Two Biostatistical Problems

by

Elizabeth Crenshaw Chase

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2023

Doctoral Committee:

Associate Professor Philip S. Boonstra, Co-Chair
Professor Jeremy M. G. Taylor, Co-Chair
Assistant Professor Yang Chen
Professor Jian Kang

Elizabeth Crenshaw Chase
ecchase@umich.edu
ORCID iD: 0000-0003-0452-2976

DEDICATION

For Pris—

who prized education, but not above all else.

ACKNOWLEDGMENTS

In the past six years, I have been surrounded by lovely people who have supported me in myriad ways. It is a gift to have so many people to thank.

First, I must thank my advisers, Phil Boonstra and Jeremy M.G. Taylor, who have been wonderful to me since Admitted Students' Day. At a job interview earlier this year, I was complimented on my realism about the limits of statistics to answer important questions. Although I am unsure if I merit that praise, if I ever do, I will know who taught me this clear-eyed approach to statistics, data, and research, and how to maintain my wonder at the potential and beauty of statistics, even when reminded daily of its shortcomings. From the start I have loved scribbling with you on the whiteboard, bickering about and becoming excited by each others' ideas. I have felt so lucky to have advisers I look forward to seeing each week, and whose meetings I always leave feeling better. Thank you for your generosity, your kindness, your ample and prompt feedback, your patience, and for teaching me so much.

I have received mentorship through a variety of channels. I have worked with Matt Schipper, Bob Dess, and Will Jackson on prostate cancer treatment decision-making since 2019. I have learned a great deal from them about effective statistical collaboration. May I be so fortunate as to have such brilliant and fun collaborators in future! I have benefited from thoughtful feedback and suggestions from Veera Baladandayuthapani, Nick Henderson, and the rest of the TaBaBooHe lab. I am grateful to my committee members, Jian Kang and Yang Chen, who have been gracious with their time and expertise. I have also received mentorship and support from older students in my department, who have listened to me, answered my questions, and provided me with resources: thank you to Lauren Beesley (the best peer mentor I could have asked for), Krithika Suresh, Emily Roberts, Holly Hartman, Emily Hector, and Tian Gu.

It is difficult to imagine my PhD without my grad school besties, Fatema Shafie Khorassani and Pedro Orozco del Pino. Whether discussing the 801 homework set, an insect infestation, a major cooking project, my love life, or how to get a job and wrap up my dissertation, you have always provided me with joy, good humor, and wise counsel. I have been lucky in my officemates (and friends): Emily, Madeline, Lam, Nate, Jess, Grant, and Rachel. It is not easy to make a windowless, fluorescent-lit, cinderblock room into a place of cheerful and cozy productivity, yet somehow you

all did it. I will miss getting to share space with you next year. There are too many friends to count, but I want to thank in particular Rachel, Emilee, Jesse, Ezra and Ilana, Megan, Amy, Sarah, Lulu, Aubrey, Nathaniel, and Sumeet for their friendship and support these past years.

Thank you to Nicole Fenech, Fatma Zohra-Nedjari, Dave Kubacki, Mandi Larson, and the other fabulous members of the U-M biostatistics administrative team for keeping the department running smoothly. I would also like to acknowledge the National Science Foundation, the NIH Biostatistics in Cancer Training Program, and the Rackham Graduate School for funding me over the years.

Last but certainly not least, I must thank my terrific family: my parents, Ben and Sarah; my siblings, Katherine and Isaac; my Aunt Beth; my beloved Pop, Pris, Granny, and Granddad; and my marvelous aunts, uncles, and cousins. I distinctly recall a phone call with my Pop, who expressed bewilderment at why I would apply for more than one job—"I'm sure everyone will hire you, and you only need one job!" This and similar sentiments from my family over the course of my PhD have always boosted my spirits and made me smile. Even when undeserved, your confidence and support has meant the world to me. I cherish every moment we have together.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

LIST OF ACRONYMS

AalJo Aalen-Johansen estimator

Adspline adaptive spline model

BBT basal body temperature

BMI body mass index

CumSum cumulative sum test

GPR Gaussian process regression

HMM hidden Markov model

HMC Hamiltonian Monte Carlo

HPR horseshoe process regression

HPR-BBT BBT-specific version of HPR

KL Kullback-Leibler

KMI Kaplan-Meier imputation

MAD Mean absolute difference

MCMC Markov Chain Monte Carlo

MedFilt median filter

NUTS No-U-Turn-Sampler

PSA prostate specific antigen

RGI Ruan and Gray imputation

RSI risk set imputation

SPMRF shrinkage prior Markov random fields

TrendFilt penalized trend filter

VI variational inference

ABSTRACT

This dissertation examines two problems in biostatistics. The first and second projects develop horseshoe process regression (HPR), a Bayesian nonparametric model that uses statistical shrinkage to capture abruptly changing associations between a continuous predictor and some outcome. We use HPR to model women's basal body temperature (BBT) across the menstrual cycle. In contrast, the third project proposes a nonparametric multiple imputation approach to estimating the cumulative incidence, a key descriptive statistic in survival analysis.

Focusing on the first project, we state the truism: biomedical data often exhibit jumps or abrupt changes. These sudden changes make these data challenging to model, as many methods will oversmooth the sharp changes or overfit in response to measurement error. We develop HPR to address this problem. We define a horseshoe process as a stochastic process in which each increment is horseshoe-distributed. We use the horseshoe process as a nonparametric Bayesian prior for modeling an association between an outcome and its continuous predictor. We provide guidance and extensions to advance HPR's use in applied practice: we introduce a Bayesian imputation scheme to allow for interpolation at unobserved values of the predictor within the HPR; include additional covariates via a partial linear model framework; and allow for monotonicity constraints. We find that HPR performs well when fitting functions that have sharp changes, and we use it to model women's BBT over the course of the menstrual cycle.

In the second project, we focus on using HPR for one particular type of abruptly changing data: BBT over the course of the menstrual cycle. Women's BBT exhibits abrupt changes at the time of ovulation and menstruation, which many methods struggle to capture. While in the first project we demonstrated that HPR had potential for modeling BBT, in the second project we tailor HPR for this setting. We re-implement HPR using variational inference to speed computation time, which we show offers comparable results to those provided by Hamiltonian Monte Carlo in the first project. We incorporate ovulation pattern into the HPR model, to provide posterior estimates of ovulation day and its uncertainty. We consider a posterior-prior passing scheme in order to share information across cycles. We use this BBT-specific version of HPR (HPR-BBT), to analyze BBT data from a large cohort of British women. Overall, HPR-BBT offers sensible estimates of ovulation day and BBT trajectory.

And now for something completely different: the third project. We propose an alternative approach to the Aalen-Johansen estimator of the cumulative incidence. Rather than calculate the cumulative incidence directly, we instead perform nonparametric multiple imputation to generate event times and types for censored individuals. Thus, on each imputation, all participants are "observed" to have an event. Calculating the cumulative incidence on each imputation is then merely estimating a proportion at each timepoint, and yields point and uncertainty estimates that can be aggregated across imputations via Rubin's Rules. The resulting multiple imputation estimator is mathematically and empirically shown to generate equivalent point estimates to the Aalen-Johansen estimator as the number of imputations increases; in addition, the multiple imputation estimator offers improved options for uncertainty estimation. We discuss connections to redistribute-to-the-right algorithms and other imputation approaches for survival analysis.

CHAPTER 1

Introduction

As the title suggests, this dissertation studies two very different biostatistical problems. In the first and second projects, we develop horseshoe process regression (HPR), a Bayesian nonparametric model that uses statistical shrinkage to capture abruptly changing associations between a continuous predictor and some outcome. In contrast, the third project proposes a nonparametric multiple imputation approach to estimating the cumulative incidence function, a key descriptive statistic in the competing risks survival analysis setting. Perhaps the only thing these two topics have in common is the word "nonparametric," which does not so much highlight their similarity as raise the question: what business do Bayesians have being nonparametric? Although largely unconnected, these two topics share the goal of analyzing complex longitudinal data with a Bayesian flair and providing statistical methodology that is prepared for the challenges of real data.

In the first and second projects, we focus on data with varying degrees of smoothness in the association between a continuous predictor and some outcome over the domain of the association. The data may have linear, smooth, or constant portions interspersed with jumps, spikes, drops, or abrupt changes in concavity. One example of such data is women's basal body temperature (BBT) data over the menstrual cycle, which we focus on in this dissertation. BBT features a sharp jump at the time of ovulation, followed by a sharp drop at menstruation. However, other examples of this kind of data might include data on human gait [35] or cancer biomarker data such as prostate specific antigen (PSA) for prostate cancer [45] or CA125 for ovarian cancer [70]. These data are challenging to model because many existing statistical methods are intended for data with a constant degree of smoothness across the domain: consistently smooth or consistently not smooth. As a result, existing methods force a choice between a fit that captures the smooth portions of the association but oversmooths the abruptly changing components (e.g. generalized linear models, splines, Gaussian process regression), or a fit that captures the abruptly changing components but introduces excess motion into the smooth components (e.g. the generalized lasso, the median filter). In either case, applying these methods to data with different levels of smoothness across the domain may yield results that fit the data poorly, have incorrect uncertainty estimates, and make it

difficult to identify the locations of the shifts between low and high variability, which is often of applied interest.

The use of shrinkage priors within a Bayesian nonparametric setting is one recent solution to this problem, first explored by Faulkner and Minin (2018) [22] and Kowal et al. (2019) [48]. Suppose we have some continuous outcome y_t observed at predictor value t, $t = 1, \ldots, n$. Then in this approach, we assume the model:

$$\begin{aligned} y_t &= f_t + \epsilon_t \\ f_1 &\sim N(a, b^2) \\ f_t - f_{t-1} &\sim G,\ t = 2, \ldots, n \\ \epsilon_t|\sigma^2 &\sim N(0, \sigma^2),\ \sigma^2 \sim H \end{aligned} \tag{1.1}$$

In this model, we assume that y_t is normally distributed about its mean, f_t. The first value of f_t, f_1, has its own prior, which for simplicity we assume is a normal distribution with mean a and variance b^2. For all timepoints after $t = 1$, f_t is a random walk in which the incremental change over t is given by some distribution G. The choice of G will dictate the shape of f_t. If we assume that G is a normal distribution, then f_t will be a Gaussian process, which provides a fairly smooth trend.

Rather than use a normal distribution for G, Faulkner and Minin (2018) suggested the use of Bayesian shrinkage priors, such as the double-exponential prior, the normal-gamma prior, or the horseshoe prior [22]. These shrinkage priors put most of their mass on either very large or very small signals, a desirable property for statistical shrinkage [12]. When G is a shrinkage prior, then f_t has an abruptly changing and dynamic shape, alternating between stretches during which f_t changes very little, and other portions when f_t may make large jumps. Faulkner and Minin (2018) found that using a horseshoe prior for G yielded particularly dynamic fits, imposing a prior that looked somewhat like a step function (Figure 1.1) [22].

When a horseshoe distribution is used for G, Equation 1.1 becomes:

$$\begin{aligned} y_t &= f_t + \epsilon_t \\ f_1 &\sim N(a, b^2) \\ f_t - f_{t-1}|\tau, \lambda_t &\sim N(0, \tau^2\lambda_t^2),\ t = 2, \ldots, n \\ \tau &\sim C^+(0, 1),\ \lambda_t \overset{iid}{\sim} C^+(0, 1),\ t = 2, \ldots, n \\ \epsilon_t|\sigma^2 &\sim N(0, \sigma^2),\ \sigma^2 \sim H \end{aligned} \tag{1.2}$$

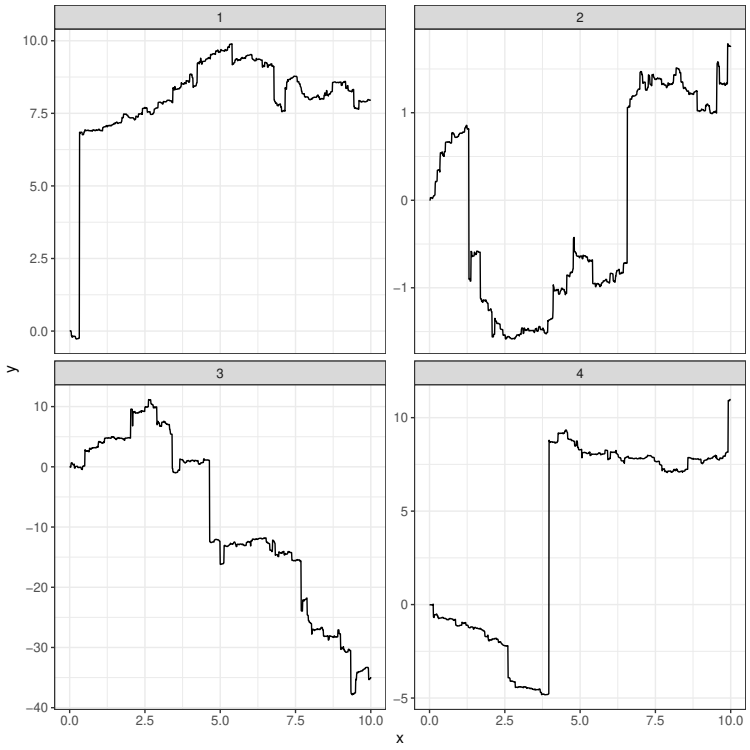

Figure 1.1: Four sample draws from the prior for a random walk in which incremental change is assumed to be horseshoe distributed.

In Equation 1.2, G is conditionally normally distributed, with a hierarchical variance term that depends on two parameters, τ and λ_t, which we call the global and local shrinkage parameters, respectively. While τ controls the overall amount of motion in f_t, each increment of change, $f_t - f_{t-1}$, gets its own local shrinkage parameter λ_t, which enables the model to capture abrupt changes in f_t.

Although these local shrinkage parameters make this type of model—which we call horseshoe process regression (HPR)—very flexible, they also cause challenges. First, they scale with sample size, as there are $n - 1$ local shrinkage parameters. This makes the model somewhat computationally intensive, particularly at larger values of n. Second, the finite set of local shrinkage parameters means that the underlying "horseshoe process" is neither infinitely divisible nor self-similar, two desirable features for a stochastic process that ensure that the probability structure of the model is the same regardless of the grid at which we observe it. As we discuss in Chapter 2, while it is possible to formulate a version of HPR using subordinated Brownian motion that does not have these limitations, mathematical and computational barriers remain high to actually implementing that model [63]. Instead, we use the discretely-observed formulation in Equation 1.2, which relies on a finite set of local shrinkage parameters, the size of which is dictated by the observed data.

This approach generally yields good performance, but it causes difficulties when we want predictions or interpolations from the HPR, and it complicates information-sharing on the local shrinkage parameters in the presence of repeated measurements.

In Chapter 2, our emphasis is on making HPR more usable in practice. We clarify the formulation of HPR and its connections to stochastic processes. We assess HPR's performance for modeling dynamically changing data and find that it offers excellent performance for step functions, piecewise linear functions, and impulse functions. We consider approaches for data interpolation within HPR and recommend a Bayesian imputation scheme to sample local shrinkage parameters and values of f at unobserved locations, which we find yields sensible results. We also provide guidance on including additional linear covariates, imposing monotonicity constraints, and on setting hyperparameters, with additional simulations and sensitivity analyses to justify our recommendations. We use HPR to model women's BBT data taken from Weschler (2015) [80]; based on those results, HPR has potential as a methodology for modeling BBT data.

In Chapter 3, we focus more deeply on using HPR for women's BBT data, building on the preliminary analysis we completed in Chapter 2. Over the course of a single menstrual cycle, BBT has a step-function shape, with lower temperatures during the first half of the menstrual cycle, followed by a sharp increase in temperature at the time of ovulation, with sustained high temperatures through the second half of the menstrual cycle. With the onset of menstruation, the BBT drops back to its lower pre-ovulation temperatures and the pattern repeats. This biphasic BBT pattern has been the subject of great interest in the medical literature because it provides a noninvasive method to detect ovulation [53]. Although this general pattern may hold in up to 90% of women [73], there can be extensive individual and cycle-to-cycle variation in the sharpness of the ovulation/menstruation BBT jump, the timing of ovulation, and the pre- and post-ovulation BBT itself. A good statistical method for analyzing BBT needs to be able to model a variety of BBT patterns; to allow information-sharing while accommodating variation across cycles; to provide an estimate of when ovulation occurred based on the modeled BBT trajectory; and be computationally speedy.

In Chapter 3, we modify the HPR model from Chapter 2 to provide these features. First, we introduce ovulation day as a parameter in the model, which allows us to provide estimates of its posterior distribution. Although we choose the prior for this parameter based on clinical knowledge of ovulation timing specifically, this general approach could extend to other changepoint detection settings outside of BBT data. Second, we re-implement HPR using variational inference to hasten estimation. Variational inference aims to estimate an approximate posterior distribution that is simpler to obtain than the true posterior [31]. The variational inference implementation that we propose offers major computational speed gains relative to full Markov Chain Monte

Carlo (MCMC); however, it requires careful choices of initial values and hyperparameters. Although the variational inference implementation's results are very similar to those of MCMC, they are often slightly worse, creating a minor tradeoff between computational speed and model performance. In most settings, we think the computational gains are worth it, but further work is needed to evaluate this [83]. Third and finally, we propose a posterior-prior passing scheme to share information across cycles for the same woman [18]. This approach, in contrast to a more traditional mixed model formulation [25], allows us to sidestep some of the challenges of information-sharing in the presence of the local shrinkage parameters. Using BBT data measured in a large cohort of British women [55, 76], we demonstrate that our ovulation-detection approach, variational inference algorithm, and posterior-prior passing scheme usually return sensible results in real data, and that including information from previous cycles improves model convergence rates and reduces the number of days of data needed to detect ovulation. However, more work is needed to further fine-tune HPR for modeling BBT.

In Chapter 4, we study a different type of complex longitudinal data: censored data in the presence of competing risks. Censoring occurs when we wish to measure a time-to-event outcome (such as time to death) but are unable to observe the event of interest. This is often caused by study termination or participant dropout. As a result, we are left with partially missing data. We know that the individual has survived event-free to the time of censoring, but not what happened afterwards. When we can assume that the censoring occurred independently of the event of interest, we have ample statistical methodology to analyze the resulting censored data and make full use of the information it contains [46]. Unfortunately, we sometimes cannot assume independent censoring. Here, we focus on one kind of dependent censoring: death from a *competing event*. If we are primarily interested in death from a particular cause (such as death from cancer), then when individuals die from another cause (such as heart attack or car accident) we are unable to observe their time to the event of interest, because they have already died of something else. This is censoring. However, assuming that a participant's death from a competing event is independent from their likelihood of dying of the event of interest is often implausible, making the independent censoring assumption untenable. We might suspect that an individual who died of a heart attack had a fundamentally different risk of dying of cancer than an individual who actually did die of cancer. Death from a competing event cannot be treated as independent censoring. Thus, the presence of competing events results in three different outcomes in our study: experiencing the event-of-interest (fully observed), independent censoring (partially observed; caused by things like dropout), and death from a competing event (partially observed; dependent censoring). Many of the methods we use for independently-censored data cannot be used in this setting, as they will return biased results [46].

The cumulative incidence function is a key descriptive statistic in the competing events setting. It is the probability of experiencing the event of interest by time t. It is usually estimated via the Aalen-Johansen estimator [2]. Although the Aalen-Johansen estimator is the most common approach to estimate the cumulative incidence, it can be difficult to adapt to unique applied problems, especially if variance estimates are desired. A more intuitive approach comes when we reformulate the Aalen-Johansen estimator as a redistribute-to-the-right algorithm, as was done by Efron (1967) [19] and Gooley et al. (1999) [33]. In a redistribute-to-the-right version of the Aalen-Johansen estimator, each participant in the study has equal weight: if our sample size is n, then each participant has weight $1/n$. If an individual is censored (from independent causes such as loss-to-follow-up or study completion), then we redistribute their weight equally among the individuals who remain alive at the time of the censoring. After redistributing the weights of the independently-censored individuals to individuals who were not censored during study follow-up (e.g. those who died of the event of interest or the competing event), it is straightforward to estimate the cumulative incidence at time t as the sum of the redistributed weights of the individuals who died of the event of interest by t, divided by the total sample size n. Working with this weighted proportion can be more straightforward than working with the Aalen-Johansen estimator.

More straightforward still is to replace the reweighting with multiple imputation, an idea used in the all-cause survival analysis setting by Taylor et al. (2002) [74]. Rather than reweighting the sample to account for the independent censoring, we would instead generate multiple imputations of an event time and type for censored individuals, so that each imputed dataset consists only of observed or imputed times for the event of interest or the competing event. We can estimate the cumulative incidence at time t on each imputed dataset as the proportion of deaths from the event of interest by time t—no weighting required. The resulting proportion estimates and their variances can be aggregated using Rubin's Rules to provide final estimates of the cumulative incidence [50].

That is the approach we take in Chapter 4, in which we propose two multiple imputation approaches to estimate the cumulative incidence function. We demonstrate mathematically that with infinite imputations, the multiple imputation approach will return equivalent results to the Aalen-Johansen estimator; we demonstrate empirically that even with a finite number of imputations the multiple imputation approach will return results that are very similar to Aalen-Johansen. If we set aside the goal of mimicking Aalen-Johansen, even at very small numbers of imputations (e.g. 5-10) the multiple imputation approach shows good performance for estimating the true underlying cumulative incidence. In addition, the use of a binomial proportion to estimate the cumulative incidence on each imputation allows for connections to longstanding work on uncertainty estimation for binomial proportions [3, 81], which motivates our proposal of two alternative variance estima-

tors for the cumulative incidence function. These alternative variance estimators give improved performance over existing options when the event rate or sample size are low.

Finally, in Chapter 5, we highlight opportunities for future work on HPR, modeling BBT, and the use of multiple imputation estimators in survival analysis. There are a variety of directions for further biostatistical research.

CHAPTER 2

Modeling Data Using Horseshoe Process Regression

2.1 Introduction

Consider a longitudinal outcome, such as a man's PSA tracked throughout treatment for prostate cancer or a woman's basal body temperature measured during the menstrual cycle. Common approaches to model these data might include a generalized linear model, splines, or a version of Gaussian process regression, all of which would recover an estimate of the underlying association between time and the outcome, provide a measure of uncertainty about that association, and enable interpolations between observed datapoints, with varying degrees of assumptions, computational speed, smoothness, and interpretability [36]. However, the examples given above—and many other types of biomedical and scientific data—share a common trait: the possible presence of jumps, kinks, or steps in the association between the outcome and predictor. In the case of PSA and prostate cancer, we might see such a jump after the removal of the prostate, when PSA plummets, and again when radiation therapy is initiated. In between these drops, we would observe more gentle increases in PSA as the prostate cancer progresses [45]. For women's basal body temperature, this jump occurs at ovulation, when body temperature increases sharply in response to progesterone release, after which point body temperature will jump or gradually decline back to the pre-ovulation temperature by the end of the menstrual cycle [66]. The number of jumps and their locations may not be known *a priori*. In the case of PSA monitoring, we would expect to know when the patient received particular treatments, but with women's basal body temperature, it is less likely that we would know the location of the jump(s) without some examination of the data. An optimal statistical method would recognize the jumps—regardless of whether their locations are known *a priori*—and allow them to be sharp, but without introducing noise in the smooth parts of the association. This is where the methods listed above struggle, failing to accommodate both the sharp jumps and the smooth portions [49].

This chapter introduces horseshoe process regression, a method to model data featuring sharp jumps and smooth portions when the locations of the jumps are not known *a priori*. To do so,

we adapt Gaussian process regression by using a horseshoe process prior rather than a Gaussian process prior. The horseshoe process prior accommodates large jumps and constant stretches, and uses information from the data to place the jumps. Extensions allow for interpolations and predictions at unobserved datapoints, the inclusion of multiple predictors (both linear and nonlinear), non-Gaussian outcomes, and monotonicity constraints. Because the model is implemented in a fully Bayesian framework, uncertainty estimates are straightforward to calculate. We provide an R package, HPR, which includes all of these features and other functions useful for applied practice.

Horseshoe process regression adds to the extensive literature on changepoint and step detection modeling, sometimes called mean or trend filtering. Early methods included cumulative sum testing approaches [66] or running median filters [49]. Other common approaches include the use of low-degree splines [49], often with some kind of penalty [20], or LASSO variants [75]. Gaussian process models, modified to be autoregressive or nonstationary, are another flexible option [68]. Many sophisticated methods have come from the econometrics literature. These include jump diffusion models, jump processes, and stochastic volatility models [14, 47]. Although these sophisticated approaches produce very flexible fits, they are often poorly suited to the biomedical setting: performance relies on large numbers of observations, usually equally-spaced, that are rarely available in patient biomarker data, and as a result, overfitting and model nonconvergence can be serious concerns in small or unequally spaced samples. There is little consideration of non-Gaussian outcomes. In addition, because time is often the predictor of interest in these settings, there is no allowance for observations at the same predictor value, as might be seen in the biomedical setting when working with dose-toxicity data, in which the predictor is dose and multiple patients could be assigned to receive the same dose.

Faulkner and Minin (2018) [22] were the first to recognize the potential of horseshoe processes for nonparametric curve fitting in their exploration of Bayesian shrinkage priors for trend filtering on k^{th} order differences, which they call shrinkage prior Markov random fields (SPMRF). They presented evidence that the horseshoe prior was well-suited to piecewise-constant curve estimation within the context of Bayesian trend filtering. Similarly, Kowal et al. (2019) [48] developed dynamic shrinkage processes for use in stochastic volatility modeling. Their model formulation resembles ours and SPMRF, but its primary focus is on a dynamically dependent variant of the model. In addition, it is targeted to econometric applications, which limits its use in the biomedical setting, with difficulties with repeated or unequally spaced predictor values. We build upon both Faulkner and Minin [22] and Kowal et al. [48] by allowing for data interpolation, additional linear predictors, and monotonicity constraints, and the insight we offer into computational performance. Another key building block of our work is that of Boonstra et al. [8], in their use of horseshoe

priors for isotonic regression with categorical predictors and binary outcomes. We extend that approach by allowing for continuous predictors through the use of a horseshoe process prior, and we consider continuous and count outcomes in addition to binary outcomes. We also consider data that are not monotonic. Our finished product is a versatile approach for fitting piecewise constant and piecewise linear functions in biomedical settings, with flexible extensions to allow for additional covariates, discrete outcomes, and monotonicity constraints.

The structure of this chapter is as follows. First, we present some technical background on the horseshoe prior and the underlying theory of horseshoe processes. Second, we present the model formulation for horseshoe process regression. We review the extensions to allow for data interpolation, additional covariates, and monotonic constraints, and provide computational details. Third, we present a simulation study to characterize horseshoe process regression's performance. We demonstrate the use of horseshoe process regression in a dataset of women's basal body temperatures from Weschler (2015) [80]. We conclude with a discussion of limitations and directions for future work.

2.2 Technical Background

2.2.1 The Horseshoe Prior

Consider the classic linear regression problem in which we have observations y_i and a length p vector of predictors $\mathbf{x_i}$ for each individual $i = 1, ..., n$. We might wish to fit a multivariable linear regression of the form $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + \epsilon_i,\ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. If p is large—possibly even larger than n—we may suspect that many of the coefficients $\beta_j,\ j = 1, ..., p$ are close to zero, with little effect on the outcome $\mathbf{y}$. Maximum likelihood estimation of the regression is not possible if $p > n$; even if $p \leq n$, the estimates of the coefficients may be highly variable. Shrinkage estimation is one approach to address this problem, in which we use a penalty or other constraint in model estimation to *shrink* the coefficients closer to zero. Ideally, this shrinkage constraint will push the coefficient estimates for predictors that have little association close to zero while leaving untouched the coefficient estimates for predictors that have a large association with the outcome [36].

The horseshoe prior is a popular Bayesian approach for shrinkage estimation [12]. It takes the form $p(\beta_j | \tau^2, \lambda_j^2) \sim N(0, \tau^2 \lambda_j^2)$, with $\tau \sim C^+(0, c)$ and, independently, $\lambda_j \overset{iid}{\sim} C^+(0, 1)$, where $C^+(a, b)$ denotes a half-Cauchy distribution with location parameter a and scale parameter b. We call τ the *global shrinkage parameter*, as it provides an overall measure of shrinkage on the β_j's. If τ is large, this prior admits many large β_j's; if τ is small, the β_j's are pushed towards zero.

However, the horseshoe prior also contains a set of *local shrinkage parameters*, $\lambda_j, j = 1, ..., p$, one for each β_j. The local shrinkage parameters allow individual β_j's to attain high values, even if τ is small. The marginal density function of the horseshoe prior takes on a distinctive shape, featuring an infinite spike at zero, along with moderately heavy tails. As a result, the horseshoe distribution approaches zero faster than a normal distribution, but while admitting large values with high probability. It has been shown that separately for each coefficient, the horseshoe favors either total shrinkage (estimate of β_j close to zero) or minimal shrinkage (leaving the estimate of β_j close to its estimate under maximum likelihood estimation) [12].

The horseshoe prior provides excellent shrinkage performance and resolves many of the computational difficulties of earlier Bayesian approaches to shrinkage estimation [12, 56]. However, it has drawbacks: the heavy tails of the half-Cauchy priors on τ, λ_j can cause problems with Bayesian model convergence. Proposed solutions to these problems include the use of decentered parameterizations to reduce *a priori* parameter correlation [59]; imposing additional regularization on the tails of the half-Cauchy priors [62]; and placing an additional hyperprior on the scale parameter c of the prior on τ [61].

2.2.2 Horseshoe Processes

We will now switch focus to stochastic processes. Suppose $y(t)$ is some outcome recorded at time t. We use time for simplicity, but note that t could be any continuous predictor. In its simplest form, a Gaussian process is defined as $y(t) - y(t - s) \sim N(0, s\tau^2)$. A Gaussian process assumes that incremental change over time is normally distributed. It relies on a single parameter, τ^2, which is the variance of the process. If τ^2 is large, the process will vary substantially over time; if τ^2 is small, the process will remain fairly constant over time. The variance between measurements increases as the elapsed time between them, s, increases.

We seek to define a horseshoe process, which takes a similar structure to a Gaussian process, except incremental change over time will be horseshoe distributed rather than normally distributed. Polson and Scott (2010) [63] demonstrated that all local-global shrinkage distributions based on scale mixtures of normals (like the horseshoe) can be extended to a stochastic process within the framework of subordinated Brownian motion. They found that a horseshoe process $H(t)$ can be represented as $H(t) \sim N(0, \exp(M(t)))$ where $M(t)$ is a Meixner process. Although this definition of a horseshoe process is the most mathematically complete, obstacles lie in the way of implementing it. The Meixner process has few workable computational implementations [54]. In addition, because the Meixner-subordination occurs on the log-variance scale for the horseshoe process (rather than the variance scale), it is challenging to understand the covariance structure and other properties of the horseshoe process [63]. In light of these challenges, we use a discrete

formulation of the horseshoe process. For a set of timepoints t_k, $k = 1, ..., m$, define a horseshoe process H_{t_k} as:

$$\begin{aligned} H_{t_k} - H_{t_{k-1}} | \tau, \lambda_k &\sim N(0, \tau^2 \lambda_k^2 (t_k - t_{k-1})),\ k = 2, ..., m \\ H_{t_1} &= 0 \\ \tau &\sim C^+(0, c) \\ \lambda_k &\overset{iid}{\sim} C^+(0, 1), k = 2, ..., m \end{aligned} \tag{2.1}$$

Under this formulation, incremental motion is horseshoe distributed, defined only at discrete observations of continuous time. Each discrete time increment has its own local shrinkage parameter, λ_k, while the overall variance is controlled by the global shrinkage parameter, τ. Variance between observations continues to scale with elapsed time.

This discrete definition of a stochastic process may pose theoretical challenges, e.g. what is the value of H_{t^*} for $t_{k-1} < t^* < t_k$? For a Gaussian process this is readily obtained, but because of the local shrinkage parameters of the horseshoe process, it is more challenging. H_{t^*} must have its own local shrinkage parameter. The only way to truly resolve this difficulty is by pursuing the Meixner-subordinated process described above, in which we have a continuously generated stochastic process of local shrinkage parameters. However, theoretical and computational development of that model remains intractable.

In the interim, one solution might be to define the value of the local shrinkage parameter to be some fixed value within each increment. This approach is somewhat unsatisfactory, though, because the carried-forward local shrinkage parameter approach blunts the horseshoe process' most unique feature: its abrupt, dynamically changing behavior with the rapidly changing local shrinkage parameters. Rather than define the local shrinkage parameter of H_{t^*} to be some function of the other local shrinkage parameters, we instead use the Bayesian imputation scheme described below, and impute a value for the missing local shrinkage parameter for any unobserved values of the horseshoe process at which predictions are desired. Therefore, the "horseshoe process" is only defined for discrete observations—but a horseshoe process can be generated for *any* set of predefined times.

2.3 Methods

2.3.1 Model Formulation

Let y_i be some outcome observed for patients $i = 1, ..., n$ at continuous predictor value x_i. Define $\mathbf{x}, \mathbf{y}$ as the corresponding length n vectors of these observations. Define $\mathbf{t}$ as a length m vector containing the unique, ordered values of $\mathbf{x}$. Suppose that $x_i = t_j$. Then we define a HPR model as:

$$\begin{aligned} g(E(y_i)) = f_j &= \alpha + \sum_{k=1}^{j} h_k \\ h_k|\tau, \lambda_k &\sim N(0, \tau^2\lambda_k^2(t_k - t_{k-1})),\ k = 2, ..., m,\ h_1 = 0 \\ \tau &\sim C^+(0, c),\ \lambda_k \overset{iid}{\sim} C^+(0, 1),\ k = 2, ..., m \\ \alpha &\sim N(a, b^2) \end{aligned} \tag{2.2}$$

In Equation 2.2, we assume that $\mathbf{f}$, the appropriately-transformed mean trajectory of $\mathbf{y}$ with respect to time, follows a horseshoe process, formulated as the sum of the discretely observed horseshoe increments. This corresponds to placing a horseshoe prior on the first order differences of $\mathbf{f}$ and approximates placing a horseshoe prior on the first derivative of the association. The HPR model assumes that the functional form is like a step function. In general, the λ_k values will be small, resulting in long stretches of near-constant values of $\mathbf{f}$, punctuated by abrupt steps which may be quite large when large values of λ_k are supported by the data. The process starts at a y-intercept α, which has a normal prior placed on it. The $g(E(y_i))$ formulation allows for non-Gaussian data through the use of an appropriate transformation g. We use a logit transformation and Bernoulli likelihood for binary outcomes and a log transformation and Poisson likelihood for count data. The model also allows for multiple observations at the same t_j value and does not require the $\mathbf{t}$ values to be evenly spaced.

2.3.2 Extension 1: Interpolation and Prediction

As is often the case, there may be values of $\mathbf{x}$ at which we wish to obtain predictions or extrapolations, or to obtain a more finely-spaced grid of increments h_k with which to approximate the horseshoe process [85]. In this case, we can augment the grid $\mathbf{t}$ with additional gridpoints that are not included in $\mathbf{x}$ so that $\mathbf{t}$ is the unique, ordered set of observed $\mathbf{x}$ values and augmentation points $\mathbf{x_{aug}}$. Let $\mathbf{y_{obs}}$ denote the observed values of $\mathbf{y}$, and let $\mathbf{f_{obs}}$ denote the estimates of the underlying transformed mean of $\mathbf{y}$ at the observed locations. Define $\mathbf{y_{aug}}, \mathbf{f_{aug}}$ as the unobserved outcome and transformed mean at the requested augmentation points. Let $\boldsymbol{\theta}$ represent the hyperparameters of the model, e.g. $\boldsymbol{\theta} = \{\alpha, \boldsymbol{h}, \boldsymbol{\lambda}, \tau\}$. Then we wish to obtain samples from the posterior distribution

of $\mathbf{f_{obs}}, \mathbf{f_{aug}}, \boldsymbol{\theta}$ [31]. A common approach to do this would be to place a prior on $\mathbf{y_{aug}}$ to reflect the additional uncertainty for the imputed outcomes, e.g. $\mathbf{y_{aug}} \sim \mathbf{N}(\mathbf{f_{aug}}, \sigma^2)$ in the case of Gaussian outcomes. However, in the case where we only care about the underlying mean trajectory $\mathbf{f_{aug}}$, we see that placing a prior on $\mathbf{y_{aug}}$ is unnecessary, because it does not contribute to the posterior of the other parameters:

$$\begin{aligned} P(\mathbf{f_{obs}}, \mathbf{f_{aug}}, \boldsymbol{\theta} | \mathbf{y_{obs}}) &\propto \int P(\mathbf{f_{obs}}, \mathbf{f_{aug}}, \mathbf{y_{obs}}, \mathbf{y_{aug}}, \boldsymbol{\theta}) \mathbf{dy_{aug}} \\ &= \int P(\mathbf{y_{obs}} | \mathbf{f_{obs}}, \boldsymbol{\theta}) \mathbf{P}(\mathbf{y_{aug}} | \mathbf{f_{aug}}, \boldsymbol{\theta}) \mathbf{P}(\mathbf{f_{obs}}, \mathbf{f_{aug}}, \boldsymbol{\theta}) \mathbf{dy_{aug}} \\ &= P(\mathbf{y_{obs}} | \mathbf{f_{obs}}, \boldsymbol{\theta}) \mathbf{P}(\mathbf{f_{obs}}, \mathbf{f_{aug}}, \boldsymbol{\theta}) \end{aligned} \tag{2.3}$$

Then for an unobserved element of $\boldsymbol{t}$, we sample a corresponding value of $\boldsymbol{f}$ according to some prior distribution for $\mathbf{f_{aug}}$, without needing to specify a prior distribution for $\mathbf{y_{aug}}$. We choose to impose the same form on $\mathbf{f_{aug}}$ as on $\mathbf{f_{obs}}$, e.g. an augmentation point $x^* = t_j$ will have $f_j = \alpha + \sum_{k=1}^{j} h_k$ with its own local shrinkage parameter λ_j sampled, where $\lambda_j \sim C^+(0, 1)$. Note that this approach assumes that $\mathbf{f_{aug}}$ and $\mathbf{f_{obs}}$ have a similar level of abrupt change. For example, if at the observed datapoints $\mathbf{f_{obs}}$ appears relatively flat, but at the augmentation points $\mathbf{f_{aug}}$ is abruptly changing, then the above augmentation approach will not be able to interpolate those jumps in the association because there is no evidence of them in the observed data. If data are observed at particular times/locations because they are more or less variable than at unobserved times/locations—for example, some biomarker that is deliberately measured in response to low or high levels of variability—then our approach will produce biased estimates of $\mathbf{f_{aug}}$. However, the estimates of $\mathbf{f_{obs}}$ should be unaffected.

This approach highlights some of the limitations of working with a discrete formulation of a horseshoe process. Every unique value of $\boldsymbol{t}$ must have its own local shrinkage parameter, whether there is an observed outcome at that location to support its estimation or not. In a truly continuous formulation of a horseshoe process, we would have posterior estimates of the parameters of the underlying Meixner process that generated the local shrinkage parameters. Thus—in an ideal world—we might be able to obtain the posterior distribution of $\boldsymbol{\lambda_{aug}} | \boldsymbol{\lambda_{obs}}, \mathbf{y_{obs}}$ from the properties of the multivariate Meixner distribution, and then use these values in combination with properties of the multivariate normal distribution to obtain the distribution of $\mathbf{f_{aug}} | \mathbf{y_{obs}}$, possibly without needing to rerun MCMC sampling [65]. The proposed Bayesian imputation approach, which samples values of $\boldsymbol{\lambda}$ at the augmentation points, may be more computationally burdensome, and we might worry about instability in the presence of large numbers of augmented datapoints or large gaps between them. The simulation study conducted below is meant to address some of these

concerns.

2.3.3 Extension 2: Partial Linear Models

We can extend Equation 2.2 to include multiple predictors using a partial linear model framework. Suppose we have a length p vector of covariates $\mathbf{z_i}$ for each subject, yielding an n x p matrix $\mathbf{Z}$ of covariates. If we assume that these predictors are linearly related to the outcome $\mathbf{y}$, then we can extend the model to be:

$$g(E(y_i)) = f_i = \alpha + \sum_{k=1}^{j} h_k + \boldsymbol{\beta}\mathbf{z_i} \tag{2.4}$$
$$\boldsymbol{\beta} \sim N(\mathbf{0}, \mathbf{d^2})$$

with all other priors unchanged and again assuming that $x_i = t_j$. We now include the parameters $\boldsymbol{\beta}$, a length p vector of coefficients, each of which have a normal prior with mean 0 and variance d_l^2, $l = 1, ..., p$. $\mathbf{Z}$ can contain either continuous or categorical predictors; categorical predictors would need to be converted to a dummy parameterization. We could also consider different priors on $\boldsymbol{\beta}$ such as a Cauchy prior rather than a normal prior if more regularization is desired [30]. It may also be possible to include multiple horseshoe smoother terms in an additive formulation, although this is not something we explore here [48].

2.3.4 Extension 3: Monotonicity Constraints

In some settings we may wish to constrain the HPR to be monotonic, i.e. for $t_j < t_k$, $f_j \leq f_k$ for monotonic increasing. Monotonicity constraints are easily introduced via a transformation of the first derivative to constrain it to be nonnegative (or nonpositive). We propose using the absolute value function as our transformation [8]. Then we modify the HPR to be:

$$g(E(y_i)) = f_j = \alpha + \sum_{k=1}^{j} |h_k| \tag{2.5}$$

All priors are unchanged from before; the only difference is that the horseshoe process sums the transformed increments. As a result, the first derivative is constrained to be nonnegative, and thus the function must be monotonic increasing. (We could similarly take $-|h_k|$ to obtain a monotonic decreasing function.)

2.3.5 Computation

There are several parameters and hyperparameters whose priors merit further discussion. We recommended placing a normal prior with mean a and variance b^2 on α, the y-intercept of the HPR. We recommend setting a to be $\bar{y}$, the sample mean of $\boldsymbol{y}$ and b to be 5 times sd_y the sample standard deviation of $\boldsymbol{y}$. When linear predictors are present, unless there is subject knowledge to further refine prior specifications, we recommend centering and scaling all continuous predictors to have a scale of 1 for continuous outcomes and 0.5 for discrete outcomes, and then setting the prior scale on the linear coefficients to be 5 for continuous outcomes and 2.5 for discrete outcomes [32].

The scale parameter of the global variance, c, is also important. Some may choose to place a hyperprior on c that favors small values, such as an inverse-Gamma or a tightly-constrained half-normal distribution. In our experience, the value of c has little effect on the resulting HPR fit unless data are very sparse. For all models, we set it to be equal to 0.01, which we find yields good performance. However, future work may wish to investigate this further in sparse data.

We implement all models using Hamiltonian Monte Carlo (HMC) via Stan and the cmdstanr package in R. Rather than explore the parameter space through random steps, HMC uses Hamiltonian dynamics to guide movement through the parameter space [5]. As implemented in Stan, HMC is combined with the No-U-Turn-Sampler (NUTS), which assists with HMC's tuning and provides additional diagnostics [38]. To interface Stan with R, we use the cmdstanr package [27].

In complex models like ours, the performance of HMC is affected by the choice of model parameterization. Decentered parameterizations are favored, in which no priors feature other parameters, to reduce *a priori* dependence between parameters. Relationships between parameters are instead obtained through transformations that are performed after parameter sampling steps [59]. We use decentered parameterizations for all parameters in the model. Thus, the full list of parameters and priors which we sample is:

- $\alpha \sim N(\bar{y}, (5sd_y)^2)$
- $\tau = c\tau_1\sqrt{\tau_2}$, where $\tau_1 \sim N(0,1)$ and $\tau_2 \sim InvGamma(\frac{1}{2}, \frac{1}{2})$. Note that this corresponds to $\tau \sim C^+(0, c)$.
- Independently for $k = 2, ..., m$, $\lambda_k = \lambda_{1k}\sqrt{\lambda_{2k}}$, where $\lambda_{1k} \sim N(0,1)$ and $\lambda_{2k} \sim InvGamma(\frac{1}{2}, \frac{1}{2})$. Note that this corresponds to $\lambda_k \sim C^+(0,1)$.
- For $k = 2, ..., m$, $h_k = \gamma_k\tau\lambda_k\sqrt{(t_k - t_{k-1})}$, where $\gamma_k \sim N(0,1)$. Note that this corresponds to $h_k|\tau, \lambda_k \sim N(0, \tau^2\lambda_k^2(t_k - t_{k-1}))$.
- In the case of Gaussian outcome data: $\sigma = \sigma_1\sqrt{\sigma_2}$, where $\sigma_1 \sim N(0, s^2)$ and $\sigma_2 \sim$

$InvGamma(\frac{1}{2}, \frac{1}{2})$. Note that this corresponds to $\sigma \sim C^+(0, s)$. In the simulations below, we use $s = 5$.

- In the case of linear covariates: a length p vector $\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{d}^2)$.

After making necessary parameter transformations, we connect these parameters to the data using an appropriate likelihood (Gaussian, Bernoulli, or Poisson).

For HMC sampling, we use 4 chains, each with a warm-up phase of 1000 samples and a sampling phase of 2000 samples, without thinning. This yields a total of 8000 samples from the posterior distribution. Despite the efforts described above to improve computational tractability, it is still common to observe occasional HMC divergences, which usually indicate poor posterior exploration. We present evidence that these sporadic divergences do not seem to harm model performance, although high numbers of divergences should still be cause for concern [5].

All of the methods described above are implemented in the R package HPR, hosted on GitHub [64]. The package is used to conduct the simulation studies described below and the data application.

2.4 Simulation Study

2.4.1 Horseshoe Process Regression

We considered four true underlying associations, each of which were observed at an equally spaced grid of n = 100 observations:

1. bigstep: $f(x) = 0 * I(x \leq 2) + 6 * I(2 < x \leq 5) + 1 * I(5 < x \leq 6) + 3 * I(6 < x \leq 8) + 10 * I(x > 8)$

2. joinpoint: $f(x) = (1.5x) * I(x < 2) + (16 - 5x) * I(2 \leq x < 3) + 1 * I(3 \leq x < 6) + (10 - x) * I(6 \leq x < 9) + (5x - 44) * (x \geq 9)$

3. impulse: $f(x) = 0 * I(x = 0) + \exp(-x) * I(0 < x < 3) + 1 * I(x = 3) + \exp(-(x-3)) * I(3 < x < 7) + \exp(-(x - 7)) * I(x = 7)$

4. bounce: $f(x) = |sin(x)|$

$I()$ denotes the indicator function; i.e. $I(x) = 1$ if condition x is true, and $I(x) = 0$ otherwise. For the Gaussian outcomes, we simulated measurement error with $\sigma = 0.5$ for the bigstep and joinpoint scenarios, $\sigma = 0.2$ for the bounce scenario, and $\sigma = 0.1$ for the impulse scenario. Plots of the true underlying curves along with a sample dataset are given in Figure 2.1 for continuous

outcomes. Based on the underlying model, we expected HPR to excel at the bigstep scenario. Joinpoint and impulse both featured abrupt changes, but not of HPR's assumed form, so it was less clear how HPR would perform in these settings. We anticipated weak performance in the bounce scenario, which was primarily smooth and without abrupt changes, but we included it to provide insight into HPR's limitations. We considered a number of comparison methods. First, we compared to Gaussian process regression (GPR), a smooth approach which places a Gaussian process prior on the function fit, using the R package mgcv [82]. Second, we compared to an adaptive spline model (Adspline), which uses the set of unique x values as knot locations for a set of P-splines, which are adaptively penalized [82]. Third, we compared to the running median filter (MedFilt) as implemented in the package FBN, which estimates the function at x as the median of some number of observations neighboring x. We set a window-size of 3, meaning that the function at x is estimated as the median of $y(x)$ and its immediate neighbors on either side [49]. Fourth, we compared to the first-order penalized trend filter (TrendFilt) of Tibshirani et al. from the package genlasso [75], which penalizes the first order differences of the function, and sets the penalty parameter using generalized cross validation. Note that MedFilt and TrendFilt do not provide uncertainty quantification and thus their performance on metrics like credible interval coverage and width is not presented.

We assessed performance with three metrics:

1. Mean absolute difference (MAD): $\frac{1}{n}\sum_{i=1}^{n}|f(x_i) - \hat{f}(x_i)|$, where $\hat{f}(x_i)$ is the predicted function's value at x_i and $f(x_i)$ is the true function's value at x_i.

2. Credible/confidence interval width (Width): $\frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_i)^{0.975} - \hat{f}(x_i)^{0.025}$, where $\hat{f}(x_i)^{0.975}$ denotes the upper bound of a 95% credible/confidence interval for $\hat{f}(x_i)$ and $\hat{f}(x_i)^{0.025}$ is the lower bound.

3. Credible/confidence interval coverage (Coverage): $\frac{1}{n}\sum_{i=1}^{n} I(\hat{f}(x_i)^{0.025} \leq f(x_i) \leq \hat{f}(x_i)^{0.975})$.

Note that these three metrics were summed across all of the observed datapoints. We present additional results on pointwise performance in Appendix B. We assessed performance on each metric across the 100 replicates of each of our 4 data-generating scenarios for each method. All simulations were implemented in R, using tools from the Dynamic Statistical Comparisons (DSC) framework [64, 28]. All code used to completely reproduce the simulations can be found on GitHub.

Results for continuous outcomes are given in Figure 2.2. As we can see, HPR performed quite well, with the smallest MAD of any of the comparison methods for the bigstep, impulse, and joinpoint scenarios. As expected, for the bounce scenario, it was surpassed by the methods that are

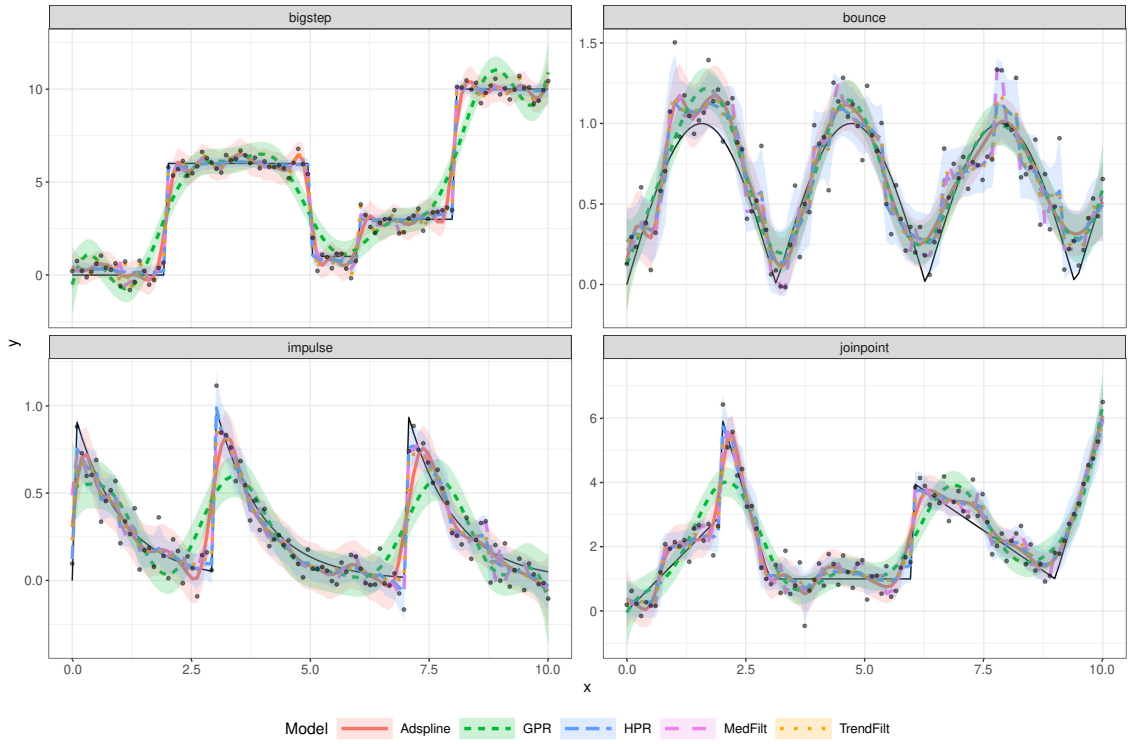

Figure 2.1: Point estimates and 95% credible/confidence intervals of horseshoe process regression (HPR) and comparison methods on sample datasets from four data-generating scenarios for continuous outcomes, each with n = 100. Comparison methods are an adaptive spline model (Adspline), a Gaussian process regression (GPR), a median filter (MedFilt), and a penalized trend filter (TrendFilt).

better-suited to smooth associations, although it provided better results than the other non-smooth methods (TrendFilt, MedFilt). Its credible interval width was slightly wider than the width of the intervals returned by the other methods that give uncertainty estimates, with the exception of the bigstep and impulse scenarios, for which HPR returned a substantially narrower credible interval across the domain than the other comparison methods. Coverage was generally good, with rates above 95% for all of the scenarios. When we examine the pointwise plots in Appendix B (Figure B.1), we see that the pointwise performance resembled that averaged over the curve. In particular, the comparison methods particularly struggled at the jumps in bigstep, joinpoint, and impulse, as we expected, with bias spiking at the jump points while credible interval coverage dropped.

Additional results are given in Appendices A-C for count outcomes (Figures A.1, A.2, and B.2), binary outcomes (Figures A.3, A.4, and B.3), and monotonicity constraints (Figures C.1, C.2, C.3). Performance was generally similar to that of continuous data. HPR continued to offer the best performance of all comparison methods for fitting step functions. For binary outcomes, though, GPR offered better performance than HPR for the joinpoint scenario in terms of MAD (Figure

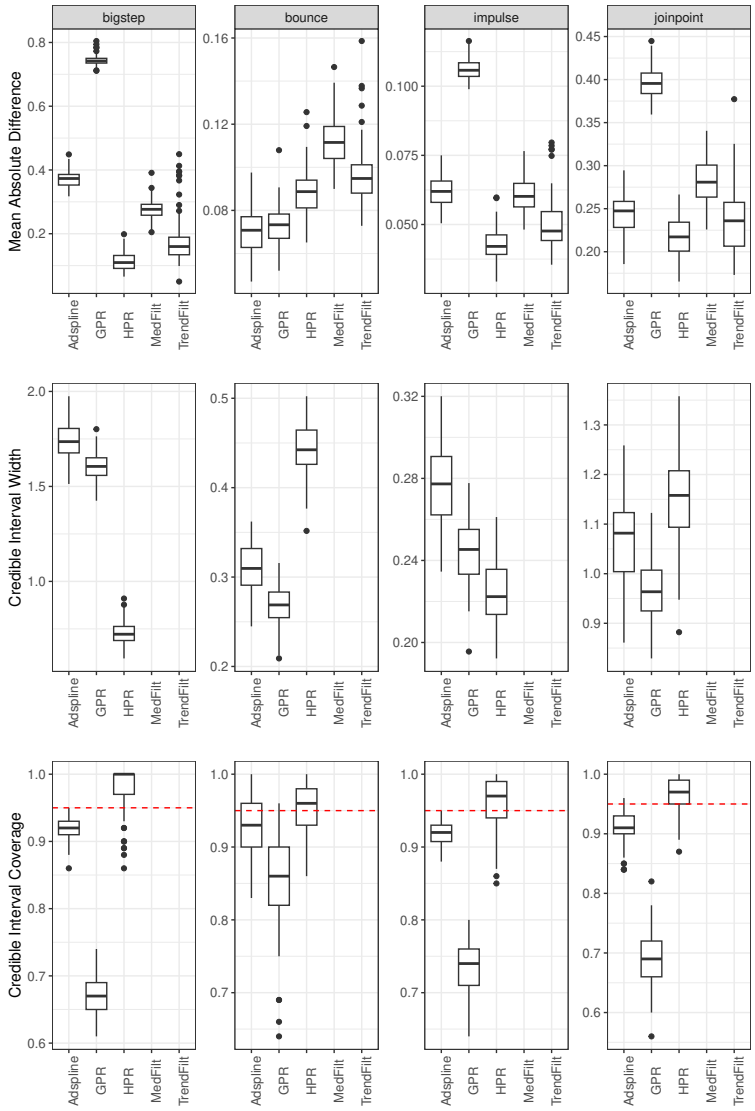

Figure 2.2: Horseshoe process regression (HPR) simulation results for continuous outcomes, based on 100 replicates on four data-generating scenarios, each with n = 100. Comparison methods are an adaptive spline model (Adspline), a Gaussian process regression (GPR), a median filter (MedFilt), and the penalized trend filter (TrendFilt). The top row gives performance for mean absolute difference (smaller is better); the second row gives performance for credible/confidence interval width; the third row gives performance for credible/confidence interval coverage (0.95 is nominal and given as a horizontal red dashed line). Each column is for one data-generating scenario; sample datasets for each scenario are shown in Figure 2.1. Note that interval coverage/width is not given for MedFilt and TrendFilt because these methods do not provide uncertainty estimation.

A.4). Coverage was nominal except for the bigstep scenario, for which it was less than nominal for binary and count outcomes—although still closer to nominal than the other comparators.

2.4.2 Data Interpolation and Prediction

Here, we focus on our data interpolation scheme, to assess whether it performs sensibly at varying grid densities. We only considered scenarios bigstep and bounce, and we restricted our focus to HPR, as the data interpolation performance of the comparison methods from section 2.4.1 have been assessed elsewhere [75, 82]. We randomly sampled 100 unevenly spaced datapoints between 0 and 10 to be our observed x locations. Then, we fit the HPR either 1) only using the observed datapoints, 2) augmented by a grid of datapoints at every 0.5 (21 augmented datapoints), and 3) augmented by a grid of datapoints at every 0.1 (101 augmented datapoints). Note that for some replicates the augmented datapoints will be extrapolations, because we did not require the inclusion of 0 and 10 in our randomly sampled observed data locations. We calculated the performance metrics above separately for the observed datapoints and the augmented datapoints, to see if predictions at the observed datapoints changed depending on the number of gridpoints, and if predictions at the augmented datapoints were fairly accurate.

Results for continuous outcomes are given in Figure 2.3; results for count and binary outcomes are given in Appendix D (Figures D.1 and D.2). Overall, we see that performance of the augmentation scheme was good, with MAD holding fairly constant, as expected, across observed and augmented datapoints regardless of grid density. MAD was slightly worse at augmented points relative to observed datapoints, and credible interval widths were wider at augmented datapoints than at observed datapoints, as we expected. Both MAD and credible interval width appeared somewhat improved with a larger number of augmentation points. This "improved performance" is misleading, because in the data generating schemes considered here—which do not feature an extremely large number of abrupt changes—the probability that an augmentation point is placed at the location of an abrupt jump is reduced in the presence of more augmentation points, artificially boosting aggregate performance. Credible interval coverage held fairly constant across grid density, with rates at or above 95% for all scenarios. Coverage rate was similar across observed and augmented datapoints. Performance was generally similar for binomial and Poisson outcomes.

2.4.3 Partial Linear Models

We conducted simulations to assess the performance of the HPR partial linear model. Full simulation set-up details and results are given in Appendix E. Performance of the partial linear model was generally good. HPR offered substantially reduced MAD and credible interval width for the latent mean $\hat{E}(y_i)$ when the nonlinear component of the partial linear model was a step function. When the nonlinear component was a smooth function, HPR's performance was worse than the comparison methods for continuous outcomes, although credible interval coverage was still very good (Figure E.1). For binary and count outcomes, HPR consistently surpassed the comparison

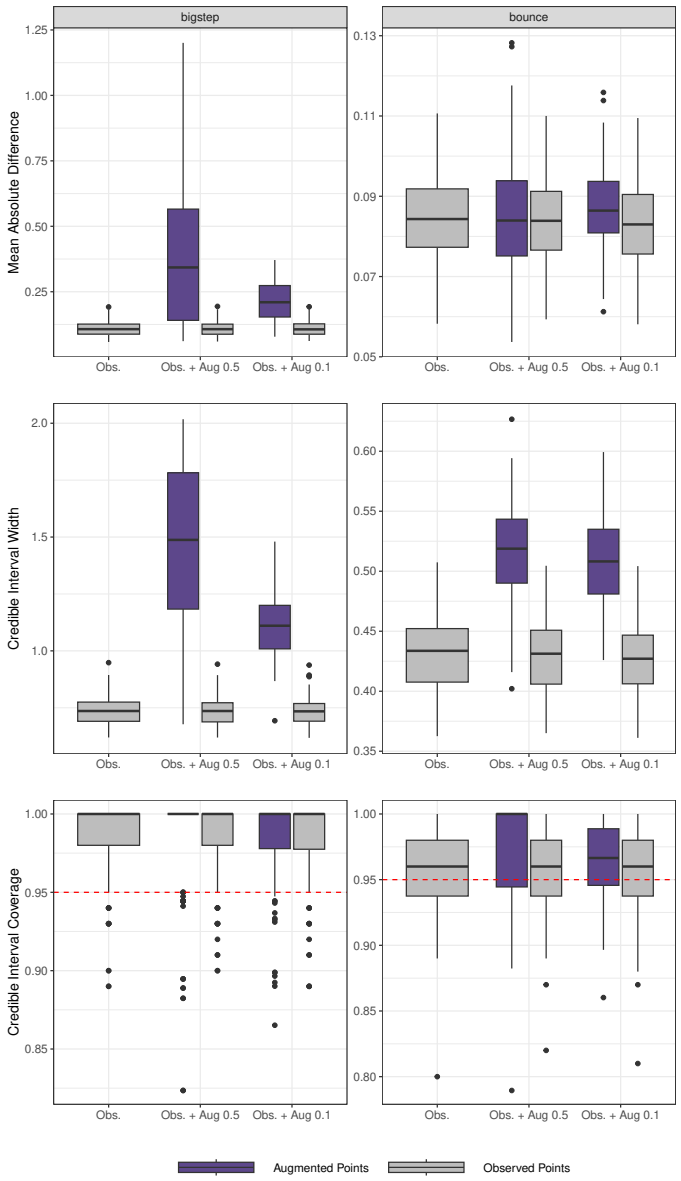

Figure 2.3: Horseshoe process regression (HPR) data augmentation simulation results for continuous outcomes, based on 100 replicates on two data-generating scenarios. We compared a HPR calculated only at n = 100 observed points to a HPR with augmentation points at a grid of every 0.5 and a HPR with augmentation points at a grid of every 0.1 (from 0 to 10). The top row gives performance for mean absolute difference calculated at both the observed and augmented points (smaller is better); the second row gives performance for credible interval width calculated at both the observed and augmented points; the third row gives credible interval coverage calculated at both the observed and augmented points (0.95 is nominal and marked as a horizontal red dashed line). Performance at observed points and augmented points are displayed separately. Each column is for one data-generating scenario.

methods, even when the nonlinear component of the partial linear model was a smooth function (Figures E.3 and E.5). The GPR particularly struggled for count outcomes (Figure E.3). Regardless of the form of the nonlinear component, performance for estimating the linear effects of the partial linear model was good (Figures E.2, E.4, and E.6).

2.4.4 Computational Assessment and Sensitivity Analyses

We provide some insight into HPR's computational performance, with results given in Appendix F (Figures F.1, F.2, F.3). Almost all of the models fitted in the simulation studies featured at least some HMC divergences. In most cases less than 5% of samples ended in a divergence. Max treedepth warnings occurred rarely. $\hat{R}$ diagnostics and effective sample sizes generally seemed adequate [77]. Although slow compared to non-Bayesian methods, computation time was generally quite reasonable, with most models finishing in less than 5 minutes. This can be made faster with parallelization, which is available in our R package. Models took longer to run as the sample size and the amount of augmentation data increased, as we would expect. For more information on these diagnostics, please see the Stan reference manual [72].

We also explored the role of sample size and prior specification in model estimation, with full results given in Appendix G (Figures G.1 and G.2). We focused these sensitivity analyses on the bigstep scenario described above, because it is HPR's recommended setting. In addition to the sample size of $n = 100$ that we used above, we also considered $n = 30$ and $n = 500$. We considered several different settings for the hyperparameters of the model (the prior mean and variance for the y-intercept α, the prior scale c on the global shrinkage parameter τ, and the prior scale s on the measurement error σ). Findings were generally stable across hyperparameter values, although at smaller sample sizes ($n = 30$), findings were more affected by hyperparameter choices. Poor choices for the prior variance on α—particularly setting it too small—negatively affected model fit. The choice of c also affected findings at small sample sizes, especially for binary outcomes. Model estimation improved with larger sample sizes, although estimation was still adequate at the $n = 30$ sample size.

2.5 Application

We use HPR to fit the trajectories of women's BBT over the menstrual cycle. BBT follows a reliable pattern in healthy, premenopausal women who are not taking hormonal birth control. Each menstrual cycle starts with the onset of the period, at which time BBT is low. With ovulation (usually around day 14 of the menstrual cycle), a woman's BBT spikes, sometimes by a full degree Fahrenheit, and will remain high until the onset of the next period, when it will drop suddenly

to the pre-ovulation temperature and the cycle will repeat. If the woman conceives a pregnancy that cycle, her BBT will instead stay at the post-ovulation high temperature for the duration of her pregnancy, sometimes even increasing in a second spike around a week after conception when the embryo implants in the uterine lining.

Although this general pattern is consistent across healthy women (sustained low temperature, spike at ovulation, sustained high temperature, drop with onset of period, repeat unless pregnant), the details vary considerably between women and even within a single woman. The date of ovulation may be as early as day 9 or as late as day 30, depending on factors like stress and other medical conditions, with the full menstrual cycle length varying from 20 to 40+ days. Some women exhibit a more gradual increase/decrease in BBT over the menstrual cycle, rather than sharp jumps and drops. However, by tracking BBT over several menstrual cycles, patterns may emerge that suggest underlying health conditions or provide guidance on how best to time sexual intercourse to improve or reduce chances of pregnancy. When used in combination with other health and fertility indicators, BBT plots can provide insight into women's health.

Here, we model BBT trajectories for several women, using data abstracted from example charts given in Toni Weschler's *Taking Charge of Your Fertility* [80]. Weschler presents example plots and gives her hypothesis for the day of ovulation for each plot, based on BBT and other predictors (cervical mucus and position, pregnancy testing information, etc.). We use Weschler's hypothesis as a best guess for the correct answer, and thus explore HPR's ability to match the results obtained from careful examination by a trained expert. In all cases, day 1 of the menstrual cycle is the day the period started, and the last observation corresponds to the day of the start of the next period (except in the plots of pregnant women, which we note). We indicate Weschler's hypothesis on each plot. Our outcome is BBT and our predictor is day of the menstrual cycle. We fit a horseshoe process regression (HPR), and for comparison, a Gaussian process regression (GPR), an adaptive spline model (Adspline), a penalized first-order trend filter (TrendFilt), and a median filter with a window size of 3 (MedFilt). Note that there is no information sharing across women; models are refit separately for each woman's BBT trajectory.

In Figure 2.4, we present the plots for four women who did not conceive a pregnancy that cycle. In panel A, we see a fairly normal pattern and observe that HPR and TrendFilt both capture the ovulation jump cleanly (although only HPR is able to provide uncertainty estimates). Note that ovulation occurred later than we would expect, at day 24 (with increased temperatures starting on day 25). GPR, by comparison, oscillates substantially in the pre-ovulation phase and oversmoothes the ovulation jump, starting it almost 3 days early. Adspline manages to smooth the pre-ovulation phase, but not the post-ovulation phase, and it also starts the ovulation jump 2 days early. In panel B, we observe no clear pattern—this subject actually did not ovulate that cycle,

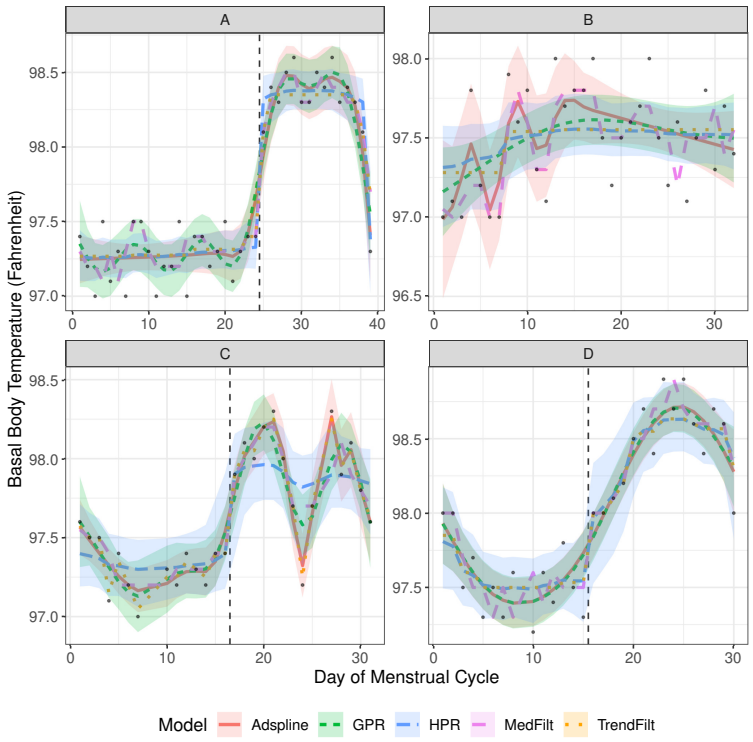

Figure 2.4: Fitted basal body temperature (BBT) trajectory and 95% credible/confidence intervals from a horseshoe process regression (HPR), adaptive spline model (Adspline), Gaussian process regression (GPR), median filter (MedFilt), and penalized trend filter (TrendFilt) for four women who did not conceive pregnancy. Observed datapoints are given as black dots, with an expert's guess of the ovulation time given by the vertical black dashed line. Note that there is no information sharing across women; models are refit separately for each woman's BBT trajectory.

which is why there is no BBT shift. HPR correctly does not detect one. In panel C, we see another fairly typical example, this one with a slight temperature drop post-ovulation around day 24, which occurs in some women and is normal. Of the three methods, HPR does the best job of capturing the ovulation jump at day 16 without being overly swayed by the minor temperature dip at day 24, while the other methods oscillate heavily in response. Finally, in panel D we see a plot from a woman with weak thermal shift, in that her BBT increases less abruptly with ovulation, making it challenging to identify the date of ovulation. Nonetheless, HPR and TrendFilt both start the jump at day 16, correctly identifying the date of ovulation as day 15. While MedFilt places the jump correctly, it introduces excess motion in the pre- and post-ovulatory portions of the cycle. GPR and Adspline over-smooth, placing the beginning of temperature increase around day 12.

In Figure 2.5, we present two plots for women who conceived that cycle. In panel A, we see a successful pregnancy which features a second temperature shift at implantation. HPR and TrendFilt identify ovulation at day 15, and then somewhat capture the slightly higher BBT post-

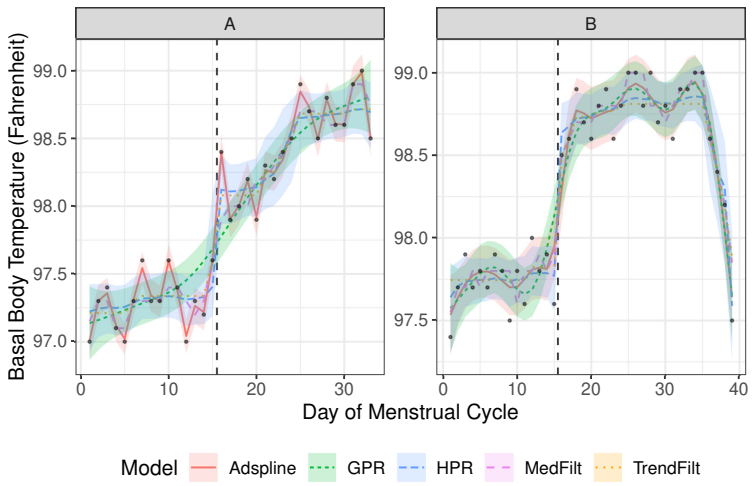

Figure 2.5: Fitted basal body temperature (BBT) trajectory and 95% credible/confidence intervals from a horseshoe process regression (HPR), adaptive spline model (Adspline), Gaussian process regression (GPR), median filter (MedFilt), and penalized trend filter (TrendFilt) for two women who conceived a pregnancy. Observed datapoints are given as black dots, with an expert's guess of the ovulation time given by the vertical black dashed line. Note that there is no information sharing across women; models are refit separately for each woman's BBT trajectory.

implantation, starting around day 23. GPR would completely smooth over the BBT trajectory for the pregnancy cycle, missing the dates of ovulation and implantation entirely, while Adspline and MedFilt overfit, making the fit difficult to interpret. In panel B, we show the plot of a pregnancy that ended in a miscarriage. HPR identifies the ovulation jump at day 15. The sustained high temperatures (lasting almost 20 days) indicate conception and implantation of an embryo. However, the dropping temperatures from day 36 are an early warning sign, resulting in a miscarriage on day 39.

Finally, we give an example of the use of covariates in HPR. We restrict our focus to HPR because the other comparison methods either cannot include linear covariates (MedFilt, TrendFilt) or do not return easily interpretable nonlinear components in the presence of linear covariates (Adspline, GPR). This woman was sick from days 8-10 of her menstrual cycle, with a high fever. By including illness as an additional categorical linear covariate in the HPR, we can adjust away its effect. That results in the estimated BBT trajectory given in Figure 2.6, in which ovulation occurs on day 21, with the illness estimated to increase BBT by 2.88 (2.44, 3.27) degrees Fahrenheit.

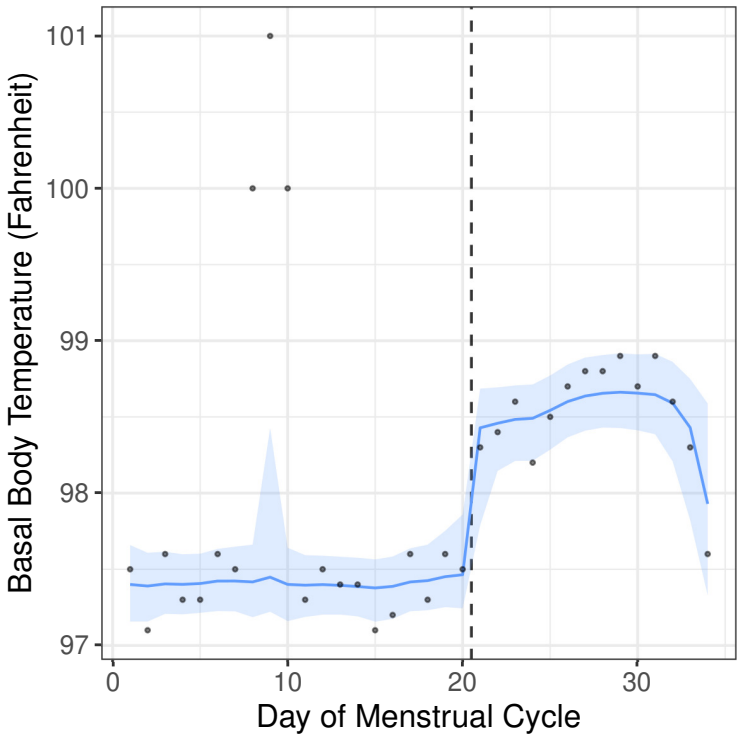

Figure 2.6: Fitted basal body temperature trajectory and 95% credible/confidence intervals from a horseshoe process regression (HPR) adjusted for the presence of fever for a woman who was ill during days 8-10 of her menstrual cycle. Observed datapoints are given as black dots, with an expert's guess of the ovulation time given by the vertical black dashed line.

2.6 Discussion

We present horseshoe process regression (HPR), a method for fitting functions that feature local changes in variability. HPR contributes to a growing literature on the use of local-global shrinkage families within a stochastic process framework [22, 48, 63]. Here, we focus on the details of implementation, filling in some of the gaps of an overarching theoretical framework that has not been fully translated to applied use. We find that HPR outperforms other existing methods for fitting step functions and other associations with abrupt changes, although we would not recommend using it to fit an association that is expected to be smooth. Other methods like Gaussian process regression would likely yield better results. We extend HPR to allow for additional linear covariates, data interpolation, and monotonicity constraints. We find that our data interpolation scheme yields good results and that HPR's superior performance for fitting step functions persists in the presence of additional linear covariates, even when correlated with the nonlinear predictor. Together, these extensions make HPR more usable for applied research on nonlinear associations with local changes in variance in small samples. HPR is available as an R package on GitHub.

Computational burden is an unresolved challenge in our implementation of HPR. We use Stan and cmdstanr, which yields good performance in our context and gives reasonable computing times and decent convergence diagnostics in a range of data-generating scenarios. However, computing time becomes prohibitive at larger sample sizes and is a limitation. Future work may wish to consider a Gibbs sampling type approach like that of Kowal et al. (2019) [48] or other approaches for scalable Bayesian inference like variational Bayes [79]. Even with extensive tuning and debugging, we were unable to fully eliminate HMC divergences from model implementation, a problem that seems common to the horseshoe prior [22, 62]. We see no evidence that these sporadic HMC divergences affect performance, although it is something of which to be aware. In general, if other convergence diagnostics (effective samples, Gelman-Rubin convergence diagnostic, max treedepth warnings) are acceptable, it seems mostly safe to ignore sporadic divergences occuring in $< 5\%$ of samples. If HMC divergences are more frequent than that or combined with other evidence of nonconvergence, we would recommend further tuning of the model or the use of a different method. Future work may also wish to examine whether the regularized horseshoe prior within HPR resolves some of these issues without sacrificing model performance [62].

Our data interpolation scheme motivated new questions about the theory and implementation of horseshoe process prior models. To date, all implementations of horseshoe process prior models have chosen to use a discrete formulation, as we do here [22, 48]. Although this simplifies mathematical derivation and computation, it requires imputation or approximations to conduct data interpolation and augmentation. We chose to use a Bayesian imputation scheme. In an ideal world, it would be possible to develop a horseshoe process in a truly continuous formulation and leverage properties of the multivariate Meixner distribution to calculate predictions and interpolations without needing to rerun the MCMC. However, current mathematical theory does not make that possible. We attempted several *ad hoc*, stop-gap measures that allowed us to impute new predictions without rerunning the MCMC, such as imputing the local shrinkage parameter for the new predictor as the mean of its two nearest local shrinkage parameter neighbors, and then using the imputed local shrinkage parameter within Gaussian kriging equations to generate a prediction [65]. We did not have success with these approximations and thus chose the more statistically rigorous Bayesian imputation approach we outline here. Future work may wish to investigate this and other approximations for horseshoe process prediction and kriging further.

We have demonstrated the utility of HPR and shown that it is usable for a variety of real-world data. We find that HPR is an excellent choice for data featuring local changes in variance, such as step functions or piecewise linear functions. Future work will further elucidate HPR's strengths and weaknesses and provide new insights into computationally efficient and stable implementations and its underlying theory.

CHAPTER 3

A Variational Inference Implementation of Horseshoe Process Regression to Model Basal Body Temperature Data

3.1 Introduction

The menstrual cycle plays a major role in the health of women of reproductive age. In addition to its obvious relationship to fertility, the menstrual cycle may also affect women's body weight, body temperature, resting heart rate, and mood [60]; metabolic rate [71]; physical performance [37]; and susceptibility to infection [42]—to give a few examples. As awareness of the menstrual cycle's importance has increased, clinicians, health researchers, and women themselves have become interested in monitoring biomarkers and symptoms throughout the menstrual cycle, in order to better understand patterns and what they suggest about underlying health.

As a result, there has been a massive increase in the use of menstrual tracking technology, with a 2019 survey finding that 1 in 3 American women have used a menstrual tracking app at least once in their life [26]. In these apps, women can log their menstrual periods and other biomarkers, such as weight, body temperature, cervical mucus, the results of ovulation tests, and symptoms like nausea, abdominal cramping, or headaches. Women's goals in using these apps are varied. Common goals include predicting the onset of the next menstrual period, predicting ovulation before it occurs (to conceive pregnancy), detecting ovulation after it has occurred (to avoid pregnancy), and monitoring patterns across the menstrual cycle to promote self-awareness and to share with health care providers [21].

In Chapter 2, we developed horseshoe process regression (HPR), a method for modeling abruptly changing data. HPR showed potential for modeling basal body temperature (BBT), a key biomarker of the menstrual cycle. In women of reproductive age not taking hormonal birth control, BBT is low during the first half of the menstrual cycle, jumps sharply immediately after

ovulation, and remains high for the second half of the menstrual cycle, dropping with the onset of the next menstrual period. Researchers have tried to use this biphasic property of BBT to detect and predict ovulation for more than a century [53]. HPR was well-suited to modeling BBT because it was able to capture the sharp changes that BBT exhibits at the time of ovulation and menstruation.

However, HPR was developed in the general setting of abruptly changing data, rather than BBT specifically, and lacked features that would be necessary in the context of menstrual tracking. Although HPR was able to capture jumps in BBT, it had no way to identify the location of jumps. It did not make use of the wealth of prior clinical knowledge about BBT and the menstrual cycle; nor could it share information across cycles. HPR was able to provide predictions of BBT at future timepoints, but these predictions did not use information from previous cycles or prior information about BBT patterns, and they only provided a predicted temperature measurement—not a predicted day of ovulation/menstruation. Finally, the Hamiltonian Monte Carlo (HMC) approach used to estimate HPR meant that computation time was a challenge, with even a single cycle's worth of data taking 10 to 20 seconds to fit.

Here, we modify HPR to provide this missing functionality. To do so, we develop a variational inference (VI) implementation of HPR, which drastically decreases computation time [79]. We explicitly incorporate ovulation day into the HPR model, and we tailor the priors to reflect clinical knowledge on BBT across the menstrual cycle. Drawing on ideas from the sequential Monte Carlo literature, we propose a posterior-prior passing scheme to share information across cycles while keeping computation times swift [18]. With these new features in place, we demonstrate how this BBT-specific version of HPR (HPR-BBT) can be used to detect ovulation. Taken together, these adaptations to HPR make it a powerful approach for modeling BBT to monitor menstrual health.

HPR adds to a rich literature on menstrual cycle prediction and modeling (which some researchers date back to 20,000 BCE [9]). Despite the extensive work done in this space, HPR still has novel strengths. Händel and Wahlström (2019) reviewed existing methods for using BBT to identify time of ovulation [34]. Many of the approaches were heuristic in nature (e.g. ovulation occurred on the day that was followed by 6 higher temperature measurements). Other early approaches relied on linear regression or Gaussian process regression, both of which are ill-suited to the biphasic BBT setting and did not yield reliable results in many women.

Händel and Wahlström (2019) found that the two best-performing approaches were hidden Markov models, followed by the cumulative sum test of Royston and Abrams (1980) [34, 66]. The cumulative sum test is a testing procedure to identify the date of ovulation after-the-fact. Although

it incorporates a variety of clinical knowledge on BBT and the menstrual cycle—and its review of these topics is an invaluable resource—it has some shortcomings. It cannot be used for prediction or to model the temperature trajectory in its entirety, which makes it difficult to use in combination with other predictors or within a larger model. In addition, it does not provide uncertainty estimates and cannot share information across cycles. A hidden Markov model (HMM) is more flexible and has been used to detect and predict ovulation from BBT with good performance, as in Chen et al. (2009) and Luo et al. (2020) [13, 53]. However, HMMs do not provide uncertainty estimates around the temperature trajectory or day of ovulation—which we think is desirable. Furthermore, although HMMs make it possible to share information across cycles, in the applications we have seen [13, 53], this information-sharing requires the assumption that the pre-ovulation temperatures are similar across cycles (and similarly for the post-ovulation temperatures). This assumption is contradicted by the data [66], forcing users to choose between discarding previous cycles' data or accepting an assumption that may yield poor results for particular cycles. Finally, although some attempt has been made to incorporate clinical knowledge into the HMM, such as making it less likely for ovulation to occur in the first 7 days of the menstrual cycle, this incorporation has been fairly limited in scope [53].

The structured nature of BBT across the menstrual cycle—and the amount of clinical information we have about it—may warrant a Bayesian approach. Some work has been done in this direction, such as by Scarpa and Dunson (2009) [69]. Although their model overcomes many of the shortcomings of the cumulative sum test and HMMs, it places a fairly smooth prior on BBT, and thus does not capture the BBT jump as sharply as HPR and HMMs. HPR could offer improvements over this and other Bayesian methods as a result. Computation time is also likely a challenge for theirs and other Bayesian methods, much as it was for HPR in Chapter 2. In this chapter, we solve this problem using variational inference (VI). With VI, we seek a "good-enough" approximation of the posterior distribution. Critically, this approximate posterior is easier to estimate than the true posterior, enabling faster computation without the use of MCMC. Although the results from VI are rarely identical to those from MCMC, they may still be very good, and the computational speed gains may justify the slight worsening in posterior estimation—provided that consistently decent performance of VI can be demonstrated [31, 79, 83].

HPR provides a close match to the underlying data-generating model for BBT over the menstrual cycle. The Bayesian approach makes it straightforward to incorporate prior clinical knowledge, obtain uncertainty estimates, combine the model with other covariates or larger models of the menstrual cycle, and modify model assumptions to accommodate different women's menstrual features or menstrual tracking priorities. With VI, we are able to resolve the computational challenges and thus make HPR usable in real application. We look forward to future work to further improve

HPR for this purpose and develop other promising Bayesian methods for menstrual health.

The remainder of the paper is as follows. First, we review some necessary background on BBT across the menstrual cycle, HPR, and VI. We then present the VI implementation of HPR and compare its performance to HMC. We describe the modifications to HPR to explicitly address the BBT setting and show how these are accommodated within the VI algorithm. We use the updated HPR-BBT model to analyze BBT trajectories from a large cohort of British women, collected by a Catholic charity between 1960-1980 [55, 76]. We close with a discussion of limitations and directions for future work.

3.2 Background

3.2.1 Basal body temperature

As was discussed above, basal body temperature (BBT) follows a biphasic pattern over the menstrual cycle. The start of the menstrual period is denoted as day 1 of the menstrual cycle. This is also the start of the follicular phase, which is the pre-ovulation portion of the menstrual cycle, characterized by lower BBT. In most women, the follicular phase is a median of 16 days long (mode 15) [23]. The timing of ovulation is sensitive to external stressors and thus the length of the follicular phase can vary substantially, even within the same woman. A third of women are estimated to regularly have their follicular phase vary by more than 7 days in length (e.g. ovulation on day 12 of one cycle and on day 20 of another cycle). However, 95% of follicular phases will be 10-22 days long [23]. The follicular phase ends with ovulation, at which point the BBT increases; the World Health Organization has recommended a minimum difference of 0.2 degrees Celsius between the pre- and post-ovulation BBT as a threshold for ovulation [1], although in many women it will be a 0.3-0.5 degree difference [66, 73]. It is estimated that 90% of women between the ages of 25 and 45 will exhibit this biphasic BBT pattern in most cycles [73]. Although the difference in pre- and post-ovulation BBT is fairly standard, the mean BBT may differ across cycles, even within the same woman. However, within one woman, the standard deviation of BBT measurements is usually similar pre- and post-ovulation and across cycles [66]. After ovulation, the luteal phase begins and continues until the end of the menstrual cycle (with the start of the next period). The luteal phase is a median of 13 days long (mode 13) and is less variable in length than the follicular phase, with 95% of luteal phases lasting between 9-16 days. Less than 10% of women will have their luteal phase length vary by more than 7 days across cycles [23]. These menstrual cycle features are summarized in the schematic given in Figure 3.1.

Follicular phase length is the primary driver of the total length of the menstrual cycle. A

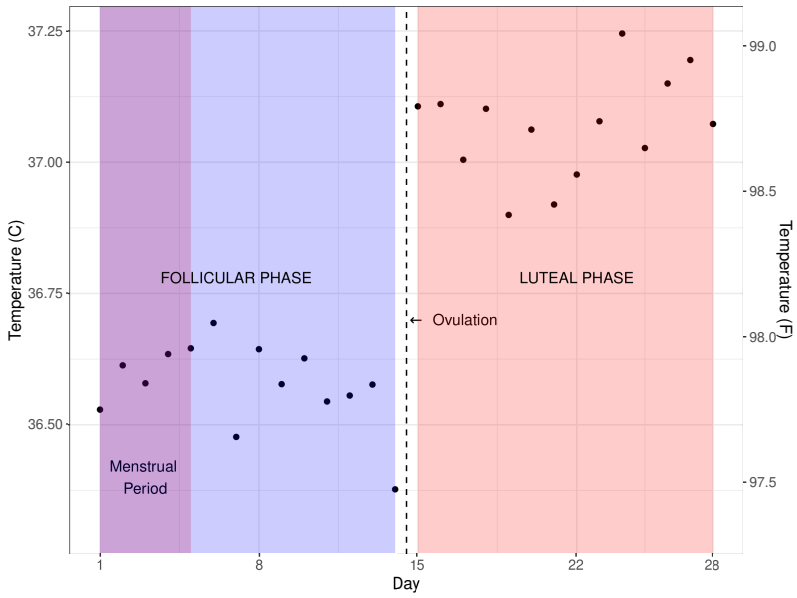

Figure 3.1: Paradigmatic menstrual cycle basal body temperature (BBT) data.

woman's age and body mass index (BMI) have both been linked with follicular phase length, while luteal phase length is generally stable across age (and, to a lesser extent, across BMI) [11, 23]. Older women have slightly shorter follicular phases (and by extension, shorter menstrual cycles) [11, 73]; women with higher BMI have more variable follicular phase/menstrual cycle lengths [11]. Very young (< 25 years) and older (> 45 years) women also had more variable cycle lengths with less clear BBT shifts [73].

Follicular and luteal phase length are both of clinical interest and may affect fertility. Generally, more attention is paid to follicular phase length, which corresponds to the timing of ovulation. Knowing when a woman is likely to ovulate enables her to time sexual intercourse to maximize/minimize her chances of conceiving pregnancy. However, luteal phase length is also of clinical interest. Women with very short luteal phases (< 10 days) may have difficulty sustaining pregnancy, as the fertilized embryo does not have time to implant in the uterine lining, a condition called luteal phase defect. Depending on the underlying cause of the luteal phase defect, there may be treatments to correct it once identified [58]. A very long luteal phase for a single cycle can be early evidence of pregnancy; if luteal phases are routinely long, that may suggest polycystic ovarian syndrome (PCOS) or some other hormonal condition [15].

3.2.2 Horseshoe process regression

In Chapter 2, we developed horseshoe process regression (HPR), a model for fitting abruptly changing data. Let y_i be the BBT measurement observed on day t_i, $i = 1, ..., m$ of a single menstrual cycle. Then a HPR for these data would be:

$$\begin{aligned} y_i &= f_i + \epsilon_i \\ f_i &= \alpha + H_i \\ H_i|\tau, \lambda_i &\sim N(H_{i-1}, \tau^2\lambda_i^2(t_i - t_{i-1})),\ i = 2, ..., m,\ H_1 = 0 \\ \tau &\sim C^+(0, 1/\sqrt{s_\tau}),\ \lambda_i \overset{iid}{\sim} C^+(0, 1),\ i = 2, ..., m \\ \alpha &\sim N(a, b^2) \\ \epsilon_i|\sigma^2 &\sim N(0, \sigma^2),\ \sigma \sim C^+(0, 1/\sqrt{s_\sigma}) \end{aligned} \tag{3.1}$$

This model has three main components. First, it assumes normally-distributed measurement error, ϵ_i, about the estimated mean BBT trajectory, f_i. This measurement error is assumed to have constant variance over the menstrual cycle. The mean BBT trajectory f_i is made up of a y-intercept α, which gives the mean BBT on the first day of the menstrual cycle, and a nonlinear component H_i, which dictates the shape of the mean BBT trajectory across the menstrual cycle. H_i is a horseshoe random walk: incremental change in H_i across time is horseshoe distributed, and thus is a mixture of near-zero change, punctuated by large, abrupt shifts [12]. The variance of each incremental change, $\tau^2\lambda_i^2(t_i - t_{i-1})$, is made up of three terms. The global shrinkage parameter, τ, gives a measurement of the overall amount of motion of the BBT trajectory. If τ is large, we expect BBT to exhibit a great deal of "jumpiness" across the menstrual cycle; if τ is small, we expect it to be fairly constant. It also features a local shrinkage parameter, λ_i, with one local shrinkage parameter per daily increment in BBT. Even if τ is very small, there may still be large jumps in BBT because of a large single value of λ_i, if supported by data. Finally, variance increases as the elapsed time between BBT measurements, $(t_i - t_{i-1})$, increases, as may happen if a woman does not collect BBT measurements for several days.

Put together, the HPR model assumes an underlying functional form that looks like a step function, with stretches of approximately flat BBT interspersed with occasional jumps or drops. This matches the biphasic BBT pattern that many women exhibit. However, HPR can also accommodate the piecewise-linear BBT trajectories that are seen in some women [69]. Considering the prior clinical information we have on BBT over the menstrual cycle, we can make several connections to HPR's existing priors. First, it is likely appropriate to assume a constant σ^2 over the cycle and even across cycles, within one woman. It is also likely sensible to share information on τ^2 across

menstrual cycles within a single woman. Separate cycles would need separate values of α, though, as there is evidence that the baseline temperature differs from cycle to cycle, even within women. However, assuming that the α's from different cycles come from a common distribution seems plausible, e.g. the distribution of a random intercept. The most delicate question is how to handle the sequence of local shrinkage parameters, $\lambda_i, i = 2, ..., m$, within the BBT context. In its present form, HPR is agnostic to the biphasic pattern we expect of BBT—it does not know that we anticipate a single large jump in BBT around the midpoint of the cycle, with otherwise flat temperatures. This information would be best incorporated via the local shrinkage parameters. In addition, it is not clear how to share information about the local shrinkage parameters across cycles. It would not make sense to require the same $\lambda_i, i = 2, ..., m$ across cycles, as for most women there is too much variability in ovulation day to support such an assumption. It may not even make sense to share the distribution of local shrinkage parameters across cycles. We consider these challenges further below in Section 3.4.

HPR is currently estimated using Hamiltonian Monte Carlo (HMC) [72]. Although this HMC implementation is effective, it can be time-consuming. For this reason, we propose to offer an alternative approach to estimating HPR via variational inference.

3.2.3 Variational inference

Denote the parameters of our model as $\boldsymbol{\theta}$, with data $\boldsymbol{X}$. Then in a Bayesian context we seek to estimate the posterior $p(\boldsymbol{\theta}|\boldsymbol{X})$, usually via Monte Chain Monte Carlo (MCMC) methods. In contrast, variational inference (VI) seeks an approximation $q(\boldsymbol{\theta})$ to $p(\boldsymbol{\theta}|\boldsymbol{X})$ that will be easier to obtain than the full posterior. (Note that $q(\boldsymbol{\theta})$ conditions on the data $\boldsymbol{X}$ and hyperparameters, but convention is to suppress that dependence in notation.) Here, we follow a standard approach to VI and seek the approximate posterior that minimizes the Kullback-Leibler (KL) divergence between the true and approximate posteriors, $\int q(\boldsymbol{\theta}) \ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{X})} d\boldsymbol{\theta}$. Note that:

$$\ln p(\boldsymbol{X}) = \int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{X})} d\boldsymbol{\theta} \tag{3.2}$$

The right-most term is the KL-divergence between the approximate and true posteriors, and we seek to minimize it. Note that it is bounded below by 0. The term on the left-hand side, $\ln p(\boldsymbol{X})$, is constant with respect to $\boldsymbol{\theta}$. Therefore, to minimize the KL-divergence between the approximate and true posteriors, it suffices to maximize the middle term, $\int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}$, which is called the variational lower bound, variational objective, or evidence lower bound (ELBO). To reduce the space of options for $q(\boldsymbol{\theta})$, we make the *mean-field assumption*: the set of parameters $\boldsymbol{\theta}$ can be separated into K subsets such that $q(\boldsymbol{\theta}) = \prod_{k=1}^{K} q(\boldsymbol{\theta}_k)$, i.e. approximate posterior independence

between $\boldsymbol{\theta}_k, \boldsymbol{\theta}_j, k \neq j$. Under the mean-field assumption, it can be shown that for $k = 1, ..., K$, the $q(\boldsymbol{\theta}_k)$ that maximizes the variational objective is given by:

$$q(\boldsymbol{\theta}_k) \propto \exp E_{q_{j \neq k}}[\ln p(\boldsymbol{X}, \boldsymbol{\theta})] \tag{3.3}$$

Therefore, we can use the joint distribution of the likelihood and priors to obtain the q-densities $q(\boldsymbol{\theta}_k), k = 1, ..., K$. Note that the expectation above is *with respect to the q-densities*, not the priors or true posterior. Because the q-densities are all assumed to be independent, taking this expectation is often straightforward. If the likelihood and priors are conjugate, this will yield closed-form q-densities that depend on the expectations of other parameters with respect to their own closed-form q-densities. This provides the structure of an iterative algorithm in which we cycle through the expectations to obtain a current estimate of each q-density, evaluating the variational objective after all expectations are updated, until the variational objective reaches some level of convergence. At this point, the final values of the expectations provide us with the closed-form versions of the q-densities, from which we can obtain approximate posterior samples, if desired. For more information on VI, we refer the reader to Section 13.7 of Gelman et al. (2013) [31] or Wand (2017) [79].

3.3 Variational Inference for Horseshoe Process Regression

3.3.1 Algorithm Implementation

In order to enable conjugacy and ease of taking expectations [79], we reparametrize the HPR model as:

$$\begin{aligned}
y_i &= f_i + \epsilon_i \\
f_i &= \alpha + H_i \\
H_i - H_{i-1} | \tau^2, \lambda_i^2 &\sim N(0, \tau^2 \lambda_i^2 (t_i - t_{i-1})),\ i = 2, ..., m \\
H_1 &= 0 \\
\alpha &\sim N(a, b^2) \\
\tau^2 | a_\tau &\sim Inv\chi^2(1, 1/a_\tau),\ a_\tau \sim Inv\chi^2(1, s_\tau) \\
\lambda_i^2 | a_{\lambda_i} &\overset{iid}{\sim} Inv\chi^2(1, 1/a_{\lambda_i}),\ a_{\lambda_i} \sim Inv\chi^2(1, 1),\ i = 2, ..., m \\
\epsilon_i | \sigma^2 &\sim N(0, \sigma^2),\ \sigma^2 | a_\sigma \sim Inv\chi^2(1, 1/a_\sigma),\ a_\sigma \sim Inv\chi^2(1, s_\sigma)
\end{aligned} \tag{3.4}$$

Note that this model is equivalent to HPR as given in Equation 3.1; we have simply re-expressed the half-Cauchy priors as the quotient of two Inverse-χ^2 distributions [79], i.e.:

$$A^2|a \sim Inv\chi^2(1, 1/a),\ a \sim Inv\chi^2(1, s) \implies A \sim C^+(0, \frac{1}{\sqrt{s}}) \tag{3.5}$$

where the probability density function of the Inverse-$\chi^2(\kappa, s)$ distribution is:

$$p(x|\kappa, s) = \frac{(s/2)^{\kappa/2}}{\Gamma(\kappa/2)} x^{-(\kappa/2+1)} \exp\{\frac{-s}{2x}\},\ x > 0 \tag{3.6}$$

Equation 3.4 implies the following joint log-likelihood:

$$\begin{aligned} \ln p(\boldsymbol{X}, \boldsymbol{\theta}) = &\ln p(\boldsymbol{y}|\boldsymbol{H}, \sigma^2, \alpha) + \ln p(\boldsymbol{H}|\boldsymbol{\Lambda}, \tau^2) + \ln p(\alpha) + \ln p(\sigma^2|a_\sigma) \\ &+ \ln p(a_\sigma) + \ln p(\tau^2|a_\tau) + \ln p(a_\tau) \\ &+ \sum_{i=2}^{m} [\ln p(\lambda_i^2|a_{\lambda_i}) + \ln p(a_{\lambda_i})] \end{aligned} \tag{3.7}$$

where $\boldsymbol{H}$ is the length m vector containing the values of $H_i, i = 1, ..., m$ and $\boldsymbol{\Lambda}$ is a length $m-1$ vector containing the values of $\lambda_i^2,\ i = 2, ..., m$. We assume the following independence structure for our variational approximation:

$$q(\alpha, \boldsymbol{H}, \tau^2, a_\tau, \boldsymbol{\Lambda}, \boldsymbol{a_\lambda}, \sigma^2, a_\sigma) = q(\alpha)q(\boldsymbol{H})q(\tau^2)q(a_\tau)q(\sigma^2)q(a_\sigma) \prod_{i=2}^{m} q(\lambda_i^2)q(a_{\lambda_i}) \tag{3.8}$$

where $\boldsymbol{a_\lambda}$ is a length $m-1$ vector containing the values of $a_{\lambda_i},\ i = 2, ..., m$. This assumes that all parameters are approximately posterior independent of each other, except the values of $\boldsymbol{H}$, which are treated jointly. After calculating $q(\boldsymbol{\theta}_k) \propto \exp E_{q_{j \neq k}}[\ln p(\boldsymbol{X}, \boldsymbol{\theta})]$ for each parameter block $k = 1, ..., K$ (derivations given in Appendix H), that yields the following set of q-densities:

- $q(\alpha) = N(\alpha|\mu = \frac{E_{\sigma^2}[\sigma^{-2}](\boldsymbol{y}^T\boldsymbol{1}_m - E_H[\boldsymbol{H}]^T\boldsymbol{1}_m) + \frac{a}{b^2}}{mE_{\sigma^2}[\sigma^{-2}] + b^{-2}}, V = (mE_{\sigma^2}[\sigma^{-2}] + b^{-2})^{-1})$
- $q(\boldsymbol{H^*}) = MVN(\boldsymbol{H^*}|\boldsymbol{\mu} = (E_{\sigma^2}[\sigma^{-2}]\boldsymbol{I}_{m-1\times m-1} + E_R[\boldsymbol{R}])^{-1}E_{\sigma^2}[\sigma^{-2}](\boldsymbol{y^{*T}} - E_\alpha[\alpha]\boldsymbol{1}_{m-1}^T)^T, \boldsymbol{\Sigma} = (E_{\sigma^2}[\sigma^{-2}]\boldsymbol{I}_{m-1\times m-1} + E_R[\boldsymbol{R}])^{-1})$
- $q(\sigma^2) = Inv\chi^2(\sigma^2|\kappa = m+1, s = E_{H,\alpha}[(\boldsymbol{y} - \alpha\boldsymbol{1}_m - \boldsymbol{H})^T(\boldsymbol{y} - \alpha\boldsymbol{1}_m - \boldsymbol{H})] + E_{a_\sigma}[\frac{1}{a_\sigma}])$
- $q(a_\sigma) = Inv\chi^2(a_\sigma|\kappa = 2, s = E_{\sigma^2}[\frac{1}{\sigma^2}] + s_\sigma)$

- $q(\tau^2) = Inv\chi^2(\tau^2|\kappa = m, s = \sum_{i=2}^{m} \frac{E_H[(H_i - H_{i-1})^2]E_\lambda[\lambda_i^{-2}]}{\delta_i} + E_{a_\tau}[\frac{1}{a_\tau}])$
- $q(a_\tau) = Inv\chi^2(a_\tau|\kappa = 2, s = E_{\tau^2}[\frac{1}{\tau^2}] + s_\tau)$
- For $i = 2, ..., m$, $q(\lambda_i^2) = Inv\chi^2(\lambda_i^2|\kappa = 2, s = E_{\tau^2}[\frac{1}{\tau^2}]\frac{E_H[(H_i - H_{i-1})^2]}{\delta_i} + E_{a_{\lambda_i}}[\frac{1}{a_{\lambda_i}}])$
- For $i = 2, ..., m$, $q(a_{\lambda_i}) = Inv\chi^2(a_{\lambda_i}|\kappa = 2, s = E_{\lambda_i^2}[\frac{1}{\lambda_i^2}] + s_{\lambda_i})$

where $\boldsymbol{H}^*$ is the elements of $\boldsymbol{H}$ without H_1, $\boldsymbol{y}^*$ is the elements of $\boldsymbol{y}$ except y_1, and $\boldsymbol{R}$ is a first-order differencing matrix of variance terms for $\boldsymbol{H}$. Full derivations for the q-densities and a detailed outline of the algorithm can be reviewed in Appendix H.

After iterating through the parameters of each q-density until the variational objective converges, we have final forms for all of the expectations listed above. Then, we obtain point estimates of $\boldsymbol{f}$, the mean BBT trajectory, as $E(\alpha) + E(\boldsymbol{H})$, with variance $Var(\alpha) + Var(\boldsymbol{H})$. We could alternatively draw B samples from $q(\alpha), q(\boldsymbol{H})$ and take the mean/median and desired percentiles from the sum of the samples to provide point and uncertainty estimates of $\boldsymbol{f}$. In our experience these two approaches return equivalent results when the mean, 2.5^{th}, and 97.5^{th} percentiles are obtained from the q-density samples, compared to $E(\alpha) + E(\boldsymbol{H})$ and a 95% confidence interval constructed using $Var(\alpha) + Var(\boldsymbol{H})$.

3.3.2 Specifying Hyperparameters and Initial Values

The HPR model has a number of hyperparameters that need to be specified. The hyperparameters are a, b (the prior mean and variance of the normal prior on α, the y-intercept of the model) and $s_\tau, s_\sigma, s_{\lambda_i}$ (the prior scale parameters on the Inverse-χ^2 priors on $a_\tau, a_\sigma, a_{\lambda_i}$, respectively). For a and b, we recommend setting $a = \bar{y}_{1:10}$, the sample mean of the first ten BBT measurements, and setting b as 5 times the standard deviation of the BBT measurements. We recommend setting all values of $s_{\lambda_i} = 1$, which corresponds to a $C^+(0, 1)$ prior on $\lambda_i, i = 2, \ldots, m$. We set $s_\tau = 10000$, which corresponds to a $C^+(0, 0.01)$ prior on τ. We recommend setting $s_\sigma = (\frac{1}{sd(\boldsymbol{y})})^2$, which corresponds to a $C^+(0, sd(\boldsymbol{y}))$ prior on σ. All of these hyperparameter settings are based on the recommendations given in Chapter 2. In the HMC implementation in Chapter 2, we found that HPR was more sensitive to the choice of s_τ and b than to $a, \boldsymbol{s_\lambda}$, and s_σ. We find the same to generally be true with the VI implementation. Although these hyperparameter recommendations are a good place to start (and yield consistently good performance in the BBT setting), researchers working with different data may need to further tune for their setting.

In addition, the VI implementation requires us to initialize $E(\boldsymbol{H})$, $E(\alpha)$, $E(\frac{1}{\sigma^2})$, $E(\frac{1}{a_\sigma})$, $E(\frac{1}{\tau^2})$, $E(\frac{1}{a_\tau})$ and $E(\frac{1}{\lambda_i^2})$, $E(\frac{1}{a_{\lambda_i}})$ for $i = 2, \ldots, m$. For $E(\alpha)$, $E(\frac{1}{a_\sigma})$, $E(\frac{1}{a_\tau})$ and $E(\frac{1}{a_{\lambda_i}})$ there are natural initial values given by the priors: we initialize $E(\alpha) = a$, $E(\frac{1}{a_\sigma}) = \frac{1}{s_\sigma}$, $E(\frac{1}{a_\tau}) = \frac{1}{s_\tau}$, and $E(\frac{1}{a_{\lambda_i}}) =$

$\frac{1}{s_{\lambda_i}}$. For the remaining expectations, we initialize $E(\boldsymbol{H}) = \vec{0}$, which corresponds to a flat line through the data; we initialize $E(\frac{1}{\sigma^2}) = \frac{1}{var(\boldsymbol{y})}$; we initialize $E(\frac{1}{\tau^2}) = \frac{1}{0.01var(\boldsymbol{y})}$; and we initialize $E(\frac{1}{\lambda_i^2}) = \frac{1}{0.1^2}$. These specifications seemed justifiable given the interpretations of these parameters and usually yield good performance.

3.3.3 Comparison to Hamiltonian Monte Carlo

To assess the performance of our VI implementation of HPR, we compared it to the HMC implementation from Chapter 2. We considered three sample sizes, $m = 28$, $m = 112 = 28 \times 4$, and $m = 420 = 28 \times 15$, where m corresponds to the number of measurements taken over the cycle, e.g. once per day, four times per day, or 15 times per day. We considered four true underlying functions, motivated by the BBT setting:

1. bigstep: $f(t) = I(t \leq 14) * 36.6 + I(t > 14) * 37.1$
2. flat: $f(t) = 36.8$
3. joinpoint1: $f(t) = I(t \leq 14) * 36.6 + I(14 < t \leq 20) * (t/12 + 37.1 - 5/3) + I(t > 20) * 37.1$
4. joinpoint2: $f(t) = I(t \leq 7) * (-t/20 + 36.95) + I(7 < t \leq 14) * 36.6 + I(14 < t \leq 22) * (t/16 + 35.725) + I(t > 22) * 37.1$

For each of our 12 data-generating scenarios (3 options for number of measurements $\times$ 4 options for function) we generated 100 sample datasets and then estimated the HPR model on each dataset using either HMC or VI, following all hyperparameter and initial values specifications given above. Sample datasets for the $m = 28$ sample size are given in Figure 3.2.

For the two estimation approaches, we compared their mean point estimates, their efficiency (the standard deviation of the point estimates), their mean credible interval width, and their credible interval coverage, aggregated pointwise at each timepoint across the 100 replicates of each data-generating scenario. These results are given in Figures I.1-I.4 in Appendix I. In addition, we present examples of the different posteriors obtained via HMC and VI in Appendix J.

Overall, VI returned fits with slightly more noise, slightly less efficiency, and slightly worse coverage than HMC. These differences were particularly acute at $t = 1$, with improved performance after $t = 1$. In general, differences were more substantial in the joinpoint scenarios. VI's credible intervals were slightly wider than HMC's, except in the joinpoint scenarios and at $t = 1$ in all scenarios, for which they were narrower. Of particular note is that VI's credible intervals did not always narrow with sample size, as can be seen when comparing the credible interval widths in the bigstep and joinpoint scenarios between the $m = 112$ and $m = 420$ settings. However, in most cases, differences between HMC and VI lessened with increasing sample size. These differences

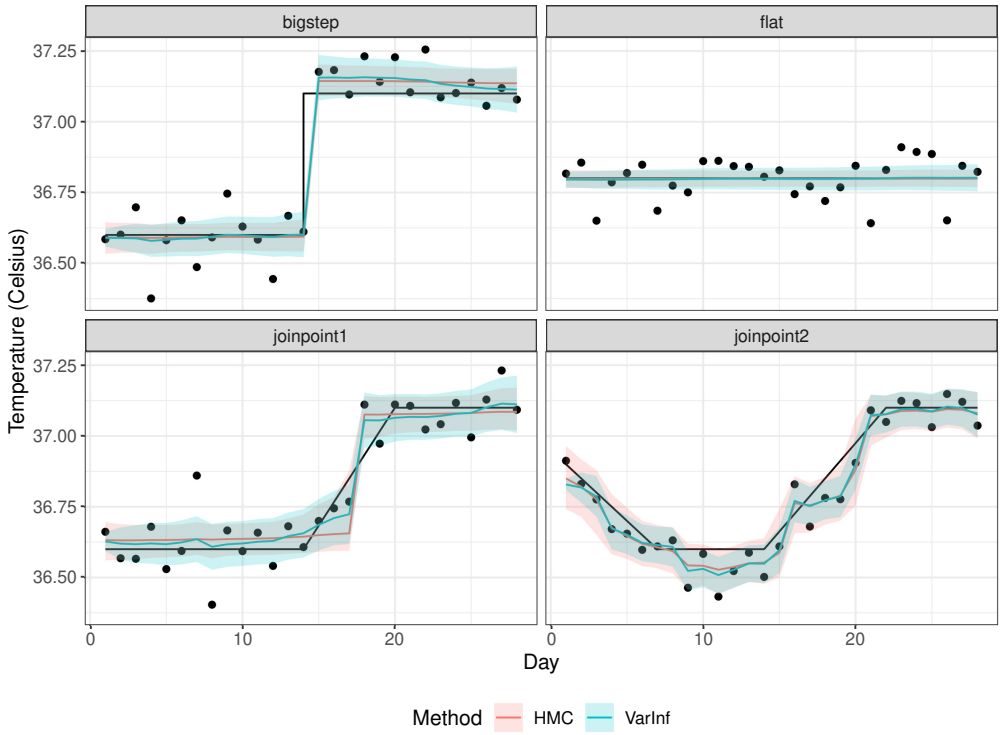

Figure 3.2: Sample datasets for variational inference (VarInf) and Hamiltonian Monte Carlo (HMC) comparison simulations, with $m = 28$. Each plot shows one sample dataset, and the point estimates and 95% credible intervals from VarInf and HMC estimation. The true underlying trajectory is given as a black line. Observed data are given as black dots.

in model fit were generally fairly minor (particularly after $t = 1$) and dwarfed by the signal in the data, as can be seen in the sample fits given in Figure 3.2. Examining the posterior densities shown in Appendix J, we note major discrepancies between the VI and HMC posterior densities for all parameters except $\boldsymbol{H}, \sigma^2$. These discrepancies do not seem to affect model performance, but they suggest that inference based on VI's estimates of the hierarchical variance parameters would be ill-advised. However, although VI returned slightly worse results than HMC, it offered markedly faster computation times (Table 3.1). While HMC would not be computationally feasible in a live-updating setting, VI likely would.

3.4 Horseshoe Process Regression for Basal Body Temperature Data

Although the VI implementation described in Section 3.3 reduced the computational challenges of using HPR to model BBT, it did not make any adjustments to HPR to better tailor it to the

Table 3.1: Median computation times (seconds) for the variational inference (VarInf) and Hamiltonian Monte Carlo (HMC) implementations of a horseshoe process regression (HPR) across 100 replicates.

	n = 28		n = 112		n = 420	
Function	VarInf	HMC	VarInf	HMC	VarInf	HMC
bigstep	0.3	40.2	0.6	185.2	2.2	1201.8
flat	0.3	7.4	0.5	32	2.3	195.5
joinpoint1	0.3	34.7	0.6	153.6	2.2	906.1
joinpoint2	0.3	33.4	0.6	155.7	2.3	858.6

BBT setting. Here, we describe two such modifications: an approach for incorporating ovulation day into the HPR model, and an *ad hoc* approach for sharing information on BBT trajectory and anticipated day of ovulation across menstrual cycles.

3.4.1 Incorporating Day of Ovulation

There are a variety of options for how to incorporate ovulation day into the HPR model. Here, we take a simple approach and add ovulation day as an explicit parameter in the model, O. The ovulation day parameter O determine the scales of the priors on the local shrinkage parameters. The updated model is then:

$$\begin{aligned}
y_i &= f_i + \epsilon_i \\
f_i &= \alpha + H_i \\
H_i - H_{i-1}|\tau^2, \lambda_i^2 &\sim N(0, \tau^2\lambda_i^2(t_i - t_{i-1})),\ i = 2, ..., m,\ H_1 = 0 \\
\alpha &\sim N(a, b^2) \\
\epsilon_i|\sigma^2 &\sim N(0, \sigma^2),\ \sigma^2|a_\sigma \sim Inv\chi^2(1, 1/a_\sigma),\ a_\sigma \sim Inv\chi^2(1, s_\sigma) \\
\tau^2|a_\tau &\sim Inv\chi^2(1, 1/a_\tau),\ a_\tau \sim Inv\chi^2(1, s_\tau) \\
\lambda_i^2|a_{\lambda_i} &\overset{iid}{\sim} Inv\chi^2(1, 1/a_{\lambda_i}),\ i = 2, ..., m \\
a_{\lambda_i} &\sim Inv\chi^2(1, 1),\ i = 2, ..., O, O+2, ..., m \\
a_{\lambda_{O+1}} &\sim Inv\chi^2(1, 1/4) \\
O &\sim Multinom(\mathbf{\Psi})
\end{aligned} \tag{3.9}$$

Thus, we assume that all of the scales of the local shrinkage parameters are *a priori* distributed according to $Inv\chi^2(1, 1)$, except one—the scale parameter $a_{\lambda_{O+1}}$, which corresponds to the scale of the local shrinkage parameter for the jump between the day of ovulation and the following day. That scale parameter is more tightly bound than the others, with an $Inv\chi^2(1, 1/4)$ prior.

Note that this is equivalent to placing $C^+(0,1)$ priors on all of the local shrinkage parameters $\lambda_i,\ i = 2, ..., m$ except λ_{O+1}, which has a $C^+(0,2)$ prior. This corresponds to our prior belief that there is a distinctly larger jump in BBT at time of ovulation, relative to the rest of the menstrual cycle. Although a $C^+(0,2)$ prior may not seem substantially different from a $C^+(0,1)$ prior, in this setting it is evidently enough of a difference to flag the day of ovulation. Making the ratio between the scale parameters larger (e.g. a $C^+(0,100)$ prior for day of ovulation vs. a $C^+(0,1)$ prior for the other days) did not affect performance; we opted to leave the ratio small.

The day of ovulation, O, must also have a prior. Here, we propose the use of a multinomial prior with prior probability vector $\mathbf{\Psi}$. The outcome in this case is which day is chosen for ovulation out of some set of candidate days (e.g. days 8-29), each of which has prior probability dictated by $\mathbf{\Psi}$. Figure 3.3 gives our proposed choice of $\mathbf{\Psi}$, distributed across days 8-29. We chose this set of days and $\mathbf{\Psi}$ based on the empirical distribution of ovulation day observed in large cohort studies [11, 23]. However, this is straightforward to modify if a multinomial distribution on a different set of days (e.g. days 10-35 rather than 8-29) or with different prior probabilities is preferred.

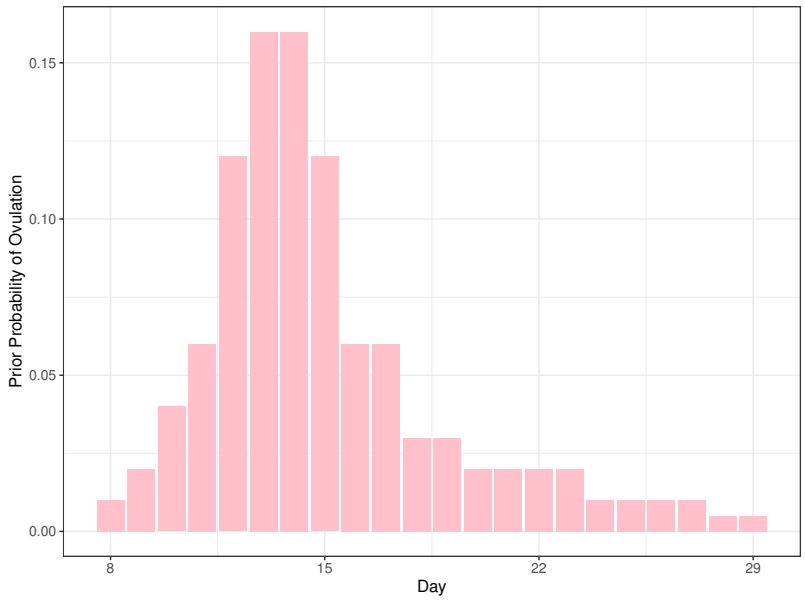

Figure 3.3: Proposed prior on day of ovulation.

Accommodating this revised model within the VI implementation is straightforward. It results in modifications to the q-density for $\boldsymbol{a}_{\boldsymbol{\lambda}}$, the creation of a new q-density for O, and slight changes to the variational objective L. More details on these changes are given in Appendix K. All other q-densities and initial value/hyperparameter specifications remain unchanged from what was described in Section 3.3.

We obtain the parameters of each q-density as output after the VI algorithm has converged. That

allows us to produce point and uncertainty estimates of the BBT trajectory, as before. In addition, we obtain approximate posterior probabilities of ovulation occurring on each day between days 8 and 29. There are a variety of ways we could translate these probabilities into a chosen day of ovulation. We propose to identify the first day with a posterior probability of ovulation greater than 0.3 as the day of ovulation; if no day has posterior probability of ovulation greater than 0.3 then we will use the posterior mode as the chosen day of ovulation. For uncertainty intervals, we sample 4000 ovulation days from a multinomial distribution with probabilities corresponding to the approximate posterior probabilities of ovulation and then take the 2.5^{th} and 97.5^{th} percentiles of these days to provide an uncertainty interval. Another good option might be to take the highest posterior density interval, in the case of multimodal posterior probabilities. A plot of the posterior probabilities of ovulation is another easily interpretable summary of when ovulation likely occurred.

3.4.2 Posterior-Prior Passing

As was discussed in Section 3.2, there is clinical evidence to support sharing information across cycles, within a woman. In particular, it seems sensible to share information on measurement error, σ^2; the global shrinkage parameter, τ^2; and the y-intercept, α. It would also be useful to share information on the local shrinkage parameters, $\mathbf{\Lambda}$, although it is less clear how to do so. Research suggests that the day of ovulation may vary by more than 7 days from cycle to cycle in a third of women [23], making it implausible to aggressively share information on the local shrinkage parameters across cycles. Here, we propose to share information on the distribution of the day of ovulation, O, across cycles.

Although it would be possible to implement HPR-BBT in a true repeated measurements fashion, treating τ^2, σ^2 as shared parameters across cycles and O, α as random effects coming from some common distribution across cycles, doing so would likely slow computational times as the number of cycles increases. In an attempt to get around this issue, we propose an *ad hoc* but computationally efficient scheme in which the posterior distributions of $\alpha, \tau^2, \sigma^2, O$ from one cycle become the prior distribution for those parameters in the next cycle. Because variational inference returns a closed form of the posterior (which is conjugate for the prior), this posterior-prior passing is straightforward to carry out. Put more formally, let $p_j(\theta|\cdot)$ be the prior for parameter θ in cycle j, conditional on some hyperparameters, and define $q_j(\theta|\cdot)$ as the q-density returned for parameter θ in cycle j, conditional on some q-density parameters. Then we propose:

$$
\begin{aligned}
q_j(\alpha|a_j, b_j) &\rightarrow p_{j+1}(\alpha|a_j, 50b_j) \\
q_j(a_\tau|\kappa_{\tau,j}, s_{\tau,j}) &\rightarrow p_{j+1}(a_\tau|\kappa_{\tau,j}, s_{\tau,j}) \\
q_j(a_\sigma|\kappa_{\sigma,j}, s_{\sigma,j}) &\rightarrow p_{j+1}(a_\sigma|\kappa_{\sigma,j}, s_{\sigma,j}) \\
q_j(O|\boldsymbol{\Psi}_j) &\rightarrow p_{j+1}(O|\boldsymbol{\gamma}_j = g(\boldsymbol{\Psi}_j, \boldsymbol{\gamma}_{j-1}))
\end{aligned}
\tag{3.10}
$$

Note that $\kappa_{\sigma,j}, \kappa_{\tau,j}$ are the estimated degrees of freedom for the Inverse-χ^2 q-densities of a_σ, a_τ, respectively, estimated for cycle j.

For the most part, we are directly passing the posterior from the previous cycle as the prior for the next cycle, with two exceptions. First, we multiply the q-density standard deviation of α from cycle j by 50 when passing it as the prior for cycle $j+1$. This is because the posterior q-density for α is often quite narrow, which can cause computational difficulties when imposed as a prior for new data. Second, we transform the q-density posterior probabilities of ovulation from cycle j using the transformation $g(\boldsymbol{\Psi}_j, \boldsymbol{\gamma}_{j-1})$. This transformation is done to improve retention of information across cycles about the distribution of day of ovulation and to smooth over multimodality. Specifically, $\gamma_{k,j}$, the k^{th} element of the length K vector $\boldsymbol{\gamma}_j$ is:

$$
\gamma_{k,j} = g(\boldsymbol{\Psi}_j, \boldsymbol{\gamma}_{j-1}) = \frac{\exp\{-|\sum_{l=1}^{k}(\Psi_{l,j} + \gamma_{l,j-1}) - \sum_{l=k+1}^{K}(\Psi_{l,j} + \gamma_{l,j-1})|\}}{\sum_{k=1}^{K}\exp\{-|\sum_{l=1}^{k}(\Psi_{l,j} + \gamma_{l,j-1}) - \sum_{l=k+1}^{K}(\Psi_{l,j} + \gamma_{l,j-1})|\}} \tag{3.11}
$$

In the above expression, note that the denominator serves to normalize the probability $\gamma_{k,j}$ so that the vector $\boldsymbol{\gamma}_j$ sums to 1, i.e. the denominator is the same as the numerator, except summed from 1 to K. Therefore, it suffices to focus on the numerator. Inside of the exponent, we have $|\sum_{l=1}^{k}(\Psi_{l,j} + \gamma_{l,j-1}) - \sum_{l=k+1}^{K}(\Psi_{l,j} + \gamma_{l,j-1})|$. First, note that the term $\Psi_{l,j} + \gamma_{l,j-1}$ is the sum of $\Psi_{l,j}$, the posterior probability of ovulation on day l from cycle j, and $\gamma_{l,j-1}$, the probability of ovulation on day l that was fed in as the prior for cycle j. With this sum, we forcibly incorporate information on ovulation from previous cycles via $\gamma_{l,j-1}$, and weight that information equally to the information on ovulation we have just obtained from the most recent cycle via $\Psi_{l,j}$. We then calculate $|\sum_{l=1}^{k}(\Psi_{l,j} + \gamma_{l,j-1}) - \sum_{l=k+1}^{K}(\Psi_{l,j} + \gamma_{l,j-1})|$—the absolute difference between the cumulative probability of ovulation occurring on days less than or equal to k and the cumulative probability of ovulation occurring on days greater than k. This serves to smooth over multimodality. The days that are closer to modes in $\Psi_{l,j} + \gamma_{l,j-1}$ will return smaller values from this expression, while days that are further from the modes will return larger values. This is the inverse of what we want (i.e. days that are closer to

modes should have larger probabilities, not smaller), so we then multiply the absolute difference by -1. Next, we exponentiate, to add smoothness. Finally, as discussed, we use the denominator to normalize the quantities from the numerator to sum to 1 and serve as a valid probability vector. To summarize, $g(\boldsymbol{\Psi}_j, \boldsymbol{\gamma}_{j-1})$ generally behaves sensibly and returns a set of probabilities which reflects the information from previous cycles, with multimodality smoothed out.

3.5 Data Application

3.5.1 Data

We use the HPR-BBT model to fit BBT data collected from a large cohort of British women. These data were gathered by the Catholic Marriage Advisory Council of England and Wales between 1955 and 1988. Women could write to the Council to obtain a free course on BBT-based natural family planning for contraception. After completing the course, a woman would collect a cycle's worth of BBT data and send it to the Council for review and annotation, a process which she would repeat until she was confident in her ability to identify her time of ovulation each cycle on her own. Most women sent between 6-12 cycles' worth of data, but some women sent far more. The Council kept photocopies of all of the BBT data they received, which were subsequently abstracted to form a data repository on BBT data across the menstrual cycle. Participants were primarily from England and Wales, although a substantial minority were from Ireland, Scotland, and other European countries. Participants were largely healthy and fertile. For more information on these data, please see Miolo et al. (1993) [55] and the University of Padua Department of Statistical Sciences Data Repository [76].

In total, the cohort consisted of BBT measurements from 36,139 cycles on 1,786 women, for a total of 779,216 BBT measurements. Women ranged in age from 16-55 years over data collection. We excluded cycles that were flagged by data abstractors as having inadequate measurements to identify the temperature jump and those that were affected by illness. We also restricted the sample to cycles whose end date could be confirmed by calendar date, to ensure that the reported end of the cycle was the true end of the cycle. We required that the cycle include at least one BBT measurement prior to day 9 of the cycle and in the last 8 days of the cycle. Finally, we restricted to women with age data available who recorded at least 3 usable cycles of BBT measurements, with the first cycle mostly complete (data collection started by day 3 and ended no more than 2 days early). More information on these sample exclusions can be reviewed in Appendix L, Figure L.1.

This resulted in a final sample of 10,017 cycles from 869 women, for a total of 266,690 BBT

measurements. Over the course of data collection, women ranged in age from 16 to 52, with a median age at cycle collection of 33. The number of cycles each woman collected ranged between 3 to 90, with women collecting a median of 7 cycles of data each (IQR: 5-14). The median cycle length was 28 days (IQR: 26-30). 65.7% of cycles had complete data; an additional 18.0% were missing 1-3 measurements. The remaining 16.3% of cycles were missing more than 3 BBT measurements, with a maximum of 21 missing measurements in a single cycle.

3.5.2 Individual Performance

To demonstrate HPR-BBT's performance for producing individual estimates, we first focus on data from a 35-year-old participant who recorded 6 cycles of usable BBT data. In Figure 3.4, we show her estimated BBT trajectory and most likely day of ovulation from each cycle. We compare the results from the HPR-BBT model without cross-cycle information sharing (HPR-NoInf), the HPR-BBT model with cross-cycle posterior-prior passing (HPR-Inf), a hidden Markov model (HMM) [78], and the cumulative sum test (CumSum) [66]. For the HMM, if multiple jumps in BBT were identified in a single cycle, we used the first jump that occurred after day 7.

In general, we see similar results from all methods, with some slight differences. The cumulative sum test (CumSum) often returned a later estimate of ovulation day than the other methods, which was particularly apparent in Cycles 2, 4, and 6. For the most part, HPR-Inf and HPR-NoInf returned very similar fits for the BBT trajectory for Cycles 2-6, with the most noticeable differences in Cycle 5. However, their estimates of the ovulation day and its uncertainty did differ occasionally, with HPR-Inf usually returning a wider uncertainty interval for ovulation day (as can be seen in every cycle except Cycle 6). In addition, by Cycle 4 HPR-Inf developed a preference for later ovulation (after estimating ovulation at day 17 for Cycle 3), which resulted in later estimates of ovulation for Cycles 4 and 5 (HPR-Inf estimated ovulation to be on day 19 for Cycle 4, while HPR-NoInf chose day 17; HPR-Inf estimated ovulation on day 14 for Cycle 5, while HPR-NoInf chose day 13).

In Figure 3.5, we give the approximate posterior probabilities that ovulation occurred on a given day for each cycle, and the resulting prior probabilities that were used by HPR-Inf. As was described in Section 3.4, the transformed probabilities were much wider than the untransformed posterior probabilities. We also note the smoothing of the multimodality seen in Cycles 2 and 4. In Cycle 5, for which there was a great deal of uncertainty about where to place the ovulation day, the transformed and untransformed probabilities were fairly similar, although the bimodality was smoothed over in the transformed probabilities.

Finally, in Figure 3.6, we show how HPR-BBT performed in the presence of real-time updating,

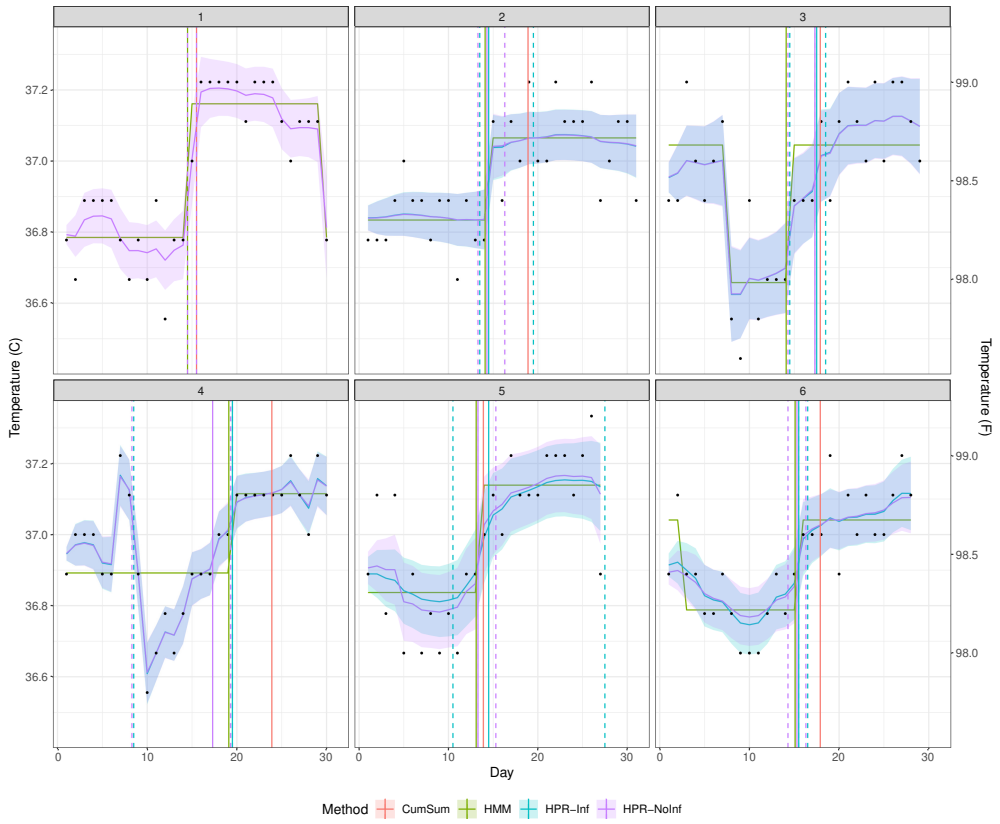

Figure 3.4: Estimated basal body temperature (BBT) trajectory and day of ovulation for 6 cycles of data from a 35-year-old woman. Point estimates of the BBT trajectory are given as solid lines, with uncertainty estimates around the trajectory given as shaded regions. Point estimates of day of ovulation are given as vertical solid lines, with uncertainty shown as vertical dashed lines. Observed data are given as black dots. We provide the results from the HPR-BBT model without information sharing (HPR-NoInf), the HPR-BBT model with information sharing (HPR-Inf), a hidden Markov model (HMM), and the cumulative sum test (CumSum). Note that only HPR-BBT provides uncertainty estimates.

by refitting the model as new data became available in the cycle. We show the model refit at days 19, 20, 21, and 22 of Cycle 5.

With the data up to day 19, we see that both models were fairly agnostic about the location of ovulation, although HPR-Inf placed more mass later in the cycle. After the adding the data from day 20, HPR-Inf shifted in recognition of a jump, which it initially placed on day 13 with uncertainty to day 14. HPR-NoInf remained agnostic. With day 21's data, HPR-NoInf also recognized a jump, which it initially placed on day 14 with uncertainty of days 13-15; HPR-Inf remained unchanged from day 20. On day 22, HPR-NoInf shifted the day of ovulation to day 13, with uncertainty still extending to day 15, while HPR-Inf remained unchanged. Overall, it took both methods 6 days post-ovulation to produce an initial guess of ovulation for this cycle.

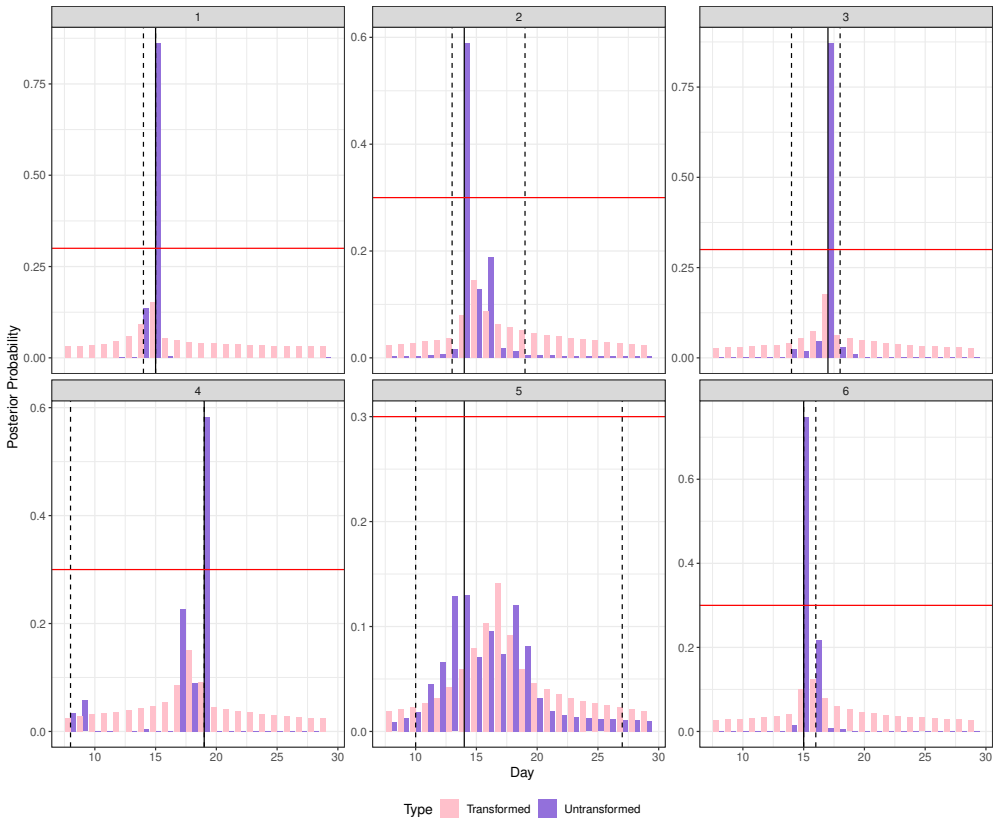

Figure 3.5: Approximate posterior probabilities of ovulation for 6 cycles of data from a 35-year-old woman estimated by the HPR-BBT model with information sharing (HPR-Inf). The approximate posterior probabilities are given in purple, while the transformed probabilities that were passed as the prior for the next cycle are shown in pink. The estimated day of ovulation and its 95% credible interval are shown as vertical black lines.

3.5.3 Population Performance

To give a sense of HPR-BBT's performance in the full cohort of 10,017 cycles from 869 women, we ran HPR-Inf and HPR-NoInf on all 10,017 cycles. Overall, HPR-Inf converged within the maximum number of iterations for 92.2% of cycles, while HPR-NoInf converged for 91.0% of cycles, suggesting that drawing on information from previous cycles may be helpful for setting priors and initial values for the VI algorithm. Among cycles that converged, HPR-Inf identified a day with posterior probability of ovulation greater than 0.3 in 76.1% of cycles, while HPR-NoInf did so in 81.6% of cycles. (Recall that in cycles where this threshold was not reached, the day with the largest posterior probability of ovulation was returned as the ovulation day estimate; however, this estimate would not be returned until cycle measurement was complete.)

When refit daily with each day's new BBT measurement, HPR-Inf would return its final estimate for day of ovulation within 3 days of its occurrence for 19.3% of cycles and within 4-7 days

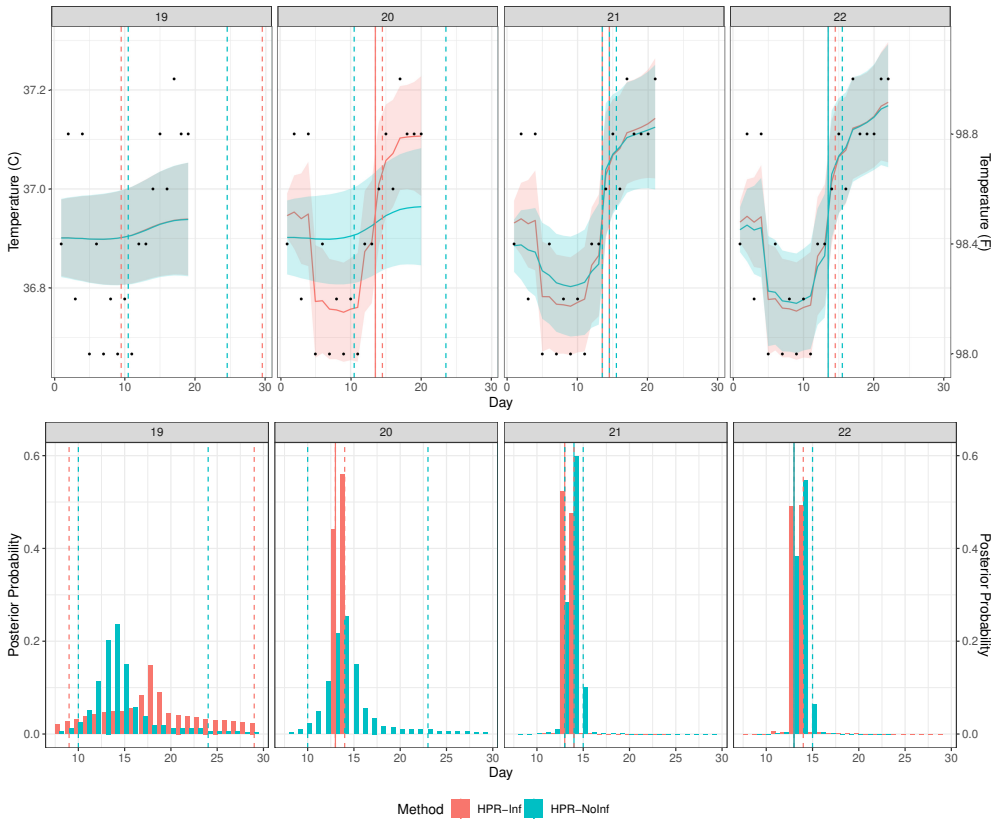

Figure 3.6: Estimated basal body temperature (BBT) trajectories and posterior probabilities of ovulation for Cycle 5 of 6 from a 35-year-old woman, refit with 19, 20, 21, and 22 days of data available. We provide the results from the HPR-BBT model without information sharing (HPR-NoInf) and the HPR-BBT model with information sharing (HPR-Inf). Uncertainty estimates for the BBT trajectory are shown as a shaded region; uncertainty for the estimated day of ovulation is given as dashed lines. Observed data are given as black dots.

of its occurrence for an additional 47.3% of cycles. 33.4% of cycles required more than a week after the estimated day of ovulation for ovulation to be identified. By comparison, HPR-NoInf would return its final estimate for day of ovulation within 3 days of its occurrence for 11.5% of cycles, within 4-7 days of occurrence for 55.5% of cycles, and more than a week after ovulation for 33.0% of cycles. This suggests that the information sharing from previous cycles enables higher posterior ovulation probabilities with less data from the current cycle, and thus HPR-Inf can produce its day of ovulation estimate more quickly than HPR-NoInf. Among cycles that produced an estimated day of ovulation before the cycle was complete, HPR-Inf produced an estimate that did not change for the remainder of the cycle for 64.5% of cycles. In 25.2% of cycles the estimate shifted once over the remainder of data collection; in 10.2% of cycles it shifted more than once. For HPR-NoInf, its estimate did not shift in 74.5% of cycles, had one shift in 20.8% of cycles, and had more than one shift in 4.7% of cycles.

Finally, as a check of the accuracy of HPR-Inf and HPR-NoInf's predictions, we consider how the population distribution of follicular and luteal phase length estimated by HPR-Inf and HPR-NoInf compares to results from Fehring et al. (2006) [23], who presented results from 1,060 cycles from 165 women, in which ovulation was confirmed by urinary estrogen levels. Across all cycles, HPR-NoInf estimated the median day of ovulation to be 15 (SD = 3.16), while HPR-Inf produced a median day of ovulation of 17 (SD = 3.43). Fehring et al. found that the median day of ovulation was day 16 (SD = 3.4), which seems generally in agreement with our findings. HPR-NoInf estimated the median luteal phase length to be 12 days (SD = 3.36), while HPR-Inf produced a median luteal phase length of 11 days (SD = 3.52). These estimates are less in agreement with Fehring et al., who found a median luteal phase length of 13 days (SD = 2.0). Taken together, this suggests that HPR-Inf, in particular, may be estimating ovulation to occur later than it actually does (and, by extension, a shorter luteal phase than reality)—or lingering data quality issues.

3.6 Discussion

Here, we have taken initial steps to make horseshoe process regression (HPR) more usable for modeling basal body temperature (BBT) data, by implementing a variational inference (VI) approach that speeds computation time and incorporates day of ovulation explicitly into the HPR model. We demonstrated that our VI re-implementation of HPR provides comparable performance to the HMC implementation from Chapter 2 in the BBT setting. We also showed that the HPR-BBT model, modified to include ovulation day and implemented via VI, returns sensible results for ovulation timing, compared to results from a large and well-regarded cohort study [23].

However, there is room for improvement. To provide computational gains, we pursued a fairly standard mean-field VI approach. Although this implementation yields good performance in our setting, there are a number of changes that might further stabilize estimation. The discrepancies between the posterior distributions obtained via Hamiltonian Monte Carlo (HMC) and VI are of particular concern (Appendix J), and resemble results found by Neville, Ormerod, and Wand (2014) [57]. These discrepancies are likely caused by lingering posterior dependence between the hierarchical parameters $a_\tau, \boldsymbol{a}_{\boldsymbol{\lambda}}$. To resolve this issue, Neville et al. (2014) recommended a reparameterization of the horseshoe distribution that does not rely on a fully conjugate parameterization, instead leaving the local shrinkage parameters as half-Cauchy and using more complex numerical approaches to obtain the q-densities [57]. This may be worthwhile to pursue in future work. In addition, we note that since we were able to obtain a conjugate representation of the HPR model for VI, it is likely possible to implement HPR using a Gibbs sampler. Further work is needed to explore the performance of a Gibbs approach, but that may offer comparable computation times

to VI while providing estimates of the true posterior, rather than an approximation.

More work remains before HPR-BBT will be truly functional for modeling BBT data. A primary challenge of the BBT setting is the complexity and messiness of the data, particularly in the context of menstrual tracking app data. We group these challenges into two categories: issues related to the study design and sampling mechanism, and contextual issues that are particular to BBT. In the study design and sampling mechanism domain, our analysis was complicated by missing data, which is common in the BBT setting. In particular, women often only recorded BBT measurements for the 10 days surrounding their anticipated ovulation date, which made it difficult for any of the methods we considered here to detect signal. It also made it difficult to confirm the true end date of the cycle, which is important to producing correct estimates of luteal phase length. We did what we could to verify cycle end date via calendar information, but we still saw evidence that some of the cycles in our cleaned data may have had incorrect end dates. We also saw evidence that some women may have been thresholding their own temperature measurements via rounding or assigning all temperature measurements below some threshold to a "low" temperature (and similarly for high temperatures above some threshold). This type of do-it-yourself data preprocessing was a primary cause of model convergence issues. Finally, one major shortcoming of the data we use here (and many large cohort studies on BBT patterns) is that our data do not have information on the true date of ovulation, making it impossible to assess the accuracy of HPR-BBT for detecting ovulation. We can compare to other algorithms for detecting ovulation based on BBT (as we have done here with the cumulative sum test and hidden Markov model), but these methods have their own shortcomings. Ideally our data would have information on the true ovulation date, confirmed by ultrasound or hormone analysis [23].

Beyond the issues with study design and sampling mechanism, the BBT data have difficult features that are inherent to the scientific problem. There is high variability in follicular and luteal phase length, and a high proportion of women do not follow the paradigmatic step function BBT trajectory. Other researchers have reduced these issues by eliminating women who show a high degree of menstrual cycle variability (e.g. those with more than 4 days' difference in cycle length across cycles [53]) or by restricting to women whose cycles fall in a particular length range [44], but these approaches exclude a large number of women whose BBT pattern is otherwise normal. Ideally, statistical methodology would be able to meet these challenges. In this regard, HPR-BBT is a step in the right direction, as it has a high degree of flexibility to accommodate non-paradigmatic cycles. It provides uncertainty estimates, which helps to distinguish between easy-to-classify and hard-to-classify cycles. In addition, its local-global shrinkage parameters provide natural quantification of different BBT trajectory patterns (e.g. piecewise linear rather than step function, triphasic rather than biphasic) and may be able to elucidate patterns of BBT trajectory.

However, HPR-BBT could be further improved for this purpose. We think adding other biomarkers and predictors beyond BBT would substantially reduce issues with unexplained variability in follicular phase length. These additional predictors are also essential to predicting time of next ovulation, rather than merely detecting previous ovulation—a key goal of menstrual tracking app technology. To make HPR-BBT more flexible, we might consider a different prior for day of ovulation that does not require specification of a finite set of days (as the multinomial prior does), to accommodate women with unusually early or unusually late ovulation. In addition, it may be useful to add further layers of hierarchy by placing priors on the scale parameters of the local shrinkage parameters and on the probabilities of ovulation, instead of treating them as fixed hyperparameters. Our current approach relies on the somewhat-arbitrary point estimate of ovulation day as being the first day with probability of ovulation greater than 0.3. Although this generally performs well, developing a point estimator that does not rely on a probability threshold may further improve performance. Finally, different methods for sharing information across cycles may provide better estimates. The posterior-prior passing scheme we use here offers modest gains in performance, but a true repeated measurements formulation (with clever computational implementation) may do even better.

Despite its challenges, BBT data are statistically fascinating and important to half the world's population—which makes the lack of methodology truly tailored to BBT surprising. We hope that our foray into this complicated realm inspires further work to develop elegant, scalable Bayesian approaches for BBT data. Of course, methodology will be (at most) half the battle: translation into actual menstrual tracking app use is where the real conundrums and rewards likely wait.

CHAPTER 4

A Multiple Imputation Approach for Cumulative Incidence Estimation

4.1 Introduction

Transitioning from the previous chapters' focus on horseshoe process regression and basal body temperature data, we now turn our attention to competing risks in survival analysis. Competing risks are a common challenge in applied survival analysis. For example, researchers may wish to study the incidence or causes of cancer-specific mortality, in which death from cancer is the primary interest. Death from other causes—such as heart attack or car accident—is of less interest. These other-cause deaths complicate estimation, however, because they create a new type of missing data: death from something else. Different from traditional censoring, in which we do not observe the outcome of interest because of study termination, participant drop-out, or loss-to-follow-up, censoring from competing events prevents us from observing the outcome of interest because the participant has already died of another cause.

Statistical discussions of this competing-risks missingness sometimes veer into the philosophical (if not theological). Can individuals who have already died from one cause of death still be considered at risk for another cause? Under what circumstances do we treat death from another cause as censoring in the traditional sense, e.g. equivalent to study dropout and similar? Is it even reasonable to consider the risk of death from a particular cause, as if other causes could be prevented? At what point is our cause of death determined? In the competing risks setting, the choice of estimator is intrinsically linked to these deeper questions about assumptions and interpretation. In this paper, we aim to give insight into these decisions within the setting of cumulative incidence estimation.

The cumulative incidence is a key descriptive statistic in the competing risks setting, and gives the probability of dying of the event of interest prior to time t (and, implicitly, of not having died of something else prior to that). It is usually estimated nonparametrically via the Aalen-Johansen

estimator, which will be described further below. Here, we instead propose to estimate the cumulative incidence using a multiple imputation scheme that imputes event times and types for censored individuals. On each of these imputed datasets, we can then estimate the cumulative incidence as a proportion. By reformulating the cumulative incidence problem as that of a proportion, it is easier to understand the differences between death from a competing event and traditional censoring, and developing estimators of the cumulative incidence in complex settings becomes more straightforward. In addition, the connection to proportions motivates new methods for uncertainty estimation for the cumulative incidence, which offer improved performance in some settings.

Our imputation approach has connections to other research that proposes methodology for or deeper understandings of cumulative incidence estimation. Efron (1967) first noted that the Kaplan-Meier survival estimator could be reformulated as a redistribution-to-the-right algorithm [19], which Gooley et al. (1999) subsequently demonstrated for the Aalen-Johansen estimator [33]. Following Taylor et al. (2002), we make this redistribution-to-the-right explicit via multiple imputation [74]. Ruan and Gray (2008) also proposed a multiple imputation approach for cumulative incidence estimation, although they opted to impute censoring times for individuals who died of a competing event—a very different approach from what we propose here [67]. We demonstrate that despite the seeming differences between our approach and that of Ruan and Gray (2008), they are functionally equivalent. Finally, the connections we make with binomial and multinomial distributions, and variance estimators motivated thereof, are reminiscent of work by Cox and Oakes (1984) [17] and Betensky and Schoenfeld (2001) [6], who used the binomial and multinomial distributions to develop nonparametric variance estimators for the survival and cumulative incidence functions, respectively. All of these approaches will be reviewed in greater depth below.

The remainder of the paper is as follows. First, we provide some necessary background on multiple imputation estimators for the survival and cumulative incidence functions. Then, we propose our multiple imputation approach for estimating the cumulative incidence function and its variance. We present simulation studies to demonstrate the empirical performance of our method. We close with a discussion of future work.

4.2 Background

4.2.1 Multiple Imputation for Survival Analysis

We will start by focusing on the simpler setting of single-cause survival analysis, i.e. no competing risks yet. Let T_i denote the time to some outcome of interest for $i = 1, ..., n$ subjects. Let C_i denote the corresponding time to censoring. Then in the all-cause survival analysis setting, the observed

data consist of $X_i = min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. Let $t_1, t_2, ..., t_l$ be the ordered, unique values of X_i, and let $d_j, j = 1, ..., l$ be the number of individuals who experienced the event of interest ($\delta_i = 1$) at time t_j. Let y_j be the number at-risk just before time t_j, e.g. $y_j = \sum_{i=1}^{n} I(X_i \geq t_j)$. The survival function $S(t) = \Pr(T > t)$ is of particular interest and is usually estimated nonparametrically with the Kaplan-Meier method [43]:

$$\hat{S}(t) = \prod_{t_j \leq t} (1 - \frac{d_j}{y_j}) \tag{4.1}$$

Efron (1967) observed that the Kaplan-Meier estimator could be reformulated as a redistribution to the right algorithm [19]. In the case of no censoring, each individual in the sample is allocated a weight $w_i = \frac{1}{n}$ and we estimate $\hat{S}(t) = 1 - \sum_{i:t_i \leq t} w_i$. In the presence of censoring, the weight of a censored individual is equally re-allocated to the weights of individuals still under observation after the time of censoring. Efron (1967) showed that this re-allocation approach is equivalent to the Kaplan-Meier estimator. Note that this approach still assumes independent censoring.

Taylor et al. (2002) made this re-allocation of weights explicit via multiple imputation [74]. They proposed to generate M imputations of the event time for censored individuals, by either sampling randomly from individuals still at risk, or by sampling from the survival distribution estimated by the Kaplan-Meier approach. These filled-in data could then be analyzed according to Efron's redistribution-to-the-right algorithm, generating $\hat{S}(t)^{(m)} = 1 - \sum_{i:t_i^{(m)} \leq t} \frac{1}{n}, m = 1, ..., M$. After analysis, the mean of the M estimates from the imputations provided the final estimate of $\hat{S}_{imp}(t)$, while Rubin's Rules provided variance estimates [50]. Taylor et al. demonstrated mathematically that with infinite imputations, these imputation estimates were equivalent to the estimates from the Kaplan-Meier method and, with finite imputations, empirically still very similar. In addition, the filled-in imputations could be analyzed using log-linear regression and—after pooling across imputations—would yield estimates equivalent to those from the Cox proportional hazards model [16]. In subsequent work, Hsu, Taylor, and Murray extended the imputation approach to allow for covariate-dependent imputation that could address dependent-censoring [39, 41] and to more complex all-cause survival settings [40].

4.2.2 Aalen-Johansen Estimation

In the presence of competing risks, we have an additional endpoint to consider. Let V_i be the time to a competing event for individuals $i = 1, ..., n$. Then the observed data are now $X_i = min(T_i, V_i, C_i)$ and $\delta_i = 1$ when $X_i = T_i$, $\delta_i = 2$ when $X_i = V_i$, and $\delta_i = 0$ when $X_i = C_i$. Let $t_1, t_2, ..., t_l$ be the ordered, unique values of the *event times*, e.g. times of individuals with

$\delta_i = 1, 2$. Let $d_j, j = 1, ..., l$ be the number of events of interest ($\delta_i = 1$) observed at time t_j, and let v_j be the number of competing events ($\delta_i = 2$) observed at time t_j. Let c_j be the number of individuals censored in the interval $[t_j, t_{j+1})$. In the setting of competing risks, the cumulative incidence function $F(t) = Pr(X \leq t, \delta = 1)$ is the most common quantity of interest, and it is usually estimated with the Aalen-Johansen estimator [2]:

$$\hat{F}(t) = \sum_{j:t_j \leq t} \hat{S}(t_{j-1}) \frac{d_j}{y_j} \tag{4.2}$$

Note that $\hat{S}(t_{j-1})$ is the estimate of *all-cause* survival at time t_{j-1}. Either the Kaplan-Meier or Nelson-Aalen estimator for $S(t)$ can be used; in the case of the Kaplan-Meier estimator, this would be $\hat{S}(t_{j-1}) = \prod_{t_k \leq t_{j-1}} (1 - \frac{d_k + v_k}{y_k})$.

Gooley et al. (1999) noted that the Aalen-Johansen estimator can be reformulated as a redistribute-to-the-right algorithm in the style of Efron (1967) [19, 33]. In the case of no censoring, the cumulative incidence is $\hat{F}(t) = \sum_{j:t_j \leq t} \frac{d_j}{n}$, and thus each individual has weight $\frac{1}{n}$. In the case of censoring, the weights of censored individuals are equally reallocated among individuals still at risk at time of censoring—exactly what we propose to do here via multiple imputation.

Variance estimation for the Aalen-Johansen estimator has been the topic of some discussion in the literature. Aalen and Johansen derived an initial variance estimator using martingale theory, which is the most commonly-used variance estimator in the applied literature [2]. However, the Aalen-Johansen variance estimator has been found to be anticonservative in small samples, particularly at later timepoints [4, 10]. More recent work has recommended the use of the variance estimators of Gaynor et al. (1993) [29] and Betensky and Schoenfeld (2001) [6], which are Greenwood-type estimators [4, 10].

4.2.3 Ruan and Gray Imputation

Ruan and Gray (2008) proposed a multiple imputation approach for survival analysis in the competing risks setting [67]. For individuals who died of a competing event ($\delta_i = 2$), they proposed to impute a *censoring* time by sampling from the censoring distribution among those still at risk, estimated using the Kaplan-Meier method. Having generated M imputations according to this scheme, each imputation thus contained only $\delta = 0, 1$ outcomes, which permits the use of traditional survival analysis techniques. To obtain estimates of cumulative incidence, it suffices to calculate $\hat{F}(t)^{(m)} = 1 - \hat{S}(t)^{(m)}$ on each imputation m, where $\hat{S}(t)^{(m)}$ is again calculated using the Kaplan-Meier method. Relying on existing theory that the Kaplan-Meier estimator is asymptotically normally distributed [46], these estimates $\hat{F}(t)^{(b)}$ are aggregated at each timepoint across

imputations using Rubin's Rules [50]. Similar to Taylor et al. (2002) [74], modeling was a key interest, and Ruan and Gray presented empirical evidence that the hazard ratio estimates from Cox proportional hazards models—when fit on each imputation and subsequently pooled across imputations—were equivalent to the subdistribution hazard ratio estimates obtained from Fine and Gray regression on the original, unimputed data [24, 67].

Similar to our own goals in this paper, Ruan and Gray (2008) aimed to make competing risks analysis—and particularly Fine and Gray regression—more easily adaptable. In this regard, they succeeded: methodology and software offerings for all-cause survival analysis are more flexible than in the multiple-cause setting, and Ruan and Gray's approach permits the use of any methodology intended for time-to-event data with $\delta = 0, 1$ outcomes.

Nonetheless, we have decided to pursue an alternative imputation scheme in this work, instead imputing event times for censored individuals, rather than censoring times for individuals who died of a competing event. We have two reasons for this. First, we find it somewhat counterintuitive to impute a censoring time for individuals who have already died of a competing event. Second, although the Kaplan-Meier estimator (which underlies their analysis strategy on each imputation) is easier to work with than the Aalen-Johansen cumulative incidence estimator, we think that a proportion (which underlies our method) is simpler still and makes further adaptation and variance estimation more straightforward.

4.3 Methods

4.3.1 Event Time and Type Imputation

As discussed, our data consist of $(\boldsymbol{X}, \boldsymbol{\delta})$, where X_i is the observed event time and δ_i is the event indicator, with $\delta_i = 1$ for the event of interest, $\delta_i = 2$ for the competing event, and $\delta_i = 0$ for censoring, for $i = 1, ..., n$ subjects. We propose to impute an event time and type for individuals who were censored ($\delta_i = 0$). We consider two different techniques for this imputation. Both of these approaches assume independent censoring such that $(T, V) \perp C$.

4.3.1.1 Risk Set Imputation (RSI)

In risk set imputation, we sample directly from individuals still at risk. In imputation $m, m = 1, ..., M$, we begin with $t_0 = 0$. Returning to our notation from before, between t_0 and t_1, there were c_0 individuals censored, who need to have event times and types imputed. We sample with replacement from the y_1 individuals still at risk at time t_1, and assign their event times/types to the c_0 censored individuals. Note that some of these c_0 censored individuals may be assigned to

individuals who are themselves censored later; in this case they will be re-imputed to still later times as we move through time.

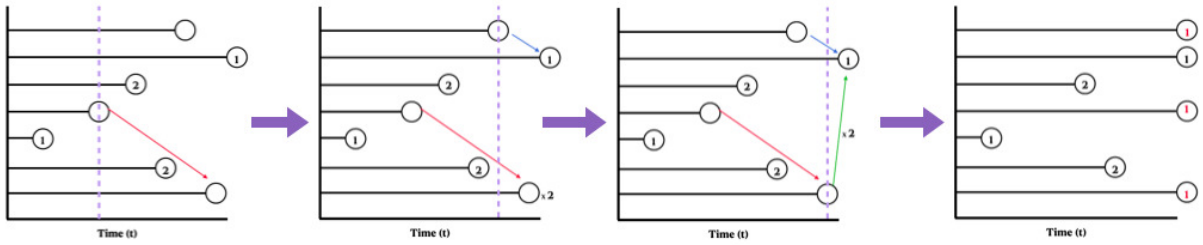

Figure 4.1: Schematic representation of risk set imputation to generate a single imputation, with the final imputed result given on the far right. Note the reallocation of the censored individual who was initially allocated to the bottom censored individual.

Having generated imputations for the c_0 individuals, we then move to generate imputations for the c_1 individuals who were censored in $[t_1, t_2)$, along with the c_1^* individuals who may have been imputed to a censoring time in $[t_1, t_2)$ in the c_0 round of imputation. We continue like this until we come to the $c_l + c_l^*$ individuals who were either censored (or imputed to be censored) on or after the final event time t_l. These individuals are left as censored—the only censored observations remaining in the dataset. We cannot impute an event time for them; therefore, we truncate the analysis at the final event time t_l and do not estimate the cumulative incidence after t_l. Having generated a single imputation m, we repeat this process until we have M imputations total. A schematic representation of this process to generate a single imputation is given in Figure 4.1.

4.3.1.2 Kaplan-Meier Imputation (KMI)

In some settings, the risk set approach may be less desirable. For example, if we wish to carry out the imputation conditional on covariates (as we envision for future work), we may run into small sample sizes when trying to sample from a covariate-restricted risk set. For this reason, we offer an approach that samples event times from the Kaplan-Meier estimate of the overall survival distribution, and conditional on that, an event type. A schematic representation of this process is given in Figure 4.2.

In imputation m, again starting with the c_0 individuals censored in $[t_0, t_1)$, we first draw an event time from the Kaplan-Meier estimate of the all-cause survival distribution among survivors, $Pr(min(T, V) > t | min(T, V) > t_0) = \hat{S}(t)/\hat{S}(t_0)$. In some datasets, there may be censored observations after the final event time. In this case, $\hat{S}(t)/\hat{S}(t_0)$ will not be a complete probability distribution (i.e. the probabilities summed across event times will not equal 1). To address this issue, when calculating $\hat{S}(t)/\hat{S}(t_0)$ we add a "shadow event time" just after the last censoring time

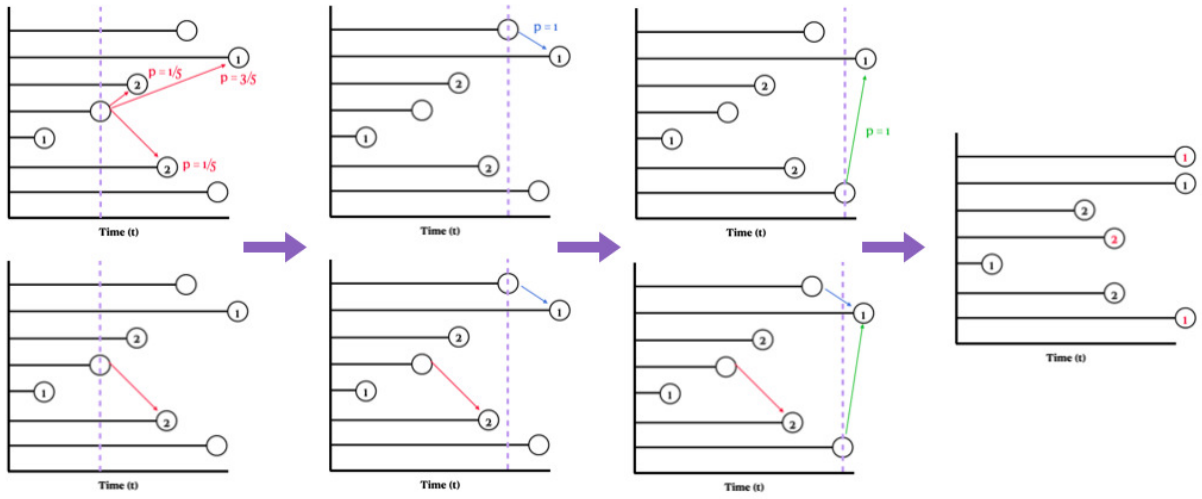

Figure 4.2: Schematic representation of Kaplan-Meier imputation to generate one imputation. The underlying imputation probabilities are given in the top row, with a sample imputed dataset shown in the bottom row. The final imputed result is given at the far right.

to make $\hat{S}(t)/\hat{S}(t_0)$ into a complete probability distribution. If the shadow event time is sampled, then the censored individual will be assigned to that time. Note, however, that the shadow event time will not be counted towards the sample size n, and that we will truncate the analysis at the last observed event time t_l. This shadow event time serves only to make $\hat{S}(t)/\hat{S}(t_0)$ complete.

Having sampled an event time from $\hat{S}(t)/\hat{S}(t_0)$, we must now assign an event type conditional on the sampled time. To do this, we sample an event type at random from the individuals who had an event at the newly-imputed event time. For example, if an individual censored in $[t_0, t_1)$ was imputed to have event time t_3, we would draw an event type at random from the $d_3 + v_3$ individuals who had an event at time t_3 (either a competing event or the event-of-interest). We then move on to the individuals censored in $[t_1, t_2)$, then $[t_2, t_3)$, and so on. In the interval $[t_j, t_{j+1})$, we sample from $\hat{S}(t)/\hat{S}(t_j)$.

Remark. For both the risk set imputation (RSI) and Kaplan-Meier imputation (KMI) approaches it would be possible to combine the imputation procedure with a bootstrap, to reflect the full uncertainty of the imputes and make RSI and KMI into proper imputation procedures. This was recommended by Taylor et al. (2002) [74]. However, in our experience, the bootstrap makes very little difference in the point and uncertainty estimates, which was also the case in Ruan and Gray (2008) [67]. For this reason, we do not recommend the use of a bootstrapped imputation, as it adds complexity for no clear gain.

4.3.2 Point Estimation

After we perform the multiple imputation, we have M copies of the dataset featuring only $\delta = 1, 2$ outcomes up to t_l, the last observed event time in the data. At this point, obtaining the cumulative incidence at time t is simple:

$$\hat{F}(t)^{(m)} = \frac{1}{n} \sum_{j:t_j \leq t} d_j + d_j^{*(m)} \tag{4.3}$$

where $d_j^{*(m)}$ is the number of individuals imputed to have the event-of-interest at time t_j in imputation m. We only calculate $\hat{F}(t)^{(m)}$ up to time t_l, the last observed event time in the dataset. We take the mean across imputations at each timepoint $t_j, j = 0, ..., l$ as our point estimator: $\hat{F}_{imp}(t_j) = \frac{1}{M} \sum_{m=1}^{M} \hat{F}(t_j)^{(m)}$. This leads to our first result:

> **Result 1.** $E[\hat{F}_{imp}(t)] = \hat{F}(t)$, where the expectation is taken across imputations and $\hat{F}(t)$ is the Aalen-Johansen estimator of the cumulative incidence.

A proof is given in Appendix M. Therefore, as the number of imputations increases, the imputation estimator will exactly reproduce the Aalen-Johansen estimator. Any difference between the imputation and Aalen-Johansen point estimates are attributable to the finite number of imputations.

4.3.3 Variance Estimation, Confidence Intervals, and Credible Intervals

4.3.3.1 Variance Estimation

To estimate the variance of $\hat{F}_{imp}(t)$, we rely on Rubin's Rules, which calculate the pooled variance of an estimator across multiple imputations as the weighted sum of the variance of the estimator between and within imputations. Rubin's Rules require that the estimator to be pooled be asymptotically normal [50]. At each timepoint and within each imputation, $\hat{F}(t)^{(m)}$ is merely a proportion: the number of individuals who have died or have been imputed to die of the event of interest by time t divided by the total sample size n. Then we can rely on the asymptotic normality of the estimate of a proportion to state:

$$\sqrt{n}[\hat{F}(t)^{(m)} - F(t)^{(m)}] \xrightarrow{d} N(0, F(t)^{(m)}[1 - F(t)^{(m)}]) \tag{4.4}$$

This echoes the approach taken by Cox and Oakes (1984) [17] and Betensky and Schoenfeld (2001) [6], who used likelihood theory to develop variance estimators for the Kaplan-Meier and Aalen-Johansen estimators, respectively. In the case of the Kaplan-Meier estimator, the likelihood is that

of a binomial distribution at each timepoint; for the Aalen-Johansen, it is a multinomial distribution. We rely on Cox and Oakes' demonstration of the asymptotic independence of the hazards at differing timepoints to alleviate any concerns about dependence between $\hat{F}(t_j)^{(m)}, \hat{F}(t_k)^{(m)}$ when $j \neq k$. Then we can apply Rubin's Rules to obtain:

$$\hat{Var}(\hat{F}_{imp}(t)) = W + (1 + \frac{1}{M})B \tag{4.5}$$

where $W = \frac{1}{M}\sum_{m=1}^{M} \hat{Var}(\hat{F}(t)^{(m)})$ and $B = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{F}(t)^{(m)} - \hat{F}_{imp}(t))^2$, with $\hat{Var}(\hat{F}(t)^{(m)}) = \frac{1}{n}\hat{F}(t)^{(m)}[1 - \hat{F}(t)^{(m)}]$.

4.3.3.2 Wald Interval

Using this pooled variance estimate, it is straightforward to obtain a Wald 95% confidence interval as $\hat{F}_{imp}(t) \pm 1.96\sqrt{\hat{Var}(\hat{F}_{imp}(t))}$. However, this simple version of the Wald interval is not guaranteed to have bounds within $[0, 1]$—a desirable property for the confidence interval of a probability. To correct this, we can use the standard complementary log-log transformation with the Delta Method:

$$\sqrt{n}[\log(-\log \hat{F}(t)^{(m)}) - \log(-\log F(t)^{(m)})] \xrightarrow{d} N\left(0, \frac{1 - F(t)^{(m)}}{F(t)^{(m)} \log(F(t)^{(m)})^2}\right) \tag{4.6}$$

Then we can again apply Equation 4.5, now with $W = \frac{1}{M}\sum_{m=1}^{M} \hat{Var}(\log(-\log \hat{F}(t)^{(m)}))$ and $B = \frac{1}{M-1}\sum_{m=1}^{M}(\log(-\log \hat{F}(t)^{(m)}) - C)^2$, with $C = \frac{1}{M}\sum_{m=1}^{M}\log(-\log \hat{F}(t)^{(m)})$ and $\hat{Var}(\log(-\log \hat{F}(t)^{(m)})) = \frac{1-\hat{F}(t)^{(m)}}{n\hat{F}(t)^{(m)}\log(\hat{F}(t)^{(m)})^2}$. A 95% confidence interval for $\hat{F}_{imp}(t)$ is $\hat{F}_{imp}(t)^{\exp(\pm 1.96\sqrt{\hat{Var}(\log(-\log \hat{F}_{imp}(t)))})}$.

This approach yields a Wald confidence interval for the cumulative incidence at each timepoint guaranteed to fall within $[0, 1]$. However, the use of the Wald interval may still cause concern, as the Wald interval's anticonservativeness and collapsed uncertainty at $\hat{p} = 0, 1$ are well known in the setting of confidence intervals for the binomial proportion [3]. We can take advantage of the rich literature on interval estimation for binomial proportions to propose two alternative uncertainty intervals for the imputation cumulative incidence estimator: a Wilson score interval and a Bayesian beta-binomial interval. As with intervals for binomial proportions, these two intervals may offer improved uncertainty estimates in small samples and when $\hat{F}_{imp}(t) = 0$.

4.3.3.3 Wilson Interval

The interval of Wilson (1927) inverts a 95% score test for the proportion to obtain the confidence interval bounds [81]:

$$\frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{[\hat{p}(1-\hat{p}) + z^2_{\alpha/2}/4n]/n}}{1 + z^2_{\alpha/2}/n} \tag{4.7}$$

This can be thought of as a shrinkage estimator that shrinks the estimate of $\hat{p}$ closer to $\frac{1}{2}$ [3]. Because it inverts the score test, the interval is naturally bounded in $[0, 1]$ and it has nonzero uncertainty at $\hat{p} = 0, 1$.

One challenge of using the Wilson interval in our setting, though, is that it does not provide an estimate of the variance, just a confidence interval. Rubin's Rules do not tell us how to pool confidence intervals across imputations. Fortunately, Lott and Reiter (2020) developed the theory needed to pool a Wilson interval across multiple imputations [52]. We implement their multiple imputation Wilson interval to provide uncertainty intervals for the cumulative incidence at each timepoint. Readers are referred to Lott and Reiter (2020) for full details [52].

4.3.3.4 Bayesian Interval

In addition to the Wilson interval, we also consider a Bayesian beta-binomial interval for the cumulative incidence. In a Bayesian beta-binomial interval for the proportion, we assume a prior $p \sim Beta(a, b)$, where a, b are hyperparameters. This yields a posterior distribution for p of $p|y \sim Beta(a + x, b + n - x)$, where x denotes the number of successes out of n trials. In our setting, we assume that $F(t_j)^{(m)} \sim Beta(a, b)$ and thus $F(t_j)^{(m)}|(\boldsymbol{X}, \boldsymbol{\delta}) \sim Beta(a + d_j + d_j^{*(m)}, b + n - d_j - d_j^{*(m)})$. Because the Beta distribution is only defined on $[0, 1]$, this interval will also naturally be bounded by $[0, 1]$ and will provide nonzero uncertainty when $\frac{d_j + d_j^{*(m)}}{n} = 0$, driven by the prior and the sample size.

To pool this interval across imputations, we follow recommendations for Bayesian inference after multiple imputation given by Zhou and Reiter (2010) [84]. On each imputation m and at each timepoint t_j, we draw S posterior samples from $F(t_j)^{(m)}|(\boldsymbol{X}, \boldsymbol{\delta})$. We then aggregate the posterior samples across all imputations to yield a full "posterior" sample of size $M \times S$ at each timepoint t_j. We can then obtain quantities of interest from this sample, such as the 2.5^{th} and 97.5^{th} percentiles, to construct a 95% credible interval of $\hat{F}_{imp}(t)$.

One important consideration is the values of the hyperparameters a, b. Setting $a = b = \frac{1}{2}$

is the most common choice, as this is the Jeffreys prior for the beta-binomial model. However, $a = b = \frac{1}{2}$ is a somewhat odd choice in the setting of cumulative incidence, as it places the bulk of its mass at $F(t) = 0, 1$, with less mass given to intermediate values, particularly those in $[0.25, 0.75]$. When competing risks are present, though, we would rarely expect $F(t) = 1$ even at the latest timepoints, and thus it seems nonsensical to place so much mass at 1. Instead, we recommend the use of $a = 0.8, b = 1.2$ (Figure 4.3).

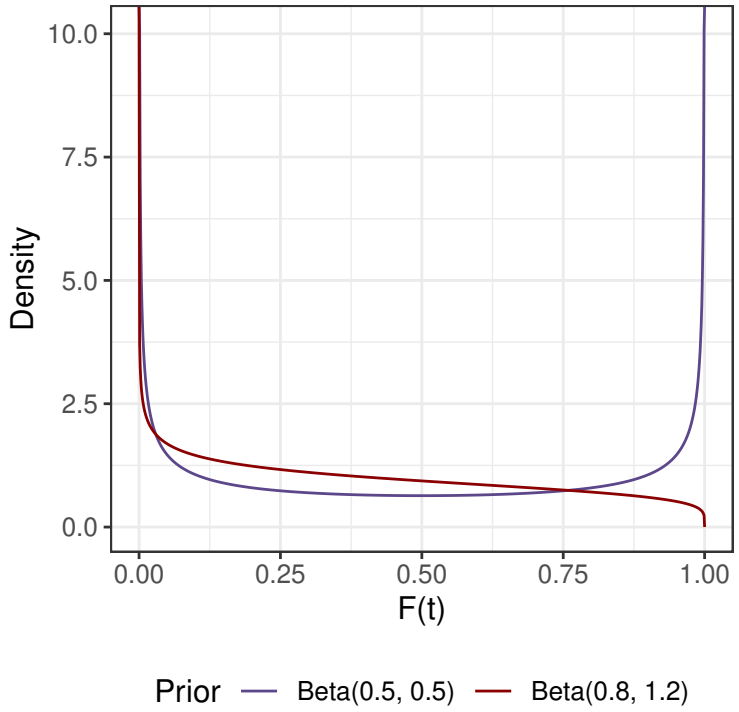

Figure 4.3: Two priors for the beta-binomial interval for the cumulative incidence imputation estimator.

Although still similar to $a = b = 0.5$, particularly with the large amount of mass placed at $F(t) = 0$, this modified prior puts more mass at values of $F(t) < 0.75$. However, if there is reason to believe that the cumulative incidence of the event of interest will be greater than 0.75 during study follow-up, it may be preferred to use the Jeffreys prior of $a = b = 0.5$, or some other modification that is suited to the applied setting. If further modification is pursued, we would recommend leaving $a < 1$, as the infinite a priori spike at $F(t) = 0$ seems important for inferential performance. More information on hyperparameter choice is given in Appendix N.

Table 4.1: Simulation parameter combinations. κ_1, κ_2 are the shape parameters of the Weibull hazards of the event of interest and competing event, respectively; u is the upper-bound of a uniformly-distributed censoring time; and the far-right column gives the timepoints at which performance is evaluated.

Scenario	κ_1	κ_2	Censoring	Timepoints
A	0.5	0.5	Low: $u = 10$	(1, 2)
B	0.5	2	Low: $u = 10$	(1, 2)
C	2	0.5	Low: $u = 10$	(2, 4)
D	0.5	0.5	Moderate: $u = 2$	(0.1, 0.5)
E	0.5	2	Moderate: $u = 2$	(0.1, 0.5)
F	2	0.5	Moderate: $u = 4.5$	(0.5, 1)
G	0.5	0.5	High: $u = 0.2$	(0.01, 0.02)

4.3.4 Synopsis

To summarize, we present two options for performing the imputation (RSI and KMI) and three options for obtaining uncertainty estimates (Wald interval, Wilson interval, or Bayesian interval). As we have shown above in Result 1 (and will confirm via simulation below), the choice of imputation approach has no effect on the point estimates. The variance estimators do produce different results; we will make recommendations below. All of these approaches are implemented in the R package `mici`, soon to be available on GitHub.

4.4 Simulation Study

We performed simulations to assess the performance of our imputation estimators. We modeled our simulation design from that used by Braun and Yuan (2007) [10]. Times to the event of interest, T, were simulated from a Weibull distributon with hazard $\lambda_1(t) = \kappa_1\rho(\rho t)^{\kappa_1 - 1}$. Two competing events were simulated, one (V_1) from an exponential distribution with mean 10 and the other (V_2) from a Weibull distribution with hazard $\lambda_2(t) = \kappa_2\rho(\rho t)^{\kappa_2 - 1}$. These two competing events were subsequently aggregated. Censoring times C were simulated as uniform over the interval $[0, u]$. u was chosen to result in either low censoring (20% censored), moderate censoring (50% censored), or high censoring (75% censored). Our observed data were then $X = min(T, V_1, V_2, C)$ with $\delta = 1$ if $X = T$, $\delta = 2$ if $X = V_1$ or $X = V_2$, and $\delta = 0$ if $X = C$.

In all cases, we set $\rho = 0.2$; we considered different combinations of values of κ_1, κ_2, u. The full list of combinations is given in Table 4.1, with sample datasets shown in Figure 4.4. We considered sample sizes of $n = 25, 100, 500$. In total, this yielded $7 \times 3 = 21$ different simulation scenarios. For each scenario we generated 1000 replicates.

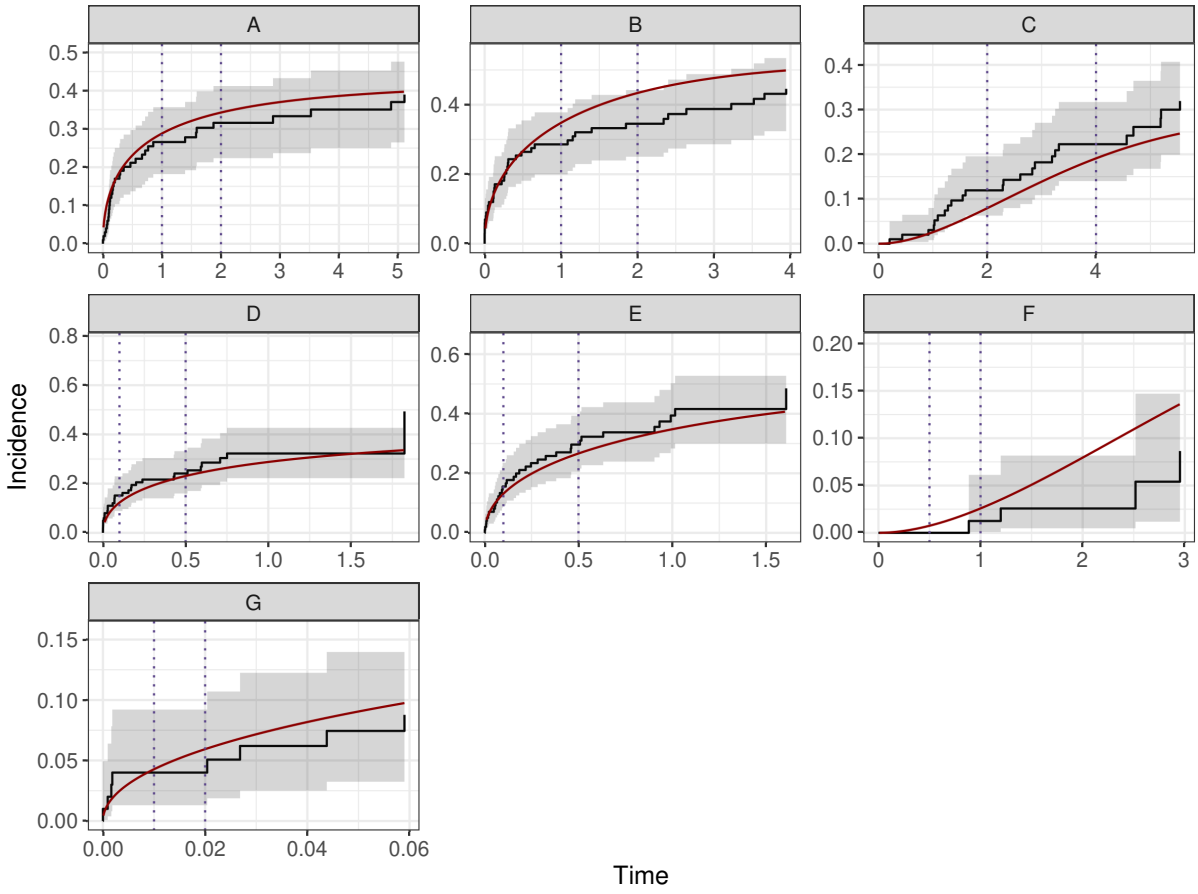

Figure 4.4: Sample datasets for the seven simulation scenarios, each generated for a sample size of $n = 100$. The Aalen-Johansen cumulative incidence estimate is given as a black line, with 95% confidence interval as the gray shaded region; the true cumulative incidence is given in dark red. Note that within each column, the only change is the rate of censoring—the true cumulative incidence curves are identical in each column. Rows correspond to rate of censoring (top is low censoring; middle is moderate censoring; bottom is high censoring). The timepoints at which performance is evaluated are given as vertical dotted purple lines.

We considered four different methods for comparison:

1. Our risk set imputation estimator (RSI), as implemented in our R package `mici`. Note that this approach offers three different uncertainty estimators: the Wald interval (RSI-Wald), the Wilson interval (RSI-Wilson), and the Bayesian interval (RSI-Bayes). For the Bayesian interval, we use our recommended prior of $Beta(0.8, 1.2)$ on the cumulative incidence and draw $S = 1000$ posterior samples on each imputation and timepoint.

2. Our Kaplan-Meier imputation estimator (KMI), as implemented in our R package `mici`. Note that this approach offers three different uncertainty estimators: the Wald interval (KMI-Wald), the Wilson interval (KMI-Wilson), and the Bayesian interval (KMI-Bayes). For the

Bayesian interval, we use our recommended prior of $Beta(0.8, 1.2)$ on the cumulative incidence and draw $S = 1000$ posterior samples on each imputation and timepoint.

3. Ruan and Gray imputation (RGI), as implemented in our R package `mici` [67].

4. The Aalen-Johansen estimator (AalJo). Note that we consider two different uncertainty estimators for the Aalen-Johansen estimator: the asymptotic variance of Aalen and Johansen as implemented in the R package `cmprsk` [24], which we abbreviate as AalJo-A, and the Greenwood-type variance of Betensky and Schoenfeld (2001) [6] as implemented in the R package `etm` [4], which we abbreviate as AalJo-G.

For the imputation estimators, we use $M = 150$ imputations in all cases. On each simulated dataset, we fit all of the methods listed above and output point estimates and 95% confidence/credible intervals (with complementary log-log transformations used when needed to obtain appropriately bounded intervals). We assessed performance at two timepoints, selected to occur at roughly the 25^{th} and 75^{th} percentiles for the observed times of the events of interest; these timepoints are given for each scenario in Table 4.1 and shown in Figure 4.4. At these two timepoints, we considered the bias, efficiency, 95% interval coverage, and 95% interval width as performance metrics. For these metrics, we use the underlying true cumulative incidence of the event of interest as our "truth", e.g., we calculate:

$$F_1(t) = Pr(X \leq t, \delta = 1) = \int_0^t \lambda_1(s) \exp(-\int_0^s \lambda_1(r) + \lambda_2(r) dr) ds \tag{4.8}$$

at the two timepoints of interest for each scenario. All simulations were performed in R [64]; code to fully reproduce the simulations will be made available on GitHub.

4.4.1 Performance for Point Estimation

Results for point estimator performance at the $n = 100$ sample size are given in Figure 4.5. Results for $n = 25$ and $n = 500$ are given in Appendix O, Figures O.1 and O.2.

As we would expect based on theory, all of the estimators produced near-identical point estimator performance for estimating the cumulative incidence, across a range of censoring rates, event rates, timepoints, and sample sizes. Bias and efficiency for estimating the true cumulative incidence were equivalent across methods; both improved with increasing sample size and decreasing censoring rates.

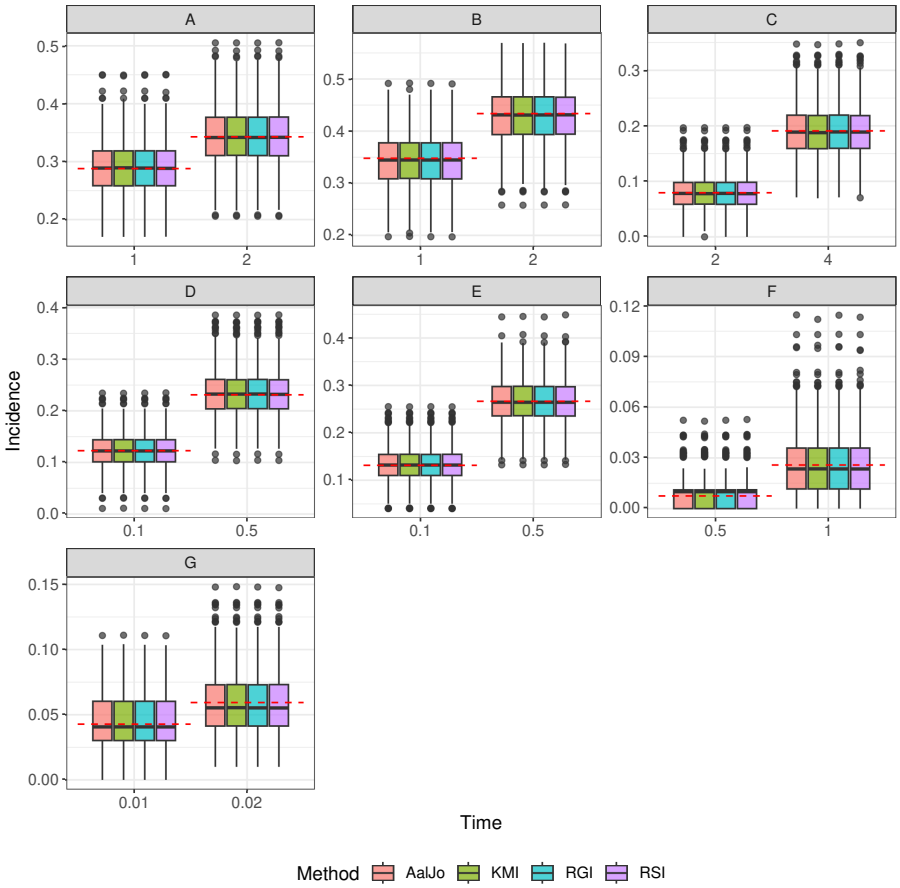

Figure 4.5: Point estimator performance for imputation and Aalen-Johansen (AalJo) estimators in seven simulation scenarios with a sample size of $n = 100$. The imputation estimators are Kaplan-Meier imputation (KMI), risk set imputation (RSI), and Ruan-Gray imputation (RGI). The true incidence is marked as a horizontal dashed line. A sample dataset for each scenario is given in Figure 4.4.

Table 4.2: Coverage rates for 95% uncertainty intervals from imputation and Aalen-Johansen estimators in Scenarios A and F. Scenario A has a low rate of censoring and similar incidence rates for the event of interest and the competing event; Scenario F has a moderate rate of censoring and the competing event is more common than the event of interest.

Method	n = 25		n = 100		n = 500	
Scenario A	Time = 1	Time = 2	Time = 1	Time = 2	Time = 1	Time = 2
AalJo-A	0.94	0.95	0.96	0.96	0.95	0.95
AalJo-G	0.96	0.95	0.97	0.95	0.95	0.95
KMI-Wald	0.93	0.94	0.96	0.95	0.95	0.95
KMI-Bayes	0.95	0.95	0.97	0.95	0.95	0.96
KMI-Wilson	0.95	0.95	0.96	0.95	0.95	0.96
RSI-Wald	0.93	0.94	0.96	0.95	0.96	0.95
RSI-Bayes	0.95	0.95	0.96	0.95	0.95	0.96
RSI-Wilson	0.95	0.95	0.96	0.95	0.95	0.96
RGI	0.93	0.94	0.96	0.95	0.95	0.96
Scenario F	Time = 0.5	Time = 1	Time = 0.5	Time = 1	Time = 0.5	Time = 1
AalJo-A	0.15	0.41	0.47	0.84	0.94	0.94
AalJo-G	0.15	0.38	0.47	0.84	0.94	0.95
KMI-Wald	0.15	0.41	0.47	0.84	0.94	0.94
KMI-Bayes	0.93	0.97	0.97	0.96	0.94	0.94
KMI-Wilson	0.90	0.95	0.97	0.95	0.96	0.94
RSI-Wald	0.15	0.41	0.47	0.84	0.94	0.94
RSI-Bayes	0.93	0.97	0.97	0.97	0.94	0.94
RSI-Wilson	0.90	0.95	0.97	0.95	0.96	0.95
RGI	0.15	0.41	0.47	0.84	0.94	0.94

4.4.2 Performance for Variance Estimation

Here, we focus on the results for Scenarios A and F; results for the other five scenarios were largely similar and are not shown. Coverage rates are shown in Table 4.2; 95% uncertainty interval widths are shown in Figure 4.6 for Scenario A and Figure 4.7 for Scenario F.

In Scenario A, in which the event rate was fairly high (about 40% at the end of follow-up) and censoring was low, interval width was largely similar across methods. The KMI and RSI estimators generally returned slightly narrower uncertainty intervals than the AalJo and RGI approaches. Coverage was approximately 95% for all estimators and at all timepoints and sample sizes.

In Scenario F, in which the event rate was lower (about 15% at the end of follow-up) and censoring was high, uncertainty estimation was much more challenging. Coverage rates were mostly nominal at the largest sample size ($n = 500$). At smaller sample sizes and earlier timepoints, we can see the benefits of the Bayesian and Wilson intervals' ability to have nonzero uncertainty prior

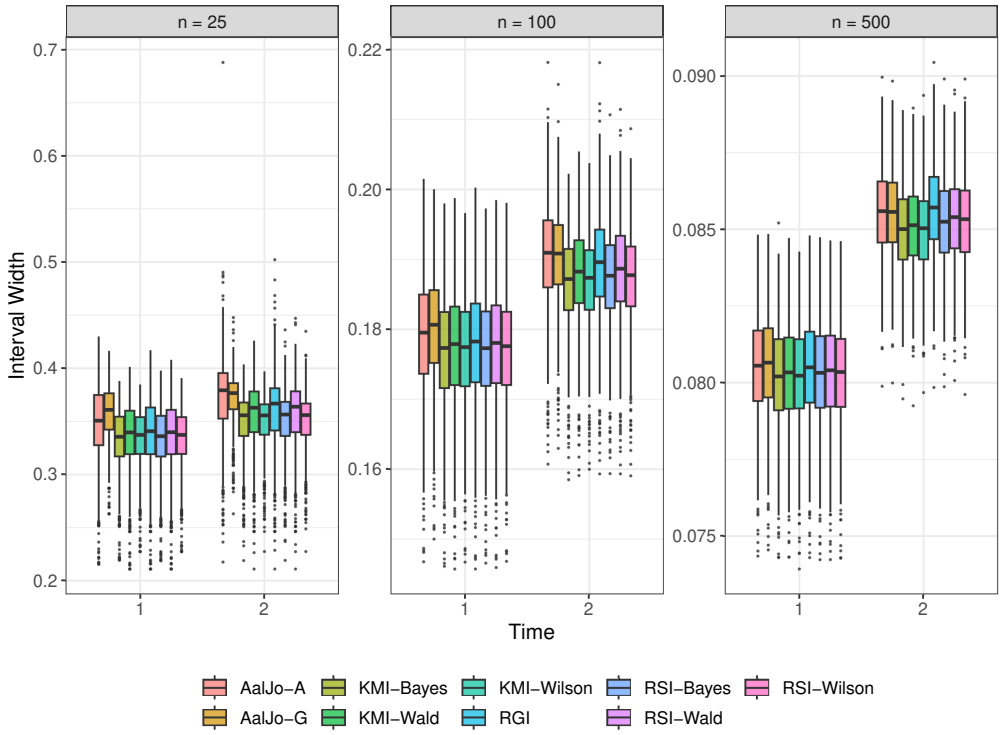

Figure 4.6: 95% uncertainty interval widths for imputation and Aalen-Johansen estimators in Scenario A at three sample sizes: $n = 25, 100, 500$. Scenario A has a low rate of censoring and similar incidence rates for the event of interest and the competing event.

to the first event time—their coverage rates at the 0.5 timepoint were consistently higher than 90%, much better than the other methods, which were closer to 15% at $n = 25$ and 47% at $n = 100$, in large part because of the high percentage of zero-width intervals returned by the AalJo, RGI, and Wald intervals. This can also be seen in Figure 4.7. Overall, we would recommend the use of either the Wilson or Bayesian intervals with either imputation approach.

4.4.3 Comment on Ruan and Gray Imputation

Our findings on the performance of Ruan and Gray's imputation scheme for competing risks endpoints confirm the results presented in their original paper: that imputing a censoring time for individuals who died of a competing event and then analyzing the imputed data as an all-cause survival endpoint is equivalent to Aalen-Johansen estimation [67].

It is critical to note that the RGI approach only permits imputation of a *censoring time* for individuals who died of a competing event, not a time for the event of interest. This connects to work by Gooley et al. (1999), who presented a redistribute-to-the-right reformulation of the Aalen-Johansen estimator [33]. In their redistribute-to-the-right formulation, they redistributed

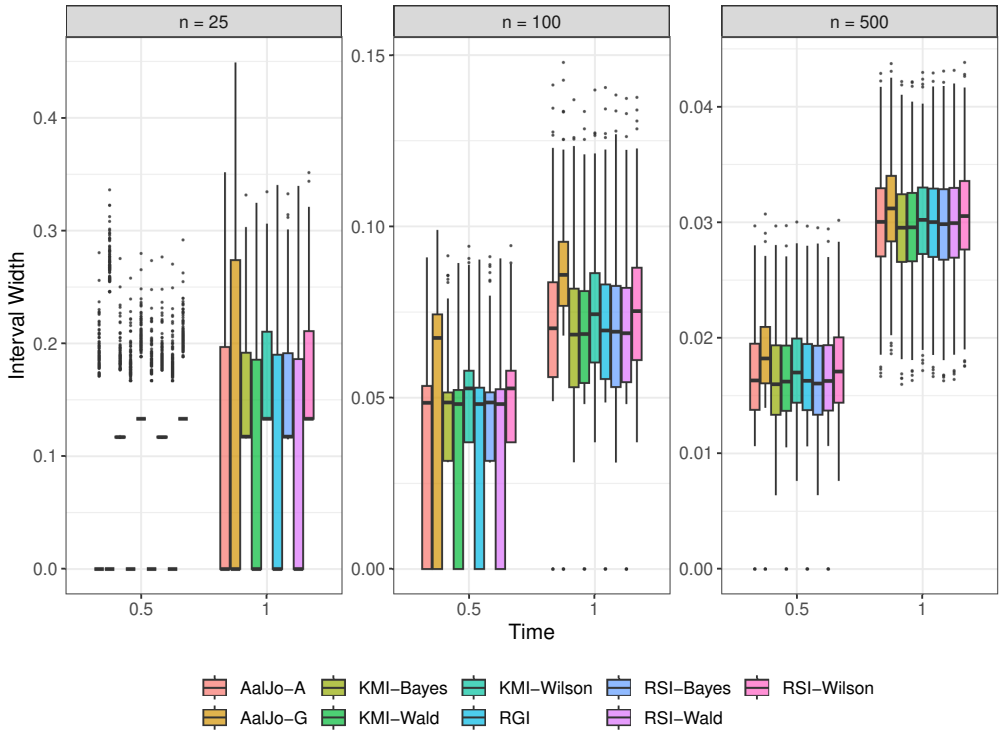

Figure 4.7: 95% uncertainty interval widths for imputation and Aalen-Johansen estimators in Scenario F at three sample sizes: $n = 25, 100, 500$. Scenario F has a moderate rate of censoring and the competing event is more common than the event of interest.

censored individuals to individuals who died of either the event of interest or a competing event, as we do here via multiple imputation. But they also considered the effects of different kinds of redistribution, such as redistributing individuals who died of a competing event. Gooley et al. found that redistributing individuals who died of a competing event to *both* censored individuals *and* individuals who died of the event of interest was equivalent to estimating the cumulative incidence as one minus the Kaplan-Meier estimate of survival—a survival analysis mistake that occurs frequently in the applied literature. As is well-documented, using one minus the Kaplan-Meier estimator to estimate the cumulative incidence returns estimates that are too high, and the sum of the cumulative incidences from each cause of death, when estimated using this approach, will sum to more than 1.

However, Gooley et al. did not consider the correctness of redistributing individuals who died of a competing event *only* to censored individuals—the approach taken by Ruan and Gray. Evidently, based on the results here and in Ruan and Gray (2008), the problem with reallocating individuals who died of a competing event is not reallocating at all, but rather reallocating to individuals who died of an event of interest. Put differently, for the purposes of Aalen-Johansen estimation, individ-

uals who died of a competing event can be viewed as censored individuals who will never have the opportunity to have the event of interest—distinct from truly censored individuals who do retain the opportunity to have the event of interest. Reallocating individuals who died of a competing event to censored individuals is equivalent to reallocating censored individuals to individuals who died of either event type—and both approaches are equivalent to Aalen-Johansen estimation.

4.4.4 Number of Imputations

A natural question for all of the imputation estimators is: how many imputations is enough? To explore this question, we compared the results using $M = 10$, $M = 50$, and $M = 150$ imputations in Scenarios A and G above. Scenarios A and G have the same underlying true cumulative incidence function, but Scenario G has a much higher rate of censoring than Scenario A. We refit all of the imputation estimators with the different options for number of imputations. Interestingly, for the purpose of estimating the true underlying cumulative incidence, the number of imputations had little effect. Even with only $M = 10$ imputations the performance on bias, efficiency, interval coverage, and interval width was the same as was reported above. This held true even in the presence of Scenario G's high censoring and at all three sample sizes (n = 25, 100, 500). It was only when the number of imputations dropped below $M = 5$ that performance on these metrics began to suffer, and even then only slightly.

However, if the goal is to reproduce the results of the Aalen-Johansen cumulative incidence estimator, then the number of imputations does matter. Figure 4.8 presents the difference between the imputation estimators and the Aalen-Johansen estimator at varying numbers of imputations in Scenario A. (Results for Scenario G are given in Appendix O, Figure O.3.) From this, we note that a large number of imputations is necessary to reproduce the Aalen-Johansen cumulative incidence point estimator—certainly more than $M = 10$, with $M = 150$ likely preferred. However, as sample size increases, the number of imputations can be reduced, as the difference between the imputation estimators and the Aalen-Johansen estimator decreases at larger sample sizes. One additional factor in the decision of how many imputations to use is computational time; results on this are also given in Appendix O.

4.5 Discussion

Here, we have reviewed existing approaches for using multiple imputation to estimate survival and cumulative incidence probabilities, and we presented a novel approach for using multiple imputation to estimate the cumulative incidence. We provided intuition for how the imputation approaches work, which offer their own insights, by extension, into how Aalen-Johansen estimation

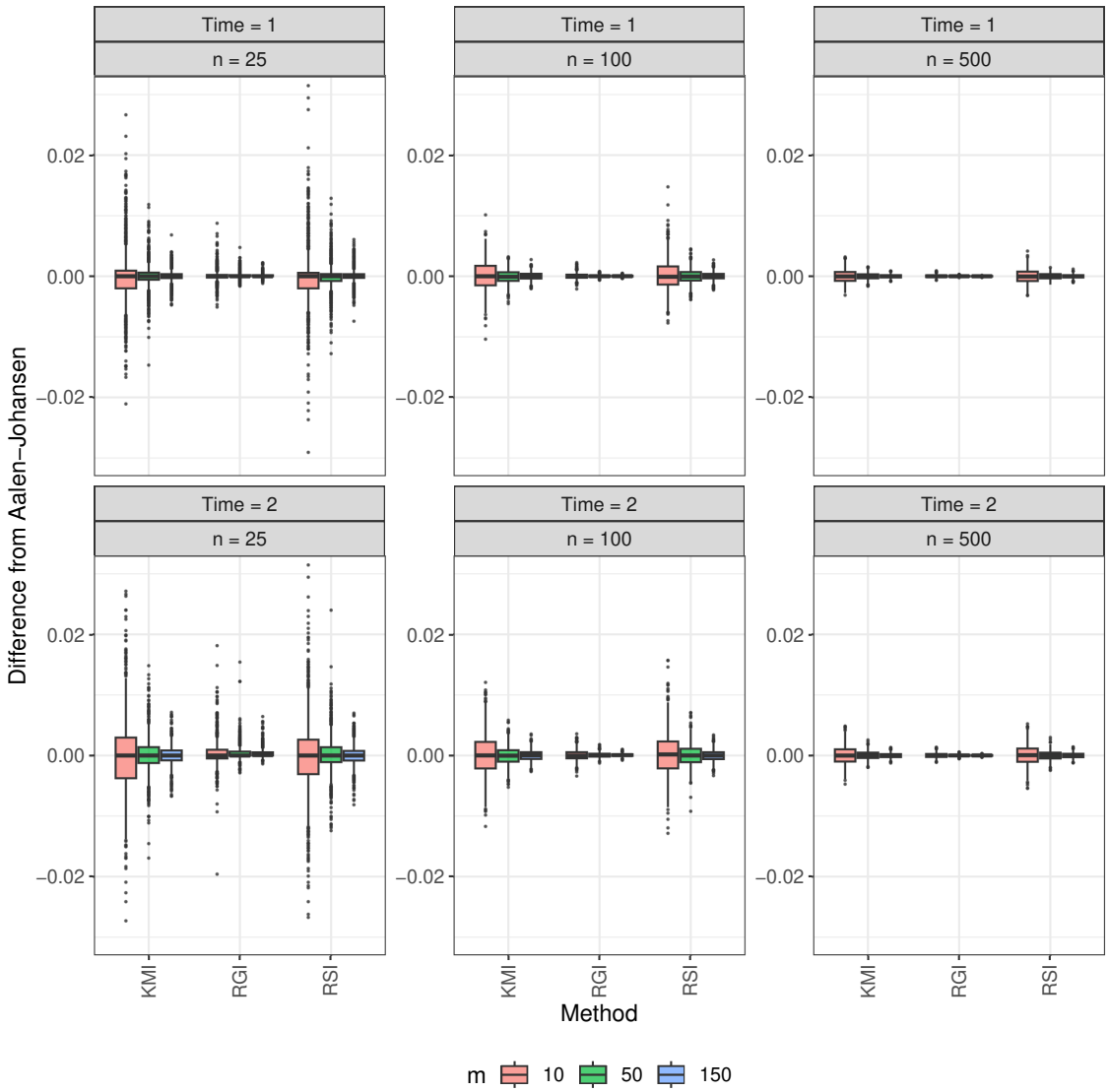

Figure 4.8: The difference between each imputation point estimator and the Aalen-Johansen point estimator at varying sample sizes and number of imputations in Scenario A. Note that the $n = 25$ plots have had their y-axis truncated to improve readability—there were additional outliers that fell outside of the range shown here.

works.

Our approach views cumulative incidence estimation as the problem of estimating a binomial proportion in the presence of partially missing data—censoring. Using multiple imputation, we account for censoring by imputing event times and types for censored individuals. As a result, estimating the cumulative incidence at each time point is merely calculating the proportion of individuals who have had the event of interest by that time. This reformulation highlights connections between existing variance estimation approaches for the cumulative incidence function—which

rely on asymptotic approximations for either martingales or the multinomial distribution—and similar asymptotic approximations for the variance of a binomial proportion, e.g. the Wald interval. In the case of binomial proportions, the Wald interval has long been seen as undesirable because the asymptotic approximation is poor in small samples [3]. Noting this connection enabled the development of two new variance estimators for the cumulative incidence function based on the Wilson score interval and Bayesian beta-binomial interval. Both return substantially better coverage rates in the presence of low event rates, as we would expect.

We provided mathematical and empirical evidence that the KMI, RSI, and RGI point estimators will exactly reproduce the Aalen-Johansen point estimator as the number of imputations increases. This may seem surprising: for the Aalen-Johansen estimator, redistributing individuals who died of a competing risk among the censored is equivalent to redistributing censored individuals among those who died of either event type. However, is it not clear that reproducing the Aalen-Johansen estimator should be our benchmark: ultimately, the goal is to correctly estimate the underlying true cumulative incidence function, which all of the imputation estimators did well, even with small numbers of imputations.

Our new uncertainty interval estimators relied on existing theory about uncertainty intervals for a binomial proportion and pooling uncertainty intervals across imputations. Of our two alternative intervals (the Wilson interval or the Bayesian interval), we would likely recommend the Wilson interval over the Bayesian interval when possible, as they perform equivalently and the Wilson interval does not require specification of hyperparameters. We saw evidence of mild sensitivity of the Bayesian interval to its hyperparameters. Future work may wish to explore this further. It may also be helpful to consider the use of time-varying hyperparameters for the Bayesian interval, as we would expect the event rate to increase over time and might like the prior to reflect that. In addition, although we only implemented these alternative intervals for our own imputation approaches, it may be possible to implement something similar for the RGI approach. The variance estimates for RGI come from pooling the Greenwood variance estimator for the Kaplan-Meier survival function across imputations. Use of a different variance estimator for the Kaplan-Meier estimator may yield better results.

Although we regard this work as impactful and thought-provoking in its own right, we believe that its true impact may be in the potential it offers for future work. Methodological offerings for estimating proportions from binary data are much richer than existing options for estimating cumulative incidence and hazard ratios for competing risks endpoints. We anticipate that our imputation approach in the competing risk setting can be extended to handle situations where there are survey or propensity weights, clustering, or dependent censoring. Regression modeling of the association between p predictors and the time to the event of interest is another interesting direc-

tion. The RGI approach had the appealing feature that after imputation a standard Cox proportional hazards model delivered results that would have been obtained from a Fine and Gray regression. The KMI and RSI approaches create datasets without any censored observations, and we expect that a broader set of regression modeling approaches could be implemented when there are no censored observations that need to be handled. Developing modeling extensions—and interpretations for the resulting quantities—may be fruitful work. Thinking even further afield, we could also imagine work that uses our multiple imputation approach to simplify analyses in the settings of semi-competing risks, multi-state models, or the restricted mean survival time.

CHAPTER 5

Closing Remarks and Directions for Future Research

In this dissertation, we have focused on two problems: modeling abruptly-changing data—particularly basal body temperature data—using horseshoe process regression, and using multiple imputation to estimate the cumulative incidence function. These two problems come from very different domains of statistics, with horseshoe process regression falling squarely in the realm of Bayesian modeling and statistical shrinkage, while our work on cumulative incidence lives in the land of nonparametric survival analysis. Nonetheless, a pragmatic, quasi-Bayesian sensibility can be seen in all projects, in our use of a discrete parameterization of the horseshoe process in Chapter 2 and variational inference in Chapter 3, and in our willingness to leverage multiple imputation and shrinkage to estimate the cumulative incidence and its uncertainty in Chapter 4. No project takes a purist approach, instead making allowances for the limitations of data, computational resources, and mathematical tractability. The result is ground-laying work on which we would like to build further in pursuit of more perfect solutions.

In Chapter 2, we presented a discrete formulation of horseshoe process regression (HPR) and showed that it has good performance for modeling abruptly changing data like step functions, piecewise linear functions, and impulse functions. We also considered how to obtain interpolations and predictions in the presence of a discrete set of local shrinkage parameters, along with partial linear modeling and modeling in the presence of monotonicity constraints. There are a variety of directions in which this project could be taken. We envision HPR as the abruptly-changing version of Gaussian process regression (GPR), and ideally it would be as easy to include a HPR term in one's larger model as it currently is to include a GPR smoothing term, if the applied setting justifies it [82]. There are a number of obstacles in the way of this; most notably, HPR is not as mathematically straightforward as GPR and does not have nice closed-form components. As a result, it is unlikely that it can ever be implemented without MCMC or approximate estimation, as we have done here. However, we still think further work to speed its computation (either through variational inference or other scalable Bayesian approaches), assess its performance in additive models with a wider range of nonlinear and linear terms, and examine its results in more diverse

data would be worthwhile. This would also help to increase awareness of the methodology and its potential (whatever that may be). Developing clever approaches for closed-form approximate kriging or posterior prediction may be another fruitful avenue of work; we ultimately recommended a fully Bayesian imputation scheme via MCMC in Chapter 2, but this is computationally burdensome.

One option which may resolve many of these issues, and which we touched upon in Chapter 2, is to pursue a more mathematically coherent formulation of HPR via subordinated Brownian motion and Meixner processes, as was described by Polson and Scott (2010) [63]. It is possible that such a model could yield analytical solutions or permit computational shortcuts; at minimum, it would likely offer more sensible ways to carry out interpolation and data augmentation. We did not pursue such an approach here because the mathematical development seemed formidable, to say nothing of implementation. However, with more time, patience, and experience, this continuous reformulation of HPR may be workable and could solve many of the problems described above that are caused by the discrete formulation.

In Chapter 3, we re-implemented HPR using variational inference (VI), and adapted HPR to detect ovulation based on women's basal body temperature (BBT) across the menstrual cycle and permit rudimentary information-sharing across cycles. Although these extensions were specifically targeted to modeling BBT, they would generalize to similar settings without much difficulty. We used this modified version of HPR, which we call HPR-BBT, to analyze data from a large cohort of British women. In the methodological space, we have a number of unanswered questions about the VI implementation of HPR. Although the approximate posterior we derived here performs well in the specific setting of BBT data, more work is needed to ensure that this approximation behaves in data other than BBT. It may be beneficial to consider more complex variational approximations than what we used here [57]. We made the mean-field assumption that all parameters were approximately posterior independent except the horseshoe trajectory estimates themselves, which we treated jointly. However, we saw some evidence of posterior correlation between the global shrinkage parameter τ^2 and measurement error σ^2, and it may be beneficial to model these parameters jointly, as well. Because the model behaves fairly well in its current form, we did not attempt a decentered parameterization of the model, although that may further improve performance [59]. In addition, our current VI implementation does not accommodate the Bayesian imputation scheme for data interpolation that we used in Chapter 2, and as a result, can only generate estimates at observed datapoints. This is a shortcoming that we would like to correct in future work, along with including the other features from Chapter 2 (additional linear predictors, monotonicity constraints) in the VI implementation.

Our approach to changepoint detection made the unknown changepoint (in our case, day of ovu-

lation) a parameter in the model and then placed a multinomial prior upon it. There are a number of ways to build on this. First, in the presence of a truly continuous predictor (unlike the BBT data, for which the predictor is discrete days) it would clearly be desirable to place a continuous prior on the changepoint. Second, it may be beneficial to introduce additional layers of hierarchy into the model. At present, we treat the scale parameters for the local shrinkage parameters as constant hyperparameters, but we could instead place priors upon them, which may improve estimation. We also treat the prior probabilities of the changepoint location (in our case, the prior probability of ovulation occurring on a given day) as fixed hyperparameters; these, too, may benefit from having their own prior. Third, our approach to information-sharing relies on a posterior-prior passing scheme that is somewhat *ad hoc*. As we discussed in Chapter 3, a mixed effects approach may yield better performance, especially if combined with careful computational implementation and a clever approach to "batching in" new data as it arrives. Fourth, although we assessed HPR-BBT's performance for live updating as new BBT data comes in each day, we did not carry out this updating in a particularly rigorous fashion, nor did we enable true posterior prediction. Further work on these topics would be beneficial, and is closely connected to methodology for prediction in the presence of discrete local shrinkage parameters, which we recommended as future work from Chapter 2.

With the BBT data specifically, there are many directions for future research. We think HPR shows great promise for quantifying different types of BBT patterns and identifying abnormal menstrual cycles. This may give insight into hormonal patterns surrounding ovulation and their implications for fertility. As we discussed in Chapter 3, including other menstrual cycle biomarkers is critical if the goal is to predict ovulation and would likely help with detecting ovulation. Such biomarkers might include age, cervical mucus, and information on illness, drinking, calendar season, and stress. However, we would need different data to make this possible, as our current cohort does not have these predictors [55, 76].

In Chapter 4, we proposed a nonparametric multiple imputation approach for estimating the cumulative incidence and demonstrated that it offered comparable performance to the traditional Aalen-Johansen approach. We also showed how it enabled two new variance estimators, which offer improved coverage rates when the event rate and/or sample size is low. However, we think the true impact of this work may be in its extensions. One immediate idea is to use this multiple imputation approach to estimate the cumulative incidence in the presence of dependent censoring. If we have reason to believe that censoring may depend on covariates which we have in our data, then for each censored individual we could calculate a similarity index between the censored individual and all of the uncensored individuals whose follow-up exceeded theirs, conditional on these covariates. After defining these similarity indices for all of the censored individuals, we

could carry out the nonparametric multiple imputation approach from Chapter 4, weighted by the similarity indices of each censored individual. In preliminary simulations, this similarity-weighted imputation approach corrects the dependent censoring and allows us to recover the true cumulative incidence—something the Aalen-Johansen estimator cannot do. There are existing methods for estimating the cumulative incidence in the presence of dependent censoring [7, 51], but the method we propose here is more straightforward than these others and flexible to further adaptation.

Another extension would be allowing propensity weights and/or clustering in the estimation of the cumulative incidence, which would likely be doable within our multiple imputation approach. We hypothesize that we would need to incorporate the propensity weights and/or clustering structure in both the imputation and analysis phases. This is different from the dependent censoring weights described above, which are only included in the imputation phase. We would carry out the imputation with the weights and/or clustering, and then analyze each imputed dataset using existing theory on estimating a weighted and/or clustered binomial proportion. The details of variance estimation in this setting may require careful attention.

We are particularly excited about using our multiple imputation approach within the context of regression modeling. In the all-cause survival setting [74] or when Ruan and Gray's alternative imputation approach for competing risks data [67] is used, there are natural extensions to regression modeling. In the all-cause survival setting, the imputed datasets can be analyzed using log-linear regression and will return estimates comparable to the Cox proportional hazards model [16, 74]; in the Ruan and Gray (2008) setting, the imputed datasets can be analyzed using the Cox proportional hazards model and will return estimates comparable to those of a Fine and Gray regression [24, 67]. With our multiple imputation approach, extensions to regression modeling are less obvious. It is not clear how we would analyze each imputed dataset to identify relationships between predictors and the time to the event of interest, and if the resulting quantities would have a meaningful interpretation. We would be interested to find out. Other uses of the multiple imputation approach might include using it to analyzing the restricted mean survival time, in the semi-competing risks setting, or for multi-state modeling.

Finally, our multiple imputation approach sits at the border between nonparametric and Bayesian methods for survival analysis. It might be interesting to take a more fully Bayesian approach to the imputation. This would require careful thought about prior specifications and model implementation. Bringing the skills honed in Chapters 2 and 3 to bear on the questions created by Chapter 4 may lead to useful work in Bayesian survival analysis. Happily, it also offers a narrow path to connect the far-flung topics of this dissertation.

APPENDIX A

Additional Simulations on HPR's Performance for Binary and Count Data

Here we explore the performance of horseshoe process regression (HPR) for binary and count outcomes via simulation. We considered four true underlying associations, each of which were observed at an equally spaced grid of n = 100 observations for count data and n = 150 observations for binary data:

1. bigstep: $f(x) = 0 * I(x \leq 2) + 6 * I(2 < x \leq 5) + 1 * I(5 < x \leq 6) + 3 * I(6 < x \leq 8) + 10 * I(x > 8)$ (divided by 10 for binary data).

2. bounce: $f(x) = |sin(x)|$ (multiplied by 10 for count data).

3. impulse: $f(x) = 0 * I(x = 0) + \exp(-x) * I(0 < x < 3) + 1 * I(x = 3) + \exp(-(x-3)) * I(3 < x < 7) + \exp(-(x-7)) * I(x = 7)$ (multiplied by 5 for count data).

4. joinpoint: $f(x) = (1.5x) * I(x < 2) + (16 - 5x) * I(2 \leq x < 3) + 1 * I(3 \leq x < 6) + (10 - x) * I(6 \leq x < 9) + (5x - 44) * (x \geq 9)$ (divided by 6 for binary data).

$I()$ denotes the indicator function; i.e. $I(x) = 1$ if condition x is true, and $I(x) = 0$ otherwise. We compared HPR to Gaussian process regression (GPR) and adaptive splines (Adspline). Unlike in Chapter 2, we did not consider the median filter (MedFilt) or trend filter (TrendFilt) because these methods are not implemented for noncontinuous outcomes.

We assessed performance with three primary metrics:

1. Mean absolute difference (MAD): $\frac{1}{n} \sum_{i=1}^{n} |g^{-1}(f(x_i)) - g^{-1}(\hat{f}(x_i))|$, where $g^{-1}(\hat{f}(x_i))$ is the predicted function's value at x_i on the mean scale and $g^{-1}(f(x_i))$ is the true function's value at x_i on the mean scale.

2. Credible/confidence interval width (Width): $\frac{1}{n} \sum_{i=1}^{n} g^{-1}(\hat{f}(x_i)^{0.975}) - g^{-1}(\hat{f}(x_i)^{0.025})$, where $\hat{f}(x_i)^{0.975}$ denotes the upper bound of a 95% credible/confidence interval for $\hat{f}(x_i)$ and $\hat{f}(x_i)^{0.025}$ is

the lower bound. Both bounds are transformed back to the mean scale using g^{-1} to assess credible interval width.

3. Credible/confidence interval coverage (Coverage): $\frac{1}{n}\sum_{i=1}^{n} I(g^{-1}(\hat{f}(x_i)^{0.025}) \leq g^{-1}(f(x_i)) \leq g^{-1}(\hat{f}(x_i)^{0.975}))$.

We assessed performance on each metric across the 100 replicates of each of our 3 data-generating scenarios for each method. All code used to completely reproduce the simulations can be found on GitHub.

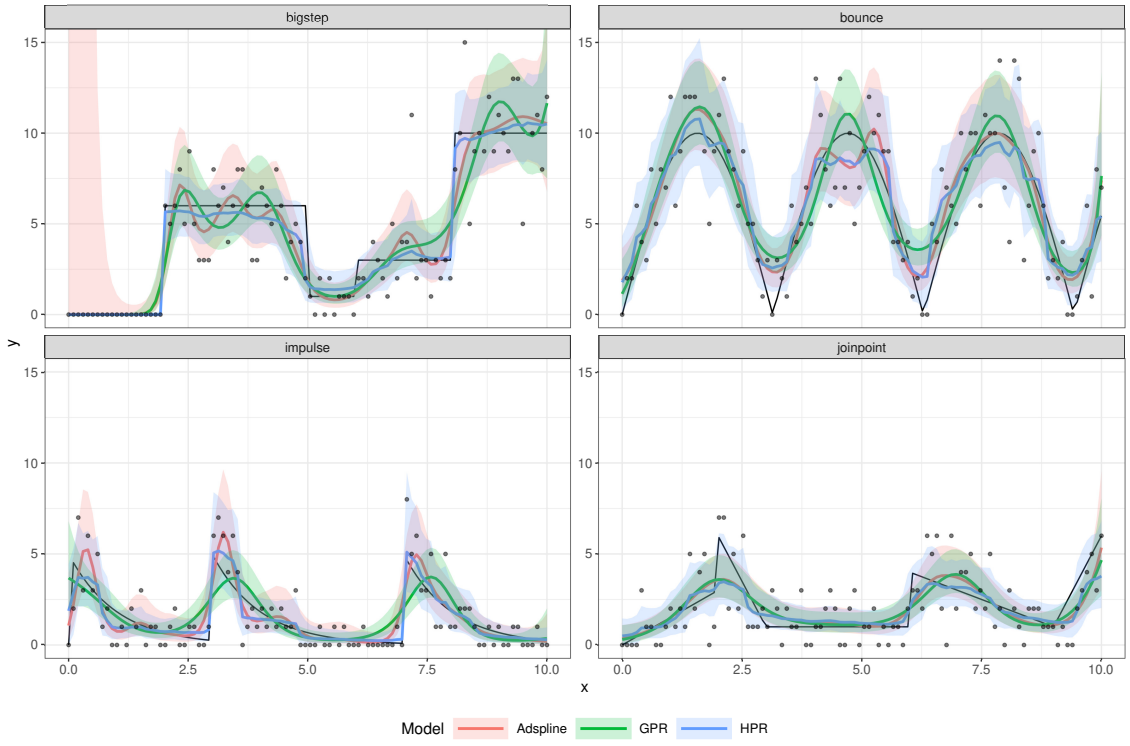

Figure A.1: Point estimates and 95% credible/confidence intervals for horseshoe process regression (HPR), adaptive splines (Adspline), and Gaussian process regression (GPR) for count data. Each sample dataset has n = 100.

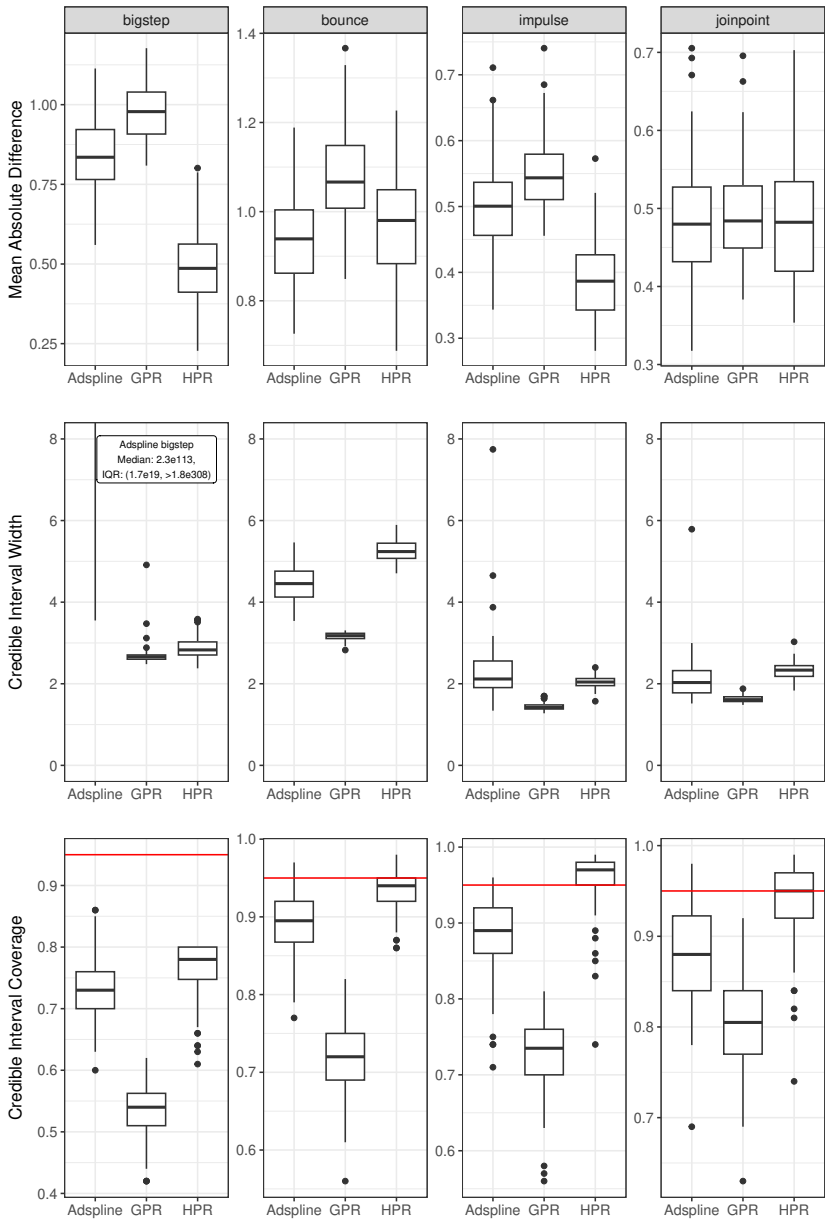

Figure A.2: Horseshoe process regression (HPR) simulation results for count data, based on 100 replicates on four data-generating scenarios, each with n = 100. Comparison methods were adaptive splines (Adspline) and Gaussian process regression (GPR). The top row gives performance for mean absolute difference (smaller is better); the second row gives performance for credible/confidence interval width; the third row gives performance for credible/confidence interval coverage (0.95 is nominal and given as a red line). Each column is for one data-generating scenario; sample datasets are given in Figure A.1.

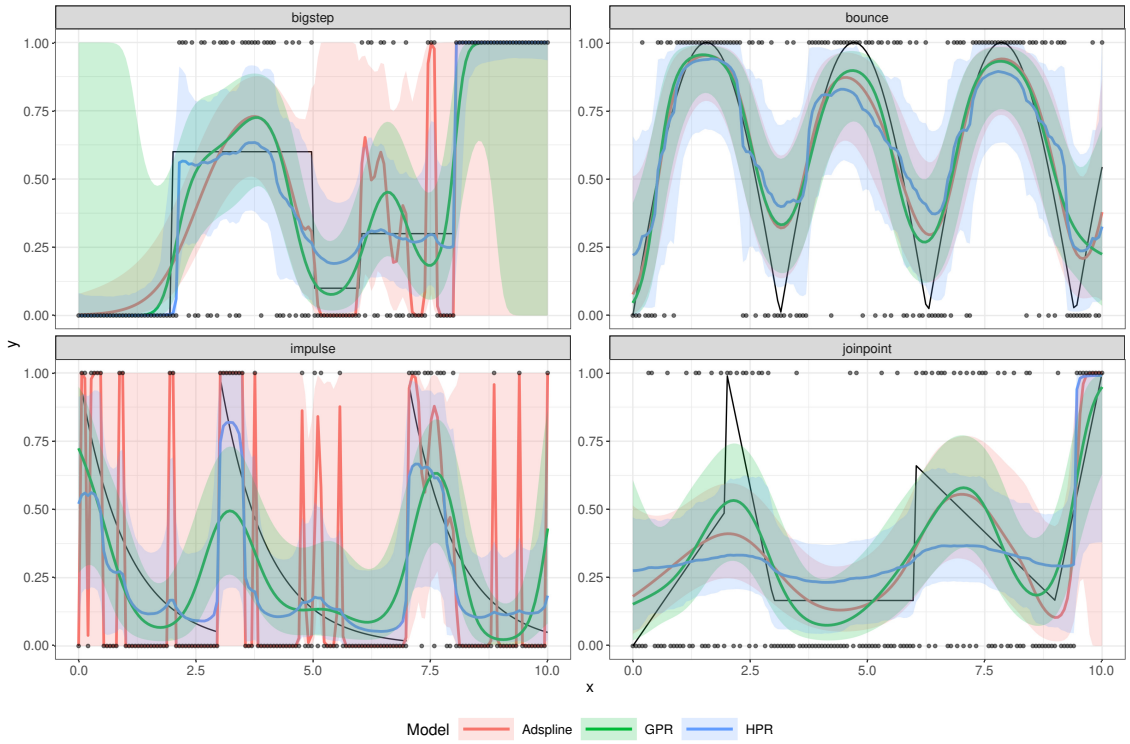

Figure A.3: Point estimates and 95% credible/confidence intervals for horseshoe process regression (HPR), adaptive splines (Adspline), and Gaussian process regression (GPR) for binary outcomes. Each sample dataset has n = 150.

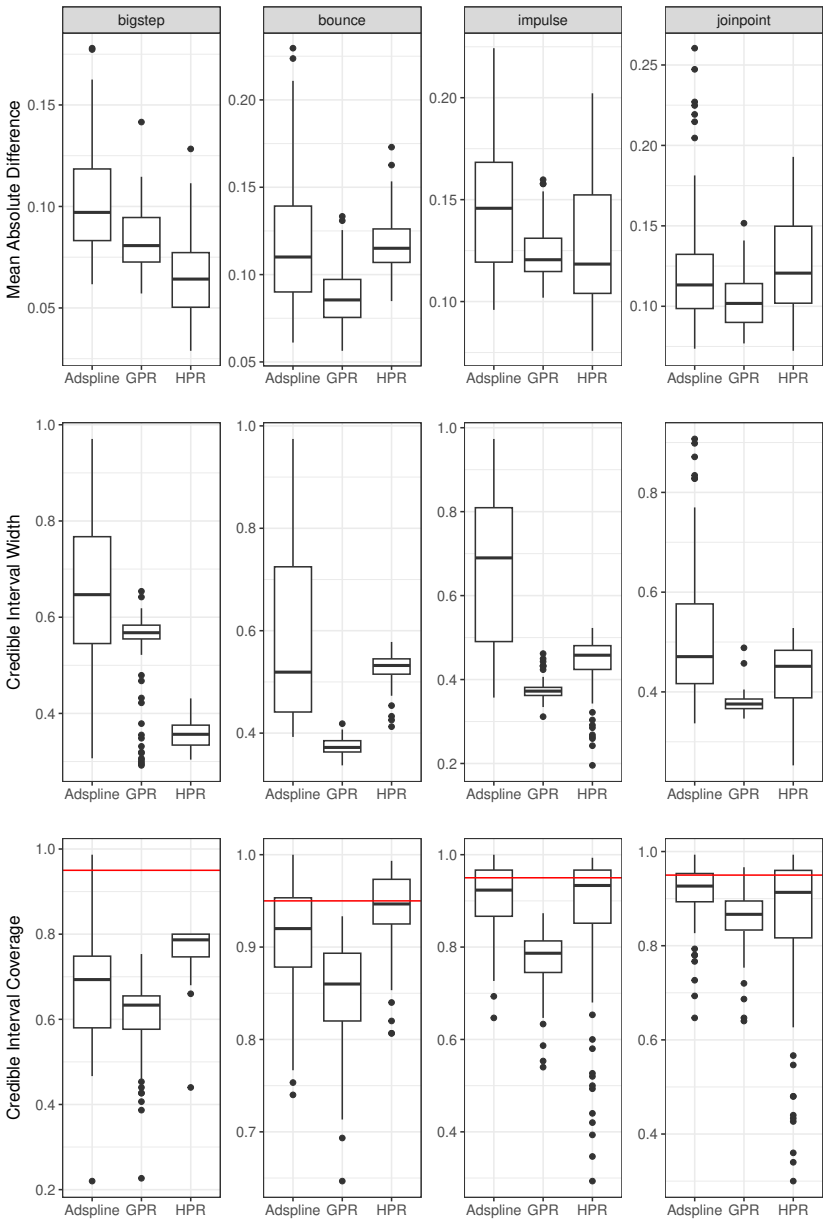

Figure A.4: Horseshoe process regression (HPR) simulation results for binary outcomes, based on 100 replicates on four data-generating scenarios, each with n = 150. Comparison methods were Gaussian process regression (GPR) and adaptive splines (Adspline). The top row gives performance for mean absolute difference (smaller is better); the second row gives performance for credible/confidence interval width; the third row gives performance for credible/confidence interval coverage (0.95 is nominal and given as a red line). Each column is for one data-generating scenario; sample datasets are given in Figure A.3.

Several findings are worthy of further note. First, Adspline had a great deal of trouble for count outcomes in the bigstep scenario, with extremely large credible intervals and highly erratic fits (Figures A.1 and A.2). It does not seem suited to this setting. GPR and HPR returned more sensible fits, with HPR excelling in the bigstep scenario. However, all methods struggled to maintain nominal credible interval coverage in the bigstep scenario; HPR was closest to nominal of the comparison methods (Figures A.2 and A.4). For binary outcomes, there was evidence that HPR was overshrinking, particularly in the impulse and joinpoint scenarios (Figure A.3). This might be improved with a higher value of c, the scale on the prior of τ.

APPENDIX B

Additional Simulations on HPR's Pointwise Performance

In addition to the aggregate outcomes presented in Section 2.4.1 and in Appendix A, we also considered pointwise performance (not summed across all observed datapoints):

1. Pointwise bias: $f(x_i) - \hat{f}(x_i)$.
2. Pointwise credible/confidence interval width: $\hat{f}(x_i)^{0.975} - \hat{f}(x_i)^{0.025}$.
3. Pointwise credible/confidence interval coverage: $I(\hat{f}(x_i)^{0.025} \leq f(x_i) \leq \hat{f}(x_i)^{0.975})$.

Results are given in Figures B.1-B.3. Pointwise performance was generally similar to aggregate performance. All methods showed the worst performance at the location of the abrupt jumps, with increased mean absolute difference and decreased coverage. Although HPR also showed this worsened performance, it did better at capturing the jumps than the other comparison methods. In the case of binary outcomes, HPR was likely to return flat-line fits, again suggesting that a larger value of c, the prior scale on τ, may be advisable for binary outcomes (Figure B.3).

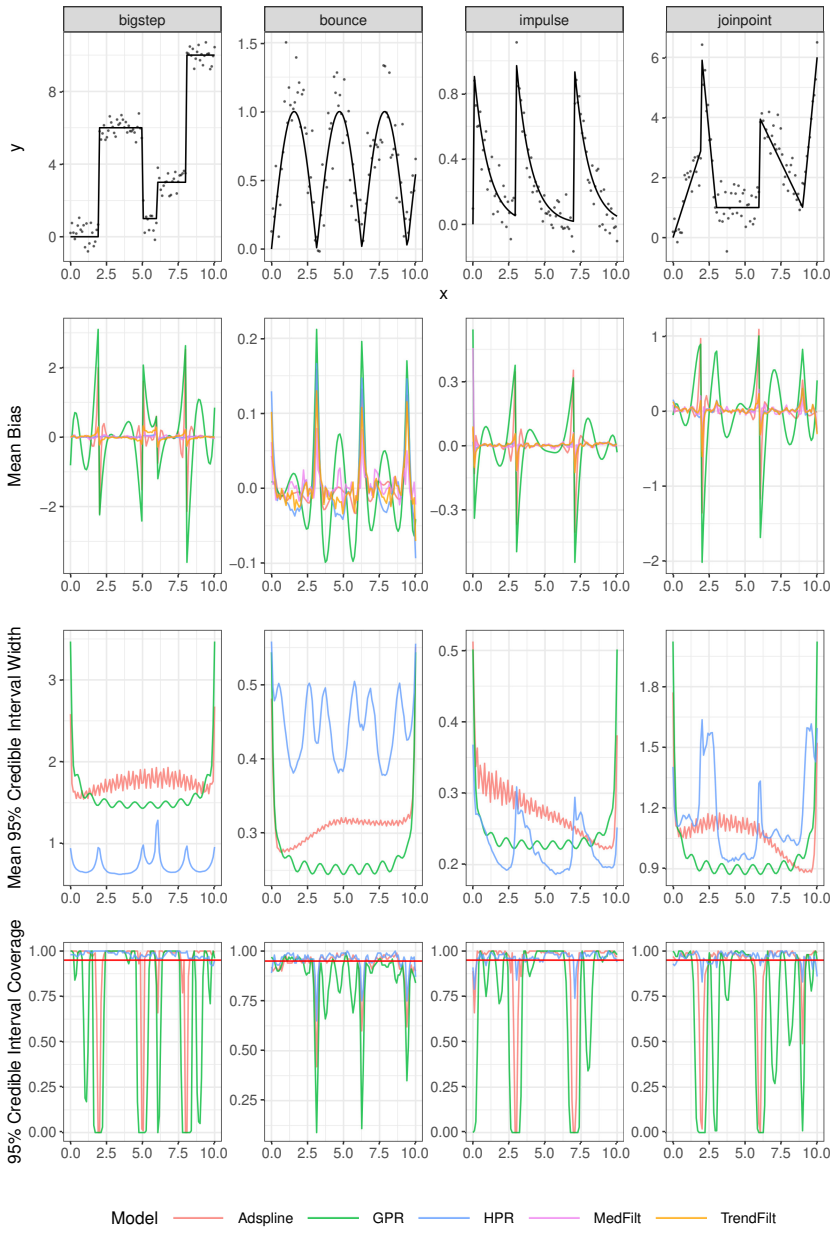

Figure B.1: Pointwise simulation results for continuous outcomes, based on 100 replicates on four data-generating scenarios, each with n = 100. Comparison methods were adaptive splines (Adspline), Gaussian process regression (GPR), median filter (MedFilt), and the penalized trend filter (TrendFilt). The top row gives a sample dataset and the true trajectory. The second row gives performance for mean bias, averaged over the 100 replicates at each point (smaller is better); the third row gives credible interval width, averaged over the 100 replicates at each point; the fourth row gives performance for credible interval coverage at each point (0.95 is nominal and given as a red line). Each column is for one data-generating scenario.

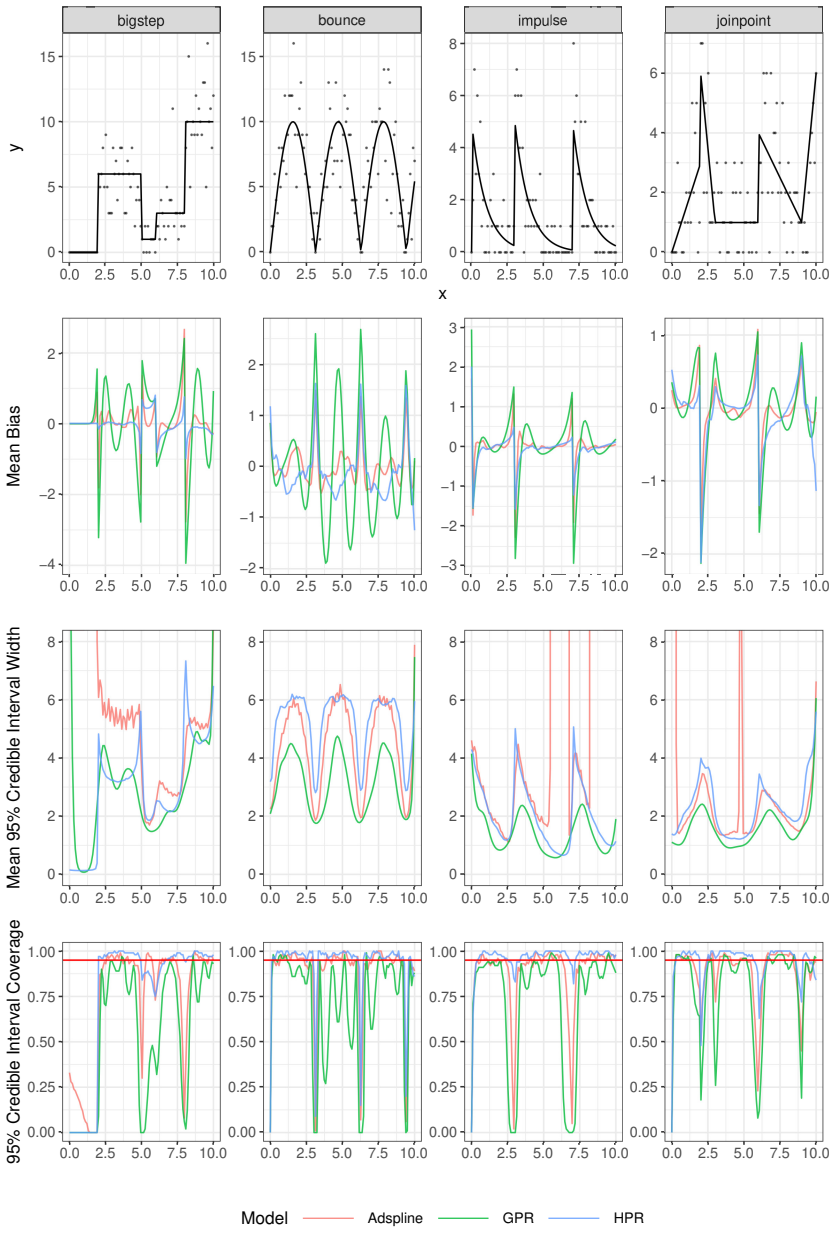

Figure B.2: Pointwise simulation results for count data, based on 100 replicates on four data-generating scenarios, each with n = 100. Comparison methods were adaptive splines (Adspline) and Gaussian process regression (GPR). The top row gives a sample dataset and the true trajectory. The second row gives performance for mean bias, averaged over the 100 replicates at each point (smaller is better); the third row gives credible interval width, averaged over the 100 replicates at each point; the fourth row gives performance for credible interval coverage at each point (0.95 is nominal and given as a red line). Each column is for one data-generating scenario.

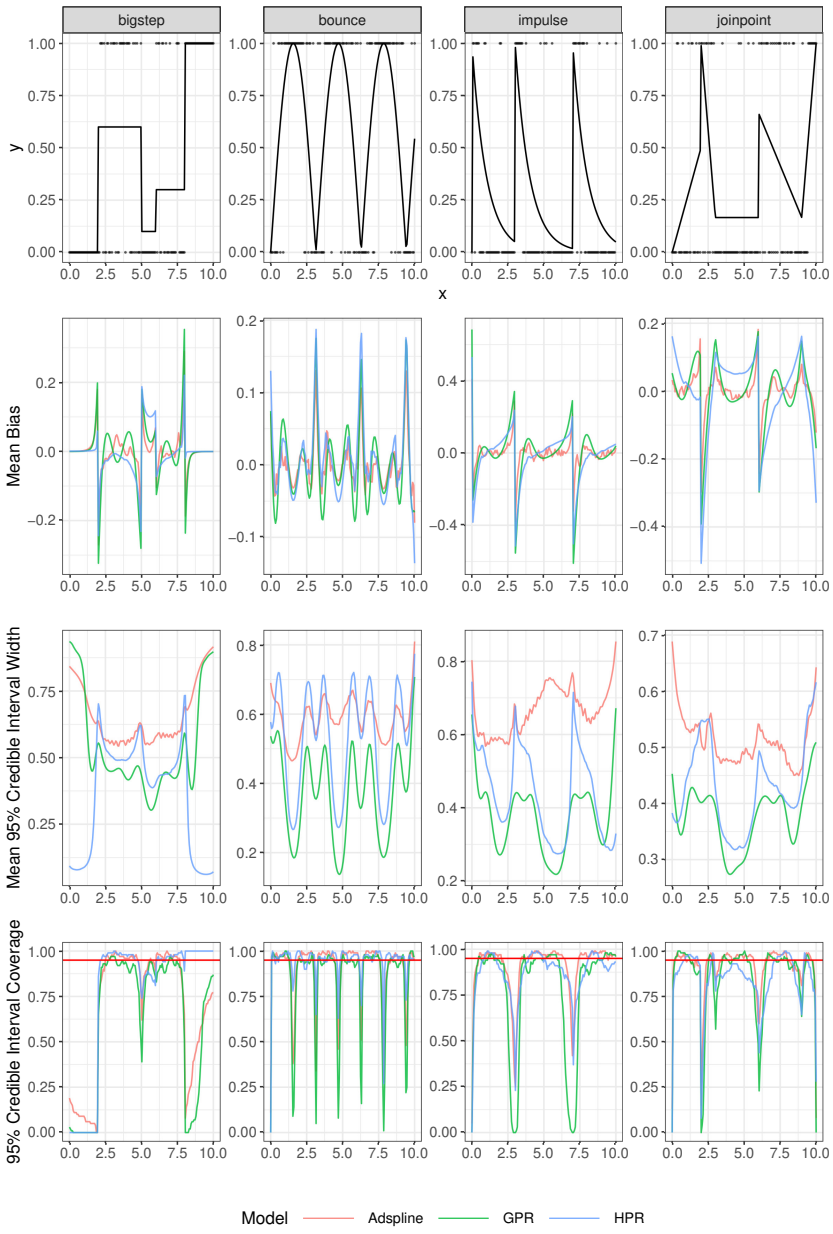

Figure B.3: Pointwise simulation results for binary outcomes, based on 100 replicates on four data-generating scenarios, each with n = 150. Comparison methods were adaptive splines (Adspline) and Gaussian process regression (GPR). The top row gives a sample dataset and the true trajectory. The second row gives performance for mean bias, averaged over the 100 replicates at each point (smaller is better); the third row gives credible interval width, averaged over the 100 replicates at each point; the fourth row gives performance for credible interval coverage at each point (0.95 is nominal and given as a red line). Each column is for one data-generating scenario.

APPENDIX C

Different Types of Monotonic-Constrained HPR

We proposed using the absolute value function to constrain the association between x and y to be monotonic. To do so, we modify our horseshoe process regression (HPR) to be:

$$g(E(y_i)) = f_j = \alpha + \sum_{k=1}^{j} |h_k| \tag{C.1}$$

Although we chose to use the absolute value, note that many other functions could be used to impose a monotonicity constraint—any function that transforms from the reals to the positive reals would be able to achieve this goal. We also considered using the $\exp$ function in lieu of the absolute value function. We present simulation results here to justify the choice of absolute value function.

We generated 100 evenly spaced predictors values between 0 and 10. Then, we considered two monotonic data-generating scenarios:

1. bigstep: $f(x) = 0 * I(x \leq 2) + 6 * I(2 < x \leq 5) + 10 * I(5 < x \leq 6) + 12 * I(6 < x \leq 8) + 20 * I(x > 8)$.

2. smooth: $f(x) = \log(\frac{x/11+0.01}{1-(x/11+0.01)})$.

We simulated Gaussian noise around each true curve with standard deviation of 1 for the bigstep scenario and standard deviation of 0.5 for the smooth scenario. Then, we compared the unconstrained HPR to a HPR constrained using the absolute value function (HPR_abs) and a HPR constrained using exponentiation (HPR_exp). We considered both estimation performance and computational metrics.

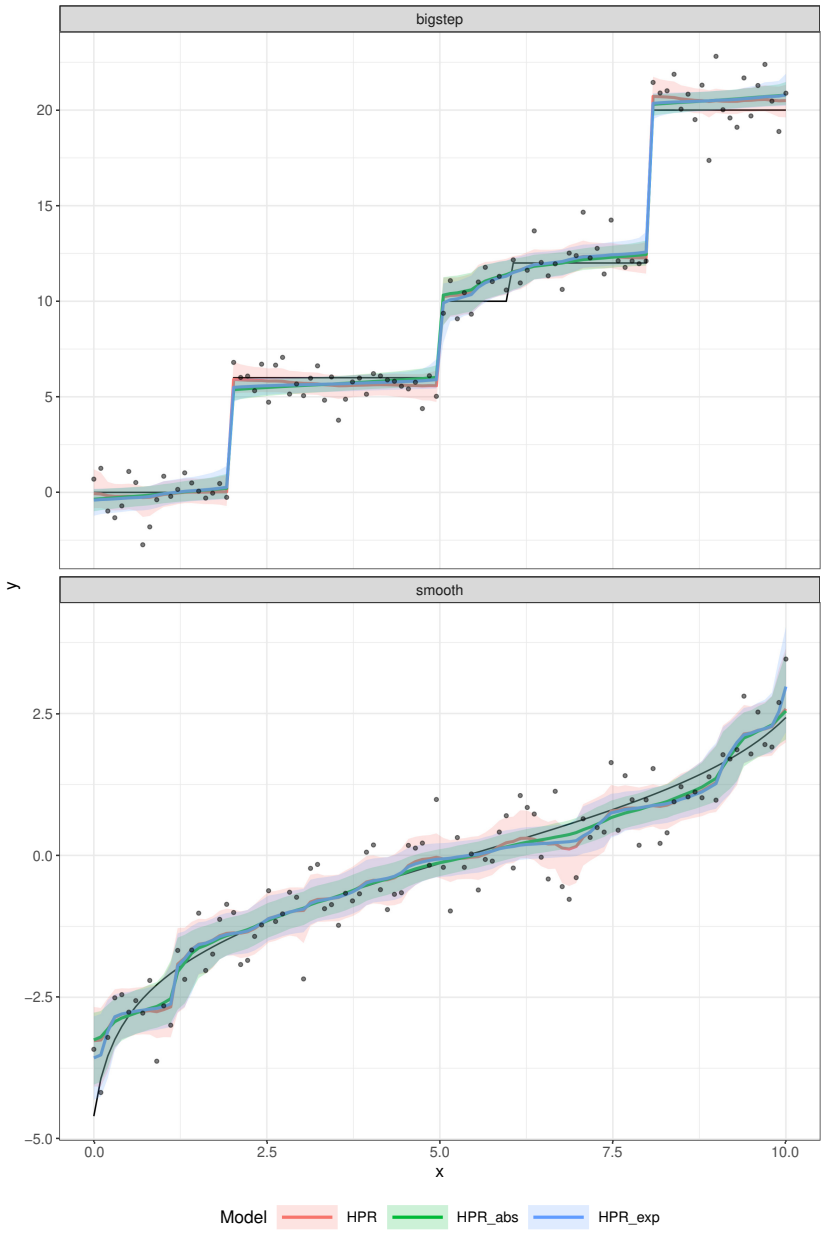

Figure C.1: Point estimates and 95% credible/confidence intervals for a HPR with no constraint (HPR), a constrained HPR via absolute value (HPR_abs), and a constrained HPR via exponentiation (HPR_exp) for continuous outcomes. Each sample dataset has n = 100.

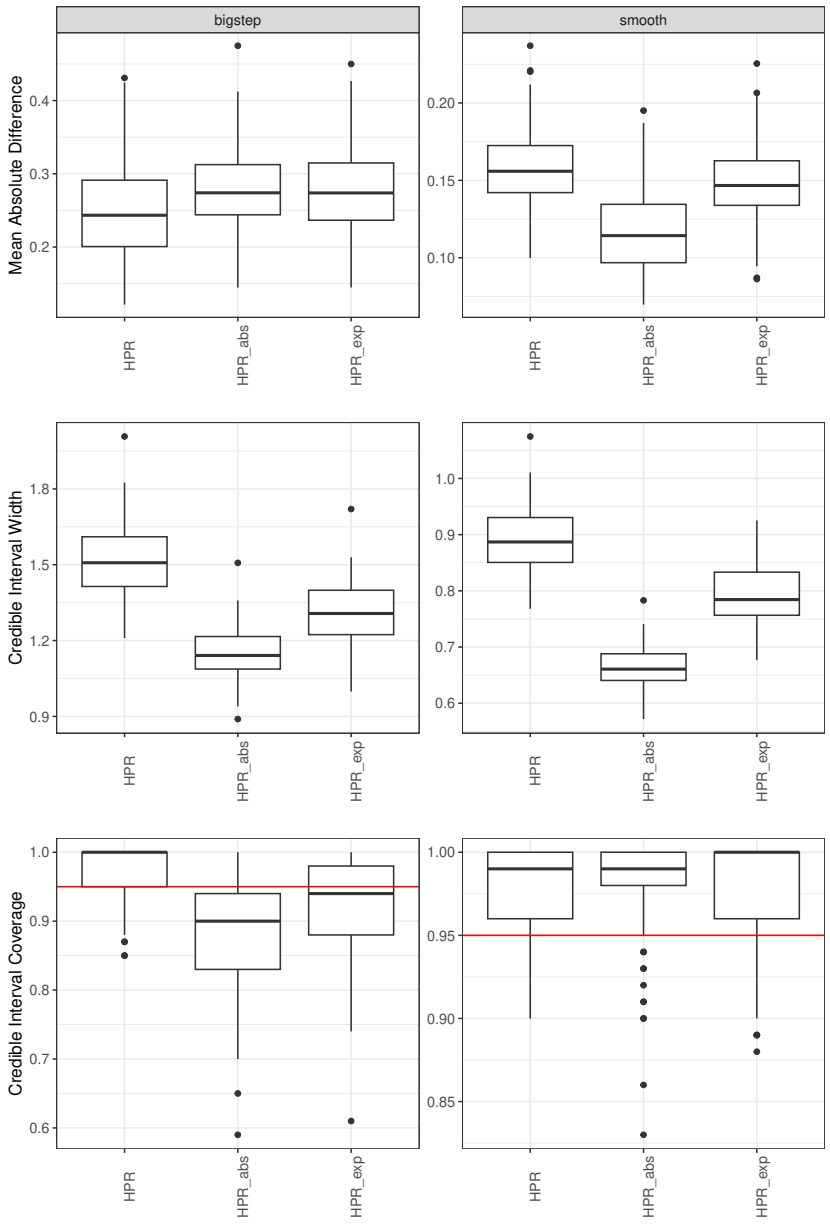

Figure C.2: Simulation results for a horseshoe process regression (HPR) constrained to be monotonic increasing, based on 100 replicates on two data-generating scenarios, each with n = 100. Comparison methods were a HPR with no constraint (HPR), a constrained HPR via absolute value (HPR_abs), and a constrained HPR via exponentiation (HPR_exp). The top row gives performance for mean absolute difference (smaller is better); the second row gives performance for credible interval width; the third row gives performance for credible interval coverage (0.95 is nominal and given as a red line). Each column is for one data-generating scenario; sample datasets can be seen in Figure C.1.

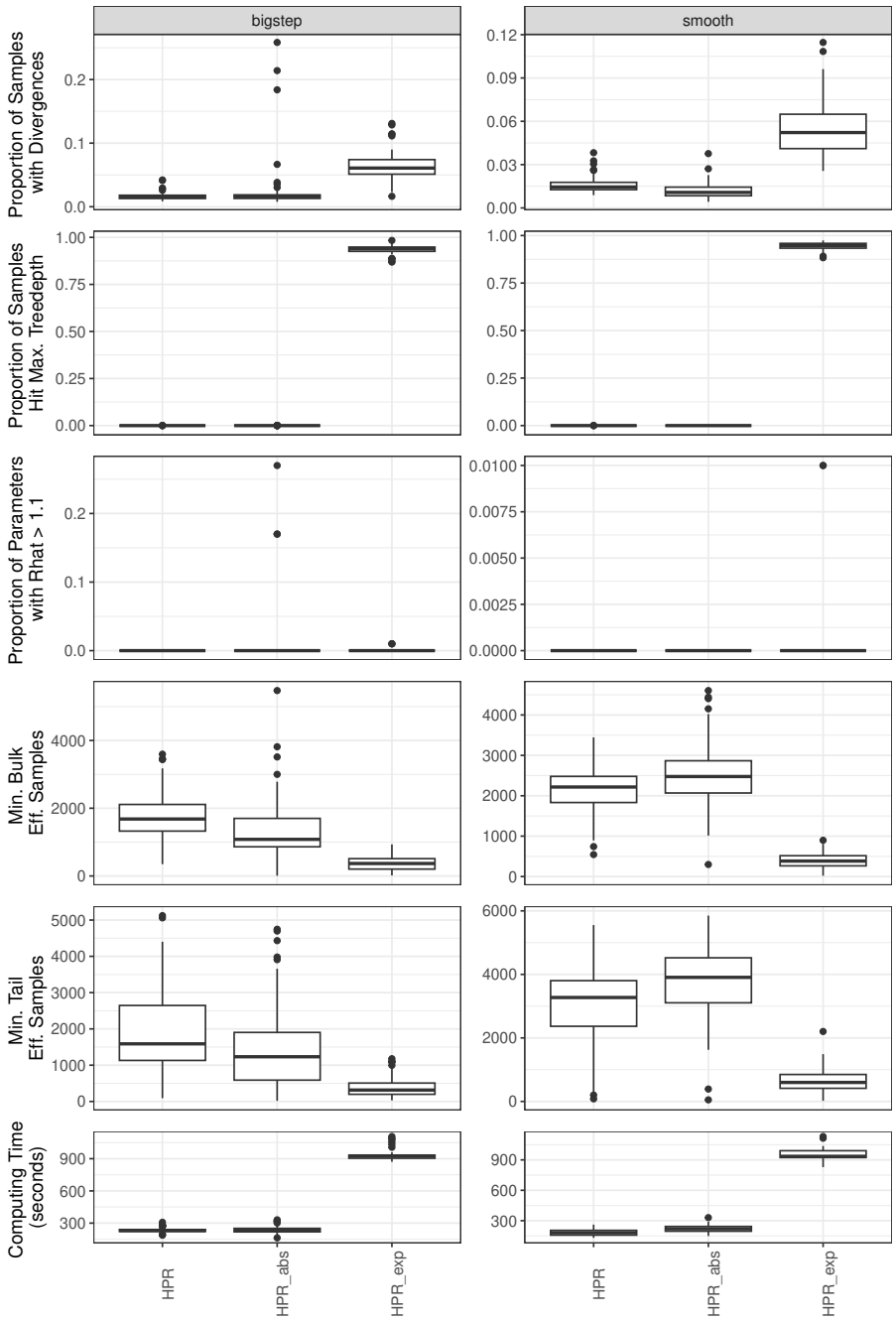

Figure C.3: Computational performance of a horseshoe process regression (HPR) constrained to be monotonic increasing using either no constraint (HPR), absolute value (HPR_abs), or exponentiation (HPR_exp) based on 100 replicates in two data-generating scenarios for Gaussian outcomes. Smaller is better for all metrics except Min. Bulk. Eff. Samples and Min. Tail. Eff. Samples (the minimum effective sample size in the bulk and tails of the posterior, respectively). Each column is for one data-generating scenario.

Results are given in Figures C.1, C.2, and C.3. From these, we see that estimation performance of the absolute-value constrained HPR was superior in terms of mean absolute difference and credible interval width, although coverage was slightly lower than nominal for the bigstep scenario. However, the computational performance is the real justification for using the absolute value transformation—the exponentiation constraint returned a substantially higher proportion of divergences, max tree-depth warnings, and other nonconvergence signals—in addition to being substantially slower.

APPENDIX D

Additional Simulations on HPR Data Augmentation for Binary and Count Outcomes

Here, we explored the performance of our data augmentation scheme for binary and count outcomes. We only considered scenarios bigstep and bounce (described in Appendix A). We randomly sampled 100 unevenly spaced datapoints between 0 and 10 to be our observed x locations (150 datapoints for binary data). Then, we fit the HPR either 1) only using the observed datapoints, 2) augmented by a grid of datapoints at every 0.5 (roughly 20 augmented datapoints), and 3) augmented by a grid of datapoints at every 0.1 (roughly 100 augmented datapoints). We calculated the performance metrics described above separately for the observed datapoints and the augmented datapoints, to see if predictions at the observed datapoints changed depending on the number of gridpoints, and if predictions at the augmented datapoints were fairly accurate.

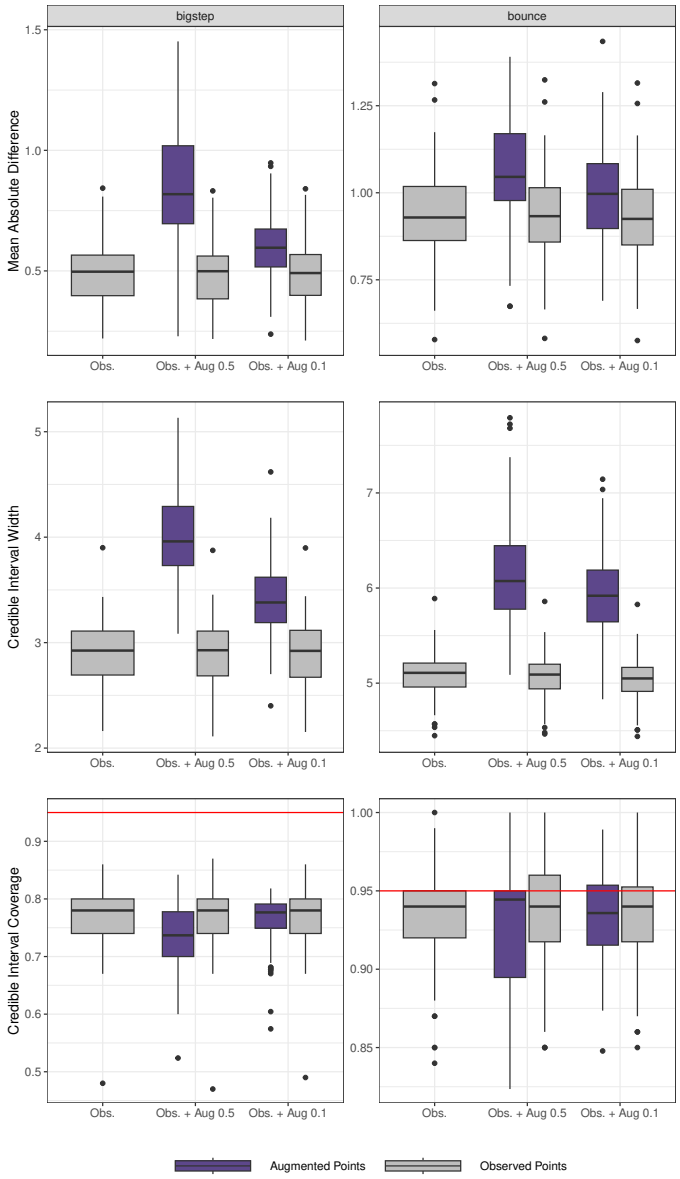

Figure D.1: Horseshoe process regression (HPR) data augmentation simulation results for count outcomes, based on 100 replicates on two data-generating scenarios. We compared a HPR calculated only at n = 100 observed points to a HPR with augmentation points at a grid of every 0.5 and a HPR with augmentation points at a grid of every 0.1 (from 0 to 10). The top row gives performance for mean absolute difference calculated at both the observed and augmented points (smaller is better); the second row gives performance for credible interval width calculated at both the observed and augmented points; the third row gives credible interval coverage calculated at both the observed and augmented points (0.95 is nominal and marked as a red line). Performance at observed points and augmented points are displayed separately. Each column is for one data-generating scenario.

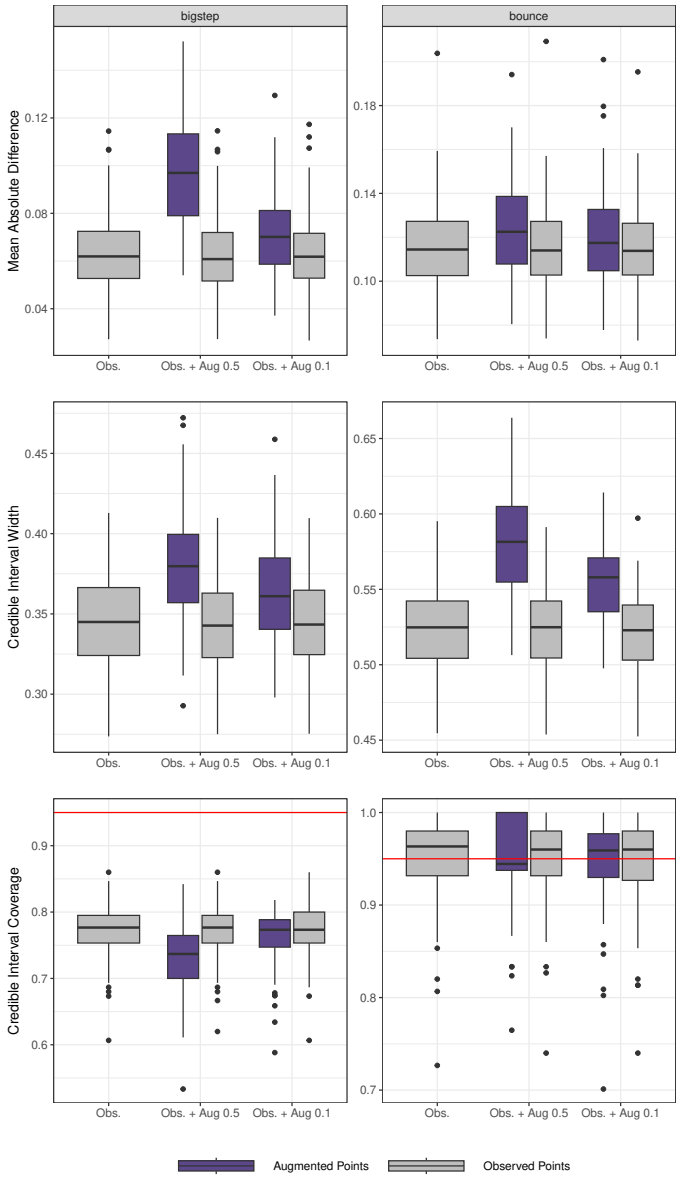

Figure D.2: Horseshoe process regression (HPR) data augmentation simulation results for binary outcomes, based on 100 replicates on two data-generating scenarios. We compared a HPR calculated only at n = 150 observed points to a HPR with augmentation points at a grid of every 0.5 and a HPR with augmentation points at a grid of every 0.1 (from 0 to 10). The top row gives performance for mean absolute difference calculated at both the observed and augmented points (smaller is better); the second row gives performance for credible interval width calculated at both the observed and augmented points; the third row gives credible interval coverage calculated at both the observed and augmented points (0.95 is nominal and marked as a red line). Performance at observed points and augmented points are displayed separately. Each column is for one data-generating scenario.

Results of the augmentation scheme for binary and count outcomes largely resembled the performance for continuous outcomes presented in Section 2.4.2. Performance at observed datapoints did not change with varying numbers of augmentation datapoints. However, aggregate performance at augmentation datapoints improved with increased grid density, with reduced mean absolute difference and narrower credible intervals. This "improved performance" is somewhat misleading, because in the data generating schemes considered here—which do not feature an extremely large number of abrupt changes—the augmentation scheme will do better with more augmentation points as a matter of probability. With more augmentation points, the probability that an augmentation point is placed at the location of an abrupt jump is reduced, artificially boosting aggregate performance. In general, the model fit estimated at an augmentation point was an interpolation of the two nearest observed datapoints, with credible intervals that reflected the increased uncertainty.

APPENDIX E

Simulation Results for HPR Partial Linear Model

We assessed the performance of our horseshoe process regression (HPR) partial linear model. We simulated five covariates. The first three of these covariates were simulated from a multivariate normal distribution with mean vector $(67, 0, 130)$ and standard deviation vector $(7, 1, 20)$ with correlation of 0.35:

$$\begin{pmatrix} X_1 \\ X_2^* \\ X_3 \end{pmatrix} \sim MVN\left(\begin{pmatrix} 67 \\ 0 \\ 130 \end{pmatrix}, \begin{pmatrix} 7^2 & 7*1*0.35 & 7*20*0.35 \\ 7*1*0.35 & 1^2 & 1*20*0.35 \\ 7*20*0.35 & 1*20*0.35 & 20^2 \end{pmatrix}\right)$$

The second of these covariates was made into a binary variable by splitting it at 24: $X_2 = I(X_2^* > 24)$.

The remaining 2 covariates were simulated independently from a multivariate normal distribution with mean vector $(80, 45)$, standard deviation vector $(20, 12)$, and correlation of 0.2:

$$\begin{pmatrix} X_4 \\ X_5 \end{pmatrix} \sim MVN\left(\begin{pmatrix} 80 \\ 45 \end{pmatrix}, \begin{pmatrix} 20^2 & 20*12*0.2 \\ 20*12*0.2 & 12^2 \end{pmatrix}\right)$$

For continuous outcomes, the first 4 of these covariates were assumed to have a linear relationship with the outcome y, with coefficient vector $(\beta_1, \beta_2, \beta_3, \beta_4) = (0, 5, 0.05, 0.1)$. We considered two different functional forms for the fifth covariate:

1. bigstep: $f(x_5) = 0 * I(x_5 < 35) + 5 * I(35 \leq x_5 < 55) + 6 * I(55 \leq x_5 < 65) + 8 * I(65 \leq x_5 < 80) + 10 * I(x_5 \geq 80)$

2. smooth: $f(x_5) = \frac{x_5^2}{100}$

Thus $E(y) = 0 * X_1 + 5 * X_2 + 0.05 * X_3 + 0.1 * X_4 + f(X_5)$. We then generated observations

for 100 subjects, and simulated Gaussian error with $\sigma = 0.5$.

For count data, the first 4 of these covariates were assumed to have a linear relationship with the outcome $\log(E(y))$, with coefficient vector $(\beta_1, \beta_2, \beta_3, \beta_4) = (0, 0.08, 0.01, 0.01)$. We considered two different functional forms for the fifth covariate:

1. bigstep: $f(x_5) = 0*I(x_5 < 35)+0.5*I(35 \leq x_5 < 55)+0.8*I(55 \leq x_5 < 65)+1*I(65 \leq x_5 < 80) + 1.3 * I(x_5 \geq 80)$

2. smooth: $f(x_5) = |sin(x_5)|$

Thus $\log(E(y)) = 0 * X_1 + 0.08 * X_2 + 0.01 * X_3 + 0.01 * X_4 + f(X_5)$. We simulated 100 datapoints as described above, calculated $\log(E(y))$, transformed to the $E(y)$ scale, and randomly sampled 100 observed outcomes from a Poisson distribution.

For binary data, we used coefficient vector $(\beta_1, \beta_2, \beta_3, \beta_4) = (0, -0.5, -0.05, 0.1)$ for the linear relationship with outcome $\log(E(y)/(1 - E(y)))$. We considered two different functional forms for the fifth covariate:

1. bigstep: $f(x_5) = -5*I(x_5 < 35)+0*I(35 \leq x_5 < 55)+1*I(55 \leq x_5 < 65)+5*I(65 \leq x_5 < 80) + 10 * I(x_5 \geq 80)$

2. smooth: $f(x_5) = sin(x_5)$

Thus $\log(E(y)/(1-E(y))) = 0*X_1-0.5*X_2-0.05*X_3+0.1*X_4+f(X_5)$. We simulated 150 datapoints as described above, calculated $\log(E(y)/(1 - E(y)))$, transformed to the $E(y)$ scale, and randomly sampled 150 observed outcomes from a Bernoulli distribution.

We compared HPR to Gaussian process regression (GPR) and an adaptive spline model (Adspline). We did not consider the median filter (MedFilt) or trend filter (TrendFilt) because these methods are not implemented for additional linear covariates. We assessed performance using the mean absolute difference between true $E(y)$ and estimated $\hat{E}(y)$, averaged over all datapoints, and width and coverage of the 95% credible intervals of $E(y)$ averaged over all datapoints. We also considered bias and coverage for the coefficients for the linear predictors.

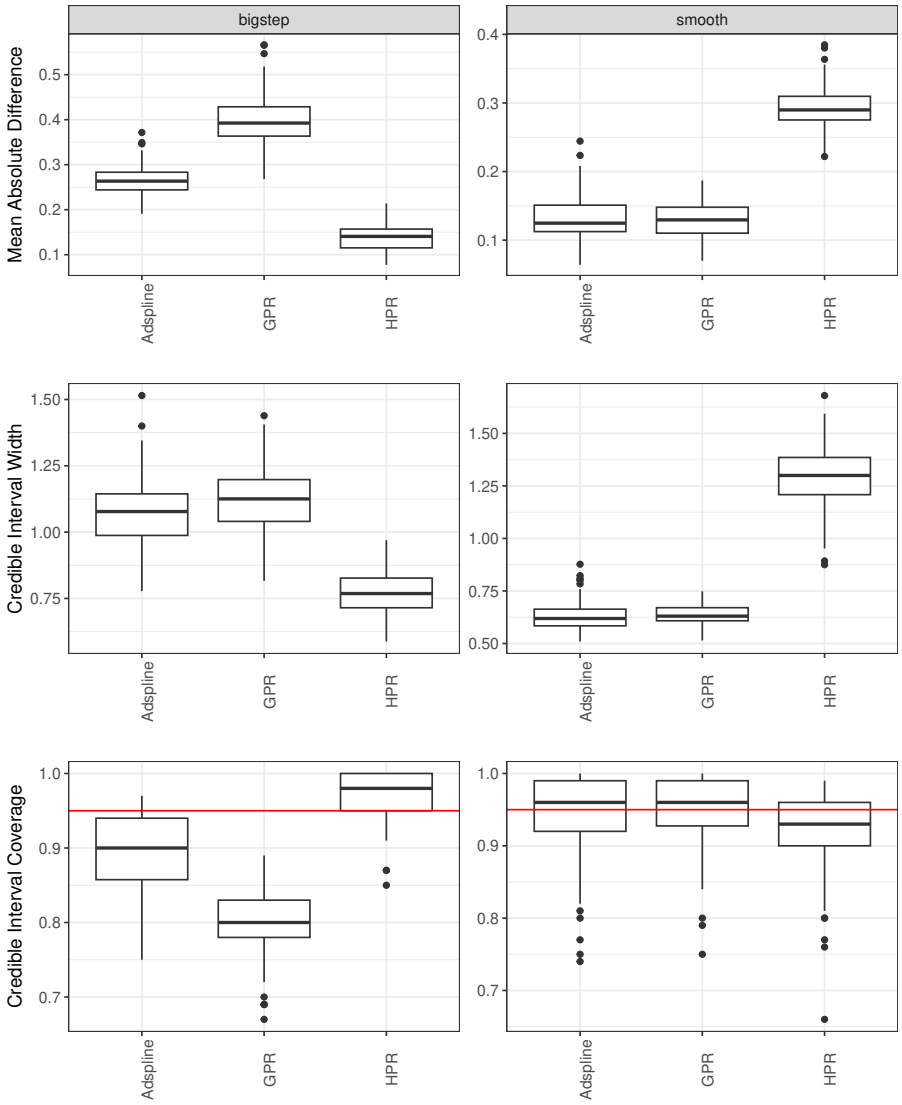

Figure E.1: Performance of a horseshoe process regression (HPR) partial linear model for estimating continuous outcomes, based on 100 replicates on two data-generating scenarios with n = 100. Comparison methods were Gaussian process regression (GPR) and adaptive splines (Adspline). The top row gives performance for mean absolute difference between the true outcome $E(y)$ and estimated outcome $\hat{E}(y)$ (smaller is better); the second row gives 95% credible interval width around $\hat{E}(y)$; the third row gives performance for credible interval coverage of $E(y)$ (0.95 is nominal and marked as a red line). Each column is for one data-generating scenario.

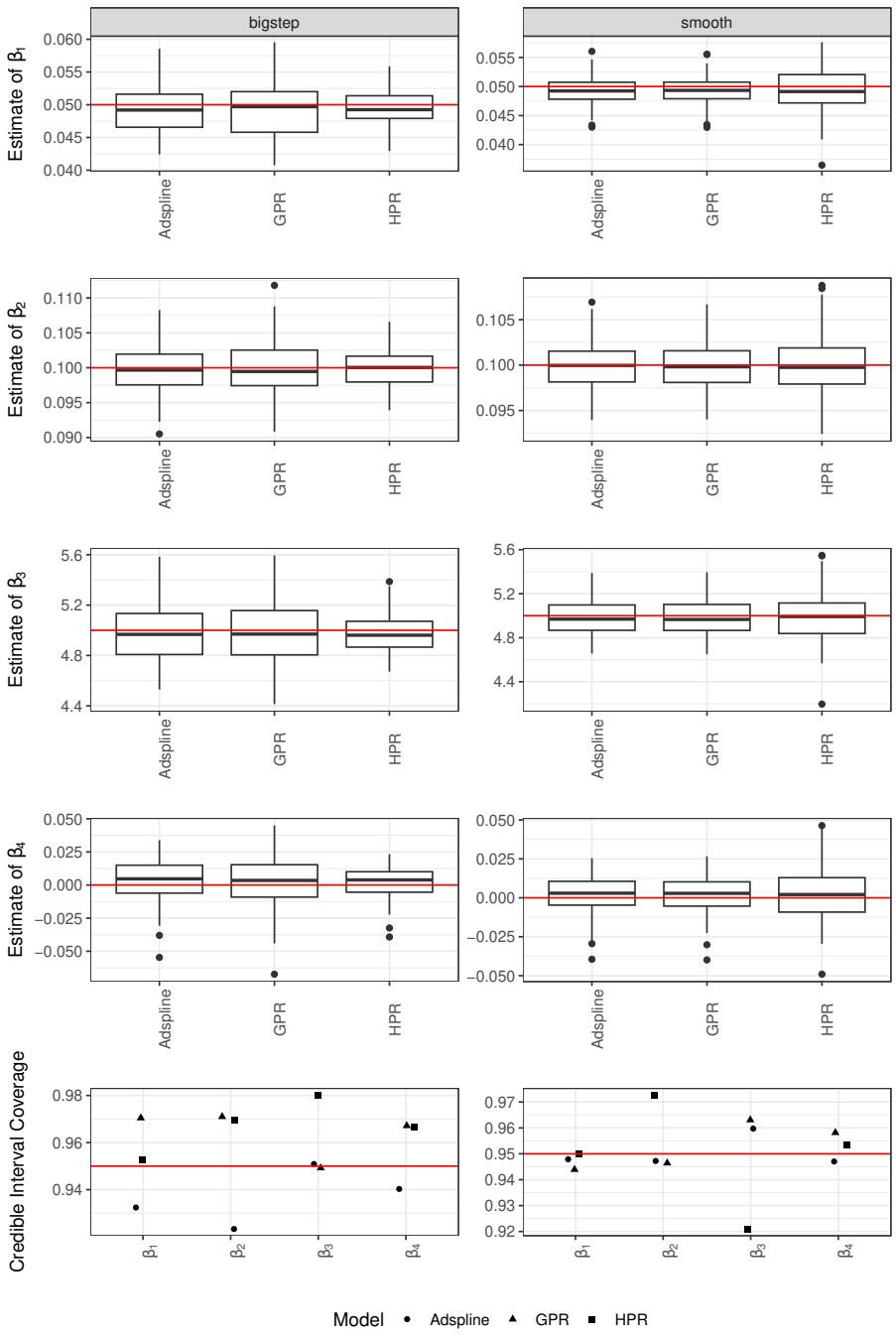

Figure E.2: Performance of a horseshoe process regression (HPR) Gaussian partial linear model for fitting four linear predictors, based on 100 replicates of two data-generating scenarios for the nonlinear predictor with n = 100. Comparison methods were Gaussian process regression (GPR) and adaptive splines (Adspline). The first four rows give the estimates of each of the coefficients, with the correct value given as a red line; the fifth row gives performance for credible interval coverage for all four coefficients (0.95 is nominal and marked as a red line). Each column is for one data-generating scenario.

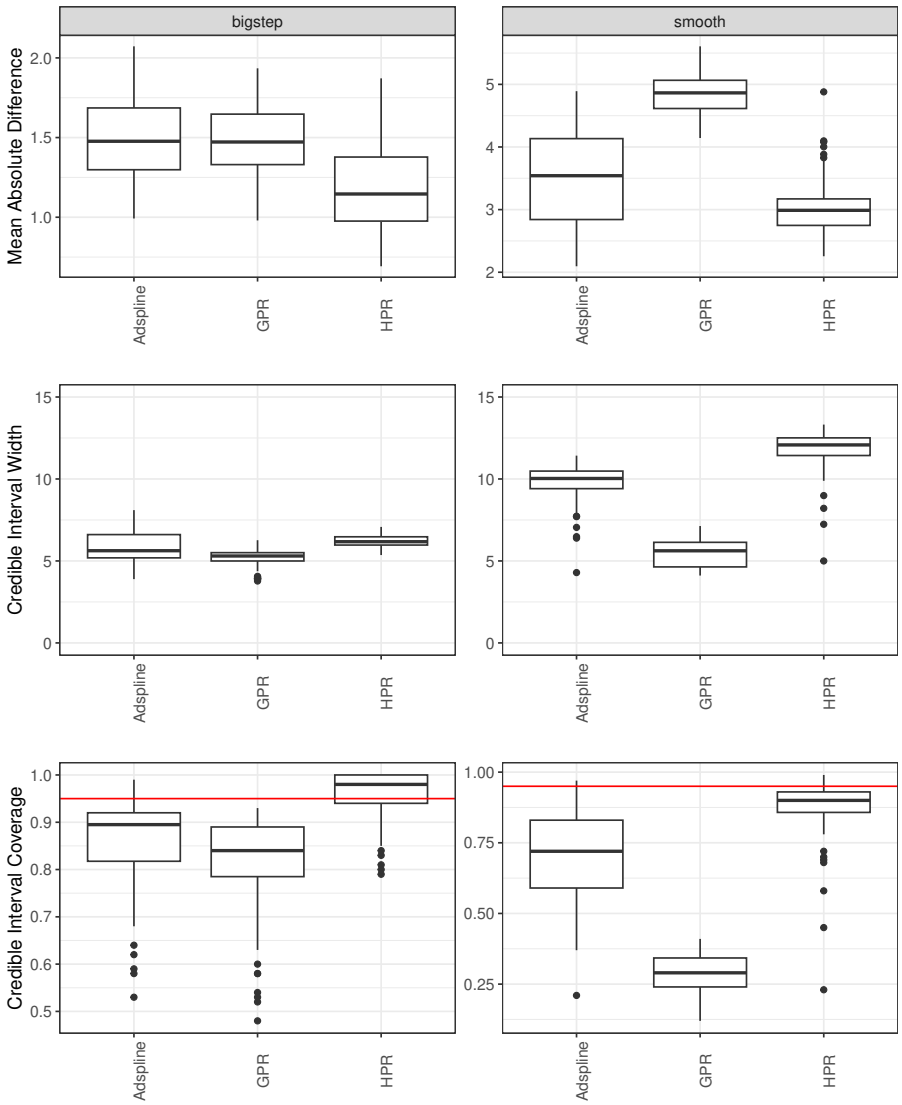

Figure E.3: Performance of a horseshoe process regression (HPR) partial linear model for estimating count outcomes, based on 100 replicates on two data-generating scenarios with n = 100. Comparison methods were Gaussian process regression (GPR) and adaptive splines (Adspline). The top row gives performance for mean absolute difference between the true outcome $E(y_i)$ and estimated outcome $\hat{E}(y_i)$ (smaller is better) averaged over 100 datapoints; the second row gives performance for credible interval width averaged over 100 datapoints; the third row gives coverage of $E(y_i)$, averaged over 100 datapoints (0.95 is nominal and marked as a red line). Each column is for one data-generating scenario. Note that the y-axis for credible interval width is truncated due to one extremely wide credible interval for HPR in the bigstep scenario.

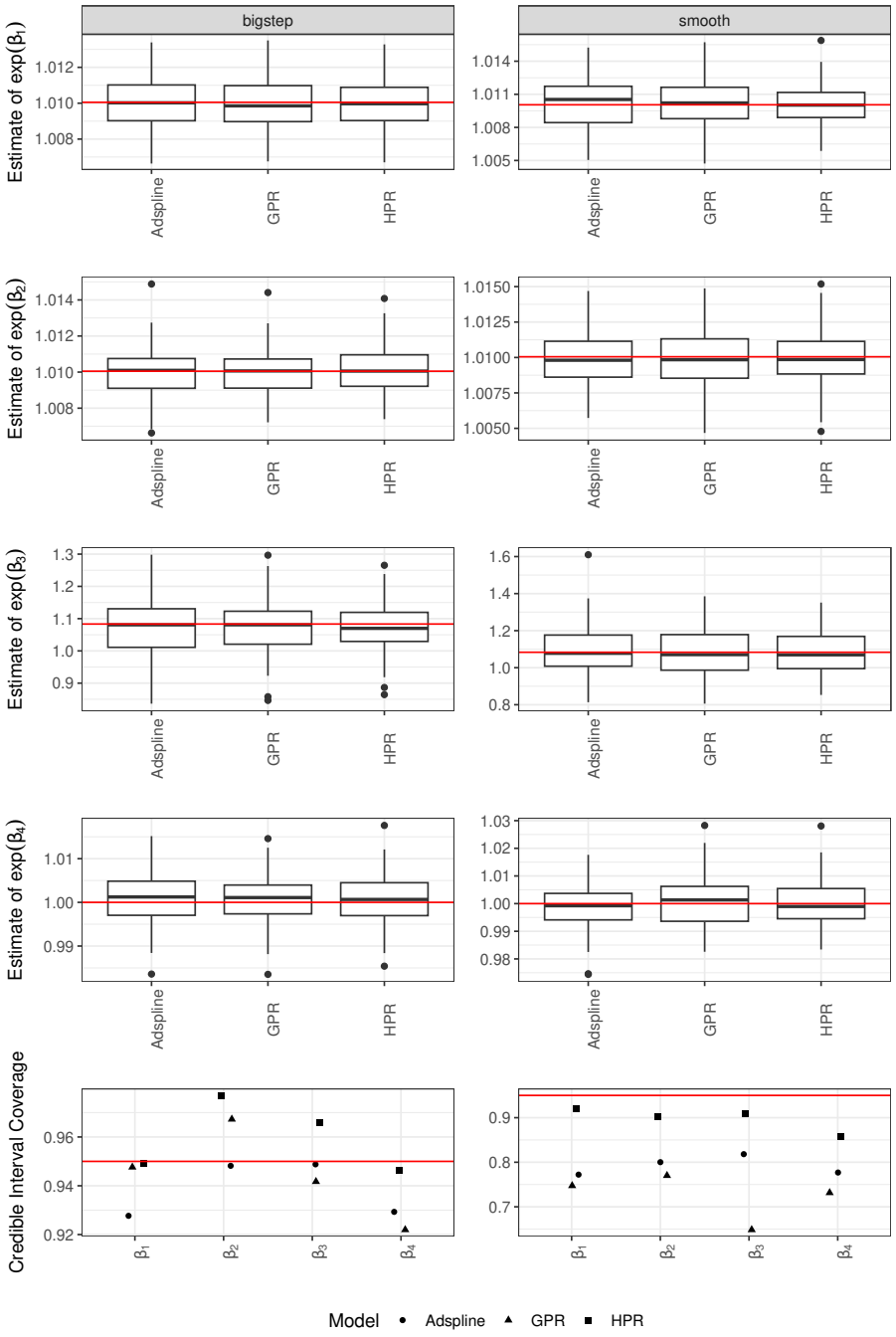

Figure E.4: Performance of a horseshoe process regression (HPR) Poisson partial linear model for fitting four linear predictors, based on 100 replicates on two data-generating scenarios for the nonlinear predictor with n = 100. Comparison methods were Gaussian process regression (GPR) and adaptive splines (Adspline). The first four rows give the estimates of each of the exponentiated coefficients, with the correct value given as a red line; the fifth row gives performance for credible interval coverage for all four exponentiated linear predictors (0.95 is nominal and marked as a red line). Each column is for one data-generating scenario.

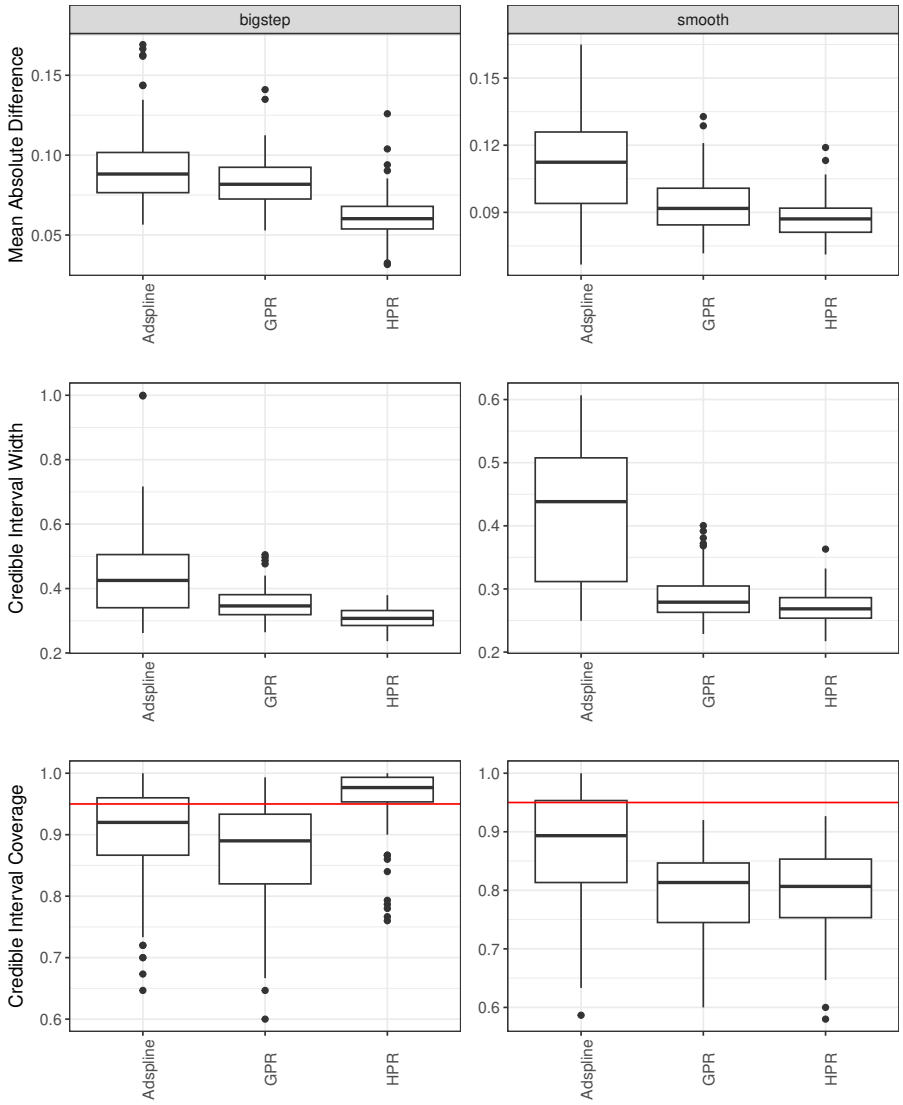

Figure E.5: Performance of a horseshoe process regression (HPR) partial linear model for estimating binary outcomes, based on 100 replicates on two data-generating scenarios with n = 150. Comparison methods were Gaussian process regression (GPR) and adaptive splines (Adspline). The top row gives performance for mean absolute difference between the true outcome $E(y_i)$ and estimated outcome $\hat{E}(y_i)$ (smaller is better) averaged over 150 datapoints; the second row gives performance for credible interval width averaged over the 150 datapoints; the third row gives coverage of $E(y_i)$, averaged over the 150 datapoints (0.95 is nominal and marked as a red line). Each column is for one data-generating scenario.

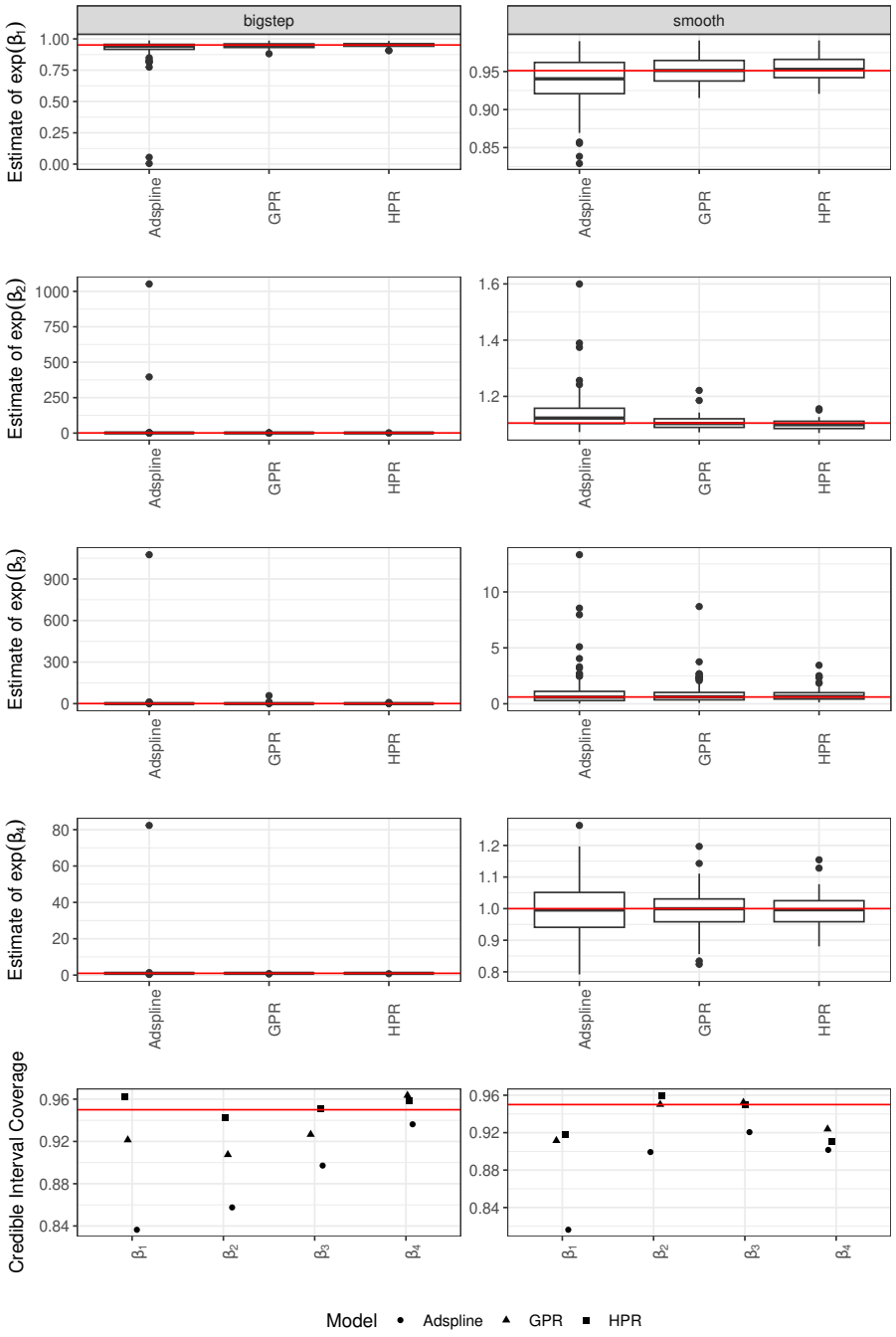

Figure E.6: Performance of a horseshoe process regression (HPR) Bernoulli partial linear model for fitting four linear predictors, based on 100 replicates on two data-generating scenarios for the nonlinear predictor with n = 150. Comparison methods were Gaussian process regression (GPR) and adaptive splines (Adspline). The first four rows give the estimates of each of the exponentiated coefficients, with the correct value given as a red line; the fifth row gives performance for credible interval coverage for all four exponentiated linear predictors (0.95 is nominal and marked as a red line). Each column is for one data-generating scenario.

Performance of the partial linear model was generally good. HPR offered substantially reduced mean absolute difference and credible interval width for the latent mean $\hat{E}(y_i)$ when $f(X_5)$ was a step function. When $f(X_5)$ was a smooth function, its performance was worse than the comparison methods for continuous outcomes (Figure E.1), although credible interval coverage was still very good. For binary and count outcomes, HPR consistently surpassed the comparison methods, even when $f(X_5)$ was a smooth function (Figures E.3 and E.5). The GPR particularly struggled for count outcomes. Regardless of the form of $f(X_5)$, performance for estimating the linear effects $(\beta_1, \beta_2, \beta_3, \beta_4)$ was good (Figures E.2, E.4, E.6).

APPENDIX F

HPR Computational Assessment

When conducting Bayesian modeling, it is important to assess model convergence and computational performance. We considered 6 computational metrics:

1. Proportion of MCMC samples that ended in a Hamiltonian Monte Carlo (HMC) divergence: This diagnostic is unique to Bayesian models fit using HMC. A divergence suggests that posterior sampling for that MCMC sample went "off the rails" and may be unreliable. In general, even a single divergence is cause for concern; however, as we discuss, in the case of HPR we think that small numbers of divergences ($<5\%$) may be unavoidable and do not negatively affect model performance. We would like this metric to be close to 0.

2. Proportion of MCMC samples that ended in a max treedepth warning: This diagnostic is also unique to Stan models and indicates whether the No-U-Turn-Sampler (NUTS) was frequently taking the maximum number of steps in Hamiltonian space without hitting a U-turn, suggesting that the step size was too small. In some cases this is indicative of model nonconvergence/poor posterior exploration. It often corresponds to slower computational times. We would like this metric to be close to 0.

3. Proportion of parameters with $\hat{R} > 1.1$: This diagnostic compares between- and within-chain estimates of each parameter of the model (all HPR models are fit with 4 chains). Larger values of $\hat{R}$ suggest that the chains have not mixed well and that posterior estimates may be unstable. Ideally $\hat{R} < 1.05$ or $\hat{R} < 1.1$; we used 1.1 here, and considered what proportion of the model's parameters have $\hat{R} > 1.1$. We used the $\hat{R}$ proposed by Vehtari et al. (2021) [77]. We would like this metric to be close to 0.

4. Minimum bulk effective sample size: This diagnostic uses rank-normalized draws to estimate the effective sample size in the bulk of the posterior, as described in Vehtari et al. (2021) [77]. It is calculated for each parameter; here, we present the minimum sample size across all parameters. We would like this metric to be large, and ideally larger than 400.

5. Minimum tail effective sample size: This diagnostic uses rank-normalized draws to estimate the effective sample size in the tails of the posterior, as described in Vehtari et al. (2021) [77]. It is calculated for each parameter; here, we present the minimum sample size across all parameters. We would like this metric to be large, and ideally larger than 400.

6. Computational time: This metric tells how long it took the model to fit, in seconds. We would like this metric to be small.

For more information on these diagnostics, please see the Stan reference manual [72].

Computational results for the basic HPR simulations are given in Figure F.1. Results for the data interpolation simulations are given in Figure F.2. Results for the partial linear model simulations are given in Figure F.3. Almost all of the models fitted in the simulation studies featured at least some HMC divergences. In most cases less than 5% of samples ended in a divergence. Max treedepth warnings occurred rarely. $\hat{R}$ diagnostics and effective sample sizes generally seemed adequate. Although slow compared to non-Bayesian methods, computation time was generally quite reasonable, with most models finishing in less than 5 minutes.

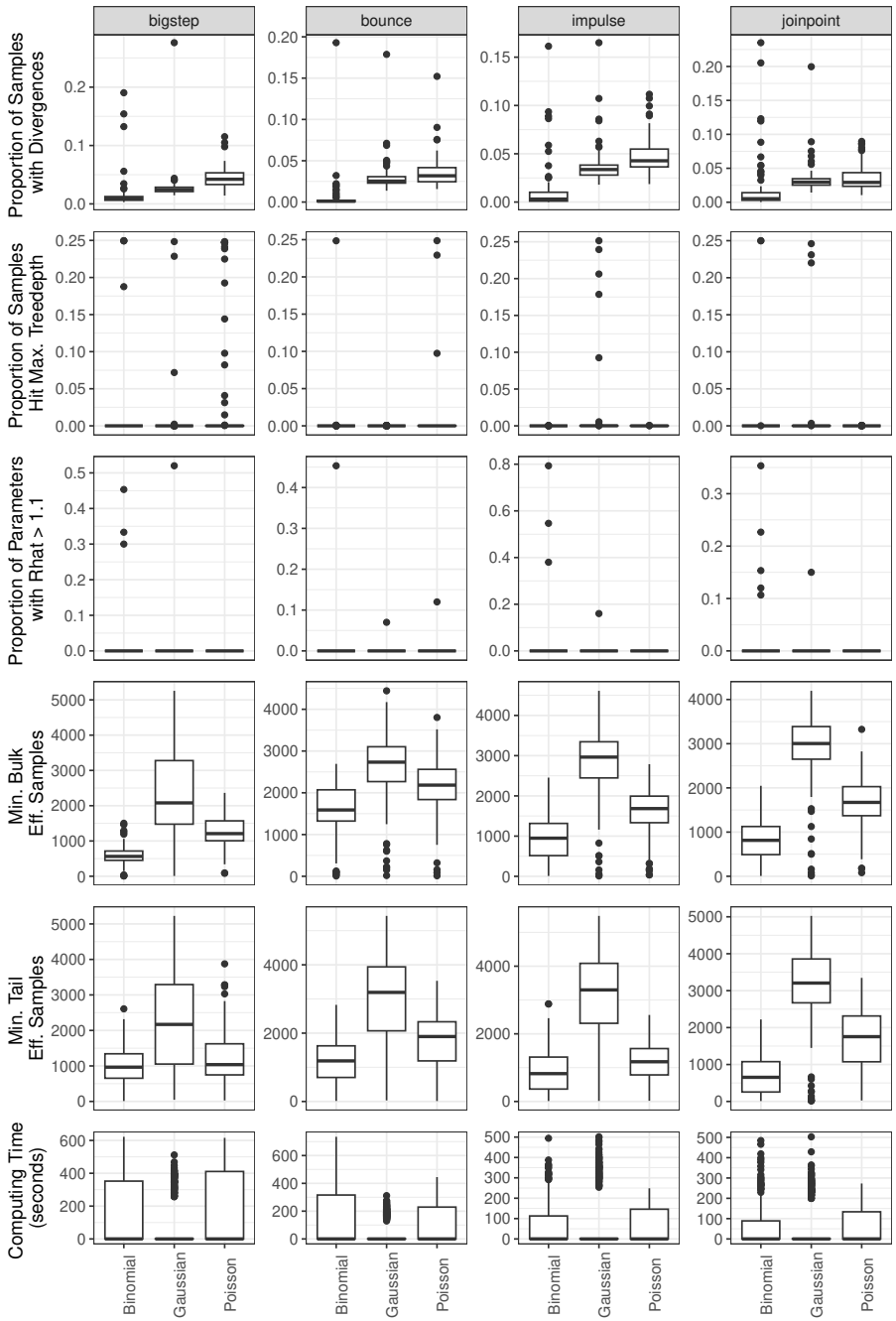

Figure F.1: Computational performance of a horseshoe process regression (HPR), based on 100 replicates in four data-generating scenarios, for continuous, binary, and count outcomes. Smaller is better for all metrics except Min. Bulk. Eff. Samples and Min. Tail. Eff. Samples (the minimum effective sample size in the bulk and tails of the posterior, respectively). Each column is for one data-generating scenario.

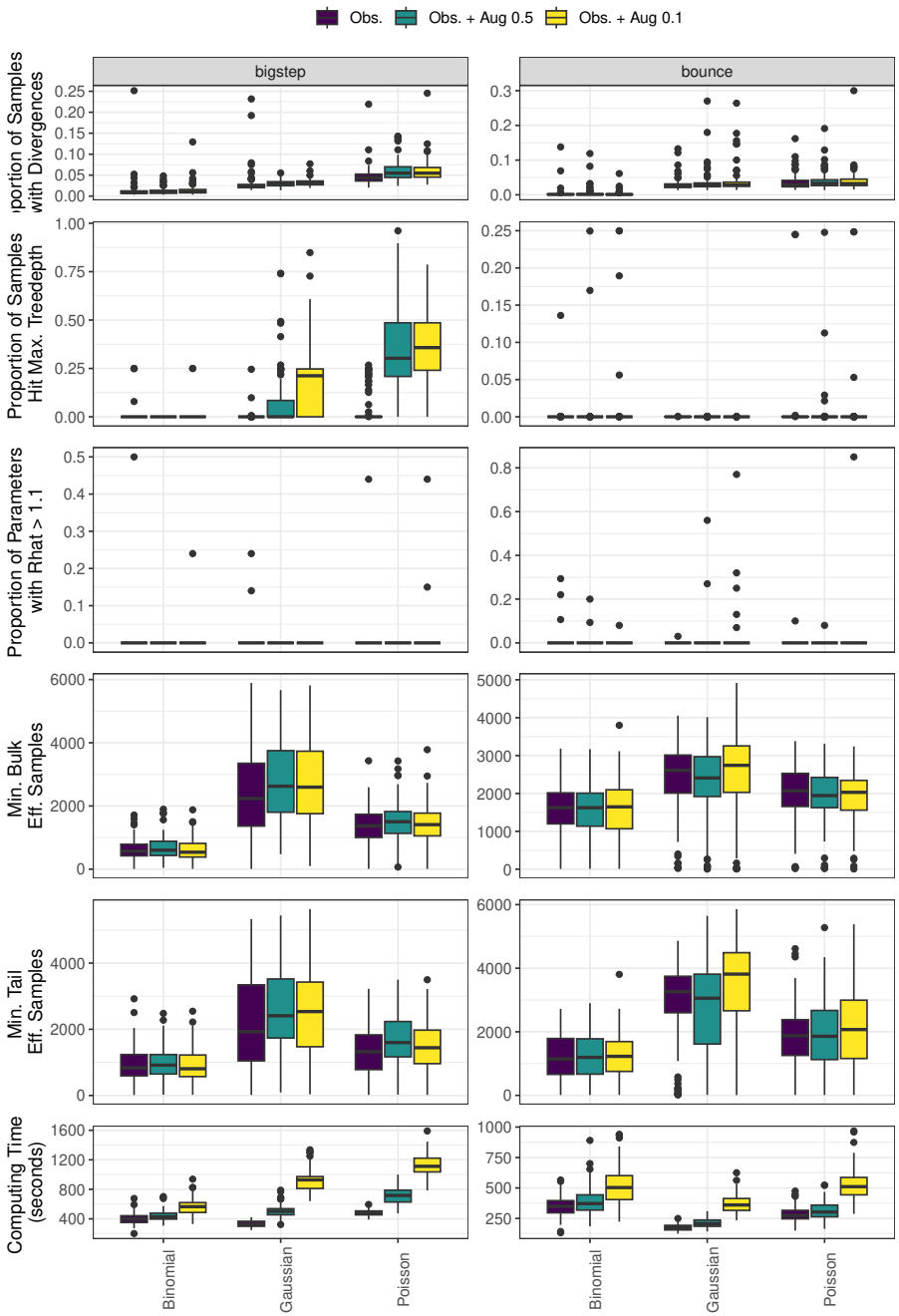

Figure F.2: Computational performance of a horseshoe process regression (HPR) in the presence of data interpolation, based on 100 replicates in two data-generating scenarios, for continuous, binary, and count outcomes. Smaller is better for all metrics except Min. Bulk. Eff. Samples and Min. Tail. Eff. Samples (the minimum effective sample size in the bulk and tails of the posterior, respectively). Each column is for one data-generating scenario.

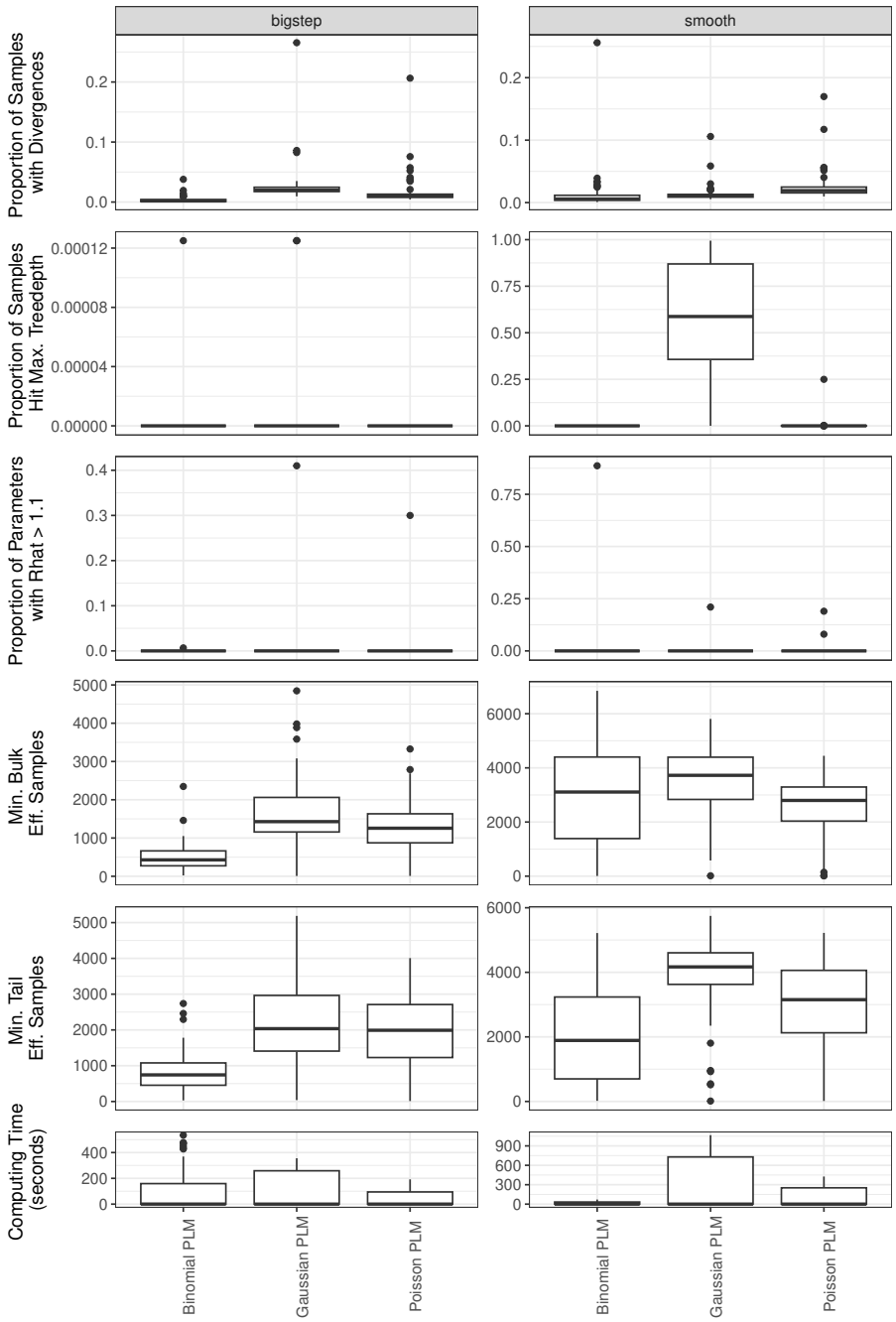

Figure F.3: Computational performance of a horseshoe process regression (HPR) partial linear model, based on 100 replicates in two data-generating scenarios, for continuous, binary, and count outcomes. Smaller is better for all metrics except Min. Bulk. Eff. Samples and Min. Tail. Eff. Samples (the minimum effective sample size in the bulk and tails of the posterior, respectively). Each column is for one data-generating scenario.

APPENDIX G

Sensitivity Analyses for HPR

We also explored the role of sample size and prior specification in model estimation. We focused these sensitivity analyses on the bigstep scenario described above, because it is horseshoe process regression (HPR)'s recommended setting. In addition to the sample size of $n = 100$ that we used above, we also considered $n = 30$ and $n = 500$. We considered several different settings for the hyperparameters of the model:

- The prior mean on the y-intercept α: We recommend setting this hyperparameter to be the sample mean of the data (using appropriate transformations for binary and count outcomes). In the sensitivity analyses below, we compare this approach to 1) setting the prior mean to be the true value of α (alpha_mean = 0) or 2) setting the prior mean to be much larger (alpha_mean = 10).
- The prior standard deviation on the y-intercept α: We recommend setting this hyperparameter to be the sample standard deviation of the data (using appropriate transformations for binary and count outcomes). In the sensitivity analyses below, we compare this approach to 1) setting the prior standard deviation to be too small (alpha_sd = 0.05) or 2) too large (alpha_sd = 50).
- The prior scale c for the global shrinkage parameter τ: We recommend using a value of $c = 0.01$ for this hyperparameter, although in some cases different values may be more suitable. We compare this approach to 1) $c = 1$ and $c = 0.0001$.
- The prior scale s for the measurement error σ, in the case of continuous outcomes: We recommend setting s to be 10 times the sample standard deviation of the data (we used $s = 5$ in these simulations). We compare this approach to 1) setting s to be too small $s = 0.05$ or 2) setting it to be the true value $s = 0.5$.

At each sample size, we generate 100 datasets from the bigstep scenario as described in Section 2.4.1 and in Appendix A. We then run a HPR using all of the recommended hyperparameter set-

tings except the one we are examining (e.g. we would change the value of c but leave the priors on α, σ according to our recommendations). Performance on the metrics of mean absolute difference, credible interval coverage, and credible interval width were compared across hyperparameter settings. We also considered the difference in point estimates and credible interval width between the recommended and alternative settings on each sample dataset.

Results are shown in Figures G.1 and G.2. Performance was generally stable across hyperparameter values, although at smaller sample sizes ($n = 30$), findings were more affected by hyperparameter choices. Poor choices for the prior variance on α—particularly setting it too small—negatively affected model fit. The choice of c also affected findings at small sample sizes, particularly for binary outcomes. Model estimation improved with larger sample sizes, although estimation was still adequate at the $n = 30$ sample size.

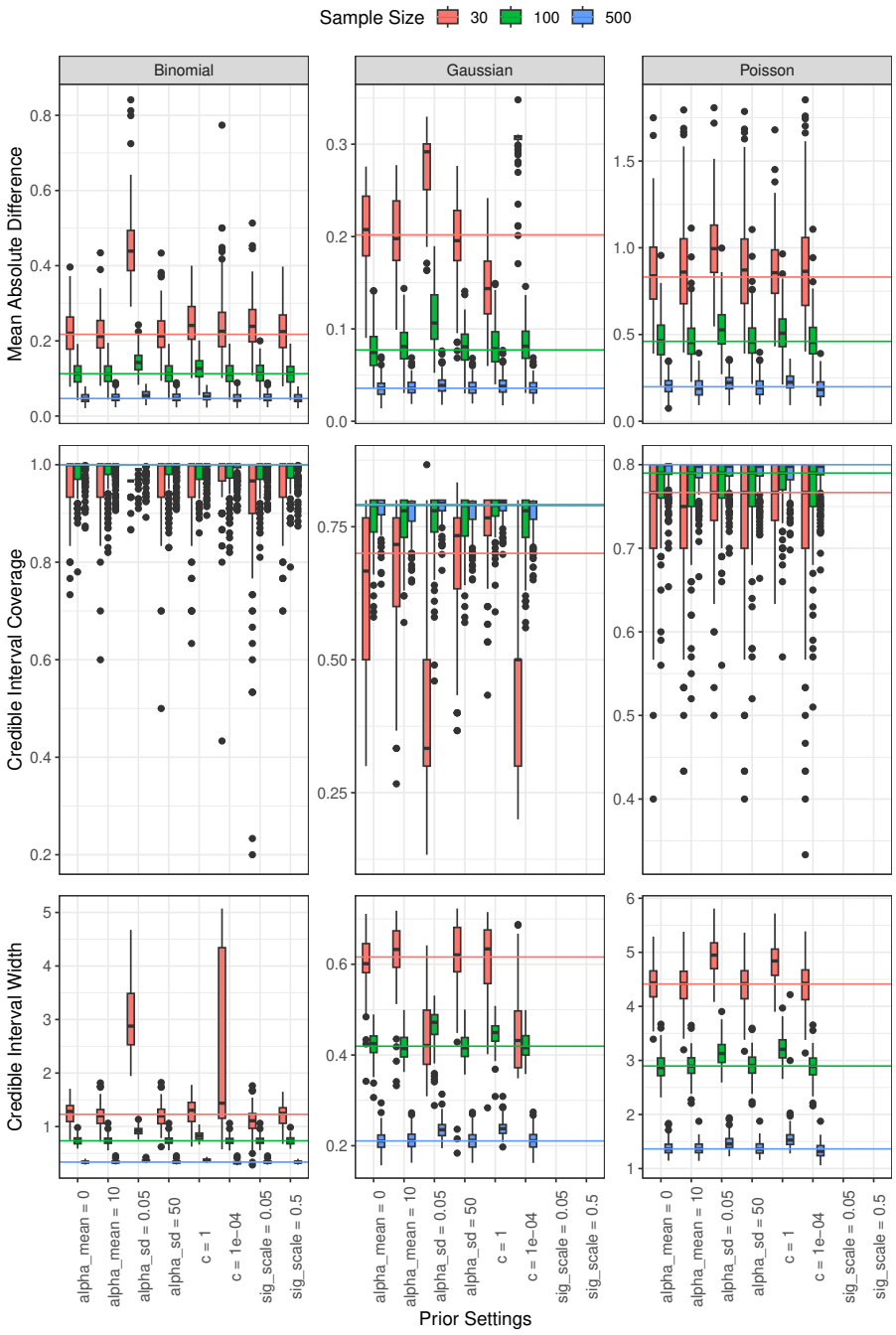

Figure G.1: Sensitivity analyses for the role of hyperparameters and sample size in horseshoe process regression (HPR), based on 100 replicates of the bigstep data generating scenario at three sample sizes ($n = 30, n = 100, n = 500$). The top row gives performance for mean absolute difference (smaller is better); the second row gives performance for credible interval coverage (0.95 is nominal); the third row gives performance for credible/confidence interval width. Each column is for one type of outcome (binary, continuous, and count). Median performance under our recommended hyperparameter settings is given as a horizontal line, with color corresponding to sample size.

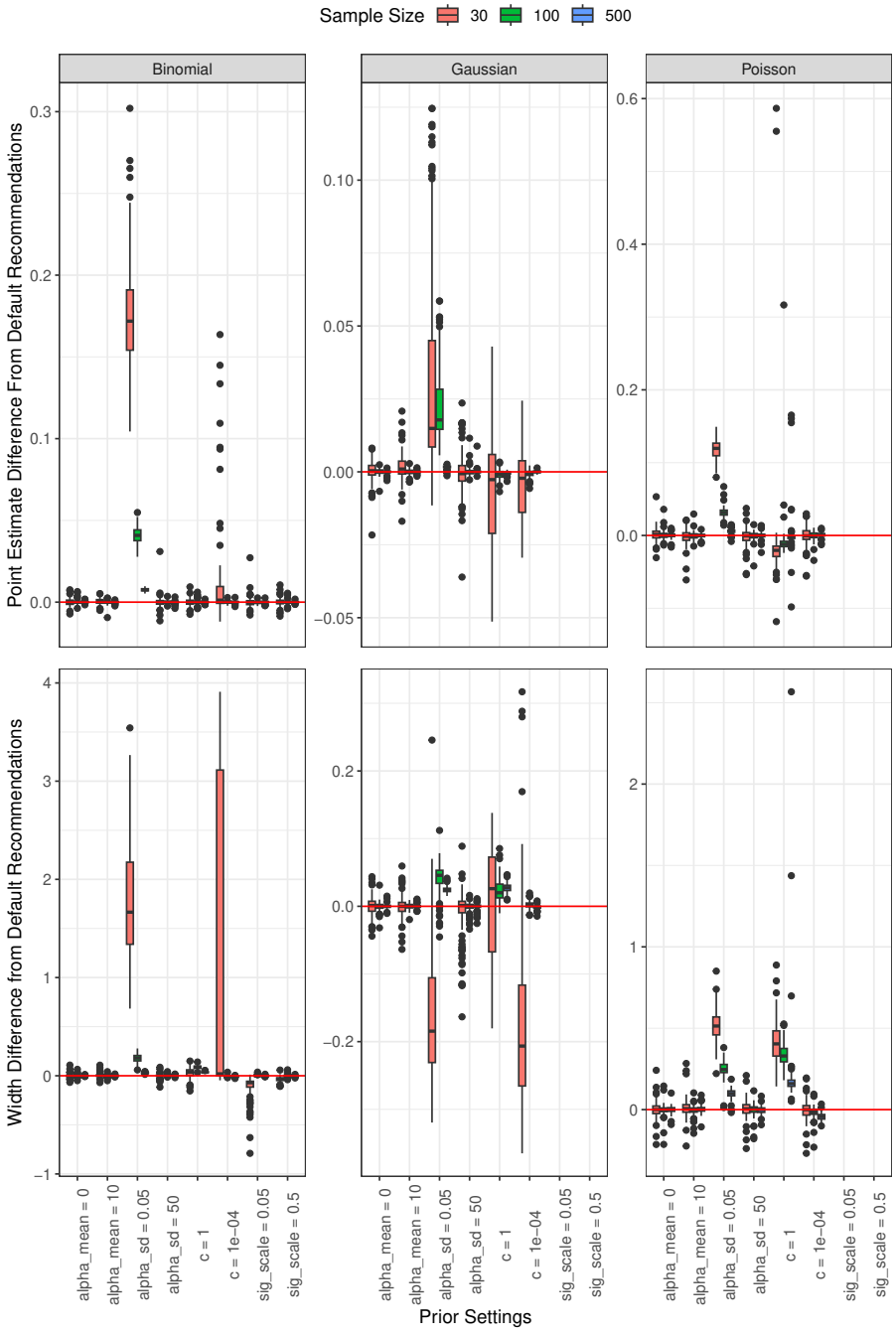

Figure G.2: Sensitivity analyses for the role of hyperparameters and sample size in horseshoe process regression (HPR), based on 100 replicates of the bigstep data generating scenario at three sample sizes ($n = 30, n = 100, n = 500$). The top row gives the difference in point estimates between each alternative hyperparameter choice and our default recommendations, aggregated across timepoints (a difference of 0 is ideal). The second row gives the difference in credible interval width between each alternative hyperparameter choice and our default recommendations, aggregated across timepoints (a difference of 0 is ideal).

APPENDIX H

Variational Inference for HPR

Let y_i be the basal body temperature (BBT) measurement observed on day t_i, $i = 1, ..., m$ of a single menstrual cycle. The model is:

$$\begin{aligned}
y_i &= f_i + \epsilon_i \\
f_i &= \alpha + H_i \\
H_i - H_{i-1} | \tau^2, \lambda_i^2 &\sim N(0, \tau^2 \lambda_i^2 (t_i - t_{i-1})),\ i = 2, ..., m \\
H_1 &= 0 \\
\alpha &\sim N(a, b^2) \\
\tau^2 | a_\tau &\sim Inv\chi^2(1, 1/a_\tau),\ a_\tau \sim Inv\chi^2(\kappa_\tau, s_\tau) \\
\lambda_i^2 | a_{\lambda_i} &\overset{iid}{\sim} Inv\chi^2(1, 1/a_{\lambda_i}),\ a_{\lambda_i} \sim Inv\chi^2(\kappa_{\lambda_i}, s_{\lambda_i}),\ i = 2, ..., m \\
\epsilon_i | \sigma^2 &\sim N(0, \sigma^2) \\
\sigma^2 | a_\sigma &\sim Inv\chi^2(1, 1/a_\sigma),\ a_\sigma \sim Inv\chi^2(\kappa_\sigma, s_\sigma)
\end{aligned} \tag{H.1}$$

Note that in Chapter 3, we have $\kappa_\tau, \kappa_\sigma, \kappa_{\lambda_i}, s_{\lambda_i} = 1$; here, we allow those hyperparameters to take on different values for flexibility. In general, we will set them to be 1.

For ease of notation, define $\boldsymbol{y}$ as the length m vector containing $y_i, i = 1, ..., m$ and $\boldsymbol{t}$ as the length m vector containing $t_i, i = 1, ..., m$. Let $\boldsymbol{\Lambda}$ be the length $m-1$ vector containing the values of $\lambda_i^2, i = 2, ..., m$; let $\boldsymbol{\lambda}$ be the length $m-1$ vector containing the values of $\lambda_i, i = 2, ..., m$. Let $\boldsymbol{a_\lambda}, \boldsymbol{\kappa_\lambda}, \boldsymbol{s_\lambda}$ be the length $m-1$ vectors containing the values of $a_{\lambda_i}, \kappa_{\lambda_i}, s_{\lambda_i}, i = 2, ..., m$, respectively. Define $\boldsymbol{H}$ as the length m vector containing the values of $H_i, i = 1, ..., m$.

In this model, $a, b^2, \kappa_\tau, s_\tau, \boldsymbol{\kappa_\lambda}, \boldsymbol{s_\lambda}, \kappa_\sigma, s_\sigma$ are hyperparameters that must be specified. Our parameters are $\boldsymbol{\theta} = (\alpha, \boldsymbol{H}, \tau^2, \tau, \boldsymbol{\Lambda}, \boldsymbol{a_\lambda}, \sigma^2, a_\sigma)$. Our data are $\boldsymbol{X} = (\boldsymbol{y}, \boldsymbol{t})$.

We seek a variational approximation $q(\boldsymbol{\theta})$ to the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{X})$. To limit the space of options for $q(\boldsymbol{\theta})$, we make the following mean field assumption:

$$q(\alpha, \boldsymbol{H}, \tau^2, a_\tau, \boldsymbol{\Lambda}, \boldsymbol{a_\lambda}, \sigma^2, a_\sigma) = q(\alpha)q(\boldsymbol{H})q(\tau^2)q(a_\tau)q(\sigma^2)q(a_\sigma)\prod_{i=2}^{m} q(\lambda_i^2)q(a_{\lambda_i}) \tag{H.2}$$

In words, we assume that all parameters are approximately posterior independent of each other, except the values of $\boldsymbol{H}$, which are treated jointly.

Under this mean-field assumption, each q-density (e.g. $q(\theta_1) = q(\alpha), q(\boldsymbol{\theta_2}) = q(\boldsymbol{H})$, etc.) is given by:

$$q(\boldsymbol{\theta}_k) \propto \exp E_{q_{j\neq k}}[\ln p(\boldsymbol{X}, \boldsymbol{\theta})] \tag{H.3}$$

for $k = 1, ..., 8$. Note that the joint log-likelihood of our model is:

$$\begin{aligned} \ln p(\boldsymbol{X}, \boldsymbol{\theta}) = &\ln p(\boldsymbol{y}|\boldsymbol{H}, \sigma^2, \alpha) + \ln p(\boldsymbol{H}|\boldsymbol{\Lambda}, \tau^2) + \ln p(\alpha) + \ln p(\sigma^2|a_\sigma) \\ &+ \ln p(a_\sigma) + \ln p(\tau^2|a_\tau) + \ln p(a_\tau) \\ &+ \sum_{i=2}^{m}[\ln p(\lambda_i^2|a_{\lambda_i}) + \ln p(a_{\lambda_i})] \end{aligned} \tag{H.4}$$

H.0.1 Finding $q(\alpha)$

We will start by finding the q-density of α. We write the log of the joint-likelihood, restricting our attention to the terms that contain α:

$$\ln p(\boldsymbol{y}|\boldsymbol{H}, \alpha, \sigma^2) + \ln p(\alpha)$$

Then the q-density is:

$$\begin{aligned}
q(\alpha) &\propto \exp\left\{ \frac{-m}{2}\ln(2\pi) - \frac{m}{2}E_{\sigma^2}[\ln\sigma^2] - \frac{1}{2}E_{\sigma^2}[\sigma^{-2}]E_H[(\boldsymbol{y}-\alpha\mathbf{1}_m-\boldsymbol{H})^T(\boldsymbol{y}-\alpha\mathbf{1}_m-\boldsymbol{H})] \right. \\
&\quad \left. - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(b^2) - \frac{1}{2b^2}(\alpha-a)^2 \right\} \\
&\propto \exp\left\{ -\frac{1}{2}E_{\sigma^2}[\sigma^{-2}]E_H[(\boldsymbol{y}-\alpha\mathbf{1}_m-\boldsymbol{H})^T(\boldsymbol{y}-\alpha\mathbf{1}_m-\boldsymbol{H})] - \frac{1}{2b^2}(\alpha-a)^2 \right\} \\
&\propto \exp\left\{ -\frac{1}{2}E_{\sigma^2}[\sigma^{-2}](m\alpha^2 - 2\alpha\boldsymbol{y}^T\mathbf{1}_m + 2\alpha E_H[\boldsymbol{H}]^T\mathbf{1}_m) - \frac{1}{2b^2}(\alpha^2-2a\alpha) \right\} \\
&\propto \exp\left\{ -\frac{1}{2}\Big(\alpha^2(mE_{\sigma^2}[\sigma^{-2}]+b^{-2}) - 2\alpha(E_{\sigma^2}[\sigma^{-2}](\boldsymbol{y}^T\mathbf{1}_m - E_H[\boldsymbol{H}]^T\mathbf{1}_m) + \frac{a}{b^2})\Big) \right\}
\end{aligned}$$

Completing the square:

$$q(\alpha) \propto \exp\left\{ -\frac{1}{2}\Big((mE_{\sigma^2}[\sigma^{-2}]+b^{-2})(\alpha - \frac{E_{\sigma^2}[\sigma^{-2}](\boldsymbol{y}^T\mathbf{1}_m - E_H[\boldsymbol{H}]^T\mathbf{1}_m) + \frac{a}{b^2}}{mE_{\sigma^2}[\sigma^{-2}]+b^{-2}})^2\Big) \right\} \quad \text{(H.5)}$$

This is a normal kernel with $\mu = \frac{E_{\sigma^2}[\sigma^{-2}](\boldsymbol{y}^T\mathbf{1}_m - E_H[\boldsymbol{H}]^T\mathbf{1}_m) + \frac{a}{b^2}}{mE_{\sigma^2}[\sigma^{-2}]+b^{-2}}$ and variance $(mE_{\sigma^2}[\sigma^{-2}] + b^{-2})^{-1}$.

H.0.2 Finding $q(\boldsymbol{H})$

Define $\boldsymbol{H}^*$ as the length $m-1$ vector containing the values of $H_i, i = 2, ..., m$, i.e. the vector $\boldsymbol{H}$ without $H_1 = 0$, a constant. Then:

$$q(\boldsymbol{H}^*) \propto \exp(E_{-\boldsymbol{H}^*}[\ln p(\boldsymbol{y}|\boldsymbol{H}, \sigma^2, \alpha) + \ln p(\boldsymbol{H}^*|\boldsymbol{\Lambda}, \tau^2)])$$

The likelihood $p(\boldsymbol{y}|\boldsymbol{H}, \sigma^2, \alpha)$ is $\boldsymbol{y} \sim N(\alpha + \boldsymbol{H}, \sigma^2 \boldsymbol{I}_{n\times n})$. The prior on $\boldsymbol{H}^*$ is more complicated because the prior is placed on its first order differences. Let $\delta_i = t_i - t_{i-1}$. This implies:

$$p(\boldsymbol{H}^*|\boldsymbol{\Lambda}, \tau^2) = (2\pi\tau^2)^{\frac{-(m-1)}{2}} \prod_{i=2}^{m} (\lambda_i^2 \delta_i)^{-1/2} \exp(\frac{-1}{2\tau^2} \sum_{i=2}^{m} \frac{(H_i - H_{i-1})^2}{\lambda_i^2 \delta_i}) \tag{H.6}$$

Let $\boldsymbol{R}$ be an $m-1 \times m-1$ matrix of the form:

$$\boldsymbol{R} = \begin{bmatrix} \omega_2^2 + \omega_3^2 & -\omega_3^2 & & & & \\ -\omega_3^2 & \omega_3^2 + \omega_4^2 & -\omega_4^2 & & & \\ & \ddots & \ddots & \ddots & & \\ & & & \omega_{m-1}^2 + \omega_m^2 & -\omega_m^2 \\ & & & -\omega_m^2 & \omega_m^2 \end{bmatrix}$$

where $\omega_i^2 = \frac{1}{\tau^2 \lambda_i^2 \delta_i}$.

Then we can rewrite the above prior as:

$$p(\boldsymbol{H}^*|\boldsymbol{\Lambda}, \tau^2) = (2\pi\tau^2)^{\frac{-(m-1)}{2}} (\prod_{i=2}^{m} \lambda_i \delta_i)^{-1/2} \exp(\frac{-1}{2} \boldsymbol{H}^{*T} \boldsymbol{R} \boldsymbol{H}^*) \tag{H.7}$$

Let $\boldsymbol{y}^*$ refer to the elements of $\boldsymbol{y}$ except y_1. Then we have:

$$\begin{aligned} q(\boldsymbol{H}^*) \propto \exp \Bigg\{ & \frac{-(2m-1)}{2} \ln(2\pi) - \frac{m}{2} E_{\sigma^2}[\ln \sigma^2] - \frac{1}{2} E_{\sigma^2}[\sigma^{-2}] E_\alpha[(\boldsymbol{y}^* - \alpha \mathbf{1}_{m-1} - \boldsymbol{H}^*)^T \\ & (\boldsymbol{y}^* - \alpha \mathbf{1}_{m-1} - \boldsymbol{H}^*)] - \frac{m-1}{2} E_{\tau^2}[\ln \tau^2] - \frac{1}{2} \sum_{i=2}^{m} E_{\lambda_i}[\ln \lambda_i^2] - \frac{1}{2} \sum_{i=2}^{m} \ln \delta_i \\ & - \frac{1}{2} (\boldsymbol{H}^{*T} E_R[\boldsymbol{R}] \boldsymbol{H}^*) \Bigg\} \end{aligned}$$

We drop all terms without $\boldsymbol{H}^*$:

$$q(\boldsymbol{H}^*) \propto \exp\left\{ -\frac{1}{2}E_{\sigma^2}[\sigma^{-2}]E_\alpha[(\boldsymbol{y}^* - \alpha\mathbf{1}_{m-1} - \boldsymbol{H}^*)^T(\boldsymbol{y}^* - \alpha\mathbf{1}_{m-1} - \boldsymbol{H}^*)] - \frac{1}{2}(\boldsymbol{H}^{*T}E_R[\boldsymbol{R}]\boldsymbol{H}^*)\right\}$$

Doing some more algebraic manipulation to clarify the multivariate Gaussian kernel:

$$\begin{aligned}
q(\boldsymbol{H}^*) &\propto \exp\left\{ -\frac{1}{2}E_{\sigma^2}[\sigma^{-2}]E_\alpha[\boldsymbol{y}^{*T}\boldsymbol{y}^* - \alpha\boldsymbol{y}^{*T}\mathbf{1}_{m-1} - \boldsymbol{y}^{*T}\boldsymbol{H}^* - \alpha\mathbf{1}_{m-1}^T\boldsymbol{y}^* + m\alpha^2 \right. \\
&\quad \left. + \alpha\mathbf{1}_{m-1}^T\boldsymbol{H}^* - \boldsymbol{H}^{*T}\boldsymbol{y}^* + \alpha\boldsymbol{H}^{*T}\mathbf{1}_{m-1} + \boldsymbol{H}^{*T}\boldsymbol{H}^*] - \frac{1}{2}(\boldsymbol{H}^{*T}E_R[\boldsymbol{R}]\boldsymbol{H}^*)\right\} \\
&\propto \exp\left\{ -\frac{1}{2}E_{\sigma^2}[\sigma^{-2}](-2\boldsymbol{y}^{*T}\boldsymbol{H}^* + 2E_\alpha[\alpha]\mathbf{1}_{m-1}^T\boldsymbol{H}^* + \boldsymbol{H}^{*T}\boldsymbol{H}^*) - \frac{1}{2}(\boldsymbol{H}^{*T}E_R[\boldsymbol{R}]\boldsymbol{H}^*)\right\} \\
&\propto \exp\left\{ -\frac{1}{2}E_{\sigma^2}[\sigma^{-2}](-2\boldsymbol{y}^{*T} + 2E_\alpha[\alpha]\mathbf{1}_{m-1}^T)\boldsymbol{H}^* - \frac{1}{2}(E_{\sigma^2}[\sigma^{-2}]\boldsymbol{H}^{*T}\boldsymbol{H}^* + \boldsymbol{H}^{*T}E_R[\boldsymbol{R}]\boldsymbol{H}^*)\right\} \\
&\propto \exp\left\{ -\frac{1}{2}(-2E_{\sigma^2}[\sigma^{-2}](\boldsymbol{y}^{*T} - E_\alpha[\alpha]\mathbf{1}_{m-1}^T)\boldsymbol{H}^* + \boldsymbol{H}^{*T}(E_{\sigma^2}[\sigma^{-2}]\boldsymbol{I}_{m-1\times m-1} + E_R[\boldsymbol{R}])\boldsymbol{H}^*)\right\}
\end{aligned}$$

Recall the formula for completing the square with matrices:

$$\boldsymbol{x}^T\boldsymbol{M}\boldsymbol{x} - 2\boldsymbol{b}^T\boldsymbol{x} = (\boldsymbol{x} - \boldsymbol{M}^{-1}\boldsymbol{b})^T\boldsymbol{M}(\boldsymbol{x} - \boldsymbol{M}^{-1}\boldsymbol{b}) - \boldsymbol{b}^T\boldsymbol{M}^{-1}\boldsymbol{b}$$

Recognizing $\boldsymbol{M} = E_{\sigma^2}[\sigma^{-2}]\boldsymbol{I}_{m-1\times m-1} + E_R[\boldsymbol{R}]$ and $\boldsymbol{b} = E_{\sigma^2}[\sigma^{-2}](\boldsymbol{y}^{*\boldsymbol{T}} - E_\alpha[\alpha]\mathbf{1}_{m-1}^T)^T$, we can finally see that:

$$\begin{aligned}
q(\boldsymbol{H}^*) = MVN(\boldsymbol{H}^*|\boldsymbol{\mu} &= (E_{\sigma^2}[\sigma^{-2}]\boldsymbol{I}_{m-1\times m-1} + E_R[\boldsymbol{R}])^{-1}E_{\sigma^2}[\sigma^{-2}](\boldsymbol{y}^{*\boldsymbol{T}} - E_\alpha[\alpha]\mathbf{1}_{m-1}^T)^T, \\
\boldsymbol{\Sigma} &= (E_{\sigma^2}[\sigma^{-2}]\boldsymbol{I}_{m-1\times m-1} + E_R[\boldsymbol{R}])^{-1})
\end{aligned} \tag{H.8}$$

H.0.3 Finding $q(\sigma^2)$

We write the log of the joint-likelihood, restricting our attention to the terms that contain σ^2:

$$\ln p(\boldsymbol{y}|\boldsymbol{H}, \alpha, \sigma^2) + \ln p(\sigma^2|a_\sigma)$$

Then the q-density is:

$$\begin{aligned}
q(\sigma^2) \propto \exp\Bigg\{ & \frac{-m}{2}\ln(2\pi) - \frac{m}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}E_{H,\alpha}[(\boldsymbol{y}-\alpha\mathbf{1}_m-\boldsymbol{H})^T(\boldsymbol{y}-\alpha\mathbf{1}_m-\boldsymbol{H})] + \frac{1}{2}\ln\frac{1}{2} \\
& - \frac{1}{2}E_{a_\sigma}[\ln a_\sigma] - \ln\Gamma(1/2) - \frac{3}{2}\ln\sigma^2 - \frac{1}{2a_\sigma\sigma^2}\Bigg\} \\
\propto \exp\Bigg\{ & -\frac{m}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}E_{H,\alpha}[(\boldsymbol{y}-\alpha\mathbf{1}_m-\boldsymbol{H})^T(\boldsymbol{y}-\alpha\mathbf{1}_m-\boldsymbol{H})] - \frac{3}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}E_{a_\sigma}[\frac{1}{a_\sigma}]\Bigg\} \\
\propto (\sigma^2)^{\frac{-m-3}{2}} \exp\Big\{ & -\frac{1}{2\sigma^2}\Big(E_{H,\alpha}[(\boldsymbol{y}-\alpha\mathbf{1}_m-\boldsymbol{H})^T(\boldsymbol{y}-\alpha\mathbf{1}_m-\boldsymbol{H})] + E_{a_\sigma}[\frac{1}{a_\sigma}]\Big)\Big\}
\end{aligned}$$

We recognize this as the kernel of an Inverse-χ^2 distribution with shape parameter $\kappa = m+1$ and scale parameter $s = E_{H,\alpha}[(\boldsymbol{y}-\alpha\mathbf{1}_m-\boldsymbol{H})^T(\boldsymbol{y}-\alpha\mathbf{1}_m-\boldsymbol{H})] + E_{a_\sigma}[\frac{1}{a_\sigma}]$.

H.0.4 Finding $q(\tau^2)$

We write the log of the joint-likelihood, restricting our attention to the terms that contain τ^2:

$$\ln p(\boldsymbol{H}|\tau^2, \boldsymbol{\Lambda}^2) + \ln p(\tau^2|a_\tau)$$

Then the q-density is:

$$\begin{aligned}
q(\tau^2) &\propto \exp\Bigg\{ -\frac{(m-1)}{2}\ln 2\pi - \frac{(m-1)}{2}\ln\tau^2 - \frac{1}{2}\sum_{i=2}^{m} E_\lambda[\ln\lambda_i^2\delta_i] - \frac{1}{2\tau^2}\sum_{i=2}^{m}\frac{E_H[(H_i-H_{i-1})^2]E_\lambda[\lambda_i^{-2}]}{\delta_i} \\
&\quad + \frac{1}{2}\ln\frac{1}{2} - \frac{1}{2}E_{a_\tau}[\ln a_\tau] - \ln\Gamma(1/2) - \frac{3}{2}\ln\tau^2 - \frac{1}{2}E_{a_\tau}[\frac{1}{a_\tau}]\frac{1}{\tau^2}\Bigg\} \\
&\propto \exp\Bigg\{ -\frac{(m-1)}{2}\ln\tau^2 - \frac{1}{2\tau^2}\sum_{i=2}^{m}\frac{E_H[(H_i-H_{i-1})^2]E_\lambda[\lambda_i^{-2}]}{\delta_i} - \frac{3}{2}\ln\tau^2 - \frac{1}{2}E_{a_\tau}[\frac{1}{a_\tau}]\frac{1}{\tau^2}\Bigg\} \\
&\propto (\tau^2)^{-\frac{(m+2)}{2}}\exp\Bigg\{ -\frac{1}{2\tau^2}\Big(\sum_{i=2}^{m}\frac{E_H[(H_i-H_{i-1})^2]E_\lambda[\lambda_i^{-2}]}{\delta_i} + E_{a_\tau}[\frac{1}{a_\tau}]\Big)\Bigg\}
\end{aligned}$$

We recognize this as the kernel of an Inverse-χ^2 distribution with shape parameter $\kappa = m$ and scale parameter $s = \sum_{i=2}^{m}\frac{E_H[(H_i-H_{i-1})^2]E_\lambda[\lambda_i^{-2}]}{\delta_i} + E_{a_\tau}[\frac{1}{a_\tau}]$.

H.0.5 Finding $q(\mathbf{\Lambda})$

We write the log of the joint-likelihood, restricting our attention to the terms that contain $\mathbf{\Lambda}$:

$$\ln p(\boldsymbol{H}|\tau^2, \mathbf{\Lambda}) + \sum_{i=2}^{m} [\ln p(\lambda_i^2|a_{\lambda_i})$$

Then the q-density is:

$$\begin{aligned}
q(\mathbf{\Lambda}) \propto \exp\Bigg\{ & -\frac{(m-1)}{2}\ln 2\pi - \frac{(m-1)}{2}E_\tau[\ln\tau^2] - \frac{1}{2}\sum_{i=2}^{m}\ln\lambda_i^2 - \frac{1}{2}\sum_{i=2}^{m}\ln\delta_i \\
& - \frac{1}{2}E_\tau[\frac{1}{\tau^2}]\sum_{i=2}^{m}\frac{E_H[(H_i - H_{i-1})^2]}{\lambda_i^2\delta_i} + \frac{m-1}{2}\ln\frac{1}{2} - \frac{1}{2}\sum_{i=2}^{m}E_{a_{\lambda_i}}[\ln a_{\lambda_i}] \\
& - (m-1)\ln\Gamma(1/2) - \frac{3}{2}\sum_{i=2}^{m}\ln\lambda_i^2 - \frac{1}{2}E_{a_{\lambda_i}}[\frac{1}{a_{\lambda_i}}]\frac{1}{\lambda_i^2}\Bigg\} \\
\propto \exp\Bigg\{ & -\frac{1}{2}\sum_{i=2}^{m}\ln\lambda_i^2 - \frac{1}{2}E_\tau[\frac{1}{\tau^2}]\sum_{i=2}^{m}\frac{E_H[(H_i - H_{i-1})^2]}{\lambda_i^2\delta_i} - \frac{3}{2}\sum_{i=2}^{m}\ln\lambda_i^2 - \frac{1}{2}E_{a_{\lambda_i}}[\frac{1}{a_{\lambda_i}}]\frac{1}{\lambda_i^2}\Bigg\}
\end{aligned}$$

Considering only a single λ_i, we obtain:

$$\begin{aligned}
q(\lambda_i^2) &\propto \exp\Bigg\{ -\frac{1}{2}\ln\lambda_i^2 - \frac{1}{2}E_\tau[\frac{1}{\tau^2}]\frac{E_H[(H_i - H_{i-1})^2]}{\lambda_i^2\delta_i} - \frac{3}{2}\ln\lambda_i^2 - \frac{1}{2}E_{a_{\lambda_i}}[\frac{1}{a_{\lambda_i}}]\frac{1}{\lambda_i^2}\Bigg\} \\
&= (\lambda_i^2)^{-2}\exp\Big\{ -\frac{1}{2\lambda_i^2}\Big(E_{\tau^2}[\frac{1}{\tau^2}]\frac{E_H[(H_i - H_{i-1})^2]}{\delta_i} + E_{a_{\lambda_i}}[\frac{1}{a_{\lambda_i}}]\Big)\Big\}
\end{aligned}$$

We recognize this as the kernel of an Inverse-χ^2 distribution with shape parameter $\kappa = 2$ and scale parameter $s = E_{\tau^2}[\frac{1}{\tau^2}]\frac{E_H[(H_i - H_{i-1})^2]}{\delta_i} + E_{a_{\lambda_i}}[\frac{1}{a_{\lambda_i}}]$. This q-density will be the same for all $\lambda_i,\ i = 2, ..., m$; multiplying them provides $q(\mathbf{\Lambda})$.

H.0.6 Finding $q(a_\sigma)$, $q(a_\tau)$, $q(\boldsymbol{a_\lambda})$

We will start with $q(a_\sigma)$. We write the log of the joint-likelihood, restricting our attention to the terms that contain a_σ:

$$\ln p(\sigma^2|a_\sigma) + \ln p(a_\sigma)$$

Then the q-density is:

$$\begin{aligned}
q(a_\sigma) &\propto \exp\left\{ \frac{1}{2}\ln\frac{1}{2} - \frac{1}{2}\ln a_\sigma - \ln\Gamma(1/2) - \frac{3}{2}E_{\sigma^2}[\ln\sigma^2] - \frac{1}{2a_\sigma}E_{\sigma^2}[\frac{1}{\sigma^2}] + \frac{\kappa_\sigma}{2}\ln\frac{s_\sigma}{2} \right. \\
&\left. - \ln\Gamma(\frac{\kappa_\sigma}{2}) - (\frac{\kappa_\sigma}{2}+1)\ln a_\sigma - \frac{s_\sigma}{2a_\sigma} \right\} \\
&\propto \exp\left\{ -\frac{1}{2}\ln a_\sigma - \frac{1}{2a_\sigma}E_{\sigma^2}[\frac{1}{\sigma^2}] - (\frac{\kappa_\sigma}{2}+1)\ln a_\sigma - \frac{s_\sigma}{2a_\sigma} \right\} \\
&\propto a_\sigma^{-\frac{\kappa_\sigma+3}{2}} \exp\left\{ -\frac{1}{2a_\sigma}\Big(E_{\sigma^2}[\frac{1}{\sigma^2}] + s_\sigma\Big) \right\}
\end{aligned}$$

We recognize this as the kernel of an Inverse-χ^2 distribution with shape parameter $\kappa = \kappa_\sigma + 1$ and scale parameter $s = E_{\sigma^2}[\frac{1}{\sigma^2}] + s_\sigma$. The derivations for $q(a_\tau)$ and $q(\boldsymbol{a_\lambda})$ are identical and are thus omitted. $q(a_\tau)$ is an Inverse-χ^2 distribution with shape parameter $\kappa = \kappa_\tau + 1$ and scale parameter $s = E_{\tau^2}[\frac{1}{\tau^2}] + s_\tau$, while $q(a_{\lambda_i})$ is an Inverse-χ^2 distribution with shape parameter $\kappa = \kappa_{\lambda_i} + 1$ and scale parameter $s = E_\lambda[\frac{1}{\lambda_i^2}] + s_{\lambda_i}$.

H.0.7 Complete list of q-densities

Summarizing the above, our complete list of q-densities is:

1. $q(\alpha) = N(\alpha|\mu = \frac{E_{\sigma^2}[\sigma^{-2}](\boldsymbol{y}^T\mathbf{1}_m - E_H[\boldsymbol{H}]^T\mathbf{1}_m) + \frac{a}{b^2}}{mE_{\sigma^2}[\sigma^{-2}] + b^{-2}}, V = (mE_{\sigma^2}[\sigma^{-2}] + b^{-2})^{-1})$
2. $q(\boldsymbol{H^*}) = MVN(\boldsymbol{H^*}|\boldsymbol{\mu} = (E_{\sigma^2}[\sigma^{-2}]\boldsymbol{I}_{m-1\times m-1} + E_R[\boldsymbol{R}])^{-1}E_{\sigma^2}[\sigma^{-2}](\boldsymbol{y^{*T}} - E_\alpha[\alpha]\mathbf{1}_{m-1}^T)^T, \boldsymbol{\Sigma} = (E_{\sigma^2}[\sigma^{-2}]\boldsymbol{I}_{m-1\times m-1} + E_R[\boldsymbol{R}])^{-1})$
3. $q(\sigma^2) = Inv\chi^2(\sigma^2|\kappa = m+1, s = E_{H,\alpha}[(\boldsymbol{y} - \alpha\mathbf{1}_m - \boldsymbol{H})^T(\boldsymbol{y} - \alpha\mathbf{1}_m - \boldsymbol{H})] + E_{a_\sigma}[\frac{1}{a_\sigma}])$
4. $q(a_\sigma) = Inv\chi^2(a_\sigma|\kappa = \kappa_\sigma + 1, s = E_{\sigma^2}[\frac{1}{\sigma^2}] + s_\sigma)$
5. $q(\tau^2) = Inv\chi^2(\tau^2|\kappa = m, s = \sum_{i=2}^m \frac{E_H[(H_i - H_{i-1})^2]E_\lambda[\lambda_i^{-2}]}{\delta_i} + E_{a_\tau}[\frac{1}{a_\tau}])$
6. $q(a_\tau) = Inv\chi^2(a_\tau|\kappa = \kappa_\tau + 1, s = E_{\tau^2}[\frac{1}{\tau^2}] + s_\tau)$
7. For $i = 2, ..., m$, $q(\lambda_i^2) = Inv\chi^2(\lambda_i^2|\kappa = 2, s = E_{\tau^2}[\frac{1}{\tau^2}]\frac{E_H[(H_i - H_{i-1})^2]}{\delta_i} + E_{a_{\lambda_i}}[\frac{1}{a_{\lambda_i}}])$
8. For $i = 2, ..., m$, $q(a_{\lambda_i}) = Inv\chi^2(a_{\lambda_i}|\kappa = \kappa_{\lambda_i} + 1, s = E_{\lambda_i^2}[\frac{1}{\lambda_i^2}] + s_{\lambda_i})$

H.0.8 Evaluating the variational objective

Recall that $L = E_q[\ln p(\boldsymbol{X}, \boldsymbol{\theta})] - \sum_{k=1}^{K} E_{q_k}[\ln q(\boldsymbol{\theta}_k)]$. This is:

$$
\begin{aligned}
L = & -\frac{1}{2} E_{\sigma^2}[\sigma^{-2}] E_{\alpha,H}[(\boldsymbol{y} - \alpha \mathbf{1}_n - \boldsymbol{H})^T (\boldsymbol{y} - \alpha \mathbf{1}_n - \boldsymbol{H})] \\
& - \frac{1}{2}(E_H[\boldsymbol{H}^{*T} E_R[\boldsymbol{R}] \boldsymbol{H}^*]) - \frac{1}{2b^2} E_\alpha[(\alpha - a)^2] \\
& - \frac{1}{2} E_{a_\sigma}[\frac{1}{a_\sigma}](E_{\sigma^2}[\frac{1}{\sigma^2}] + s_\sigma) - \frac{1}{2} E_{a_\tau}[\frac{1}{a_\tau}](E_{\tau^2}[\frac{1}{\tau_h^2}] + s_\tau) - \frac{1}{2} \sum_{i=2}^{n} E_{a_\lambda}[\frac{1}{a_{\lambda_i}}](E_\lambda[\frac{1}{\lambda_i^2}] + s_{\lambda_i}) \\
& - \frac{1}{2} \ln(n E_{\sigma^2}[\frac{1}{\sigma^2}] + \frac{1}{b^2}) - \frac{1}{2} \ln |E_{\sigma^2}[\frac{1}{\sigma^2}] \boldsymbol{I}_{n-1 \times n-1} + E_R[\boldsymbol{R}]| \\
& - \frac{n}{2} \ln \frac{1}{2} \Big(E_{a_\tau}[\frac{1}{a_\tau}] + \sum_{i=2}^{n} E_H[(H_i - H_{i-1})^2] E_\lambda[1/\lambda_i^2] 1/\delta_i \Big) \\
& - \frac{(\kappa_\tau + 1)}{2} \ln(\frac{1}{2}(s_\tau + E_{\tau_h^2}[\frac{1}{\tau_h^2}])) - \sum_{i=2}^{n} \ln \frac{1}{2} \Big(E_{a_{\lambda_i}}[\frac{1}{a_{\lambda_i}}] + \frac{E_H[(H_i - H_{i-1})^2]}{\delta_i} E_{\tau_h^2}[\frac{1}{\tau_h^2}] \Big) \\
& - \sum_{i=2}^{n} \frac{(\kappa_{\lambda_i} + 1)}{2} \ln \frac{1}{2}(s_{\lambda_i} + E_\lambda[\frac{1}{\lambda_i^2}]) - \frac{(\kappa_\sigma + 1)}{2} \ln \frac{1}{2}(s_\sigma + E_{\sigma^2}[\frac{1}{\sigma^2}]) \\
& - \frac{n+1}{2} \ln \frac{1}{2} \Big(E_{a_\sigma}[\frac{1}{a_\sigma}] + E_{\alpha,H}[(\boldsymbol{y} - \alpha \mathbf{1}_n - \boldsymbol{H})^T (\boldsymbol{y} - \alpha \mathbf{1}_n - \boldsymbol{H})] \Big)
\end{aligned}
$$

H.0.9 Algorithm structure

This yields the algorithm:

Algorithm 1 Variational inference algorithm for horseshoe process regression.

Inputs:

$\boldsymbol{X} = (\boldsymbol{y}, \boldsymbol{t})$

Initialize:

$E(\alpha)_0 \leftarrow \bar{y}$, $E(\boldsymbol{H})_0 \leftarrow \vec{0}$

$E(\frac{1}{\sigma^2})_0 \leftarrow \frac{1}{var(\boldsymbol{y})}$, $E(\frac{1}{a_\sigma})_0 \leftarrow 1$

$E(\frac{1}{\tau^2})_0 \leftarrow \frac{100}{var(\boldsymbol{y})}$, $E(\frac{1}{a_\tau})_0 \leftarrow 1$

$E(\frac{1}{\lambda_i^2})_0 \leftarrow 100$, $E(\frac{1}{a_{\lambda_i}})_0 \leftarrow 1,\ i = 2, \ldots, m$

$j \leftarrow 1$

while $j \leq 1000 \quad \& \quad \frac{|L_j - L_{j-1}|}{|L_j|} > 0.0001$ **do**

$E(\boldsymbol{H})_j \leftarrow (E(\frac{1}{\sigma^2})_{j-1}\boldsymbol{I}_{m-1\times m-1} + E[\boldsymbol{R}]_{j-1})^{-1}E(\frac{1}{\sigma^2})_{j-1}(\boldsymbol{y^{*T}} - E(\alpha)_{j-1}\mathbf{1}_{m-1}^T)^T$

$E(\alpha)_j \leftarrow [E(\frac{1}{\sigma^2})_{j-1}(\boldsymbol{y}^T\mathbf{1}_m - E(\boldsymbol{H})_j^T\mathbf{1}_m) + \frac{a}{b^2}][mE(\frac{1}{\sigma^2})_{j-1} + b^{-2}]^{-1}$

$E(\frac{1}{a_\sigma})_j \leftarrow (\kappa_\sigma + 1)/(E(\frac{1}{\sigma^2})_{j-1} + s_\sigma)$

$E(\frac{1}{\sigma^2})_j \leftarrow (m+1)(E[(\boldsymbol{y} - \alpha\mathbf{1}_m - \boldsymbol{H})^T(\boldsymbol{y} - \alpha\mathbf{1}_m - \boldsymbol{H})]_j + E(\frac{1}{a_\sigma})_j)$

$E(\frac{1}{a_\tau})_j \leftarrow (\kappa_\tau + 1)/(E(\frac{1}{\tau^2})_{j-1} + s_\tau)$

$E(\frac{1}{\tau^2})_j \leftarrow m/(\sum_{i=2}^m \frac{E[(H_i - H_{i-1})^2]_j E(\frac{1}{\lambda_i^2})_{j-1}}{\delta_i} + E(\frac{1}{a_\tau})_j)$

$E(\frac{1}{a_{\lambda_i}})_j \leftarrow (\kappa_{\lambda_i} + 1)/(E(\frac{1}{\lambda_i^2})_{j-1} + s_{\lambda_i}),\ i = 2, \ldots, m$

$E(\frac{1}{\lambda_i^2})_j \leftarrow 2/(\frac{1}{\delta_i}E(\frac{1}{\tau^2})_j E[(H_i - H_{i-1})^2]_j + E(\frac{1}{a_{\lambda_i}})_j),\ i = 2, \ldots, m$

$L_j \leftarrow L[E(\boldsymbol{H})_j, E(\alpha)_j, E(\frac{1}{a_\sigma})_j, E(\frac{1}{\sigma^2})_j, E(\frac{1}{a_\tau})_j, E(\frac{1}{\tau^2})_j, E(\frac{1}{a_{\lambda_2}})_j, \ldots, E(\frac{1}{a_{\lambda_m}})_j, E(\frac{1}{\lambda_2^2})_j, \ldots, E(\frac{1}{\lambda_m^2})_j]$

$j \leftarrow j + 1$

end while

APPENDIX I

Simulation Results Comparing Variational Inference and Hamiltonian Monte Carlo

As discussed in Section 3.3, we compared the variational inference (VI) and Hamiltonian Monte Carlo (HMC) implementations. We considered three sample sizes ($m = 28$, $m = 112 = 28 \times 4$, $m = 420 = 28 \times 15$) and four true underlying functions, motivated by the BBT setting. These functions were:

1. bigstep: $f(t) = I(t \leq 14) * 36.6 + I(t > 14) * 37.1$
2. flat: $f(t) = 36.8$
3. joinpoint1: $f(t) = I(t \leq 14) * 36.6 + I(14 < t \leq 20) * (t/12 + 37.1 - 5/3) + I(t > 20) * 37.1$
4. joinpoint2: $f(t) = I(t \leq 7) * (-t/20 + 36.95) + I(7 < t \leq 14) * 36.6 + I(14 < t \leq 22) * (t/16 + 35.725) + I(t > 22) * 37.1$

For each of our 12 data-generating scenarios (3 sample sizes $\times$ 4 functions) we generated 100 sample datasets and then estimated the HPR model on each dataset using either HMC or VI. For the two estimation approaches, we compared their mean point estimates, their efficiency (the standard deviation of the point estimates), their mean credible interval width, and their credible interval coverage, aggregated pointwise at each timepoint across the 100 replicates of each data-generating scenario. These results are given in Figures I.1-I.4.

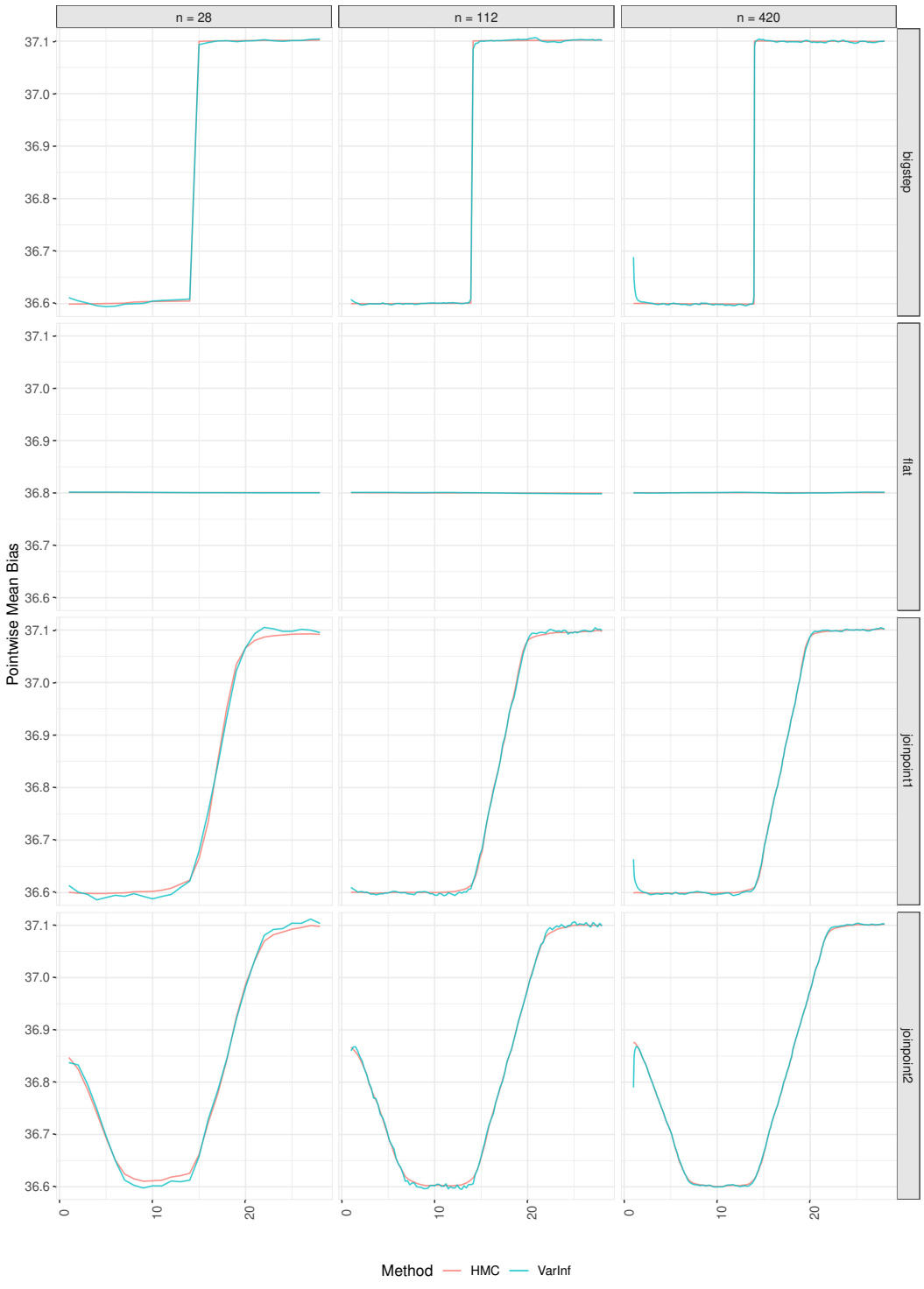

Figure I.1: Mean of the Hamiltonian Monte Carlo (HMC) and variational inference (VarInf) point estimators for horseshoe process regression at each timepoint, aggregated across 100 replicates.

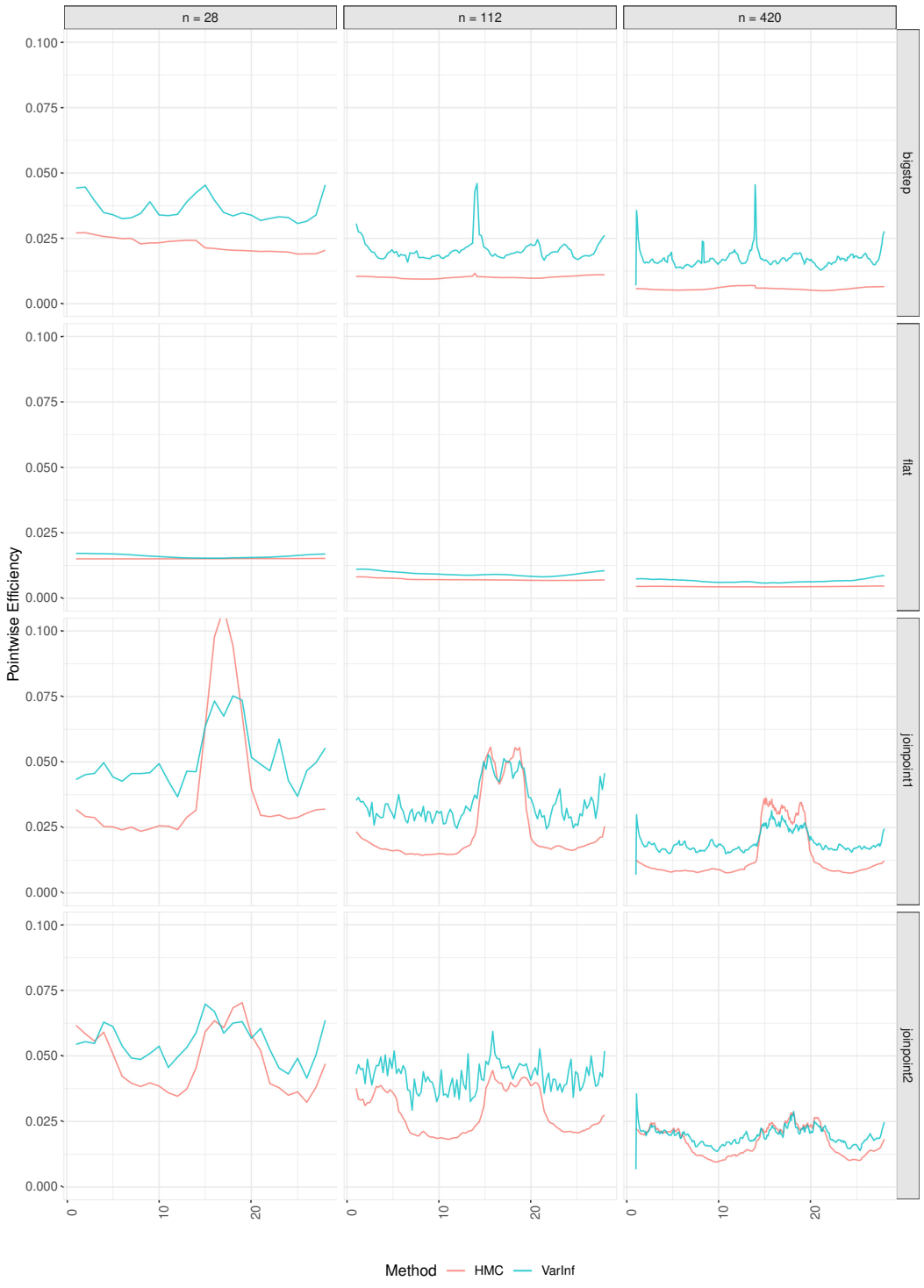

Figure I.2: Standard deviation of the Hamiltonian Monte Carlo (HMC) and variational inference (VarInf) point estimators for horseshoe process regression at each timepoint, aggregated across 100 replicates.

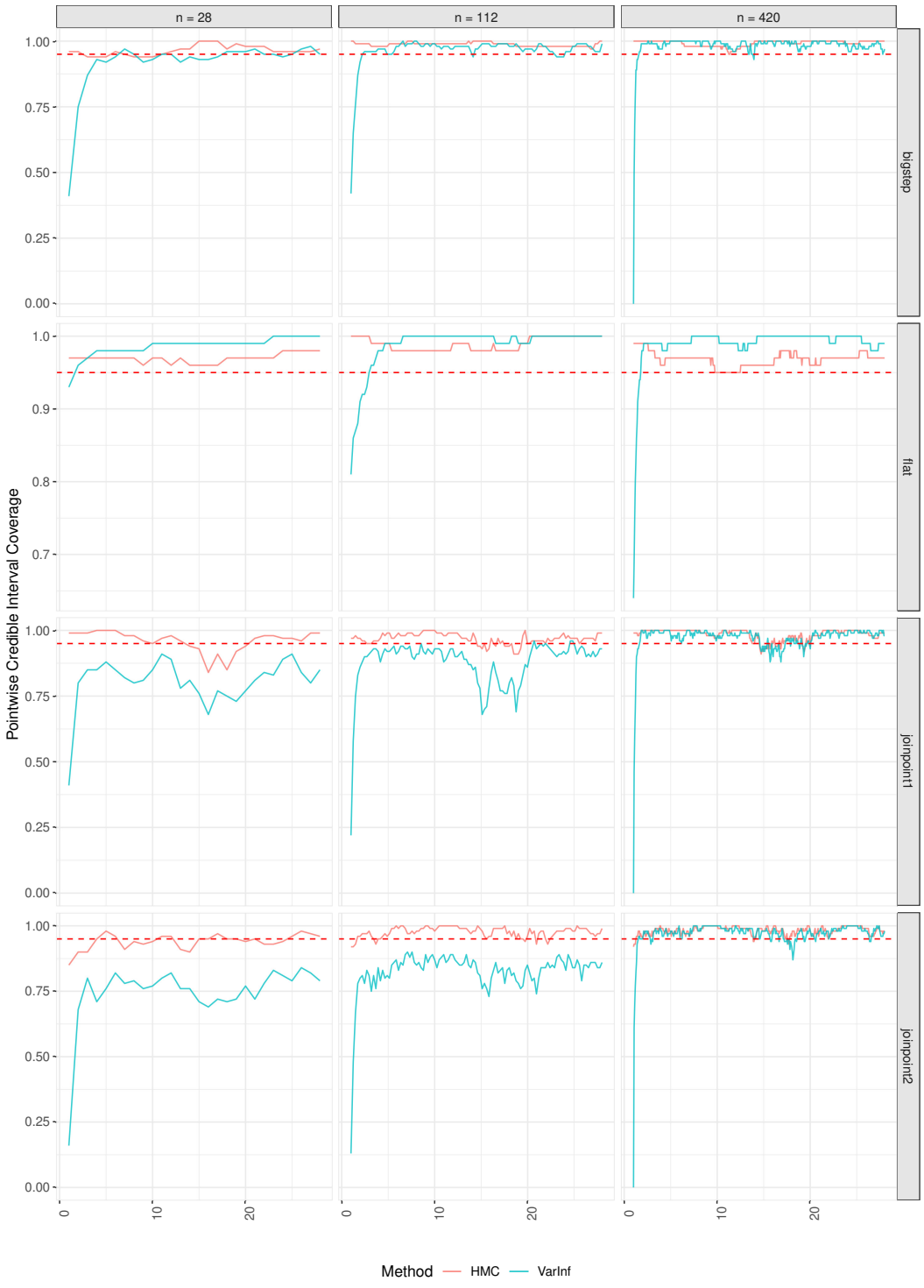

Figure I.3: Coverage of the Hamiltonian Monte Carlo (HMC) and variational inference (VarInf) 95% credible intervals for horseshoe process regression at each timepoint, aggregated across 100 replicates.

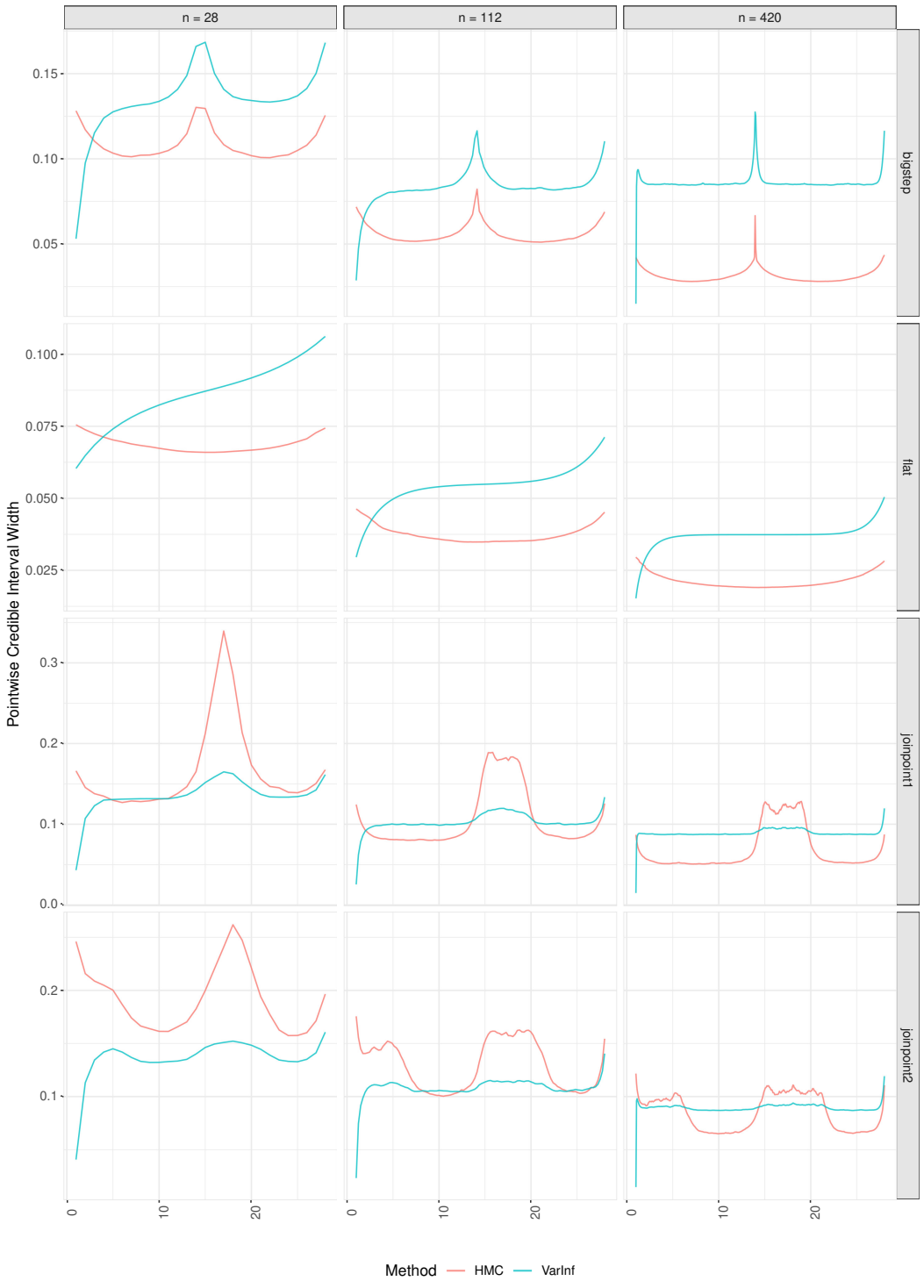

Figure I.4: Width of the Hamiltonian Monte Carlo (HMC) and variational inference (VarInf) 95% credible intervals for horseshoe process regression at each timepoint, aggregated across 100 replicates.

APPENDIX J

Variational Inference and Hamiltonian Monte Carlo Posterior Comparison

In Figure J.1, we show the posterior densities of α, σ^2, τ^2 obtained via either Hamiltonian Monte Carlo (HMC) or variational inference (VI). We show a single simulated dataset from each of our 4 data-generating scenarios (a step function, "bigstep"; a flat line, "flat"; and two piecewise linear functions, which we denote "joinpoint1" and "joinpoint2") with a sample size of $n = 112$. As we can see, there are major discrepancies between the posteriors obtained via HMC and VI. These differences are fairly minor for σ^2, but they are severe for τ^2 and, to a lesser extent, α. Although these differences in the posteriors do not seem to affect the estimates of the temperature trajectory itself (as was shown in Appendix I), they do limit our ability to consider posterior summaries of the parameters themselves.

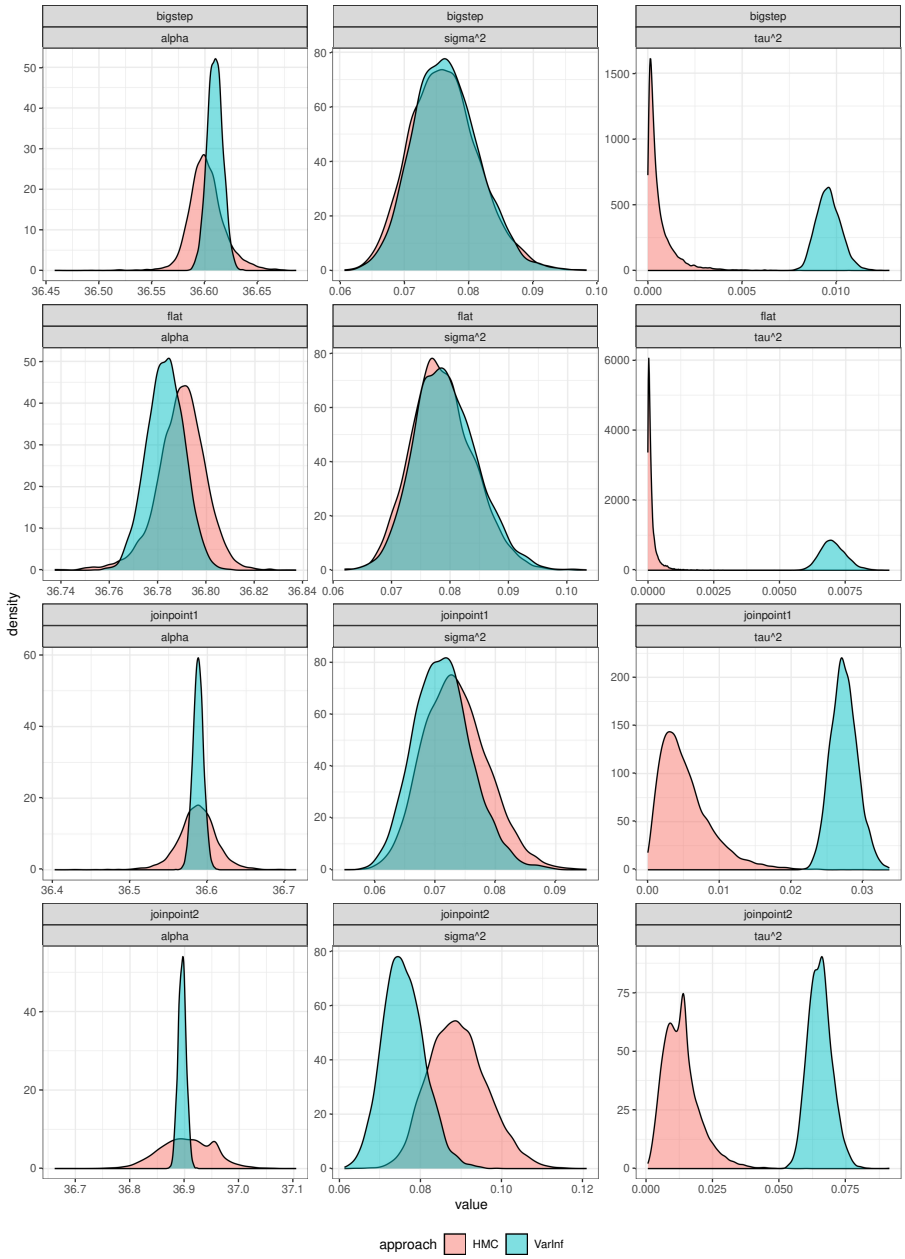

Figure J.1: Comparison of posterior densities obtained from variational inference (VarInf) and Hamiltonian Monte Carlo (HMC) algorithms.Posteriors are shown for three parameters (α, σ^2, τ^2; the columns) in a single replicate of simulated data from four data-generating scenarios (the rows) with sample size $n = 112$.

APPENDIX K

Variational Inference for HPR-BBT

Let y_i be the basal body temperature (BBT) measurement observed on day t_i, $i = 1, ..., m$ of a single menstrual cycle. The updated model is:

$$\begin{aligned}
y_i &= f_i + \epsilon_i \\
f_i &= \alpha + H_i \\
H_i - H_{i-1}|\tau^2, \lambda_i^2 &\sim N(0, \tau^2\lambda_i^2(t_i - t_{i-1})),\ i = 2, ..., m,\ H_1 = 0 \\
\alpha &\sim N(a, b^2) \\
\tau^2|a_\tau &\sim Inv\chi^2(1, 1/a_\tau),\ a_\tau \sim Inv\chi^2(\kappa_\tau, s_\tau) \\
\lambda_i^2|a_{\lambda_i} &\overset{iid}{\sim} Inv\chi^2(1, 1/a_{\lambda_i}) \\
a_{\lambda_i}|O &\sim Inv\chi^2(1, 1),\ i = 2, ..., O, O + 2, ..., m \\
a_{\lambda_{O+1}}|O &\sim Inv\chi^2(1, \frac{1}{4}) \\
\epsilon_i|\sigma^2 &\sim N(0, \sigma^2),\ \sigma^2|a_\sigma \sim Inv\chi^2(1, 1/a_\sigma),\ a_\sigma \sim Inv\chi^2(\kappa_\sigma, s_\sigma) \\
O &\sim Multinom(\boldsymbol{\Psi})
\end{aligned} \tag{K.1}$$

In this model, $a, b^2, \kappa_\tau, s_\tau, \kappa_\sigma, s_\sigma, \boldsymbol{\Psi}$ are hyperparameters that must be specified. Our parameters are $\boldsymbol{\theta} = (\alpha, \boldsymbol{H}, \tau^2, \tau, \boldsymbol{\Lambda}, \boldsymbol{a_\lambda}, \sigma^2, a_\sigma, O)$. Our data are $\boldsymbol{X} = (\boldsymbol{y}, \boldsymbol{t})$. The joint log-likelihood of our model is:

$$\begin{aligned}
\ln p(\boldsymbol{X}, \boldsymbol{\theta}) = &\ln p(\boldsymbol{y}|\boldsymbol{H}, \sigma^2, \alpha) + \ln p(\boldsymbol{H}|\boldsymbol{\Lambda}, \tau^2) + \ln p(\alpha) + \ln p(\sigma^2|a_\sigma) \\
&+ \ln p(a_\sigma) + \ln p(\tau^2|a_\tau) + \ln p(a_\tau) \\
&+ \sum_{i=2}^{m} [\ln p(\lambda_i^2|a_{\lambda_i}) + \ln p(a_{\lambda_i}|O)] + \ln p(O)
\end{aligned} \tag{K.2}$$

We make the following mean field assumption:

$$q(\alpha, \boldsymbol{H}, \tau^2, a_\tau, \boldsymbol{\Lambda}, \boldsymbol{a_\lambda}, \sigma^2, a_\sigma, O) = q(\alpha)q(\boldsymbol{H})q(\tau^2)q(a_\tau)q(\sigma^2)q(a_\sigma)\prod_{i=2}^{m} q(\lambda_i^2)q(a_{\lambda_i})q(O) \quad \text{(K.3)}$$

Reviewing the updated likelihood and mean-field assumption, it is clear that the addition of O, the ovulation day parameter, to the model will only affect the q-densities of $q(\boldsymbol{a_\lambda})$ and $q(O)$. This will also cause changes to the variational objective, L. However, all other parameters' q-densities will be unchanged from Appendix H above.

K.1 Finding $q(O)$

We write the log of the joint-likelihood, restricting our attention to O:

$$\sum_{i=2}^{m} \ln p(a_{\lambda_i}|O) + \ln p(O)$$

Then the q-density is:

$$\begin{aligned} q(O) &\propto \exp\left\{ \sum_{i=2}^{m} I(i \neq O+1)\Big[-\frac{3}{2}E_{a_{\lambda_i}}[\ln a_{\lambda_i}] - \frac{1}{2}E_{a_{\lambda_i}}[\frac{1}{a_{\lambda_i}}]\Big] \right. \\ &\quad \left. + I(i = O+1)\Big[-\frac{3}{2}E_{a_{\lambda_i}}[\ln a_{\lambda_i}] - \frac{1}{8}E_{a_{\lambda_i}}[\frac{1}{a_{\lambda_i}}]\Big] + \ln \Psi_O \right\} \\ &\propto \Psi_O \exp\left\{ -\sum_{i=2}^{m} E_{a_{\lambda_i}}[\frac{1}{a_{\lambda_i}}]\Big(\frac{1}{2}I(i \neq O+1) + \frac{1}{8}I(i = O+1)\Big) \right\} \end{aligned}$$

We recognize this as the kernel of a multinomial distribution in which the probability that $O = o$ is given by $\Psi_o \exp\Big\{ -\sum_{i=2}^{m} E_{a_{\lambda_i}}[\frac{1}{a_{\lambda_i}}]\Big(\frac{1}{2}I(i \neq o+1) + \frac{1}{8}I(i = o+1)\Big)\Big\}$, with all probabilities subsequently normalized to sum to 1.

K.2 Modified $q(a_{\lambda_i})$

We write the log of the joint-likelihood, restricting our attention to a_{λ_i}:

$$\ln p(\lambda_i^2|a_{\lambda_i}) + \ln p(a_{\lambda_i}|O)$$

Then the q-density is:

$$\begin{aligned}
q(a_{\lambda_i}) \propto \exp\Bigg\{ & \frac{1}{2}\ln\frac{1}{2a_{\lambda_i}} - \ln\Gamma(\frac{1}{2}) - \frac{3}{2}E_\lambda[\ln\lambda_i^2] - \frac{1}{2a_{\lambda_i}}E_\lambda[\frac{1}{\lambda_i^2}] \\
& + E_O[I(i \neq O+1)]\Big[\frac{1}{2}\ln\frac{1}{2} - \ln\Gamma(\frac{1}{2}) - \frac{3}{2}\ln a_{\lambda_i} - \frac{1}{2a_{\lambda_i}}\Big] \\
& + E_O[I(i = O+1)]\Big[\frac{1}{2}\ln\frac{1}{8} - \ln\Gamma(\frac{1}{2}) - \frac{3}{2}\ln a_{\lambda_i} - \frac{1}{8a_{\lambda_i}}\Big]\Bigg\} \\
\propto \exp\Bigg\{ & -\frac{1}{2}\ln a_{\lambda_i} - \frac{1}{2a_{\lambda_i}}E_\lambda[\frac{1}{\lambda_i^2}] + Pr(O \geq i, O \leq i-2)\Big[-\frac{3}{2}\ln a_{\lambda_i} - \frac{1}{2a_{\lambda_i}}\Big] \\
& + Pr(O = i-1)\Big[-\frac{3}{2}\ln a_{\lambda_i} - \frac{1}{8a_{\lambda_i}}\Big]\Bigg\} \\
\propto a_{\lambda_i}^{-2}\exp\Bigg\{ & \frac{-1}{2a_{\lambda_i}}[E_\lambda(\frac{1}{\lambda_i^2}) + Pr(O \geq i, O \leq i-2) + \frac{1}{4}Pr(O = i-1)]\Bigg\}
\end{aligned}$$

We recognize this as the kernel of an Inverse-χ^2 distribution with shape parameter $\kappa = 2$ and scale parameter $s = E_\lambda(\frac{1}{\lambda_i^2}) + Pr(O \geq i, O \leq i-2) + \frac{1}{4}Pr(O = i-1)$.

K.3 Modified variational objective

Under this new model, the variational objective is largely the same as before, with some modifications to incorporate O and the changes to $\boldsymbol{a_\lambda}$. The updated L is (with the main modification marked in red):

$$
\begin{aligned}
L = & -\frac{1}{2}E_{\sigma^2}[\sigma^{-2}]E_{\alpha,H}[(\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{H})^T(\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{H})] \\
& -\frac{1}{2}(E_H[\boldsymbol{H}^{*T}E_R[\boldsymbol{R}]\boldsymbol{H}^*])-\frac{1}{2b^2}E_\alpha[(\alpha-a)^2] \\
& -\frac{1}{2}E_{a_\sigma}[\frac{1}{a_\sigma}](E_{\sigma^2}[\frac{1}{\sigma^2}]+s_\sigma)-\frac{1}{2}E_{a_\tau}[\frac{1}{a_\tau}](E_{\tau^2}[\frac{1}{\tau_h^2}]+s_\tau)-\frac{1}{2}\sum_{i=2}^{n}E_{a_\lambda}[\frac{1}{a_{\lambda_i}}]E_\lambda[\frac{1}{\lambda_i^2}] \\
& -\frac{1}{2}\ln(nE_{\sigma^2}[\frac{1}{\sigma^2}]+\frac{1}{b^2})-\frac{1}{2}\ln|E_{\sigma^2}[\frac{1}{\sigma^2}]\boldsymbol{I}_{n-1\times n-1}+E_R[\boldsymbol{R}]| \\
& -\frac{n}{2}\ln\frac{1}{2}\Big(E_{a_\tau}[\frac{1}{a_\tau}]+\sum_{i=2}^{n}E_H[(H_i-H_{i-1})^2]E_\lambda[1/\lambda_i^2]1/\delta_i\Big) \\
& -\frac{(\kappa_\tau+1)}{2}\ln(\frac{1}{2}(s_\tau+E_{\tau_h^2}[\frac{1}{\tau_h^2}]))-\sum_{i=2}^{n}\ln\frac{1}{2}\Big(E_{a_{\lambda_i}}[\frac{1}{a_{\lambda_i}}]+\frac{E_H[(H_i-H_{i-1})^2]}{\delta_i}E_{\tau_h^2}[\frac{1}{\tau_h^2}]\Big) \\
& -\frac{(\kappa_\sigma+1)}{2}\ln\frac{1}{2}(s_\sigma+E_{\sigma^2}[\frac{1}{\sigma^2}])-\frac{n+1}{2}\ln\frac{1}{2}\Big(E_{a_\sigma}[\frac{1}{a_\sigma}]+E_{\alpha,H}[(\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{H})^T(\boldsymbol{y}-\alpha\mathbf{1}_n-\boldsymbol{H})]\Big) \\
& -\sum_{i=2}^{m}\ln\frac{1}{2}\Big[E_\lambda(\frac{1}{\lambda_i^2})+Pr(O\geq i,O\leq i-2)+\frac{1}{4}Pr(O=i-1)\Big]
\end{aligned}
$$

K.4 Modified algorithm structure

The modified algorithm is (with modifications marked in red):

Algorithm 2 Variational inference algorithm for HPR for BBT (HPR-BBT).

Inputs:

$\boldsymbol{X} = (\boldsymbol{y}, \boldsymbol{t})$

Initialize:

$E(\alpha)_0 \leftarrow \bar{y}_{1:10}$, $E(\boldsymbol{H})_0 \leftarrow \vec{0}$

$E(\frac{1}{\sigma^2})_0 \leftarrow \frac{1}{var(\boldsymbol{y})}$, $E(\frac{1}{a_\sigma})_0 \leftarrow 1$

$E(\frac{1}{\tau^2})_0 \leftarrow \frac{100}{var(\boldsymbol{y})}$, $E(\frac{1}{a_\tau})_0 \leftarrow 1$

$E(\frac{1}{\lambda_i^2})_0 \leftarrow 100$, $E(\frac{1}{a_{\lambda_i}})_0 \leftarrow 1,\ i = 2, \ldots, m$

$j \leftarrow 1$

while $j \leq 1000 \quad \& \quad \frac{|L_j - L_{j-1}|}{|L_j|} > 0.0001$ **do**

$E(\boldsymbol{H})_j \leftarrow (E(\frac{1}{\sigma^2})_{j-1}\boldsymbol{I}_{m-1\times m-1} + E[\boldsymbol{R}]_{j-1})^{-1}E(\frac{1}{\sigma^2})_{j-1}(\boldsymbol{y}^{*\boldsymbol{T}} - E(\alpha)_{j-1}\boldsymbol{1}_{m-1}^T)^T$

$E(\alpha)_j \leftarrow [E(\frac{1}{\sigma^2})_{j-1}(\boldsymbol{y}^T\boldsymbol{1}_m - E(\boldsymbol{H})_j^T\boldsymbol{1}_m) + \frac{a}{b^2}][mE(\frac{1}{\sigma^2})_{j-1} + b^{-2}]^{-1}$

$E(\frac{1}{a_\sigma})_j \leftarrow (\kappa_\sigma + 1)/(E(\frac{1}{\sigma^2})_{j-1} + s_\sigma)$

$E(\frac{1}{\sigma^2})_j \leftarrow (m+1)(E[(\boldsymbol{y} - \alpha\boldsymbol{1}_m - \boldsymbol{H})^T(\boldsymbol{y} - \alpha\boldsymbol{1}_m - \boldsymbol{H})]_j + E(\frac{1}{a_\sigma})_j)$

$E(\frac{1}{a_\tau})_j \leftarrow (\kappa_\tau + 1)/(E(\frac{1}{\tau^2})_{j-1} + s_\tau)$

$E(\frac{1}{\tau^2})_j \leftarrow m/(\sum_{i=2}^m \frac{E[(H_i - H_{i-1})^2]_j E(\frac{1}{\lambda_i^2})_{j-1}}{\delta_i} + E(\frac{1}{a_\tau})_j)$

$Pr(O = o)_j \leftarrow \Psi_o \exp\{-\sum_{i=2}^m E[\frac{1}{a_{\lambda_i}}](\frac{1}{2}I(i \neq o+1) + \frac{1}{8}I(i = o+1))\}$

$E(\frac{1}{a_{\lambda_i}})_j \leftarrow 2/E(\frac{1}{\lambda_i^2}) + Pr(O \geq i, O \leq i-2) + \frac{1}{4}Pr(O = i-1),\ i = 2, \ldots, m$

$E(\frac{1}{\lambda_i^2})_j \leftarrow 2/(\frac{1}{\delta_i}E(\frac{1}{\tau^2})_j E[(H_i - H_{i-1})^2]_j + E(\frac{1}{a_{\lambda_i}})_j),\ i = 2, \ldots, m$

$L_j \leftarrow L[E(\boldsymbol{H})_j, E(\alpha)_j, E(\frac{1}{a_\sigma})_j, E(\frac{1}{\sigma^2})_j, E(\frac{1}{a_\tau})_j, E(\frac{1}{\tau^2})_j, E(\frac{1}{a_{\lambda_2}})_j, \ldots, E(\frac{1}{a_{\lambda_m}})_j, E(\frac{1}{\lambda_2^2})_j, \ldots, E(\frac{1}{\lambda_m^2})_j, Pr(O)_j]$

$j \leftarrow j + 1$

end while

APPENDIX L

More Information on Data Analysis

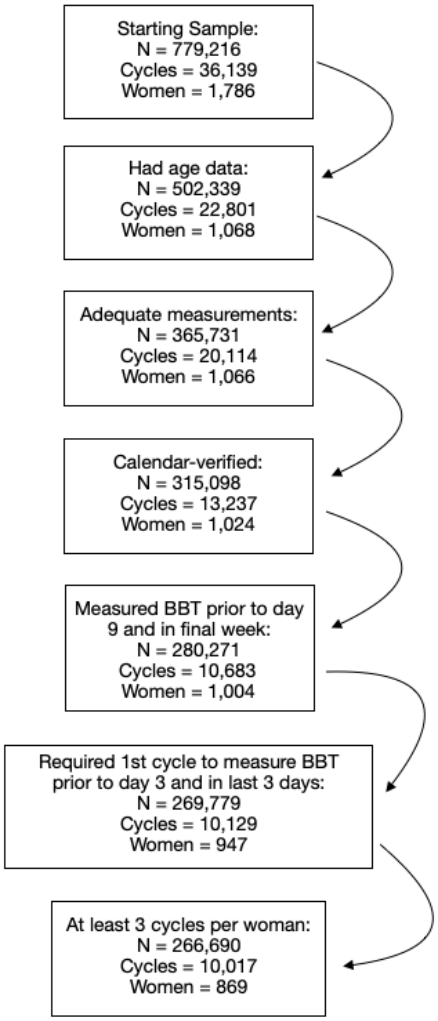

Figure L.1: Sample exclusions to move from a starting sample of 779,216 basal body temperature (BBT) measurements to a final sample of 266,690 BBT measurements.

APPENDIX M

Proof of Unbiasedness of Kaplan-Meier and Risk Set Imputation Estimators

We will now show that our two imputation estimators are unbiased for traditional cumulative incidence estimation.

First, a review of the notation: Let T_i denote the time to some outcome of interest for $i = 1, ..., n$ subjects. Let C_i denote the corresponding time to censoring. Let V_i be the time to a competing event for individuals $i = 1, ..., n$. Then the observed data are $X_i = min(T_i, V_i, C_i)$ with $\delta_i = 1$ when $X_i = T_i$, $\delta_i = 2$ when $X_i = V_i$, and $\delta_i = 0$ when $X_i = C_i$. Let $t_1, t_2, ..., t_l$ be the ordered, unique values of the *event times*, e.g. times of individuals with $\delta_i = 1, 2$. Let $d_j, j = 1, ..., l$ be the number of events of interest ($\delta_i = 1$) observed at time t_j, and let v_j be the number of competing events ($\delta_i = 2$) observed at time t_j. Let c_j be the number of individuals censored in the interval $[t_j, t_{j+1})$. Let y_j be the number at-risk just before time t_j, e.g. $y_j = \sum_{i=1}^{n} I(X_i \geq t_j)$.

Then we wish to show:

$$E[\hat{F}_{imp}(t)] = \hat{F}(t) \tag{M.1}$$

where $\hat{F}_{imp}(t)$ is our imputation estimator of the cumulative incidence function, $\hat{F}(t)$ is the Aalen-Johansen estimator of the cumulative incidence, and the expectation is taken over the imputations. We will start with $\hat{F}_{imp}(t)$ standing in for the estimator from our Kaplan-Meier imputation (KMI) approach, based on M imputations.

Note that:

$$\begin{aligned} E[\hat{F}_{imp}(t)] &= E[\frac{1}{M}\sum_{m=1}^{M}\hat{F}_{imp}^{(m)}(t)] \\ &= E[\hat{F}_{imp}^{(m)}(t)] \\ &= E[\frac{1}{n}\sum_{t_i\leq t} d_i + d_i^{*(m)}] \\ &= \frac{1}{n}\sum_{t_i\leq t} d_i + E[d_i^{*(m)}] \end{aligned} \tag{M.2}$$

Based on our imputation scheme, the expected value of $d_i^{*(m)}$ is

$$\frac{d_i[S(t_{i-1}) - S(t_i)]}{d_i + v_i}\sum_{j=0}^{i-1}\frac{c_j}{S(t_j)} \tag{M.3}$$

where $\frac{d_i}{d_i+v_i}$ is $Pr(\boldsymbol{\delta}|\boldsymbol{X})$, while $\sum_{j=0}^{i-1}\frac{c_j}{S(t_j)}$ counts the number of individuals censored before time t_i, scaling their contribution by the number of individuals still alive at the time of censoring. $[S(t_{i-1}) - S(t_i)]$ gives the probability of being imputed to event time t_i.

Therefore, to demonstrate that our imputation estimator is unbiased, we must show:

$$\hat{F}(t) = E[\hat{F}_{imp}(t)] = \frac{1}{n}\sum_{t_i\leq t} d_i + \frac{d_i[S(t_{i-1}) - S(t_i)]}{d_i + v_i}\sum_{j=0}^{i-1}\frac{c_j}{S(t_j)} \tag{M.4}$$

We will focus on the term within the sum. Leveraging the recursive property of the Kaplan-Meier estimator, we note that:

$$\begin{aligned} d_i + \frac{d_i[S(t_{i-1}) - S(t_i)]}{d_i + v_i}\sum_{j=0}^{i-1}\frac{c_j}{S(t_j)} &= d_i + \frac{d_i S(t_{i-1})[1 - (1 - \frac{d_i+v_i}{y_i})]}{d_i + v_i}\sum_{j=0}^{i-1}\frac{c_j}{S(t_j)} \\ &= d_i + \frac{d_i S(t_{i-1})}{y_i}\sum_{j=0}^{i-1}\frac{c_j}{S(t_j)} \end{aligned} \tag{M.5}$$

We pull out the term $\frac{d_i S(t_{i-1})}{y_i}$, i.e. the summand for the Aalen-Johansen estimator. This yields:

$$\frac{d_i S(t_{i-1})}{y_i}\Big[\frac{y_i}{S(t_{i-1})} + \sum_{j=0}^{i-1}\frac{c_j}{S(t_j)}\Big] \tag{M.6}$$

Reviewing our work, we now have:

$$E[\hat{F}_{imp}(t)] = \frac{1}{n}\sum_{t_i \le t} \frac{d_i S(t_{i-1})}{y_i}\Big[\frac{y_i}{S(t_{i-1})} + \sum_{j=0}^{i-1}\frac{c_j}{S(t_j)}\Big] \tag{M.7}$$

Studying this, we see that this will collapse to the Aalen-Johansen estimator if we can show that $\left[\frac{y_i}{S(t_{i-1})} + \sum_{j=0}^{i-1}\frac{c_j}{S(t_j)}\right] = n$. Let us examine this bracketed term further. First, note that $y_i = y_{i-1} - d_{i-1} - v_{i-1} - c_{i-1}$. So $c_j = y_j - d_j - v_j - y_{j+1}$. Plugging this in, we have:

$$\begin{aligned}
\frac{y_i}{S(t_{i-1})} + \sum_{j=0}^{i-1}\frac{c_j}{S(t_j)} &= \frac{y_i}{S(t_{i-1})} + \sum_{j=0}^{i-1}\frac{y_j - d_j - v_j - y_{j+1}}{S(t_j)} \\
&= \frac{y_i}{S(t_{i-1})} - \sum_{j=0}^{i-1}\frac{y_{j+1}}{S(t_j)} + \sum_{j=0}^{i-1}\frac{y_j - d_j - v_j}{S(t_j)} \\
&= -\sum_{j=0}^{i-2}\frac{y_{j+1}}{S(t_j)} + \sum_{j=0}^{i-1}\frac{y_j - d_j - v_j}{S(t_j)}
\end{aligned} \tag{M.8}$$

Note that $S(t_j) = \prod_{k=1}^{j}(1 - \frac{d_k+v_k}{y_k}) = \prod_{k=1}^{j}(\frac{y_k - d_k - v_k}{y_k})$. Plugging this in, we have:

$$\begin{aligned}
-\sum_{j=0}^{i-2}\frac{y_{j+1}}{S(t_j)} + \sum_{j=0}^{i-1}\frac{y_j - d_j - v_j}{S(t_j)} &= -\sum_{j=0}^{i-2} y_{j+1}\prod_{k=1}^{j}(\frac{y_k}{y_k - d_k - v_k}) + \sum_{j=0}^{i-1}(y_j - d_j - v_j)\prod_{k=1}^{j}(\frac{y_k}{y_k - d_k - v_k}) \\
&= -\sum_{j=0}^{i-2} y_{j+1}\prod_{k=1}^{j}(\frac{y_k}{y_k - d_k - v_k}) + \sum_{j=0}^{i-1} y_j\prod_{k=1}^{j-1}(\frac{y_k}{y_k - d_k - v_k})
\end{aligned} \tag{M.9}$$

Doing some re-indexing, we have:

$$-\sum_{j=1}^{i-1} y_j\prod_{k=1}^{j-1}(\frac{y_k}{y_k - d_k - v_k}) + \sum_{j=0}^{i-1} y_j\prod_{k=1}^{j-1}(\frac{y_k}{y_k - d_k - v_k}) = y_0 S(t_0) = n \tag{M.10}$$

So, to finish, we have:

$$\begin{aligned} E[\hat{F}_{imp}(t)] &= \frac{1}{n}\sum_{t_i \le t} \frac{d_i S(t_{i-1})}{y_i}\Big[\frac{y_i}{S(t_{i-1})} + \sum_{j=0}^{i-1}\frac{c_j}{S(t_j)}\Big] \\ &= \frac{1}{n}\sum_{t_i \le t} \frac{n d_i S(t_{i-1})}{y_i} \\ &= \sum_{t_i \le t} \frac{d_i S(t_{i-1})}{y_i} \\ &= \hat{F}(t) \end{aligned} \tag{M.11}$$

which completes the proof of the unbiasedness of the Kaplan-Meier imputation scheme. We note that this provides an alternative to the proof presented in Taylor et al. (2002) when v_j is set to 0 [74].

To show the unbiasedness of the risk set imputation estimator (RSI), it suffices to show its equivalence to the Kaplan-Meier imputation estimator. Recall that the expected value of the Kaplan-Meier imputation estimator is:

$$E[\hat{F}_{imp}(t)] = \frac{1}{n}\sum_{t_i \le t} d_i + \frac{d_i[S(t_{i-1}) - S(t_i)]}{d_i + v_i}\sum_{j=0}^{i-1}\frac{c_j}{S(t_j)} \tag{M.12}$$

The only part of this expectation that differs between the risk set imputation estimator and Kaplan-Meier imputation estimator is $[S(t_{i-1}) - S(t_i)]$. Whereas the Kaplan-Meier imputation scheme uses the survival probabilites to allocate censored individuals directly to event times, the risk-set imputation estimator works indirectly. Censored individuals can be allocated to other censored individuals and then reallocated until they ultimately arrive at an individual who had an event. As discussed in Efron (1967), though, the risk set imputation estimator's redistribute-to-the-right approach is equivalent to the Kaplan-Meier estimator in the setting of all-cause survival [19]. Thus the two estimators are the same. Therefore, both the risk set imputation estimator and the Kaplan-Meier imputation estimator will return equivalent results to the Aalen-Johansen cumulative incidence estimator as the number of imputations increases.

APPENDIX N

Sensitivity Analyses for the Hyperparameter of the Bayesian Interval

As was discussed in Section 4.3.3.4, the choice of prior for the Bayesian beta-binomial interval does play a small role in the interval's performance. Here, we present sensitivity analyses motivating our recommendation of a $Beta(0.8, 1.2)$ prior on the cumulative incidence at each time-point.

Following the simulation design described in Section 4.4, we present results here from Scenarios A and F. Recall that Scenario A had largely equal event rates with low censoring, while in Scenario F the competing event was more common and censoring was higher. We also examined performance in the other five scenarios (B-E, G) but the results were similar to what we present here and are thus omitted for the sake of brevity.

On each simulated dataset, we compared the performance of three different priors on the cumulative incidence at each time point:

1. $Beta(0.5, 0.5)$: the Jeffreys prior for a beta-binomial interval.

2. $Beta(0.8, 1.2)$: similar to the Jeffreys prior, but with more mass on the range $[0, 0.75]$ and less in $(0.75, 1]$. We think this is likely sensible in the competing risks setting.

3. A data driven approach: $Beta(a, b)$ in which the first shape parameter a was set to be two times the observed prevalence of the event of interest at the end of follow-up (omitting censored individuals). We bounded this to be between 0.5 and 1. In our existing notation, this would be $a = min(1, max(0.5, \frac{2d_l}{d_l+v_l}))$, recalling that d_l is the number of individuals who died of the event of interest by the last observed time t_l and v_l is the number of individuals who died of the competing event by the last observed time t_l. The second shape parameter b was set to be two minus the first shape parameter, e.g. $b = 2 - min(1, max(0.5, \frac{2d_l}{d_l+v_l}))$.

Results for interval width are given in Figures N.1 and N.2. Results for interval coverage are

given in Table N.1. From this, we note that in general the prior has minimal effect, particularly at larger sample sizes. However, in some settings it does play a role. The $Beta(0.5, 0.5)$ prior returned a slightly wider interval than the other two priors except when the event rate was low, in which case its interval was narrower. The $Beta(0.8, 1.2)$ prior did the exact opposite. Interval coverage was generally good, although in Scenario F, $n = 25$ and $n = 100$, we note that the $Beta(0.5, 0.5)$ and data-driven priors struggled, with excessive conservatism in the $n = 25$ columns and anti-conservatism in the $T = 1, n = 100$ column.

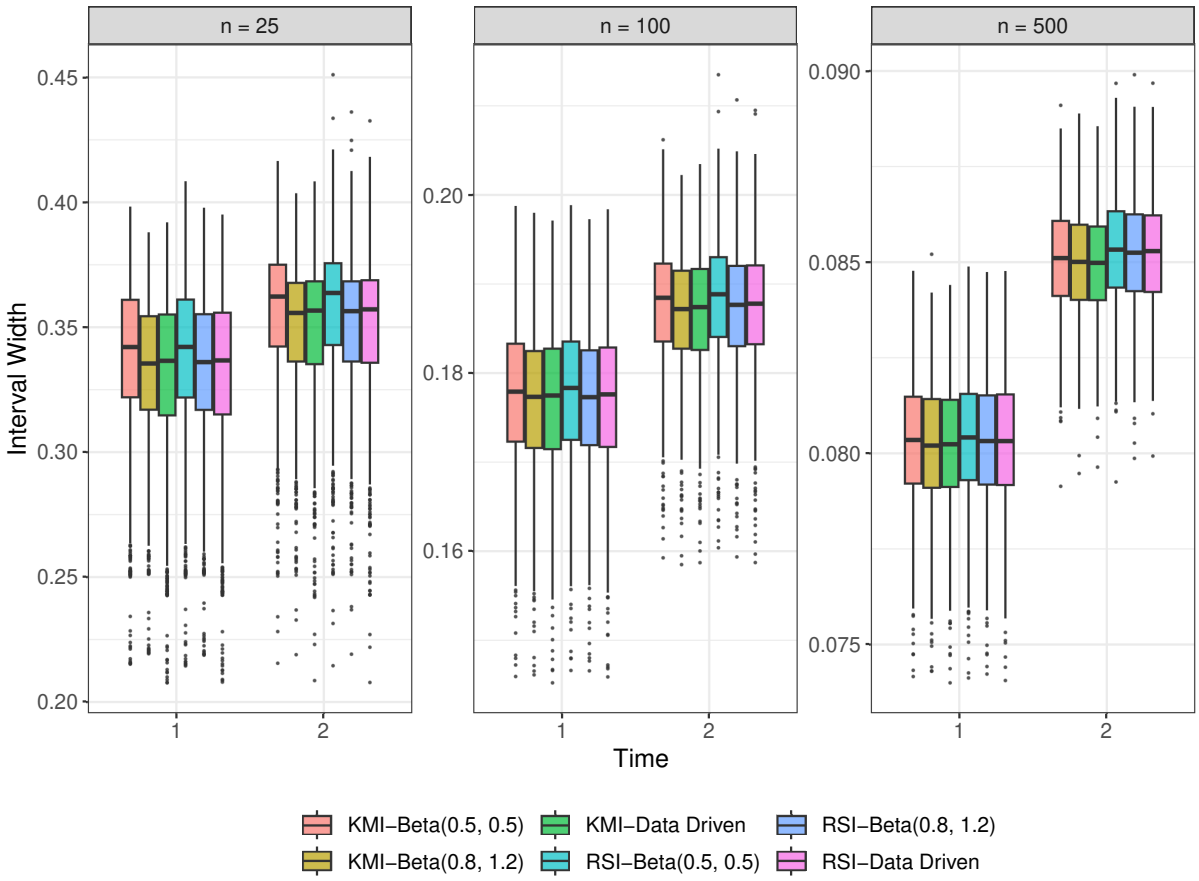

Figure N.1: 95% uncertainty interval widths for the Kaplan-Meier imputation (KMI) and risk set imputation (RSI) Bayesian intervals under three different priors in Scenario A at three sample sizes: $n = 25, 100, 500$.

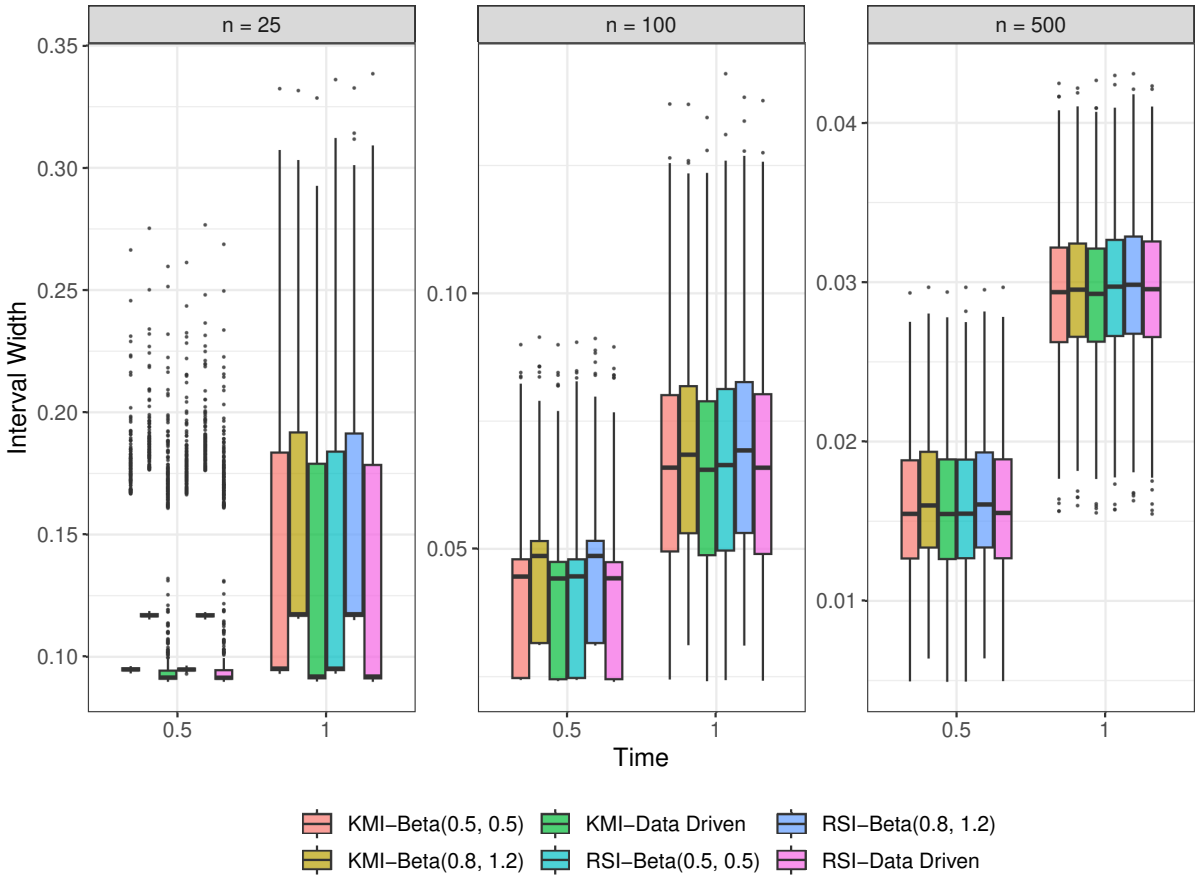

Figure N.2: 95% uncertainty interval widths for the Kaplan-Meier imputation (KMI) and risk set imputation (RSI) Bayesian intervals under three different priors in Scenario F at three sample sizes: $n = 25, 100, 500$.

Table N.1: Coverage rates for 95% uncertainty intervals for the Kaplan-Meier imputation (KMI) and risk set imputation (RSI) Bayesian intervals under three different priors in Scenarios A and F. Scenario A has a low rate of censoring and similar incidence rates for the event of interest and the competing event; Scenario F has a moderate rate of censoring and the competing event is more common than the event of interest.

Method	n = 25		n = 100		n = 500	
Scenario A	Time = 1	Time = 2	Time = 1	Time = 2	Time = 1	Time = 2
KMI-Beta(0.5, 0.5)	0.92	0.94	0.97	0.95	0.95	0.95
KMI-Beta(0.8, 1.2)	0.95	0.95	0.97	0.95	0.95	0.96
KMI-Data Driven	0.92	0.92	0.96	0.94	0.95	0.95
RSI-Beta(0.5, 0.5)	0.92	0.94	0.96	0.95	0.95	0.95
RSI-Beta(0.8, 1.2)	0.95	0.95	0.96	0.95	0.95	0.96
RSI-Data Driven	0.92	0.93	0.96	0.95	0.95	0.96
Scenario F	Time = 0.5	Time = 1	Time = 0.5	Time = 1	Time = 0.5	Time = 1
KMI-Beta(0.5, 0.5)	0.99	0.98	0.97	0.84	0.94	0.93
KMI-Beta(0.8, 1.2)	0.93	0.97	0.97	0.96	0.94	0.94
KMI-Data Driven	0.99	0.98	0.97	0.84	0.94	0.93
RSI-Beta(0.5, 0.5)	0.99	0.98	0.97	0.84	0.94	0.94
RSI-Beta(0.8, 1.2)	0.93	0.97	0.97	0.97	0.94	0.94
RSI-Data Driven	0.99	0.98	0.97	0.84	0.94	0.94

APPENDIX O

Additional Simulation Results for Multiple Imputation Cumulative Incidence Estimator

Figures O.1 and O.2 give the performance of the point estimators at sample sizes of $n = 25$ and $n = 500$, respectively. Results for $n = 100$ were presented in Section 4.4. Generally, results were similar across sample sizes. Performance of all estimators was worse in the $n = 25$ sample size and better at $n = 500$, as we would expect.

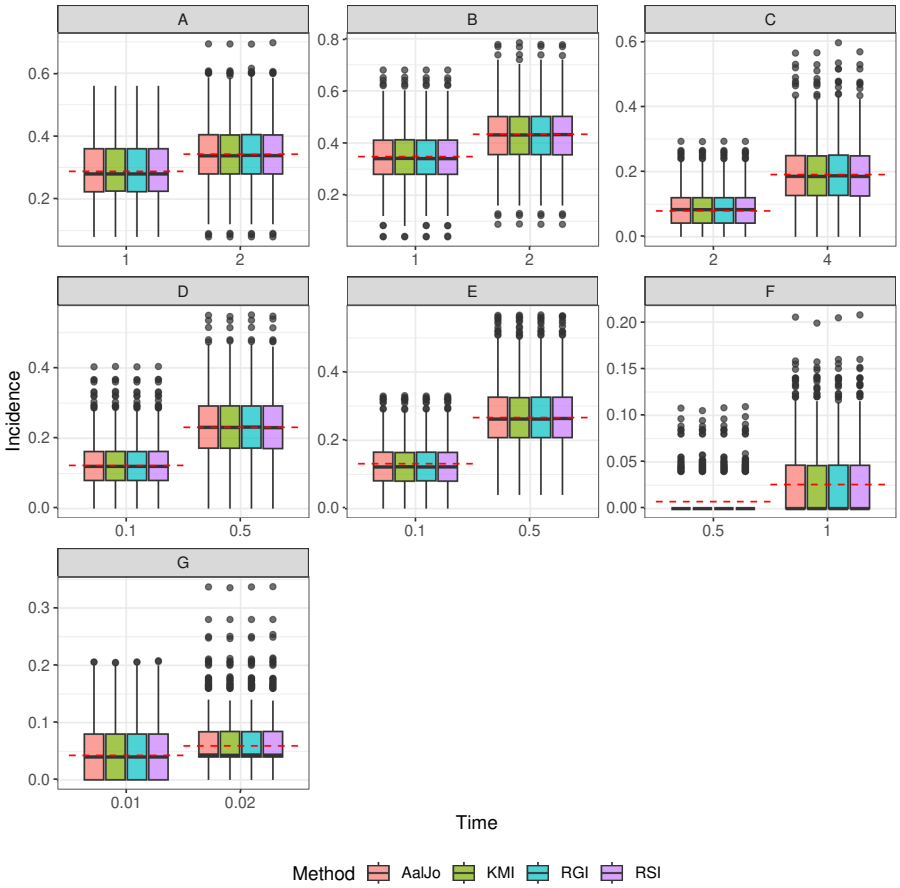

Figure O.1: Point estimator performance for imputation and Aalen-Johansen (AalJo) estimators in seven simulation scenarios with a sample size of $n = 25$. The imputation estimators are Kaplan-Meier imputation (KMI), risk set imputation (RSI), and Ruan-Gray imputation (RGI). The true incidence is marked as a horizontal dashed line. A sample dataset for each scenario is given in Figure 4.4.

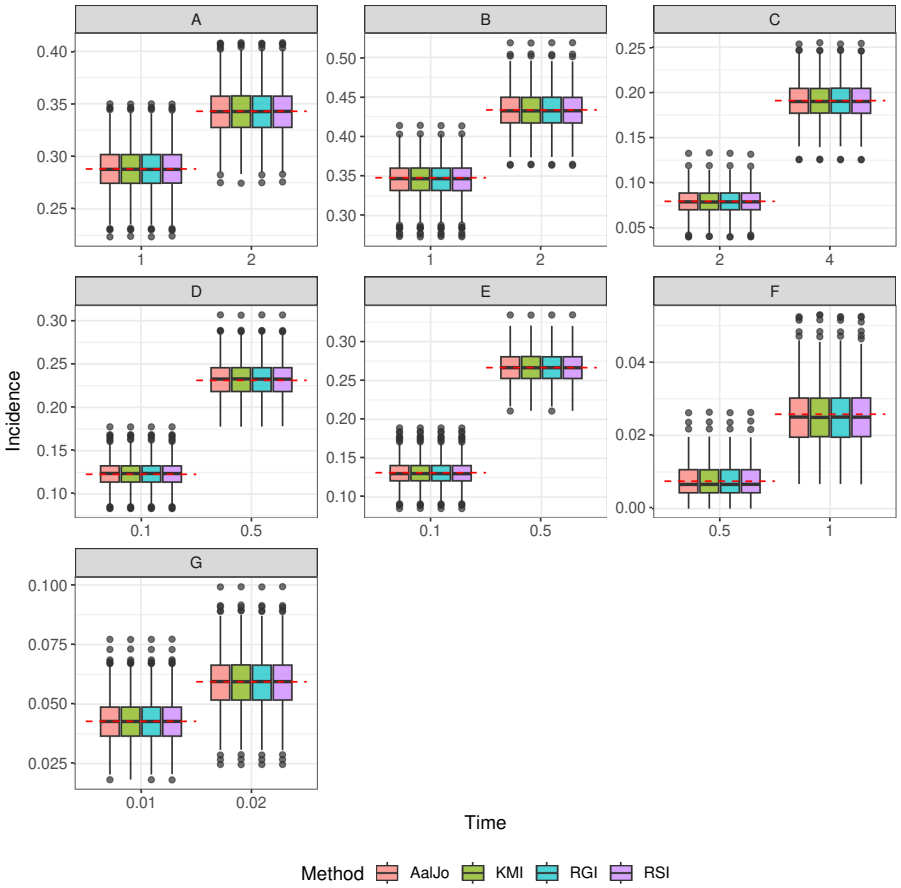

Figure O.2: Point estimator performance for imputation and Aalen-Johansen (AalJo) estimators in seven simulation scenarios with a sample size of $n = 500$. The imputation estimators are Kaplan-Meier imputation (KMI), risk set imputation (RSI), and Ruan-Gray imputation (RGI). The true incidence is marked as a horizontal dashed line. A sample dataset for each scenario is given in Figure 4.4.

Figure O.3 gives the effect of number of imputations in Scenario G, to complement the results given for Scenario A in Section 4.4. Scenario G has more censoring than Scenario A, which is part of why the number of outliers in performance has increased (particularly at $n = 25$ and with smaller numbers of imputations). Still, we see that with $M = 150$ imputations, performance is generally similar between the imputation estimators and Aalen-Johansen at larger sample sizes ($n = 100, 500$) even in the presence of high rates of censoring. For small sample sizes and high rates of censoring it may be advisable to use even more imputations, such as $M = 300$.

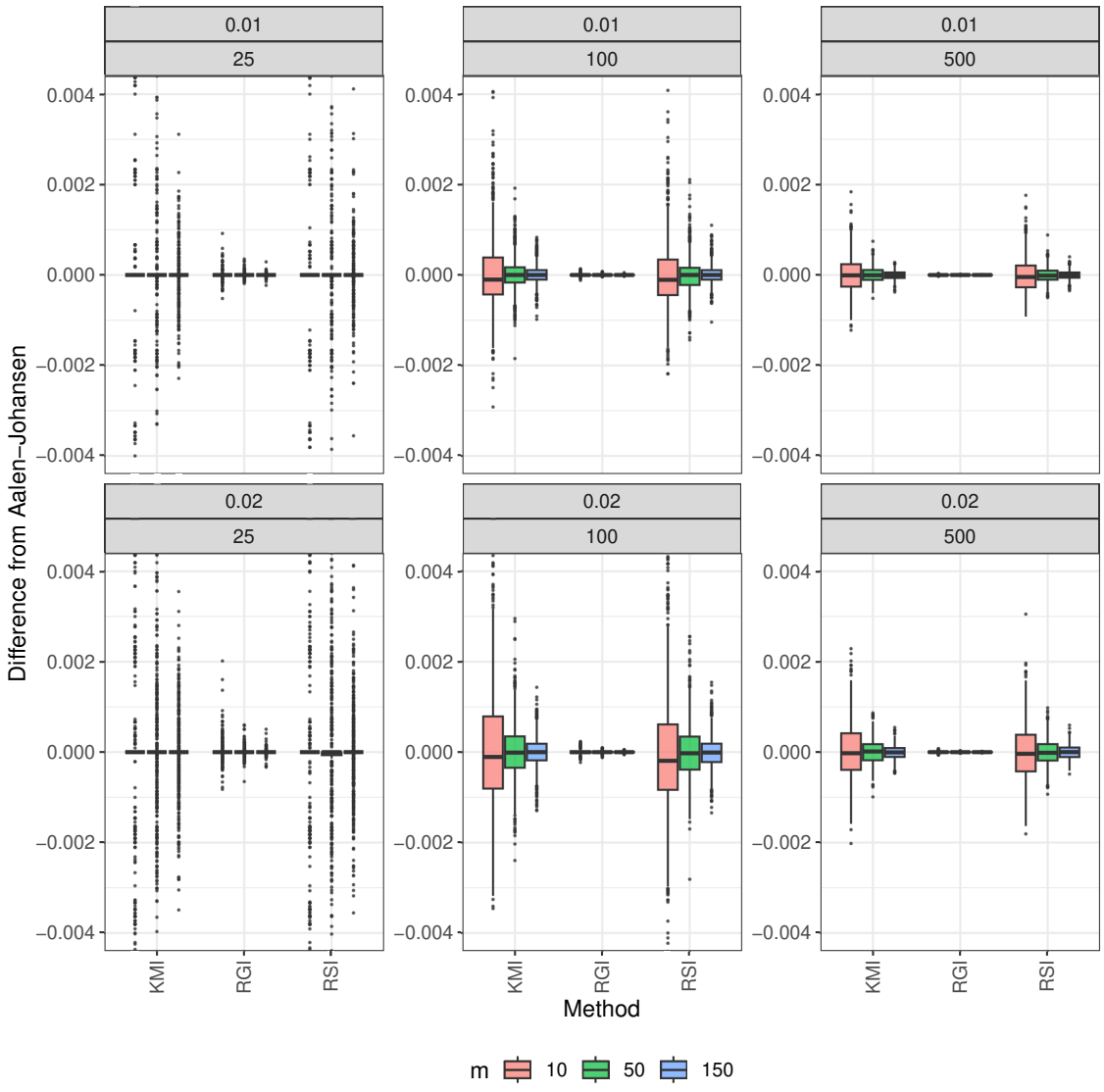

Figure O.3: The difference between each imputation point estimator and the Aalen-Johansen point estimator at varying sample sizes and number of imputations in Scenario G. Note that the $n = 25$ plots have had their y-axis truncated to improve readability—there were additional outliers that fell outside of the range shown here.

We also considered the computational time of the imputation approaches, and how they are affected by the number of imputations. These results are given in Figure O.4.

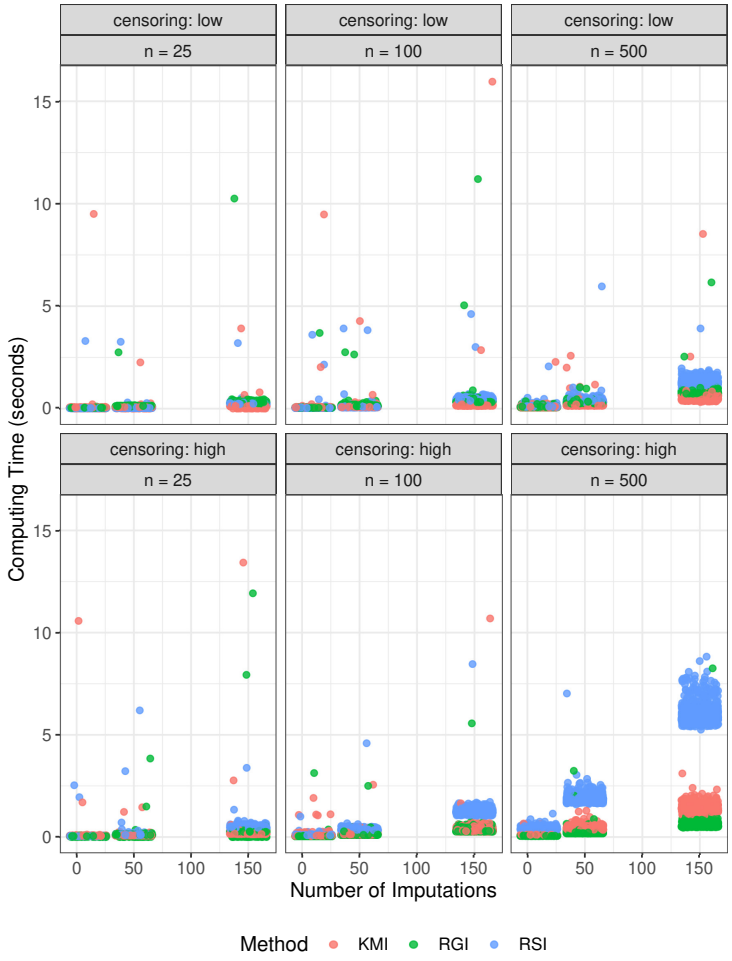

Figure O.4: Computing time for the imputation estimators. The top row presents results for Scenario A, which had low censoring; the bottom row presents results for Scenario G, which had high censoring. Two outliers with computation time of about 60 seconds from the RGI approach were omitted from this plot for readability.

Although increasing the number of imputations did increase computational time, overall the methods still ran quite quickly, even at large sample sizes and with large numbers of imputations. We also note that of the three imputation approaches, RGI is the fastest computationally, followed by KMI, with RSI as the slowest. However, this may be caused by the data-generating mechanisms we consider here—none of the simulated datasets feature extremely high rates of the competing event, which is where RGI is most likely to struggle. Regardless, the differences in computing time between the three approaches were still very modest, and we do not think computing time would be a barrier for any of the approaches.

BIBLIOGRAPHY

[1] Technical report, series no. 1967, 12. Technical report, World Health Organization, Geneva, 1976.

[2] OO Aalen and S Johansen. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5(3):141–150, 1978.

[3] A Agresti and BA Coull. Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, 1998.

[4] A Allignol, M Schumacher, and J Beyersmann. A note on variance estimation of the Aalen-Johansen estimator of the cumulative incidence function in competing risks, with a view towards left-truncated data. *Biometrical Journal*, 52(1):126–137, 2010.

[5] M Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv: 1701.02434*, 2017.

[6] RA Betensky and DA Schoenfeld. Nonparametric estimation in a cure model with random cure times. *Biometrics*, 57:282–286, 2001.

[7] N Binder, TA Gerds, and PK Andersen. Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime Data Analysis*, 20:303–315, 2014.

[8] PS Boonstra, DR Owen, and J Kang. Shrinkage priors for isotonic probability vectors and binary data modeling. *In press*, 2021.

[9] R Boyle. Ancient humans used the moon as a calendar in the sky. *Science News*, July 2019.

[10] TM Braun and Z Yuan. Comparing the small sample performance of several variance estimators under competing risks. *Statistics in Medicine*, 26:1170–1180, 2007.

[11] JR Bull, SP Rowland, E Berglund Scherwitzl, R Scherwitzl, K Gemzell Danielsson, and J Harper. Real-world menstrual cycle characteristics of more than 600,000 menstrual cycles. *npj Digital Medicine*, 2(83), 2019.

[12] CM Carvalho, NG Polson, and JG Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[13] W Chen, M Kitazawa, and T Togawa. Estimation of the biphasic property in a female's menstrual cycle from cutaneous temperature measured during sleep. *Annals of Biomedical Engineering*, 37(9):1827–1838, 2009.

[14] YT Chen, WN Lai, and EW Sun. Jump detection and noise separation by a singular wavelet method for predictive analytics of high-frequency data. *Computational Economics*, 54:809–844, 2019.

[15] Cleveland Clinic. Luteal phase. https://my.clevelandclinic.org/health/articles/24417-luteal-phase, 2022.

[16] DR Cox. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, 34(2):187–220, 1972.

[17] DR Cox and D Oakes. *Analysis of Survival Data*. Chapman Hall/CRC: Monographs on Statistics and Applied Probability 21, 1984.

[18] C Dai, J Heng, PE Jacob, and N Whiteley. An invitation to sequential Monte Carlo samplers. *Journal of the American Statistical Association*, 117(539):1587–1600, 2022.

[19] B Efron. The two sample problem with censored data. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume IV, pages 831–853. University of California Press, 1967.

[20] PHC Eilers, BD Marx, and M Durbán. Twenty years of P-splines. *Statistics and Operations Research Transactions*, 39(2):149–186, 2015.

[21] DA Epstein, NB Lee, JH Kang, E Agapie, J Schroeder, LR Pina, J Fogarty, JA Kientz, and SA Munson. Examining menstrual tracking to inform the design of personal informatics tools. In *Conference on Human Factors in Computing Systems (CHI)*, Denver, CO, May 2017.

[22] JR Faulkner and VN Minin. Locally adaptive smoothing with Markov random fields and shrinkage priors. *Bayesian Analysis*, 13(1):225–252, 2018.

[23] RJ Fehring, M Schneider, and K Raviele. Variability in the phases of the menstrual cycle. *Journal of Obstetric, Gynecologic, and Neonatal Nursing*, 35(3):376–384, 2006.

[24] JP Fine and RJ Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509, 1999.

[25] G Fitzmaurice, N Laird, and J Ware. *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics, 2004.

[26] Kaiser Family Foundation. Health apps and information survey. https://files.kff.org/attachment/Topline-Health-Apps-and-Information-Survey-September-2019, September 2019.

[27] J Gabry and R Cesnovar. CmdStanR: the R interface to CmdStan. https://mc-stan.org/cmdstanr/reference/cmdstanr-package.html, 2021.

[28] W Gao. A brief introduction to dynamical statistical comparisons. https://stephenslab.github.io/dsc-wiki/first_course/Intro_DSC.html, 2021.

[29] JJ Gaynor, EJ Feuer, CC Tan, DH Wu, CR Little, DJ Straus, BD Clarkson, and MF Brennan. On the use of cause-specific failure and conditional failure probabilities: examples from clinical oncology data. *Journal of the American Statistical Association*, 88(422):400–409, 1993.

[30] A Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–534, 2006.

[31] A Gelman, JB Carlin, HS Stern, DB Dunson, A Vehtari, and DB Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science, 3rd edition edition, 2013.

[32] A Gelman, A Jakulin, M Grazia Pittau, and YS Su. A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2(4):1360–1383, 2008.

[33] TA Gooley, W Leisenring, J Crowley, and BE Storer. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in Medicine*, 18:695–706, 1999.

[34] P Händel and J Wahlström. Digital contraceptives based on basal body temperature measurements. *Biomedical Signal Processing and Control*, 52:141–151, 2019.

[35] EJ Harris, IH Khoo, and E Demircan. A survey of human gait-based artificial intelligence applications. *Frontiers in Robotics and AI*, 8:749274, 2021.

[36] T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2017.

[37] AL Hirschberg. Challenging aspects of research on the influence of the menstrual cycle and oral contraceptives on physical performance. *Sports Medicine*, 52:1453–1456, 2022.

[38] MD Hoffman and A Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014.

[39] CH Hsu and JMG Taylor. Nonparametric comparison of two survival functions with dependent censoring via nonparametric multiple imputation. *Statistics in Medicine*, 28:462–475, 2009.

[40] CH Hsu, JMG Taylor, and C Hu. Analysis of accelerated failure time data with dependent censoring using auxiliary variables via nonparametric multiple imputation. *Statistics in Medicine*, 34(19):2768–2780, 2015.

[41] CH Hsu, JMG Taylor, S Murray, and D Commenges. Survival analysis using auxiliary variables via non-parametric multiple imputation. *Statistics in Medicine*, 25:3503–3517, 2006.

[42] SM Hughes, CN Levy, R Katz, EM Lokken, MN Anahtar, M Barousse Hall, F Bradley, PE Castle, V Cortez, GF Doncel, R Fichorova, PL Fidel, KR Fowke, SC Francis, M Ghosh, LY Hwang, M Jais, V Jespers, V Joag, R Kaul, J Kyongo, T Lahey, H Li, and J Makinde. Changes in concentrations of cervicovaginal immune mediators across the menstrual cycle: a systematic review and meta-analysis of individual patient data. *BMC Medicine*, 20(353), 2022.

[43] EL Kaplan and P Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[44] A Kawamori, K Fukaya, M Kitazawa, and M Ishiguro. A self-excited threshold autoregressive state-space model for menstrual cycles: forecasting menstruation and identifying within-cycle stages based on basal body temperature. *Statistics in Medicine*, 38:2157–2170, 2019.

[45] JS Kim, JG Ryu, JW Kim, EC Hwang, SI Jung, TW Kang, D Kwon, and K Park. Prostate-specific antigen fluctuation: what does it mean in diagnosis of prostate cancer? *International Brazilian Journal of Urology*, 41(2):258–264, 2015.

[46] JP Klein and ML Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 2nd edition, 2003.

[47] M Kostrzewski. The Bayesian methods of jump detection: The example of gas and EUA contract prices. *Polska Akademia Nauk*, 11:107–131, 2019.

[48] DR Kowal, DS Matteson, and D Ruppert. Dynamic shrinkage processes. *Journal of the Royal Statistical Society, Series B*, 81(4):781–804, 2019.

[49] MA Little and NS Jones. Generalized methods and solvers for noise removal from piecewise constant signals. I. Background theory. *The Royal Society Proceedings: Mathematical, Physical and Engineering Sciences*, 467(2135):3088–3114, 2011.

[50] RJA Little and DB Rubin. *Statistical Analysis with Missing Data*. Wiley, 3rd edition, 2019.

[51] JJ Lok, S Yang, B Sharkey, and MD Hughes. Estimation of the cumulative incidence function under multiple dependent and independent censoring mechanisms. *Lifetime Data Analysis*, 24:201–223, 2018.

[52] A Lott and JP Reiter. Wilson confidence intervals for binomial proportions with multiple imputation for missing data. *The American Statistician*, 74(2):109–115, 2020.

[53] L Luo, X She, J Cao, Y Zhang, Y Li, and PXK Song. Detection and prediction of ovulation from body temperature mreasured by an in-ear wearable thermometer. *IEEE Transactions on Biomedical Engineering*, 67(2):512–522, 2020.

[54] E Mazzola and P Muliere. Reviewing alternative characterizations of Meixner processes. *Probability Surveys*, 8:127–154, 2011.

[55] L Miolo, B Colombo, and J Marshall. A data base for biometrics research on changes in basal body temperature in the menstrual cycle. *Statistica*, 53(4):563–572, 1993.

[56] TJ Mitchell and JT Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988.

[57] SE Neville, JT Ormerod, and MP Wand. Mean field variational Bayes for continuous sparse signal shrinkage: pitfalls and remedies. *Electronic Journal of Statistics*, 8:1113–1151, 2014.

[58] The Practice Committee of the American Society for Reproductive Medicine. The clinical relevance of luteal phase deficiency: a committee opinion. *Fertility and Sterility*, 98(5):1112–1117, 2012.

[59] O Papaspiliopoulos, GO Roberts, and M Sköld. Non-centered parameterisations for hierarchical models and data augmentation. *Bayesian Statistics*, 7:307–326, 2003.

[60] E Pierson, T Althoff, D Thomas, P Hillard, and J Leskovec. Daily, weekly, seasonal and menstrual cycles in women's mood, behaviour and vital signs. *Nature Human Behaviour*, 5:716–725, 2021.

[61] J Piironen and A Vehtari. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, Fort Lauterdale, Florida, 2017.

[62] J Piironen and A Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051, 2017.

[63] NG Polson and JG Scott. Shrink globally, act locally: sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9, 2010.

[64] R Core Team. R: A Language and Environment for Statistical Computing. https://www.R-project.org/, 2021.

[65] CE Rasmussen and CKI Williams. *Gaussian Processes for Machine Learning*. Massachusetts Institute of Technology Press, 2006.

[66] JP Royston and RM Abrams. An objective method for detecting the shift in basal body temperature. *Biometrics*, 36(2):217–224, 1980.

[67] PK Ruan and RJ Gray. Analyses of cumulative incidence functions via non-parametric multiple imputation. *Statistics in Medicine*, 27:5709–5724, 2008.

[68] Y Saatci, R Turner, and CE Rasmussen. Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

[69] B Scarpa and DB Dunson. Bayesian hierarchical functional data analysis via contaminated informative priors. *Biometrics*, 65(3):772–780, 2009.

[70] SJ Skates, DK Pauler, and IJ Jacobs. Screening based on the risk of cancer calculation from Bayesian hierarchical changepoint and mixture models of longitudinal markers. *Journal of the American Statistical Association*, 96(454):429–439, 2001.

[71] SJ Solomon, MS Kurzer, and DH Calloway. Menstrual cycle and basal metabolic rate in women. *American Journal of Clinical Nutrition*, 36(4):611–616, 1982.

[72] Stan Development Team. Stan reference manual, version 2.28. mc-stan.org, 2021.

[73] T Tatsumi, M Sampei, K Saito, Y Honda, Y Okazaki, N Arata, K Narumi, N Morisaki, T Ishikawa, and S Narumi. Age-dependent and seasonal changes in menstrual cycle length and body temperature based on big data. *Obstetrics & Gynecology*, 136(4):666–674, 2020.

[74] JMG Taylor, S Murray, and CH Hsu. Survival estimation and testing via multiple imputation. *Statistics & Probability Letters*, 58:221–232, 2002.

[75] R Tibshirani and P Wang. Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics*, 9(1):18–29, 2008.

[76] Dipartimento di Scienze Statistiche Universitá degli Studi di Padova. Data repository. https://doi.org/10.25430/DATAREPOSITORY-STATISTICALSCIENCES, January 2023.

[77] A Vehtari, A Gelman, D Simpson, B Carpenter, and PC Bürkner. Rank-normalization, folding, and localization: an improved R-hat for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, 16(2):667–718, 2021.

[78] I Visser and M Speekenbrink. depmixS4: an R package for hidden Markov models. *Journal of Statistical Software*, 36(7):1–21, 2010.

[79] MP Wand. Fast approximate inference for arbitrarily large semiparametric regression models via message passing. *Journal of the American Statistical Association*, 112(517):137–168, 2017.

[80] T Weschler. *Taking Charge of Your Fertility, 20th Anniversary Edition: The Definitive Guide to Natural Birth Control, Pregnancy Achievement, and Reproductive Health*. William Morrow Paperbacks, 2015.

[81] EB Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[82] SN Wood, N Pya, and B Safken. Smoothing parameter and model selection for general smooth models (with discussion). *Journal of the American Statistical Association*, 111:1548–1575, 2016.

[83] Y Yao, A Vehtari, D Simpson, and A Gelman. Yes, but did it work? Evaluating variational inference. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80. Stockholm, Sweden, 2018.

[84] X Zhou and JP Reiter. A note on Bayesian inference after multiple imputation. *The American Statistician*, 64(2):159–163, 2010.

[85] B Zhu, PXK Song, and JMG Taylor. Stochastic functional data analysis: a diffusion model-based approach. *Biometrics*, 67(4):1295–1304, 2011.