example, certain proposed estimators (e.g. Cao et al., 2009) may have a non-zero probability of lying outside of the parameter space, leading potentially to suboptimal finite sample performance (Rotnitzky et al., 2012). This can be rectified by simply restricting the support of the prior, producing a new estimator which is population bounded in accordance with the variation of its predecessor and has identical asymptotic behaviour. Having a posterior distribution also allows the user to have a choice of estimators such as the mean, median, or the maximum a posteriori estimator, depending on the situation or target loss function.

3.2 Proposal

3.2.1 Exponentially tilted empirical likelihood

Suppose that D is a random vector drawn from a distribution P_0. The objective is to estimate $\theta_0 \in \Theta \subset \mathbb{R}^m$, which is assumed to satisfy the moment condition $\mathbb{E}_{P_0}\{g(D,\theta_0)\} = 0$, where g is a function mapping into $\mathbb{R}^m$. Thus, the dimension of the moment condition is assumed to match the dimension of the target quantity. The observed data D_i $(i = 1,\ldots,n)$ are independent and identically distributed replicates of D with realized values d_i. A Z-estimator $\hat{\theta}_n$ solves the estimating equation $n^{-1}\sum_{i=1}^n g(d_i,\theta) = 0$ for $\theta \in \Theta$. Many proposed estimators for unequal probability sampling problems take this form, accompanied with a set of regularity assumptions similar to the following.

Assumption 3.1. *(i) The parameter space Θ of θ is compact, θ_0 lies in the interior of Θ and is the unique solution to $\mathbb{E}_{P_0}\{g(D,\theta)\} = 0$ (ii) with probability 1, there is a unique solution $\hat{\theta}_n$ to $n^{-1}\sum_{i=1}^n g(D_i,\theta) = 0$ for each n (iii) $\Omega_0 = var_{P_0}\{g(D,\theta_0)\}$ is non-singular (iv) $\mathbb{E}_{P_0}\{\sup_{\theta\in\Theta}\|g(D,\theta)\|_2^2\} < \infty$ (v) with probability one, $g(D,\theta)$ is continuous at each $\theta \in \Theta$ (vi) with probability one, $g(D,\theta)$ is continuously differentiable with respect to θ in a neighbourhood Θ' of θ_0 and $\mathbb{E}_{P_0}\{\sup_{\theta'\in\Theta'}\|\partial_\theta g(D,\theta')\|_F\} < \infty$, where ∂_θ denotes the partial derivative with respect to θ and F refers to the Frobenius norm (vii) $G_0 = \mathbb{E}_{P_0}\{\partial_\theta g(D,\theta_0)\}$ is invertible.*

Assumption 3.1 is sufficient for the Z-estimator $\hat{\theta}_n$ to be consistent and asymptotically normally distributed (van der Vaart, 1998)

$$n^{1/2}(\hat{\theta}_n - \theta_0) \to \mathcal{N}(0,\Sigma_0),$$

with convergence in distribution, where $\Sigma_0 = (G_0^{\mathrm{T}}\Omega_0^{-1}G_0)^{-1}$, and G_0 and Ω_0 can be consistently estimated by

$$\hat{G}_n = n^{-1}\sum_{i=1}^n \partial_\theta g(D_i,\hat{\theta}_n) \quad \text{and} \quad \hat{\Omega}_n = n^{-1}\sum_{i=1}^n g(D_i,\hat{\theta}_n)g(D_i,\hat{\theta}_n)^{\mathrm{T}} \tag{3.1}$$

respectively.

The moment condition g can also define a semiparametric model by restricting to distributions P which satisfy $\mathbb{E}_P\{g(D,\theta)\} = 0$ for $\theta \in \Theta$. For values of θ such that the origin lies in the convex hull of $\{g(d_i,\theta) : i = 1,\ldots,n\}$, the exponentially tilted empirical likelihood (Jing and Wood, 1996, Corcoran, 1998, Lee and Young, 1999, Schennach, 2005) is defined,

up to a constant factor, as

$$L_n(\theta) = \prod_{i=1}^{n} n p_i(\theta)$$

where the probabilities $p_1(\theta), \ldots, p_n(\theta)$ solve the optimization problem

$$\max_{p_1,\ldots,p_n} \sum_{i=1}^{n} (-p_i \log p_i) \tag{3.2}$$

subject to

$$\sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i g(d_i, \theta) = 0, \quad p_i \geqslant 0 \quad (i = 1, \ldots, n). \tag{3.3}$$

For other values of θ, $L_n(\theta)$ is set to 0. The function $p(\theta) = (p_1(\theta), \ldots, p_n(\theta))^{\mathrm{T}}$ is well-defined since: (i) for each value of θ, the constraint set is compact and the objective function is continuous, so if the constraint set is non-empty, the objective function attains the maximum and (ii) the objective function is strictly concave, so there is a unique maximizer.

One may interpret this likelihood function as being derived from a θ-parameterized set of categorical distributions supported on the observed data values. For each value of θ, the solution minimizes the Kullback-Leibler divergence to the empirical distribution subject to the constraint $\mathbb{E}_P\{g(D, \theta)\} = 0$. More precisely, the Kullback-Leibler divergence is minimized with the empirical distribution as the second argument; the opposite direction corresponds to the empirical likelihood (Owen, 2001), which replaces (3.2) with

$$\max_{p_1,\ldots,p_n} \sum_{i=1}^{n} \log p_i.$$

This connection mirrors the relationship between variational Bayesian methods and expectation propagation (Gelman et al., 2013). The exponentially tilted empirical likelihood is connected to Z-estimation as follows.

Proposition 3.1. *The Z-estimator $\hat{\theta}_n$ maximizes the exponentially tilted empirical likelihood.*

Furthermore, we show that Assumption 3.1 is sufficient to establish the following separation property, which illustrates that L_n decays exponentially to 0 outside of any ball around $\hat{\theta}_n$.

Theorem 3.1. *If Assumption 3.1 is satisfied, then for any $\delta > 0$, there exists an $\varepsilon > 0$ such that*

$$\sup_{\|\theta - \hat{\theta}_n\|_2 \geqslant \delta} \frac{L_n(\theta)}{L_n(\hat{\theta}_n)} \leqslant \exp\{-\varepsilon (n-1)^{1/2}\}$$

occurs with probability approaching 1.

3.2.2 Bayesian exponentially tilted empirical likelihood

From a Bayesian perspective, Schennach (2005) proposed that the exponentially tilted empirical likelihood can be combined with a prior $p(\theta)$ to form a posterior

$$p(\theta \mid d_1, \ldots, d_n) \propto L_n(\theta) p(\theta)$$

and referred to this approach as Bayesian exponentially tilted empirical likelihood. Schennach justified this by proving that, if all observed data values are distinct, $L_n(\theta)$ can be represented as a limit

$$L_n(\theta) = \lim_{\varepsilon \to 0} \lim_{B \to \infty} \int \left\{ \prod_{i=1}^{n} p(d_i \mid \xi_B) \right\} p(\xi_B \mid \theta; \varepsilon) \, d\xi_B,$$

suggesting that it has a proper probabilistic interpretation as a likelihood derived from a semiparametric model after marginalizing an infinite dimensional nuisance parameter. The prior for the nuisance parameter $\xi_B = (\xi_{B,1}, \ldots, \xi_{B,B})^{\mathrm{T}}$ conditional on θ and a positive real number ε is a distribution on a grid of values such that the induced mixture of uniform densities centred on the components of ξ_B satisfy the moment restrictions within a tolerance ε, favouring mixtures with small support. Conditional on ξ_B, D_i is distributed according to the corresponding mixture of uniform densities. As $B \to \infty$, the spacing of the grid of values tends to zero and the range tends to infinity. Chib et al. (2018) further proved Bernstein–von Mises results, showing that the total variation distance between the posterior distribution of $n^{1/2}(\theta - \theta_0)$ and the normal distribution $\mathcal{N}(0, \Sigma_0)$ tends to zero under correctly specified moment constraints.

We specialize to the domain of Z-estimation, and establish a Bernstein-von Mises theorem with centring point equal to the Z-estimator $\hat{\theta}_n$. This implies that the posterior is not only consistent and asymptotically normal, but frequentist coverage of any credible set will be approximately equal to its credibility, extending the properties implied by Chib et al. (2018). We specify a distinct set of further assumptions.

If $L_n(\theta)$ is non-zero, the optimization problem specified by (3.2) and (3.3) can be solved (Schennach, 2007) by considering the dual problem

$$p_i(\theta) = \frac{\exp\{\hat{\lambda}_n(\theta)^{\mathrm{T}} g(d_i, \theta)\}}{\sum_{j=1}^{n} \exp\{\hat{\lambda}_n(\theta)^{\mathrm{T}} g(d_j, \theta)\}} \tag{3.4}$$

where $\hat{\lambda}_n(\theta)$ solves

$$\sum_{i=1}^{n} \exp\{\lambda^{\mathrm{T}} g(d_i, \theta)\} g(d_i, \theta) = 0. \tag{3.5}$$

Assumption 3.2. *There exists a neighbourhood $\mathscr{B}$ of θ_0 on which, with probability approaching 1, the exponentially tilted empirical likelihood is non-zero, or equivalently, there exists a function $\hat{\lambda}_n : \mathscr{B} \to \mathbb{R}^m$ satisfying, for all $\theta \in \mathscr{B}$,*

$$\sum_{i=1}^{n} exp\{\hat{\lambda}_n(\theta)^{\mathrm{T}} g(d_i, \theta)\} g(d_i, \theta) = 0.$$

Assumption 3.3. *For almost all values of d, $g(d, \theta)$ is twice differentiable with respect to θ in a neighbourhood of θ_0, and the second derivative satisfies a Lipschitz condition*

$$\left\| \partial_\theta^2 g(d, \theta) - \partial_\theta^2 g(d, \theta') \right\|_{op} \leqslant \psi(d) \left\| \theta - \theta' \right\|_2$$

for an integrable function ψ, where op refers to the operator norm.

Assumption 3.4. *For almost all values of d, there exists a neighbourhood of $(0, \theta_0)$ contained in $\mathbb{R}^m \times \Theta$ in which the function*

$$f(\lambda, \theta) = \exp\{\lambda^{\mathrm{T}} g(d, \theta)\} g(d, \theta)$$

and all of its first and second partial derivatives are dominated by an integrable function.

These allow us to establish the following intermediate result.

Proposition 3.2. *If Assumptions 3.3 and 3.4 are satisfied, on a neighbourhood of θ_0, there exists a unique function λ_0 mapping into $\mathbb{R}^m$ satisfying*

$$\mathbb{E}_{P_0}[\exp\{\lambda_0(\theta)^{\mathrm{T}} g(D, \theta)\} g(D, \theta)] = 0$$

and λ_0 is twice Lipschitz differentiable.

Consequently, one can generate an exponential family $\{P_\theta\}$ from P_0

$$\frac{dP_\theta}{dP_0}(d) = \exp\{\lambda_0(\theta)^{\mathrm{T}} g(d, \theta) - \kappa(\theta)\}$$

locally around θ_0, where $\kappa(\theta) = \log \mathbb{E}_{P_0}[\exp\{\lambda_0(\theta)^{\mathrm{T}} g(D, \theta)\}]$ and $\mathbb{E}_{P_\theta}\{g(D, \theta)\} = 0$. The exponentially tilted distribution P_θ is the I-projection (Csiszár, 1975) of P_0 onto the set

$\{P : \mathbb{E}_P\{g(D,\theta)\} = 0\}$, i.e. the closest element to P_0 in the set in terms of Kullback-Leibler divergence. In this local region of θ_0, the exponentially tilted empirical likelihood is approximately equal to the likelihood generated by this exponential family. This suggests that the exponentially tilted empirical likelihood is a plug-in estimate of a least favourable family of distributions aimed at reducing the original semiparametric model to a parametric model in a minimally informative way. This offers a general interpretation of the Bayesian exponentially tilted empirical likelihood methodology which holds even in certain situations where the Schennach (2005) interpretation does not apply, such as the design setting in §2.3 where the set of observed data values may not be distinct.

Theorem 3.2. *Suppose that Assumptions 3.1–3.4 hold. Suppose also that the prior $p(\theta)$ admits a continuous density with respect to the Lebesgue measure and is positive at θ_0. Then*

$$\int_\Theta \left| p(\theta \mid D_1,\ldots,D_n) - p_{\hat{\theta}_n, n^{-1}\Sigma_0}(\theta) \right| d\theta \to 0$$

with convergence in probability, where $p_{\hat{\theta}_n, n^{-1}\Sigma_0}$ is the density of $\mathcal{N}(\hat{\theta}_n, n^{-1}\Sigma_0)$.

By centring and scaling, and using an alternative form of the total variation distance (Tsybakov, 2009), we have the equivalent representation of

$$\sup_B \left| \mathbb{P}\{n^{1/2}(\theta - \hat{\theta}_n) \in B \mid D_1,\ldots,D_n\} - \mathcal{N}(0,\Sigma_0)(B) \right| \to 0$$

with convergence in probability, where B ranges over all elements of the Borel sigma-algebra on $\mathbb{R}^m$. Theorem 3.2 implies both posterior consistency and asymptotically correct frequentist coverage of credible sets. The following result confirms the first-order equivalence of the posterior mean and $\hat{\theta}_n$, establishing the validity of the methodology as a shrinkage estimation framework that can produce finite sample gains, while matching the asymptotic performance of the standard estimator.

Theorem 3.3. *Suppose that Assumptions 3.1–3.4 hold and $\int \|\theta\|_2 \, p(\theta)\, d\theta < \infty$. Let $\theta_n^* = \int \theta \, p(\theta \mid d_1,\ldots,d_n)\, d\theta$ be the Bayesian exponentially tilted empirical likelihood posterior mean. Then*

$$n^{1/2}(\hat{\theta}_n - \theta_n^*) \to 0 \quad \textit{and} \quad n^{1/2}(\theta_n^* - \theta_0) \to \mathcal{N}(0,\Sigma_0).$$

with convergence in probability and distribution respectively.

3.2.3 Design setting

We first consider a design setting where the selection probabilities are known for the sampled individuals. The data $D_i = (R_i Z_i, R_i, R_i \pi_i)$ $(i = 1, \ldots, n)$ are independent and identically distributed from P_0; R_i is the selection indicator which is equal to 1 if Z_i is observed and 0 otherwise, and $\pi_i = \mathbb{P}(R_i = 1 \mid W_i)$, where Z_i and R_i are conditionally independent given W_i. The variables $W_1, \ldots, W_n$ are the design variables chosen by the data collector to assign sampling probabilities to individuals in the target population, but are not included in the dataset. We make the positivity assumption that there exists a $\delta > 0$ such that $\pi_i \geqslant \delta$ with probability 1. The target quantity θ_0 is the unique solution to $\mathbb{E}_{P_0}\{u(Z, \theta)\} = 0$ for a function u and $\theta \in \Theta \subset \mathbb{R}^m$. The full data estimating function u is adapted below to the estimating function g for the observed data, allowing us to apply Theorem 3.2.

Example 3.1 (Outcome mean)**.** $Z = Y$, $u(Z, \theta) = Y - \theta$.

Example 3.2 (Linear regression)**.** $Z = (Y, X)$, $u(Z, \theta) = X^{\mathrm{T}}(Y - X\theta)$.

Consider the estimator $\hat{\theta}_n$ which solves the estimating equation

$$\sum_{i=1}^{n} \frac{R_i}{\pi_i} u(Z_i, \theta) = 0.$$

To address the technicality that the sampling probabilities are provided as $R_i \pi_i$ in the notation rather than just π_i, we set $R_i/(R_i\pi_i) = 0$ when $R_i = 0$, so that $R_i/(R_i\pi_i)$ is equivalent to R_i/π_i. In the case of estimating the population outcome mean, this estimator specializes to the Hájek estimator (Hájek, 1971). For $D = (RY, R, R\pi) \sim P_0$ and $g(D, \theta) = Ru(Z, \theta)/\pi$,

$$\begin{aligned}
\mathbb{E}_{P_0}\{g(D, \theta)\} &= \mathbb{E}_W \mathbb{E}_{P_0|W} \left\{ \frac{R}{\pi} u(Z, \theta) \mid W \right\} \\
&= \mathbb{E}_W \left[\frac{\mathbb{E}_{P_0|W}(R \mid W)}{\pi} \mathbb{E}_{P_0|W}\{u(Z, \theta) \mid W\} \right] \\
&= \mathbb{E}_{P_0}\{u(Z, \theta)\}
\end{aligned}$$

where we have used the conditional independence of R and Z conditional on W and the equality of $\mathbb{E}_{P_0|W}\{R \mid W\}$ and π. This shows that θ_0 is the unique solution to $\mathbb{E}_{P_0}\{g(D, \theta)\} = 0$. Let $L_n(\theta)$ be the exponentially tilted empirical likelihood function corresponding to the moment conditions $\mathbb{E}_{P_0}\{g(D, \theta)\} = 0$, for $\theta \in \Theta$. The likelihood function is combined with a user-specified prior $p(\theta)$ to form a posterior

$$p(\theta \mid d_1, \ldots, d_n) \propto L_n(\theta) p(\theta). \tag{3.6}$$

If Assumptions 3.1–3.4 are satisfied and $p(\theta)$ is continuous and non-zero around θ_0, Theorem 3.2 implies that

$$\int_\Theta \left| p(\theta \mid D_1, \ldots, D_n) - p_{\hat{\theta}_n, n^{-1}\Sigma_0}(\theta) \right| d\theta \to 0$$

with convergence in probability, where $p_{\hat{\theta}_n, n^{-1}\Sigma_0}$ is the density of $\mathcal{N}(\hat{\theta}_n, n^{-1}\Sigma_0)$ and $\Sigma_0 = \lim_{n\to\infty} \mathrm{var}_{P_0}(n^{1/2}\,\hat{\theta}_n)$. Since $\hat{\theta}_n$ is a consistent estimator of θ_0, the posterior will concentrate around θ_0 as n gets large. Furthermore, since Σ_0 is equal to the asymptotic variance of $n^{1/2}\,\hat{\theta}_n$, the frequentist coverage of any credible set will be approximately equal to its credibility.

3.2.4 Observational setting

In this subsection, we work in a setting where the selection mechanism is unknown. The observed data $D_i = (R_i Z_i, R_i, W_i)$ $(i = 1, \ldots, n)$ are independent and identically distributed from P_0; Z_i and R_i are as before, and W_i is a vector of covariates observed for each i such that Z_i and R_i are conditionally independent given W_i. The target quantity γ_0 is the unique solution to $\mathbb{E}_{P_0}\{u(Z, \gamma)\} = 0$ for a function u and values of γ belonging to a compact real subset Γ. In a missing data context, the conditional independence of Z_i and R_i is sometimes referred to as a *missing at random* assumption. This set-up may also be viewed as one arm of a point exposure causal inference problem in the potential outcomes framework, with the conditional independence corresponding to an assumption of no unmeasured confounders.

Let $\pi_0(W) = \mathbb{P}(R = 1 \mid W)$ be the true propensity score and let $\phi_0(W, \gamma) = \mathbb{E}_{P_0}\{u(Z, \gamma) \mid W\}$. We make the positivity assumption that there exists a $\delta > 0$ such that $\pi_0(W) \geqslant \delta$ with probability 1. Solving

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{R_i u(Z_i, \gamma)}{\hat{\pi}(W_i)} - \hat{\phi}(W_i, \gamma)\left\{\frac{R_i}{\hat{\pi}(W_i)} - 1\right\}\right] = 0, \tag{3.7}$$

where $\hat{\pi}$ and $\hat{\phi}$ are estimators of π_0 and ϕ_0 respectively, leads to a doubly robust estimator of γ; that is, it is consistent and asymptotically normal as long as at least one of $\hat{\pi}$ and $\hat{\phi}$ is consistent. There is a significant body of work regarding choices for $\hat{\pi}$ and $\hat{\phi}$, particularly for population outcome mean estimation (previously discussed in detail in Chapter 1), which lead to various favourable efficiency properties. See Kang and Schafer (2007) and Rotnitzky and Vansteelandt (2014) for comprehensive reviews.

If $\hat{\pi}$ and $\hat{\phi}$ are derived from the solutions to unbiased estimating equations, as is often the case in practice, we can exploit this to formulate a set of nested moment constraints for an exponentially tilted empirical likelihood model. We show in Theorem 3.4 that the

resulting marginal posterior distribution of γ is calibrated asymptotically to the behaviour of the selected estimator.

We restrict our attention to parametric working models $\pi(W;\alpha)$ and $\phi(W,\gamma;\beta)$ for real-valued parameters α and β. Suppose $(\hat{\alpha}_n, \hat{\beta}_n, \hat{\rho}_n)$ solve the unbiased estimating equation

$$\frac{1}{n}\sum_{i=1}^{n} U_{\alpha,\beta,\rho}(D_i, \alpha, \beta, \rho) = 0$$

where ρ is a set of additional auxiliary parameters, possibly empty (Rotnitzky and Vansteelandt, 2014). The two parameters (α, β) can be estimated either separately or together. For example, in the case of mean estimation, Robins et al. (1994) estimate α with maximum likelihood for a logistic regression model, and estimate β separately using ordinary least squares. Scharfstein et al. (1999) also use maximum likelihood estimation for α, but include the reciprocal of the propensity score as a covariate in the outcome regression model.

Let $\hat{\gamma}_n$ be the solution to

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{R_i u(Z_i,\gamma)}{\pi(W_i,\hat{\alpha}_n)} - \phi(W_i,\gamma;\hat{\beta}_n)\left\{\frac{R_i}{\pi(W_i,\hat{\alpha}_n)} - 1\right\}\right] = 0. \tag{3.8}$$

Let $\theta = (\alpha, \beta, \rho, \gamma)$ and define $g(D,\theta) = (U_{\alpha,\beta,\rho}(D,\alpha,\beta,\rho)^{\mathrm{T}}, h(D,\alpha,\beta,\gamma)^{\mathrm{T}})^{\mathrm{T}}$, where

$$h(D,\alpha,\beta,\gamma) = \frac{Ru(Z,\gamma)}{\pi(W;\alpha)} - \phi(W,\gamma;\beta)\left\{\frac{R}{\pi(W;\alpha)} - 1\right\}.$$

In accordance with Assumption 3.1(i), we assume that there exists a value $\theta_0 = (\alpha_0, \beta_0, \rho_0, \gamma^*)$ which is the unique solution to $\mathbb{E}_{P_0}\{g(D,\theta)\} = 0$. We say that the working model for the propensity score is correctly specified if $\pi_0(W) = \pi(W;\alpha_0)$ and similarly that the model for ϕ is correctly specified if $\phi_0(W,\gamma) = \phi(W,\gamma;\beta_0)$. If at least one is correctly specified, $\gamma^* = \gamma_0$ and the Z-estimator $\hat{\gamma}_n$ consistently estimates the truth.

Let $L_n(\theta)$ be the exponentially tilted empirical likelihood function corresponding to the moment conditions $\mathbb{E}_P\{g(D,\theta)\} = 0$. The likelihood function is combined with a user-specified prior $p(\theta)$ to form a posterior

$$p(\theta \mid d_1, \ldots, d_n) \propto L_n(\theta) p(\theta).$$

Let $p(\gamma \mid d_1, \ldots, d_n)$ be the marginal posterior for γ.

Theorem 3.4. *Suppose that Assumptions 3.1–3.4 hold and that the prior $p(\theta)$ admits a continuous density with respect to the Lebesgue measure and is positive at θ_0. Then as $n \to \infty$,*

$$\int_\Gamma \left| p(\gamma \mid D_1, \ldots, D_n) - p_{\hat{\gamma}_n, n^{-1}V_0}(\gamma) \right| d\gamma \to 0$$

with convergence in P_0-probability, where $p_{\hat{\gamma}_n, n^{-1}V_0}$ is the density of $\mathcal{N}(\hat{\gamma}_n, n^{-1}V_0)$ and $V_0 = \lim_{n\to\infty} var_{P_0}(n^{1/2}\hat{\gamma}_n)$.

As stated earlier, $\hat{\gamma}_n$ is, by construction, consistent for estimating γ_0 provided either $\pi_0(W) = \pi(W;\alpha_0)$ or $\phi_0(W,\gamma) = \phi(W,\gamma;\beta_0)$ for all γ or both. Therefore, Theorem 3.4 implies that the exponentially tilted empirical likelihood posterior shares this double robustness property; the posterior will concentrate around the true value as long as one of the working models is correctly specified. Furthermore, credible sets for γ will asymptotically have nominal frequentist coverage if consistency holds, even if one of the working models is misspecified. If both models are misspecified, the credible sets will have approximately nominal coverage for the probability limit γ^* of $\hat{\gamma}_n$, which is possibly different from γ_0.

3.2.5 Implementation

We describe below how one can compute $L_n(\theta)$ for a fixed value of θ. To simplify notation, let $g_i = g(d_i, \theta)$ for each $i = 1, \ldots, n$, suppressing dependence on θ. To check whether the feasible set of the optimization problem specified by (3.2) and (3.3) is non-empty, it is sufficient and computationally convenient, via an R package like lpSolve (Berkelaar, 2015) for example, to check whether there exists a feasible solution to the linear programming problem

$$\begin{aligned} &\text{maximize:} \quad 0 \ \text{ over } \ \{x \in \mathbb{R}^n : \ 0 \leqslant x_i \leqslant 1, \ i = 1, \ldots, n\} \\ &\text{subject to:} \quad g^{\mathrm{T}} x = 0 \quad \text{and} \quad c^{\mathrm{T}} x = 1 \end{aligned} \tag{3.9}$$

where $g = (g_1, \ldots, g_n)$ and $c = (1, \ldots, 1)^{\mathrm{T}}$. The objective 0 is suggested here for computational simplicity, but can be replaced by $b^{\mathrm{T}}x$ for any arbitrary $b \in \mathbb{R}^n$ as we are only concerned with the feasible set. If the feasible set is empty, $L_n(\theta)$ is set to zero. Otherwise, assuming the solution to (3.2) and (3.3) lies in the interior of the simplex, i.e. all of the values of p_i are non-zero, the optimization problem can be solved by considering the dual problem described by (3.4) and (3.5).

Assuming that $\sum_{i=1}^n g_i g_i^{\mathrm{T}}$ is strictly positive definite, a unique solution to (3.5) exists and it can be found using the Newton–Raphson method. This requires specifying a small convergence tolerance value with respect to a norm of choice. Pseudo-code for evaluating

$L_n(\theta)$ is provided in §C.1. Once we are able to evaluate L_n pointwise, we can perform posterior inference using standard Bayesian computational machinery such as Markov chain Monte Carlo or importance sampling.

3.3 Simulations

3.3.1 Mean estimation for binary outcomes

In this simulation, we consider estimating the population mean of binary outcomes in a design setting. In the notation of §3.2.3: $Z = Y$, $u(Z,\theta) = Y - \theta$ and $\theta_0 = \mathbb{E}_{P_0}(Y)$. The design variables W_i $(i = 1,\dots,n)$ are independent and identically distributed according to the beta distribution $\text{Beta}(1.5, 3.5)$, and the outcomes $Y_i \mid W_i \sim \text{Ber}(W_i)$ so that $\theta_0 = 0.3$. The selection variables R_i $(i = 1,\dots,n)$ are independent and identically distributed according to $R_i \mid \pi_i \sim \text{Ber}(\pi_i)$, where $\text{logit}(\pi_i) = W_i$. Thus, Y_i and the selection probability π_i are positively correlated, and the selection must be adjusted for to estimate θ_0. The data available for analysis are $D_i = (R_i Y_i, R_i, R_i \pi_i)$ $(i = 1,\dots,n)$, so that the design variables are excluded.

Following the approach in §3.2.3, the Z-estimator $\hat{\theta}_n$ is the Hájek estimator which solves

$$\sum_{i=1}^{n} g(D_i,\theta) = \sum_{i=1}^{n} \frac{R_i}{\pi_i}(Y_i - \theta) = 0.$$

We use g to define the exponentially tilted empirical likelihood $L_n(\theta)$, which we combine with three different priors for θ: $\theta \sim \text{Beta}(0.5, 0.5)$, $\theta \sim U(0,1)$ and $\theta \sim \text{Beta}(1.5, 3.5)$. The first is Jeffrey's prior. The mean of the $\text{Beta}(1.5, 3.5)$ prior is equal to θ_0, so we consider this prior as informative, while the first two are considered noninformative.

We compare this approach with the proposal in §2 of Wang et al. (2017). In a survey inference context, they suggest a Bayesian approach using an approximate normal likelihood

$$\hat{\theta} \mid \theta \sim \mathcal{N}(\theta, \hat{V})$$

where $\hat{\theta}$ is a consistent and asymptotically normal estimator of θ_0 and $\hat{V}$ is a robust estimator of the variance of $\hat{\theta}$. The estimator $\hat{\theta}$ acts as a summary statistic for the data, such that the posterior is

$$p(\theta \mid \hat{\theta}) \propto p(\hat{\theta} \mid \theta) p(\theta)$$

where $p(\hat{\theta} \mid \theta)$ is defined by the normal model above, and $p(\theta)$ is a prior for θ. We choose $\hat{\theta}$ to be the Hájek estimator defined above and we estimate its variance with the nonparametric bootstrap. We also use the same priors as defined above.

Table 3.1 compares the frequentist estimator $\hat{\theta}_n$ with the Bayesian methods. Each setting was replicated 2000 times. The Bayes point estimators are the posterior means. Coverage rates were computed based on central 95% credible regions. The Bayesian computation was carried out using importance sampling with 5000 particles for each replication and the tolerance for computing the exponentially tilted empirical likelihood was 10^{-4}.

The Bayesian exponentially tilted empirical likelihood estimators generally have a higher magnitude of bias than the normal approximation when a noninformative prior is used, but lower with the informative prior. This reflects the fact that the exponentially tilted empirical likelihood is less informative than the normal likelihood, resulting in higher shrinkage towards the prior mean. In the case of the two noninformative priors, this causes an upward bias towards the prior mean 0.5. This conservative characteristic leads to the Bayesian exponentially tilted empirical likelihood approach having superior performance in terms of root mean squared error and coverage rate across almost all settings, and particularly with the smaller sample sizes when the normal approximation is less accurate.

3.3.2 Doubly robust mean estimation with missing data

This simulation scenario works under the observational setting described in §3.2.4 and follows the design of Kang and Schafer (2007). For each i $(i = 1, \ldots, n)$, the vector of covariates $W_i = (W_{i1}, W_{i2}, W_{i3}, W_{i4}) \sim \mathcal{N}(0, I_4)$, where I_4 is the 4×4 identity matrix, and the selection indicator $R_i \mid W_i \sim \text{Ber}\{\pi_0(W_i)\}$ where

$$\pi_0(W_i) = \text{expit}(\alpha_{0,1} + \alpha_{0,2}^{\mathrm{T}} W_i), \quad \alpha_{0,1} = 0,\ \alpha_{0,2} = (-1, 0.5, -0.25, -0.1)^{\mathrm{T}}$$

and $Z = Y$, the outcome, with $Y_i \mid W_i \sim \mathcal{N}\{m_0(W_i), 1\}$ where

$$m_0(W_i) = \beta_{0,1} + \beta_{0,2}^{\mathrm{T}} W_i, \quad \beta_{0,1} = 210,\ \beta_{0,2} = (27.4, 13.7, 13.7, 13.7)^{\mathrm{T}}.$$

We have assumed that Y_i and R_i are conditionally independent given W_i. The data are $D_i = (R_i Y_i, R_i, W_i)$ $(i = 1, \ldots, n)$. In addition to the correctly specified models:

(a) $\pi(w; \alpha) = \mathbb{P}(R = 1 \mid W = w; \alpha) = \text{expit}(\alpha_1 + \alpha_2^{\mathrm{T}} w)$

(b) $m(w; \beta) = \mathbb{E}_{P_0}(Y \mid W = w; \beta) = \beta_1 + \beta_2^{\mathrm{T}} w,$

Table 3.1 Bias, root mean squared error and coverage rate from 2000 Monte Carlo simulations using the Hájek estimator, the Wang et al. normal approximation and the Bayesian exponentially tilted empirical likelihood approach. RMSE, root mean squared error; CR, coverage rate; BETEL, Bayesian exponentially tilted empirical likelihood.

Population size	Prior	Method	Bias ($\times 100$)	RMSE ($\times 100$)	CR (%)
$n = 25$		Hájek	0.17	11.67	
	Jeffrey's	Normal	-1.57	13.06	88.6
		BETEL	1.19	11.79	92.9
	Uniform	Normal	0.72	11.55	91.5
		BETEL	2.30	11.06	94.5
	Beta$(1.5, 3.5)$	Normal	-1.62	10.22	92.3
		BETEL	-0.11	9.18	96.2
$n = 50$		Hájek	0.08	8.27	
	Jeffrey's	Normal	-0.69	8.92	92.1
		BETEL	0.88	8.29	94.7
	Uniform	Normal	0.39	8.28	91.7
		BETEL	1.45	8.18	94.6
	Beta$(1.5, 3.5)$	Normal	-1.06	7.32	92.6
		BETEL	0.11	7.33	95.7
$n = 100$		Hájek	-0.08	6.01	
	Jeffrey's	Normal	-0.23	6.24	92.1
		BETEL	0.51	6.03	94.8
	Uniform	Normal	-0.11	6.03	92.5
		BETEL	0.57	5.87	94.6
	Beta$(1.5, 3.5)$	Normal	-0.75	5.75	92.8
		BETEL	-0.08	5.56	94.9

we also consider the misspecified models:

(c) $\pi(w';\alpha) = \mathbb{P}(R = 1 \mid W' = w';\alpha) = \text{expit}(\alpha_1 + \alpha_2^{\mathrm{T}} w')$

(d) $m(w';\beta) = \mathbb{E}_{P_0}(Y \mid W' = w';\beta) = \beta_1 + \beta_2^{\mathrm{T}} w'$,

where $W_i' = (W_{i1}', W_{i2}', W_{i3}', W_{i4}')$ are transformed covariates with

$$\begin{aligned} W_{i1}' &= \exp(W_{i1}/2) \\ W_{i2}' &= W_{i2}/\{1 + \exp(W_{i1})\} + 10 \\ W_{i3}' &= \{(W_{i1}W_{i3})/25 + 0.6\}^3 \\ W_{i4}' &= (W_{i2} + W_{i4} + 20)^3. \end{aligned}$$

The target quantity is $\mu_0 = \mathbb{E}_{P_0}(Y) = 210$. We adopt the notation m and μ instead of ϕ and γ used in §3.2.4 to match the Kang and Schafer (2007) paper. For the sake of brevity, the estimators and methods described in the rest of this section will be represented in terms of the correct covariates W_i. Under misspecification, the covariates W_i are replaced with W_i' as appropriate.

The doubly robust augmented inverse probability weighted estimator (Robins et al., 1994), sometimes referred to as the standard double robust estimator, is

$$\hat{\mu}_{\mathrm{DR}} = \sum_{i=1}^{n} \frac{1}{n} \left[\frac{R_i Y_i}{\pi(W_i;\hat{\alpha}_n)} - m(W_i;\hat{\beta}_n) \left\{ \frac{R_i}{\pi(W_i;\hat{\alpha}_n)} - 1 \right\} \right]. \tag{3.10}$$

where $\hat{\alpha}_n$ and $\hat{\beta}_n$ are estimated via maximum likelihood estimation, or equivalently, by solving

$$\frac{1}{n}\sum_{i=1}^{n} U_\alpha(D_i,\alpha) = 0, \quad \frac{1}{n}\sum_{i=1}^{n} U_\beta(D_i,\beta) = 0. \tag{3.11}$$

where U_α and U_β are the score equations for the logistic and linear regression models respectively. In this case, the set of additional auxiliary parameters ρ referred to in §3.2.4 is empty.

Saarela et al. (2016) proposed a Bayesian doubly robust approach using the Bayesian bootstrap (Rubin, 1981). A Dirichlet process model is specified for D_i in the limit of the base measure tending to 0. Inference for μ is based on a posterior predictive distribution induced by maximizing expected utility functions. Here, we follow the approach detailed in §6·2 of their paper and choose the utility functions to match the specification of $\hat{\mu}_{\mathrm{DR}}$. More

explicitly, the parameters α and β are linked to the Bayesian bootstrap model via

$$\alpha = \arg\max_{\alpha} \mathbb{E}\{R(\alpha_1 + \alpha_2^{\mathrm{T}} W) - \log[1 + \exp(\alpha_1 + \alpha_2^{\mathrm{T}} W)]\},$$
$$\beta = \arg\min_{\beta} \mathbb{E}\{R(Y - \beta_1 - \beta_2^{\mathrm{T}} W)^2\},$$

corresponding to the maximization of the expected log-likelihoods of the propensity score and outcome regression models respectively under the posterior. The target quantity μ is defined by

$$\mu = \mathbb{E}\left[\frac{RY}{\pi(W,\alpha)} - m(W,\beta)\left\{\frac{R}{\pi(W,\alpha)} - 1\right\}\right].$$

In practice, we sample from the posterior predictive distribution by repeatedly generating uniform Dirichlet weights $\omega = (\omega_1, \ldots, \omega_n)$ and computing $\hat{\mu}_{\mathrm{DR}}$ with the fixed uniform weights $(1/n, \ldots, 1/n)$ replaced with ω in (3.10) and (3.11). Define $\hat{\mu}_{\mathrm{Sa}}$ to be the posterior predictive mean of μ for this method. The Bayesian exponentially tilted empirical likelihood posterior for $\theta = (\alpha, \beta, \mu)$ is obtained by setting $u(Z, \mu) = Y - \mu$ and following the approach described in §3.2.4. We compare this to the doubly robust augmented inverse probability weighted estimator and the Saarela et al. (2016) proposal.

In Table 3.2, "OR correct" refers to use of the correct outcome regression model (a), while "OR incorrect" refers to the use of model (c). Similarly, "PS correct" refers to use of the correct propensity score model (b), while "PS incorrect" refers to the use of model (d). For both Bayesian exponentially tilted empirical likelihood estimators, we use independent flat priors for all working model parameters across all settings. For $\hat{\mu}_{\mathrm{BETEL},1}$, a flat prior is specified for the target quantity μ, while $\hat{\mu}_{\mathrm{BETEL},2}$ is equipped with a weakly informative prior $t_3(210, 1)$. The three parameters (α, β, μ) are a priori independent across all settings. Sampling from the Saarela et al. (2016) posterior can be implemented directly, as described above. The exponentially tilted empirical likelihood was computed with a tolerance of 10^{-4} and posterior samples were drawn using a Metropolis-Hastings algorithm with 2000 iterations, along with an initial 500 burn-in iterations.

The results in Table 3.2 show that $\hat{\mu}_{\mathrm{BETEL},1}$ performs similarly to $\hat{\mu}_{\mathrm{Sa}}$ in all settings. This is expected since the flat prior on all parameters was chosen to be as noninformative as possible. These similarities provide further confirmation that our asymptotic theory is relevant for finite samples. Both $\hat{\mu}_{\mathrm{BETEL},1}$ and $\hat{\mu}_{\mathrm{Sa}}$ significantly outperform $\hat{\mu}_{\mathrm{DR}}$ when both working models are misspecified, suggesting that a Bayesian approach for this problem offers helpful shrinkage even when designed to be noninformative.

In all settings, $\hat{\mu}_{\text{BETEL},2}$ outperforms both $\hat{\mu}_{\text{DR}}$ and $\hat{\mu}_{\text{Sa}}$ in root mean squared error and median absolute error. This illustrates that when substantial prior knowledge of the target quantity is available, the use of this information in our proposed approach leads to better overall performance than the other estimators evaluated.

OR correct, PS correct					OR incorrect, PS correct				
Estimator	Bias	RMSE	MAE	ESD	Estimator	Bias	RMSE	MAE	ESD
$\hat{\mu}_{\text{DR}}$	-0.01	2.55	1.73	2.55	$\hat{\mu}_{\text{DR}}$	0.27	3.61	2.32	3.60
$\hat{\mu}_{\text{Sa}}$	0.01	2.57	1.71	2.57	$\hat{\mu}_{\text{Sa}}$	0.57	3.44	2.31	3.39
$\hat{\mu}_{\text{BETEL},1}$	-0.15	2.55	1.76	2.55	$\hat{\mu}_{\text{BETEL},1}$	0.49	3.81	2.25	3.78
$\hat{\mu}_{\text{BETEL},2}$	-0.14	2.40	1.63	2.40	$\hat{\mu}_{\text{BETEL},2}$	0.48	3.27	2.01	3.24
OR correct, PS incorrect					OR incorrect, PS incorrect				
Estimator	Bias	RMSE	MAE	ESD	Estimator	Bias	RMSE	MAE	ESD
$\hat{\mu}_{\text{DR}}$	-0.01	2.59	1.73	2.59	$\hat{\mu}_{\text{DR}}$	-6.44	38.52	3.64	37.97
$\hat{\mu}_{\text{Sa}}$	-0.09	2.60	1.73	2.60	$\hat{\mu}_{\text{Sa}}$	-4.81	15.41	3.38	14.64
$\hat{\mu}_{\text{BETEL},1}$	-0.22	2.90	1.76	2.89	$\hat{\mu}_{\text{BETEL},1}$	-8.21	18.61	4.21	16.71
$\hat{\mu}_{\text{BETEL},2}$	-0.15	2.43	1.66	2.43	$\hat{\mu}_{\text{BETEL},2}$	-3.51	6.71	3.38	5.72

Table 3.2 Monte Carlo simulations based on 1000 replicates using the standard doubly robust estimator, the Saarela et al. method and the Bayesian exponentially tilted empirical likelihood approach. RMSE, root mean squared error; MAE, median of absolute errors; ESD, empirical standard deviation; DR, double robust; Sa, Saarela et al. (2016) proposal, BETEL, Bayesian exponentially tilted empirical likelihood; OR, outcome regression; PS, propensity score.

3.4 Application

We examine the association between blood pressure and sodium and potassium consumption using data from the National Health and Nutrition Examination Survey 2003–2006. The dataset includes 13957 individuals with full data on the relevant information, and is drawn from the US civilian population from 2003–2006, which we have assumed to be constant during the time period and equal to 300 million. Each observation is associated with a weight variable assumed to be proportional to the reciprocal of the sampling probability of the individual. This follows the example found in §5.2.4 in Lumley (2010).

We work in the design setting described in §3.2.3. The aim is to fit a linear regression model for blood pressure Y on sodium X_1 and potassium X_2 consumption. Age X_3 is also included for deconfounding. The moment condition is

$$g(D,\theta) = RW(Y - \theta_{\text{int}} - X_1\theta_1 - X_2\theta_2 - X_3\theta_3)X$$

where R is the selection indicator variable, W is the weight variable and $X = (1, X_1, X_2, X_3)^{\mathrm{T}}$. We consider the frequentist Z-estimator with standard errors estimated using the sandwich estimator (C.3). For our Bayesian exponentially tilted empirical likelihood proposal, each regression parameter is assigned an independent prior: $\theta_{\text{int}} \sim t_3(100, 1)$, θ_1 and θ_2 follow half-normal distributions on the positive and negative reals respectively, each with scale parameter 1, and $\theta_3 \sim t_3(0, 1)$. The priors for θ_1 and θ_2 reflect the substantial prior evidence that sodium raises blood pressure in humans, and potassium does the opposite. The likelihood was computed with a tolerance of 10^{-4} and posterior samples were drawn using a Metropolis-Hastings algorithm.

Table 3.3 compares the frequentist estimates with the Bayesian exponentially tilted empirical likelihood posterior mean estimates. In addition to the analysis of the full dataset, an analysis of a random sample of 300 samples was also carried out. With the smaller dataset, the frequentist approach leads to a positive estimated value for the effect of potassium on blood pressure. On the other hand, the Bayesian exponentially tilted empirical likelihood approach gives an estimated value much closer to the ones obtained from the full dataset. This is a clear illustration of the significant impact that the use of an informative prior can offer in small-sample inference, as previously argued in §1.2.2.2. The priors specified were not meticulously constructed to reflect all available substantive knowledge. We simply restricted the sign of the regression coefficients and imposed mild shrinkage towards 0 to protect against overestimation of effect sizes, which is known to be a common occurrence in small samples (van Zwet and Cator, 2020). This was sufficient to produce significantly better estimates compared to the frequentist approach. The results of both approaches converge with the increase in sample size, in accordance with theory.

3.5 Discussion

Our contributions in this chapter can be grouped into two areas: practical and conceptual. From the practical perspective, our Bernstein-von Mises-type result provides an asymptotic frequentist justification of BETEL akin to that of regular parametric Bayes—namely, that

Table 3.3 Frequentist estimates and standard errors and the Bayesian exponentially tilted empirical likelihood posterior means and posterior standard deviations. BETEL, Bayesian exponentially tilted empirical likelihood; s.d., standard deviation.

Sample size	Method		θ_{int}	θ_1	θ_2	θ_3
$n = 300$	Frequentist	Estimate	95.10	0.39	0.85	0.54
		Standard error	2.85	0.63	0.94	0.04
	BETEL	Posterior mean	99.31	0.51	-0.56	0.52
		Posterior s.d.	1.22	0.29	0.39	0.03
$n = 13957$	Frequentist	Estimate	99.74	0.80	-0.91	0.50
		Standard error	0.80	0.15	0.19	0.01
	BETEL	Posterior mean	99.82	0.78	-0.89	0.49
		Posterior s.d.	0.39	0.09	0.12	0.01

large-sample BETEL posterior credible regions are approximately confidence regions. This, however, tells us little about how to interpret the BETEL posterior in finite-sample inference. To this end, we suggested that BETEL can be viewed as an approximate Bayes procedure that uses a plug-in estimate of a least favourable parametric family to form the likelihood function.

Empirical likelihood estimators for missing data problems have previously been proposed by Qin and Zhang (2007) and Chan and Yam (2014). Their work provides a convenient framework for integrating multiple working models into a single analysis, extending the doubly robust property to a multiply robust one. The methods are based on maximizing the conditional empirical likelihood of the outcomes and covariates given selection, and thus differs from ours which uses the marginal exponentially tilted empirical likelihood.

As suggested in §3.2.1, the empirical distribution may be viewed as an initial estimate of the true data generating distribution in the exponentially tilted empirical likelihood. From this interpretation, it is natural to ask whether this initial estimate can be improved. While the empirical distribution can be applied very generally, its use may disregard additionally known or assumed structure about the data distribution, such as its support, conditional independencies and smoothness. Nonparametric techniques such as density estimation may offer a way to incorporate this information into the initial estimate. Investigating whether such replacements are advantageous is a topic of further research.

One might argue, however, that plugging in an initial estimate of the data generating distribution should be avoided altogether in a Bayesian modelling framework. Besides the unappealing double-use of the data, one may be concerned that the uncertainty in the initial estimate has not been taken into account in the BETEL posterior. Indeed, as we will see in the next chapter, the failure to propagate this uncertainty can lead to uncalibrated inference for functionals of the data distribution that are not completely determined by the moment conditions. Our solution will be to instead treat this "pre-tilted" distribution as a nuisance parameter and specify a prior for it, allowing the uncertainty to flow through the Bayesian update.

Chapter 4

Moment condition inference with the exponentially tilted Bayesian bootstrap

4.1 Introduction

In the previous chapter, we argued that the method of Bayesian exponentially tilted empirical likelihood (BETEL) is effective at tackling two popular classes of unequal probability sampling problems. Unfortunately, BETEL is still not completely satisfactory, both practically and conceptually. To illustrate our reasons concretely, let us recall our general set-up. We are motivated by the problem of estimating a quantity defined by a moment condition. Let $\mathscr{P}$ be a set of probability measures on a measurable space $(\mathscr{D}, \mathscr{A})$ and let $\theta : \mathscr{P} \to \Theta \subset \mathbb{R}^m$ $(m \in \mathbb{N})$ be a functional defined as the solution to $\mathbb{E}_P\{g(D, \theta(P))\} = 0$ for each $P \in \mathscr{P}$, where $D \sim P$ and g is a real function. We observe an i.i.d. sample $D_1, \ldots, D_n$ drawn from P.

Consider the simplest example where $D = X$, a one-dimensional real-valued random variable, and we wish to estimate the mean[1] of X. In this case, $\mathscr{P}$ is the set of all probability measures on $(\mathbb{R}, \mathscr{B})$ with mean in Θ, where Θ is assumed to be a compact subset of $\mathbb{R}$, and $\mathscr{B}$ is the Borel σ-algebra on $\mathbb{R}$. We define $\theta(P)$ to be the mean of $P \in \mathscr{P}$, or equivalently, $\theta(P)$ is the solution to $\mathbb{E}_P\{g(X, \theta(P))\} = 0$, where $g(X, \theta) = X - \theta$. Suppose that we have informative prior beliefs about θ, which we incorporate into a prior distribution $\pi(\theta)$. The BETEL posterior is

$$p(\theta \mid X_1, \ldots, X_n) \propto L_n(\theta)\pi(\theta),$$

[1] Note that this example differs from the problem of estimating an outcome mean with incomplete data that we have studied in detail in Chapters 1 and 3. Here, we have complete data, so the moment condition (and the overall problem) is far simpler.

where L_n is the exponentially tilted empirical likelihood function with respect to the moment condition defined by g. This posterior allows us to carry out inference on θ with the knowledge that we are correctly calibrated in a frequentist sense, due to the Bernstein-von Mises theorem (Theorem 3.3). But in standard Bayes, the posterior quantifies our uncertainty for *all* aspects of P, not just a particular finite-dimensional parameter. It is quite plausible, for example, that we only have informative prior beliefs about θ, but we are also interested in some other quantity e.g. $P(X < 0)$.

In order to carry out inference on quantities other than θ, our only option is to refer back to the plug-in family $\{P_\theta\}$ that we estimated from the data by exponentially tilting the empirical distribution. For any functional $\eta(P)$, we can induce its posterior distribution with the mapping $\theta \mapsto \eta(P_\theta)$. However, the posterior for η will not generally exhibit the asymptotic frequentist calibration enjoyed by θ, as we illustrate in the following experiment.

Suppose that the true data-generating distribution is $\mathcal{N}(0,1)$, for which $P(X < 0) = 0.5$. We specified a weakly informative prior for θ: $\theta \sim \mathcal{N}(0,16)$. To obtain posterior samples for $P(X < 0)$, we used the mapping $\theta \mapsto P_\theta(X < 0)$, i.e. the sum of the weights of P_θ corresponding to the data points that are less than 0. Across 1000 iterations for different values of n, we recorded the proportion of central 95% credible intervals that contained the truth 0.5. The results can be found in Table 4.1.

Table 4.1 Coverage of $P(X < 0)$ central 95% credible intervals

	$n = 30$	$n = 50$	$n = 70$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
Coverage	88.4%	87.7%	87.2%	86.4%	86.3%	88.1%	87.0%

The results suggest that the coverages of the BETEL intervals do not converge to the nominal level. We can attribute this to the plug-in estimate of the likelihood that is treated as an a priori truth; the uncertainty of this estimate is not taken into account in the posterior. This calibration problem is avoided in the case of θ as a result of the exponential tilting in its least favourable direction, but this does not necessarily provide any guarantees for other quantities.

Even if we are only interested in θ and are therefore unconcerned by the above dilemma, we may find the idea of a plug-in likelihood conceptually unattractive from a Bayesian point of view. The plug-in principle is inherently frequentist, and applying it in a Bayesian framework involves using the data twice. A more principled Bayesian approach for handling nuisance parameters is to specify priors for the nuisance parameters and integrate them out.

The dependence of $\{P_\theta\}$ on the data also means that BETEL lacks *coherence* in the sense of Bissiri et al. (2016); that is, the form of the BETEL posterior depends on the order that we update the data in. This may in fact have important practical consequences, as previously discussed in §1.2.2.3 in the context of methods that use estimated weights in a Bayesian set-up, e.g. Ray and van der Vaart (2018), Hahn et al. (2020). The form of the BETEL posterior will depend on the order that we receive the data, which can possibly lead to contradictory conclusions in settings where the data are generated sequentially.

In this chapter, we introduce the *exponentially tilted Bayesian bootstrap* (ETBB), a method closely related to BETEL that avoids the aforementioned issues. We conjecture that it can be derived as the limit of a sequence of fully Bayesian procedures, yielding the interpretation that we have replaced the plug-in family in BETEL with a full nuisance parameter model that is integrated out. To support this conjecture, we carry out extensive simulations that illustrate this convergence both numerically and graphically.

However, the ETBB involves a nuisance parameter of dimension equal to the sample size. We develop two computational approaches to handle the difficulties associated with high-dimensional parameters. The first is based on the pre-conditioned Crank-Nicolson proposal (Cotter et al., 2013) previously introduced in §2.2.2. The second is based on Hamiltonian Monte Carlo (Neal, 2011).

We expand our scope beyond unequal probability sampling to tackle a range of practically relevant problems that lie within the moment restriction framework. The ETBB follows the projection-based perspective outlined in §1.2.2.4, where quantities are defined as functionals of the data generating distribution, rather than as components of a parametric model. In this respect, the ETBB has wider applicability than BETEL, which is restricted to inference for quantities that can be defined as the solutions to moment conditions. The utility of ETBB is more comparable to a fully nonparametric Bayesian model such as the Dirichlet process (Ferguson, 1973), but we are able to directly specify informative priors for the parameters that we tilt across, which we have emphasized as being a crucial aspect of performing Bayesian inference (§1.2.2.2).

4.2 Proposal

Our proposal is inspired primarily by unpublished work by Yuichi Kitamura and Taisuke Otsu, of which descriptions can be found in Bornn et al. (2019) and Florens and Simoni (2019). Kitamura and Otsu introduced a nonparametric prior for P that tilts a Dirichlet process prior to satisfy the moment condition for each value of θ.

Let Π be a chosen marginal prior for θ. Let $\tilde{P}$ be another parameter, with independent prior distribution $DP(\alpha)$, where α is a user-specified finite measure. Given $(\theta, \tilde{P})$, P solves the optimization problem

$$\min_P \int \log\left(\frac{dP}{d\tilde{P}}\right) d\tilde{P} \quad \text{s.t.} \quad \mathbb{E}_P\{g(D,\theta)\} = 0 \tag{4.1}$$

The solution to the optimization problem, if it exists, is unique, and it is the information projection (Csiszár, 1975) of $\tilde{P}$ onto the space of distributions satisfying the moment condition for θ. Similar to the approach of Kessler et al. (2015), the resulting prior for P combines an arbitrary prior for θ with a nonparametric prior. By construction, the implicit marginal prior of θ is still the user-specified Π.

The parameter $\tilde{P}$ could be interpreted as an initial estimate of P that disregards the moment condition and functional of interest. Given such an estimate, if one is provided with the information that the true value of P in fact satisfies the moment condition for a particular θ, it would be coherent to replace $\tilde{P}$ with the closest value—with respect to the KL-divergence in this case—within the constraint set.

In this set-up, we can recover BETEL by replacing the Dirichlet process prior for $\tilde{P}$ with a point mass prior on the empirical distribution. The Kitamura and Otsu model avoids the data-dependence of BETEL and incorporates the uncertainty in $\tilde{P}$. However, the computation is problematic. In order to solve the exponential tilting optimization (4.1), the distribution $\tilde{P}$ must be fully known. This precludes the possibility of modifying existing exact sampling methods that are used for Dirichlet mixture models such as slice sampling (Walker, 2007), where the distribution is adaptively truncated to a finite number of features. It seems likely that the only option is an approximate sampling method; for example, one that is based on the blocked Gibbs sampler (Ishwaran and James, 2001). We propose such a sampler in §D.2 and argue that it will scale very poorly with the sample size.

Our proposal is motivated by the aim of producing a computationally tractable method with similar inferential advantages to the Kitamura and Otsu approach. Of central importance is the following conjecture, based on how the standard Dirichlet process converges weakly to the Bayesian bootstrap.

Conjecture: *For any sequence of finite measures $\{\alpha_t\}_{t=0}^{\infty}$ such that $|\alpha_t| \to 0$ as $t \to \infty$, the sequence of posteriors for P under the Kitamura and Otsu model converges weakly to a proper distribution, which we call the* ***exponentially tilted Bayesian bootstrap*** *(ETBB) posterior. Moreover, this limiting distribution assigns probability 1 to a set of discrete distributions supported only on the observed data.*

We provide evidence to support this in §4.4. If the conjecture is true, there is a symmetry between the relationship of ETEL and the ETBB, and the relationship of the nonparametric bootstrap and the Bayesian bootstrap. As stated earlier, BETEL can be derived within the Kitamura & Otsu framework by replacing the Dirichlet process prior for $\tilde{P}$ with a point-mass prior on the empirical distribution $\mathbb{P}_n$, which follows the plug-in principle of the nonparametric bootstrap. The ETBB instead specifies a Bayesian bootstrap prior for $\tilde{P}$. This symmetry is summarized in Table 4.2.

Table 4.2 Comparison of bootstrap and exponential tilting methods.

	Frequentist	Bayesian
Bootstrap	Nonparametric bootstrap	Bayesian bootstrap
Model/likelihood	ETEL	ETBB

The conjecture implies that we can approximate the limiting posterior with any base probability measure that contains the observed data in its support. In particular, we can choose a discrete distribution supported only on the observed data, so that $\tilde{P}$ can now be finitely parameterized by the vector of probabilities $\tilde{q} = (\tilde{q}_1, \ldots, \tilde{q}_n)$ with Dirichlet prior $\mathrm{Dir}(a_1, \ldots, a_n)$, where $\tilde{q}_i = \tilde{P}(D = D_i)$ and $a_i \geqslant 0$ for all i. This leads to the joint posterior $(\theta, \tilde{P})$ defined by

$$p(\theta, \tilde{q} \mid D_1, \ldots, D_n) \propto \pi(\theta) \left\{ \prod_{i=1}^{n} \frac{q_i(\theta, \tilde{q})}{\tilde{q}_i^{1-a_i}} \right\}, \tag{4.2}$$

where $(q_1(\theta, \tilde{q}), \ldots, q_n(\theta, \tilde{q}))$ is the vector of probabilities on $\{D_1, \ldots, D_n\}$ that solves the optimization problem (4.1) for θ and $\tilde{P}$. For values of θ with no solutions to (4.1), the posterior density is set to 0. We will assume that Π admits a density $\pi(\theta)$ with respect to the Lebesgue measure, but we could generalize this in the usual way.

To approximate the ETBB posterior, the Dirichlet parameters $(a_1, \ldots, a_n)$ should all be set to values close to 0. While we have conjectured that the posterior for P converges to a proper distribution, we do not make the same claim for $\tilde{P}$. If all of $a_1, \ldots, a_n$ are equal to 0, the prior for $\tilde{q}$ is improper, and the posterior described by (4.2) may be as well. To resolve this, we propose setting all the parameters to be equal to 0 and truncating the parameter space, i.e. specify a small positive constant ε such that $\tilde{q}_i > \varepsilon > 0$ for all i. By doing this, the prior density for $\tilde{q}$ remains finite on the truncated space, guaranteeing propriety. In practice, some

degree of truncation will always be enforced regardless of this due to numerical limitations. More details regarding the computation can be found in §4.3.

BETEL can also be recovered from (4.2) by setting $a_1 = \ldots = a_n$ and taking the limit as $a_1 \to \infty$; the Dirichlet prior on $\tilde{P}$ converges to a point mass on the empirical distribution (Ghosal and van der Vaart, 2017). We illustrate the contrasts between BETEL, ETBB and different Dirichlet priors in Figures 4.3 and 4.4 for the following example.

Example 4.1. We return to the example of mean estimation. For illustrative purposes, we set $n = 3$. In this setting, the data is (x_1, x_2, x_3) and the moment condition is

$$g(X, \theta) = X - \theta.$$

For all of the simulations we performed, we fixed $x_1 = -1$ and $x_3 = 1$, and specified a uniform prior for θ. We varied the values of x_2 across different values between -1 and 1.

In the special case of $x_2 = 0$, the following proposition provides the explicit form of the conditional posterior of (q_1, q_2, q_3) given $\theta = 0$. The proof is provided in §D.3.

Proposition 4.1. *For $(x_1, x_2, x_3) = (-1, 0, 1)$, the conditional posterior of (q_1, q_2, q_3) given $\theta = 0$ is*

$$\begin{aligned} q_1 \mid (\theta = 0; x_1, x_2, x_3) &\sim \frac{1}{2} Beta(2, 1) \\ q_2 \mid (\theta = 0; x_1, x_2, x_3) &\sim Beta(1, 2) \\ q_3 \mid (\theta = 0; x_1, x_2, x_3) &= q_1 \mid (\theta = 0; x_1, x_2, x_3). \end{aligned}$$

Although Proposition 4.1 only concerns a special case, it provides some evidence that the ETBB posterior is well-defined and proper, even though the prior is improper.

Figure 4.1 compares the posterior densities for θ between the ETBB and the Bayesian bootstrap. The posterior densities shown for the Bayesian bootstrap have been calculated exactly using

$$p_{BB}(\theta) = \begin{cases} \frac{1+\theta}{1+x_2}, & \text{for } -1 \leqslant \theta \leqslant x_2 \\ \frac{1-\theta}{1-x_2}, & \text{for } x_2 \leqslant \theta \leqslant 1. \end{cases}$$

The details can be found in the §D.1.

Figure 4.2 displays the ETBB posterior densities for q. We recall that the Bayesian bootstrap posterior for q is uniform on the simplex, regardless of the values of the observed data. In contrast, the posterior densities for the ETBB, while being close to uniform, are higher in certain regions of the simplex depending on the value of x_2.

Figures 4.3 and 4.4 contain the posterior densities for q as we change the values of $a = (a_1, a_2, a_3)$. As stated earlier, we can interpret BETEL as the limit as $a_1 \to \infty$ with $a_1 = a_2 = a_3$. The posterior densities for BETEL are concentrated in small regions of the simplex. As the values of $\{a_1, a_2, a_3\}$ decrease, the posterior densities become gradually more diffuse.

4.3 Computation of the ETBB posterior

4.3.1 Optimization

Where there exists a solution, the vector $q = (q_1(\theta, \tilde{q}), \ldots, q_n(\theta, \tilde{q}))$ can be found by solving the dual optimization problem

$$q_i(\theta, \tilde{q}) = \frac{\tilde{q}_i \exp\{\lambda^{\mathrm{T}} g(D_i, \theta)\}}{\sum_{j=1}^{n} \tilde{q}_j \exp\{\lambda^{\mathrm{T}} g(D_j, \theta)\}}$$

where λ satisfies

$$\sum_{j=1}^{n} \tilde{q}_j \exp\{\lambda^{\mathrm{T}} g(D_j, \theta)\} g(D_j, \theta) = 0.$$

For each value of θ and $\tilde{q}$, λ can be approximated using a Newton-Raphson algorithm similar to the one used for ETEL in the previous chapter; we have provided pseudo-code in Algorithm 4.1. Since θ and $\tilde{q}$ are fixed for each optimization, there is no ambiguity in using the shorthand notation

$$\begin{aligned} g_i &= g(D_i, \theta) \\ f(\lambda) &= \sum_{j=1}^{n} \tilde{q}_j \exp\{\lambda^{\mathrm{T}} g_j\} g_j \\ H(\lambda) &= \sum_{j=1}^{n} \tilde{q}_j \exp\{\lambda^{\mathrm{T}} g_j\} g_j g_j^{\mathrm{T}}. \end{aligned}$$

As in the previous chapter, we should first check whether there exist solutions to the linear programming problem

$$\sum_{j=1}^{n} p_j g(D_j, \theta) = 0; \quad \{(p_1, \ldots, p_n) : 0 \leqslant p_i \leqslant 1, \sum_{j=1}^{n} p_j = 1\}. \tag{4.3}$$

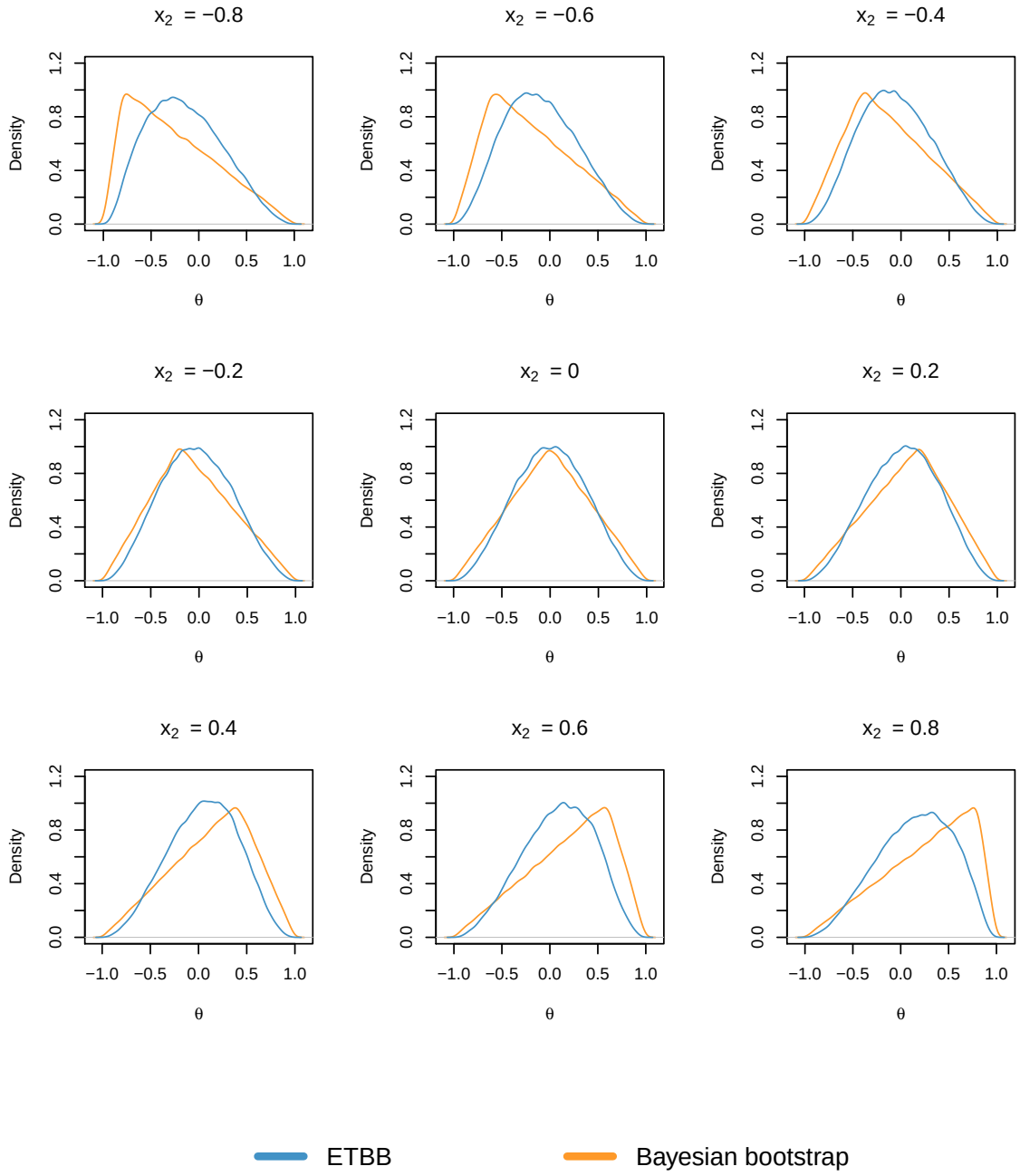


Fig. 4.1 The ETBB and Bayesian bootstrap posterior densities for θ across different values of x_2.

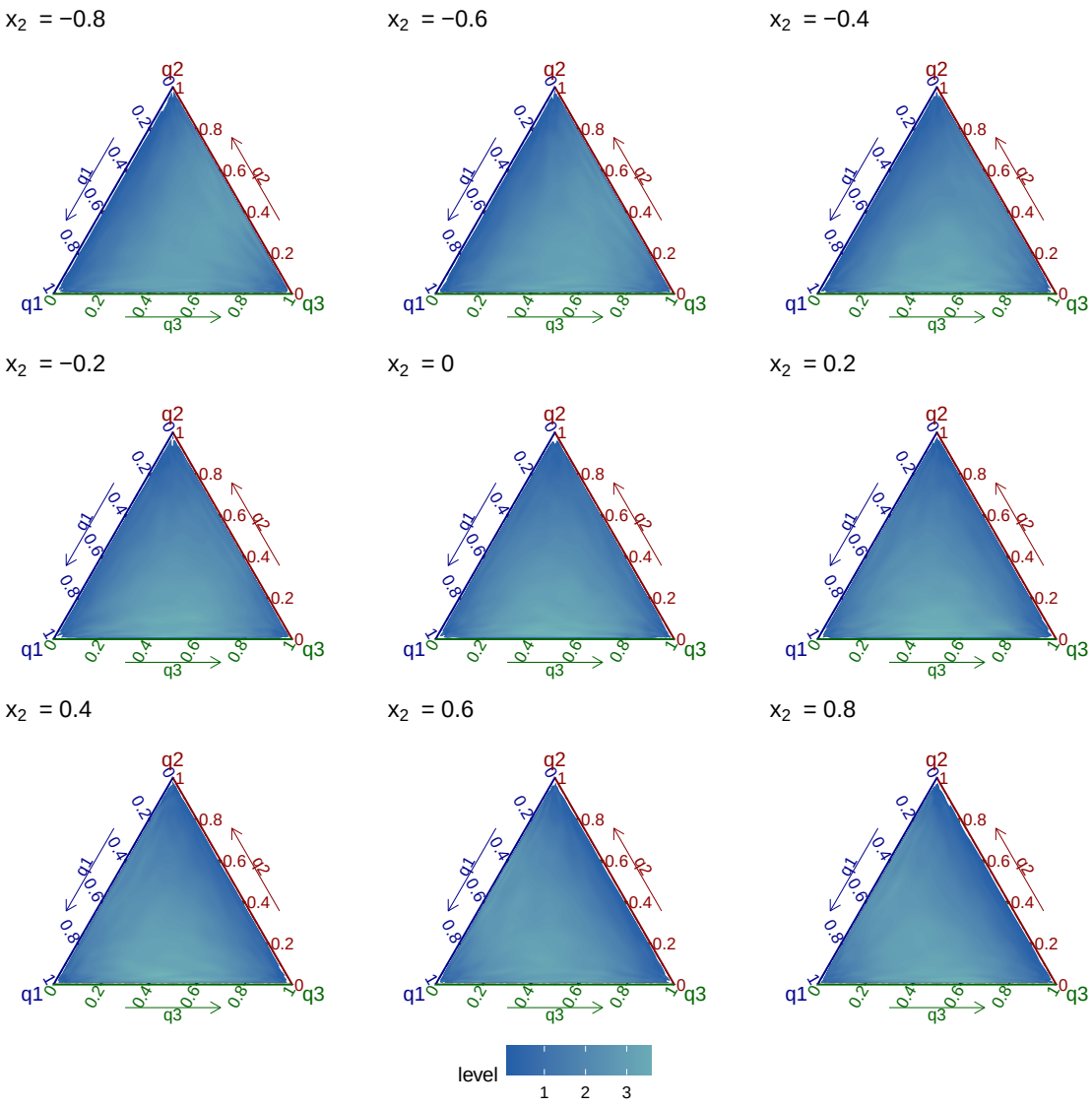


Fig. 4.2 The ETBB posterior density for q across different values of x_2.

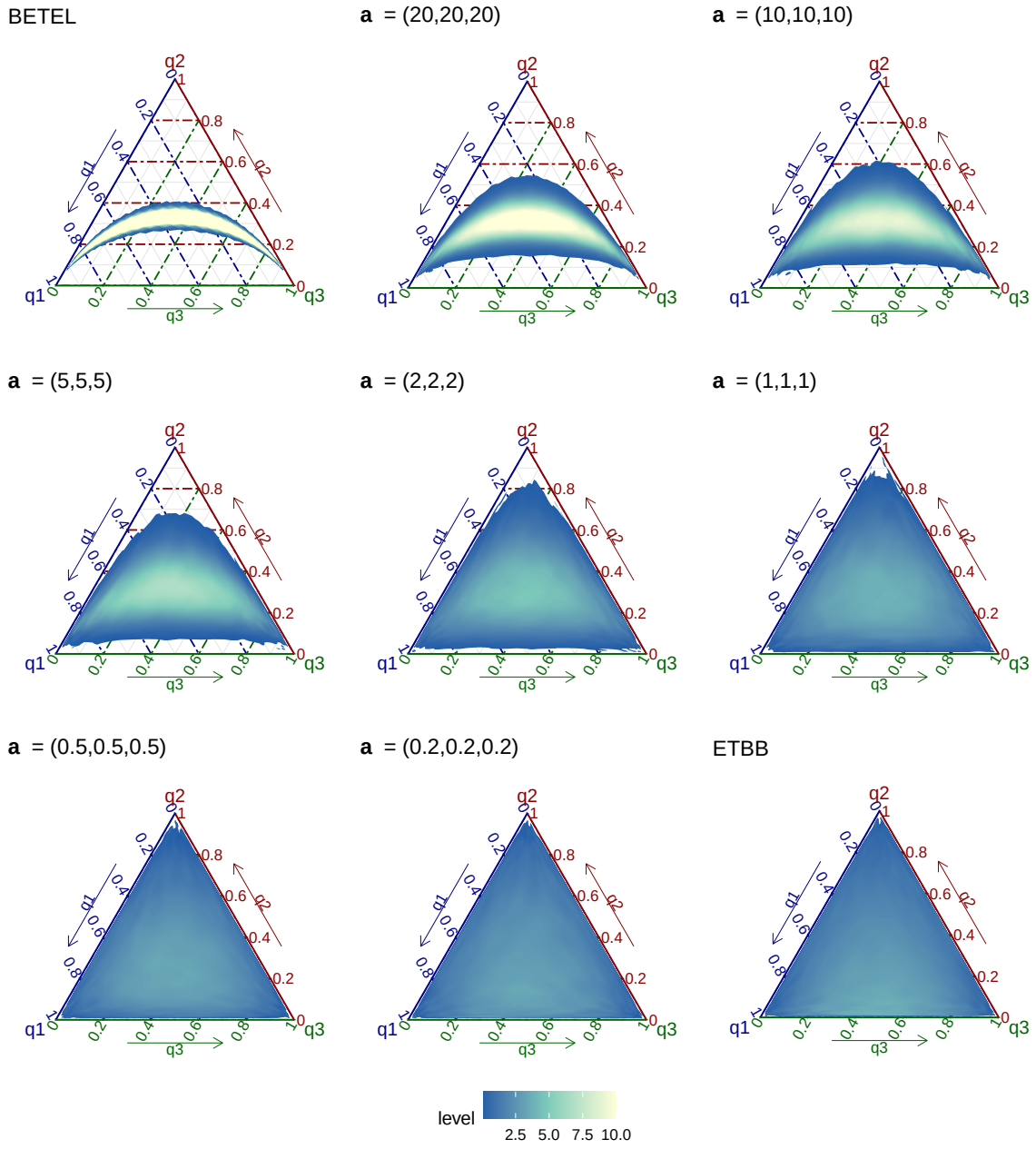


Fig. 4.3 The BETEL, Dirichlet and ETBB posterior density for q across different values of a with $x_2 = 0$.

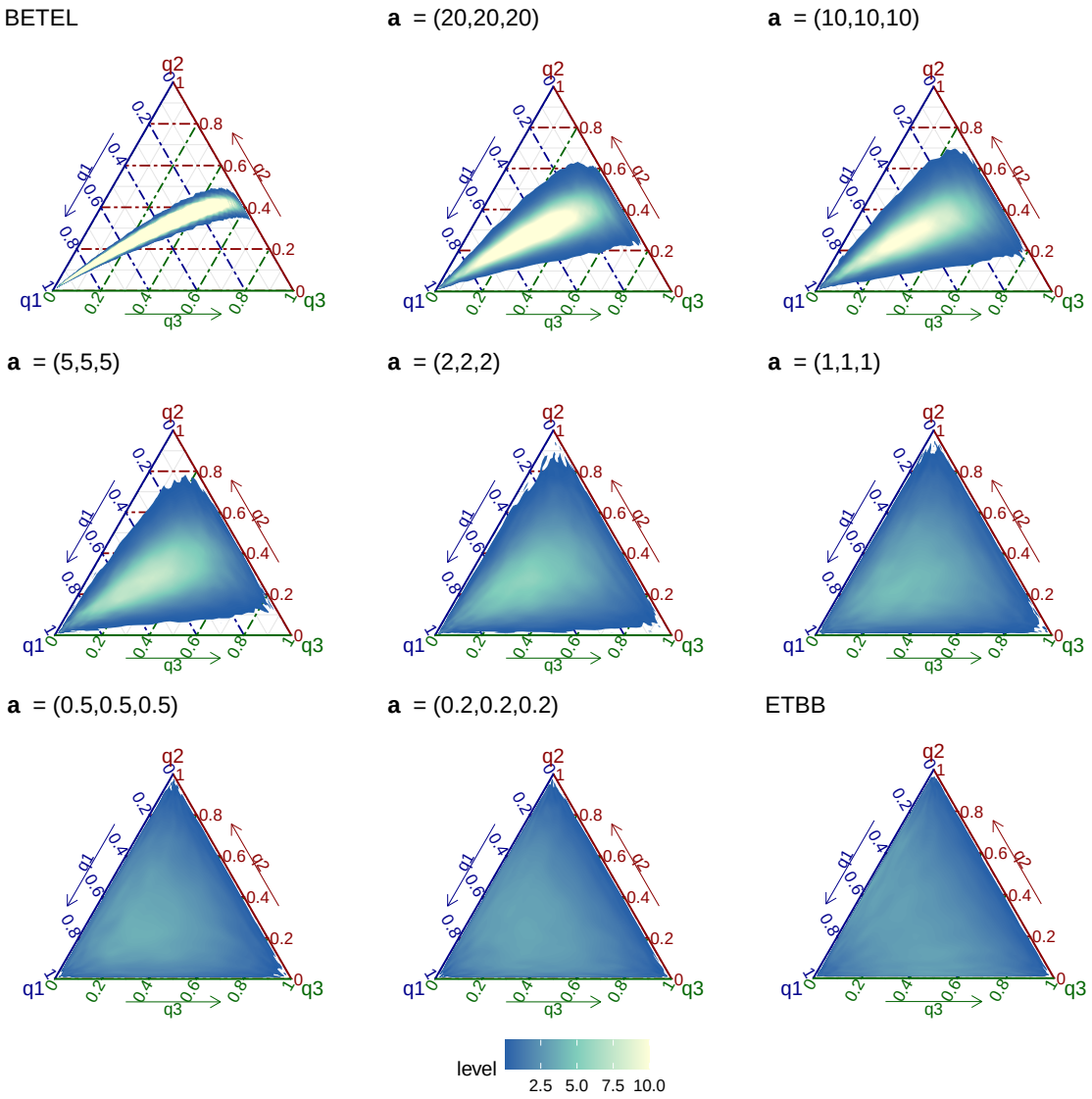


Fig. 4.4 The BETEL, Dirichlet and ETBB posterior density for q across different values of a with $x_2 = 0.8$.

Algorithm 4.1: Newton algorithm: optimizing q

Input θ, $\tilde{q}$ and tolerance $\tau_0 > 0$.
Solve linear programming problem described by (4.3).
If no feasible solutions exist
    output 0
else
    $\lambda \leftarrow (0, \ldots, 0)$
    $\tau \leftarrow \tau_0 + 1$
    while $\tau > \tau_0$
        $s \leftarrow H(\lambda)^{-1} f(\lambda)$
        $r \leftarrow 0$
        $\lambda' \leftarrow \lambda - s$
        while $f(\lambda) > f(\lambda')$
            $r \leftarrow r + 1$
            $\lambda' \leftarrow \lambda - 2^{-r} s$
        $\tau \leftarrow \|\lambda' - \lambda\|$
        $\lambda \leftarrow \lambda'$
    for $i = 1$ to $i = n$
        $q_i \leftarrow \tilde{q}_i \exp(\lambda^{\mathrm{T}} g_i) / \{\sum_{j=1}^{n} \tilde{q}_j \exp(\lambda^{\mathrm{T}} g_j)\}$
    output $(q_1, \ldots, q_n)$

For moderate to high dimensional θ, inverting $H(\lambda)$ may be computationally expensive. An alternative to Newton's method is gradient descent: solving $f(\lambda) = 0$ is equivalent to minimizing the objective

$$\Upsilon(\lambda) = \sum_{j=1}^{n} \tilde{q}_j \exp\{\lambda^{\mathrm{T}} g_j\}.$$

This is due to convexity; the unique global minimum lies at the point where $f(\lambda)$—the derivative of $\Upsilon(\lambda)$ with respect to λ—is equal to 0. An example of a gradient procedure is provided in Algorithm 4.2. Since gradient descent is a first-order method—using only the first derivative of the objective—it will likely take more iterations to converge than Newton's method. This will nevertheless result in a favourable trade-off for gradient descent if each iteration of Newton's method is much more expensive, particularly due to the inversion of $H(\lambda)$.

Algorithm 4.2: Gradient descent: optimizing q

Input θ, $\tilde{q}$ and tolerance $\tau_0 > 0$.
Solve linear programming problem described by (4.3).
If no feasible solutions exist
    output 0
else
    $\lambda \leftarrow (0, \ldots, 0)$
    $\tau \leftarrow \tau_0 + 1$
    while $\tau > \tau_0$
        $t \leftarrow 1$
        while $\Upsilon(\lambda - tf(\lambda)) > \Upsilon(\lambda) - (t/2)\|f(\lambda)\|_2^2$
            $t \leftarrow t/2$
        $\lambda' \leftarrow \lambda - tf(\lambda))$
        $\tau \leftarrow \|\lambda' - \lambda\|$
        $\lambda \leftarrow \lambda'$
    for $i = 1$ to $i = n$
        $q_i \leftarrow \tilde{q}_i \exp(\lambda^{\mathrm{T}} g_i) / \{\sum_{j=1}^{n} \tilde{q}_j \exp(\lambda^{\mathrm{T}} g_j)\}$
    output $(q_1, \ldots, q_n)$

4.3.2 Sampling $\tilde{q} \mid \theta$

We propose sampling from the ETBB posterior using a Gibbs sampler, alternating between updating θ and $\tilde{q}$. In both cases, exact sampling is infeasible and we suggest several options to tackle this.

There are two main challenges for sampling $\tilde{q}$ conditional on θ. First, we may encounter posterior multimodality. Second, the dimension of $\tilde{q}$ can be high since it scales with the sample size n. The two methods we propose in this subsection were designed to handle these issues to varying extents.

4.3.2.1 Pre-conditioned Crank-Nicolson proposal

The simpler of our two proposals uses pre-conditioned Crank-Nicolson proposals (Cotter et al., 2013) with Metropolis-Hastings. We had previously introduced this proposal in §2.2.2 and applied it to resolve the issues caused by high-dimensionality in the case-cohort design.

First, we perform an auxiliary transformation from the n-simplex to the unit hypercube $[0,1]^{n-1} \subset \mathbb{R}^{n-1}$. This transformation φ and its inverse are described in Algorithms 4.3 and 4.4 respectively.

Algorithm 4.3: Hypercube transformation $\tilde{q} \mapsto \varphi(\tilde{q})$

Input $\tilde{q}$.
Set $z_1 = \tilde{q}_1$.
For $i = 2$ to $i = n-1$
    Set $z_i = \tilde{q}_i \{\prod_{j<i}(1-z_j)\}^{-1}$.
Output $z = (z_1, \ldots, z_{n-1})$.

Algorithm 4.4: Inverse hypercube transformation $z \mapsto \varphi^{-1}(z)$

Input z.
Set $\tilde{q}_1 = z_1$.
For $i = 2$ to $i = n-1$
    Set $\tilde{q}_i = z_i \prod_{j<i}(1-z_j)$.
Set $\tilde{q}_n = \prod_{j=1}^{n-1}(1-z_j)$.
Output $\tilde{q} = (\tilde{q}_1, \ldots, \tilde{q}_n)$.

The inverse transformation can be interpreted as a stick-breaking construction. Suppose that $z = (z_1, \ldots, z_{n-1})$ lies inside the unit hypercube $[0,1]^{n-1}$. In the first step, we take a stick of length 1 and break off a piece of proportion z_1. In the second step, we take what is left and break off a piece of proportion z_2. We repeat this process until the n-th step, where we simply keep the remaining piece.

These transformations are similar to those proposed in Betancourt (2012); they only differ by a reflection in the hypercube, i.e. $(z_1, \ldots, z_{n-1}) \mapsto (1-z_1, \ldots, 1-z_{n-1})$. From the analysis in Betancourt (2012), we can immediately deduce that if $\tilde{q} \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_n)$, then

$$\varphi(\tilde{q}) \sim \prod_{i=1}^{n-1} \mathrm{Beta}(\alpha_i, \tilde{\alpha}_i), \tag{4.4}$$

where $\tilde{\alpha}_i = \sum_{k=i+1}^{n} \alpha_i$.

To overcome multimodality, we suggest selecting a proposal distribution for $\tilde{q}$ that has a disproportionate amount of probability mass at the boundaries of the n-simplex, where the modes are likely to be located. A straightforward choice is $\mathrm{Dir}(\alpha, \ldots, \alpha)$, where $0 < \alpha < 1$. To ensure good mixing, it may be necessary to correlate successive proposals; for this, we use the pre-conditioned Crank-Nicolson proposal (Cotter et al., 2013), which we had previously introduced in §2.2.2.

Let Φ and $F_{a,b}$ be the cumulative distribution functions of $\mathcal{N}(0,1)$ and $\mathrm{Beta}(a,b)$ respectively. We can invert $\tilde{q} \sim \mathrm{Dir}(\alpha, \ldots, \alpha)$ into a standard multivariate normal variable as

follows: map $\tilde{q}$ to $z = \varphi(\tilde{q})$ in the unit hypercube and set

$$U = (\Phi^{-1}\{F_{\alpha,(n-1)\alpha}(z_1)\}, \Phi^{-1}\{F_{\alpha,(n-2)\alpha}(z_2)\}, \ldots, \Phi^{-1}\{F_{\alpha,\alpha}(z_{n-1})\}).$$

By (4.4), $U \sim \mathcal{N}_{n-1}(0, I_{n-1})$. The pre-conditioned Crank-Nicolson proposal is $U' = \rho U + \sqrt{1-\rho^2}\varepsilon$, where $\varepsilon \sim \mathcal{N}_{n-1}(0, I_{n-1})$ and $\rho \in [0, 1)$. This proposal can be mapped back into the n-simplex, and accepted/rejected with a Metropolis-Hastings step.

The pseudo-code for a single update for $\tilde{q}$ conditional on θ is presented in Algorithm 4.5. Along with ρ and α, we also input the current values $\tilde{q}^{(\text{cur})}$ and $\theta^{(\text{cur})}$ of the sampler. We recommend starting with a value of α that is close to 1 and reducing the value to try to increase the acceptance rate if necessary. When n far exceeds the dimension of θ, we hypothesize that the marginal posterior of $\tilde{q}$ will be close to uniform and a value of α approximately equal to 1 will be optimal. If the sampler appears to get stuck at modes, it may help to try setting $\rho = 0$, which corresponds to proposing independent samples from $\text{Dir}(\alpha, \ldots, \alpha)$; this could increase the frequency for the sampler to jump from one mode to another.

Algorithm 4.5: Pre-conditioned Crank-Nicolson update for $\tilde{q}$

Input $\tilde{q}^{(\text{cur})}$, $\theta^{(\text{cur})}$, ρ, α.
Set $z = \varphi(\tilde{q}^{(\text{cur})})$.
Set $U = (\Phi^{-1}\{F_{\alpha,(n-1)\alpha}(z_1)\}, \Phi^{-1}\{F_{\alpha,(n-2)\alpha}(z_2)\}, \ldots, \Phi^{-1}\{F_{\alpha,\alpha}(z_{n-1})\})$.
Sample $\varepsilon \sim \mathcal{N}_{n-1}(0, I_{n-1})$ and set $U' = \rho U + \sqrt{1-\rho^2}\varepsilon$.
Set $z' = (F^{-1}_{\alpha,(n-1)\alpha}\{\Phi(U'_1)\}, \ldots, F^{-1}_{\alpha,\alpha}\{\Phi(U'_{n-1})\})$.
Set $\tilde{q}' = \varphi^{-1}(z')$.
With probability

$$\min\left[1, \left\{\prod_{i=1}^{n} \frac{q_i(\theta^{(\text{cur})}, \tilde{q}')}{q_i(\theta^{(\text{cur})}, \tilde{q}^{(\text{cur})})}\right\} \left\{\prod_{i=1}^{n} \frac{\tilde{q}'_i}{\tilde{q}_i^{(\text{cur})}}\right\}^{-\alpha}\right],$$

set $\tilde{q}^{(\text{cur})} = \tilde{q}'$.
Output $\tilde{q}^{(\text{cur})}$.

4.3.2.2 Hamiltonian Monte Carlo

For problems with very large sample sizes—and thus, very high-dimensional $\tilde{q}$—we propose a Hamiltonian Monte Carlo procedure for updating $\tilde{q}$. For high-dimensional problems, this procedure may exhibit substantial sampling efficiency gains, but it requires careful tuning of several parameters to be effective.

First, we need to map values in the n-simplex to an unconstrained space. We use φ—defined in Algorithm 4.3—as an auxiliary mapping into the unit hypercube $[0,1]^{n-1}$. We define a further mapping $\psi : [0,1]^{n-1} \to \mathbb{R}^{n-1}$ by $z \mapsto y$, where

$$y_k = \log\left[\frac{(n-k)z_k}{1-z_k}\right]$$

for $k = 1,\ldots,n-1$. The above inverts into

$$z_k = \frac{e^{y_k}}{n-k+e^{y_k}}.$$

A key ingredient of Hamiltonian Monte Carlo is computing the gradient of the log-target density. In this case, our target is the conditional posterior density of $\tilde{q}$ given θ:

$$p(\tilde{q} \mid \theta, D_1, \ldots, D_n) \propto \left\{\prod_{i=1}^{n} \frac{q_i(\theta,\tilde{q})}{\tilde{q}_i}\right\}$$

We provide the gradient with respect to $\tilde{q}$ in Proposition 4.2, which is proved in §D.3. Since θ is fixed while updating $\tilde{q}$, there is no ambiguity in using the shorthand notation $g_i = g(D_i, \theta)$.

Proposition 4.2. *For θ lying inside the convex hull defined by the moment condition g and the data, the gradient of the log-conditional posterior density is*

$$\frac{\partial \log p(\tilde{q} \mid \theta, D_1, \ldots, D_n)}{\partial \tilde{q}_j} = -\frac{q_j(\theta,\tilde{q})}{\tilde{q}_j}\left\{\left[\sum_{i=1}^{n} g_i\right]^{\mathrm{T}}\left[\sum_{i=1}^{n} q_i(\theta,\tilde{q}) g_i g_i^{\mathrm{T}}\right]^{-1} g_j + n\right\}$$

for $j = 1,\ldots,n$.

The chain rule allows us to find the gradient with respect to y:

$$\frac{\partial \log p(\tilde{q} \mid \theta, D_1, \ldots, D_n)}{\partial y_k} = \sum_{j=1}^{n} \frac{\partial \tilde{q}_j}{\partial y_k} \frac{\partial \log p(\tilde{q} \mid \theta, D_1, \ldots, D_n)}{\partial \tilde{q}_j},$$

where

$$\frac{\partial \tilde{q}_j}{\partial y_k} = \begin{cases} 0, & \text{for } k > j \\ \tilde{q}_j(1-z_j), & \text{for } k = j \\ -\tilde{q}_j z_k, & \text{for } k < j. \end{cases}$$

In the above, z_j and z_k are associated with the value $z = \psi^{-1}(y)$ in the unit hypercube. We will use the shorthand notation

$$S(y') = \left.\frac{\partial \log p(\tilde{q} \mid \theta, D_1, \ldots, D_n)}{\partial y}\right|_{y=y'}.$$

The gradient is involved in numerically integrating Hamilton's equations to propose a new value of $\tilde{q}$. The integrator we use is called the *leapfrog integrator*. We augment our parameter space with an additional momentum variable $\eta \in \mathbb{R}^{n-1}$ that has the same dimension as y. We also require specification of the following tuning parameters: positive-definite mass matrix $M \in \mathbb{R}^{(n-1)\times(n-1)}$, step-size $\varepsilon > 0$, and the number of leapfrog steps L. Suggestions on how to select and tune these parameters can be found further below. The leapfrog integrator is described in Algorithm 4.6. Starting at initial values of y and η, we update each in turn for L steps and output their final values.

Algorithm 4.6: Leapfrog integrator

Input y, η, M, ε, L.
Set $y_0 = y$ and $\eta_0 = \eta$.
For $l = 0$ to $l = L - 1$
  Set $\eta_{l+1/2} = \eta_l - (\varepsilon/2)S(y_l)$
  Set $y_{l+1} = y_l + \varepsilon M^{-1}\eta_{l+1/2}$
  Set $\eta_{l+1} = \eta_{l+1/2} - (\varepsilon/2)S(y_{l+1})$
Output (y_L, η_L).

The output from the leapfrog integrator is accepted/rejected with a Metropolis-Hastings step. The final ingredient is the Jacobian factor that takes into account the change of variables from $\tilde{q}$ to y. The Jacobian factor at y is

$$J(y) = \prod_{i=1}^{n-1} z_i(1 - z_i)\left(1 - \sum_{j=1}^{i-1} \tilde{q}_j\right),$$

where $z = \psi^{-1}(y)$ and $\tilde{q} = \varphi^{-1}(z)$. The details of the derivation can be found at https://mc-stan.org/docs/2_24/reference-manual/simplex-transform-section.html.

The Hamiltonian Monte Carlo update procedure is described in Algorithm 4.7. For each update, we sample a new inital value of the momentum η_0 from $\mathcal{N}_{n-1}(0, M)$. As a result, the density function $\phi_{0,M}$ of $\mathcal{N}_{n-1}(0, M)$ enters the acceptance ratio at the Metropolis-Hastings

step. Regardless of whether the proposal is accepted or rejected, the current value of the momentum is discarded and a new value is sampled in the next update.

Algorithm 4.7: Hamiltonian Monte Carlo update for $\tilde{q}$

Input $\tilde{q}^{(\text{cur})}$, $\theta^{(\text{cur})}$, M, ε, L.
Set $y_0 = \psi(\varphi(\tilde{q}^{(\text{cur})}))$.
Sample $\eta_0 \sim \mathcal{N}_{n-1}(0, M)$.
Set $(y_L, \eta_L) = \text{leapfrog}(y_0, \eta_0, M, \varepsilon, L)$ using Algorithm 4.6.
Set $\tilde{q}_L = \varphi^{-1}(\psi^{-1}(y_L))$.
With probability
$$\min\left[1, \frac{p(\tilde{q}_L|\theta^{(\text{cur})}, D_1, \ldots, D_n)}{p(\tilde{q}^{(\text{cur})}|\theta^{(\text{cur})}, D_1, \ldots, D_n)} \frac{J(y_L)}{J(y_0)} \frac{\phi_{0,M}(\eta_L)}{\phi_{0,M}(\eta_0)}\right],$$
set $\tilde{q}^{(\text{cur})} = \tilde{q}_L$.
Output $\tilde{q}^{(\text{cur})}$.

The most difficult aspect of Hamiltonian Monte Carlo is tuning the parameters involved. A systematic way of tuning M is to estimate the covariance matrix of y

$$M^{-1} = \mathbb{E}\{(y - \mathbb{E}[y])(y - \mathbb{E}[y])^{\mathrm{T}}\}$$

where the expectation is taken with respect to the target distribution of y. One can iteratively refine this estimate by repeatedly running the chain for a small number of iterations. An initial estimate could be obtained by using the sampling method described in §4.3.2.1 and transforming the posterior sample of $\tilde{q}$ to obtain a sample of y. Tuning L and ε is trickier to prescribe. One option is to start by fixing a value of L and adjusting ε to achieve the optimal acceptance rate of about 65.1% (Neal, 2011). The choice of $L = 1$ results in the Metropolis-adjusted Langevin algorithm.

4.3.3 Sampling $\theta \mid \tilde{q}$

Designing a sampler for $\theta \mid \tilde{q}$ that targets

$$p(\theta \mid \tilde{q}, D_1, \ldots, D_n) \propto \pi(\theta) \left\{ \prod_{i=1}^{n} \frac{q_i(\theta, \tilde{q})}{\tilde{q}_i} \right\}$$

is more conventional and straightforward. For low-dimensional θ, it may suffice to use random walk Metropolis-Hastings

$$\theta' \sim \mathcal{N}(\theta^{(\text{cur})}, V_{\text{prop}}),$$

where V_{prop} could be a scalar multiple of an estimate of the target distribution covariance. An initial value could be obtained by using the nonparametric bootstrap or Bayesian bootstrap.

For high-dimensional θ, gradient-based methods like the Metropolis-adjusted Langevin algorithm or Hamiltonian Monte Carlo may be necessary to improve sampling efficiency. The gradient of the log target density is provided in Proposition 4.3, which is proved in §D.3.

Proposition 4.3. *Suppose that $\pi(\theta)$ and $g(D,\theta)$ are both differentiable with respect to θ. For θ lying inside the convex hull defined by the moment condition g and the data, the gradient of the log-conditional posterior density is*

$$\frac{\partial \log p(\theta \mid \tilde{q}, D_1, \ldots, D_n)}{\partial \theta} = \frac{\nabla \pi(\theta)}{\pi(\theta)} + \frac{\partial \lambda^{\mathrm{T}}}{\partial \theta}\left(\sum_{i=1}^{n} g_i\right) + \sum_{i=1}^{n}(1 - nq_i)\frac{\partial g_i^{\mathrm{T}}}{\partial \theta}\lambda,$$

where λ is the solution to the dual optimization problem for θ and $\tilde{q}$, and

$$\frac{\partial \lambda}{\partial \theta} = -\left(\sum_{i=1}^{n} q_i(\theta, \tilde{q}) g_i g_i^{\mathrm{T}}\right)^{-1} \left\{\sum_{j=1}^{n} q_j(\theta, \tilde{q})(I_n + g_j \lambda^{\mathrm{T}})\frac{\partial g_j}{\partial \theta}\right\}.$$

4.4 Comparison with Kitamura & Otsu

In this section, we compare the Kitamura & Otsu proposal and the exponentially tilted Bayesian bootstrap with the aim of providing support for our claim that the ETBB can be viewed as a limit of the Kitamura & Otsu posterior as the base measure of the Dirichlet process tends to zero. First, we must develop a sampler for the Kitamura & Otsu approach. This is challenging because draws from a Dirichlet process are infinite-dimensional. Exact computational methods involving the Dirichlet process exploit either conjugacy or the ability to adaptively truncate samples (Walker, 2007). Neither option is possible in our setting; the ETBB model is not conjugate, and samples from a Dirichlet process cannot be truncated without introducing approximation error in the tilting step. Thus, we will settle with an approximate sampling method using truncated stick-breaking (Ishwaran and James, 2001).

Fix a whole number $K \geqslant n$ that determines the size of the support of any randomly drawn distribution. Let $\alpha > 0$ and let G_0 be a probability measure. Let

$$p_1 = V_1, \quad p_K = \prod_{i=1}^{K-1}(1-V_i), \quad p_k = V_k \prod_{i=1}^{k-1}(1-V_i) \quad \text{for } 2 \leqslant k \leqslant K-1$$

where $V_1, \ldots, V_{K-1} \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, \alpha)$ and let $A_1, \ldots, A_K \overset{\text{i.i.d.}}{\sim} G_0$. The random distribution defined by $\mathbb{P}(D = A_k) = p_k$ for $k = 1, \ldots, K$ has distribution approximately equal to the Dirichlet process $DP(\alpha, G_0)$. The approximation becomes increasingly exact as $K \to \infty$.

It is evident that $p_1, \ldots, p_K$ are not identically distributed; p_1 will be larger than p_2 on average, p_2 larger than p_3 on average etc. This means that the ordering of $A_1, \ldots, A_K$ is important. We split the parameterization of $A_1, \ldots, A_K$ into $B_1, \ldots, B_K$—just the set of values of the atoms—and the ordering indicators $I_1, \ldots, I_K$, which take values in $\{1, \ldots, K\}$; if $I_j = k$, then $A_j = B_k$. A priori, $B_1, \ldots, B_K \overset{\text{i.i.d.}}{\sim} G_0$ and $I_1, \ldots, I_K$ is uniform over the set of permutations on $\{1, \ldots, K\}$.

Given $p_1, \ldots, p_K, A_1, \ldots, A_K$, let $\tilde{P}$ be the probability measure[2] defined by $\tilde{P}(D = A_k) = p_k$ for $k = 1, \ldots, K$. For each value of θ, we define the tilted probability measure

$$\mathbb{P}_{\text{tilt}}(D = \cdot \mid \theta, p_1, \ldots, p_K, A_1, \ldots, A_K)$$

as the solution to the optimization problem (4.1). As before, this solution—if it exists—will be a discrete distribution supported only on $A_1, \ldots, A_K$. The likelihood function is thus

$$\prod_{i=1}^{n} \mathbb{P}_{\text{tilt}}(D = D_i \mid \theta, p_1, \ldots, p_K, A_1, \ldots, A_K),$$

the product of the exponentially tilted probabilities of observing the data. As a result, $\{B_1, \ldots, B_K\}$ must contain the values $\{D_1, \ldots, D_n\}$ a posteriori. Without loss of generality, we will fix $B_i = D_i$ for $i = 1, \ldots, n$. In §D.2, we develop a blocked Gibbs sampler (Ishwaran and James, 2001) that cycles through updating θ, $\{I_1, \ldots, I_K\}$, $\{B_1, \ldots, B_K\}$ and $\{V_1, \ldots, V_{K-1}\}$.

We perform two sets of simulations to support our conjecture given in §4.2; that is, the posterior in the Kitamura & Otsu model converges to the ETBB posterior for any sequence of base measures that tends to 0. The requirement that this holds for *any* sequence is crucial for ensuring that the ETBB posterior is well-defined. Thus, we will illustrate this convergence

[2]This $\tilde{P}$ plays the same role as the $\tilde{P}$ in the ETBB but is supported on the atoms $A_1, \ldots, A_K$, rather than just the observed data points.

for different sequences of base measures in each setting. Throughout, the number of sticks for the Kitamura & Otsu implementation was fixed at 10.

Example 4.1 (continued)**.** We return to the mean estimation example in §4.2 with three data points. The data is (x_1, x_2, x_3) and the moment condition function is

$$g(X, \theta) = X - \theta.$$

We fixed $x_1 = -1$ and $x_3 = 1$, and specified a $\mathscr{U}(-1, 1)$ prior for θ. For the Kitamura & Otsu proposal, we consider three sequences of base measures. Each sequence has the following structure: we fix a base probability distribution $\bar{\alpha}$ and set $\alpha = |\alpha|\bar{\alpha}$ for different values of $|\alpha| > 0$. In the first and second sequences, we have $\bar{\alpha} = \mathscr{N}(0, 2^2)$ and $\bar{\alpha} = \mathscr{N}(0, 0.1^2)$ respectively. In the third sequence, $\bar{\alpha}$ is the skew normal distribution with density function

$$\phi\left(\frac{x}{2}\right)\Phi(-x),$$

where ϕ and Φ are the density and cumulative distribution functions of $\mathscr{N}(0, 1)$ respectively.

Table 4.3 presents the posterior mean estimates of the marginal distributional probabilities. We see that as $|\alpha|$ tends to 0, the estimates for the Kitamura & Otsu method appear to converge to those of the ETBB. Furthermore, the total probability mass on $\{x_1, x_2, x_3\}$ tends to 1, supporting our claim that the full nonparametric posterior will concentrate on the set of distributions that are supported only on the observed data.

Figures 4.5 and 4.6 compare the posterior density plots for θ for $x_2 = 0$ and $x_2 = 0.8$ respectively. In each setting, we can see that as $|\alpha|$ tends to zero, the density functions for the Kitamura and Otsu method appear to converge to the density for the ETBB for both values of x_2. For $\bar{\alpha} = \mathscr{N}(0, 2^2)$, the base distribution is more dispersed than the ETBB posterior. As a result, the Kitamura and Otsu posterior becomes tighter as $|\alpha|$ decreases. We observe the opposite effect when $\bar{\alpha} = \mathscr{N}(0, 0.1^2)$. For clarity of presentation, we omitted the curves for $|\alpha| = 0.05$, which followed the patterns described above but were very close to the curves for $|\alpha| = 0.1$.

Example 4.2 (Logistic regression)**.** In the second set of simulations, we investigate logistic regression. The data are fixed to be the six values

$$\{d_1, \ldots, d_6\} = \{(-1, 0), (-0.5, 0), (0.2, 0), (-0.2, 1), (0.5, 1), (1, 1)\}.$$

Table 4.3 Mean estimation comparison of posterior mean probabilities between Kitamura & Otsu and ETBB.

$\mathcal{N}(0,2^2)$	$x_2 = 0$				$x_2 = 0.8$			
Method	$\mathbb{P}(X = x_1)$	$\mathbb{P}(X = x_2)$	$\mathbb{P}(X = x_3)$	Total	$\mathbb{P}(X = x_1)$	$\mathbb{P}(X = x_2)$	$\mathbb{P}(X = x_3)$	Total
$\lvert\alpha\rvert = 1$	0.243	0.223	0.245	0.710	0.252	0.227	0.235	0.714
$\lvert\alpha\rvert = 0.5$	0.285	0.247	0.287	0.818	0.303	0.256	0.260	0.819
$\lvert\alpha\rvert = 0.2$	0.328	0.256	0.327	0.911	0.346	0.276	0.285	0.907
$\lvert\alpha\rvert = 0.1$	0.336	0.263	0.333	0.932	0.357	0.279	0.292	0.928
$\lvert\alpha\rvert = 0.05$	0.347	0.265	0.347	0.959	0.370	0.283	0.296	0.948
ETBB	0.362	0.278	0.361	1	0.398	0.293	0.309	1

$\mathcal{N}(0,0.1^2)$	$x_2 = 0$				$x_2 = 0.8$			
Method	$\mathbb{P}(X = x_1)$	$\mathbb{P}(X = x_2)$	$\mathbb{P}(X = x_3)$	Total	$\mathbb{P}(X = x_1)$	$\mathbb{P}(X = x_2)$	$\mathbb{P}(X = x_3)$	Total
$\lvert\alpha\rvert = 1$	0.279	0.218	0.279	0.777	0.282	0.241	0.256	0.779
$\lvert\alpha\rvert = 0.5$	0.315	0.248	0.315	0.878	0.334	0.262	0.279	0.875
$\lvert\alpha\rvert = 0.2$	0.342	0.259	0.344	0.945	0.364	0.281	0.288	0.932
$\lvert\alpha\rvert = 0.1$	0.350	0.269	0.350	0.969	0.374	0.287	0.291	0.952
$\lvert\alpha\rvert = 0.05$	0.351	0.267	0.351	0.970	0.377	0.285	0.295	0.957
ETBB	0.362	0.278	0.361	1	0.398	0.293	0.309	1

Skew normal	$x_2 = 0$				$x_2 = 0.8$			
Method	$\mathbb{P}(X = x_1)$	$\mathbb{P}(X = x_2)$	$\mathbb{P}(X = x_3)$	Total	$\mathbb{P}(X = x_1)$	$\mathbb{P}(X = x_2)$	$\mathbb{P}(X = x_3)$	Total
$\lvert\alpha\rvert = 1$	0.223	0.235	0.337	0.795	0.231	0.270	0.294	0.795
$\lvert\alpha\rvert = 0.5$	0.273	0.256	0.347	0.876	0.295	0.277	0.305	0.876
$\lvert\alpha\rvert = 0.2$	0.311	0.262	0.357	0.930	0.337	0.284	0.309	0.929
$\lvert\alpha\rvert = 0.1$	0.324	0.269	0.357	0.950	0.348	0.283	0.309	0.940
$\lvert\alpha\rvert = 0.05$	0.329	0.263	0.362	0.954	0.355	0.282	0.314	0.951
ETBB	0.362	0.278	0.361	1	0.398	0.293	0.309	1

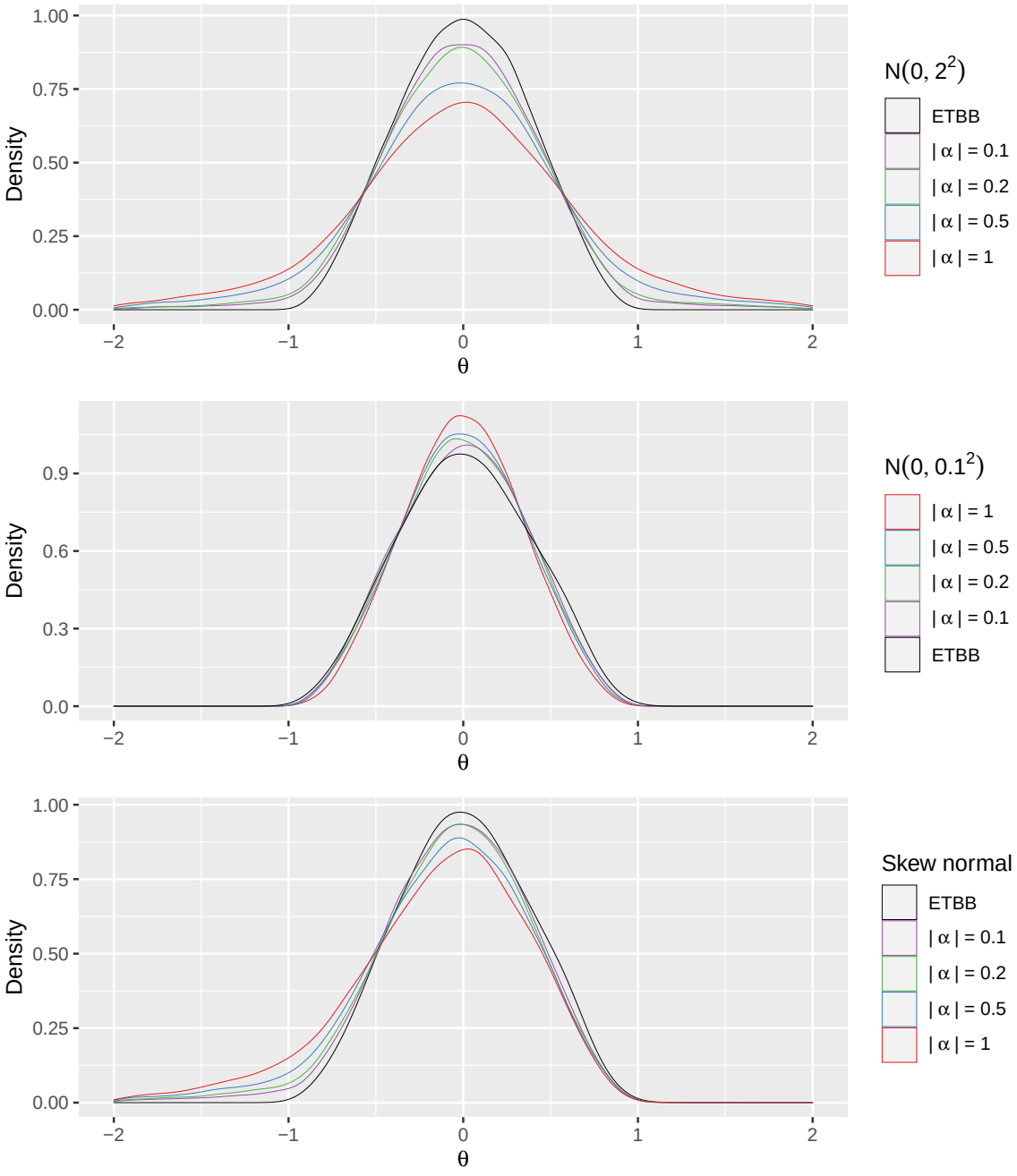


Fig. 4.5 Mean estimation posterior densities for θ $(x_2 = 0)$.

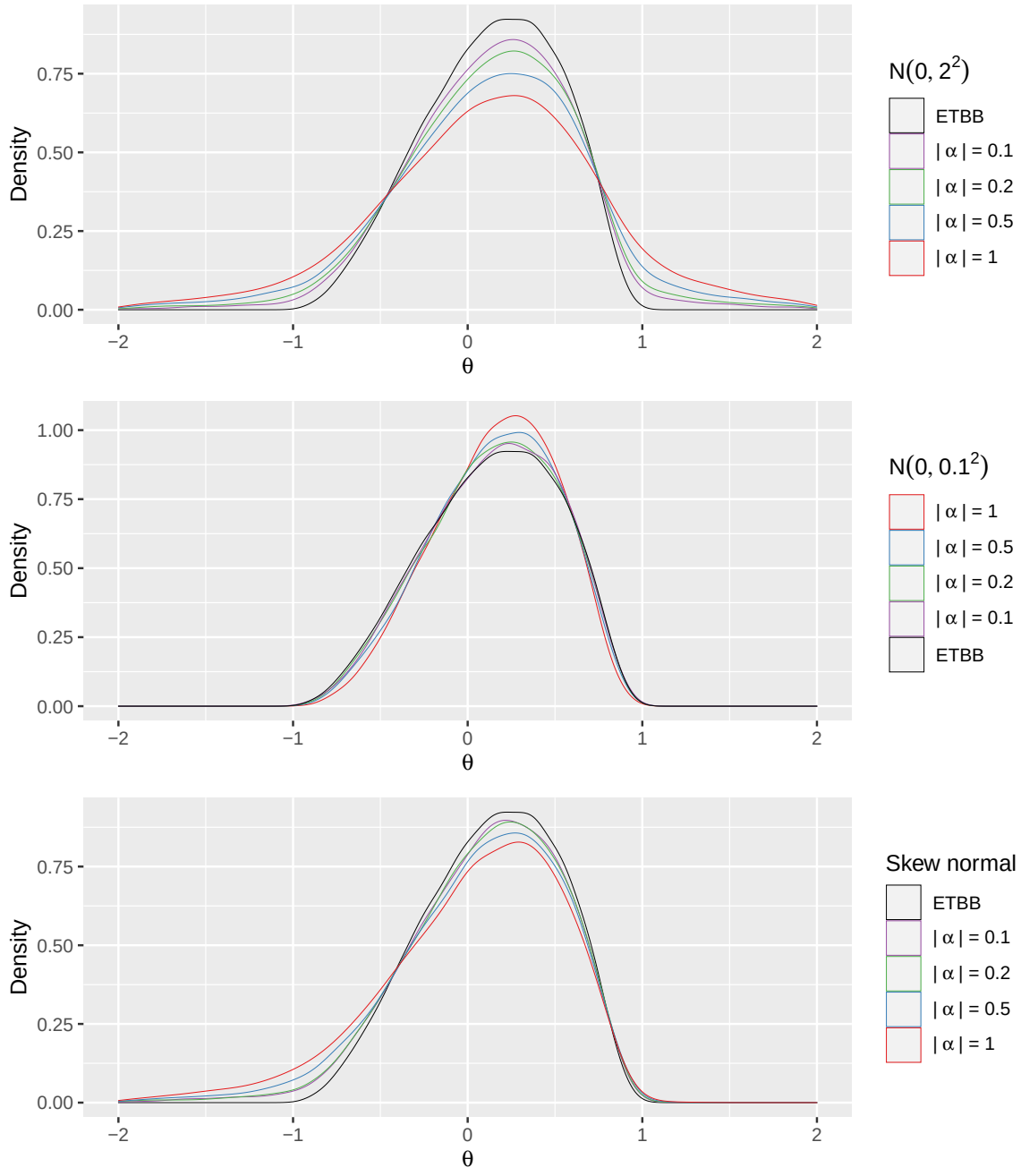


Fig. 4.6 Mean estimation posterior densities for θ $(x_2 = 0.8)$.

The target quantity θ satisfies the moment condition defined by

$$g(D,\theta) = X\left(Y - \frac{e^{\theta X}}{1+e^{\theta X}}\right),$$

where $D = (X,Y)$, i.e. θ is the regression coefficient for X with no intercept. We consider two sequences of base measures for the Kitamura and Otsu proposal, both with the same structure as before—we fix the base distribution and vary the size of the base measure. In setting 1, we consider a base distribution having independent X and Y: $X \sim \mathcal{N}(0,1)$ and $Y \sim \text{Ber}(0.5)$. Equivalently, this is the logistic regression model with $\theta = 0$. In setting 2, the base distribution is defined by $X \sim \mathcal{N}(0,0.5^2)$ and $Y \mid X \sim \text{Ber}(\text{expit}(8X))$. In both settings, we specify a flat prior for θ.

Table 4.4 contains the posterior mean estimates of the marginal distributional probabilities. Similar to the mean estimation example, we can see in both settings that the estimates from the Kitamura and Otsu method converge to the estimates from the ETBB as $|\alpha|$ tends to 0. We also see the total of the Kitamura and Otsu estimates converging to 1, again supporting our conjecture that the limit of the Kitamura and Otsu posterior puts all of its mass on the set of distributions supported only on the observed data.

Figure 4.7 presents the posterior densities for θ. In setting 1, we can see that the density for $|\alpha| = 1$ is—relatively speaking—shrunk towards the value 0. This is expected because X and Y are independent under the base distribution, corresponding to $\theta = 0$. As $|\alpha|$ decreases, we see the curves converge towards the ETBB density. The opposite occurs in setting 2. For $|\alpha| = 1$, the density is shrunk towards $\theta = 8$ due to the influence of the base distribution. We then see the curves converge towards the ETBB density from the opposite direction to setting 1.

4.5 Further examples

4.5.1 Robust linear regression

Consider the linear model

$$Y = X\beta + \varepsilon,$$

where $\mathbb{E}[\varepsilon \mid X] = 0$. Bayesian linear regression conventionally operates under the assumption of *homoscedasticity*; that is, $\mathbb{E}[\varepsilon^2 \mid X] = \mathbb{E}[\varepsilon^2]$. Such a modelling choice is more likely to be

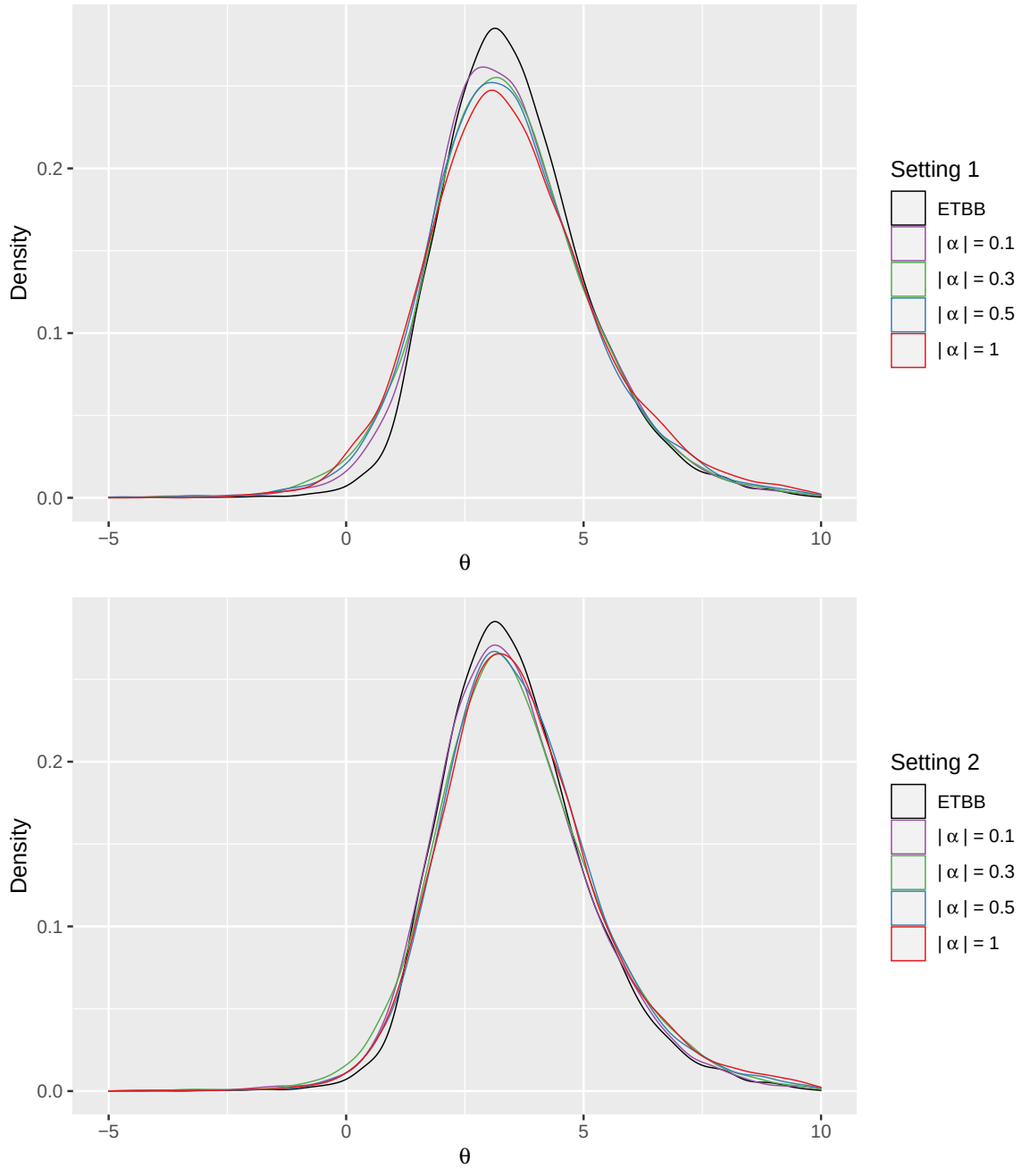


Fig. 4.7 Logistic regression posterior densities for θ.

Table 4.4 Logistic regression comparison of posterior mean probabilities between Kitamura & Otsu and ETBB.

	Setting 1						
Method	$\mathbb{P}(D=d_1)$	$\mathbb{P}(D=d_2)$	$\mathbb{P}(D=d_3)$	$\mathbb{P}(D=d_4)$	$\mathbb{P}(D=d_5)$	$\mathbb{P}(D=d_6)$	Total
$\lvert\alpha\rvert=1$	0.144	0.164	0.127	0.128	0.168	0.147	0.879
$\lvert\alpha\rvert=0.5$	0.145	0.169	0.148	0.145	0.175	0.150	0.931
$\lvert\alpha\rvert=0.3$	0.148	0.173	0.154	0.153	0.175	0.154	0.957
$\lvert\alpha\rvert=0.1$	0.155	0.179	0.158	0.154	0.180	0.153	0.980
ETBB	0.155	0.184	0.156	0.158	0.190	0.157	1
	Setting 2						
Method	$\mathbb{P}(D=d_1)$	$\mathbb{P}(D=d_2)$	$\mathbb{P}(D=d_3)$	$\mathbb{P}(D=d_4)$	$\mathbb{P}(D=d_5)$	$\mathbb{P}(D=d_6)$	Total
$\lvert\alpha\rvert=1$	0.138	0.151	0.145	0.147	0.146	0.141	0.868
$\lvert\alpha\rvert=0.5$	0.148	0.166	0.147	0.148	0.172	0.146	0.926
$\lvert\alpha\rvert=0.3$	0.146	0.169	0.158	0.153	0.172	0.154	0.951
$\lvert\alpha\rvert=0.1$	0.150	0.180	0.162	0.163	0.177	0.154	0.986
ETBB	0.155	0.184	0.156	0.158	0.190	0.157	1

for the sake of convenience than a true reflection of the analyst's beliefs. In the simplest case, conjugate models allow one to forgo MCMC and sample from the posterior directly.

If homoscedasticity is violated, a Bayesian model that does not take this into account will generally produce credible regions for β that do not achieve nominal coverage as the number of samples goes to infinity. Even if heteroscedasticity is detected, handling the resulting model selection issue from a Bayesian standpoint is not straightforward. On the other hand, a frequentist can remain agnostic with the use of heteroscedastic-consistent standard errors (White, 1980).

In this subsection, we demonstrate that the ETBB can provide a solution to these problems. In the first scenario, we generate $X \sim \mathcal{N}(0,1)$ and $\varepsilon \mid X \sim \mathcal{N}(0,0.5^2)$. In the second scenario, we generate $X \sim \mathcal{N}(0,1)$ and $\varepsilon \mid X \sim \mathcal{N}(0,0.5^2|X|)$. In both cases, the true value of β is 0.4, and we generate $n = 100$ samples.

We compare the ETBB and BETEL with the following standard Bayesian model:

$$\begin{aligned} \varepsilon|X,\sigma^2 &\sim \mathcal{N}(0,\sigma^2) \\ \beta,\sigma^2 &\sim p(\beta,\sigma^2) \propto \sigma^{-2}. \end{aligned}$$

This is a special case of the conjugate normal-inverse gamma model with a Jeffrey's prior for (β,σ^2). The model is correctly specified in scenario 1 but fails to address the heteroscedasticity present in scenario 2.

For the ETBB and BETEL, we investigate two sets of moment conditions. The first is the ordinary least squares condition $\mathbb{E}[X(Y - X\beta)] = 0$ with a weakly informative $\beta \sim \mathcal{N}(0, 4^2)$ prior. The second combines the ordinary least squares condition with $\mathbb{E}[\sigma^2 - (Y - X\beta)^2] = 0$ with the same prior as the standard Bayesian model above. This set of conditions can be interpreted as the expected score equations for the normal linear model; we seek the pseudo-true values of β and σ^2 that minimize the KL-divergence between the normal linear model and the true data-generating distribution.

Table 4.5 and 4.6 summarize the results for both scenarios, each run for 500 iterations. Values for biases, empirical standard deviations, and root mean squared errors are given with respect to posterior mean estimates. Central 95% credible intervals were used for β, while one-sided intervals were used for σ^2 (from 0 to the upper 95% quantile).

In the first scenario, the standard Bayesian model is correctly specified, and as expected, we see from Table 4.5 that the central 95% credible intervals for both β and σ^2 achieve nominal coverage. In the second scenario, however, the model misspecification leads to the standard Bayesian credible intervals undercovering substantially, as shown in Table 4.6. BETEL with ordinary least squares performs well for β in both settings but poorly for σ^2, which is not included in the moment condition. This is fixed by using the score conditions instead; in that case, BETEL performs well in each category, albeit with slight undercoverage, similar to both ETBB methods.

Our simulations illustrate that the ETBB can allow the Bayesian to remain as agnostic to homoscedasticity/heteroscedasticity as a frequentist using ordinary least squares. There is little loss in performance relative to a fully parametric model under correct specification, and there are significant gains when homoscedasticity is violated.

The advantages of using ETBB with ordinary least squares over BETEL with the score conditions are more evident if we consider extensions to multivariate outcomes. This is because specifying a prior for a covariance matrix is well-known to be difficult (Barnard et al., 2000), particularly if the user wishes to be noninformative in their beliefs. The Jeffreys prior that we used in §2.4.2 is convenient in the conjugate normal-inverse Wishart model but lacks this benefit when coupled with BETEL. The number of parameters involved in a covariance matrix induces a large computational burden in the optimization step, limiting the scalability of the method.

Above, we have only considered a semiparametric model where the linearity of the regression function is assumed to be true. Since our moment condition does not impose any model restrictions, we can also use the ETBB under a fully nonparametric setting, where linearity may not necessarily hold. In that case, β can be viewed as a statistical functional

Table 4.5 (Scenario 1) Comparison of standard Bayes, BETEL, and ETBB for the homoscedastic errors model. OLS, ordinary least squares; Sco, score equation; N-IG, normal-inverse gamma; ESD, empirical standard deviation; RMSE, root mean squared error, Wid, mean width of central 95% credible intervals, Cov, coverage of central 95% credible intervals.

	β				
Method	Bias	ESD	RMSE	Wid	Cov
N-IG	0.004	0.050	0.050	0.197	95.6%
BETEL-OLS	-0.002	0.049	0.049	0.189	92.2%
ETBB-OLS	0.003	0.051	0.051	0.188	94.0%
BETEL-Sco	0.003	0.050	0.050	0.188	94.0%
ETBB-Sco	0.003	0.051	0.051	0.186	93.4%
	σ^2				
Method	Bias	ESD	RMSE	Wid	Cov
N-IG	0.002	0.033	0.033	0.318	94.8%
BETEL-OLS	-0.001	0.035	0.035	0.255	52.0%
ETBB-OLS	-0.003	0.033	0.033	0.307	91.8%
BETEL-Sco	-0.002	0.033	0.033	0.308	91.2%
ETBB-Sco	-0.002	0.033	0.033	0.307	91.4%

Table 4.6 (Scenario 2) Comparison of standard Bayes, BETEL, and ETBB for the heteroscedastic errors model. N-IG, normal-inverse gamma; ESD, empirical standard deviation; RMSE, root mean squared error, Wid, mean width of central 95% credible intervals, Cov, coverage of central 95% credible intervals.

	β				
Method	Bias	ESD	RMSE	Wid	Cov
N-IG	0.003	0.063	0.063	0.175	83.6%
BETEL-OLS	0.003	0.062	0.062	0.227	93.2%
ETBB-OLS	0.003	0.062	0.062	0.229	93.8%
BETEL-Sco	0.003	0.062	0.062	0.229	93.8%
ETBB-Sco	0.004	0.062	0.062	0.222	93.0%

derived from a nonparametric projection of the true data-generating distribution onto the space of linear models (Buja et al., 2019a,b).

4.5.2 Quantile regression

Standard regression generally focuses on estimating the conditional mean function $\mathbb{E}[Y \mid X = x]$. The aim of quantile regression is to provide a more detailed summary of the data by using several regression curves corresponding to different quantiles of the conditional distribution of Y given X.

Consider the following model (Koenker and Bassett, 1978): the τ-th conditional quantile function of $Y \in \mathbb{R}$ given $X \in \mathbb{R}^p$ is specified to be

$$Q_\tau(Y \mid X) := \inf\{t \mid \mathbb{P}(Y \leqslant t \mid X) \geqslant \tau\} = X\beta(\tau),$$

where $\tau \in (0,1)$. This model specification is not generative; additional structure is required to perform standard Bayesian inference. Given data $\{(X_i, Y_i) \mid i = 1, \ldots, n\}$, $\beta(\tau)$ can be estimated from a frequentist perspective by

$$\hat{\beta}(\tau) = \arg\min_\beta \sum_{i=1}^n \rho_\tau(Y_i - X_i\beta)$$

where $\rho_\tau(y - x\beta) = (\tau - \mathbb{1}\{y < x\beta\})(y - x\beta)$ (Cameron and Trivedi, 2010).

Chamberlain and Imbens (2003) suggested an approach based on the Bayesian bootstrap that repeatedly solves

$$\beta(\tau)^{(l)} = \arg\min_\beta \sum_{i=1}^n w_i^{(l)} \rho_\tau(Y_i - X_i\beta)$$

where $(w_1^{(l)}, \ldots, w_n^{(l)})$ are uniform Dirichlet weights. Lancaster and Jun (2010) and Yang and He (2012) studied approaches based on BETEL and Bayesian empirical likelihood (BEL) respectively[3]. Following Yang and He (2012), we consider the moment condition

$$\mathbb{E}[\psi_\tau(Y - X\beta(\tau))X] = 0, \tag{4.5}$$

where

$$\psi_\tau(u) = \begin{cases} \mathbb{1}\{u < 0\} - \tau, & \text{for } u \neq 0 \\ 0, & \text{for } u = 0. \end{cases}$$

[3]The empirical likelihood was previously defined in §3.2.1

It is likely that we are interested in multiple quantiles $(\tau_1, \ldots, \tau_k)$. Using BETEL or Bayesian empirical likelihood would require kp estimating equations, which may be computationally expensive. Moreover, it is plausible that the user only has informative beliefs about some of the quantiles of interest (particularly, the median). In contrast, the ETBB can be used to estimate all quantile coefficients simultaneously while allowing the user to specify moment conditions for a selected subset of the quantiles.

We assess the performance of the ETBB with the following simulation study by Yang and He (2012). The outcomes are generated by $Y_i = \beta_I + \beta_S(X_i - 2) + \varepsilon_i$ $(i = 1, \ldots, n)$, where $\beta_I = 2$, $\beta_S = 1$, and X_i and ε_i are independently generated from the chi-squared distribution with 2 degrees of freedom and $\mathcal{N}(0, 2^2)$ respectively. We specify the moment conditions (4.5) for $\tau = 0.5$ only; that is, the conditions for the median regression coefficients $\beta_I(0.5)$ and $\beta_S(0.5)$. Independent priors of $\mathcal{N}(0, 100^2)$ are specified for both parameters. Unlike Yang and He (2012), we will also investigate the results for $\tau = 0.25$ and $\tau = 0.75$, even though we have not used the corresponding moment conditions.

Table 4.7 presents the results for ETBB across 500 iterations in each setting for $n = 100$ and $n = 200$. We have included the results for BEL, BETEL and Chamberlain and Imbens (2003) for comparison. The coverage and (average) widths are given with respect to the central 95% credible intervals. We can see that the ETBB intervals perform similarly to Chamberlain and Imbens (2003), achieving approximate calibration for all parameters with similar widths. This is reassuring because the Bayesian bootstrap is known to satisfy the nonparametric Bernstein-von Mises theorem (Ghosal and van der Vaart, 2017) and is therefore asymptotically efficient in the nonparametric model.

BEL and BETEL achieve nominal coverage for the median parameters, but the intervals for the other parameters exhibit substantial undercoverage as expected. We obtain further evidence that the plug-in approach fails to fully propagate the uncertainty in the parameters that are not determined by the moment conditions.

4.5.3 Doubly robust estimation

We revisit the experiment in §3.3.2 derived from Kang and Schafer (2007). Similar to our earlier investigation of BETEL, we use two ETBB models in each setting. The first specifies a flat prior for all parameters. The second specifies a flat prior for all parameters aside from μ, for which we specify a $t_3(210, 1)$ prior. We compare the ETBB with the approaches described previously in §3.3.2: the standard doubly robust estimator (Robins et al., 1994), the Saarela et al. (2016) proposal, and the two BETEL models with the same priors as above.

Table 4.7 Comparison of ETBB with BEL, BETEL and Chamberlain and Imbens (2003) for quantile regression. CI, Chamberlain and Imbens.

$n = 100$	Coverage					
Method	$\beta_I(0.25)$	$\beta_S(0.25)$	$\beta_I(0.5)$	$\beta_S(0.5)$	$\beta_I(0.75)$	$\beta_S(0.75)$
BEL	90.2%	79.0%	97.8%	98.8%	76.2%	77.2%
BETEL	88.4%	77.6%	97.2%	98.0%	76.2%	75.0%
CI	97.5%	95.7%	96.6%	96.2%	96.4%	94.7%
ETBB	98.4%	93.8%	95.2%	96.4%	95.2%	93.8%
$n = 100$	Width					
Method	$\beta_I(0.25)$	$\beta_S(0.25)$	$\beta_I(0.5)$	$\beta_S(0.5)$	$\beta_I(0.75)$	$\beta_S(0.75)$
BEL	0.66	0.37	1.06	0.59	0.66	0.36
BETEL	0.65	0.35	1.05	0.56	0.66	0.34
CI	1.10	0.58	1.01	0.53	1.12	0.58
ETBB	1.07	0.54	1.02	0.56	1.09	0.56
$n = 200$	Coverage					
Method	$\beta_I(0.25)$	$\beta_S(0.25)$	$\beta_I(0.5)$	$\beta_S(0.5)$	$\beta_I(0.75)$	$\beta_S(0.75)$
BEL	88.8%	75.2%	96.8%	97.4%	72.6%	78.6%
BETEL	88.2%	74.2%	97.0%	96.4%	72.8%	76.8%
CI	96.6%	95.2%	96.4%	95.4%	95.2%	95.8%
ETBB	97.4%	91.2%	95.6%	92.6%	93.4%	92.2%
$n = 200$	Width					
Method	$\beta_I(0.25)$	$\beta_S(0.25)$	$\beta_I(0.5)$	$\beta_S(0.5)$	$\beta_I(0.75)$	$\beta_S(0.75)$
BEL	0.45	0.24	0.73	0.39	0.46	0.24
BETEL	0.45	0.23	0.73	0.38	0.46	0.24
CI	0.77	0.39	0.71	0.37	0.77	0.40
ETBB	0.76	0.37	0.69	0.35	0.75	0.38

Table 4.8 presents the results; the values for the Bayesian approaches are given with respect to the posterior mean estimates. The performance of the ETBB estimators are similar to their corresponding BETEL estimators, which suggests that the ETBB is also an effective approach for doubly robust estimation. As before, the performance of BETEL and ETBB are similar to the Bayesian bootstrap approach of Saarela et al. (2016) when using the flat prior for all parameters. With the weakly informative prior, BETEL and ETBB substantially outperform both the standard doubly robust estimator and Saarela et al. (2016), particularly

when both models are misspecified, demonstrating the protective effect of such priors in the presence of model misspecification.

OR correct, PS correct					OR incorrect, PS correct				
Estimator	Bias	RMSE	MAE	ESD	Estimator	Bias	RMSE	MAE	ESD
$\hat{\mu}_{\text{DR}}$	-0.01	2.55	1.73	2.55	$\hat{\mu}_{\text{DR}}$	0.27	3.61	2.32	3.60
$\hat{\mu}_{\text{Sa}}$	0.01	2.57	1.71	2.57	$\hat{\mu}_{\text{Sa}}$	0.57	3.44	2.31	3.39
$\hat{\mu}_{\text{BETEL},1}$	-0.15	2.55	1.76	2.55	$\hat{\mu}_{\text{BETEL},1}$	0.49	3.81	2.25	3.78
$\hat{\mu}_{\text{ETBB},1}$	-0.15	2.57	1.71	2.56	$\hat{\mu}_{\text{ETBB},1}$	0.51	3.95	2.33	3.92
$\hat{\mu}_{\text{BETEL},2}$	-0.14	2.40	1.63	2.40	$\hat{\mu}_{\text{BETEL},2}$	0.48	3.27	2.01	3.24
$\hat{\mu}_{\text{ETBB},2}$	-0.15	2.40	1.67	2.40	$\hat{\mu}_{\text{ETBB},2}$	0.47	3.30	2.01	3.27
OR correct, PS incorrect					OR incorrect, PS incorrect				
Estimator	Bias	RMSE	MAE	ESD	Estimator	Bias	RMSE	MAE	ESD
$\hat{\mu}_{\text{DR}}$	-0.01	2.59	1.73	2.59	$\hat{\mu}_{\text{DR}}$	-6.44	38.52	3.64	37.97
$\hat{\mu}_{\text{Sa}}$	-0.09	2.60	1.73	2.60	$\hat{\mu}_{\text{Sa}}$	-4.81	15.41	3.38	14.64
$\hat{\mu}_{\text{BETEL},1}$	-0.22	2.90	1.76	2.89	$\hat{\mu}_{\text{BETEL},1}$	-8.21	18.61	4.21	16.71
$\hat{\mu}_{\text{ETBB},1}$	-0.16	2.74	1.76	2.74	$\hat{\mu}_{\text{ETBB},1}$	-7.73	16.67	4.25	14.77
$\hat{\mu}_{\text{BETEL},2}$	-0.15	2.43	1.66	2.43	$\hat{\mu}_{\text{BETEL},2}$	-3.51	6.71	3.38	5.72
$\hat{\mu}_{\text{ETBB},2}$	-0.14	2.51	1.63	2.51	$\hat{\mu}_{\text{ETBB},2}$	-3.86	8.18	3.52	7.22

Table 4.8 Monte Carlo simulations based on 1000 replicates using the standard doubly robust estimator, the Saarela et al. method, BETEL and ETBB. RMSE, root mean squared error; MAE, median of absolute errors; ESD, empirical standard deviation; DR, double robust; Sa, Saarela et al. (2016) proposal; OR, outcome regression; PS, propensity score.

4.6 Discusssion

In this chapter, we have developed a nonparametric Bayesian framework for moment condition inference. The motivation for the ETBB was to address several shortcomings of BETEL, namely, the practical and conceptual issues that arise from using the empirical distribution as a plug-in. Our solution was to replace the plug-in with a prior, inspired by a proposal by Kitamura and Otsu. Compared to the Kitamura and Otsu approach, however, it is relatively straightforward to sample from the ETBB posterior. We have developed two computational options: one based on the pre-conditioned Crank-Nicolson proposal, and

one using Hamiltonian Monte Carlo. The ETBB was shown to be an effective approach for various problems that are often handled using estimating equations.

The most important direction for future work is the development of theory to confirm our hypotheses. First, we have specified an improper prior, so it is necessary to verify that the ETBB posterior is proper; currently, we only have confirmation for a special case of a simple example (Proposition 4.1). Next, we would like to confirm that the ETBB posterior is indeed a noninformative limit of the Kitamura and Otsu proposal. This conjecture was supported by our extensive simulation results. And finally, to justify the use of the ETBB for estimating general functionals, it is of interest to prove a nonparametric Bernstein-von Mises theorem for the ETBB posterior of P, similar to what is known to hold for the Dirichlet process (Ghosal and van der Vaart, 2017).

As we have already discussed in §1.2.2.4, the key deficiency of the Bayesian bootstrap—and Dirichlet processes in general—is the inability to directly specify an informative prior on the target quantity, which is something that we have resolved with the ETBB. Bornn et al. (2019) provided a Bayesian justification for combining the Bayesian bootstrap with an informative prior through importance sampling. First, one draws a sample from the Bayesian bootstrap posterior; then, the importance weight assigned to each draw is proportional to its value under the prior density function. Finally, one obtains a new posterior sample by resampling according to the weights. However, the justification for this approach relies on the support of the data generating distribution being known and discrete, which is restrictive in practice. If this does not hold, one would have to make the assumption that the observed data forms the entire support, which suffers from the same type of data-dependence as BETEL and BEL.

A limitation of our work is the assumption that the dimension of the moment conditions equals the dimension of the parameter; that is, the parameter is exactly identified. There is substantial interest—-particularly in econometrics—in moment conditions that overidentify the parameter. The standard approach for such problems is *generalized method of moments* (Hansen, 1982), which minimizes a weighted version of the sample moment conditions, rather than setting them exactly to 0. Some developments in this direction have been established for BETEL by Chib et al. (2018). We conjecture that similar extensions can also be made for the ETBB.

Chapter 5

Conclusions and future work

In this thesis, we have developed Bayesian modelling approaches for an array of problems that are often handled using weights and/or estimating equations. A common thread that links our work is the ability to directly specify a prior on the quantity of interest while using nonparametric modelling. Although we recommend that the user specifies an informative prior based on subject matter knowledge, we acknowledge that there are situations where an objective prior might be appropriate. Investigating effective choices of objective priors for our proposed methodology is a topic of future research. It would be of interest to compare these choices with more standard nonparametric Bayesian methods such as Dirichlet processes.

Our work in Chapter 2 exploited the fact that the Cox partial likelihood can be derived from a Bayesian perspective by specifying a Bayesian bootstrap prior on the baseline cumulative hazard function, which is restricted to be a step function with jumps only at the failure times. We noted that the same restriction can also motivate the Cox partial likelihood as a profile likelihood function by maximizing the jump sizes. Indeed, this was the source of Breslow's estimator of the cumulative baseline hazard function (Breslow, 1972). A similar phenomenon that links the Bayesian bootstrap and profile likelihood was discovered by Seaman and Richardson (2004) in the context of logistic regression analysis of case-control studies. We believe that this connection may be far more general. An exploration of this hypothesis could potentially lead to a Bayesian framework for wide classes of problems where profile likelihood estimation is the norm. As with our method in Chapter 2, the Bayesian paradigm offers an attractive alternative due to the marginalization of nuisance parameters, and computational techniques such as the pseudo-marginal algorithm.

The approaches in Chapters 3 and 4 can be viewed as Bayesian analogues of estimating equation methodology. As discussed in Chapter 1, the state-of-the-art for some semiparametric problems involves the use of flexible machine learning methods to estimate nuisance

parameters, rather than estimating equations. It is not immediately clear whether our methods can be generalized to incorporate machine learning. One possibility is to fix estimates of the nuisance parameters beforehand—using machine learning, sample-splitting and cross-fitting—and then tilt only in the direction of the target quantity. The asymptotic theory in the frequentist setting suggests that neglecting to account for the uncertainty in the nuisance parameter estimates will not affect the asymptotic posterior variance, i.e. posterior credible intervals will still attain nominal coverage asymptotically. However, this compromise of using plug-in estimators would diminish the conceptual appeal of ETBB over BETEL and other alternatives.

Our approach in Chapter 2 differs from that of Chapters 3 and 4 due to the use of a more orthodox Bayesian model, albeit with an improper Bayesian bootstrap prior on the baseline cumulative hazard function. With BETEL and ETBB, we were in fact deliberately avoiding a more conventional Bayesian approach to circumvent problems such as the ones posed by the Robins-Ritov example (§1.2.1). Conversely, it is natural to wonder whether BETEL and ETBB could be effective for case-cohort Cox regression. Recall from §2.2.1 that the status quo for this problem is weighted Cox regression, which involves solving weighted versions of the Cox partial score equation (2.3). These equations do not take the form of a sum of independent and identically distributed terms. Thus, in order to apply BETEL or ETBB, we would have to use nested estimating equations, similar to our doubly robust estimation approach in Chapter 3, which would produce a set of estimating equations for every failure time point. For even moderately sized datasets, the corresponding set of moment conditions would likely be too high-dimensional to enable computation of the posterior.

Throughout this thesis, we have assumed that the dimension of the data does not exceed the sample size, enabling standard regression models and estimating equations to be used. However, this may preclude the use of our proposed methodology for many modern data science applications. As argued in §1.2.2.2, we believe that it is dangerous to use inverse probability weighting coupled with assumptions such as sparsity, due to the potential instability caused by model misspecification and practical violations of positivity. In such high-dimensional settings, it may be preferable to choose a more conventional outcome regression approach (Bayesian or otherwise) and settle with just prediction, rather than attempt to perform frequentist-calibrated inference under false pretenses.

We have discussed how BEL, BETEL and ETBB are all based on finding the distribution that minimizes the KL-divergence to some "initial estimate" of the data generating distribution subject to moment constraints. The empirical likelihood inputs the initial estimate as the first argument of the KL-divergence; this leads to a maximum likelihood/profile likelihood

interpretation (see §3.2.1). BETEL and ETBB do the opposite (the initial estimate is the second argument), which has information geometric justifications (Csiszár, 1975). The asymmetry of the KL-divergence may be considered unappealing. A possible direction for future work is to investigate the effectiveness of a symmetric difference measure, such as the total variation distance or the Hellinger distance, which are also examples of f-divergences (Csiszár and Shields, 2004).

With the exception of the "design setting" in Chapter 3, we have only considered sampling mechanisms that are ignorable. In the design setting, we implicitly assumed that the sampling mechanism is ignorable for the data collector given their possession of additional design information since they are able to provide weights that adjust for the selection bias. However, there are settings where neither of these scenarios can reasonably be assumed to be true. For example, some believe that Republican voters were less likely to respond to 2020 US presidential election polls and that the pollsters did not collect sufficient information to adjust for this, leading to systematic underestimation of Donald Trump's support (Panagopoulos, 2021). If so, the missing data were *not missing at random* (NMAR). In such cases, the target quantities are not identifiable without further assumptions specifiying the dependence of the sampling on the missing observations. Of course, such assumptions are untestable, so one must proceed with caution. A detailed discussion of NMAR can be found on pages 10-11 of Fitzmaurice et al. (2014). Nevertheless, if one is comfortable with making such assumptions and is able to fit the selection model with estimating equations, we speculate that both BETEL and ETBB can be used to perform inference.

To conclude, we hope that the projection-based view of statistical estimation will gain wider adoption in the Bayesian community. We have demonstrated that this perspective creates opportunities to apply Bayesian methodology to a wide range of problems that are considered to be difficult or unattractive to handle in a standard Bayesian set-up. It is also important that researchers should not feel the need to trade-off flexibility with interpretability. We can combine the many practical benefits of Bayesian inference with robust, frequentist-valid measures of uncertainty asymptotically, attaining practical and interpretable "Calibrated Bayes".

References

A. E. Ades and A. J. Sutton. Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. *Journal of the Royal Statistical Society, Series A*, 161:5–35, 2006.

J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B*, 44:139–177, 1982.

C. Andrieu and G. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*, 37:697–725, 2009.

W. E. Barlow. Robust variance estimation for the case-cohort design. *Biometrics*, 50: 1064–1072, 1994.

J. Barnard, R. McCulloch, and X.-L. Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10: 1281–1311, 2000.

D. Basu. An essay on the logical foundations of survey sampling, part I. In *Foundations of Statistical Inference (Proc. Sympos., Univ. Waterloo, Waterloo, Ont., 1970)*, pages 203–242. Holt, Rinehart and Winston, Toronto, 1971.

M. Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164:1139–1160, 2003.

M. Berkelaar. *lpSolve: Interface to 'Lp_solve' v. 5.5 to Solve Linear/Integer Programs*, 2015. R package.

M. Betancourt. Cruising the simplex: Hamiltonian Monte Carlo and the Dirichlet distribution. *AIP Conference Proceedings*, 1443, 2012.

P. Bissiri, C. Holmes, and S. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society, Series B*, 78:1103–1130, 2016.

D. Boos and J. Monahan. Bootstrap methods using prior information. *Biometrika*, 73:77–83, 1986.

Ø. Borgan and S. Samuelson. Cohort sampling for time-to-event data: an overview. In Ø. Borgan, N. Breslow, N. Chatterjee, M. Gail, A. Scott, and C. Wild, editors, *Handbook of Statistical Methods for Case-Control studies*, pages 285–301. CRC Press, Boca Raton, 2017.

Ø. Borgan, B. Langholz, S. Samuelson, L. Goldstein, and J. Pogoda. Exposure stratified case-cohort designs. *Lifetime data analysis*, 6:39–58, 2000.

L. Bornn, N. Shephard, and R. Solgi. Moment conditions and Bayesian nonparametrics. *Journal of the Royal Statistical Society, Series B*, 81:5–43, 2019.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

N. Breslow. Discussion of: Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, 34:216–218, 1972.

A. Buja et al. Models as approximations I: consequences illustrated with linear regression. *Statistical Science*, 34:523–544, 2019a.

A. Buja et al. Models as approximations II: a model-free theory of parametric regression. *Statistical Science*, 34:545–565, 2019b.

A. Cameron and P. Trivedi. *Microeconometrics Using Stata* . Stata Press, College Station, TX, 2010.

W. Cao, A. Tsiatis, and M. Davidian. Improving efficiency and robustness of the double robust estimator for a population mean with incomplete data. *Biometrika*, 96:723–734, 2009.

C. Cassel, C. Särndal, and J. Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63:615–620, 1976.

G. Chamberlain and G. Imbens. Nonparametric applications of Bayesian inference. *Journal of Business Economic Statistics*, 21:12–18, 2003.

K. Chan and S. Yam. Oracle, multiple robust and multipurpose calibration in a missing response problem. *Statistical Science*, 29:380–396, 2014.

K. Chen and S. Lo. Case-cohort and case-control analysis with Cox's model. *Biometrika*, 86: 755–764, 1999.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21:C1–C68, 2018.

S. Chib, M. Shin, and A. Simoni. Bayesian estimation and comparison of moment condition models. *Journal of the American Statistical Association*, 113:1656–1668, 2018.

H. Chipman, E. Geroge, and R. McCulloch. Bayesian ensemble learning. *Neural Information Processing Systems*, 19:265–272, 2007.

H. Chipman, E. Geroge, and R. McCulloch. BART: Bayesian Additive Regression Trees. *Annals of Applied Statistics*, 4:266–298, 2010.

M. Clyde and E. George. Model uncertainty. *Statistical Science*, 19:81–94, 2004.

C. Coombs. *A Theory of Data*. Wiley, New York, 1964.

S. Corcoran. Bartlett adjustment of empirical discrepancy statistics. *Biometrika*, 85:967–972, 1998.

S. Cotter, G. Roberts, A. Stuart, and D. White. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28:424–446, 2013.

D. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, 34:187–220, 1972.

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3:146–158, 1975.

I. Csiszár and P. Shields. Information theory and statistics: a tutorial. *Foundations and Trends in Communications and Information Theory*, 1:417–528, 2004.

A. Davison, D. Hinkley, and B. Worton. Bootstrap likelihoods. *Biometrika*, 79:113–130, 1992.

N. Day et al. EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. *British Journal of Cancer*, 80:95–103, 1999.

G. Deligiannidis, A. Doucet, and M. Pitt. The correlated pseudomarginal method. *Journal of the Royal Statistical Society, Series B*, 80:839–870, 2018.

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

V. Dorie, J. Hill, U. Shalit, M. Scott, and D. Cervone. Automated versus do-it-yourself methods for causal inference: lessons learned from a data analysis competition. *Statistical Science*, 34:43–68, 2019.

B. Efron. Bayes and likelihood calculations from confidence intervals. *Biometrika*, 80:3–26, 1993.

B. Efron and C. Morris. Stein's estimation rule and its competitors—an Empirical Bayes approach. *Journal of the American Statistical Association*, 68:117–130, 1973.

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York, 1994.

T. Ferguson. A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, 2: 209–230, 1973.

D. Firth and K. Bennett. Robust models in probability sampling. *Journal of the Royal Statistical Society, Series B*, 60:3–21, 1998.

G. Fitzmaurice et al. Introduction and preliminaries. In G. Molenberghs, G. Fitzmaurice, M. G. Kenward, A. Tsiatis, and G. Verbeke, editors, *Handbook of Missing Data Methodology*, pages 3–22. CRC Press, Boca Raton, 2014.

J. Florens and A. Simoni. Gaussian processes and Bayesian moment estimation. *Journal of Business Economic Statistics*, 2019.

N. Forouhi et al. Differences in the prospective association between individual plasma phospholipid saturated fatty acids and incident type 2 diabetes: the epic-interact case-cohort study. *Lancet Diabetes and Endocrinology*, 2:810–818, 2014.

J. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2001.

W. Fuller. *Sampling Statistics*. Wiley, Hoboken, NJ, 2009.

A. Gelman and C. Shalizi. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66:8–38, 2013.

A. Gelman, J. Carlin, H. Stern, and A. Rubin. *Bayesian Data Analysis*. CRC Press, Boca Raton, 2013.

S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, Cambridge, 2017.

J. Ghosh and R. Ramamoorthi. *Bayesian Nonparametrics*. Springer, New York, 2003.

D. Graham, E. McCoy, and D. Stephens. Approximate Bayesian inference for doubly robust estimation. *Bayesian Analysis*, 11:47–69, 2016.

M. Greenacre. *Compositional Data Analysis In Practice*. Chapman & Hall/CRC, New York, 2019.

P. Hahn, J. Murray, and C. Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15:965–1056, 2020.

J. Hájek. Discussion of: An essay on the logical foundations of survey sampling, part I. In *Foundations of Statistical Inference (Proc. Sympos., Univ. Waterloo, Waterloo, Ont., 1970)*. Holt, Rinehart and Winston, Toronto, 1971.

L. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054, 1982.

M. Henmi and S. Eguchi. A paradox concerning nuisance parameters and projected estimating functions. *Biometrika*, 91:929–941, 2004.

P. Hoff and J. Wakefield. Bayesian sandwich posteriors for pseudo-true parameters. *Journal of Statistical Planning and Inference*, 143:1638–1642, 2013.

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

L. Huang et al. Circulating saturated fatty acids and incident type 2 diabetes: A systematic review and meta-analysis. *Nutrients*, 11:5, 2019.

H. Ishwaran and L. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173, 2001.

P. Jacob, L. Murray, C. Holmes, and C. Robert. Better together? Statistical learning in models made of modules. *arXiv*, page 1708.08719, 2017.

B.-Y. Jing and A. Wood. Exponential empirical likelihood is not Bartlett correctable. *Annals of Statistics*, 24:365–369, 1996.

J. Kalbfleisch. Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B*, 40:214–221, 1978.

J. Kalbfleisch and J. Lawless. Likelihood analysis for multi-state models for disease incidence and mortality. *Statistics in Medicine*, 7:149–160, 1988.

J. Kang and J. Schafer. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science*, 22:523–39, 2007.

R. Keogh and I. White. Using full-cohort data in nested case-control and case-cohort studies by multiple imputation. *Statistics in Medicine*, 32:4021–4043, 2013.

D. Kessler, P. Hoff, and D. Dunson. Marginally specified priors for nonparametric Bayesian estimation. *Journal of the Royal Statistical Society, Series B*, 77:35–58, 2015.

Y. Kim and J. Lee. Bayesian bootstrap for proportional hazards model. *Annals of Statistics*, 31:1905–1922, 2003.

B. Kleijn and A. van der Vaart. The Bernstein-von Mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

R. Koenker and G. J. Bassett. Regression quantiles. *Econometrica*, 46:33–50, 1978.

S. Kulathinal and E. Arjas. Bayesian inference from case-cohort data with multiple end-points. *Scandinavian Journal of Statistics*, 33:25–36, 2006.

M. Kulich and D. Y. Lin. Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association*, 99:832–844, 2004.

T. Lancaster and S. Jun. Bayesian quantile regression methods. *Journal of Applied Econometrics*, 25:287–307, 2010.

C. Langenberg et al. Design and cohort description of the interact project: an examination of the interaction of genetic and lifestyle factors on the incidence of type 2 diabetes in the epic study. *Diabetologia*, 54:2272–2282, 2011.

S. Lee and G. Young. Nonparametric likelihood ratio confidence intervals. *Biometrika*, 86: 107–118, 1999.

J. Lin. Erythrocyte saturated fatty acids and incident type 2 diabetes in Chinese men and women: A prospective cohort study. *Nutrients*, 10, 2018.

L. Lin, K. F. Liu, and J. Sloan. A noisy Monte Carlo algorithm. *Physical Review D*, 61: 074505, 2000.

R. Little. Calibrated Bayes, for statistics in general, and missing data in particular. *Statistical Science*, 26:162–174, 2011.

Y. Liu and R. Goudie. Stochastic approximation cut algorithm for inference in modularized Bayesian models. *arXiv*, page 2006.01584, 2020.

W. Lu and A. Tsiatis. Semiparametric transformation models for the case-cohort study. *Biometrika*, 93:207–214, 2006.

Y. Lu et al. Serum lipids in association with type 2 diabetes risk and prevalence in a Chinese population. *Journal of Clinical Endocrinology and Metabolism*, 103:671–680, 2018.

T. Lumley. *Complex Surveys: A Guide to Analysis using R*. Wiley, Hoboken, NJ, 2010.

S. Lyddon, S. Walker, and C. Holmes. Nonparametric learning from Bayesian models with randomized objective functions. *Advances in Neural Information Processing Systems*, 31: 265–272, 2018.

L. McCandless, I. Douglas, S. Evans, and L. Smeeth. Cutting feedback in Bayesian regression adjustment for the propensity score. *The International Journal of Biostatistics*, 6, 2010.

D. McGregor, J. Palarea-Albaladejo, P. Dall, K. Hron, and S. Chastin. Cox regression survival analysis with compositional covariates: application to modelling mortality risk from 24-h physical activity patterns. *Statistical Methods in Medical Research*, 29:1386–1402, 2020.

X.-L. Meng. Statistical paradises and paradoxes in big data (I): law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics*, 12: 685–726, 2018.

J. Monahan and D. Boos. Proper likelihoods for Bayesian analysis. *Biometrika*, 79:271–278, 1992.

S. Morgan and C. Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, Cambridge, 2007.

T. Morris, I. White, P. Royston, S. Seaman, and A. Wood. Multiple imputation for an incomplete covariate that is a ratio. *Statistics in Medicine*, 33:88–104, 2013.

B. Nan, M. Emond, and J. Wellner. Information bounds for Cox regression models with missing data. *Annals of Statistics*, 32:723–753, 2004.

R. Neal. MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*, pages 116–162. CRC Press, Boca Raton, 2011.

P. Newcombe, S. Connolly, S. Seaman, S. Richardson, and S. Sharp. A two-step method for variable selection in the analysis of a case-cohort study. *International Journal of Epidemiology*, 47:597–604, 2018.

M. Newton and A. Raftery. Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B*, 56:3–48, 1994.

A. Ni, J. C, and D. Zeng. Variable selection for case-cohort studies with failure time outcome. *Biometrika*, 103:547–562, 2016.

A. Owen. *Empirical Likelihood*. Chapman & Hall/CRC, New York, 2001.

C. Panagopoulos. Accuracy and bias in the 2020 U.S. general election polls. *Presidential Studies Quarterly*, 51:214–227, 2021.

K. Pearson. Mathematical contributions to the theory of evolution on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London LX*, pages 489–502, 1897.

D. Pfeffermann, F. A. S. Moura, and P. L. Nascimento-Silva. Multi-level modelling under informative sampling. *Biometrika*, 93:943–959, 2006.

M. Plummer. Cuts in Bayesian graphical models. *Statistics and Computing*, 25:37–43, 2015.

R. Prentice. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73:1–11, 1986.

J. Qin and B. Zhang. Empirical-likelihood-based inference in missing response problems and its application in observational studies. *Journal of the Royal Statistical Society, Series B*, 69:101–122, 2007.

K. Ray and A. van der Vaart. Semiparametric Bayesian causal inference using Gaussian process priors. *arXiv*, page 1808.04246v1, 2018.

J. Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*. Springer, New York, 2004.

J. Robins and Y. Ritov. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16:285–319, 1997.

J. Robins and L. Wasserman. Robins and Wasserman respond to a Nobel prize winner, 2012a. URL https://normaldeviate.wordpress.com/2012/08/28/.

J. Robins and L. Wasserman. Robins and Wasserman respond to a Nobel prize winner continued: a counterexample to Bayesian inference?, 2012b. URL https://normaldeviate.wordpress.com/2012/09/02/.

J. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89: 846–866, 1994.

J. Robins, A. Rotnitzky, and M. van der Laan. Comment: On profile likelihood. *Journal of the American Statistical Association*, 95:477–482, 2000.

J. Robins, M. Hernán, and L. Wasserman. Discussion of: On Bayesian estimation of marginal structural models. *Biometrics*, 71:296–299, 2015.

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

A. Rotnitzky and S. Vansteelandt. Double-robust methods. In G. Molenberghs, G. Fitzmaurice, M. G. Kenward, A. Tsiatis, and G. Verbeke, editors, *Handbook of Missing Data Methodology*, pages 185–212. CRC Press, Boca Raton, 2014.

A. Rotnitzky, Q. Lei, M. Sued, and J. Robins. Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99:439–456, 2012.

D. Rubin. The Bayesian bootstrap. *Annals of Statistics*, 9:130–134, 1981.

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

O. Saarela, L. Belzile, and D. Stephens. A Bayesian view of doubly robust causal inference. *Biometrika*, 103:667–681, 2016.

C. Särndal, B. Swensson, and J. Wretman. *Model-assisted Survey Sampling*. Springer-Verlag, New York, 1992.

D. Scharfstein, A. Rotnitzky, and J. Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association*, 94:1096–1146, 1999.

T. Scheike and A. Juul. Maximum likelihood estimation for Cox's regression model under nested case-control sampling. *Biostatistics*, 5:193–206, 2004.

T. Scheike and T. Martinussen. Maximum likelihood estimation for Cox's regression model under case–cohort sampling. *Scandinavian Journal of Statistics*, 31:283–293, 2004.

S. Schennach. Bayesian exponentially tilted empirical likelihood. *Biometrika*, 92:31–46, 2005.

S. Schennach. Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics*, 35:634–672, 2007.

S. Seaman and S. Richardson. Equivalence of prospective and retrospective models in the bayesian analysis of case-control studies. *Biometrika*, 91:15–25, 2004.

S. Self and R. Prentice. Asymptotic distribution theory and efficiency results for case-cohort studies. *Annals of Statistics*, 16:64–81, 1988.

S. Sharp, M. Poulaliou, S. Thompson, I. White, and A. Wood. A review of published analyses of case-cohort studies and recommendations for future reporting. *PLoS One*, 9:e101176, 2014.

C. Sherlock, A. Thiery, and A. Lee. Pseudo-marginal Metropolis-Hastings sampling using averages of unbiased estimators. *Biometrika*, 104:727–734, 2017.

Y. Si, N. Pillai, and A. Gelman. Bayesian nonparametric weighted sampling inference. *Bayesian Analysis*, 10:605–625, 2015.

C. Sims. On an example of Larry Wasserman, Round 2. http://sims.princeton.edu/yftp/WassermanExmpl/WassermanR2.pdf, 2012a. accessed 2021-02-02.

C. Sims. On an example of Larry Wasserman, Round 3. http://sims.princeton.edu/yftp/WassermanExmpl/WassermanR3.pdf, 2012b. accessed 2021-02-02.

C. Sims. On an example of Larry Wasserman, Round 4. http://sims.princeton.edu/yftp/WassermanExmpl/WassermanR4.pdf, 2012c. accessed 2021-02-02.

C. Sims. Robins-Wasserman, Round N. http://sims.princeton.edu/yftp/WassermanExmpl/WassermanR4a.pdf, 2012d. accessed 2021-02-02.

D. Sinha, J. Ibrahim, and M. Chen. A Bayesian justification of Cox's partial likelihood. *Biometrika*, 90:629–641, 2003.

C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Proc. Third Berkeley Symp. Math. Statist. Prob.*, 1:197–206, 1956.

J. Steingrimsson and R. Strawderman. Estimation in the semiparametric accelerated failure time model with missing covariates: Improving efficiency through augmentation. *Journal of the American Statistical Association*, 112:1221–1235, 2017.

M. Tanner and W. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540, 1987.

D. Thomas. Addendum to: "Methods of cohort analysis: appraisal by application to asbestos mining," by Liddell, F.D.K., McDonald, J.C.& Thomas, D.C. *Journal of the Royal Statistical Society, Series A*, 140:469–491, 1977.

A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer-Verlag, New York, 2006.

A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

M. van der Laan and J. Robins. *Unified Methods for Censored Longitudinal Data and Causality.* Springer-Verlag, New York, 2003.

M. van der Laan and S. Rose. *Targeted Learning.* Springer-Verlag, New York, 2011.

M. van der Laan and S. Rose. *Targeted Learning in Data Science.* Springer-Verlag, New York, 2018.

A. van der Vaart. Asymptotic linearity of minimax estimators. *Statistica Neerlandica*, 2–3: 179–194, 1992.

A. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

A. van der Vaart. Semiparametric statistics. In P. Bernard, editor, *Lectures on Probability Theory and Statistics*, pages 331–457. Springer Verlag, Berlin, 2002.

E. van Zwet and E. Cator. The significance filter, the winner's curse and the need to shrink. *arXiv*, page 2009.09440, 2020.

S. Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics - Simulation and Computation*, 36:45–54, 2007.

Z. Wang, J. Kim, and S. Yang. Approximate Bayesian inference under informative sampling. *Biometrika*, 105:91–102, 2017.

H. White. A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica*, 48:817–838, 1980.

A. Wilson. The case for Bayesian deep learning. *arXiv*, page 2001.10995, 2020.

Y. Yang and X. He. Bayesian empirical likelihood for quantile regression. *Annals of Statistics*, 40:1102–1131, 2012.

S. Zanganeh and R. Little. Bayesian inference for the finite population total from a heteroscedastic probability proportional to size sample. *Journal of Survey Statistics and Methodology*, 3:91–102, 2015.

A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243, 1986.

D. Zeng and D. Lin. Efficient estimation of semiparametric transformation models for two-phase cohort studies. *Journal of the American Statistical Association*, 109:371–383, 2014.

Q. Zhao, P. Ju, S. Bacallado, and R. Shah. BETS: The dangers of selection bias in early analyses of the coronavirus disease (COVID-19) pandemic. *Annals of Applied Statistics*, 2021. (to appear).

C. Zigler, K. Watts, R. Yeh, Y. Wang, and B. Coull. Model feedback in Bayesian propensity score estimation. *Biometrics*, 69:263–273, 2013.

Appendix A

Appendix for Chapter 1

A.1 Proof of Theorem 1.1

By Lemma 1.6 of van der Vaart (2002), a parametric submodel $\{P_{t,g}\}$ with score function $g \in \dot{\mathscr{P}}_P$ that is differentiable in quadratic mean admits the expansion

$$\log \prod_{i=1}^{n} \frac{dP_{1/\sqrt{n}}}{dP}(D_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(D_i) - \frac{1}{2} Pg^2 + o_P(1),$$

and therefore converges in distribution under P to $\mathscr{N}(-\frac{1}{2}Pg^2, \frac{1}{2}Pg^2)$. By applying the continuous mapping theorem, $dP^n_{1/\sqrt{n}}/dP^n$ converges in distribution under P to a log-normal distribution, which is strictly positive with probability 1. Le Cam's first lemma (Lemma 6.4 in van der Vaart (1998)) implies that the sequence $P^n_{1/\sqrt{n}}$ is contiguous with respect to P^n.

Now consider

$$\begin{pmatrix} \sqrt{n}(\hat{\mu} - \mu(P)) \\ \log \prod_{i=1}^{n} \frac{dP_{1/\sqrt{n}}}{dP}(D_i) \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(D_i) \\ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(D_i) - \frac{1}{2} Pg^2 \end{pmatrix} + o_P(1).$$

By the multivariate central limit theorem and Slutsky's lemma, the left hand side converges in distribution under P to

$$\mathscr{N}\left(\begin{pmatrix} 0 \\ -\frac{1}{2}Pg^2 \end{pmatrix}, \begin{pmatrix} P\psi^2 & P[\psi g] \\ P[\psi g] & Pg^2 \end{pmatrix} \right).$$

By Le Cam's third lemma (Example 6.7 of van der Vaart (1998)),

$$\sqrt{n}(\hat{\mu} - \mu(P)) \overset{P_{1/\sqrt{n},g}}{\rightsquigarrow} \mathscr{N}(P[\psi g], P\psi^2).$$

We deduce that

$$\sqrt{n}(\hat{\mu} - \mu(P_{1/\sqrt{n},g})) = \sqrt{n}(\hat{\mu} - \mu(P)) - \sqrt{n}(\mu(P_{1/\sqrt{n},g}) - \mu(P))$$
$$\overset{P_{1/\sqrt{n},g}}{\rightsquigarrow} \mathcal{N}(0, P\psi^2) + P[\psi g] - \frac{\partial \mu(P_{t,g})}{\partial t}|_{t=0}.$$

Thus, $\hat{\mu}$ is regular if and only if

$$P[\psi g] = \frac{\partial \mu(P_{t,g})}{\partial t}|_{t=0}. \quad \square$$

A.2 Characterizing the mean-zero gradients for known π

Let Q be the joint distribution of Y and X induced by P. In the nonparametric complete data model, where Y and X are always observed, the unique mean-zero gradient for μ is $Y - \mu$. That Y is a gradient can be seen from

$$\frac{\partial \mu(Q_{t,g})}{\partial t}|_{t=0} = \frac{\partial}{\partial t} \int y dQ_{t,g}(x,y)|_{t=0} = \int y g(x,y) dQ(x,y).$$

If f is orthogonal to the space, then $P[(f - Pf)^2] = 0$, since $f - Pf$ is contained in the tangent space, i.e. f is a constant. Thus, $Y - \mu$ is the unique mean-zero gradient.

Let $r \cdot s(y \mid x) + s(x)$ be a score for the incomplete data model. Then

$$R \cdot s(Y \mid X) + s(X) = \mathbb{E}_{Q,\pi}[s(Y \mid X) + s(X) \mid D].$$

This follows from writing

$$\mathbb{E}_{Q,\pi}[s(Y \mid X) \mid D] = \underbrace{\mathbb{E}_{Q,\pi}[R \cdot s(Y \mid X) \mid D]}_{=R \cdot s(Y|X)} + \underbrace{\mathbb{E}_{Q,\pi}[(1-R) \cdot s(Y \mid X) \mid D]}_{=(1-R)\mathbb{E}_P[s(Y|X)|X]=0}.$$

We can show that $RY/\pi(X) - \mu$, the influence function of the Horvitz-Thompson estimator, is a gradient. Using iterated expectations,

$$\begin{aligned}
\mathbb{E}_P\left[\left(\frac{RY}{\pi(X)} - \mu\right)\{R \cdot s(Y \mid X) + s(X)\}\right] &= \mathbb{E}_P\left[\left(\frac{RY}{\pi(X)} - \mu\right)\mathbb{E}_P\{s(Y \mid X) + s(X) \mid D\}\right] \\
&= \mathbb{E}_Q\left[\mathbb{E}_\pi\left\{\left(\frac{RY}{\pi(X)} - \mu\right) \mid Y, X\right\}[s(Y \mid X) + s(X)]\right] \\
&= \mathbb{E}_Q[(Y - \mu)\{s(Y \mid X) + s(X)\}].
\end{aligned}$$

We claim that any mean-zero gradient for the complete data model can be written in the form

$$\frac{RY}{\pi(X)} - \mu + b(D),$$

where b is a one-dimensional measurable function satisfying $Pb^2 < \infty$ and $\mathbb{E}_\pi[b(D) \mid Y,X] = 0$, and conversely, any function of this form is a mean-zero gradient.

We prove the converse first. It suffices to show that such a b is orthogonal to the tangent space. Keeping the same notation as above,

$$\begin{aligned}\mathbb{E}_P[b(D)\{R \cdot s(Y \mid X) + s(X)\}] &= \mathbb{E}_P[b(D)\mathbb{E}_{Q,\pi}\{s(Y \mid X) + s(X) \mid D\}] \\ &= \mathbb{E}_Q[\mathbb{E}_\pi[b(D) \mid Y,X]\{s(Y \mid X) + s(X)\}] \\ &= 0.\end{aligned}$$

Now let ψ be a mean-zero gradient of the incomplete data model. It can be written in the form

$$\psi(D) = \left(\frac{RY}{\pi(X)} - \mu\right) + \left[\psi(D) - \left(\frac{RY}{\pi(X)} - \mu\right)\right]. \tag{A.1}$$

Since

$$\mathbb{E}_Q[\mathbb{E}_\pi[\psi(D) \mid Y,X]\{s(Y \mid X) + s(X)\}] = \mathbb{E}_P[\psi(D)\{R \cdot s(Y \mid X) + s(X)\}],$$

the conditional expectation $\mathbb{E}_\pi[\psi(D) \mid Y,X]$ must be a mean-zero gradient of the complete data model. Thus, it must be equal to $Y - \mu$ by uniqueness. The same is true for the conditional expectation of $RY/\pi(X) - \mu$, since it is also a mean-zero gradient of the incomplete data model. Therefore, the term in square brackets in (A.1) qualifies as a function b.

It remains to characterize the set of functions b. We can write

$$b(D) = R \cdot b_1(Y,X) + (1-R) \cdot b_2(X), \tag{A.2}$$

where b_2 cannot depend on Y because Y is unobserved if $R = 0$. Thus,

$$0 = \mathbb{E}[b(D) \mid Y,X] = \pi(X) \cdot b_1(Y,X) + (1 - \pi(X)) \cdot b_2(X).$$

By rearranging, we see that

$$b_1(Y,X) = \frac{\pi(X) - 1}{\pi(X)} b_2(X).$$

Thus,

$$b(D) = \frac{\pi(X) - R}{\pi(X)} b_2(X),$$

ranging over arbitrary b_2. □

A.3 Calculating the influence function for an estimator with estimated weights

Let $\hat{\mu}_c$ be the estimator that solves

$$\mathbb{P}_n \left(\frac{R(Y - c(X))}{\pi(X)} + c(X) - \mu \right) = 0$$

and therefore has influence function

$$\psi_c(D) = \frac{R(Y - c(X))}{\pi(X)} + c(X) - \mu(P).$$

With a slight abuse of notation, let

$$\psi_c(D;\alpha) = \frac{R(Y - c(X;\alpha))}{\pi(X;\alpha)} + c(X;\alpha) - \mu(P)$$

so that $\psi_c(D) = \psi_c(D;\alpha_0)$. Let $\hat{\mu}_c^*$ be the estimator that solves

$$\mathbb{P}_n \left(\frac{R(Y - c(X;\hat{\alpha}))}{\pi(X;\hat{\alpha})} + c(X;\hat{\alpha}) - \mu \right) = 0$$

where $\hat{\alpha}$ solves

$$\mathbb{P}_n\{S_\alpha(R,X;\alpha)\} = 0.$$

Since $(\hat{\mu}_c^*, \hat{\alpha})$ jointly solve a set of unbiased estimating equations, the theory of Z-estimation (van der Vaart, 1998) implies that the influence function of $\hat{\mu}_c^*$ is

$$\psi_c^*(D) = \psi_c(D) + P\left(\frac{\partial \psi_c}{\partial \alpha}(D;\alpha)|_{\alpha=\alpha_0} \right)^{\mathrm{T}} i_{\alpha_0}^{-1} S_\alpha(R,X,\alpha_0),$$

where i_{α_0} is the Fisher information

$$i_{\alpha_0} = P\{S_\alpha(R,X,\alpha_0) S_\alpha(R,X,\alpha_0)^{\mathrm{T}}\}.$$

Let P_α be the probability measure derived from replacing $\pi(X)$ in P with $\pi(X;\alpha)$, keeping all else fixed. The product rule implies that

$$\frac{\partial}{\partial\alpha}P_\alpha\{\psi_c(D;\alpha)\}|_{\alpha=\alpha_0} = P\left(\frac{\partial\psi_c}{\partial\alpha}(D;\alpha)|_{\alpha=\alpha_0}\right) + P\{\psi_c(D)S_\alpha(R,X,\alpha_0)\}.$$

The first term on the right arises from differentiating ψ_c while keeping P_α fixed; the second is the other way round. But $P_\alpha\{\psi_c(D;\alpha)\} = 0$ for all α, so the left-hand side is 0. Thus,

$$\psi_c^*(D) = \psi_c(D) - P\{\psi_c(D)S_\alpha(R,X,\alpha_0)\}^{\mathrm{T}} i_{\alpha_0}^{-1} S_\alpha(R,X,\alpha_0);$$

that is, ψ_c^* is equal to ψ_c minus its orthogonal projection onto the linear space spanned by the components of $S_\alpha(R,X,\alpha_0)$.

A.4 The unique mean-zero gradient in the nonparametric model is ψ_{eff}

Any one-dimensional score $b(D)$ for the $p(r \mid x)$ model must satisfy $\mathbb{E}_\pi[b(D) \mid Y,X] = 0$ and $Pb^2 < \infty$. In §A.2, we already showed that such functions are characterized by

$$b(D) = \frac{R-\pi(X)}{\pi(X)}b_2(X), \tag{A.3}$$

ranging over arbitrary b_2. They are all scores; we can see this by considering the one-dimensional parametric submodels defined by

$$\pi(X;t) \equiv \pi(X) + tb_2(X)[1-\pi(X)].$$

Recall the decomposition (1.12) of ψ_c. If we take $b_2 \equiv m - c$, then b is exactly equal to term ②. Since ② is orthogonal to terms ① and ③, its covariance with ψ_c is the variance of ②. This is zero if and only if $c \equiv m$. Furthermore, ψ_{eff} is indeed orthogonal to any b of the form (A.3). Hence, ψ_{eff} is the unique element of (1.10) that is orthogonal to all scores for π. Since ψ_{eff} is a gradient in the model where π is known, and μ does not depend on π, we deduce that ψ_{eff} is still a gradient in the nonparametric model.

A.5 Sample splitting

Let $\hat{\psi}_{\text{eff}}^{(1)}$ be the estimated efficient influence function with $\hat{m} \equiv \hat{m}^{(1)}$ and $\hat{\pi} \equiv \hat{\pi}^{(1)}$, which are estimated from $D_{(n/2)+1}, \ldots, D_n$. Term ③ in (1.18) for samples $D_1, \ldots, D_n$ satisfies

$$\mathbb{E}[\sqrt{n}(\mathbb{P}_{n/2} - P)[\hat{\psi}_{\text{eff}}^{(1)} - \psi_{\text{eff}}] \mid D_{(n/2)+1}, \ldots, D_n] = 0$$

and

$$\begin{aligned}\text{var}[\sqrt{n}(\mathbb{P}_{n/2} - P)[\hat{\psi}_{\text{eff}}^{(1)} - \psi_{\text{eff}}] \mid D_{(n/2)+1}, \ldots, D_n] &= 2\text{var}[\hat{\psi}_{\text{eff}}^{(1)}(D_1) - \psi_{\text{eff}}(D_1) \mid D_{(n/2)+1}, \ldots, D_n] \\ &\leqslant 2\|\hat{\psi}_{\text{eff}}^{(1)} - \psi_{\text{eff}}\|^2.\end{aligned}$$

By Chebyshev's inequality, we deduce that

$$\sqrt{n}(\mathbb{P}_{n/2} - P)[\hat{\psi}_{\text{eff}}^{(1)} - \psi_{\text{eff}}] = O_P(\|\hat{\psi}_{\text{eff}}^{(1)} - \psi_{\text{eff}}\|) = o_P(1),$$

and hence,

$$\sqrt{n}(\hat{\mu}_{\text{DR}}^{(1)} - \mu(P)) \equiv \frac{2}{\sqrt{n}} \sum_{i=1}^{n/2} \hat{\psi}_{\text{eff}}^{(1)}(D_i) = \frac{2}{\sqrt{n}} \sum_{i=1}^{n/2} \psi_{\text{eff}}(D_i) + o_P(1).$$

By swapping the roles of the two halves, we similarly obtain:

$$\sqrt{n}(\hat{\mu}_{\text{DR}}^{(2)} - \mu(P)) \equiv \frac{2}{\sqrt{n}} \sum_{i=(n/2)+1}^{n} \hat{\psi}_{\text{eff}}^{(2)}(D_i) = \frac{2}{\sqrt{n}} \sum_{i=(n/2)+1}^{n} \psi_{\text{eff}}(D_i) + o_P(1).$$

Finally, letting $\check{\mu}_{\text{DR}} = (\hat{\mu}_{\text{DR}}^{(1)} + \hat{\mu}_{\text{DR}}^{(2)})/2$ yields

$$\sqrt{n}(\check{\mu}_{\text{DR}} - \mu(P)) = \sqrt{n}\mathbb{P}_n[\psi_{\text{eff}}] + o_P(1).$$

Appendix B

Appendix for Chapter 2

B.1 Derivation of the marginal posterior of β

We provide a more detailed derivation of expression (2.8). First, (2.6) is proportional to

$$\left[\prod_{i\in\mathscr{S}} \exp(\beta_1^{\mathrm{T}} Z_i + \beta_2^{\mathrm{T}} W_i)^{\Delta_i} \exp\left\{-e^{\beta_1^{\mathrm{T}} Z_i + \beta_2^{\mathrm{T}} W_i} \Lambda_0(Y_i)\right\}\right]$$
$$\left[\prod_{j\in\bar{\mathscr{S}}} \int \exp\left\{-e^{\beta_1^{\mathrm{T}} z_j + \beta_2^{\mathrm{T}} W_j} \Lambda_0(Y_j)\right\} p(z_j \mid W_j, X_j, \gamma) dz_j\right] p(\gamma \mid D_{\mathscr{S}}) p(\beta),$$

where we have incorporated the restricted posterior of γ. Then, we integrate with respect to Λ_0 and apply Fubini's theorem to bring the Λ_0 integral inside:

$$\int_{\{z_j : j\in\bar{\mathscr{S}}\}} \int_{\Lambda_0} \left[\prod_{i\in\mathscr{S}} \exp(\beta_1^{\mathrm{T}} Z_i + \beta_2^{\mathrm{T}} W_i)^{\Delta_i} \exp\left\{-e^{\beta_1^{\mathrm{T}} Z_i + \beta_2^{\mathrm{T}} W_i} \Lambda_0(Y_i)\right\}\right]$$
$$\left[\prod_{j\in\bar{\mathscr{S}}} \exp\left\{-e^{\beta_1^{\mathrm{T}} z_j + \beta_2^{\mathrm{T}} W_j} \Lambda_0(Y_j)\right\}\right] d\Lambda_0 \left[\prod_{k\in\bar{\mathscr{S}}} p(z_k \mid W_k, X_k, \gamma) dz_k\right] p(\gamma \mid D_{\mathscr{S}}) p(\beta). \tag{B.1}$$

The Λ_0 integral on the inside can be rewritten as

$$\int_{\Lambda_0} \prod_{k=1}^{n} \left[\exp(\beta_1^{\mathrm{T}} \tilde{Z}_k + \beta_2^{\mathrm{T}} W_k) \exp\left\{-\Delta\Lambda_0(Y_k) \sum_{l=1}^{n} R_l(T_k) e^{\beta_1^{\mathrm{T}} \tilde{Z}_l + \beta_2^{\mathrm{T}} W_l}\right\}\right]^{\Delta_k} d\Lambda_0$$

where $\tilde{Z}_k$ equals Z_k if $k \in \mathscr{S}$ and equals z_k otherwise. Integrating out each $\Delta\Lambda_0(Y_k)$ yields

$$\prod_{k=1}^{n} \left\{ \frac{\exp(\beta_1^{\mathrm{T}} \tilde{Z}_k + \beta_2^{\mathrm{T}} W_k)}{\sum_{l=1}^{n} R_l(T_k) \exp(\beta_1^{\mathrm{T}} \tilde{Z}_l + \beta_2^{\mathrm{T}} W_l)} \right\}^{\Delta_k}.$$

Substituting this back into (B.1) and then integrating with respect to γ yields (2.8).

B.2 Justification of Algorithm 2.2

Let $\gamma = (\gamma_1, \ldots, \gamma_B)$. To justify Algorithm 2.2, it is sufficient to check that detailed balance holds for (U, γ). This amounts to showing that

$$\phi(u; 0_M, I_M) p(\gamma \mid \{D_i : i \in \mathscr{S}\}) K\{(u,\gamma),(\tilde{u},\tilde{\gamma})\} = \phi(\tilde{u}; 0_M, I_M) p(\tilde{\gamma} \mid \{D_i : i \in \mathscr{S}\}) K\{(\tilde{u},\tilde{\gamma}),(u,\gamma)\} \tag{B.2}$$

where $\phi(\cdot; \mu, \Sigma)$ is the density function of $\mathscr{N}(\mu, \Sigma)$ and $K\{(u,\gamma),(\tilde{u},\tilde{\gamma})\} = \phi(\tilde{u}; \rho u, (1-\rho^2) I_M) p(\tilde{\gamma} \mid \{D_i : i \in \mathscr{S}\})$. Clearly, the terms involving γ on both sides of (B.2) match. Furthermore,

$$\begin{aligned}
\phi(u; 0_M, I_M)\phi(\tilde{u}; \rho u, (1-\rho^2) I_M) &= (2\pi)^{-M} (1-\rho^2)^{-M/2} \exp\left\{ \frac{1}{2} \left[u^{\mathrm{T}} u + \frac{(\tilde{u} - \rho u)^{\mathrm{T}} (\tilde{u} - \rho u)}{1-\rho^2} \right] \right\} \\
&= (2\pi)^{-M} (1-\rho^2)^{-M/2} \exp\left\{ \frac{1}{2} \left[\tilde{u}^{\mathrm{T}} \tilde{u} + \frac{(u - \rho \tilde{u})^{\mathrm{T}} (u - \rho \tilde{u})}{1-\rho^2} \right] \right\} \\
&= \phi(\tilde{u}; 0_M, I_M)\phi(u; \rho \tilde{u}, (1-\rho^2) I_M),
\end{aligned}$$

which establishes (B.2).

B.3 Application computation

We set $B = 1$. First, consider sampling ξ given $(\Sigma, Z_{\mathscr{S}}, W_{\mathscr{S}}, X_{\mathscr{S}})$. Let $C = (V_{\mathscr{S}}^{\mathrm{T}} V_{\mathscr{S}})^{-1}$. Since C and Σ are both positive definite, they possess unique positive definite square roots $C^{1/2}$ and $\Sigma^{1/2}$ respectively. Let $U_\xi \sim \mathscr{MN}(0_{13\times 9}, I_{13\times 13}, I_{9\times 9})$—or equivalently, let U_ξ be a 13×9 matrix where the entries are independent $\mathscr{N}(0,1)$ variables—independent of $(\Sigma, Z_{\mathscr{S}}, W_{\mathscr{S}}, X_{\mathscr{S}})$. Then,

$$\varphi_\xi(U_\xi, \Sigma, Z_{\mathscr{S}}, W_{\mathscr{S}}, X_{\mathscr{S}}) = \hat{\xi} + C^{1/2} U_\xi \Sigma^{1/2}$$

has the conditional distribution (2.12).

Next, consider sampling Z^{mis} given $(W_{\bar{\mathscr{S}}}, X_{\bar{\mathscr{S}}}, \xi, \Sigma)$. With $U_Z \sim \mathscr{N}(0_9, I_{9\times 9})$ independent of $(W_{\bar{\mathscr{S}}}, X_{\bar{\mathscr{S}}}, \xi, \Sigma)$,

$$\varphi_Z(U_Z, W_{\bar{\mathscr{S}}}, X_{\bar{\mathscr{S}}}, \xi, \Sigma) = \Sigma^{1/2} U_Z + \xi^{\mathrm{T}} V_{\bar{\mathscr{S}}}$$

has conditional distribution equal to (2.11) for the missing values of Z.

The sampling algorithm is described in Algorithm B.1. The correlation parameters ρ_ξ and ρ_Z were both set to 0.995. For both the synthetic data experiment and the real application dataset, we used a normal proposal for β: $q(\cdot \mid \beta) = \mathscr{N}(\beta, V_{\text{prop}})$. Our initial parameter values $\beta^{(0)}$ and proposal variances V_{prop} are provided in the supplementary code.

Algorithm B.1: Correlated sampling algorithm for the application

Select an initial parameter value $\beta^{(0)}$.
Draw an initial value $(U_\xi^{(0)}, U_Z^{(0)}, \Sigma^{(0)})$.
Compute $\xi^{(0)} = \varphi_\xi(U_\xi^{(0)}, \Sigma^{(0)}, Z_{\mathscr{S}}, W_{\mathscr{S}}, X_{\mathscr{S}})$.
Compute $Z^{\text{mis}}_{(0)} = \varphi_Z(U_Z^{(0)}, W_{\bar{\mathscr{S}}}, X_{\bar{\mathscr{S}}}, \xi^{(0)}, \Sigma^{(0)})$.
For $r = 1$ to $r = N$
(a) Propose $\tilde{\beta}$ from $q(\beta \mid \beta^{(r-1)})$.
(b) Propose $\tilde{\Sigma}$ from (2.13).
(c) Sample $\varepsilon_\xi \sim \mathscr{MN}(0_{13\times 9}, I_{13\times 13}, I_{9\times 9})$ and set $\tilde{U}_\xi = \rho_\xi U_\xi^{(r-1)} + \sqrt{(1-\rho_\xi^2)}\varepsilon_\xi$.
(d) Compute $\tilde{\xi} = \varphi_\xi(\tilde{U}_\xi, \tilde{\Sigma}, Z_{\mathscr{S}}, W_{\mathscr{S}}, X_{\mathscr{S}})$.
(e) Sample $\varepsilon_Z \sim \mathscr{N}(0_9, I_{9\times 9})$ and set $\tilde{U}_Z = \rho_Z U_Z^{(r-1)} + \sqrt{(1-\rho_Z^2)}\varepsilon_Z$.
(f) Compute $\tilde{Z}^{\text{mis}} = \varphi_Z(\tilde{U}_Z, W_{\bar{\mathscr{S}}}, X_{\bar{\mathscr{S}}}, \tilde{\xi}, \tilde{\Sigma})$.
(g) With probability $\min\left\{1, \frac{q(\beta^{(r-1)}\mid\tilde{\beta})p(\tilde{\beta})h(\tilde{\beta},\tilde{Z}^{\text{mis}})}{q(\tilde{\beta}\mid\beta^{(r-1)})p(\beta^{(r-1)})h(\beta^{(r-1)},Z^{\text{mis}}_{(r-1)})}\right\}$,
set $(\beta^{(r)}, U_\xi^{(r)}, U_Z^{(r)}) = (\tilde{\beta}, \tilde{U}_\xi, \tilde{U}_Z)$.
Otherwise, set $(\beta^{(r)}, U_\xi^{(r)}, U_Z^{(r)}) = (\beta^{(r-1)}, U_\xi^{(r-1)}, U_Z^{(r-1)})$.
Output $(\beta^{(1)}, \ldots, \beta^{(N)})$.

B.4 Convergence diagnostics

We provide convergence diagnostics for the sampling computation in §2.4.4. Figure B.1 contains the trace plots for the log-hazard ratios of the nine saturated fatty acids for 3 separate chains, each run for 1000000 iterations.

In §2.4.4, we discarded the first 200000 iterations of the sampler and used the subsequent 800000 iterations for analysis. Using the final 800000 iterations for each of the 3 chains, we

computed the Gelman-Rubin statistics (Gelman et al., 2013) for the log-hazard ratios of the 9 saturated fatty acids to be: 1.000019, 1.000053, 1.000005, 1.000056, 1.000052, 1.000021, 1.000082, 1.000057, 1.000033 for C15:0, C17:0, C14:0, C16:0, C18:0, C20:0, C22:0, C23:0 and C24:0 respectively.

B.5 Investigating the results for C18:0

In this section, we provide an informal calculation to demonstrate how increasing the relative concentration of C18:0 could indicate a positive association with type 2 diabetes, even when one does not exist on the transformed scale.

Suppose that our initial saturated fatty acid proportions are equal to the mean values in Table 2.4. This implies that the initial proportion of non-saturated fatty acids is 54.01%. If we increase the proportion of the fatty acid C18:0 by 1 standard deviation—1.32%—while keeping the other saturated fatty acid proportions fixed, the proportion of non-saturated fatty acids decreases to 52.69%. As a result, all logratios apart from the one corresponding to C18:0 increase by $\log(54.01) - \log(52.69) = 0.025$ (3 decimal places). Setting the posterior mean estimates of the hazard ratios in Table 2.4 as the truths, we can compute the change in risk as follows:

$$\begin{aligned}(0.97)^{\frac{0.025}{0.27}} \cdot (0.86)^{\frac{0.025}{0.26}} \cdot (1.18)^{\frac{0.025}{0.26}} \cdot (1.39)^{\frac{0.025}{0.07}} \cdot (0.91)^{\frac{0.025}{0.31}} \cdot (1.11)^{\frac{0.025}{0.24}} \cdot (0.99)^{\frac{0.025}{0.70}} \cdot (0.78)^{\frac{0.025}{0.26}} \\ = 1.00 \cdot 0.99 \cdot 1.02 \cdot 1.12 \cdot 0.99 \cdot 1.01 \cdot 1.00 \cdot 0.98 \\ = 1.10.\end{aligned}$$

We observe in particular that the effect is dominated by the factor of 1.12 from C16:0 due to its small standard deviation (0.07) on the transformed scale. For reference, Forouhi et al. (2014) estimated the hazard ratio of C18:0 across 6 different models to be (1.25, 1.06, 1.06, 1.12, 1.12, 1.07).

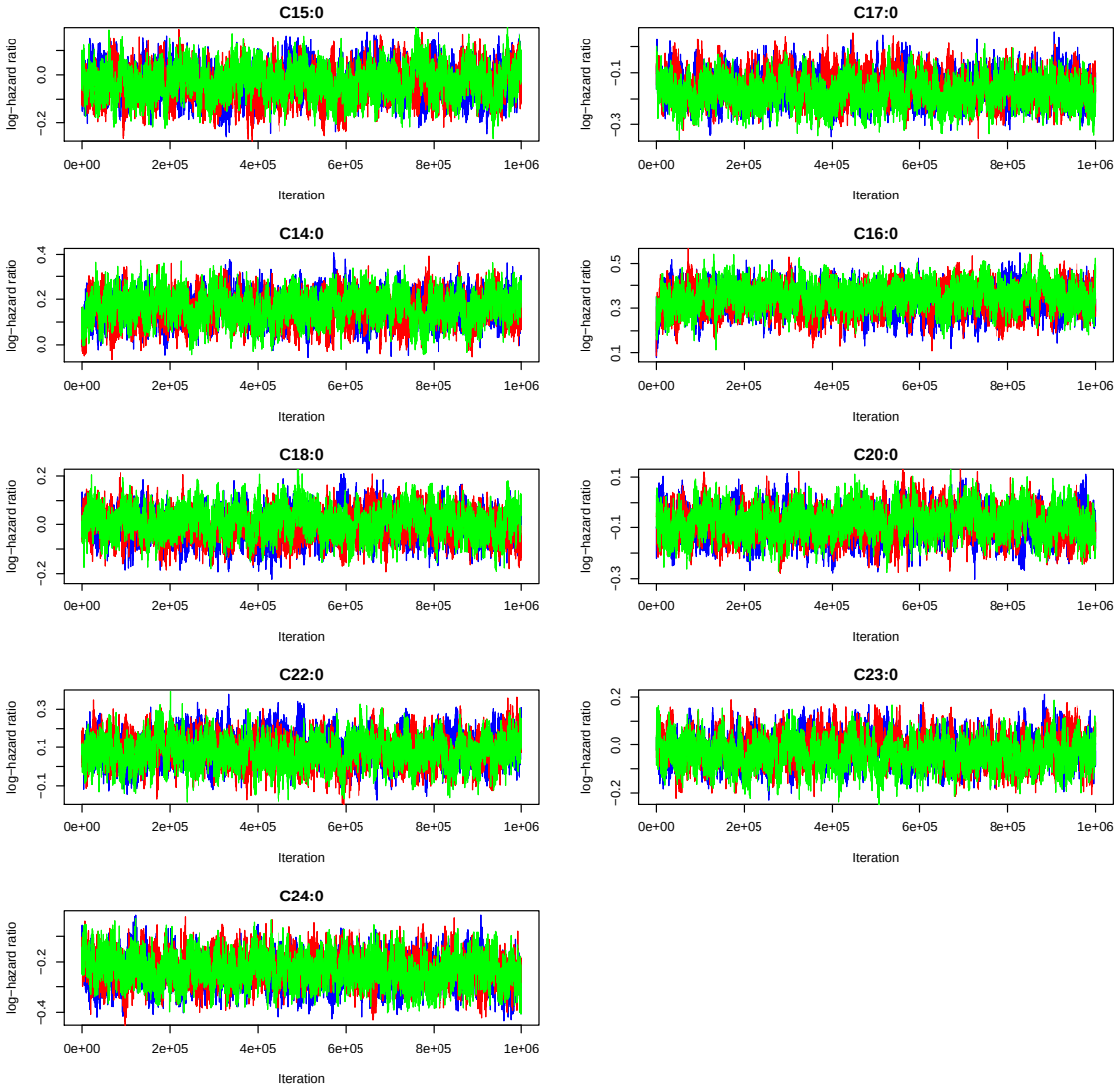


Fig. B.1 Trace plots for the log-hazard ratios of the nine saturated fatty acids.

Appendix C

Appendix for Chapter 3

C.1 Implementation pseudo-code

Define

$$f(\lambda) = \sum_{i=1}^{n} \exp\{\lambda^{\mathrm{T}} g_i\} g_i$$
$$H(\lambda) = \sum_{i=1}^{n} \exp\{\lambda^{\mathrm{T}} g_i\} g_i g_i^{\mathrm{T}}.$$

Further to our description in §2.5, we provide pseudo-code of the likelihood computation algorithm (Algorithm C.1) to assist users in implementing the method.

C.2 Notation

To reduce the amount of notational clutter in the proofs, we introduce the notation (i) $l_n(\theta) = \log L_n(\theta)$ and (ii) $g_i(\theta) = g(d_i, \theta)$.

C.3 Proofs

Proof of Proposition 3.1. The optimization problem

$$\max_{p_1,\ldots,p_n} \sum_{i=1}^{n} \{-p_i \log p_i\}$$

Algorithm C.1: Computing the exponentially tilted empirical likelihood

Input θ and tolerance $\tau_0 > 0$.
Solve linear programming problem described by (3.9).
If no feasible solutions exist
    output 0
else
    $\lambda \leftarrow (0, \ldots, 0)$
    $\tau \leftarrow \tau_0 + 1$
    while $\tau > \tau_0$
        $s \leftarrow H(\lambda)^{-1} f(\lambda)$
        $r \leftarrow 0$
        $\lambda' \leftarrow \lambda - s$
        while $f(\lambda) > f(\lambda')$
            $r \leftarrow r + 1$
            $\lambda' \leftarrow \lambda - 2^{-r} s$
        $\tau \leftarrow \|\lambda' - \lambda\|$
        $\lambda \leftarrow \lambda'$
    for $i = 1$ to $i = n$
        $p_i \leftarrow \exp(\lambda^{\mathrm{T}} g_i) / \sum_{j=1}^{n} \exp(\lambda^{\mathrm{T}} g_j)$
    $L \leftarrow \prod_{i=1}^{n} n p_i$
    Output L.

subject to

$$\sum_{i=1}^{n} p_i = 1$$

is solved uniquely by $p_i = 1/n$ for each $i = 1, \ldots, n$ (using the method of Lagrange multipliers for example). If the additional constraint

$$\sum_{i=1}^{n} p_i g(d_i, \hat{\theta}_n) = 0$$

is imposed, it follows that $p_i = 1/n$ for each $i = 1, \ldots, n$ is still the unique solution since it satisfies the constraint. By the AM-GM inequality,

$$L_n(\theta) = \prod_{i=1}^{n} n p_i(\theta) \leqslant 1$$

with equality if and only if each $p_i(\theta)$ is equal to $1/n$, attained at $\theta = \hat{\theta}_n$. □

Proof of Theorem 3.1. From the proof of Proposition 3.1, $L_n(\hat{\theta}_n) = 1$. Furthermore, by consistency of $\hat{\theta}_n$, θ_0 will lie in the ball $\{\theta : \|\theta - \hat{\theta}_n\| \leqslant \delta/2\}$ with probability approaching one. Hence,

$$\sup_{\|\theta-\hat{\theta}_n\|_2 \geqslant \delta} \frac{L_n(\theta)}{L_n(\hat{\theta}_n)} \leqslant \sup_{\|\theta-\theta_0\| \geqslant \delta/2} \sup_{p \in \Phi(\theta)} \prod_{i=1}^{n} np_i \tag{C.1}$$

occurs with probability approaching 1, where $\Phi(\theta) = \{p : \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i g_i(\theta) = 0, p_i \geqslant 0, i = 1, \ldots, n\} \cup \{0\}$, and it is therefore sufficient to establish the upper-bound for the right-hand side.

By a similar argument to the proof of Lemma C.1, $\mathbb{E}_{P_0}\{g(D,\theta)\}$ is continuous in θ and we have assumed that it has a unique zero at θ_0. By the compactness of Θ, there exists some $\varepsilon > 0$ such that

$$\inf_{\|\theta-\theta_0\| \geqslant \delta/2} \|\mathbb{E}_{P_0}\{g(D,\theta)\}\|_1 > \varepsilon.$$

By Assumption 3.1(iv), $n^{-1}\sum_{i=1}^n g_i(\theta)$ and $n^{-1}\sum_{i=1}^n \|g_i(\theta)\|_2^2$ converge uniformly in probability to $\mathbb{E}_{P_0}\{g(D,\theta)\}$ and $\mathbb{E}_{P_0}\{\|g(D,\theta)\|_2^2\}$ respectively. Therefore,

$$\sup_{\theta\in\Theta} \left\| \frac{1}{n}\sum_{i=1}^n g_i(\theta) - \mathbb{E}_{P_0}\{g(D,\theta)\} \right\|_1 < \varepsilon/2, \quad \sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \|g_i(\theta)\|_2^2 < 2\mathbb{E}_{P_0}\left\{ \sup_{\theta\in\Theta} \|g(D,\theta)\|_2^2 \right\}$$

occur with probability approaching 1. On this event,

$$\inf_{\|\theta-\theta_0\| \geqslant \delta/2} \left\| \frac{1}{n}\sum_{i=1}^n g_i(\theta) \right\|_1^2 = \inf_{\|\theta-\theta_0\| \geqslant \delta/2} \inf_{p \in \Phi(\theta)} \left\| \frac{1}{n}\sum_{i=1}^n g_i(\theta) - \sum_{i=1}^n p_i g_i(\theta) \right\|_1^2 > \frac{\varepsilon^2}{4}.$$

By the Cauchy-Schwarz inequality, the left hand side is bounded above by

$$\inf_{\|\theta-\theta_0\| \geqslant \delta/2} \inf_{p \in \Phi(\theta)} \left\{ \frac{1}{n}\sum_{i=1}^n (np_i - 1)^2 \right\} \left\{ \frac{1}{n}\sum_{i=1}^n \|g_i(\theta)\|_2^2 \right\}.$$

Hence, there exists a strictly positive constant $\tilde{\varepsilon}$ such that

$$\inf_{\|\theta-\theta_0\| \geqslant \delta/2} \inf_{p \in \Phi(\theta)} \left\{ \frac{1}{n}\sum_{i=1}^n (np_i - 1)^2 \right\} \geqslant \tilde{\varepsilon}.$$

Consider the optimization problem of maximizing $\prod_{i=1}^n np_i$ subject to

$$\sum_{i=1}^n p_i = 1, \quad \sum_{i=1}^n (np_i - 1)^2 \geqslant n\tilde{\varepsilon}, \quad p_i \geqslant 0 \text{ for each } i = 1, \ldots, n.$$

For an element $p = (p_1, \ldots, p_n)$ in the constraint set, if p_i, p_j both exceed n^{-1} for some i, j and are unequal, replacing both with $(p_i + p_j)/2$ would strictly increase the objective while remaining in the constraint set. We deduce that for any solution to the the optimization problem, all values of p_i exceeding n^{-1} must be equal. At least one value exceeds n^{-1} due to the inequality constraint. A similar argument applies for values below n^{-1}.

For fixed $m \in \{1, \ldots, n-1\}$, we consider maximizing the objective when m values of p_i are equal to $p_+ > n^{-1}$, and the remaining $n-m$ values are equal to $p_- < n^{-1}$. We can further write $np_+ = 1 + a$, $np_- = 1 - b$, where $0 \leqslant a \leqslant n-1$, $0 \leqslant b \leqslant 1$. By taking the logarithm of the objective, we seek to maximize $m\log(1+a) + (n-m)\log(1-b)$ subject to

$$am = (n-m)b, \quad ma^2 + (n-m)b^2 \geqslant n\tilde{\varepsilon}.$$

For $(n\tilde{\varepsilon})/\{(n-1)^2 + \tilde{\varepsilon}\} < m < n/(1+\tilde{\varepsilon})$, the constraint set is non-empty and the solution is

$$a = \left\{ \frac{\tilde{\varepsilon}(n-m)}{m} \right\}^{1/2}, \quad b = \left[\frac{\tilde{\varepsilon} m}{n-m} \right]^{1/2}.$$

We consider sufficiently large n such that $m = 1$ lies in the permissible range. We claim that for fixed n, $m = 1$ is the value which maximizes the objective, which can now be written as

$$\left[1 + \left\{ \frac{\tilde{\varepsilon}(n-m)}{m} \right\}^{1/2} \right]^m \left\{ 1 - \left(\frac{\tilde{\varepsilon} m}{n-m} \right)^{1/2} \right\}^{n-m}.$$

Letting $x = \{(n-m)/m\}^{1/2}$, which is strictly decreasing in m, and taking the logarithm of the objective, it is sufficient to show that the function

$$\frac{n}{1+x^2}\log(1 + x\tilde{\varepsilon}^{1/2}) + \frac{nx^2}{1+x^2}\log\left(1 - \frac{\tilde{\varepsilon}^{1/2}}{x}\right)$$

is increasing in x. By differentiating with respect to x and simplifying, it is sufficient to show that

$$2x\left\{ \log(1 + x\tilde{\varepsilon}^{1/2}) - \log\left(1 - \frac{\tilde{\varepsilon}^{1/2}}{x}\right) \right\} - (1+x^2)\left(\frac{\tilde{\varepsilon}^{1/2}}{1 + x\tilde{\varepsilon}^{1/2}} + \frac{\tilde{\varepsilon}^{1/2}}{1 - \tilde{\varepsilon}^{1/2}/x} \right) < 0. \quad \text{(C.2)}$$

The first term is equal to

$$\begin{aligned} 2x\log\left(\frac{1+x\tilde{\varepsilon}^{1/2}}{1-\tilde{\varepsilon}^{1/2}/x}\right) &= 2x\log\left\{1+\frac{\tilde{\varepsilon}^{1/2}(x^2+1)}{x(1-\tilde{\varepsilon}^{1/2}/x)}\right\} \\ &\leqslant \frac{2\tilde{\varepsilon}^{1/2}(x^2+1)}{1-\tilde{\varepsilon}^{1/2}/x}\left(\frac{1+x\tilde{\varepsilon}^{1/2}}{1-\tilde{\varepsilon}^{1/2}/x}\right)^{-1/2} \\ &= \frac{2\tilde{\varepsilon}^{1/2}(x^2+1)}{\{(1-\tilde{\varepsilon}^{1/2}/x)(1+x\tilde{\varepsilon}^{1/2})\}^{1/2}} \end{aligned}$$

where we have used the inequality $\log(1+z) \leqslant z(z+1)^{-1/2}$. Therefore, the left-hand side of (C.2) is upper-bounded by

$$\frac{\tilde{\varepsilon}^{1/2}(x^2+1)}{\{(1-\tilde{\varepsilon}^{1/2}/x)(1+x\tilde{\varepsilon}^{1/2})\}^{1/2}}\left\{2-\left(\frac{1-\tilde{\varepsilon}^{1/2}/x}{1+x\tilde{\varepsilon}^{1/2}}\right)^{1/2}-\left(\frac{1+x\tilde{\varepsilon}^{1/2}}{1-\tilde{\varepsilon}^{1/2}/x}\right)^{1/2}\right\}. \tag{C.3}$$

For positive z, $z+z^{-1}$ is lower-bounded by 2, with equality if and only if $z=1$. But $\tilde{\varepsilon}$ is strictly greater than 0, so

$$\left(\frac{1-\tilde{\varepsilon}^{1/2}/x}{1+x\tilde{\varepsilon}^{1/2}}\right)^{1/2}$$

cannot equal 1. Therefore, (C.3) is strictly less than 0, as required.

Returning to (C.1), we conclude that

$$\begin{aligned} \sup_{\|\theta-\theta_0\|\geqslant\delta/2}\sup_{p\in\Phi(\theta)}\prod_{i=1}^{n}np_i &\leqslant [1+\{\tilde{\varepsilon}(n-1)\}^{1/2}]\left\{1-\left(\frac{\tilde{\varepsilon}}{n-1}\right)^{1/2}\right\}^{n-1} \\ &= [1+\{\tilde{\varepsilon}(n-1)\}^{1/2}]\exp\left[(n-1)\log\left\{1-\left(\frac{\tilde{\varepsilon}}{n-1}\right)^{1/2}\right\}\right] \\ &\leqslant [1+\{\tilde{\varepsilon}(n-1)\}^{1/2}]\exp\{-\tilde{\varepsilon}(n-1)^{1/2}\}. \end{aligned}$$

For $0<\varepsilon^*<\tilde{\varepsilon}$, and sufficiently large n, we have a further upper-bound of $\exp\{-\varepsilon^*(n-1)^{1/2}\}$. □

Proof of Proposition 3.2. We work in a neighbourhood of $(0,\theta_0)$ in $\mathbb{R}^m\times\Theta$ in which Assumptions 3.3 and 3.4 hold. The function

$$\mathbb{E}_{P_0}[\exp\{\lambda^{\mathrm{T}}g(D,\theta)\}g(D,\theta)]$$

is 0 at $(0, \theta_0)$ and the domination condition of Assumption 3.4 allows us to differentiate under the integral sign twice and deduce that the function is twice continuously differentiable. By the implicit function theorem, there exist a neighbourhood $\mathscr{U} \subset \Theta$ of θ_0 and a neighbourhood of $\mathscr{W} \subset \mathbb{R}^m$ of 0 such that there exists a unique twice continuously differentiable function $\lambda_0 : \mathscr{U} \to \mathscr{W}$ satisfying

$$\lambda_0(\theta_0) = 0, \quad \mathbb{E}_{P_0}[\exp\{\lambda_0(\theta)^{\mathrm{T}} g(D, \theta)\} g(D, \theta)] = 0$$

for all $\theta \in \mathscr{U}$. The second part of Theorem 3.1 in Csiszár (1975) implies that λ_0 is in fact the unique mapping into $\mathbb{R}^m$ which satisfies the above properties. The implicit function theorem also implies that the second derivative $\partial^2 \lambda_0$ of λ_0 can be expressed as the sum and products of expectations of expressions involving λ_0, $\partial \lambda_0$, g, $\partial_\theta g$, which are all continuously differentiable in θ, and $\partial_\theta^2 g$, which satisfies the Lipschitz condition from Assumption 3.3, defined on a bounded set. Therefore, $\partial^2 \lambda_0$ is Lipschitz continuous. □

Lemma C.1. *The function*

$$\mathbb{E}_{P_0}\{g(D, \theta) g(D, \theta)^{\mathrm{T}}\}$$

is continuous in θ.

Proof of Lemma C.1. For a fixed value $\theta^* \in \Theta$, consider a sequence $\theta_n \to \theta^*$. Define

$$f_n(d) = g(d, \theta_n) g(d, \theta_n)^{\mathrm{T}}, \quad f(d) = g(d, \theta^*) g(d, \theta^*)^{\mathrm{T}}$$

such that f_n converges pointwise to f P_0-almost everywhere and

$$\|f_n(d)\|_F \leqslant \sup_{\theta \in \Theta} \left\| g(d, \theta) g(d, \theta)^{\mathrm{T}} \right\|_F$$

for all n and for all values of d, where F refers to the Frobenius norm. The upper-bound is integrable, since

$$\mathbb{E}_{P_0} \left\{ \sup_{\theta \in \Theta} \left\| g(d, \theta) g(d, \theta)^{\mathrm{T}} \right\|_F \right\} = \mathbb{E}_{P_0} \left\{ \sup_{\theta \in \Theta} \|g(d, \theta)\|^2 \right\} < \infty$$

by Assumption 3.1(iv). Therefore, we can apply the dominated convergence theorem to deduce that

$$\lim_{n \to \infty} \mathbb{E}_{P_0} \left\{ g(d, \theta_n) g(d, \theta_n)^{\mathrm{T}} \right\} = \mathbb{E}_{P_0} \left\{ g(d, \theta^*) g(d, \theta^*)^{\mathrm{T}} \right\},$$

which establishes continuity. □

Lemma C.2. *Under Assumptions 3.1–3.4, there exists a value of $\delta > 0$ such that the δ-ball around $\hat{\theta}_n$ satisfies the following properties with probability approaching 1:*

(i) contained in a neighbourhood of θ_0 satisfying the conditions of Assumptions 3.2 and 3.3 and Proposition 3.2.

(ii) the set of vectors $\{g_1(\theta), \ldots, g_n(\theta)\}$ span $\mathbb{R}^m$ for all values of θ.

(iii) the function $\hat{\lambda}_n(\theta)$ from Assumption 3.2 is the unique function mapping into $\mathbb{R}^m$ which satisfies

$$\sum_{i=1}^{n} \exp\{\hat{\lambda}_n(\theta)^{\mathrm{T}} g_i(\theta)\} g_i(\theta) = 0.$$

(iv) $\hat{\lambda}_n$ is twice continuously differentiable and

$$\partial \hat{\lambda}_n(\theta) = -\left\{ \sum_{i=1}^{n} p_i(\theta) g_i(\theta) g_i(\theta)^{\mathrm{T}} \right\}^{-1} \left[\sum_{j=1}^{n} p_j(\theta) \{I + g_j(\theta) \hat{\lambda}_n(\theta)^{\mathrm{T}}\} \partial_\theta g_j \right]. \quad \text{(C.4)}$$

(v) l_n is twice differentiable with $\partial l_n(\hat{\theta}_n) = 0$ and $n^{-1} \partial^2 l_n(\hat{\theta}_n) = -\hat{\Sigma}_n^+ = -\hat{G}_n^{\mathrm{T}} \hat{\Omega}_n^{-1} \hat{G}_n$.

Proof of Lemma C.2. Consider a ball around θ_0 satisfying the conditions of Assumptions 3.2 and 3.3 and Proposition 3.2. By the consistency of $\hat{\theta}_n$, with probability approaching one, $\hat{\theta}_n$ is within half the radius from θ_0. Thus, we can take the ball around $\hat{\theta}_n$ of half the radius.

Assumption 3.1(iii) and Lemma C.1 imply that there exists a neighbourhood of θ_0 where the determinant of $\mathbb{E}_{P_0}\{g(D,\theta) g(D,\theta)^{\mathrm{T}}\}$ is bounded away from 0. By the uniform law of large numbers implied by Assumption 3.1(iv), $n^{-1} \sum_{i=1}^{n} g_i(\theta) g_i(\theta)^{\mathrm{T}}$ is positive definite for all θ in this neighbourhood with probability approaching 1. This is equivalent to the set $\{g_1(\theta), \ldots, g_n(\theta)\}$ spanning $\mathbb{R}^m$. If necessary, we shrink the ball around $\hat{\theta}_n$ to be contained in here.

The function $f_n(\lambda, \theta) = \sum_{i=1}^{n} \exp\{\lambda^{\mathrm{T}} g_i(\theta)\} g_i(\theta)$ is differentiable with respect to λ with partial derivative

$$\partial_\lambda f_n(\lambda, \theta) = \sum_{i=1}^{n} \exp\{\lambda^{\mathrm{T}} g_i(\theta)\} g_i(\theta) g_i(\theta)^{\mathrm{T}}$$

which is positive definite by the previous property. Thus, for fixed θ, $f_n(\lambda, \theta)$ is an injective mapping of λ and $\hat{\lambda}_n(\theta)$ the unique value which maps to 0.

By the uniqueness of $\hat{\lambda}_n$ and the application of the implicit function theorem to f_n at each value of $(\hat{\lambda}_n(\theta), \theta)$, $\hat{\lambda}_n$ is equal to the implicit function and is thus twice continuously

differentiable. The first derivative is

$$\begin{aligned}\partial\hat{\lambda}_n(\theta) &= -\left\{\sum_{i=1}^n \exp\hat{\lambda}_n(\theta)^{\mathrm{T}} g_i(\theta) g_i(\theta) g_i(\theta)^{\mathrm{T}}\right\}^{-1}\left[\sum_{j=1}^n \exp\hat{\lambda}_n(\theta)^{\mathrm{T}} g_j(\theta)\{I_m + g_j(\theta)\hat{\lambda}_n(\theta)^{\mathrm{T}}\}\partial_\theta g_j\right] \\ &= -\left\{\sum_{i=1}^n p_i(\theta) g_i(\theta) g_i(\theta)^{\mathrm{T}}\right\}^{-1}\left[\sum_{j=1}^n p_j(\theta)\{I_m + g_j(\theta)\hat{\lambda}_n(\theta)^{\mathrm{T}}\}\partial_\theta g_j\right].\end{aligned}$$

We can express the log exponentially tilted empirical likelihood as

$$\begin{aligned}l_n(\theta) &= \log\prod_{i=1}^n \frac{\exp\{\hat{\lambda}_n(\theta)^{\mathrm{T}} g_i(\theta)\}}{\sum_{j=1}^n \exp\{\hat{\lambda}_n(\theta)^{\mathrm{T}} g_j(\theta)\}} \\ &= \sum_{i=1}^n \{\hat{\lambda}_n(\theta)^{\mathrm{T}} g_i(\theta)\} - n\log\sum_{j=1}^n \exp\{\hat{\lambda}_n(\theta)^{\mathrm{T}} g_j(\theta)\}\end{aligned}$$

and we differentiate with respect to θ to obtain

$$\begin{aligned}\partial l_n(\theta) &= \sum_{i=1}^n \partial(\hat{\lambda}_n^{\mathrm{T}} g_i) - n\sum_{i=1}^n \frac{\partial(\hat{\lambda}_n^{\mathrm{T}} g_i)\exp\{\hat{\lambda}_n(\theta)^{\mathrm{T}} g_i(\theta)\}}{\sum_{j=1}^n \exp\{\hat{\lambda}_n(\theta)^{\mathrm{T}} g_j(\theta)\}} \\ &= \sum_{i=1}^n \partial(\hat{\lambda}_n^{\mathrm{T}} g_i)\{1 - np_i(\theta)\}.\end{aligned}$$

But $p_i(\hat{\theta}_n) = 1/n$ for each $i = 1,\ldots,n$, so

$$\partial l_n(\hat{\theta}_n) = 0.$$

The second derivative of l_n is

$$\partial^2 l_n(\theta) = \sum_{i=1}^n \partial^2(\hat{\lambda}_n^{\mathrm{T}} g_i)(\theta)\{1 - np_i(\theta)\} - n\sum_{i=1}^n \{\partial(\hat{\lambda}_n^{\mathrm{T}} g_i)^{\mathrm{T}}\partial(p_i)\}(\theta).$$

Since $p_i(\hat{\theta}_n) = 1/n$ for each $i = 1,\ldots,n$, the first sum is zero at $\theta = \hat{\theta}_n$. Furthermore,

$$\begin{aligned}\partial(\hat{\lambda}_n^{\mathrm{T}} g_i)(\hat{\theta}_n) &= \left(g_i^{\mathrm{T}}\partial\hat{\lambda}_n + \hat{\lambda}_n^{\mathrm{T}}\partial_\theta g_i\right)(\hat{\theta}_n) \\ &= (g_i^{\mathrm{T}}\partial\hat{\lambda}_n)(\hat{\theta}_n)\end{aligned}$$

since $\hat{\lambda}_n(\hat{\theta}_n) = 0$ by part (ii) and

$$\begin{aligned}\partial p_i(\hat{\theta}_n) &= \left\{ p_i \partial(\hat{\lambda}_n^{\mathrm{T}} g_i) - p_i \sum_{j=1}^{n} p_j \partial(\hat{\lambda}_n^{\mathrm{T}} g_j) \right\}(\hat{\theta}_n) \\ &= n^{-1}(g_i^{\mathrm{T}} \partial\hat{\lambda}_n)(\hat{\theta}_n) - n^{-2} \left\{ \sum_{j=1}^{n} g_j(\hat{\theta}_n) \right\}^{\mathrm{T}} \partial\hat{\lambda}_n(\hat{\theta}_n)\end{aligned}$$

where the second term is zero since $\hat{\theta}_n$ is the Z-estimator. We deduce from part (iii) that

$$\begin{aligned}\partial\hat{\lambda}_n(\hat{\theta}_n) &= -\left\{ n^{-1} \sum_{i=1}^{n} g_i(\hat{\theta}_n) g_i(\hat{\theta}_n)^{\mathrm{T}} \right\}^{-1} \left\{ n^{-1} \sum_{j=1}^{n} \partial_\theta g_j(\hat{\theta}_n) \right\} \\ &= -\hat{\Omega}_n^{-1} \hat{G}_n.\end{aligned}$$

Putting everything together,

$$\begin{aligned}n^{-1}\partial^2 l_n(\hat{\theta}_n) &= -\hat{G}_n^{\mathrm{T}} \hat{\Omega}_n^{-1} \left\{ n^{-1} \sum_{i=1}^{n} g_i(\hat{\theta}_n) g_i(\hat{\theta}_n)^{\mathrm{T}} \right\} \hat{\Omega}_n^{-1} \hat{G}_n^{\mathrm{T}} \\ &= -\hat{G}_n^{\mathrm{T}} \hat{\Omega}_n^{-1} \hat{G}_n,\end{aligned}$$

as required. □

Proof of Theorem 3.2. This proof is based on the proof of Theorem 1.4.2 in Ghosh and Ramamoorthi (2003).

We make a change of variables $s = n^{1/2}(\theta - \hat{\theta}_n)$

$$\int_{\mathbb{R}^m} \left| p^*(s \mid D_1, \ldots, D_n) - (2\pi)^{-m/2} |\Sigma_0|^{-1/2} \exp(-0.5 s^{\mathrm{T}} \Sigma_0^{-1} s) \right| ds$$

where

$$\begin{aligned}p^*(s \mid D_1, \ldots, D_n) &= \frac{p(\hat{\theta}_n + s/n^{1/2}) L_n(\hat{\theta}_n + s/n^{1/2})}{\int p(\hat{\theta}_n + t/n^{1/2}) L_n(\hat{\theta}_n + t/n^{1/2})\, dt} \\ &= \frac{p(\hat{\theta}_n + s/n^{1/2}) \exp\{l_n(\hat{\theta}_n + s/n^{1/2}) - l_n(\hat{\theta}_n)\}}{\int p(\hat{\theta}_n + t/n^{1/2}) \exp\{l_n(\hat{\theta}_n + t/n^{1/2}) - l_n(\hat{\theta}_n)\}\, dt}\end{aligned}$$

and is extended to all of $\mathbb{R}^m$ by taking the value zero outside of its original domain. Writing $C_n = \int_{\mathbb{R}^m} p(\hat{\theta}_n + t/n^{1/2}) \exp\{l_n(\hat{\theta}_n + t/n^{1/2}) - l_n(\hat{\theta}_n)\}\, dt$, we are required to show that

$$C_n^{-1} \int_{\mathbb{R}^m} \left| p(\hat{\theta}_n + s/n^{1/2}) \exp\{l_n(\hat{\theta}_n + s/n^{1/2}) - l_n(\hat{\theta}_n)\} - C_n (2\pi)^{-m/2} |\Sigma_0|^{-1/2} \exp\left(-s^{\mathrm{T}} \Sigma_0^{-1} s/2\right) \right| ds \tag{C.5}$$

tends in probability to 0. It is sufficient to show that

$$\mathscr{I}_1 = \int_{\mathbb{R}^m} \left| p(\hat{\theta}_n + s/n^{1/2}) \exp\{l_n(\hat{\theta}_n + s/n^{1/2}) - l_n(\hat{\theta}_n)\} - p(\theta_0) \exp\left(-s^{\mathrm{T}} \Sigma_0^{-1} s/2\right) \right| ds \to 0$$

with convergence in probability, since it implies that C_n converges to $p(\theta_0)(2\pi)^{m/2}|\Sigma_0|^{1/2}$ in probability and the integral in (C.5) is bounded above by $\mathscr{I}_1 + \mathscr{I}_2$, where

$$\begin{aligned} \mathscr{I}_2 &= \int_{\mathbb{R}^m} \left| p(\theta_0) \exp\left(-s^{\mathrm{T}} \Sigma_0^{-1} s/2\right) - C_n (2\pi)^{-m/2} |\Sigma_0|^{-1/2} \exp\left(-s^{\mathrm{T}} \Sigma_0^{-1} s/2\right) \right| ds \\ &= \left| p(\theta_0) - C_n (2\pi)^{-m/2} |\Sigma_0|^{-1/2} \right| \int_{\mathbb{R}^m} \exp\left(-s^{\mathrm{T}} \Sigma_0^{-1} s/2\right) ds \end{aligned}$$

which also converges to 0 in probability.

Let $\delta > 0$ be small enough to satisfy the conditions of Lemma C.2. Let $c > 0$. We separate $\mathscr{I}_1$ into the three regions $A_1 = \{s : \|s\|_2 < c \log n^{1/2}\}$, $A_2 = \{s : c \log n^{1/2} < \|s\|_2 < \delta n^{1/2}\}$, $A_3 = \{s : \|s\|_2 > \delta n^{1/2}\}$.

We begin with A_3.

$$\begin{aligned} &\int_{A_3} \left| p(\hat{\theta}_n + s/n^{1/2}) \exp\{l_n(\hat{\theta}_n + s/n^{1/2}) - l_n(\hat{\theta}_n)\} - p(\theta_0) \exp\left(-s^{\mathrm{T}} \Sigma_0^{-1} s/2\right) \right| ds \\ &\qquad \leqslant \int_{A_3} p(\hat{\theta}_n + s/n^{1/2}) \frac{L_n(\hat{\theta}_n + s/n^{1/2})}{L_n(\hat{\theta}_n)}\, ds + \int_{A_3} p(\theta_0) \exp\left(-s^{\mathrm{T}} \Sigma_0^{-1} s/2\right) ds. \end{aligned}$$

The first integral goes to zero by Theorem 3.1. The second goes to zero by the tail properties of the multivariate normal distribution.

By Taylor's theorem,

$$\begin{aligned} l_n(\hat{\theta}_n + s/n^{1/2}) - l_n(\hat{\theta}_n) &= \frac{1}{2n} \hat{c}^2 l_n(\hat{\theta}_n)(s,s) + \frac{1}{2n} \{\hat{c}^2 l_n(\theta_s)(s,s) - \hat{c}^2 l_n(\hat{\theta}_n)(s,s)\} \\ &= -\frac{1}{2} s^{\mathrm{T}} \hat{\Sigma}_n^{+} s + R_n(s) \end{aligned}$$

where $\theta_s = \hat{\theta}_n + (\eta s)/n^{1/2}$ for some $\eta \in [0,1]$, with the first order term vanishing due to Lemma C.2. By the domination conditions of Assumption 3.4 and the uniqueness of λ_0 from

Proposition 3.2, all of

$$\sup_{\theta\in\mathscr{B}_\delta(\hat\theta_n)}\left\|\hat\lambda_n(\theta)-\lambda_0(\theta)\right\|,\quad \sup_{\theta\in\mathscr{B}_\delta(\hat\theta_n)}\left\|\partial\hat\lambda_n(\theta)-\partial\lambda_0(\theta)\right\|,\quad \sup_{\theta\in\mathscr{B}_\delta(\hat\theta_n)}\left\|\partial^2\hat\lambda_n(\theta)-\partial^2\lambda_0(\theta)\right\|$$

converge to 0 in probability. For the following, let $h_i(\theta)=\lambda_0(\theta)^{\mathrm{T}}g_i(\theta)$ and $h(D,\theta)=\lambda_0(\theta)^{\mathrm{T}}g(D,\theta)$ and we suppress dependence on θ for presentational clarity

$$\frac{1}{n}\partial^2 l_n=\frac{1}{n}\sum_{i=1}^n\left(\partial^2 h_i\left[1-\frac{\exp(h_i)}{\mathbb{E}_{P_0}\{\exp h(D)\}}\right]-\frac{\exp(h_i)\partial h_i^{\mathrm{T}}}{\mathbb{E}_{P_0}\{\exp h(D)\}}\left[\partial h_i-\frac{\mathbb{E}_{P_0}\{\exp h(D)\partial h(D)\}}{\mathbb{E}_{P_0}\{\exp h(D)\}}\right]\right)+o_{P_0}(1).$$

From Assumption 3.3 and Proposition 3.2, we know that for each i, $\partial^2 h_i$ satisfies a Lipschitz condition, and all other terms are continuously differentiable in θ, thus

$$\sup_{s\in A_1\cup A_2}\frac{n^{-1}\left\|\partial^2 l_n(\theta_s)-\partial^2 l_n(\hat\theta_n)\right\|_{op}}{\|\theta_s-\hat\theta_n\|_2}\leqslant O_{P_0}(1).$$

Now consider

$$\int_{A_1}\left|p(\hat\theta_n+s/n^{1/2})\exp\{l_n(\hat\theta_n+s/n^{1/2})-l_n(\hat\theta_n)\}-p(\theta_0)\exp\left(-\frac{1}{2}s^{\mathrm{T}}\Sigma_0^{-1}s\right)\right|ds\leqslant J_1+J_2$$

where

$$\begin{aligned}J_1&=\int_{A_1}p(\hat\theta_n+s/n^{1/2})\left|\exp\left(-\frac{1}{2}s^{\mathrm{T}}\hat\Sigma_n^{+}s+R_n(s)\right)-\exp\left(-\frac{1}{2}s^{\mathrm{T}}\Sigma_0^{-1}s\right)\right|ds\\ J_2&=\int_{A_1}\left|p(\hat\theta_n+s/n^{1/2})-p(\theta_0)\right|\exp\left(-\frac{1}{2}s^{\mathrm{T}}\Sigma_0^{-1}s\right)ds.\end{aligned}$$

By consistency of $\hat\theta_n$ and continuity of $p(\theta)$ at θ_0, J_2 converges to 0 in probability. Furthermore,

$$\sup_{s\in A_1}R_n(s)\leqslant\sup_{s\in A_1}\|s\|_2^2\|\theta_s-\hat\theta_n\|_2\,O_{P_0}(1)\leqslant c^3\frac{(\log n^{1/2})^3}{n^{1/2}}O_{P_0}(1)=o_{P_0}(1)$$

and $\hat\Sigma_n^{+}$ converges to Σ_0^{-1} in probability by Assumption 3.1. Therefore, J_1 converges in probability to zero.

Next consider

$$\int_{A_2} \left| p(\hat{\theta}_n + s/n^{1/2}) \exp\{l_n(\hat{\theta}_n + s/n^{1/2}) - l_n(\hat{\theta}_n)\} - p(\theta_0) \exp\left(-\frac{1}{2} s^{\mathrm{T}} \Sigma_0^{-1} s\right) \right| ds$$
$$\leqslant \int_{A_2} p(\hat{\theta}_n + s/n^{1/2}) \exp\left\{-\frac{1}{2} s^{\mathrm{T}} \hat{\Sigma}_n^{+} s + R_n(s)\right\} ds + \int_{A_2} p(\theta_0) \exp\left(-\frac{1}{2} s^{\mathrm{T}} \Sigma_0^{-1} s\right) ds.$$

The second integral is bounded above by $p(\theta_0) \exp\{-\zeta (c \log n^{1/2})^2/2\} \mathrm{vol}(A_2)$ where $\zeta > 0$ is the smallest eigenvalue of Σ_0^{-1}. For $n^{1/2} > e$, $(\log n^{1/2})^2 > \log n^{1/2}$, so we can further upper-bound the second integral by

$$K p(\theta_0) \frac{n^{m/2}}{n^{\zeta c^2/4}}$$

where $K > 0$ is a constant. For sufficiently large c, this tends to 0 as n tends to infinity.

For the first integral, since $\|\theta_s - \hat{\theta}_n\|_2 < \delta$ for all $s \in A_2$, we have

$$\sup_{s \in A_2} \frac{|R_n(s)|}{\|s\|_2^2} \leqslant \delta \, O_{P_0}(1).$$

Therefore, for any $\varepsilon > 0$, we can choose sufficiently small δ to ensure that

$$\mathrm{pr}\left\{ |R_n(s)| < \frac{1}{4} s^{\mathrm{T}} \hat{\Sigma}_n^{+} s \text{ for all } s \in A_2 \right\} > 1 - \varepsilon$$

for all sufficiently large n. Hence, with probability greater than $1 - \varepsilon$,

$$\int_{A_2} p(\hat{\theta}_n + s/n^{1/2}) \exp\left\{-\frac{1}{2} s^{\mathrm{T}} \hat{\Sigma}_n^{+} s + R_n(s)\right\} ds$$
$$\leqslant \sup_{s \in A_2} p(\hat{\theta}_n + s/n^{1/2}) \int_{A_2} \exp\left(-\frac{1}{4} s^{\mathrm{T}} \hat{\Sigma}_n^{+} s\right) ds$$

which converges to zero in probability. □

Proof of Theorem 3.3. Using the same notation as the proof of Theorem 3.2, we claim that

$$\int_{\mathbb{R}^m} \|s \{p^*(s \mid D_1, \ldots, D_n) - (2\pi)^{-m/2} |\Sigma_0|^{-1/2} \exp(-0.5 s^{\mathrm{T}} \Sigma_0^{-1} s)\}\|_2 \, ds \to 0$$

with convergence in probability. This is similar to what was proved in Theorem 3.2, but there is now an additional factor of $\|s\|_2$ in the integrand. The claim implies that

$$\left\| \int_{\mathbb{R}^m} s \{p^*(s \mid D_1, \ldots, D_n) - (2\pi)^{-m/2} |\Sigma_0|^{-1/2} \exp(-0.5 s^{\mathrm{T}} \Sigma_0^{-1} s)\} \, ds \right\|_2 \to 0$$

with convergence in probability, but the second term within the norm is equal to the mean of a mean zero multivariate normal distribution. Thus,

$$n^{1/2}(\theta_n^* - \hat{\theta}_n) = \int_{\mathbb{R}^m} s \, p^*(s \mid d_1, \ldots, d_n) \, ds \to 0$$

with convergence in probability. The second assertion follows from this along with the asymptotic normality of $\hat{\theta}_n$ stated in §2.1.

It remains to prove the initial claim. Since $\int_{\mathbb{R}^m} \|s\|_2 \exp(-0.5 s^{\mathrm{T}} \Sigma_0^{-1} s) \, ds < \infty$, we can argue similarly to the proof of Theorem 3.2 that it is sufficient to show

$$\int_{\mathbb{R}^m} \|s\{p(\hat{\theta}_n + s/n^{1/2}) \exp\{l_n(\hat{\theta}_n + s/n^{1/2}) - l_n(\hat{\theta}_n)\} - p(\theta_0) \exp(-s^{\mathrm{T}} \Sigma_0^{-1} s/2)\}\|_2 \, ds \to 0$$

with convergence in probability. As before, we decompose the integral into the three regions A_1, A_2 and A_3. For A_3,

$$\begin{aligned} &\int_{A_3} \|s\{p(\hat{\theta}_n + s/n^{1/2}) \exp\{l_n(\hat{\theta}_n + s/n^{1/2}) - l_n(\hat{\theta}_n)\} - p(\theta_0) \exp(-s^{\mathrm{T}} \Sigma_0^{-1} s/2)\}\|_2 \, ds \\ &\leqslant \int_{A_3} \|s\|_2 \, p(\hat{\theta}_n + s/n^{1/2}) \frac{L_n(\hat{\theta}_n + s/n^{1/2})}{L_n(\hat{\theta}_n)} \, ds + \int_{A_3} \|s\|_2 \, p(\theta_0) \exp(-s^{\mathrm{T}} \Sigma_0^{-1} s/2) \, ds. \end{aligned}$$

Changing variables back to θ, the first integral on the right hand side is equal to

$$\int_{\|\theta - \hat{\theta}_n\|_2 > \delta} n^{(m+1)/2} \|\theta - \hat{\theta}_n\|_2 \, p(\theta) \frac{L_n(\theta)}{L_n(\hat{\theta}_n)} \, d\theta.$$

But

$$\int_{\|\theta - \hat{\theta}_n\|_2 > \delta} \|\theta - \hat{\theta}_n\|_2 \, p(\theta) \, d\theta \leqslant \int_{\|\theta - \hat{\theta}_n\|_2 > \delta} (\|\theta\|_2 + \|\hat{\theta}_n\|_2) \, p(\theta) \, d\theta,$$

and the right hand side is stochastically bounded by the finite moment assumption. Thus, by applying Theorem 3.1, the first integral tends to zero in probability. The second integral also tends to zero in probability by the tail properties of the multivariate normal distribution.

Furthermore,

$$\int_{A_1} \|s\|_2 \exp(-s^{\mathrm{T}}\Sigma_0^{-1}s/2)\,ds = O_{P_0}(1) \quad \text{and} \quad \int_{A_2} \|s\|_2 \exp\left(-s^{\mathrm{T}}\Sigma_0^{-1}s/4\right) ds \to 0$$

with convergence in probability, from which we can deduce that the integrals for A_1 and A_2 will also converge to 0 in probability using the same arguments as the proof of Theorem 3.2. □

Proof of Theorem 3.4. Theorem 3.2 implies L^1 convergence of the full posterior as $n \to \infty$

$$\int_\Theta \left| p(\theta \mid D_1, \ldots, D_n) - p_{\hat{\theta}_n, n^{-1}\Sigma_0}(\theta) \right| d\theta \to 0$$

with convergence in probability, where $\Theta \subset \mathbb{R}^m$ is the parameter space of θ, $p_{\hat{\theta}, n^{-1}\Sigma_0}$ is the density of $\mathcal{N}(\hat{\theta}_n, n^{-1}\Sigma_0)$, $\hat{\theta}_n = (\hat{\alpha}_n, \hat{\beta}_n, \hat{\rho}_n, \hat{\gamma}_n)$ and $\Sigma_0 = \lim_{n\to\infty} \mathrm{var}_{P_0}(n^{1/2}\hat{\theta}_n)$. It remains to show the corresponding result for the marginal posterior. Let $m_1 = \dim(\alpha) + \dim(\beta) + \dim(\rho)$, so that $(\alpha, \beta, \rho) \in \mathbb{R}^{m_1}$, and let $m_2 = \dim(\gamma)$, so $m_1 + m_2 = m$. The posterior density $p(\theta \mid d_1, \ldots, d_n)$ is assigned the value 0 outside of Θ.

$$\begin{aligned}
\int_\Gamma \left| p(\gamma \mid D_1, \ldots, D_n) - p_{\hat{\gamma}_n, n^{-1}V_0}(\gamma) \right| d\gamma &= \int_\Gamma \left| \int_{\mathbb{R}^{m_1}} p(\theta \mid D_1, \ldots, D_n) - p_{\hat{\theta}_n, n^{-1}\Sigma_0}(\theta)\, d\alpha\, d\beta\, d\rho \right| d\gamma \\
&\leqslant \int_\Gamma \int_{\mathbb{R}^{m_1}} \left| p(\theta \mid D_1, \ldots, D_n) - p_{\hat{\theta}_n, n^{-1}\Sigma_0}(\theta) \right| d\alpha\, d\beta\, d\rho\, d\gamma \\
&\leqslant \int_{\mathbb{R}^{m_2}} \int_{\mathbb{R}^{m_1}} \left| p(\theta \mid D_1, \ldots, D_n) - p_{\hat{\theta}_n, n^{-1}\Sigma_0}(\theta) \right| d\alpha\, d\beta\, d\rho\, d\gamma \\
&= \int_\Theta \left| p(\theta \mid D_1, \ldots, D_n) - p_{\hat{\theta}_n, n^{-1}\Sigma_0}(\theta) \right| d\theta \\
&\quad + \int_{\mathbb{R}^m \setminus \Theta} p_{\hat{\theta}_n, n^{-1}\Sigma_0}(\theta)\, d\theta.
\end{aligned}$$

The first term tends in probability to 0 by Theorem 3.2. This implies that

$$\int_\Theta p_{\hat{\theta}_n, n^{-1}\Sigma_0}(\theta)\, d\theta \to 1$$

with convergence in probability, so the second term also tends in probability to 0. □

Appendix D

Appendix for Chapter 4

D.1 Posterior density of θ for the Bayesian bootstrap

For the mean estimation example in Section 4.2, we derive the posterior density of θ for the Bayesian bootstrap. We can write $\theta = -q_1 + x_2 q_2 + q_3$ where the probability weights q_1, q_2, q_3 are distributed according to the uniform Dirichlet distribution. Since the weights sum to 1, we can alternatively write $\theta = 1 - 2q_1 + (x_2 - 1)q_2$ by removing q_3. The joint distribution of (q_1, q_2) is given by the uniform density

$$p_{BB}(q_1, q_2) = 2\,\mathbb{1}\{0 \leqslant q_1, 0 \leqslant q_2, q_1 + q_2 \leqslant 1\}. \tag{D.1}$$

One way to derive this is to note that the marginal distribution of q_1 is Beta$(1,2)$, which has density $2(1-q_1)\mathbb{1}\{0 \leqslant q_1 \leqslant 1\}$, and the conditional distribution of q_2 given q_1 is $U[0, 1-q_1]$, which has density $(1-q_1)^{-1}\mathbb{1}\{0 \leqslant q_2 \leqslant 1-q_1\}$.

Let $\phi = q_1$. We perform a change-of-variables from (q_1, q_2) to (θ, ϕ). The Jacobian factor satisfies

$$|J|^{-1} = \begin{vmatrix} 1 & -2 \\ 0 & x_2 - 1 \end{vmatrix} = |x_2 - 1| = 1 - x_2.$$

Thus, the joint density of (θ, ϕ) is

$$p_{BB}(\theta, \phi) = \frac{2}{1-x_2}\mathbb{1}\left\{0 \leqslant \phi, 0 \leqslant \frac{1-\theta-2\phi}{1-x_2}, \frac{1-\theta-\phi-x_2\phi}{1-x_2} \leqslant 1\right\} \tag{D.2}$$

by substitution of (θ, ϕ) into D.1. Let us rearrange the three inequalities inside the indicator function in terms of ϕ:

$$\begin{aligned}\phi &\geqslant 0\\ \phi &\leqslant \frac{1-\theta}{2}\\ \phi &\geqslant \frac{x_2-\theta}{1+x_2}.\end{aligned}$$

If $x_2 \leqslant \theta$, the third inequality is redundant; if $x_2 \geqslant \theta$, the first inequality is redundant.

To derive the marginal density of θ, we integrate ϕ out of D.2. For $x_2 \leqslant \theta$, we integrate between 0 and $(1-\theta)/2$ to get

$$p_{BB}(\theta) = \frac{1-\theta}{1-x_2}\mathbb{1}\{x_2 \leqslant \theta \leqslant 1\}.$$

For $x_2 \geqslant \theta$, we integrate between $(x_2-\theta)/(1+x_2)$ and $(1-\theta)/2$ to get

$$\begin{aligned}p_{BB}(\theta) &= \frac{2}{1-x_2}\left\{\frac{1-\theta}{2} - \frac{x_2-\theta}{1+x_2}\right\}\mathbb{1}\{-1 \leqslant \theta \leqslant x_2\}\\ &= \frac{1+\theta}{1+x_2}\mathbb{1}\{-1 \leqslant \theta \leqslant x_2\}.\end{aligned}$$

D.2 Computation for Kitamura & Otsu

In this section, we develop a blocked Gibbs sampler for the Kitamura and Otsu proposal using the truncated stick-breaking set-up introduced in §4.4.

D.2.1 Updating θ

The conditional posterior density of θ given everything else is proportional to

$$\pi(\theta)\prod_{i=1}^{n}\mathbb{P}_{\text{tilt}}(D = D_i \mid \theta, p_1, \ldots, p_K, A_1, \ldots, A_K).$$

Similar to our discussion in §4.3.3, there is considerable flexibility in selecting a proposal for θ, including simple options like random walk Metropolis-Hastings.

D.2.2 Updating $\{I_1, \ldots, I_K\}$

In terms of mixing, this is the most challenging block of parameters to update due to the size of the set of permutations on $\{1, \ldots, K\}$. We propose sampling J uniformly on $\{1, \ldots, K\}$ and proposing $\{I'_1, \ldots, I'_K\}$, where $I'_J = I_{J+1}$, $I'_{J+1} = I_J$, and $I'_k = I_k$ otherwise. In other words, we swap I_J with I_{J+1}. The conditional posterior density is proportional to the likelihood function.

D.2.3 Updating $\{B_1, \ldots, B_K\}$

As stated above, we have fixed $B_i = D_i$ for $i = 1, \ldots, n$. Hence, we only need to update $B_{n+1}, \ldots, B_K$. The conditional posterior density is proportional to

$$\left\{ \prod_{k=n+1}^{K} g_0(B_k) \right\} \left\{ \prod_{i=1}^{n} \mathbb{P}_{\text{tilt}}(D = D_i \mid \theta, p_1, \ldots, p_K, A_1, \ldots, A_K) \right\} \tag{D.3}$$

where g_0 is the probability density/mass function of G_0. We suggest proposing $B'_{n+1}, \ldots, B'_K$ directly from G_0—an independence sampler. In the Metropolis-Hastings step, the product on the left of (D.3) will drop out of the acceptance ratio.

D.2.4 Updating $\{V_1, \ldots, V_{K-1}\}$

The conditional posterior density of $\{V_1, \ldots, V_{K-1}\}$ is proportional to

$$\left\{ \prod_{k=1}^{K-1} (1 - V_k)^{\alpha - 1} \right\} \left\{ \prod_{i=1}^{n} \mathbb{P}_{\text{tilt}}(D = D_i \mid \theta, p_1, \ldots, p_K, A_1, \ldots, A_K) \right\}.$$

We suggest using pre-conditioned Crank-Nicolson proposals. Let Φ be the cumulative distribution function of the standard normal distribution, and let $\rho \in [0, 1)$. For each $k = 1, \ldots, K-1$, we propose

$$V'_k = \Phi(\rho \Phi^{-1}(V_k) + \sqrt{1 - \rho^2}\varepsilon_k)$$

where $\varepsilon_1, \ldots, \varepsilon_{K-1} \sim^{\text{i.i.d.}} \mathcal{N}(0, 1)$.

D.3 Proofs

Proof of Proposition 4.1: We will use the shorthand $q_i = q_i(0,\tilde{q})$ for $i = 1,2,3$. The conditional posterior $p(\tilde{q}_1,\tilde{q}_2,\tilde{q}_3 \mid \theta = 0, x_1, x_2, x_3)$ is proportional to

$$\prod_{i=1}^{3} \frac{q_i}{\tilde{q}_i} = \prod_{i=1}^{3} \frac{e^{\lambda x_i}}{\sum_{j=1}^{3} \tilde{q}_j e^{\lambda x_j}}$$

where λ satisfies

$$-\tilde{q}_1 e^{-\lambda} + \tilde{q}_3 e^{\lambda} = 0.$$

Thus, $e^{\lambda} = \sqrt{\tilde{q}_1/\tilde{q}_3}$ and

$$\prod_{i=1}^{3} \frac{q_i}{\tilde{q}_i} = (\tilde{q}_2 + 2\sqrt{\tilde{q}_1\tilde{q}_3})^{-3}.$$

We further deduce that

$$q_1 = q_3 = \frac{\sqrt{\tilde{q}_1\tilde{q}_3}}{\tilde{q}_2 + 2\sqrt{\tilde{q}_1\tilde{q}_3}}, \quad q_2 = \frac{\tilde{q}_2}{\tilde{q}_2 + 2\sqrt{\tilde{q}_1\tilde{q}_3}}.$$

It is therefore sufficient to derive the form of the posterior for q_1; we subsequently have $q_3 = q_1$ and $q_2 = 1 - 2q_1$. The cumulative distribution function $\mathbb{P}(q_1 \leqslant t)$ for q_1 is proportional to

$$\int_{\frac{\sqrt{\tilde{q}_1\tilde{q}_3}}{\tilde{q}_2+2\sqrt{\tilde{q}_1\tilde{q}_3}} \leqslant t} (\tilde{q}_2 + 2\sqrt{\tilde{q}_1\tilde{q}_3})^{-3}\, d\tilde{q}_1 d\tilde{q}_2 d\tilde{q}_3$$

for $t \in [0, 0.5]$. We make the change of variables

$$\begin{aligned}
\tilde{q}_2 &= V_1 \\
\tilde{q}_1 &= V_2(1 - V_1) \\
\tilde{q}_3 &= (1 - V_1)(1 - V_2)
\end{aligned}$$

that maps the simplex to the unit square in $\mathbb{R}^2$ via stick-breaking; the integral is now equal to

$$\int_{0 \leqslant \frac{(1-V_1)\sqrt{V_2(1-V_2)}}{V_1+2(1-V_1)\sqrt{V_2(1-V_2)}} \leqslant t} (1 - V_1)\{V_1 + 2(1 - V_1)\sqrt{V_2(1 - V_2)}\}^{-3}\, dV_1 dV_2$$

where $(1-V_1)$ is the Jacobian factor. Rearranging the parameter set in terms of V_1, we get

$$\rho(t,V_2) := \frac{(1-2t)\sqrt{V_2(1-V_2)}}{(1-2t)\sqrt{V_2(1-V_2)}+t} \leqslant V_1 \leqslant 1.$$

We begin by integrating with respect to V_1 over the above interval. For the time being, we use the shorthand $W = \sqrt{V_2(1-V_2)}$ to reduce clutter. Integrating by parts (differentiating $(1-V_1)$ and integrating $\{V_1+2(1-V_1)W\}^{-3}$), we get

$$\begin{aligned}
\left[\frac{(1-V_1)\{V_1+2(1-V_1)W\}^{-2}}{-2\{1-2W\}}\right]_{V_1=\rho(t,V_2)}^{V_1=1} &- \int_{V_1=\rho(t,V_2)}^{1} \frac{\{V_1+2(1-V_1)W\}^{-2}}{2\{1-2W\}}dV_1 \\
&= \frac{t\{(1-2t)W+t\}}{2W(1-2W)} + \left[\frac{\{V_1+2(1-V_1)W\}^{-1}}{2(1-2W)^2}\right]_{V_1=\rho(t,V_2)}^{V_1=1} \\
&= \frac{t\{(1-2t)W+t\}}{2W(1-2W)} + \frac{2t-t/W}{2(1-2W)^2} \\
&= \frac{t^2}{2W^2} \\
&= \frac{t^2}{2V_2(1-V_2)}.
\end{aligned}$$

Integrating $\{V_2(1-V_2)\}^{-1}$ directly over the interval $V_2 \in [0,1]$ yields infinity. Instead, we consider the limit of

$$\int_{a_k}^{b_k} \frac{t^2}{2V_2(1-V_2)}\,dV_2$$

for $0 < a_k < b_k < 1$ as $a_k \to 0$, $b_k \to 1$. For each term in this sequence, we have

$$\mathbb{P}(q_1 \leqslant t) \propto t^2$$

for $t \in [0,0.5]$, i.e. $q_1 \sim 0.5 \cdot \text{Beta}(1,2)$. Therefore, we can define the conditional posterior of (q_1,q_2,q_3) to be this limit, which is proper and matches the required result. □

Proof of Proposition 4.2: We use the shorthand notation $g_i = g(D_i,\theta)$ and $q_i = q_i(\theta,\tilde{q})$. Using the notation of the dual optimization problem, λ satisfies

$$\sum_{i=1}^{n} \tilde{q}_i \exp(\lambda^{\mathrm{T}} g_i) g_i = 0.$$

Differentiating both sides by $\tilde{q}_j$,

$$\frac{\partial \lambda}{\partial \tilde{q}_j} \sum_{i=1} \tilde{q}_i \exp(\lambda^{\mathrm{T}} g_i) g_i g_i^{\mathrm{T}} + \exp(\lambda^{\mathrm{T}} g_j) g_j = 0,$$

which rearranges to

$$\frac{\partial \lambda}{\partial \tilde{q}_j} = -\left(\sum_{i=1}^{n} q_i g_i q_i^{\mathrm{T}} \right)^{-1} \left(\frac{q_j}{\tilde{q}_j} g_j \right). \tag{D.4}$$

The log-posterior density is

$$\begin{aligned} \log p(\tilde{q} \mid \theta, D_1, \ldots, D_n) &= \sum_{i=1}^{n} \log \left(\frac{q_i}{\tilde{q}_i} \right) + c(\theta) \\ &= \left(\sum_{i=1}^{n} \lambda^{\mathrm{T}} g_i \right) - n \log \left\{ \sum_{i=1}^{n} \tilde{q}_i \exp(\lambda^{\mathrm{T}} g_i) \right\} + c(\theta), \end{aligned}$$

where $c(\theta)$ is independent of $\tilde{q}$. Hence,

$$\begin{aligned} \frac{\partial \log p(\tilde{q} \mid \theta, D_1, \ldots, D_n)}{\partial \tilde{q}_j} &= \frac{\partial \lambda^{\mathrm{T}}}{\partial \tilde{q}_j} \left(\sum_{i=1}^{n} g_i \right) - \frac{n \exp(\lambda^{\mathrm{T}} g_j)}{\sum_{i=1}^{n} \tilde{q}_i \exp(\lambda^{\mathrm{T}} g_i)} - \frac{n \frac{\partial \lambda^{\mathrm{T}}}{\partial \tilde{q}_j} \sum_{i=1} \tilde{q}_i \exp(\lambda^{\mathrm{T}} g_i) g_i}{\sum_{k=1}^{n} \tilde{q}_k \exp(\lambda^{\mathrm{T}} g_k)} \\ &= \frac{\partial \lambda^{\mathrm{T}}}{\partial \tilde{q}_j} \left(\sum_{i=1}^{n} g_i \right) - n \frac{q_j}{\tilde{q}_j} - n \frac{\partial \lambda^{\mathrm{T}}}{\partial \tilde{q}_j} \left(\sum_{i=1}^{n} q_i g_i \right) \\ &= \frac{\partial \lambda^{\mathrm{T}}}{\partial \tilde{q}_j} \left(\sum_{i=1}^{n} g_i \right) - n \frac{q_j}{\tilde{q}_j}, \end{aligned}$$

where the last equality is due to $\sum_{i=1}^{n} q_i g_i = 0$. The required result is obtained by substituting (D.4) into the right-hand side. □

Proof of Proposition 4.3: As with the proof of Proposition 4.2, we use the shorthand notation $g_i = g(D_i, \theta)$ and $q_i = q_i(\theta, \tilde{q})$, and using the notation of the dual optimization problem, λ satisfies

$$\sum_{i=1}^{n} \tilde{q}_i \exp(\lambda^{\mathrm{T}} g_i) g_i = 0.$$

Differentiating both sides by θ,

$$\sum_{i=1}^{n} \tilde{q}_i \frac{\partial \lambda}{\partial \theta} \exp(\lambda^{\mathrm{T}} g_i) g_i g_i^{\mathrm{T}} + \sum_{i=1}^{n} \tilde{q}_i \exp(\lambda^{\mathrm{T}} g_i)(I_n + g_i \lambda^{\mathrm{T}}) \frac{\partial g_i}{\partial \theta} = 0.$$

By rearranging, we obtain

$$\frac{\partial \lambda}{\partial \theta} = -\left(\sum_{i=1}^{n} q_i g_i g_i^{\mathrm{T}} \right)^{-1} \left\{ \sum_{j=1}^{n} q_j (I_n + g_j \lambda^{\mathrm{T}}) \frac{\partial g_j}{\partial \theta} \right\}.$$

The log-posterior density is

$$\begin{aligned} \log p(\tilde{q} \mid \theta, D_1, \ldots, D_n) &= \log \pi(\theta) + \sum_{i=1}^{n} \log \left(\frac{q_i}{\tilde{q}_i} \right) + c(\tilde{q}) \\ &= \log \pi(\theta) + \left\{ \sum_{i=1}^{n} (\lambda^{\mathrm{T}} g_i) \right\} - n \log \left\{ \sum_{j=1}^{n} \tilde{q}_j \exp(\lambda^{\mathrm{T}} g_j) \right\} + c(\tilde{q}), \end{aligned}$$

where $c(\tilde{q})$ is independent of θ. Hence,

$$\begin{aligned} \frac{\partial \log p(\tilde{q} \mid \theta, D_1, \ldots, D_n)}{\partial \theta} &= \frac{\nabla \pi(\theta)}{\pi(\theta)} + \sum_{i=1}^{n} \frac{\partial (\lambda^{\mathrm{T}} g_i)}{\partial \theta} - n \frac{\sum_{j=1}^{n} \frac{\partial (\lambda^{\mathrm{T}} g_j)}{\partial \theta} \tilde{q}_j \exp(\lambda^{\mathrm{T}} g_j)}{\sum_{k=1}^{n} \tilde{q}_k \exp(\lambda^{\mathrm{T}} g_k)} \\ &= \frac{\nabla \pi(\theta)}{\pi(\theta)} + \sum_{i=1}^{n} (1 - n q_i) \frac{\partial (\lambda^{\mathrm{T}} g_i)}{\partial \theta} \\ &= \frac{\nabla \pi(\theta)}{\pi(\theta)} + \sum_{i=1}^{n} (1 - n q_i) \left(\frac{\partial \lambda^{\mathrm{T}}}{\partial \theta} g_i + \frac{\partial g_i^{\mathrm{T}}}{\partial \theta} \lambda \right) \\ &= \frac{\nabla \pi(\theta)}{\pi(\theta)} + \frac{\partial \lambda^{\mathrm{T}}}{\partial \theta} \left(\sum_{i=1}^{n} g_i \right) + \sum_{i=1}^{n} (1 - n q_i) \frac{\partial g_i^{\mathrm{T}}}{\partial \theta} \lambda, \end{aligned}$$

where we have used the fact that $\sum_{i=1}^{n} q_i g_i = 0$ for the last equality. □