Prediction-powered machine learning for model selection and uncertainty

Vik Shirvaikar

St. Peter's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2025

In the midst of this perplexity, I received from Oxford the manuscript you have just examined. I lingered, naturally, on the sentence: "I leave to various future times, but not to all, my garden of forking paths."

— *Jorge Luis Borges, 1941*

Acknowledgements

My research has greatly benefited from the creative supervision of Chris Holmes. I'm grateful to Choudur Lakshminarayan for inspiring me to be a statistician, and Stephen Walker for his guidance and insight.

I've had the honor of working with several excellent collaborators who have challenged and motivated me. In particular, thank you to Andrea Storås, Xi Lin, and Nic Steyn for helping me grow as a researcher.

My doctorate was supported by the Engineering and Physical Sciences Research Council, through the StatML CDT, and Novo Nordisk. A special thanks to the team at the Department of Statistics, especially Joanna, Emma, Beverley, and Frédérique, for making it a warm and welcoming place to work.

I'm grateful to the many friends who have supported me along the way, and to my partner Julia for being a constant source of joy and strength.

Finally, thank you to my brother and parents, Vinny, Anjali, and Mukul Shirvaikar, for always believing in and encouraging me. I could never have reached this point without you.

Abstract

This thesis explores model selection and uncertainty through the lens of prediction. Building on recent developments in Bayesian predictive inference, we approach uncertainty as a missing data problem, extending the logic of the bootstrap by treating future observations as the basis for inference. This framework offers a complementary perspective to conventional frequentist and Bayesian methods, as it supports probabilistic uncertainty quantification without requiring the subjective specification of a prior distribution. We first apply this lens to model uncertainty and hypothesis testing, proposing a novel procedure where uncertainty is propagated via the recursive imputation of new data, using a one-step-ahead model selection criterion. We then broaden this view, arguing that a model's sequential predictive behavior — specifically, its production of conditionally identically distributed updates — can be used to characterize its coherence, allowing for Bayesian-style uncertainty even in plug-in or frequentist settings. Finally, we apply predictive model ensembling to causal treatment effect estimation, introducing a random forest method that targets relative risk heterogeneity, and demonstrating its application to data from a major cardiovascular clinical trial. Taken together, these results argue that predictive resampling methods, grounded in bootstrap principles, can provide flexible and principled tools for model evaluation and validation.

Contents

1 Introduction **1**

1.1 The eternal debate: inference or prediction? 1

1.2 The puzzle of inverse probability 2

1.2.1 Bayes, Laplace, and uniform priors 2

1.2.2 Fisher and fiducial inference 4

1.2.3 Modern attempts at unification 5

1.3 Uncertainty through a predictive lens 6

1.3.1 Bayesian inference as a missing data problem 6

1.3.2 Martingale posterior distributions 9

1.3.3 Conditionally identically distributed sequences 12

1.3.4 Related work 14

1.4 From parameters to models 15

1.4.1 Why model uncertainty matters 15

1.4.2 Bayesian model averaging and exploration 16

1.4.3 Bagging, ensembling, and mixtures 17

1.5 Thesis outline 18

2 A general framework for probabilistic model uncertainty **21**

2.1 Introduction 21

2.2 Model uncertainty via predictive resampling 26

2.2.1 The standard Bayesian approach 26

2.2.2 Recursive model updating 27

2.2.3 Point hypothesis example 29

2.2.4 Consistent model selection . . . 33
2.3 Convergence of one-step updates . . . 35
2.4 Illustrations . . . 38
2.4.1 Density estimation . . . 38
2.4.2 Variable selection . . . 45
2.5 Conclusion . . . 51

3 Hypothesis testing via predictive resampling 53
3.1 Introduction . . . 53
3.2 Review of existing approaches . . . 53
3.2.1 Frequentist testing . . . 53
3.2.2 Bayesian testing . . . 55
3.2.3 Recent developments in e-values . . . 56
3.3 Illustrations . . . 57
3.3.1 Two-sided testing . . . 58
3.3.2 Multi-level testing . . . 65

4 Bayesian prediction without parameter uncertainty 68
4.1 Introduction . . . 68
4.2 Related work . . . 70
4.2.1 Generalized Bayesian inference . . . 70
4.2.2 Bayesian neural networks . . . 71
4.2.3 Model uncertainty in deep learning . . . 72
4.3 Rethinking Bayesian prediction . . . 73
4.4 Evaluating predictive coherence . . . 76
4.4.1 Consistency under sequential updating . . . 77
4.4.2 Exchangeability via log-joint variance . . . 78
4.5 Illustrations . . . 79
4.5.1 Consistency under sequential updating results . . . 82
4.5.2 Exchangeability via log-joint variance results . . . 82

4.6 Conclusion 83

5 Predictive resampling with conditionally identically distributed parametric models 84

5.1 Introduction 84

5.2 Review of existing approaches 84

5.3 Basic constructions for conditionally identically distributed sequences . . 86

5.3.1 Univariate location parameter 86

5.3.2 Simple linear model 88

5.4 Illustrations 90

5.4.1 Nonparametric curve fitting 90

5.4.2 Time-to-event survival analysis 91

6 Targeting relative risk heterogeneity with causal forests 93

6.1 Introduction 93

6.2 Problem setting 95

6.3 The importance of relative risk 96

6.4 Review of existing approaches 98

6.4.1 Causal forests 99

6.4.2 Model-based forests 101

6.5 Relative risk causal forests 102

6.5.1 Forest construction 103

6.5.2 Treatment effect estimation 106

6.5.3 Omnibus testing 107

6.6 Simulation study 108

6.6.1 Data generation 108

6.6.2 Results 109

6.7 Conclusion 110

7 Exploring relative risk heterogeneity in the LEADER clinical trial 114

7.1 Introduction 114

7.2 Trial details 114

7.3 Results 116

7.3.1 Baseline risk screening 116

7.3.2 Omnibus testing 117

7.3.3 Covariate analysis 117

8 Discussion and future work 121

A Supplementary figures and tables 123

A.1 Density estimation (Chapter 2) 123

A.2 Two-sided hypothesis testing (Chapter 3) 128

A.3 Causal forest simulation (Chapter 6) 133

Bibliography 134

Chapter 1

Introduction

1.1 The eternal debate: inference or prediction?

Modern machine learning and artificial intelligence are characterized by large datasets and substantial computational resources, making it possible to develop highly flexible models that can capture complex hidden patterns. However, this progress has sharpened the long-standing tension between two core statistical objectives: *inference*, or understanding the underlying mechanisms which generated the data, and *prediction*, or accurately forecasting future outcomes without requiring an explanation. This distinction has been widely discussed (Breiman, 2001b; Shmueli, 2010), with an emphasis on how the methods suited to each task can differ substantially.

The subject of this thesis is how prediction-centric methodologies, particularly those leveraging bootstrapping and related resampling techniques, can address essential inference problems such as model selection and uncertainty quantification. Related ideas have long appeared under the broad umbrella of "predictive inference" (Geisser, 1993), but recent developments provide a more formal framework for this perspective, in which missing data is treated as the fundamental source of statistical uncertainty, and the prediction of future observations becomes the central mechanism for inference (Fong et al., 2024; Fortini & Petrone, 2025). We aim to demonstrate that these approaches offer practical, flexible, and interpretable tools which can complement more traditional paradigms.

Another long-standing tension in statistics concerns the distinction between frequentist and Bayesian approaches. Classical frequentist confidence intervals are based on the concept of repeated sampling, yielding probability statements that do not directly

apply to an observed dataset and can be difficult to interpret in practice. In contrast, Bayesian credible intervals provide more intuitive probabilistic reasoning but require a prior distribution, which may be arbitrary or hard to elicit. In this thesis, we explore how prediction-based methods offer an alternative perspective, supporting prior-free probability statements directly on quantities of interest, while sidestepping some of the philosophical and practical challenges associated with both conventional frameworks.

As a specific area of focus, this thesis looks at questions related to models, including selection, validation, evaluation, and uncertainty quantification. In applied settings, these tasks often reduce to retrospectively summarizing the observed data given a model — for example, via likelihood-based information criteria, or cross-validation scores on held-out subsets. Through the predictive lens, we aim to show that these questions can instead be approached as forward-looking problems about the future or unseen data. This includes model uncertainty, by resampling different possible imputations of the complete population (Chapters 2 and 3); model characterization, by examining the step-by-step distribution of future predictive updates (Chapters 4 and 5); and model ensembling, by comparing predictive fits to construct and combine weak learners into a mixture (Chapters 6 and 7).

1.2 The puzzle of inverse probability

Before introducing the primary work of this thesis, we provide a brief overview of some historical foundations and core debates that have shaped the theory of statistical inference.

1.2.1 Bayes, Laplace, and uniform priors

The foundations of Bayesian inference trace back to a simple question about uncertainty in a binomial setting. If we observe an event with x successes and y failures (for example, a coin landing on heads or tails), what can we say about the true probability of success p? In a posthumously published letter, Thomas Bayes introduced a method for updating beliefs about this probability using binary outcome data (Bayes, 1764). Bayes illustrated his idea with a thought experiment involving a billiard table: the first ball establishes an unknown threshold, and each subsequent ball is counted as a success or failure depending on whether

it falls to the left or right of that point. This construction assumes a fixed but unknown parameter, but Bayes' key innovation was to treat that parameter probabilistically, placing a distribution over it to express uncertainty.

Pierre-Simon Laplace independently derived similar results and formalized them into a "rule of succession," which we would now recognize as a conjugate beta-binomial update with a Uniform$[0, 1]$ or Beta$(1, 1)$ prior distribution (Laplace, 1774; Stigler, 1986). In terms of the coin toss example, we add one to each count and divide, so the final probability of heads becomes $p = \frac{x+1}{x+1+y+1} = \frac{x+1}{x+y+2}$. At the time, probability was primarily applied to describe possible variation in observable outcomes. Together with Augustus de Morgan, these thinkers extended its use to what became known as *inverse probability*: the idea that one could reason in the other direction, starting with the outcomes and working backwards to learn about the underlying process that generated the data.

Through the 19th century, the Bayesian solution to the so-called "inverse problem" became widely accepted. A central point of contention, however, was the use of the uniform prior — adding one to each count in the coin toss calculation. This returns a posterior mean of $1/2$ before observing any data, which seems like a neutral choice, but clearly introduces external information into the analysis. Philosophers such as George Boole and John Venn raised further objections, pointing out examples in which a uniform prior led to clearly unreasonable conclusions (Zabell, 1989).

A deeper technical concern soon followed: the uniform prior is not invariant under reparameterization, meaning that what appears "uninformative" in one parameterization becomes informative in another. In the binomial example, we are often interested in the log-odds $\theta = \log(\frac{p}{1-p})$, but placing a Uniform$[0, 1]$ prior on the success probability necessarily implies a non-uniform prior on the log-odds, specifically a logistic distribution peaked at $\theta = 0$. As the 20th century began, these concerns sparked renewed efforts to formalize and refine the foundations of Bayesian inference, and opened the door to alternative updating principles.

Stigler (1982) offers an intriguing historical clarification, observing that the common interpretation of Bayes' prior distribution as reflecting "equal ignorance" about a parameter

may misread Bayes' original reasoning. In fact, Bayes justified the uniform prior not by appealing to ignorance about the parameter itself, but by assuming equal ignorance about the outcomes of future observations. For example, if a random binary event occurs ten times, and no information is available, each count from zero to ten successes should be considered equally likely. In the binomial case, this happens to lead to the same result — a uniform prior over the parameter — but the reasoning is predictive rather than parametric. This subtle distinction avoids some of the issues associated with reparameterization and reinforces an interpretation of uncertainty centered on observable quantities.

1.2.2 Fisher and fiducial inference

In the early 20th century, R.A. Fisher, originally trained in the Bayesian tradition, sought a novel alternative to the uniform prior, motivated in part by personal and intellectual disagreements with contemporaries such as Karl Pearson. In a series of papers on inverse probability, Fisher introduced what became known as the *fiducial argument* (Edwards, 1997). Although the formulation evolved over time, the core idea remained the same.

Given a sufficient statistic, we know that no other statistic can provide additional information on a parameter of interest. Fisher proposed inverting the cumulative distribution function of this sufficient statistic to yield a distribution over the parameter itself, thereby assigning probabilistic uncertainty to parameters without relying on prior beliefs (Zabell, 1992). The fiducial argument worked nicely in simple cases, such as the binomial parameter problem that had originally motivated Bayes and Laplace, but Fisher was never able to extend it into a coherent general theory (Savage, 1976). From the outset, critics identified logical inconsistencies and pathological behaviors, especially when applying the method to more complex models. Zabell (2022) summarizes the resulting impasse:

> "...the happy accident that in the case of a single parameter, fiducial percentiles can be spliced together to form a distribution function in the purely mathematical sense (that is, a function increasing from 0 to 1) led [Fisher] to regard this construct as a viable and principled probabilistic replacement for a Bayesian posterior distribution. His inability to extend this construction

to the multi-parameter setting in a way that won general acceptance never caused him to waver from this view."

Even in the cases where the fiducial argument could be applied, the resulting distributions were found to be theoretically or practically equivalent to frequentist confidence intervals (Cox, 1958), Bayesian posteriors (Lindley, 1958), or both. By 1956, Fisher's last-ditch effort was to argue that fiducial methods (unlike Bayesian ones) were fundamentally defined by the absence of prior information, based on a concept of "recognizable subsets", but this was also ultimately shown to be imprecise (Savage, 1976). Despite some later efforts to salvage aspects of the theory (Fraser, 1961), the fiducial approach largely faded from mainstream statistical discourse following Fisher's death in 1962.

Instead, contemporary statistics operates largely within the well-known frequentist paradigm, where observed data is viewed as one realization of a data-generating process with a fixed but unknown underlying parameter. This approach was formalized in the early 1900s by figures such as Jerzy Neyman and Egon Pearson, alongside contributions from Fisher himself (Fisher, 1925; Neyman & Pearson, 1933), and gained widespread adoption throughout the 20th century, eventually becoming the dominant framework in both statistical research and education. However, frequentist methods face challenges in interpretation: hypothesis tests and confidence intervals are necessarily defined in terms of hypothetical repeated experiments, not the actual data. As a result, they only permit probability statements about long-run sampling frequencies, not the parameters themselves, meaning their conclusions can be non-intuitive in applied settings.

1.2.3 Modern attempts at unification

Statisticians in the late 20th and early 21st centuries have explored a number of strategies to reconcile the Bayesian, fiducial, and frequentist perspectives.

Under the umbrella of objective Bayesian inference (Berger, 2006), several methods have been proposed to specify principled, data-driven prior distributions. An important starting point for this line of work is the Jeffreys prior, which transforms the Fisher information matrix into a density function that is invariant under reparameterization

(Jeffreys, 1961). This resolves one of the classical concerns with uniform priors, but is not a universal solution, leading to improper or atypical results in certain settings. Objective Bayesian methods include a number of alternatives to p-values and Bayes factors that seek to mitigate the influence of the prior distriubtion by aligning it with the observed data, through posterior predictive checks, data-splitting strategies, and so forth (Bayarri & Berger, 2000, 2004; Meng, 1994; Zhang, 2014).

Other approaches adopt a fiducial-style procedure within a frequentist framework, using inversion to target a distribution over the parameter space while avoiding some of the interpretive challenges that troubled Fisher's original formulation. Key contributions in this area include confidence distributions (Xie & Singh, 2013) and generalized fiducial inference (Hannig et al., 2016). These methods typically define a "distribution estimator", rather than a point or interval estimator, enabling uncertainty to be represented probabilistically while emphasizing strong empirical coverage properties.

1.3 Uncertainty through a predictive lens

While the approaches discussed above provide valuable insights, we argue that they may overlook a more fundamental point: the source of statistical uncertainty is missing data. If we directly target this source, and focus on modeling the distribution of unobserved outcomes given the observed data, it becomes possible to make meaningful probabilistic statements about parameters without requiring a prior distribution. In this section, we formalize this idea through a predictive representation of Bayesian inference, and describe the conditions under which such an approach yields valid and coherent results.

1.3.1 Bayesian inference as a missing data problem

Let $Y_1, Y_2, \ldots$ denote a sequence of random variables, which we assume are drawn from an unknown sampling distribution $P^\star$. Upon observing $y_{1:n}$, suppose we wish to conduct inference on an associated parameter θ. The typical Bayesian approach is to specify a likelihood or sampling model $\{P_\theta : \theta \in \Theta\}$, which is a family of probability distributions indexed by a parameter θ from a parameter space Θ, with associated densities $f(y \mid \theta)$.

We elicit a prior density $\pi(\theta)$ over the parameter space, then apply Bayes' theorem to yield the posterior density

$$\pi(\theta \mid y_{1:n}) \;\propto\; \pi(\theta) \prod_{i=1}^{n} f(y_i \mid \theta). \tag{1.1}$$

If, however, the entire infinite sequence of random variables $Y_{n+1:\infty}$ was observed, any identifiable quantity would be completely determined. This quantity could be a parameter θ (e.g., the mean or variance) or a functional of $P^\star$ such as its density. The statistical uncertainty in the posterior distribution therefore arises entirely from the fact that we only observe a sample of size n and are missing the remaining observations.

Motivated by this insight, we adopt a predictive reformulation of statistical learning, where the goal is to directly model the conditional distribution of future observations given the past. In other words, our target is $p(y_{n+1:\infty} \mid y_{1:n})$, the true conditional density of $Y_{n+1:\infty}$ given $Y_{1:n}$, as determined by the unknown data-generating process $P^\star$.

Rather than modeling the data we already have, we focus our modeling effort on what is required to resolve our uncertainty — namely, the distribution of the data we do not observe. Our target quantity is thus defined as a functional of the complete random sequence $Y_{1:\infty}$, with uncertainty indexed over different possible realizations of its unobserved tail $Y_{n+1:\infty}$. This frames inference entirely in terms of observables, avoiding the need for subjective prior specification on a latent parameter.

To put this into practice, several recent papers (Fong et al., 2024; Holmes & Walker, 2023) have explored a direct specification of uncertainty over the conditional density $p(y_{n+1:N} \mid y_{1:n})$ via the sequential factorization

$$p(y_{n+1:N} \mid y_{1:n}) = \prod_{i=n+1}^{N} p(y_i \mid y_{1:i-1}), \tag{1.2}$$

which follows directly from the chain rule for conditional densities. In the limit $N \to \infty$, this formulation defines the full joint distribution $p(y_{1:\infty})$, allowing inference on any identifiable quantity of interest. In many real-world applications, we can also view this as a finite imputation problem, where N represents the known total population size.

Echoing the "prequential" (predictive sequential) perspective of Dawid (1982, 1984), this approach views Bayesian learning entirely through the sequential updating of predictive

density functions. The core inferential task reduces to two alternating recursive steps, a process which Fong et al. (2024) refer to as *predictive resampling*, to simulate a complete future data sequence one observation at a time. First, given the current data $y_{1:i-1}$ at each stage, we require a mechanism to sample the next observation y_i. Second, upon imputing y_i, we must update the predictive density to reflect the new information. This update can be represented as a general one-step mapping,

$$\{p(\cdot \mid y_{1:i-1}), y_i\} \to p(\cdot \mid y_{1:i}), \tag{1.3}$$

which denotes a function that transforms the current predictive density, together with the newly sampled data point, into the predictive density for the next observation. The process then progressively evolves as new observations become available.

As a concrete illustration of the update rule in Equation 1.3, we note that its simplest form will be familiar to readers as the Bayesian bootstrap of Rubin (1981). The Bayesian bootstrap arises within the framework described above when using the simple non-parametric predictive rule where we duplicate one of the existing data points at random, also known as a Pólya urn scheme with replacement. The one-step predictive density is defined as a draw from the empirical distribution,

$$Y_{n+1} \mid y_{1:n} \;\sim\; \text{Uniform}\{y_1, \ldots, y_n\}. \tag{1.4}$$

This is equivalent to sampling from the discrete measure $\frac{1}{n}\sum_{i=1}^{n} \delta_{y_i}$, with mass $1/n$ at each observed y_i. In the limit, this process defines a randomized empirical distribution

$$F_\infty = \sum_{i=1}^{n} w_i \, \delta_{y_i}, \;\; w_{1:n} \sim \text{Dirichlet}(1, \ldots, 1)$$

with random weights $w_{1:n}$ that are positive and sum to one. This is precisely the Bayesian bootstrap, in contrast to the frequentist bootstrap (Efron, 1979), which approximates a sampling distribution by repeatedly drawing resamples of size n with replacement from the observed data. The following discussion can therefore be viewed as a generalization of the Bayesian bootstrap, in which predictive updates are not limited to copies of existing data drawn from the empirical distribution.

1.3.2 Martingale posterior distributions

The predictive formulation described in the previous section is intentionally general: it specifies a framework in which uncertainty is represented through the sequence of predictive distributions $p(y_i \mid y_{1:i-1})$, but without imposing any particular form on the predictive rule. In principle, many possible rules could be constructed, including trivial or degenerate cases that would not yield meaningful uncertainty. A key research question is therefore what conditions the predictive specification must satisfy in order to return a valid notion of uncertainty that is coherent with standard Bayesian inference (Battiston & Cappello, 2025).

To understand these conditions, we first examine the learning properties of parametric Bayesian inference, with a specific focus on consistency — the property that the posterior distribution concentrates on the true parameter value as data accumulate. The key result in this setting comes from Doob (1949), and provides a benchmark for coherent learning in the parametric case. We then extend these ideas to suggest the conditions under which a predictive representation can reproduce the same learning behavior and yield meaningful uncertainty quantification.

Specifically, we return to the Bayesian updating framework of Equation 1.1, with data generated conditionally on a latent parameter $\Theta \sim \Pi$ according to

$$Y_i \mid \Theta = \theta \ \sim \ P_\theta, \ i = 1, 2, \ldots$$

independently and identically. Here, P_θ is the sampling model distribution, with associated density $f(y \mid \theta)$. This induces the joint density

$$p(\theta, y_{1:n}) = \pi(\theta) \prod_{i=1}^{n} f(y_i \mid \theta), \tag{1.5}$$

for the parameter θ and observations $y_{1:n}$ at any given sample size n. Let $\mathcal{F}_n = \sigma(Y_1, \ldots, Y_n)$ denote the filtration generated by the first n observations, representing the information available after observing the data. To estimate the parameter with this information, an appropriate starting point is the posterior mean, given by $\bar{\theta}_n = E[\Theta \mid \mathcal{F}_n]$.

The following result then establishes that Bayesian updating is consistent: as more data are observed, the posterior distribution concentrates on the true parameter value.

Theorem 1 (Doob, 1949)**.** *Assume that Θ takes values in a linear space with $E[|\Theta|] < \infty$, and that $(\Theta, Y_1, Y_2, \ldots)$ is distributed according to Equation 1.5, so that $\Theta \sim \Pi$. Additionally, assume that the model $\{P_\theta : \theta \in \Theta\}$ is identifiable, meaning that $P_\theta = P_{\theta'}$ implies $\theta = \theta'$, and that Θ is measurable with respect to the σ-algebra $\mathcal{F}_\infty$ generated by the infinite data sequence. Then the sequence of posterior means satisfies*

$$\bar{\theta}_n = E[\Theta \mid \mathcal{F}_n] \to \Theta$$

as $n \to \infty$ almost surely.

In other words, although Bayesian updating is typically interpreted as refining belief about the latent parameter Θ, it can equivalently be viewed as sequentially resolving uncertainty about the unobserved data. Each new observation incrementally reduces uncertainty, and in the limit, the a priori variability in Θ is explained entirely by the missing information in $Y_{1:\infty}$.

The key criterion underlying this result is that the sequence of posterior means forms a *martingale*. A martingale is a stochastic process whose expected next value, given the current information, equals its present value; intuitively, it represents a learning process that evolves without systematic drift, where the best forecast for the next step is simply the current estimate. With respect to the filtration $\{\mathcal{F}_n\}$ generated by the observations, this property is written as

$$E[\bar{\theta}_n \mid \mathcal{F}_{n-1}] = \bar{\theta}_{n-1}. \tag{1.6}$$

Doob's martingale convergence theorem states that any integrable martingale $\{\xi_n\}$ converges almost surely to a limiting random variable ξ_∞. Applying this result, the sequence $\{\bar{\theta}_n\}$ therefore converges almost surely to some limit $\bar{\theta}_\infty$. Under the identifiability condition, the limit is uniquely determined by the infinite sequence of data, ensuring that $\bar{\theta}_\infty = \Theta$. In this sense, Doob's theorem ensures that Bayesian learning proceeds as a drift-free updating process that eventually resolves all parametric uncertainty through the step-by-step incorporation of information from observed data.

How does this extend to the predictive framework? Recall that the predictive resampling pipeline aims to directly specify uncertainty on the missing data, applying the

sequential factorization of Equation 1.2. The algorithm alternates between imputing the next data point and updating the predictive density via the one-step mapping in Equation 1.3. As described previously, this amounts to a generalized Bayesian bootstrap where predictive updates are not required to be copies of existing data.

Under the result from Doob, we know that coherent Bayesian learning rests on the construction of a martingale for the sequence of posterior means of the quantity of interest. For the bootstrap predictive update, the data-generating procedure can therefore also be constrained by this martingale requirement. Specifically, any predictive rule we apply should ensure that the sequence of posterior means forms a martingale, meaning that its conditional expectation before incorporating each new imputed observation equals its previous value. This condition enables a predictive update scheme to propagate uncertainty without introducing systematic bias.

Fong et al. (2024) bring these insights together to define the *martingale posterior distribution* as the posterior uncertainty resulting from a valid predictive resampling process. After resampling the unobserved sequence $y_{n+1:N}$, the quantity of interest is evaluated using the completed sample $y_{1:N}$. For simple functionals such as the mean or variance, this evaluation is straightforward. More generally, Fong et al. (2024) construct a random limiting empirical distribution, denoted F_N, from which any parameter or functional $\theta(F_N)$ can be derived.

Across repeated predictive resampling trials, the resulting distribution of $\theta(F_N)$ then serves as the martingale posterior, representing the uncertainty induced purely by the predictive specification. Crucially, when the predictive rule satisfies the martingale condition discussed above, the limiting distribution F_N converges to a random measure F_∞ that coincides almost surely with the posterior distribution obtained under the standard prior–likelihood formulation. In this sense, the martingale posterior provides a predictive characterization of Bayesian learning that reproduces the same asymptotic uncertainty while avoiding explicit reference to a prior distribution on parameters.

1.3.3 Conditionally identically distributed sequences

The previous section established a benchmark for coherent Bayesian learning, rooted in results from Doob: the sequence of posterior means for a parameter of interest must form a martingale, satisfying Equation 1.6. A core goal of our framework, however, is to ground Bayesian inference entirely in terms of observable data. This motivates the search for a parallel coherence condition which is defined not on a latent parameter, but rather on the sequence of future observables directly.

The foundation for this condition comes from the concept of *conditionally identically distributed (c.i.d.)* sequences (Berti et al., 2004). To develop this argument, we first revisit the role of exchangeability in standard Bayesian inference, then introduce the c.i.d. condition and show that it can be understood as a weakening of exchangeability. Ultimately, we define "predictive coherence" as a martingale property that applies directly to the sequence of predictive distribution functions in a resampling scheme.

In the Bayesian literature, a typical starting point to justify the adoption of a prior-likelihood model is to assume *exchangeability*. An infinite sequence $y_1, y_2, \ldots$ is defined as exchangeable if, for any finite n, its joint distribution is invariant to permutation, i.e., $p(y_1, \ldots, y_n) = p(y_{\sigma(1)}, \ldots, y_{\sigma(n)})$ for any permutation σ. The representation theorem of de Finetti (1937) demonstrates the consequences of this assumption for the foundational case of a binary sequence.

Theorem 2 (de Finetti, 1937)**.** *Let* $(Y_1, Y_2, \ldots)$*, with* $Y_i \in \{0, 1\}$*, be an infinite sequence of binary random variables. If this sequence is exchangeable, then there exists some cumulative distribution function* Π *such that the joint probability of any finite subsequence* $Y_{1:n}$ *has the form*

$$p(y_{1:n}) = \int \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}\, d\Pi(\theta).$$

Additionally, there exists a random variable Θ *distributed according to* Π *which is the limit of the empirical mean, with*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Y_i = \Theta$$

almost surely.

This theorem provides a powerful justification for the traditional prior-likelihood specification. It shows that for a binary sequence, the subjective assumption of exchangeability — an assumption purely about the structure of the observable data — is mathematically equivalent to postulating the standard Bayesian generative model. Specifically, exchangeability implies a model where the observations y_i are independent Bernoulli trials conditional on a parameter θ, and this unobservable parameter θ is itself treated as a random variable drawn from a prior distribution Π. This has been described as a "shortcut to infinity", as it allows us to induce the familiar generative structure of Equation 1.5 using nothing more than an intuitive assumption on the symmetry of the data.

When applying the predictive factorization of Equation 1.2, we directly specify the sequence of one-step-ahead conditional distributions. Critically, this formulation does not require the infinite sequence $(Y_1, Y_2, \ldots)$ to be fully exchangeable. This departure means Theorem 2 no longer applies, so we cannot invoke the standard likelihood-prior construction of Equation 1.5 or the corresponding coherence of parametric belief updates. However, the key insight is that we can still conduct Bayesian inference under the c.i.d. assumption, as it represents a weaker but sufficient form of exchangeability.

To formalize this, let $P_i(y) = P(Y_{i+1} \leq y \mid y_{1:i})$ denote the one-step-ahead predictive cumulative distribution function (CDF) given the history $y_{1:i}$. As described by Berti et al. (2004), our sequence of future observations $Y_{n+1}, Y_{n+2}, \ldots$ is c.i.d. if

$$P(Y_{i+k} \leq y \mid y_{1:i}) = P_i(y) \;\; \forall k \geq 1. \tag{1.7}$$

In other words, given any history $y_{1:i}$, all future observations $Y_{i+1}, Y_{i+2}, \ldots$ are identically distributed according to the current one-step-ahead predictive distribution $P_i(y)$. This c.i.d. property is a weakening of full exchangeability (Mlodozeniec et al., 2024), as it does not assume the entire sequence is exchangeable, but rather that only the future is exchangeable, conditional on the past.

Fong et al. (2024) show that if the c.i.d. condition in Equation 1.7 is satisfied, then the predictive updating scheme satisfies two key properties:

1. Existence: The sequence of predictive CDFs $P_{n+1}(y), P_{n+2}(y), \ldots$ converges almost surely to a random probability distribution function $P_\infty(y)$ for all y.

2. Unbiasedness: The posterior expectation of this limiting distribution, conditional on the information at step n, is the current predictive distribution, or $\mathbb{E}[P_\infty(y) \mid \mathcal{F}_n] = P_n(y)$ almost surely where $\mathcal{F}_n = \sigma(y_{1:n})$.

These properties are collectively termed *predictive coherence*, and provide a sufficient basis for a valid martingale posterior distribution. The existence property ensures that the predictive resampling process converges to a well-defined random measure, while the unbiasedness property ensures that no new information or bias is introduced, and the process only serves to propagate uncertainty.

As a further connection with the previous section, Fong et al. (2024) also show that the c.i.d. condition can be equivalently formulated as a martingale condition on the sequence of predictive distributions, with

$$\mathbb{E}[P_i(y) \mid y_{1:i-1}] = \int P_i(y) dP_{i-1}(y_i) = P_{i-1}(y).$$

The construction of a c.i.d. sequence therefore serves as a predictive analogue to the martingale property in the parametric setting, giving us a coherence condition defined purely in predictive terms. The martingale property is not always straightforward to assess for parameters in complex models, but this condition suggests an alternative route for practical model validation based solely on observable predictions.

1.3.4 Related work

Beyond de Finetti, several previous works have identified the value of uncertainty quantification through observables rather than parameters. Roberts (1965) presents an early version of the predictive argument for finite data, noting that any statement about a finite population parameter can be reinterpreted as a predictive statement about the unobserved part of the population. Geisser (1982, 1993) concedes that the "lurking parameter" may be a relevant construct in modeling, but argues that prediction is more relevant than parametric inference since results are expressed in terms of actual observables. Dawid (1982, 1984) goes a step further, arguing at a fundamental level that "the purpose of statistical inference is to make sequential forecasts for future observations rather than to express information about parameters."

In Bayesian nonparametrics, Fortini and Petrone (2012, 2020) analyze the predictive construction of underlying models, where resampling is applied to retrieve the prior law of the mixing distribution. Berti et al. (2021) discuss a general class of models building on c.i.d. sequences, while Battiston and Cappello (2025) introduce asymptotically conditionally identically distributed (a.c.i.d.) sequences as a generalization of the c.i.d. condition. Hahn et al. (2018) provide a fast online approach for sequential updates that makes use of bivariate copulas.

Instructive reviews of the predictive approach to Bayesian inference are given by Berti et al. (2023) and Fortini and Petrone (2025); as above, these rely on the overall view of Bayes as a sequential learning problem, where the goal is to specify one-step marginal predictive updates without having to rely on the usual prior-posterior pipeline.

1.4 From parameters to models

So far, we have focused on quantifying parameter uncertainty by bootstrapping the unobserved data. However, another fundamental challenge in statistical inference is model uncertainty — accounting for uncertainty not just within the parameters of a given model, but also across the broader space of plausible models. Interestingly, many popular techniques in this space also rely on bootstrap-based principles, resampling and ensembling several models to yield uncertainty. This provides an underlying thematic link between predictive inference and modeling practices in modern machine learning.

1.4.1 Why model uncertainty matters

Model uncertainty can be of interest for a variety of reasons: for instance, to assess the sensitivity of results to modeling assumptions, to quantify ambiguity in which model best fits the data, or to combine competing models for improved robustness. This challenge is illustrated by the *garden of forking paths* metaphor from Gelman and Loken (2019), which highlights that the seemingly arbitrary modeling choices made by analysts — such as variable selection, data transformation, or treatment of missing data — can have a greater influence on the results than the data itself.

In the social sciences, this has led to the development of *multiverse analysis*, where researchers systematically conduct the main analysis across a range of plausible data collection and processing regimes, and then examine the variation in outcomes (Harder, 2020; Steegen et al., 2016). This allows for a form of empirical sensitivity analysis, ensuring that reported results are not purely the byproduct of a single selected model structure. Riha et al. (2024) provide a computational iterative filtering workflow to support multiverse analysis in more complex settings, enabling the efficient exploration and assessment of several alternative models.

1.4.2 Bayesian model averaging and exploration

More formally, let $\{\mathcal{M}_k\}_{k=1}^K$ be a set of candidate models for observed data $\mathcal{D}$. Bayesian approaches to model uncertainty typically target the posterior probability of each individual model given the data,

$$\pi(\mathcal{M}_k \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathcal{M}_k)\pi(\mathcal{M}_k)}{P(\mathcal{D})}, \tag{1.8}$$

where $\pi(\mathcal{M}_k)$ is the prior probability of model $\mathcal{M}_k$ and $P(\mathcal{D} \mid \mathcal{M}_k)$ is the marginal likelihood of the data under model $\mathcal{M}_k$. These posterior probabilities may be of interest in their own right, or as an intermediate step for Bayesian model averaging (BMA),

$$\pi(\xi \mid \mathcal{D}) = \sum_{k=1}^{K} P(\xi \mid \mathcal{M}_k)\pi(\mathcal{M}_k \mid \mathcal{D}) \tag{1.9}$$

where ξ is a separate target quantity, such as a prediction or parameter, that could vary considerably across different models (Hoeting et al., 1999; Leamer, 1978). BMA has been found to empirically improve predictive performance across a variety of settings, by protecting against overfitting and providing robustness to model misspecification.

Central to this calculation is the marginal likelihood or evidence, given by

$$P(\mathcal{D} \mid \mathcal{M}_k) = \int P_{\mathcal{M}_k}(\mathcal{D} \mid \theta_k)\pi(\theta_k)d\theta_k. \tag{1.10}$$

This returns the probability of the data under each model by integrating over the model-specific parameters. The marginal likelihood has been described as the Bayesian encoding of "Occam's razor" (MacKay, 1992a), since simpler models place a greater probability

weight on a narrower range of possible datasets, and therefore return a greater likelihood if consistent with the observed data.

In the case of two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$, the standard approach is to examine the ratio of their posterior probabilities, known as the posterior odds. Applying Equation 1.8, this can be decomposed as

$$\underbrace{\frac{\pi(\mathcal{M}_1 \mid \mathcal{D})}{\pi(\mathcal{M}_2 \mid \mathcal{D})}}_{\text{Posterior Odds}} = \underbrace{\frac{P(\mathcal{D} \mid \mathcal{M}_1)}{P(\mathcal{D} \mid \mathcal{M}_2)}}_{\text{Bayes Factor}} \times \underbrace{\frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)}}_{\text{Prior Odds}} \tag{1.11}$$

where the normalizing constant $P(\mathcal{D})$ cancels out. Note that the ratio of the marginal likelihoods is also known as the *Bayes factor*, and quantifies the evidence from the data alone (Kass & Raftery, 1995). As this decomposition shows, the final posterior odds are sensitive to the prior distributions on both the parameters within each $P(\mathcal{D} \mid \mathcal{M}_k)$ and the models themselves, motivating the development of several variations that use data-dependent prior distributions (Berger & Pericchi, 1996; O'Hagan, 1995).

To compute posterior model probabilities directly, a number of alternative approaches have been proposed that rely on sampling. Standard Markov Chain Monte Carlo (MCMC) approaches aim to construct a random walk through parameter space that converges to the desired posterior distribution. Once convergence is reached, the resulting samples can be used to estimate expectations and quantiles, or to visualize the posterior distribution by plotting a histogram. To extend this to model uncertainty, the sampling protocol needs to also allow for jumps between models. For example, Model Composition MCMC (MC^3) achieves this by sampling not only parameters but also model indices (Madigan et al., 1995). Perhaps the most well-known variant is Reversible Jump MCMC (RJ-MCMC), which incorporates an adjustment for "model size" to allow for transitions between models of differing dimensions (Green, 1995). The resulting chain provides uncertainty not only across individual parameter values but also across model structures.

1.4.3 Bagging, ensembling, and mixtures

While methods like RJ-MCMC offer principled ways to explore model space, they can be computationally intensive and difficult to tune. In practice, particularly for supervised

learning tasks, a more flexible and scalable alternative is often to train and combine the results of several models directly.

Breiman (2001b) discusses the *Rashomon effect*: the observation that many problems admit a multiplicity of models with comparable predictive performance. Subsequent work on "Rashomon sets" has formalized this idea, defining a class of near-optimal models — for example, those whose predictive risk lies within ϵ of the best model — and using this set to answer questions such as variable importance, by analyzing all of the models collectively (Fisher et al., 2019).

These insights motivate the idea of *bagging* (bootstrap aggregating), where predictions from several models trained on different bootstrap samples are combined. By averaging across multiple models, this effectively reduces model variance while maintaining low bias. Bagging is exemplified by random forests (Breiman, 2001a), one of the most widely used ensemble methods, in which multiple decision trees are independently trained on subsets of the data, then combined via majority vote or average prediction. Over time, random forests have been extended to handle a variety of statistical tasks, including survival analysis, causal inference, and conditional density estimation.

Beyond decision trees, the general strategy of fitting multiple weak learners and combining their predictions appears across other ensemble frameworks. These include boosting, which gradually improves performance by sequentially fitting models to the residuals or errors of previous models; and stacking, which fits a meta-model to learn how to best combine the predictions of a set of base learners, often using cross-validation to assign optimal weights (Yao et al., 2018). Together, these strategies illustrate an emphasis on prediction as a primary avenue for evaluating and combining models.

1.5 Thesis outline

This thesis investigates novel applications of predictive inference and bootstrapping to questions of model selection and uncertainty quantification. Here, we provide a brief overview of the upcoming chapters.

Chapters 2 and 3 Existing approaches to model uncertainty typically either compare models using a retrospective model selection criterion or evaluate posterior model probabilities having set a prior distribution. In these chapters, we propose an alternative strategy which views missing observations as the source of model uncertainty, where the true model would be identified with the complete data. To quantify model uncertainty, it is then necessary to provide a probability distribution for the missing observations conditional on what has been observed. This can be set sequentially using one-step-ahead predictive densities, which recursively sample from the best model according to some consistent model selection criterion. This approach bypasses the need for subjective prior specification or integration over parameter spaces, addressing issues with standard methods such as the Bayes factor. In Chapter 2, we introduce the framework, with illustrations from density estimation and variable selection. In Chapter 3, we focus on the question of hypothesis testing, contrasting predictive resampling with existing approaches in greater detail.

Chapters 4 and 5 Bayesian methods typically frame uncertainty through prior and posterior distributions on parameters, a framework which becomes increasingly strained as model dimensions grow. On the other hand, uncertainty can also be understood as arising from an incomplete dataset, and quantified through the construction of a predictive model for the missing observations. To ensure convergence of the resulting sequence of one-step-ahead distributions, we restrict the future samples to be c.i.d. given the observed data. In Chapter 4, we argue that if this key condition is satisfied, then a model can still provide coherent Bayesian predictions even without a traditional prior-posterior update. We propose and implement diagnostics for the c.i.d. property, showing that a wider class of machine learning methods, including those based on plug-in or frequentist estimation, can be understood as Bayesian through a predictive lens. In Chapter 5, we focus on a direct approach for predictive resampling with parametric c.i.d. sequences. We provide illustrations from the areas of nonparametric curve fitting with splines and time-to-event survival analysis.

Chapters 6 and 7 The identification of heterogeneous treatment effects across subgroups is of significant interest in clinical trial analysis. Several state-of-the-art heterogeneity estimation methods, including causal random forests, apply recursive partitioning for non-parametric identification of relevant covariates and interactions. However, the partitioning criterion typically targets differences in absolute risk. This can dilute statistical power by masking variation in the relative risk, which is often a more appropriate quantity of clinical interest. In these chapters, we propose and implement a methodology for modifying causal forests to target relative risk, using a novel node-splitting procedure based on exhaustive generalized linear model comparison. In Chapter 6, we present the method, along with results from simulation studies that suggest relative risk causal forests can capture undetected sources of heterogeneity. In Chapter 7, we provide an application case study on real-world data from a major cardiovascular clinical trial.

Each chapter contains further technical background details as needed. Chapters 2 and 3 are currently under review as a journal submission. Chapters 4 and 5 consist of work not yet submitted for publication. Chapters 6 and 7 are currently under revision as a journal submission.

Chapter 2

A general framework for probabilistic model uncertainty

2.1 Introduction

In Section 1.3, we established a predictive framework for inference, viewing missing observations as the fundamental source of statistical uncertainty. This approach centers on specifying a joint distribution for the missing data $p(y_{n+1:N} \mid y_{1:n})$ via the sequential factorization in Equation 1.2, a process termed predictive resampling.

This framework has been applied in the context of parameter estimation and conditional prediction (Fong et al., 2024; Holmes & Walker, 2023), but in this chapter, we extend it to provide a general construction for probabilistic model uncertainty. We propose a novel approach where the one-step-ahead predictive density $p(y_i \mid y_{1:i-1})$ used in the resampling process is itself determined by a model selection step. Specifically, our method recursively alternates between selecting the best model for the current data $y_{1:i-1}$, according to some consistent selection criterion, and sampling the next imputation y_i from that model's predictive distribution.

Ultimately, this allows us to retrieve Monte Carlo uncertainty on the "true" model as identified with the complete missing data using a specified criterion for model selection. This predictive approach is fundamentally different from the usual frequentist view, based on repeated experiments of size n, but also deviates from the usual Bayesian requirement of eliciting a subjective prior distribution. We motivate this approach in detail, and demonstrate its application to several statistical decision problems.

To formalize this, consider a set of candidate models $\{\mathcal{M}_k\}_{k=1}^K$ for the observed data $y_{1:n}$. As established, the uncertainty in which model is optimal arises from the missing $Y_{n+1:\infty}$. We require a generative model $p(y_{n+1:\infty} \mid y_{1:n})$ for the missing data given the observed data, and at any given point in the resampling process, the most natural choice for predicting new information is the best model available given the current information.

To impute the missing data, we therefore adapt the one-step predictive update from Equation 1.3 with respect to the candidate models. In particular, we use a model selection criterion C to determine the best current model for $y_{1:n}$, which we write as $\mathcal{M}_{\widehat{k}(n)}$, along with associated parameter(s) $\widehat{\theta}_{\widehat{k}(n)}$. (We discuss the specification of C below.) We sample a new observation from $p(\cdot \mid \mathcal{M}_{\widehat{k}(n)}, \widehat{\theta}_{\widehat{k}(n)})$, add it to the data, and repeat the process recursively up to some sufficiently large N. The update is defined by

$$\begin{aligned} Y_{n+1} \mid y_{1:n} &\sim p(\cdot \mid \mathcal{M}_{\widehat{k}(n)}, \widehat{\theta}_{\widehat{k}(n)}), \\ \widehat{k}(n+1) &= \operatorname*{argmax}_k C(\mathcal{M}_k, y_{1:n+1}). \end{aligned} \tag{2.1}$$

Rather than evaluating the evidence for models under limited data, our focus shifts to the distribution over the "true" model as determined by different possible completions of the dataset. Under the martingale posterior framework (Fong et al., 2024), the target of our inference for each model $\mathcal{M}_k$ is $P(\mathcal{M}_{\widehat{k}(\infty)} = \mathcal{M}_k \mid y_{1:n})$: the probability that it would be selected given infinite data, a quantity induced by the predictive resampling process. We denote this as $\Pi_{\text{pred}}(\mathcal{M}_k \mid y_{1:n})$, deliberately using the Π_{pred} notation to distinguish this predictive probability from a standard Bayesian posterior probability $\pi(\mathcal{M}_k \mid y_{1:n})$.

In practice, we approximate Π_{pred} using a Monte Carlo estimator. After B independent resampling trials, our estimate $\widehat{\Pi}_{\text{pred}}$ is given by the empirical frequency,

$$\widehat{\Pi}_{\text{pred}}(\mathcal{M}_k \mid y_{1:n}) = B^{-1} \sum_{b=1}^{B} \mathbb{1}\left(\mathcal{M}^{(b)}_{\widehat{k}(\infty)} = \mathcal{M}_k\right).$$

Here, b indexes the repeated replications of $y_{n+1:\infty}$. (As a practical matter, we assess the behavior of C as the sample size grows and stop at some sufficiently large $N < \infty$ when the choice of model has clearly converged, using $\mathcal{M}^{(b)}_{\widehat{k}(N)}$ as our proxy for $\mathcal{M}^{(b)}_{\widehat{k}(\infty)}$.) The resulting estimates express uncertainty over the space of candidate models, without requiring the elicitation of prior distributions on either the models or their parameters.

A key ingredient in this procedure is the one-step model selection criterion C. Since the eventual goal is to make inferences based on the complete data, we require a criterion which is consistent, or guaranteed to select the correct model as $n \to \infty$ (Claeskens & Hjort, 2008). The above pipeline can then be viewed as a way of converting our consistent criterion directly into a statement of posterior uncertainty over the space of models. As we discuss in Section 2.2.4, the requirement of consistency admits the use of the Bayesian information criterion (BIC), but not the Akaike information criterion (AIC) or leave-one-out cross-validation (LOO-CV). The AIC and LOO-CV do not necessarily select the correct model as $n \to \infty$, and so should not be used in a setting where our goal is to understand uncertainty in decisions based on the complete data.

The work of Draper (1995) on model expansion provides a useful context for our approach. Draper denotes a model $\mathcal{M} = (S, \theta)$ as a set of "structural assumptions" S (such as the assumed linear structure, the link function in a GLM, etc.) along with parameter(s) θ. He notes that statistical applications typically assume a best structure S^* and then discuss parametric uncertainty on θ, but that this equates to placing an overly concentrated prior point mass of one on S^*, leading to overconfident conclusions. A preferable approach would be to propagate structural uncertainty by placing a more diffuse prior distribution across a wider space of models. Draper suggests that this wider space could be determined by "starting with a single structural choice S^* and expanding it in directions suggested by context", though the specific prescription for this expansion is determined on a case-by-case basis.

Our approach is also a form of model expansion around the initial $\mathcal{M}_{\widehat{k}(n)}$, where the progressive search for new models is guided by the imputation of unseen data. Figure 2.1 represents this idea with a sampling trajectory diagram based on a simple hypothesis test, discussed further in Section 2.2.3. Each individual path tracks the value of a sufficient summary statistic (in this case, the mean) for one possible realization of the complete data as new samples are imputed. All paths start at a common initial best model — in this case, the alternative hypothesis H_1 — but by recursively sampling new observations and then updating the choice of best model, we introduce uncertainty over the model

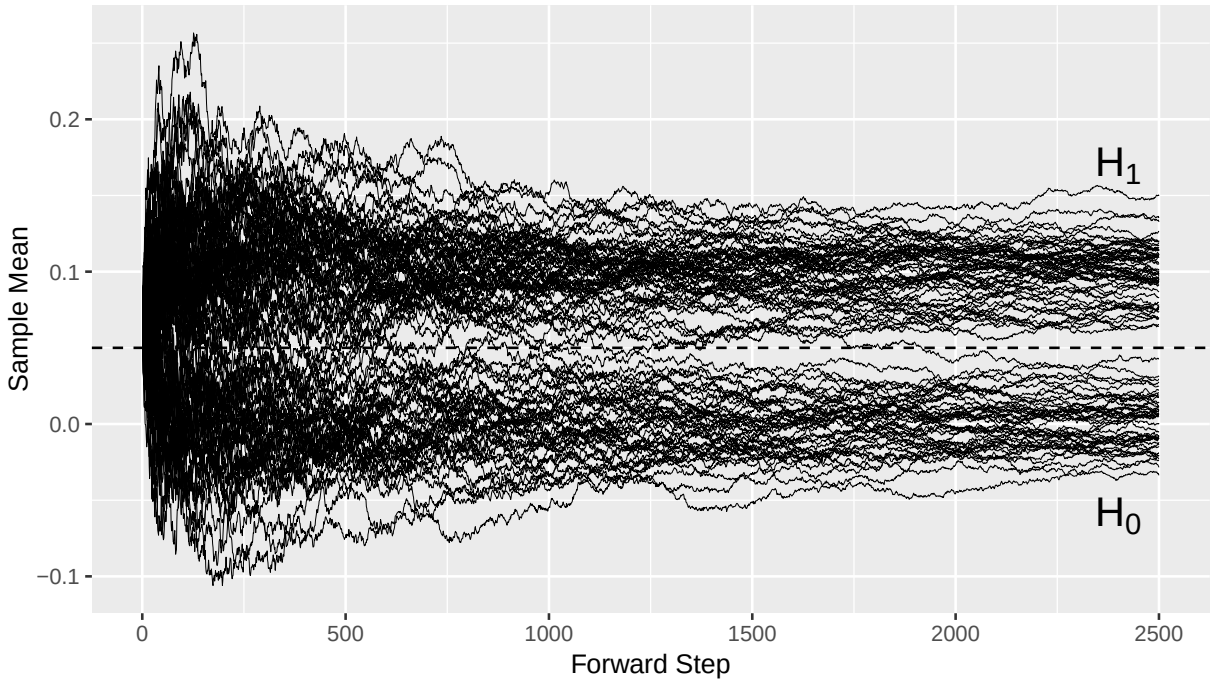


Figure 2.1: Trajectory diagram showing propagation of uncertainty through sampling of missing observations, where different possible realizations of the complete data start at the same $\mathcal{M}_{\widehat{k}(n)} = H_1$, but individually converge to either H_0 or H_1.

space. Several of the sample paths cross back and forth over the dotted line, representing the decision boundary between the models, but each one ultimately converges to a single model choice once the sample size becomes sufficiently large.

As noted previously, the practical implications of our work are similar to those expressed in the "prequential" approach of Dawid (1984), which reframes statistical inference as a sequential forecasting problem. Dawid often employs a plug-in predictive density, such as one based on the maximum likelihood estimate (MLE) $\widehat{\theta}_{i-1}$. This specifies the joint distribution sequentially as

$$\prod_{i=1}^{n} p(y_i \mid y_{i-1}, \widehat{\theta}_{i-1}),$$

which is then used to evaluate the model's likelihood based on observed data. We extend this plug-in strategy, with the one-step-ahead prediction $p(\cdot \mid \mathcal{M}_{\widehat{k}(n)}, \widehat{\theta}_{\widehat{k}(n)})$ determined by the model and parameters that optimize a selection criterion C, typically a dimension-penalized likelihood such as the BIC. (We concede that the penalty term could be viewed as a form of effective prior distribution, a point we discuss further below.) However, our overarching motivation goes one step further: the sequential forecasting of unseen data is

valuable not only in its own right, but also for what it progressively reveals about the corresponding model uncertainty.

What distinguishes this framework from traditional approaches to model uncertainty? Standard calculations using Bayes factors (Kass & Raftery, 1995) exhibit a known over-dependence on subjective model priors, and entail the computational complexity of integrating over the complete parameter space within a model. There have been several "objective" variations on the Bayes factor theme, such as fractional (O'Hagan, 1995) and intrinsic (Berger & Pericchi, 1996) Bayes factors, but these still require the loss of some training data or the selection of an arbitrary calibration weight. Alternative approaches apply sampling strategies to explore the space of models (George & McCulloch, 1993; Madigan et al., 1995), of which RJ-MCMC is perhaps the most well-known (Green, 1995). MCMC-style methods require the difficult construction of an ergodic Markov chain that can eventually visit all parts of the model space.

In contrast, by focusing exclusively on the prediction of observable data points to propagate uncertainty, we bypass the need for any subjective prior distribution over the space of candidate models or their parameters. Predictive resampling makes complete use of the information contained in the training data, with a fully pre-specified protocol to define $\Pi_{\text{pred}}(\mathcal{M}_k \mid y_{1:n})$. This procedure's only fundamental dependence is on the chosen model selection criterion C. The penalty term within that criterion (such as the $\log(n)$ term in BIC) admittedly introduces an external preference for model parsimony. Our approach, therefore, should be understood not as an entirely "objective" process, but as a mechanism that converts a model selection criterion directly into a probabilistic measure of uncertainty.

Additionally, the procedure is pragmatic and straightforward. We reduce uncertainty quantification to two simple steps, model comparison and simulation of a new observation, which can be easily parallelized over the B trials, and applied as a wrapper for any model selection rule. The first step, which requires optimization using a consistent model selection criterion, can leverage modern methods for efficient model search, and is computationally simpler than the transformations required to jump between dimensions

in sampling methods such as RJ-MCMC.

The remainder of this chapter is organized as follows. In Section 2.2, we formally present our framework for model uncertainty in the context of traditional approaches. Section 2.3 contains further theory related to the necessary conditions for model convergence. Section 2.4 contains illustrations, and Section 2.5 provides a discussion and conclusions. Code to replicate all experiments is available at `https://github.com/vshirvaikar/MPModel`.

2.2 Model uncertainty via predictive resampling

Before detailing our approach to quantifying model uncertainty, we outline the usual Bayesian strategies.

2.2.1 The standard Bayesian approach

As introduced in Section 1.4, Bayesian model uncertainty typically targets posterior model probabilities $\pi(\mathcal{M}_k \mid \mathcal{D})$ according to Equation 1.8, which can then be used for BMA (Equation 1.9). The comparison of two models is formalized by the posterior odds, as shown in Equation 1.11, which decompose prior beliefs about models (the prior odds) from the evidence contained in the data (the Bayes factor).

A well-known problem with this framework is its sensitivity to prior specification at multiple levels. First, the final posterior odds depend directly on the model priors $\pi(\mathcal{M}_k)$. An arbitrary choice, such as a "uniform" prior distribution with constant $\pi(\mathcal{M}_k)$ for all models, may have a substantial and unintended impact on the result, especially as the number of candidate models grows. Scott and Berger (2010) highlight that this can lead to significant issues with multiple testing and false positive control.

Second, as seen in Equation 1.10, the marginal likelihood is sensitive to the parameter priors $\pi(\theta_k)$. Efforts to mitigate this with "objective" priors have led to an array of Bayes factor alternatives, such as intrinsic and fractional Bayes factors. These methods use the observed data to help specify the prior distribution in some way, such as by weighting the likelihood (O'Hagan, 1995) or setting aside a portion of data for "training" (Berger &

Pericchi, 1996). However, this still requires an element of user choice, and can result in the loss of some information contained within the observed data.

A separate, practical problem with the Bayes factor is the computational complexity of integrating over the complete parameter space for marginal likelihood calculation. In practice, this integral is often sidestepped by using the BIC, which provides an asymptotic approximation to the negative log-evidence (Schwarz, 1978). This is given by

$$\text{BIC} = -2 \log \widehat{\mathcal{L}} + d \log n \tag{2.2}$$

where $\widehat{\mathcal{L}}$ is the maximum likelihood of the model at the optimal parameter values, d is the dimension of the model, and n is the sample size. (The BIC is sometimes denoted as the negative of the above; in our case, a lower value indicates a better model fit.)

A tempting idea is then to treat this approximation as an exact quantity — for example, by computing $\exp(-\frac{1}{2}\text{BIC})$ as a direct substitute for the marginal likelihood. However, Kass and Raftery (1995) show that this approximation has a relative error of $O(1)$, meaning it does not improve with n and should not be used to evaluate exact posterior probabilities. In Section 3.3.1, we provide a further discussion of this approximation in the context of two-sided hypothesis testing and e-values.

2.2.2 Recursive model updating

We now formally present a predictive resampling approach that emphasizes missing data as the source of model uncertainty. With observed $y_{1:n}$, the guiding principle is that uncertainty quantification for any statistical task, including model selection, requires the construction of a model for the data we have not observed, given what has been observed. Algorithm 1 outlines the procedure, alternating between the selection of the best available model at any given point and the imputation of a new observation. This is presented for finite N, and the choice of model stabilizes as $N \to \infty$.

Following the logic of Fong et al. (2024), we express uncertainty over the final choice of model in light of different possible realizations of the complete dataset. Rather than targeting a parameter, our inferential goal is the martingale posterior probability $\Pi_{\text{pred}}(\mathcal{M}_k \mid \mathcal{D})$ for each model, which we approximate with our Monte Carlo estimator

Algorithm 1 Predictive resampling for model uncertainty

Specify search space of candidate models $\{\mathcal{M}_k\}_{k=1}^K$
Specify consistent model selection criterion C
Set number of trials B and final sample size $N \gg n$
for b from 1 to B **do**
    for i from $n+1$ to N **do**
        Calculate $C(\mathcal{M}_k, y_{1:i-1})$ for $k \in \{1, \ldots, K\}$
        Optimize and identify best model index $\widehat{k}(i-1) = \text{argmax}_k\, C(\cdot, \cdot)$
        Set the chosen model to $\mathcal{M}_{\widehat{k}(i-1)}$ (with possible MLE(s) $\widehat{\theta}_{\widehat{k}(i-1)}$)
        Sample $Y_i \mid y_{1:i-1} \sim p(\cdot|\mathcal{M}_{\widehat{k}(i-1)}, \widehat{\theta}_{\widehat{k}(i-1)})$ and add to training data
    end for
    Calculate and record final model $\mathcal{M}^{(b)} = \mathcal{M}_{\widehat{k}(N)}$
end for
Return final probabilities $\widehat{\Pi}_{\text{pred}}(\mathcal{M}_k \mid \mathcal{D}) = B^{-1} \sum_{b=1}^B \mathbb{1}(\mathcal{M}^{(b)} = \mathcal{M}_k)$

$\widehat{\Pi}_{\text{pred}}$. The benefit of this approach is that uncertainty arises from actual observables. The only required inputs are a routine to optimize a consistent model selection criterion C and a method to generate new samples from the selected model. This inverts the inferential process: we go from decision to uncertainty, making a fully determined choice for each replicated dataset, rather than from uncertainty to decision, making a single choice based on the evidence from our one observed dataset.

For supervised data, with observed covariates $\mathbf{X}$ and outcomes $\mathbf{y}$, we adapt the framework by viewing $\mathbf{X}$ as a fixed set of support points. The target for predictive resampling is then the imputation of new outcome vectors, conditional on the fixed design matrix. In other words, at each iterative step, we replicate $\mathbf{X}$ and use the optimal model fit to the current data to sample a new $n \times 1$ outcome vector. The new vector is appended to the set of outcomes, and the process repeats. This "block resampling" keeps with the idea of "repeating the experiment", and ensures that the process does not introduce out-of-distribution bias in the covariate space. We defer further discussion of this setup to the illustration of variable selection in Section 2.4.2.

We can also consider the result of performing predictive resampling via a standard Bayesian update. In other words, rather than the deterministic optimization in Algorithm

1, we would draw y_i at each step from the full posterior predictive mixture,

$$p(\cdot \mid y_{1:i-1}) = \sum_k p(\cdot \mid \mathcal{M}_k)\pi(\mathcal{M}_k \mid y_{1:i-1}).$$

In this model-averaged resampling, the relative proportions of the final models would converge to the posterior model probabilities $\pi(\mathcal{M}_k \mid \mathcal{D})$ as $B \to \infty$.

Our approach can therefore be understood as a re-interpretation of standard Bayesian model uncertainty, where the Bayesian mixture is replaced by a plug-in selection step at each stage. The benefit of this re-interpretation is that it provides a direct, operational link between a consistent selection criterion C and the resulting model uncertainty $\widehat{\Pi}_{\text{pred}}$, bypassing the computationally complex and prior-sensitive calculation of the marginal likelihood required for the standard posterior distribution.

2.2.3 Point hypothesis example

To demonstrate, suppose we wish to compare two point hypothesis models

$$\mathcal{M}_0 : \theta = \theta_0$$

$$\mathcal{M}_1 : \theta = \theta_1$$

for the unknown mean parameter of a normal distribution with a known variance of 1. Let $\phi(y \mid \theta, 1)$ denote the probability density function for this normal distribution. Since we are comparing two simple hypotheses with fixed parameters, there is no penalty for model complexity. For observed data $y_{1:n}$, we can directly select the model $\mathcal{M}_{\widehat{k}(n)}$ that maximizes the log-likelihood criterion C, denoted as

$$\widehat{k}(n) = \operatorname*{argmax}_{k \in \{0,1\}} C(k, y_{1:n}) \text{ where } C(k, y_{1:n}) = \sum_{i=1}^{n} \log\, \phi(y_i \mid \theta_k, 1).$$

The parameter associated with the chosen model is then $\theta_{\widehat{k}(n)}$ (i.e. either θ_0 or θ_1.) In this setting, the choice of model and choice of parameter are functionally identical, but we distinguish them here to establish notation that can later be generalized.

Under the predictive resampling approach, we generate a new data point

$$Y_{n+1} \mid y_{1:n} \sim \phi(\cdot \mid \theta_{\widehat{k}(n)}, 1)$$

from the chosen model and add its realization y_{n+1} to the observed data. We then find the choice of model $\mathcal{M}_{\widehat{k}(n+1)}$ that maximizes the likelihood of the augmented data $y_{1:n+1}$, and repeat the process. Once we have imputed a sufficiently large number of additional samples $N \gg n$, we record our final choice of model. We repeat this pipeline several times and index our uncertainty over replications of the "completed" population.

Theoretical framework

We now illustrate the convergence properties of the model choice in this simple two-model setting. This serves as a demonstration of the key concepts, with a discussion for more general settings provided in Section 2.3.

Our proof strategy is to construct a likelihood ratio process that tracks the evidence for any given model relative to the model chosen at the previous step. We then show that this process is a supermartingale, meaning that its expected value at the next step, given the present, is always less than or equal to its current value. Applying a result from Doob (1953), any non-negative supermartingale must converge to a finite limit. This implies that the model choice itself eventually stabilizes as more data are imputed.

Let $\mathcal{F}_n = \sigma(y_{1:n})$ be the filtration generated by the data. For a given $k \in \{0, 1\}$, define the likelihood ratio process $L_n(k)$ as

$$L_n(k) = \frac{\prod_{i=1}^{n} \phi(y_i \mid \theta_k, 1)}{\prod_{i=1}^{n} \phi(y_i \mid \theta_{\widehat{k}(n-1)}, 1)}.$$

Note that the denominator uses the parameter $\theta_{\widehat{k}(n-1)}$ selected at the previous step. While $\theta_{\widehat{k}(n-1)}$ is a random variable depending on $y_{1:n-1}$, it is $\mathcal{F}_{n-1}$-measurable and thus treated as fixed when conditioning on the past. $L_n(k)$ is therefore the ratio of the likelihood of $\mathcal{M}_k$ to the likelihood of the previously selected model $\mathcal{M}_{\widehat{k}(n-1)}$, both evaluated on the full data $y_{1:n}$. By construction, if $\mathcal{M}_k$ is the previous best model (with $k = \widehat{k}(n-1)$), then the numerator and denominator are identical, and $L_n(k) = 1$.

We now examine the expectation of $L_n(k)$ given $\mathcal{F}_{n-1}$. Since the terms up to $n-1$ are constants conditional on $\mathcal{F}_{n-1}$, they can be factored out, yielding

$$\mathbb{E}[L_n(k) \mid \mathcal{F}_{n-1}] = \left(\frac{\prod_{i=1}^{n-1} \phi(y_i \mid \theta_k, 1)}{\prod_{i=1}^{n-1} \phi(y_i \mid \theta_{\widehat{k}(n-1)}, 1)} \right) \cdot \mathbb{E}\left[\frac{\phi(Y_n \mid \theta_k, 1)}{\phi(Y_n \mid \theta_{\widehat{k}(n-1)}, 1)} \mid \mathcal{F}_{n-1} \right].$$

The expectation in the second term is taken over the data-generating distribution

$$Y_n \mid \mathcal{F}_{n-1} \sim \phi(\cdot \mid \theta_{\widehat{k}(n-1)}, 1).$$

The expectation term thus integrates to one:

$$\mathbb{E}[\cdot] = \int_{-\infty}^{\infty} \frac{\phi(y \mid \theta_k, 1)}{\phi(y \mid \theta_{\widehat{k}(n-1)}, 1)} \phi(y \mid \theta_{\widehat{k}(n-1)}, 1)\, dy = \int_{-\infty}^{\infty} \phi(y \mid \theta_k, 1)\, dy = 1.$$

Substituting this back, we find the conditional expectation is

$$\mathbb{E}[L_n(k) \mid \mathcal{F}_{n-1}] = \frac{\prod_{i=1}^{n-1} \phi(y_i \mid \theta_k, 1)}{\prod_{i=1}^{n-1} \phi(y_i \mid \theta_{\widehat{k}(n-1)}, 1)}.$$

Now, we compare this to $L_{n-1}(k)$, which is defined recursively by the same process as

$$L_{n-1}(k) = \frac{\prod_{i=1}^{n-1} \phi(y_i \mid \theta_k, 1)}{\prod_{i=1}^{n-1} \phi(y_i \mid \theta_{\widehat{k}(n-2)}, 1)}.$$

By definition, $\theta_{\widehat{k}(n-1)}$ maximizes the likelihood given $y_{1:n-1}$. Therefore, this must be greater than or equal to the likelihood when using $\theta_{\widehat{k}(n-2)}$ from the previous step:

$$\prod_{i=1}^{n-1} \phi(y_i \mid \theta_{\widehat{k}(n-1)}, 1) \geq \prod_{i=1}^{n-1} \phi(y_i \mid \theta_{\widehat{k}(n-2)}, 1).$$

Since the likelihoods are positive, this implies

$$\frac{1}{\prod_{i=1}^{n-1} \phi(y_i \mid \theta_{\widehat{k}(n-1)}, 1)} \leq \frac{1}{\prod_{i=1}^{n-1} \phi(y_i \mid \theta_{\widehat{k}(n-2)}, 1)}.$$

Multiplying by the positive numerator $\prod_{i=1}^{n-1} \phi(y_i \mid \theta_k, 1)$ preserves the inequality:

$$\underbrace{\frac{\prod_{i=1}^{n-1} \phi(y_i \mid \theta_k, 1)}{\prod_{i=1}^{n-1} \phi(y_i \mid \theta_{\widehat{k}(n-1)}, 1)}}_{\mathbb{E}[L_n(k) \mid \mathcal{F}_{n-1}]} \leq \underbrace{\frac{\prod_{i=1}^{n-1} \phi(y_i \mid \theta_k, 1)}{\prod_{i=1}^{n-1} \phi(y_i \mid \theta_{\widehat{k}(n-2)}, 1)}}_{L_{n-1}(k)}.$$

Thus, we have shown $\mathbb{E}[L_n(k) \mid \mathcal{F}_{n-1}] \leq L_{n-1}(k)$, confirming that $L_n(k)$ is a non-negative supermartingale. Under Doob's martingale convergence theorem, $L_n(k)$ therefore converges almost surely to a finite limit $L_\infty(k)$ for both $k \in \{0, 1\}$ (Doob, 1953). The parameter is selected which maximizes $L_\infty(k)$; in the limit, the model selection process will eventually converge to a single choice.

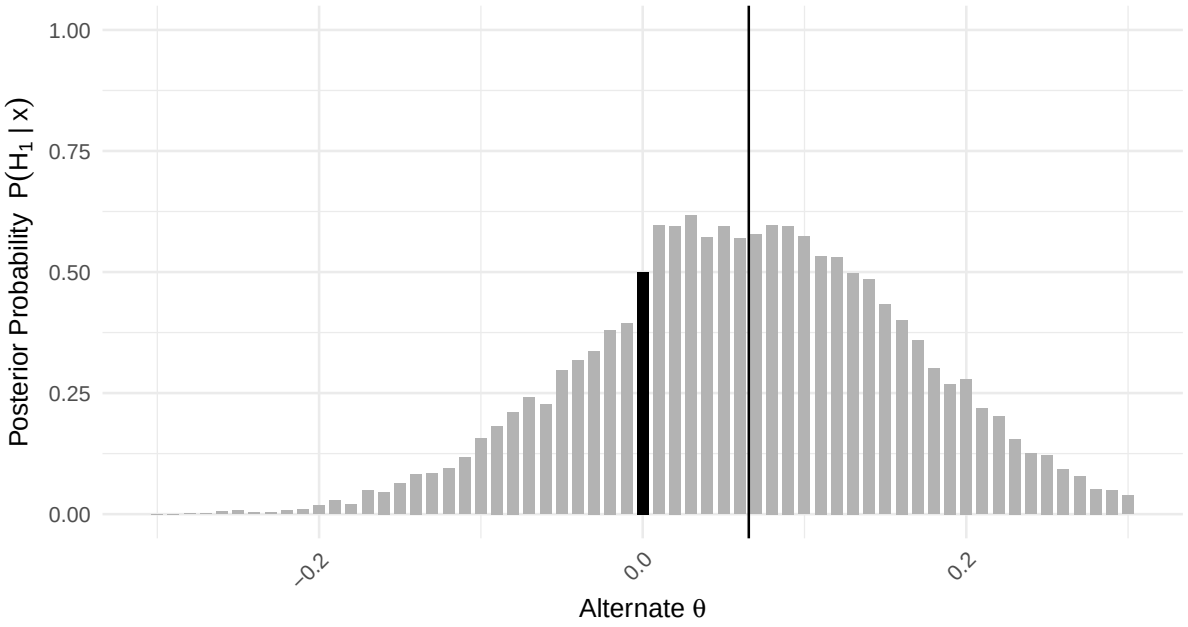


Figure 2.2: Proportion of trials in which alternate model H_1 is selected as alternate mean θ_1 varies from -0.3 to 0.3. The observed sample mean $\bar{x}$ is denoted by the vertical line, while the baseline of $\theta_0 = \theta_1 = 0$ is denoted by the black bar.

Experimental results

Experimentally, we generate $n = 100$ data points from $\mathcal{N}(0, 1)$ as our observed data, which gives $\bar{x} = 0.066$. We set $\theta_0 = 0$ and vary θ_1 in the range {-0.3, -0.29, -0.28, ..., 0.3}, and impute an additional $N = n + 2500$ samples over $B = 1000$ trials. Figure 2.1 demonstrates that this is a sufficiently large sample size for the final model choice to converge, in the example case where $\theta_1 = 0.1$. For the initial iteration, we will always select the model where θ_k is closer to the observed sample mean of 0.066; however, from the second iteration forwards, as the imputed data points introduce uncertainty, it becomes possible to switch between models. We would expect to see that H_1 is chosen most frequently in cases where the alternative mean θ_1 is closer to $\bar{x}$. Figure 2.2 confirms this: with the baseline case of $\theta_0 = \theta_1 = 0$ marked at 50% in light blue, H_1 is selected more frequently in cases where θ_1 is closer to $\bar{x}$, and less frequently otherwise.

As an alternative means of explanation, Figure 2.1 captures the idea that model uncertainty can often be expressed in terms of a decision rule on some population-level summary statistic. In this case, when comparing $H_0 : \theta_0 = 0$ against $H_1 : \theta_1 = 0.1$, we

would expect this decision threshold to be at $\bar{x} = 0.05$. We in fact see a clear divergence between populations with $\bar{x}_N > 0.05$, where we select H_1, and populations with $\bar{x}_N < 0.05$, where we select H_0. By targeting a probability distribution over the complete population $p(y_{n+1:\infty} \mid y_{1:n})$, what we are ultimately attempting to retrieve is a probability distribution over this summary statistic computed on the infinite population. This partitions the space of possible observable datasets into critical regions, which we can then interpret in the context of model uncertainty.

2.2.4 Consistent model selection

This framework requires the specification of a model selection criterion, which will be used to select the best model at each step for the generation of y_{n+1} given observed $y_{1:n}$, and ultimately to select the best final model for each possible realization of the complete data. The key requirement for this criterion is therefore that it be consistent, i.e., that the probability of selecting the correct model converges to 1 as $N \to \infty$ (Claeskens & Hjort, 2008). The BIC provides consistency, as well as a clear connection to Bayesian inference, and therefore serves as our preferred one-step model selection rule (Schwarz, 1978).

To further motivate the BIC, we can consider its more general interpretation in the context of predictive evaluation. A scoring rule is a general summary measure for a probabilistic forecast (Matheson & Winkler, 1976). Consider the score S which is defined as the sum of the individual log-predictive probabilities

$$S(y_{1:n}, \mathcal{M}) = \sum_{i=1}^{n} \log p_{\mathcal{M}}(y_i \mid y_{1:i-1})$$

with the data evaluated as if they had appeared in sequence. This is a proper scoring rule (Gneiting & Raftery, 2007) in that its expected value is maximized when the true distribution is used for forecasting. In fact, Fong and Holmes (2020) show that it is the unique scoring rule which guarantees coherent model evaluation.

Dawid (1984) shows that this scoring rule is equivalent to the Bayes factor, which the BIC approximates. This means that the difference in scores between models $\mathcal{M}_1$ and $\mathcal{M}_2$,

$$\log BF = S(y_{1:n}, \mathcal{M}_1) - S(y_{1:n}, \mathcal{M}_2),$$

can be interpreted as the "weight of evidence" in favor of the first model (Good, 1952). The BIC therefore provides a general method for comparing forecasting rules (Gneiting & Raftery, 2007), supporting its use as a tool to select the best model for a one-step predictive update $p(y_{n+1} \mid y_{1:n})$.

We now briefly discuss other common approaches for model comparison. Cross-validation splits the observed data (of size n) into two parts, uses the first $n-p$ samples to train the model, and then evaluates predictive performance on the held-out validation set of p samples. Full leave-p-out cross-validation requires fitting the model $\binom{n}{p}$ times, which can be computationally infeasible, and so simplifications such as k-fold cross-validation (dividing the data into blocks of size n/k for testing) are often preferred (Geisser, 1975). The marginal likelihood is equivalent to exhaustive leave-p-out cross-validation averaged over all values of p and all possible test sets, providing an underlying link between these approaches (Fong & Holmes, 2020; Gneiting & Raftery, 2007).

One of the most common simplifications is LOO-CV, where the model is fit and evaluated n times. This is equivalent to k-fold cross-validation with $k = n$. However, Shao (1993) shows that LOO-CV is asymptotically inconsistent for linear models, meaning it should not be applied in a context where our eventual target is a decision based on the complete data. This can be rectified by using leave-p-out cross-validation with $p/n \to 1$ as $n \to \infty$, where the size of the validation set grows alongside the sample size, though this again faces computational constraints as n grows. Intuitively, this is necessary because a larger validation set provides a smoother assessment of prediction error; optimizing the fit on a single observation at a time can select unnecessarily large models. Yang (2007) extends these findings to nonparametric models, while Vehtari and Lampinen (2002) motivate cross-validation in the Bayesian context. Vehtari et al. (2017) and Sivula et al. (2023) discuss Bayesian LOO-CV; it is noted to be unreliable in certain common use cases, such as comparing similar models or misspecified models.

The AIC approximates predictive fit, given by

$$\mathrm{AIC} = 2d - 2\log\widehat{\mathcal{L}}$$

where the penalty for model complexity is fixed, unlike the variable penalty of $\log n$ applied in the BIC (Akaike, 1974). However, the AIC is asymptotically equivalent to LOO-CV, and so it is also asymptotically inconsistent (Shao, 1993), for the similar reason that it tends to overfit to models that are too complex.

We acknowledge that the consistency of the BIC is limited to the $\mathcal{M}$-closed setting, where the true model is assumed to be contained in the space of candidate models (Bernardo & Smith, 2004). In the $\mathcal{M}$-open setting, where the true data-generating process lies outside the specified model class, predictive performance can be improved through the use of stacking or other methods based on cross-validation or AIC (Yao et al., 2018). However, our aim here is to develop a principled, general approach for computing posterior model probabilities within a given class, rather than optimizing out-of-sample prediction for the resulting Bayesian model average.

Overall, our approach is grounded in the idea that statistical uncertainty stems from missing data, and if complete data were available, we could reliably identify the correct model. For well-calibrated uncertainty propagation, we need a criterion that can consistently choose the right model once the missing observations are recovered. While this does not necessarily have to be the BIC, it should not be methods like LOO-CV or AIC, as they lack the guarantee of asymptotic consistency. In Section 2.4.2, we demonstrate that the AIC selects overly complex models as expected when used for predictive resampling in the context of variable selection.

2.3 Convergence of one-step updates

We now consider the convergence of the model choices returned by predictive resampling in a general setting. Following the specification in Algorithm 1, we start with a finite set of candidate models $\{\mathcal{M}_k\}_{k=1}^K$ and observed data $y_{1:n}$. We apply the update from Equation 2.1 to recursively select models and sample new observations. At step i, the chosen model is denoted by $\mathcal{M}_{\widehat{k}(i)}$, and any associated parameter MLEs are $\widehat{\theta}_{\widehat{k}(i)}$.

Our aim is to explore how the model choice $\widehat{k}(n)$ converges to some $k(\infty)$ as the sample size grows from $n \to \infty$. To extend the proof strategy applied in Section 2.2.3,

we construct a likelihood ratio process $L_n(k)$ that measures the evidence for any model relative to the model selected at the previous step, then aim to show that this process is a supermartingale. Specifically, let $\mathcal{F}_n = \sigma(y_{1:n})$ be the filtration generated by the data. For a given $k \in \{1, \ldots, K\}$, define $L_n(k)$ as

$$L_n(k) = \frac{\prod_{i=1}^{n} p(y_i \mid \mathcal{M}_k, \widehat{\theta}_{k(n)})}{\prod_{i=1}^{n} p(y_i \mid \mathcal{M}_{\widehat{k}(n-1)}, \widehat{\theta}_{\widehat{k}(n-1)})}.$$

As before, the denominator uses the parameters $\widehat{\theta}_{\widehat{k}(n-1)}$ estimated at the previous step. The process $L_n(k)$ represents the ratio of the maximum likelihood of $\mathcal{M}_k$ to the maximum likelihood of the previously selected model $\mathcal{M}_{\widehat{k}(n-1)}$.

The next step is to examine the conditional expectation of $L_n(k)$ given $\mathcal{F}_{n-1}$. Here, we encounter a technical obstacle regarding the interdependency of the parameter re-estimation. In particular, the updated parameter $\widehat{\theta}_{k(n)}$ is a function of Y_n, meaning it is not $\mathcal{F}_{n-1}$-measurable and cannot be factored out of the conditional expectation. To proceed with the derivation using the strategy from Section 2.2.3, we must introduce two substantial simplifying assumptions: first, that the parameter estimate does not shift instantaneously with the new data point, or $\widehat{\theta}_{k(n)} = \widehat{\theta}_{k(n-1)}$, and second, that the candidate models possess equal parameter dimensions.

Consequently, the remainder of this proof only formally demonstrates the convergence mechanism for the restricted setting where the parameter estimates are treated as fixed and dimensions are equal. This isolates the behavior of the model selection criterion from the variance of the parameter estimation. Rigorously handling the parameter variance to extend this result to the general setting remains an ongoing priority for future research. One potential strategy involves constructing a joint likelihood ratio process $L_n(k, \theta)$ to show (nested) joint convergence. A complementary approach relies on the asymptotic stability of the MLE. In practice, parameter estimates for common models typically stabilize once the sample size is sufficiently large, implying that the stepwise difference $\widehat{\theta}_{k(n)} - \widehat{\theta}_{k(n-1)}$ effectively vanishes. Asymptotic arguments may therefore be able to bound the error and validate the convergence result for large n.

Continuing the argument for the restricted case, we have

$$\mathbb{E}[L_n(k) \mid \mathcal{F}_{n-1}] = \left(\frac{\prod_{i=1}^{n-1} p(y_i \mid \mathcal{M}_k, \widehat{\theta}_{k(n-1)})}{\prod_{i=1}^{n-1} p(y_i \mid \mathcal{M}_{\widehat{k}(n-1)}, \widehat{\theta}_{\widehat{k}(n-1)})} \right) \cdot \mathbb{E}\left[\frac{p(Y_n \mid \mathcal{M}_k, \widehat{\theta}_{k(n-1)})}{p(Y_n \mid \mathcal{M}_{\widehat{k}(n-1)}, \widehat{\theta}_{\widehat{k}(n-1)})} \mid \mathcal{F}_{n-1} \right].$$

The expectation in the second term is taken over the data-generating distribution

$$Y_n \mid \mathcal{F}_{n-1} \sim p(\cdot \mid \mathcal{M}_{\widehat{k}(n-1)}, \widehat{\theta}_{\widehat{k}(n-1)}).$$

As in our simple proof, this integral evaluates to one:

$$\mathbb{E}[\cdot] = \int \frac{p(y \mid \mathcal{M}_k, \widehat{\theta}_{k(n-1)})}{p(y \mid \mathcal{M}_{\widehat{k}(n-1)}, \widehat{\theta}_{\widehat{k}(n-1)})} p(y \mid \mathcal{M}_{\widehat{k}(n-1)}, \widehat{\theta}_{\widehat{k}(n-1)})\, dy = 1.$$

Substituting this back, we find the conditional expectation is

$$\mathbb{E}[L_n(k) \mid \mathcal{F}_{n-1}] = \frac{\prod_{i=1}^{n-1} p(y_i \mid \mathcal{M}_k, \widehat{\theta}_{k(n-1)})}{\prod_{i=1}^{n-1} p(y_i \mid \mathcal{M}_{\widehat{k}(n-1)}, \widehat{\theta}_{\widehat{k}(n-1)})}.$$

We now compare this to $L_{n-1}(k)$, which is defined recursively as

$$L_{n-1}(k) = \frac{\prod_{i=1}^{n-1} p(y_i \mid \mathcal{M}_k, \widehat{\theta}_{k(n-1)})}{\prod_{i=1}^{n-1} p(y_i \mid \mathcal{M}_{\widehat{k}(n-2)}, \widehat{\theta}_{\widehat{k}(n-2)})}.$$

By definition, $\mathcal{M}_{\widehat{k}(n-1)}$ (with $\widehat{\theta}_{\widehat{k}(n-1)}$) maximizes the likelihood at step $n-1$, so its likelihood is greater than or equal to the likelihood of the model chosen at step $n-2$:

$$\prod_{i=1}^{n-1} p(y_i \mid \mathcal{M}_{\widehat{k}(n-1)}, \widehat{\theta}_{\widehat{k}(n-1)}) \geq \prod_{i=1}^{n-1} p(y_i \mid \mathcal{M}_{\widehat{k}(n-2)}, \widehat{\theta}_{\widehat{k}(n-2)}).$$

Since the likelihoods are positive, this implies

$$\frac{1}{\prod_{i=1}^{n-1} p(y_i \mid \mathcal{M}_{\widehat{k}(n-1)}, \widehat{\theta}_{\widehat{k}(n-1)})} \leq \frac{1}{\prod_{i=1}^{n-1} p(y_i \mid \mathcal{M}_{\widehat{k}(n-2)}, \widehat{\theta}_{\widehat{k}(n-2)})}.$$

Multiplying by the positive numerator preserves the inequality:

$$\underbrace{\frac{\prod_{i=1}^{n-1} p(y_i \mid \mathcal{M}_k, \widehat{\theta}_{k(n-1)})}{\prod_{i=1}^{n-1} p(y_i \mid \mathcal{M}_{\widehat{k}(n-1)}, \widehat{\theta}_{\widehat{k}(n-1)})}}_{\mathbb{E}[L_n(k) \mid \mathcal{F}_{n-1}]} \leq \underbrace{\frac{\prod_{i=1}^{n-1} p(y_i \mid \mathcal{M}_k, \widehat{\theta}_{k(n-1)})}{\prod_{i=1}^{n-1} p(y_i \mid \mathcal{M}_{\widehat{k}(n-2)}, \widehat{\theta}_{\widehat{k}(n-2)})}}_{L_{n-1}(k)}.$$

Thus, we have shown $\mathbb{E}[L_n(k) \mid \mathcal{F}_{n-1}] \leq L_{n-1}(k)$, meaning that $L_n(k)$ is a non-negative supermartingale. The model $\mathcal{M}_{\widehat{k}(n)}$ selected by sequential maximum likelihood with fixed parameter values and equal dimensions therefore converges almost surely as $n \to \infty$.

In practice, a penalty term is usually applied to the maximum likelihood due to the risk of overfitting. We can extend the logic of the restricted proof above to characterize the convergence behavior of such penalized objectives. Generally, a penalized likelihood criterion can be written using the form

$$c(n, d_k) \prod_{i=1}^{n} p(y_i \mid \mathcal{M}_k, \theta_k)$$

where c is a penalty function in terms of the sample size n and model dimension d_k. For the AIC, this penalty is $c(n, d_k) = e^{-d_k}$ and remains constant with respect to sample size. For the BIC, this penalty is $c(n, d_k) = e^{-\frac{d_k}{2} \log n}$. The strength of the penalty increases with the sample size, meaning that the multiplicative term $c(n, d_k)$ decreases.

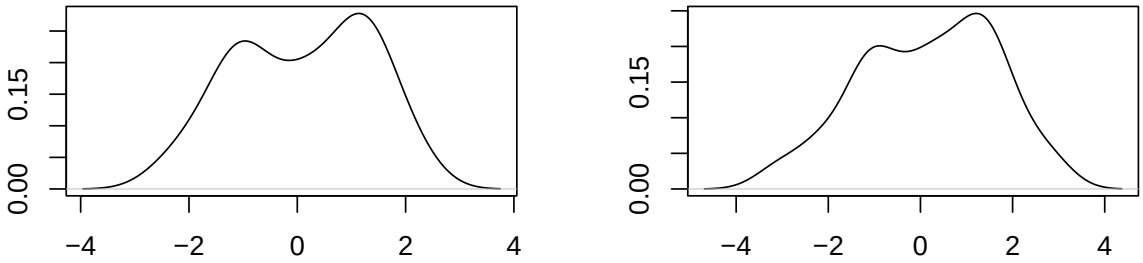
For convergence under these conditions, we consider the penalized likelihood ratio process. The conditional expectation of this process includes the ratio of the penalty terms at the current and previous steps, $\frac{c(n,d_k)}{c(n-1,d_k)}$. If we continue with the assumption of equal dimensions, then this ratio is one for the AIC. For the BIC, the ratio is less than one, because $c(n, d_k)$ decreases as $n \to \infty$. The unpenalized process is already a supermartingale as shown previously, so multiplying the expectation by a factor less than or equal to one preserves the inequality $\mathbb{E}[L_n(k) \mid \mathcal{F}_{n-1}] \leq L_{n-1}(k)$. Thus, the model choice stabilizes almost surely even when these standard penalties are applied.

2.4 Illustrations

In this section, we demonstrate model uncertainty via predictive resampling on example problems from density estimation and variable selection. Code for all illustrations can be found at `https://github.com/vshirvaikar/MPModel`.

2.4.1 Density estimation

A typical model selection question is the number of components required in a finite Gaussian mixture model (GMM) for density estimation.



(a) Density with $\sigma = 0.6$ for components (b) Density with $\sigma = 0.9$ for components

Figure 2.3: Kernel density plots for data generated from GMM with 2 components.

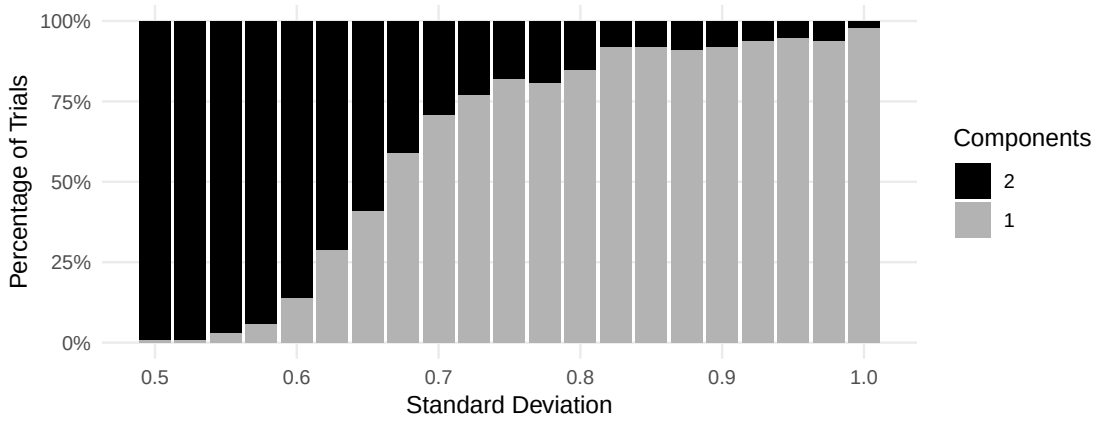


Figure 2.4: Posterior uncertainty over number of components G via resampling.

Background

To begin, we demonstrate the predictive resampling framework on a simple univariate example. We simulate $n = 50$ samples from a GMM with $G = 2$ components

$$f_0(y) = \frac{1}{2}\mathcal{N}(y \mid -1, \sigma^2) + \frac{1}{2}\mathcal{N}(y \mid 1, \sigma^2)$$

and rescale the individual data points to vary the standard deviation between $\sigma = 0.5$ and $\sigma = 1$. Figure 2.3 shows sample kernel density plots with a fixed bandwidth of 0.5 for the generated data. The two separate peaks are clearly visible when $\sigma = 0.6$ (Figure 2.3a), for example, but begin to merge together when $\sigma = 0.9$ (Figure 2.3b).

Under the predictive resampling framework, the goal is to identify an optimal GMM at each step, which is then used to recursively predict one additional data point. Following Fraley and Raftery (2002), we apply the expectation-maximization (EM) algorithm for clustering to our observed data of size n. The convergence properties of EM for Gaussian mixtures have been widely studied (Xu & Jordan, 1996), and while its consistency is not

guaranteed in all conditions, it is well-established for simple and correctly specified models such as the ones we explore here (Balakrishnan et al., 2017).

We vary the number of components G across a specified range and select the model with the lowest BIC. We assume equal variances, with a dimension of

$$d = G \text{ means} + (G-1) \text{ proportions} + 1 \text{ common variance}$$

in the BIC calculation for a total of $2G$; if we allow unequal variances, the final term becomes G as well for a total of $3G - 1$. We simulate a new data point from this model, augmenting our dataset to size $n + 1$, and repeat the above process. This continues for several additional points, yielding a final resampled dataset of size N, after which we record the final selected model for the "complete" data. We replicate this across several trials and then index our uncertainty over the distribution of final models.

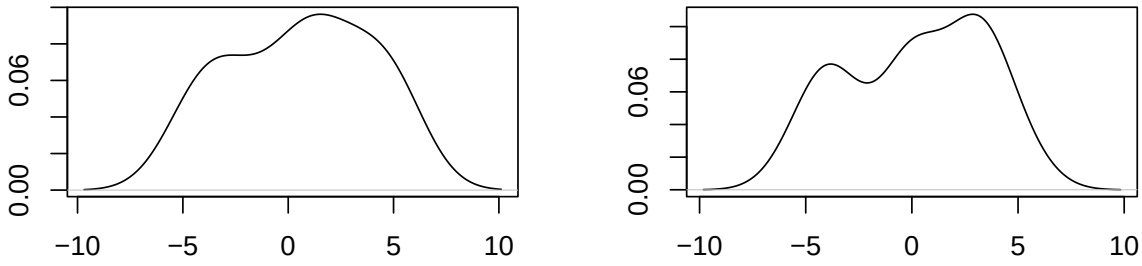
We implement this pipeline using the `mclust` package in R (Scrucca et al., 2023), with candidate models containing either 1 or 2 components, and recursively simulate $N = n + 200$ new observations per trial across a total of $B = 100$ trials for each value of σ. We empirically find that this value of N is sufficiently large for the model $\mathcal{M}_{\widehat{k}(N)}$ to closely approximate the final $\mathcal{M}_{k(\infty)}$; convergence diagrams can be found in the appendix. Figure 2.4 shows the distribution of the final number of components across all trials, as σ increases from 0.5 to 1. As expected, the posterior probability of selecting only $G = 1$ component increases as the observed data becomes more unimodal.

Simulated example

As a more complex example, we generate $n = 20$ and $n = 50$ data points from a GMM with $G = 3$ components

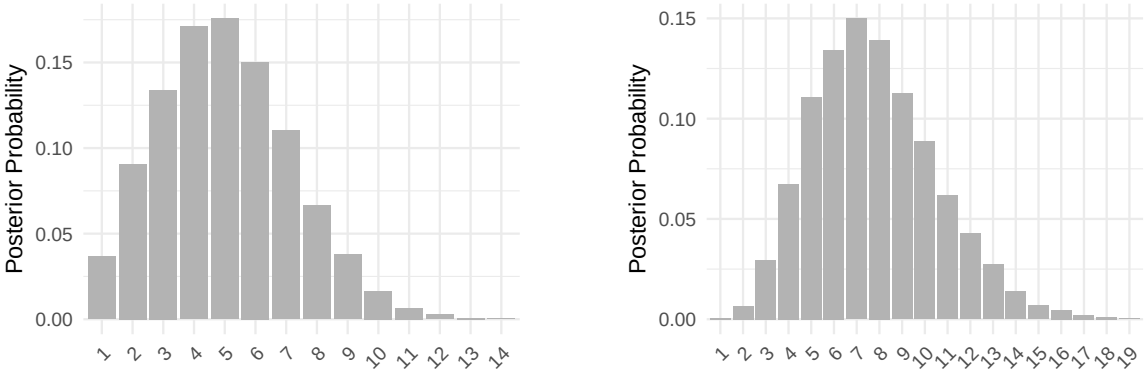
$$f_0(y) = 0.4\mathcal{N}(y \mid -3, 1) + 0.3\mathcal{N}(y \mid 0, 1) + 0.3\mathcal{N}(y \mid 4, 1)$$

where the goal is to identify and return uncertainty around the true value of G. Figure 2.5 shows kernel density plots for the data, where the three peaks are less clear in the $n = 20$ case (Figure 2.5a) than the $n = 50$ case (Figure 2.5b).



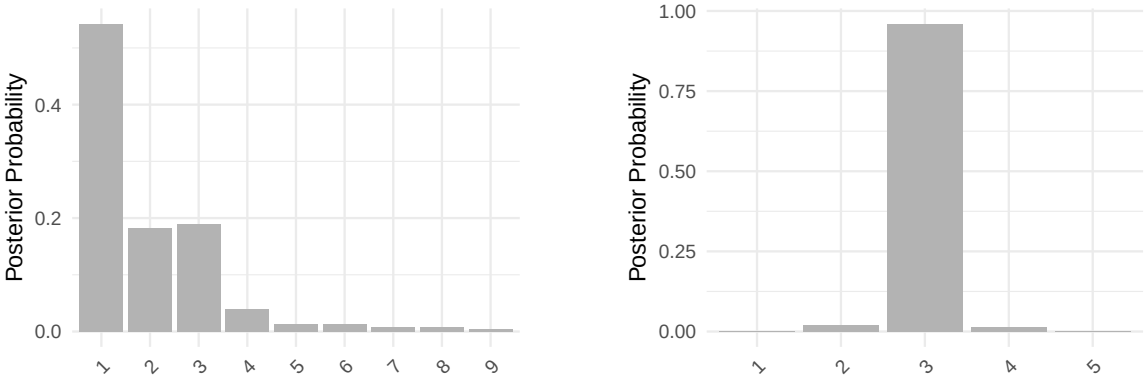
(a) Density for $n = 20$ observations (b) Density for $n = 50$ observations

Figure 2.5: Kernel density plots for data generated from GMM with 3 components.



(a) Components for $n = 20$ observations (b) Components for $n = 50$ observations

Figure 2.6: Posterior uncertainty over number of components G sampled in DPMM.



(a) Components for $n = 20$ observations (b) Components for $n = 50$ observations

Figure 2.7: Posterior uncertainty over number of components G via resampling.

Following Fong et al. (2024), we apply Dirichlet process mixture modeling (DPMM) as our baseline for comparison. The DPMM (Escobar & West, 1995) is often considered the gold standard for Bayesian nonparametric density estimation. It samples from a Dirichlet process distribution $DP(H, \alpha)$ where each mixture component has its underlying parameters drawn from the base distribution H, and the probability of forming new components is determined by the concentration parameter α. This generative scheme is often described metaphorically as a "Chinese restaurant process", where new data points are "customers" who are seated at 'tables" (clusters) with a probability proportional to the table's popularity, or at a new, empty table with a probability proportional to α.

In the following examples, we focus on comparing the number of components G estimated under each model. We acknowledge the key theoretical consideration that the DPMM is not a suitable tool for estimating a "true" number of components. Specifically, it is a known result that the posterior on G asymptotically tends towards ∞ as n increases (Cai et al., 2020; Yang et al., 2019).

Our comparative aim is to highlight a different point: the DPMM's results, as with many standard Bayesian methods, are highly sensitive to the prior distribution. In applied analysis, prior specification (for both H and α) can be a key feature, as it allows an expert to encode rich, suitable prior information for the specific dataset at hand. However, this flexibility also means that meaningful estimation is dependent on careful, subjective tuning, as different prior choices can lead to vastly different outcomes. Predictive resampling, in contrast, is designed for a different and more constrained task; it replaces the multi-dimensional, subjective prior specification of the DPMM with a single, transparent assumption in the form of a principled model selection criterion. We therefore focus on how predictive resampling transforms this criterion into a probabilistic measure of uncertainty, while bypassing the need for complex prior elicitation and tuning.

We implement DPMM with the `dirichletprocess` package in R (Ross & Markwick, 2019). For Gaussian mixtures, the package default assumes a Normal-Inverse-Gamma base distribution H with hyperparameters $(\mu_0, \kappa_0, \alpha_0, \beta_0) = (0, 1, 1, 1)$, and places a Gamma$(2, 4)$ prior distribution on α. In practical implementations, directly evaluating

the relevant joint distribution is not possible, so DPMM uses Gibbs sampling to return uncertainty over the number of components, their means, variances, and weights. Over eight sampling chains, we discard the first 500 iterations as burn-in and retain the next 2,000 iterations. Figure 2.6 displays the distribution of the number of components sampled in the DPMM. For $n = 20$ (Figure 2.6a), the mode is 5 components, and for $n = 50$ (Figure 2.6b) the mode is 7 components. Without subjective prior tuning, the default DPMM returns a component structure more complex than the "true" $G = 3$.

For the resampling approach, we implement EM clustering with candidate models ranging from 1 to 9 components. Models with both equal and unequal variances are tested, with differing dimension penalties in the BIC calculation as noted previously. We recursively simulate $N = n + 600$ new observations per trial across a total of $B = 400$ trials; this value of N is again sufficiently large that $\mathcal{M}_{k(N)}$ closely approximates $\mathcal{M}_{k(\infty)}$, with convergence diagrams available in the appendix.

Figure 2.7 shows the distribution of the final number of components across all trials. In the $n = 20$ case, the initial model with the best BIC has 1 component, but our method reveals that this choice is highly uncertain. The final model transitions to the true $G = 3$ in 19% of trials, with the posterior probability spread across values up to $G = 9$. In the $n = 50$ case, the initial model with the best BIC is $G = 3$, and our method confirms this choice with high confidence, as 96% of trials converge to this model. These results demonstrate that predictive resampling is not just applying a penalty; it is translating the selection criterion's own confidence into a coherent probability. The method accurately reflects that the BIC's choice is uncertain at $n = 20$ (yielding a wide posterior) but decisive at $n = 50$ (yielding a sharp posterior).

Real-world example

To demonstrate on a real-world example, we consider the galaxies dataset from Roeder (1990), which contains velocity measurements for $n = 82$ galaxies in the Corona Borealis region. Figure 2.8 shows a kernel density plot of the data, where the goal is to group similar galaxies by velocity. This dataset has been used for univariate clustering analysis across several previous works (Richardson & Green, 1997; Rodríguez et al., 2025).

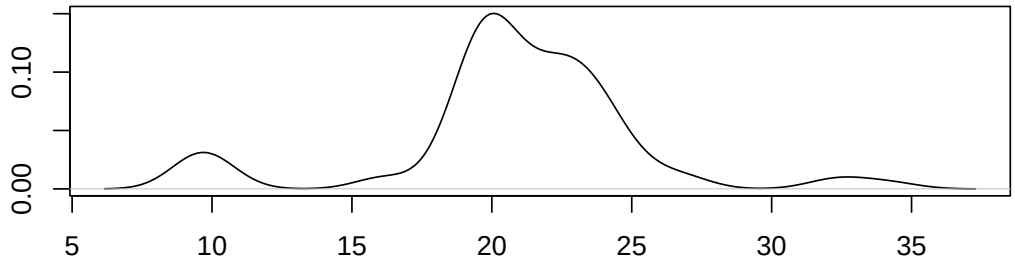


Figure 2.8: Kernel density plot for galaxies dataset.

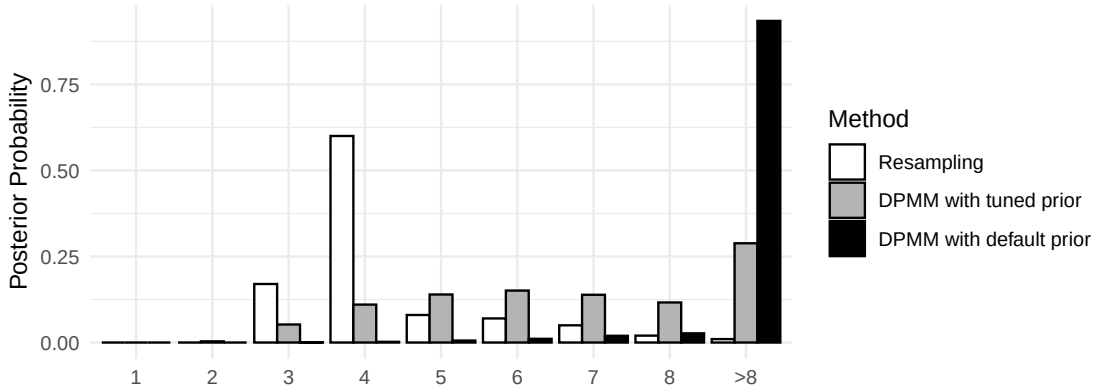


Figure 2.9: Posterior uncertainty over number of components G via resampling (white); DPMM with tuned prior distribution $\gamma(1, 8)$ for concentration parameter α (gray); and DPMM with default prior distribution $\gamma(2, 4)$ for concentration parameter (black).

We again use the `dirichletprocess` package in R, implemented via eight Gibbs sampling chains. To illustrate the effect of prior specification, we compare two different choices of prior distribution. First, we set the prior cluster mean μ_0 to the observed mean of 20.83, but otherwise use the package defaults, including $\alpha \sim \text{Gamma}(2, 4)$. Second, we apply a more conservative prior distribution on the concentration parameter, changing it to $\alpha \sim \text{Gamma}(1, 8)$.

Figure 2.9 shows the results. The default prior distribution (shaded in black) overestimates the number of components as expected, with a mode of 15. Simply tuning the α prior distribution to the more conservative Gamma$(1, 8)$ (shaded in gray) significantly changes the result, shifting the new mode to 6 components. This demonstrates that the DPMM is not providing a single answer, but rather a flexible estimate that is highly dependent on careful, subjective prior specification.

In contrast, the resampling approach returns a probabilistic answer to the model selection question based on a single, transparent criterion. We again use the `mclust`

package for candidate models from 1 to 9 components (with equal and unequal variances) and the BIC. We recursively simulate $N = n + 1500$ new observations per trial across $B = 100$ trials; this sample size is sufficient for the model choice to converge, with a diagram available in the appendix. Figure 2.9 shows the resulting martingale posterior (shaded in white), which has a realistic mode of 4 components.

This illustration showcases how predictive resampling can convert any general model selection technique into a probabilistic quantification of model uncertainty. We demonstrate here using density estimation with `mclust`, but the above framework would apply for any other package or method that allows the user to compare models using a consistent criterion and then sample a new observation from the best model.

2.4.2 Variable selection

Another common model uncertainty question is variable selection in regression, where the goal is to identify the relevant covariates in a given $n \times p$ design matrix $\mathbf{X}$ with respect to an observed $n \times 1$ outcome vector $\mathbf{y}$.

Background

As briefly discussed in Section 2.2.2, the predictive resampling framework must be adapted for supervised learning. Rather than imputing a single new observation y_i at a time (as in Algorithm 1), we treat the observed design matrix $\mathbf{X}$ as a fixed set of support points. We then sequentially resample new $n \times 1$ outcome vectors $\mathbf{Y}$, conditional on this fixed $\mathbf{X}$. This results in a block-iterative process, which we refer to as "block resampling". At each step m of a single resampling trial, we:

1. Define the current dataset as consisting of the fixed design matrix $\mathbf{X}$ replicated m times and the set of m previously sampled outcome vectors, $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$, where $\mathbf{y}_1$ is the original observed outcome vector.

2. Find the best model $\mathcal{M}_{\widehat{k}(m)}$ for the current data, according to a pre-specified model selection criterion such as the BIC.

3. Sample a new $n \times 1$ outcome vector $\mathbf{Y}_{m+1} \sim p(\cdot \mid \mathbf{X}, \mathcal{M}_{\widehat{k}(m)})$ and append its realization to the data.

This is repeated for M steps, resulting in a final sample size of $N = n \times (M+1)$. The logic is identical to Algorithm 1, but the object being resampled is an n-dimensional vector $\mathbf{y}$ conditional on $\mathbf{X}$ instead of a 1-dimensional scalar y_i. By only resampling outcomes at these fixed support points, we condition on the observed covariate distribution and avoid introducing out-of-distribution bias.

A key ingredient in the block resampling process is identification of the optimal model at each step m. With a pre-specified criterion such as the BIC, a full optimization would require evaluating 2^p candidate models at each iteration, a task that is computationally infeasible for even moderate p. In the implementation below, we therefore apply forward stepwise regression to approximate the BIC-optimal model at each step, using the `MASS` package in R (Venables & Ripley, 2002). We begin with the null model (intercept only) and greedily add terms one at a time that most improve the model BIC until no further improvements can be found (Efroymson, 1960). We acknowledge the known limitations of this approach: greedy search is not guaranteed to find the global BIC-optimal model, and different stepwise procedures (e.g., forward versus backward selection) are not always consistent with each other. A full, non-greedy optimization of the model selection criterion at each resampling step remains a significant area for future improvement.

A potential path to more efficient optimization lies in a direct analysis of the BIC calculation. The BIC is a penalized likelihood criterion, which for a linear model is a direct function of the Residual Sum of Squares (RSS). For a candidate model with d parameters (where $d \leq p$ depending on the selected variables), the calculation is

$$\text{BIC} = N \log(\text{RSS}/N) + d \log(N)$$

where N is the total sample size (here, $N = nm$ at step m). The computational bottleneck is thus the need to calculate the RSS for all 2^p models.

At the first step with $n \times 1$ outcome vector $\mathbf{y}_1$, the RSS is given by

$$\text{RSS} = \mathbf{y}_1'(\mathbf{I} - \mathbf{H}_d)\mathbf{y}_1$$

where $\mathbf{H}_d$ is the hat matrix, $\mathbf{H}_d = \mathbf{X}_d(\mathbf{X}_d'\mathbf{X}_d)^{-1}\mathbf{X}_d'$, which projects the outcome vector onto its fitted values $\hat{\mathbf{y}}_1 = \mathbf{H}_d\mathbf{y}_1$. The $n \times d$ design matrix $\mathbf{X}_d$ in this context contains only the selected covariates. As we proceed to step m of predictive resampling, the augmented data now consists of the stacked $nm \times 1$ outcome vector $\mathbf{y}^{(m)} = [\mathbf{y}_1, \ldots, \mathbf{y}_m]$ and the $nm \times d$ stacked design matrix (formed by m copies of $\mathbf{X}_d$). To compute the RSS for this stacked system, one would need to construct its $nm \times nm$ hat matrix. The computational expense of this operation — specifically, forming and inverting the $d \times d$ cross-product matrix for the large stacked design matrix — makes a brute-force search over every combination of covariates prohibitive.

However, the key insight is that this large matrix operation is unnecessary. The $d \times d$ cross-product matrix for the stacked system simplifies algebraically as

$$(\mathbf{X}_d'\mathbf{X}_d + \ldots + \mathbf{X}_d'\mathbf{X}_d)^{-1} = (m \cdot \mathbf{X}_d'\mathbf{X}_d)^{-1} = \frac{1}{m}(\mathbf{X}_d'\mathbf{X}_d)^{-1}.$$

This allows the full RSS for the nm data points to be calculated efficiently using only the small, original $n \times n$ hat matrix $\mathbf{H}_d$,

$$\text{RSS}_m = \left(\sum_{l=1}^{m} \mathbf{y}_l'\mathbf{y}_l\right) - m(\bar{\mathbf{y}}_m'\mathbf{H}_d\bar{\mathbf{y}}_m),$$

where $\bar{\mathbf{y}}_m = \frac{1}{m}\sum_{l=1}^{m}\mathbf{y}_l$ is the element-wise average of the m outcome vectors. This demonstrates that the BIC for any candidate model can be computed using only operations on $n \times n$ matrices. Future work could leverage this property to develop a computationally practical, non-greedy search algorithm.

Simulated example

To demonstrate, we conduct a simulation study across 100 independent trials. For each trial, we first generate a design matrix with $p = 20$ independent covariates, $X_{i,j} \sim \mathcal{N}(0, 1)$, for sample sizes $n \in \{10, 20, 50, 100\}$. We then generate the outcomes Y_i from a sparse linear model where only the first five covariates are active,

$$Y_i = X_{i,1} + X_{i,2} + X_{i,3} + X_{i,4} + X_{i,5} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, 1).$$

We expect the model uncertainty in variable selection to be high for $n = 10$ and to become more concentrated on the true five-variable model as n increases.

As a baseline for comparison, we apply Gibbs sampling (George & McCulloch, 1993) with binary inclusion indicators $\gamma_j \in \{0, 1\}$ for each covariate. The resulting Markov chain can iteratively explore different subsets of predictors by adding and removing covariates from the model specification. This approach is modeled after RJ-MCMC (Green, 1995), but avoids the computational complexity of modifying the Metropolis-Hastings acceptance probability to adjust for the change in "volume" of the parameter space when moving between models of different sizes.

We implement our Gibbs sampler with the `rjags` package in R (Plummer et al., 2023). For regression, we model the linear predictor term as

$$\mu_i = \beta_0 + \sum_{j=1}^{p} (\beta_j \cdot \gamma_j \cdot X_{i,j})$$

where the outcomes are then $y_i \sim \mathcal{N}(\mu_i, 1/\tau)$. The prior distributions are specified as $\beta_j \sim \mathcal{N}(0, 0.01)$ for the intercept term and all regression coefficients; $\gamma_j \sim \text{Bernoulli}(0.5)$ for the variable inclusion indicators; and $\tau \sim \text{Gamma}(0.01, 0.01)$ for the response precision. In the context of model selection, our target parameters of interest are the variable inclusion indicators γ_j that indicate whether each covariate is excluded or included.

In each trial, for three separate Gibbs sampling chains, we discard the first 5,000 iterations and retain the next 10,000 iterations. Figure 2.10 displays the mean posterior selection frequency for each covariate across the $B = 100$ trials, while Figure 2.11 displays the proportion of trials for which $\gamma_j > 0.5$, indicating the variable is more likely than not to be included. Barbieri and Berger (2004) refer to this as the "median probability model", and demonstrate that it often yields better predictive performance than the maximum a posteriori model. In both cases, we observe a similar pattern, with significant uncertainty across all variables for $n = 10$; roughly 50-50 identification of the correct variables for $n = 20$; and near-certainty for $n = 50$ and $n = 100$.

For the resampling approach, we apply the algorithm as previously described, using forward stepwise regression with BIC. This process is repeated for $M = 10$ block-sampling steps, yielding a final resampled dataset of size $n \times (M + 1)$, after which we record the final selected model. We then replicate this entire procedure across $B = 100$ trials to return uncertainty over the final model choice.

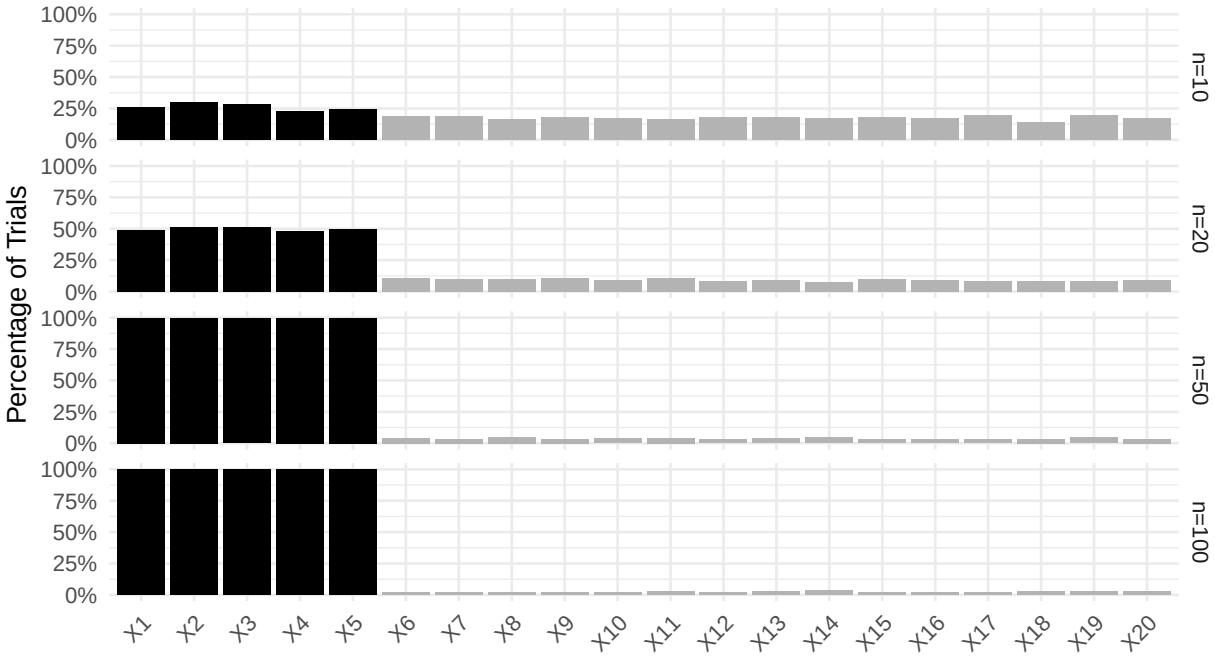


Figure 2.10: Mean posterior selection frequencies $\sum \gamma_j / B$ for Gibbs sampling as observed sample size increases, with correct variables in black.

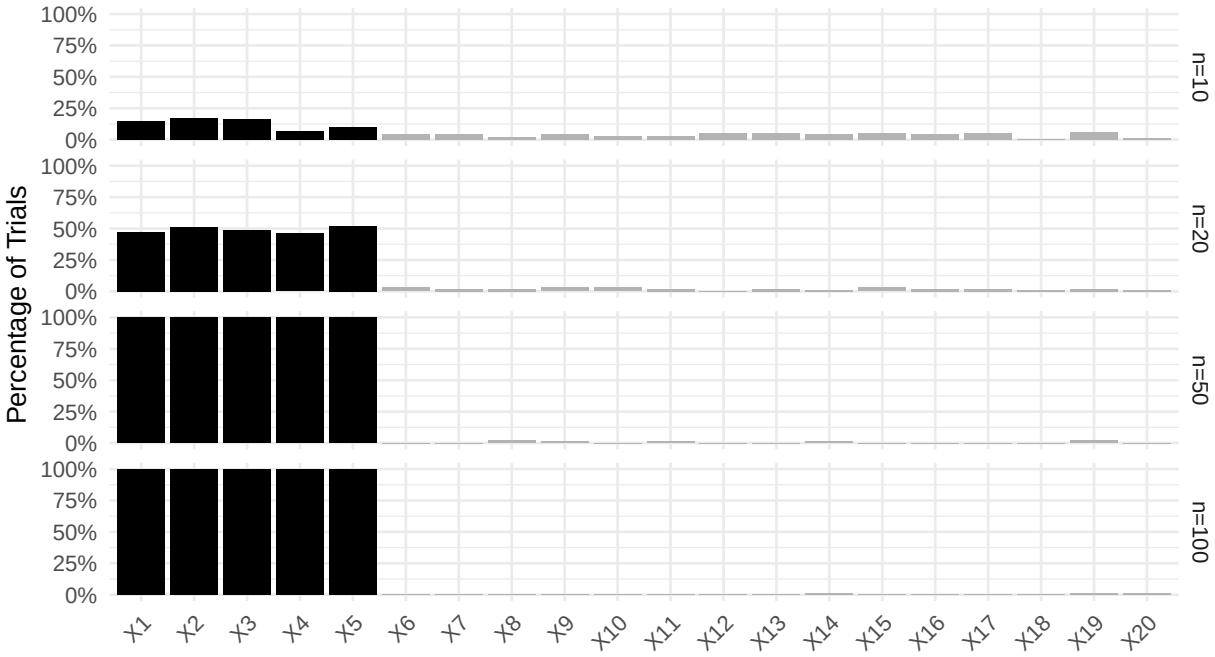


Figure 2.11: Proportion of trials with posterior selection frequency $\gamma_j > 0.5$ for Gibbs sampling as observed sample size increases, with correct variables in black.

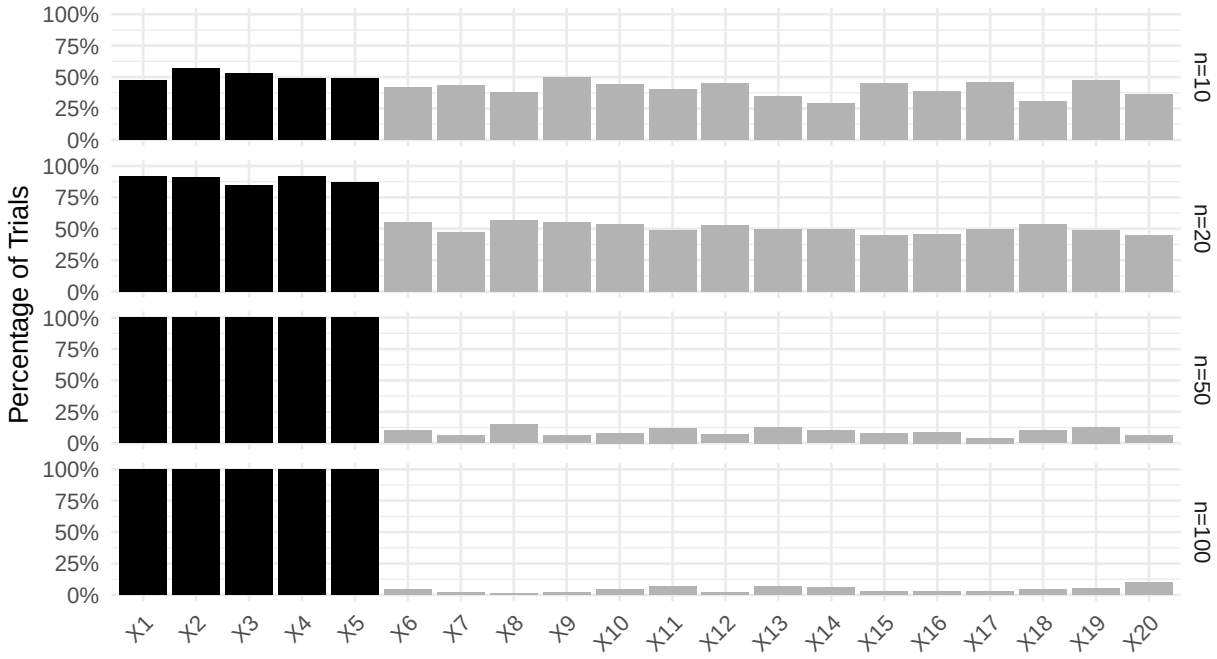


Figure 2.12: Proportion of trials with x_j in final model for forward stepwise regression with BIC as observed sample size increases, with correct variables in black.

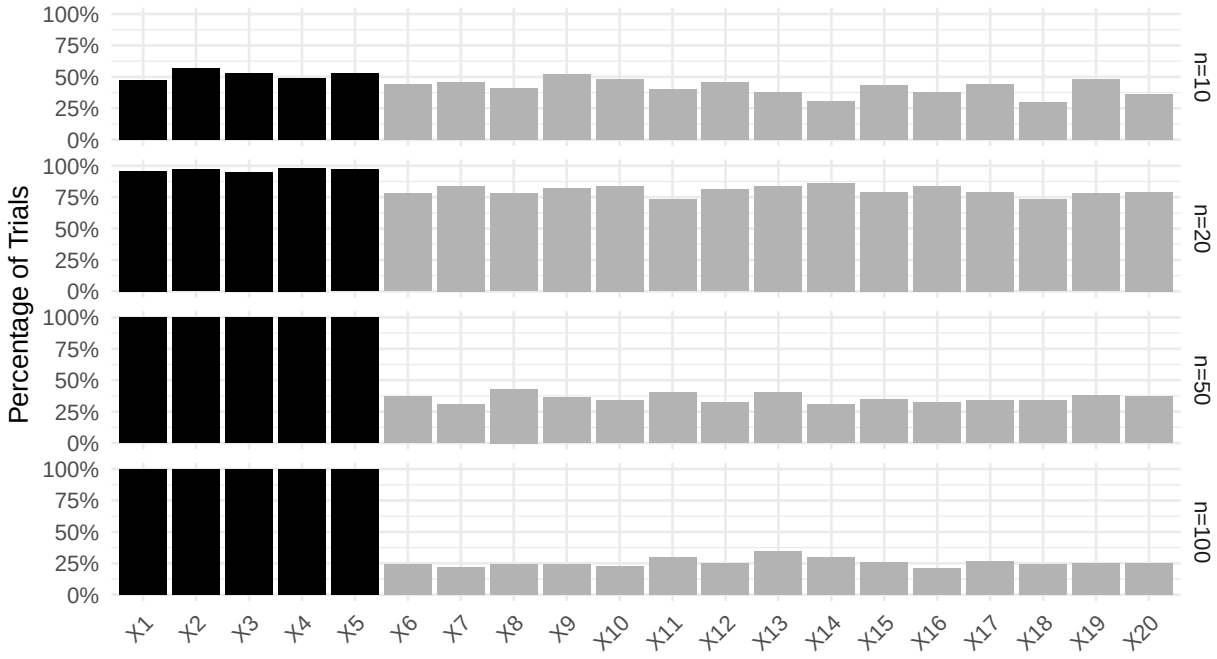


Figure 2.13: Proportion of trials with x_j in final model for forward stepwise regression with AIC as observed sample size increases, with correct variables in black.

Figure 2.12 displays the posterior selection frequency of each variable as a function of the observed sample size, using BIC as the optimization criterion. We observe similar patterns to the Gibbs sampling approach. In the $n = 10$ case, the procedure is highly uncertain and distributes weight across all available covariates, while in the $n = 20$ case, the true variables clearly receive more weight. In the $n = 50$ and $n = 100$ cases, variables x_1 to x_5 are correctly identified in almost all trials.

For comparison, Figure 2.13 displays the posterior selection frequencies with AIC as the optimization criterion. As discussed previously, AIC is asymptotically equivalent to LOO-CV, meaning it is also asymptotically inconsistent since it tends to select overly complex models (Shao, 1993). This pattern is reflected in the observed results, with the AIC including more variables on average across all sample sizes.

2.5 Conclusion

We view model uncertainty through the lens of missing information. With a complete data sequence in hand, we would be able to reliably identify the correct model, meaning that inferential uncertainty entirely arises from only observing a finite sample. Through predictive resampling (Fong et al., 2024), we can sequentially impute different possible realizations of the missing data. This allows us to convert a model selection criterion directly into a probability distribution over the space of candidate models.

Our approach serves as a form of model expansion around the initial best model for the observed data, as discussed by Draper (1995), and also echoes the prequential argument of Dawid (1984) with its focus on step-by-step prediction as the fundamental object of statistical modeling. The framework can be applied as a wrapper for any general method or package, as long as it allows for models to be compared and a new observation to then be sampled from the best model.

We acknowledge that the method has known failure cases, typically related to a high degree of model separation, where the data-generating distributions of competing models have little "overlap". For instance, consider the point hypothesis example of Section 2.2.3, but with $\theta_0 = -1$, $\theta_1 = 1$, and a small known variance such as $\sigma^2 = 0.1$. Suppose we

observe $n = 10$ with mean $\bar{y} = 0.001$. A deterministic criterion like the BIC will marginally prefer $\mathcal{M}_1$, but new samples will then come from $\mathcal{N}(1, 0.1)$, pulling subsequent model choices further toward $\mathcal{M}_1$. The algorithm will return $\widehat{\Pi}_{\text{pred}}(\mathcal{M}_1) \approx 1$, in contrast to a standard Bayesian approach, which would correctly return posterior probabilities near 0.5 for both models. This discrepancy is due to the "path-dependent" nature of Algorithm 1, which selects the single best model at each step. In cases with such poor model overlap, improvements could be made by incorporating a more stochastic selection mechanism, rather than a deterministic optimization at each step.

Computationally, we avoid the complexities of Bayes factor calculations and MCMC-style sampling methods, with the additional benefit of bypassing the need for subjective prior specification. The method, however, does introduce challenges related to Monte Carlo uncertainty quantification. Specifically, our estimator $\widehat{\Pi}_{\text{pred}}$ relies on indexing uncertainty over several resampling trials to obtain precise results, which becomes increasingly difficult as the model space grows. The ability to parallelize the independent resampling trials is a valuable asset that can alleviate some of these computational demands.

A complementary strategy is to implement early stopping, terminating a sampling trajectory once a particular model has clearly been selected. Fong and Yiu (2024a) show promising evidence that most uncertainty is captured within the initial stages of resampling, meaning that a limited number of iterations can suffice when combined with a suitable approximation. Recent theoretical developments frame this as a form of "martingale central limit theorem" (Fortini & Petrone, 2025), which could provide significant computational benefits in practice.

Another question that warrants attention is the requirement for models to be easily updated and for the consistent model selection criterion to be rapidly optimized at each step. While we have not specifically explored this aspect in the current work, it presents a potential area for future research. Efficient iteration of the model search process — for example, through an online search process that maintains a working model and uses a gradient-based update at each step — will be critical for scaling our method to more complex or higher-dimensional model spaces.

Chapter 3

Hypothesis testing via predictive resampling

3.1 Introduction

In this chapter, we conduct a deeper investigation into hypothesis testing, perhaps the most prevalent model uncertainty question in the statistical literature. We begin with a brief discussion of certain issues in both frequentist and Bayesian hypothesis testing, along with recent work on e-values. In particular, we focus on how the interpretation of these results can often be unclear or counter-intuitive. We then demonstrate how predictive resampling frames hypothesis testing as a decision problem on a population statistic, propagating uncertainty through the missing data to directly quantify the probability of competing hypotheses, without requiring the specification of a prior distribution.

3.2 Review of existing approaches

3.2.1 Frequentist testing

Suppose we have an unknown parameter θ and wish to test

$$H_0 : \theta \in \Theta_0$$

$$H_1 : \theta \in \Theta_1.$$

Generally, we would define our test such that $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 = \Theta$. A classical (frequentist) procedure constructs a critical region C such that the null hypothesis H_0 is rejected if the observed sample $Y_{1:n}$ falls within C and not rejected if $Y \notin C$ (Berger, 2003).

Typically, this is done by calculating a test statistic $T = t(Y)$ as a function of the observed sample, then specifying the critical region as $C = \{y \mid t(y) > t_0\}$. The critical value t_0 is set using a pre-specified significance level of $1 - \alpha$ such that $\alpha = P_0(T > t_0)$ is the selected Type I error rate, or the probability of erroneously rejecting the null hypothesis when it is actually true.

A key aspect of this approach is that it focuses solely on the observed $Y_{1:n}$ to calculate the test statistic, which is then treated as a random variable. Within this construction, the p-value $p = P_0(T > t(y))$ can be understood as the probability under the null hypothesis of obtaining a test statistic which is "at least as extreme" as the observed value. This probability is interpreted with respect to the variation across all possible samples of size n in the population from which Y is drawn. For a specific observed sample, having $t(y) > t_0$ is then equivalent to having $p < \alpha$, and so hypothesis tests are often discussed in terms of whether the p-value falls below the chosen significance threshold.

However, a well-known issue with rejecting the null hypothesis based on $p < \alpha$ at sample size n is that this conflates the effect size with the sample size (Gelman & Stern, 2006; Nickerson, 2000). For example, suppose we are testing

$$\begin{aligned} H_0 &: \theta = 0 \\ H_1 &: \theta \neq 0 \end{aligned} \tag{3.1}$$

for the unknown mean parameter of a normal distribution with known variance $\sigma^2 = 1$. For an observed sample $y_{1:n}$, the effect size is the absolute mean difference from the null hypothesis value of zero, or $|\bar{y}|$ exactly.

The test rejects H_0 if the effect size exceeds the critical value $t_0 = z_{\alpha/2} \cdot (\sigma/\sqrt{n})$. As the sample size n increases, this rejection boundary shrinks toward zero. Consequently, a vanishingly small effect size becomes sufficient to achieve $p < \alpha$, even when that difference is far below any threshold of practical relevance. Figure 3.1 demonstrates this relationship: as n grows, the magnitude of the effect size required to reject H_0 shrinks. This leads to a well-known paradox: for any fixed alternative $\theta_A \neq 0$, no matter how practically insignificant (e.g., $\theta_A = 0.0001$), the power of the test to reject H_0 will approach 1 as $n \to \infty$. Jeffreys (1961) famously summarizes this result, emphasizing why p-value

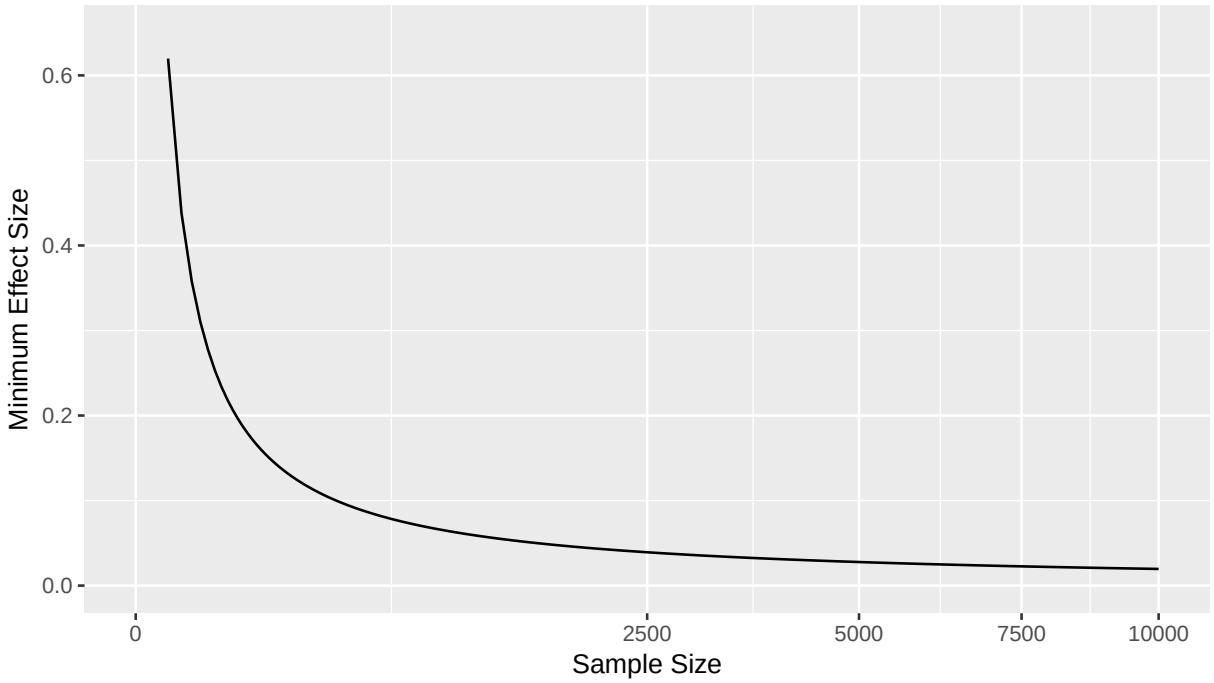


Figure 3.1: Minimum effect size $\bar{y} - \theta$ sufficient with sample size n to reach "significant" conclusion of $p < 0.05$ for simple one-sample z-test.

computations based on tail area are illogical: "...a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred."

3.2.2 Bayesian testing

Meanwhile, Bayesian hypothesis testing faces challenges related to the precise specification of prior distributions. This occurs at two levels: the prior probabilities on the models themselves (π_0 and π_1), and the parameter prior distributions for the parameters (θ) within each model. A famous illustration is the Jeffreys-Lindley paradox (Jeffreys, 1961; Lindley, 1957), which arises when testing a point null hypothesis, such as

$$H_0 : \theta = 0.5$$

$$H_1 : \theta \neq 0.5$$

for a binomial proportion θ. With equal prior model probabilities ($\pi_0 = \pi_1 = 0.5$) and a diffuse parameter prior distribution for θ under H_1, such as $\theta \sim \text{Uniform}(0, 1)$, the Bayesian approach may strongly prefer H_0 even when a frequentist p-value strongly rejects it. Intuitively, this occurs because the alternative model spreads its prior mass thinly across $\theta \in [0, 1]$, diminishing its marginal likelihood regardless of the actual data.

The unfortunate consequence is that a minimally informative prior distribution on the parameter, specified in an apparently objective manner, becomes maximally informative for model selection, in favor of the null.

However, Johnson and Rossell (2010) highlight a contrasting issue that occurs when the alternative prior distribution $\pi_1(\theta)$ assigns significant mass "locally" (i.e., very near the null value θ_0). In this case, the null and alternative hypotheses are not well-separated, meaning that evidence is accumulated much more rapidly to support true alternative models than to support true null models. As a result, it can become asymmetrically difficult to accumulate strong evidence for H_0 even when it is true.

3.2.3 Recent developments in e-values

In recent years, an alternative approach for hypothesis testing has emerged under the name of e-values (Grünwald et al., 2024; Vovk & Wang, 2021). As the e-value framework increases in popularity, we include it in our discussion to provide a comparison of predictive resampling with this emerging approach. We highlight that the e-value provides several concrete benefits but still faces challenges in interpretation.

An e-value is a non-negative random variable E (which may depend on the sample y) whose expected value under the null hypothesis H_0 is at most one. In other words, it is defined by the property that $\mathbb{E}_0[E] \leq 1$. This property provides two considerable benefits for accumulating evidence against H_0. First, the product of e-values from independent experiments is also an e-value (i.e., if E_1 and E_2 are e-values, then $\mathbb{E}_0[E_1 E_2] \leq 1$). Second, an e-value can be updated as data accumulates, $y_1, y_2, \ldots$, to form a test supermartingale. This is a sequence of non-negative random variables $E_n = E(y_{1:n})$ where the expected value, given the past, does not increase under H_0. Together, this allows for evidence to be continuously monitored and accumulated while maintaining strict control of the Type I error rate, avoiding the "p-hacking" issues of sequential p-value analysis.

The e-value is closely linked to the likelihood ratio, and many e-values are identical to Bayes factors with particular prior distributions. This provides a direct frequentist justification (Type I error control) for a Bayesian object, bridging a significant gap between

the two paradigms. Shafer (2021) discusses a very similar concept using the terminology of "betting scores", arguing that an e-value is most naturally motivated and understood as the monetary winnings from making a bet against the null hypothesis.

Despite these powerful properties, a challenge remains in the direct interpretation of an e-value's magnitude. This contrasts with classical test statistics, which are often designed to diverge. For example, a classical statistic T_n for the normal mean problem might be constructed to diverge in opposite directions: $T_n \to -\infty$ under H_0 and $T_n \to \infty$ under H_1. This allows for a fixed critical value (such as zero) to separate the hypotheses. One such test statistic (which amounts to a modified version of the BIC) is

$$T_n = \log(\bar{y}_n^2 \sqrt{n}) = \begin{cases} \frac{1}{2}\log n + \log(\bar{y}_n^2) & H_1 \colon \theta \neq 0 \\ -\frac{1}{2}\log n + \log(n\bar{y}_n^2) & H_0 \colon \theta = 0. \end{cases}$$

Under H_1, $\bar{y}_n^2$ converges to a non-zero finite constant, so T_n goes to ∞ at rate $\frac{1}{2}\log n$. Under H_0, the term $n\bar{y}_n^2$ (assuming $\sigma^2 = 1$) converges in distribution to a χ_1^2 random variable. Its logarithm is therefore a finite random variable, so T_n is dominated by the $-\frac{1}{2}\log n$ term and goes to $-\infty$. Since these convergences are at the same rate, zero is a suitable critical value.

An e-value, by contrast, does not diverge to $-\infty$ under the null; it must remain non-negative, and converges to a finite random variable (with an expected value less than 1). This provides the strong Type I error guarantee, but it means that the practical strength of evidence for an e-value with a given magnitude (say, $E_n = 5$) is unclear. Lacking an intuitive scale, Vovk and Wang (2021) resort to the heuristic "rule of thumb" originally proposed by Jeffreys (1961) for Bayes factors (values between 1 and $\sqrt{10}$ are "not worth more than a bare mention", between $\sqrt{10}$ and 10 are "substantial", etc.)

3.3 Illustrations

Meanwhile, predictive resampling provides a path to a direct probability for each hypothesis, under a different set of interpretations and assumptions. Unlike the frequentist p-value (a tail-area probability) or the e-value (a betting score), resampling targets a genuine probability $\Pi_{\text{pred}}(\mathcal{H}_k \mid \mathcal{D})$, but unlike the Bayesian approach, this probability is derived without specifying subjective prior distributions over parameters or models.

Having observed data $y_{1:n}$, we focus on the missing $Y_{n+1:\infty}$ as the source of uncertainty, where the decision rule for selecting a hypothesis is based on some summary statistic of the complete population. Applying Algorithm 1, we compare the null and alternative hypotheses at each step using a consistent model selection criterion (such as BIC). We then impute a new observation from the currently preferred hypothesis, including any parameter MLE(s) as needed. By indexing the final model choice over many replications of this imputation process, we approximate the martingale posterior probability of each hypothesis. Code to replicate all experiments is available at `https://github.com/vshirvaikar/MPModel`.

3.3.1 Two-sided testing

To demonstrate, consider the null hypothesis test from Equation 3.1, and suppose we observe $y_{1:n}$ with sample mean $\bar{y}$.

Methodology

The classical p-value is computed using a two-sided z-test, with test statistic

$$Z = \frac{\bar{y}}{1/\sqrt{n}} = \sqrt{n}\,\bar{y},$$

and corresponding p-value $p = 2\,\Phi(-|Z|)$, where Φ denotes the cumulative distribution function of the standard normal distribution.

Following Vovk and Wang (2021), we calculate the e-value as the likelihood ratio between the null and the most favorable point alternative (at $\bar{y}$), resulting in

$$E = \frac{\prod_{i=1}^{n} e^{-(y_i-\bar{y})^2/2}}{\prod_{i=1}^{n} e^{-y_i^2/2}}.$$

After simplifying, this reduces to

$$E = \exp\left(n\bar{y}^2 - \frac{1}{2}\sum_{i=1}^{n}(y_i-\bar{y})^2 + \frac{1}{2}\sum_{i=1}^{n} y_i^2 - \frac{n\bar{y}^2}{2}\right) = \exp\left(\frac{n\bar{y}^2}{2}\right).$$

As mentioned previously, this has the appealing property of closure under multiplication, where e-values from independent samples can be easily combined.

As discussed in Section 2.2.1, a common approach for approximating posterior model probabilities is to compare BIC values, since $\exp(-\frac{1}{2}\text{BIC})$ serves as a large-sample approximation to the marginal likelihood. We also include this approach for comparison.

The null model $H_0 : \theta = 0$ is fully determined and has no free parameters ($d = 0$) in the evaluation of Equation 2.2. The alternative model $H_1 : \theta \neq 0$ has a penalty term of $d = 1$, with one free parameter θ which is estimated by the sample mean $\bar{y}$.

Under this approach, the posterior probability of the null model is given by

$$p(H_0 \mid y_{1:n}) = \frac{\exp(-\frac{1}{2}\mathrm{BIC}_0)}{\exp(-\frac{1}{2}\mathrm{BIC}_0) + \exp(-\frac{1}{2}\mathrm{BIC}_1)}.$$

Substituting

$$\exp\left(-\frac{1}{2}\mathrm{BIC}_k\right) = \widehat{\mathcal{L}}_k \cdot n^{-d_k/2},$$

with the same maximized likelihoods under each model, we get

$$p(H_0 \mid y_{1:n}) = \frac{\prod_{i=1}^n e^{-y_i^2/2}}{\prod_{i=1}^n e^{-y_i^2/2} + \prod_{i=1}^n e^{-(y_i-\bar{y})^2/2} \cdot n^{-1/2}}.$$

Dividing by $\prod_{i=1}^n e^{-y_i^2/2}$ and simplifying gives us

$$p(H_0 \mid y_{1:n}) = \frac{1}{1 + \exp\left(\frac{n\bar{y}_n^2}{2}\right) \cdot n^{-1/2}} = \frac{1}{1 + E \cdot n^{-1/2}},$$

where E is the e-value from the likelihood ratio calculation above. This shows that the BIC approximation to the posterior probability of H_0 is closely related to a scaled inverse of the e-value; in fact, Vovk and Wang (2021) suggest using the simpler "e-to-p calibrator" function $p = \min(1, 1/e)$ to transform e-values into the $[0, 1]$ window.

Finally, for the predictive resampling approach, we directly compare the BIC given the observed data at each step, which reduces to rejecting H_0 if $n\bar{y}_n^2 > \log n$, and vice versa. We therefore sample y_m from $\mathcal{N}(0, 1)$ if $(m-1)\bar{y}_{m-1}^2 < \log(m-1)$ and from $\mathcal{N}(\bar{y}_{m-1}, 1)$ otherwise. The resulting sequence

$$h_{m+1} = 1(m\bar{y}_m^2 > \log m), \quad m > n$$

converges to 0 or 1. We repeat this up to the total sample size of $N = n + 20n$ across several Monte Carlo iterations (in this case, 1000) to index our uncertainty between the two hypotheses. In the supplementary material, we provide convergence diagrams showing that the specified value of N is sufficiently large for the choice of model to converge in this particular setting. While this determination is currently based on visual inspection, a key area for future work is the development of a formal convergence diagnostic.

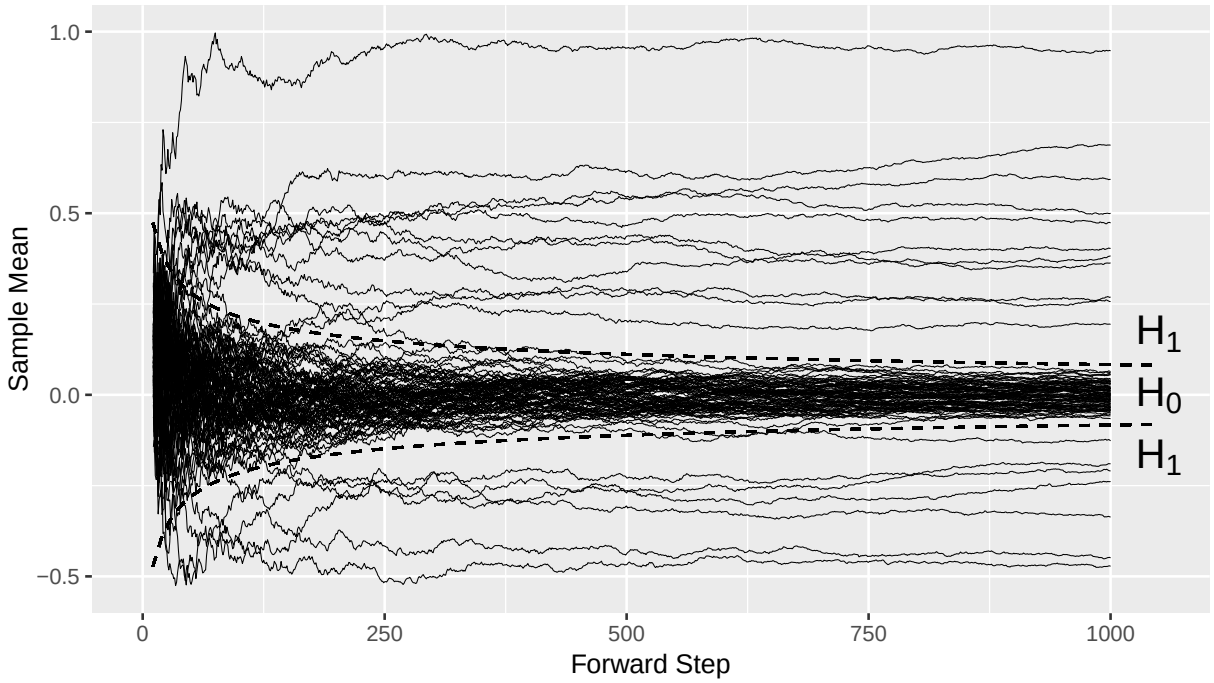


Figure 3.2: Trajectory diagram showing uncertainty propagation via resampling for two-sided hypothesis testing. The dotted lines indicate the BIC-based decision boundary $\bar{y}_m = \sqrt{(\log m)/m}$, which contracts toward zero as $n \to \infty$, reflecting the increasing penalty for model complexity.

Figure 3.2 illustrates the resampling procedure for a simple example with $n = 10$ observed data points and a total sample size of $N = 1000$. The initial data are drawn from $\mathcal{N}(0, 1)$ with an observed mean of $\bar{y}_n = 0.132$, meaning that H_0 is initially preferred. As the remaining $N - n$ observations are imputed, some sample paths cross over into the H_1 critical region. The curved funnel defined by the dotted lines represents the decision boundary at $\bar{y}_m = \sqrt{(\log m)/m}$, reflecting the BIC penalty for model complexity. As n increases, this boundary narrows around zero, requiring stronger evidence to justify selection of the more complex model. We ultimately retrieve a Monte Carlo posterior probability of $\widehat{\Pi}_{\text{pred}}(H_0 \mid y_{1:n}) = 0.846$ for this particular example.

Results

To compare the approaches, we first generate data under the null, i.e., from $\mathcal{N}(0, 1)$, for sample sizes $n = \{30, 100, 300, 1000\}$ across 400 random seeds. As expected, the classical p-values under the null are uniformly distributed across $[0, 1]$ regardless of the observed sample size. Since the p-values cannot be directly interpreted as probabilities, we interpret

True model	Summary metric	Sample size			
		30	100	300	1,000
H_0 ($\mu = 0$)	Prop. of tests with $p < 0.05$ (Type I error)	6%	4%	4%	6%
	Prop. of tests with $e > 10$ (Type I error)	4%	3%	3%	5%
	Average resampling posterior prob. of H_1	0.14	0.08	0.05	0.04
	Prop. of tests with $\widehat{\Pi}_{\text{pred}}(H_1 \mid \mathcal{D}) > 0.05$	80%	33%	17%	11%
	Prop. of tests with $\widehat{\Pi}_{\text{pred}}(H_1 \mid \mathcal{D}) > 0.1$	33%	15%	8%	7%
H_1 ($\mu = 0.1$)	Prop. of tests with $p < 0.05$ (Power)	9%	17%	41%	89%
	Prop. of tests with $e > 10$ (Power)	5%	13%	34%	85%
	Average resampling posterior prob. of H_1	0.17	0.18	0.29	0.69
	Prop. of tests with $\widehat{\Pi}_{\text{pred}}(H_1 \mid \mathcal{D}) > 0.5$	11%	13%	25%	71%
	Prop. of tests with $\widehat{\Pi}_{\text{pred}}(H_1 \mid \mathcal{D}) > 0.9$	4%	5%	11%	50%

Table 3.1: Metrics for two-sided hypothesis testing across 400 random seeds

them as a binary decision using a pre-specified Type I error rate, usually $\alpha = 0.05$. In the first row of Table 3.1, we see that the proportion of tests with $p < 0.05$ is approximately 5% for all values of n. For the e-values, as recommended by Vovk and Wang (2021), we apply Jeffreys' rule of thumb, under which $e > 10$ indicates that the evidence against the null hypothesis is "strong". The proportion of tests with $e > 10$ also remains at approximately the same level for all values of n.

In contrast, predictive resampling allows us to accumulate evidence *in favor of the null* as the observed sample size grows. In the third row of Table 3.1, we see that the average posterior probability of H_1 returned by resampling decreases with n. This reflects the underlying principle that missing data is the source of statistical uncertainty, and that observing additional data consistent with a given model (in this case, H_0) should result in a greater degree of probabilistic certainty about that model.

If desired, we can also apply a decision rule, and record the proportion of trials for which the posterior probability of H_1 exceeds a certain level. In the fourth and fifth rows of Table 3.1, we see that under the null hypothesis, the proportion of tests with $\widehat{\Pi}_{\text{pred}}(H_1 \mid \mathcal{D})$ above certain thresholds (in this case, 0.05 and 0.1) decreases with n. Conversely, this means that increasing the sample size results in a greater proportion of tests with at least

0.95 and 0.9 posterior probability on H_0 respectively.

As a different mode of visualization, in Figure 3.3 we plot the classical p-values for 100 of the 400 seeds on the horizontal axis, and the resampling posterior probabilities of H_0 on the vertical axis. The X marks on the plots indicate tests with $p < 0.05$ where classical testing would reject H_0, and the O marks indicate $p > 0.05$ for which classical testing would fail to reject H_0. We again see that the p-values are invariant to sample size, but that the overall level of the resampling probabilities increases towards 1 as the sample size grows. In the supplementary material, we provide a corresponding plot comparing e-values with predictive resampling; the interpretation is largely similar.

Alternatively, in the bottom half of Table 3.1, we generate the same set of sample sizes from $\mathcal{N}(0.1, 1)$ under H_1. All three methods successfully accumulate evidence against the null as n increases; this can be seen visually in Figure 3.4, where the points gradually migrate towards the bottom and left as we observe more data. (The corresponding plot for e-values is in the supplementary material.)

In this context, since H_1 is the "true" underlying model, we would interpret the p-value and e-value results in terms of statistical power. As expected, both increase in power as n grows, based on the proportion of tests with $p < 0.05$ or $e > 10$ respectively. However, this calculation is only possible because we are simulating several datasets. The concept of power does not directly translate to a single trial, where we only get one chance to observe the outcome; it is inherently probabilistic, estimating the likelihood that a study will detect a certain effect across many repetitions of the same trial. This can be observed in applied statistical practice, where power calculations generally have to resort to simulation.

In contrast, the average posterior probability of H_1 returned by predictive resampling also grows as n increases, but we can directly interpret the results for a single trial, rather than having to interpret our outcomes in the context of long-run Type I and Type II error rates. However, as before, we can also apply a decision rule if desired, and consider the proportion of individual trials for which $\widehat{\Pi}_{\text{pred}}(H_1 \mid \mathcal{D})$ exceeds a given level. To "accept" the alternative hypothesis, we might specify a higher threshold (such as 0.5 or 0.9); at the bottom of Table 3.1, we see that these proportions increase with n.

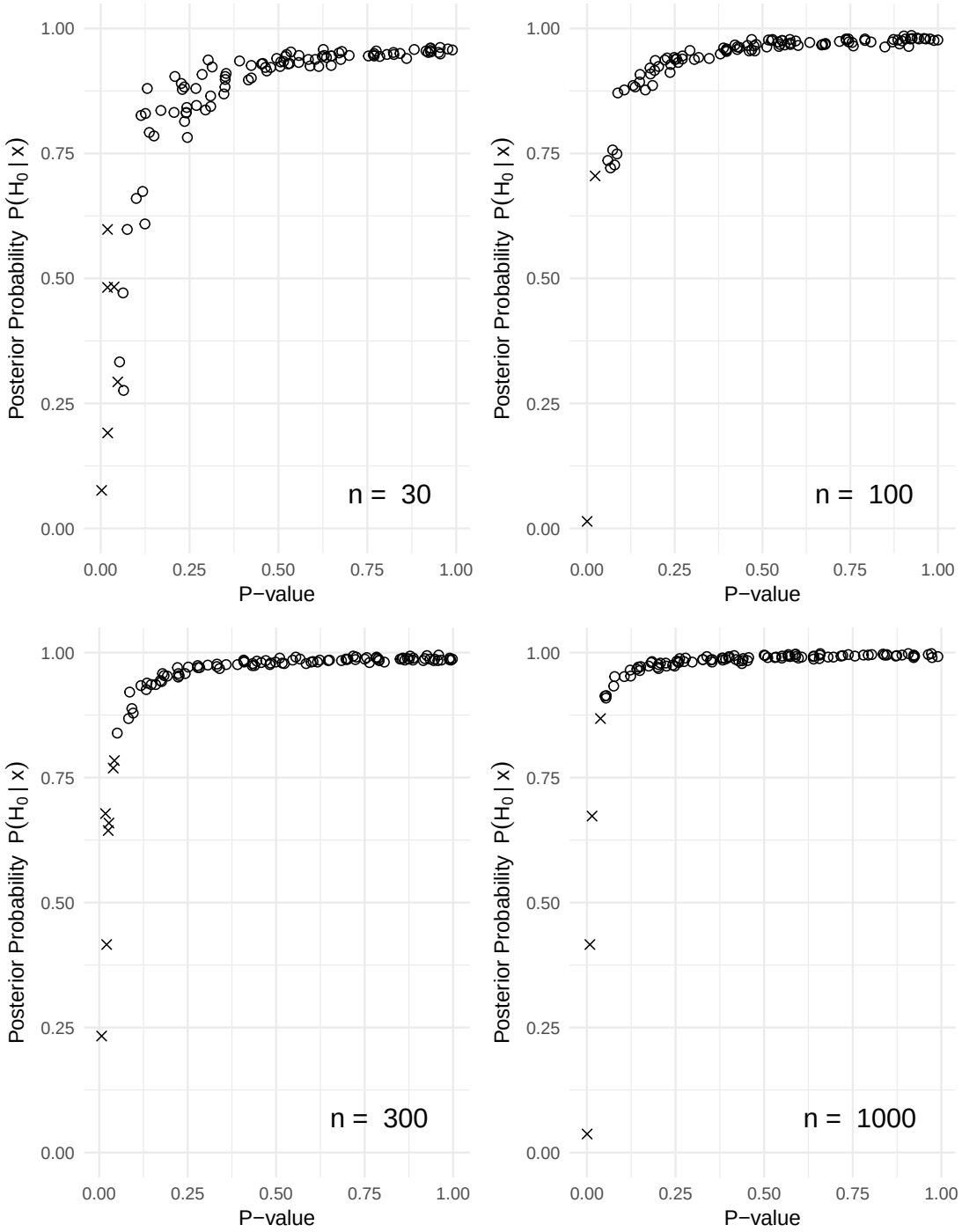


Figure 3.3: Observed p-value vs. resampling posterior probability of H_0 for data generated under the null $\mathcal{N}(0,1)$ across 100 out of 400 random seeds. X denotes tests with $p < 0.05$ where classical testing rejects H_0, and O denotes tests with $p > 0.05$.

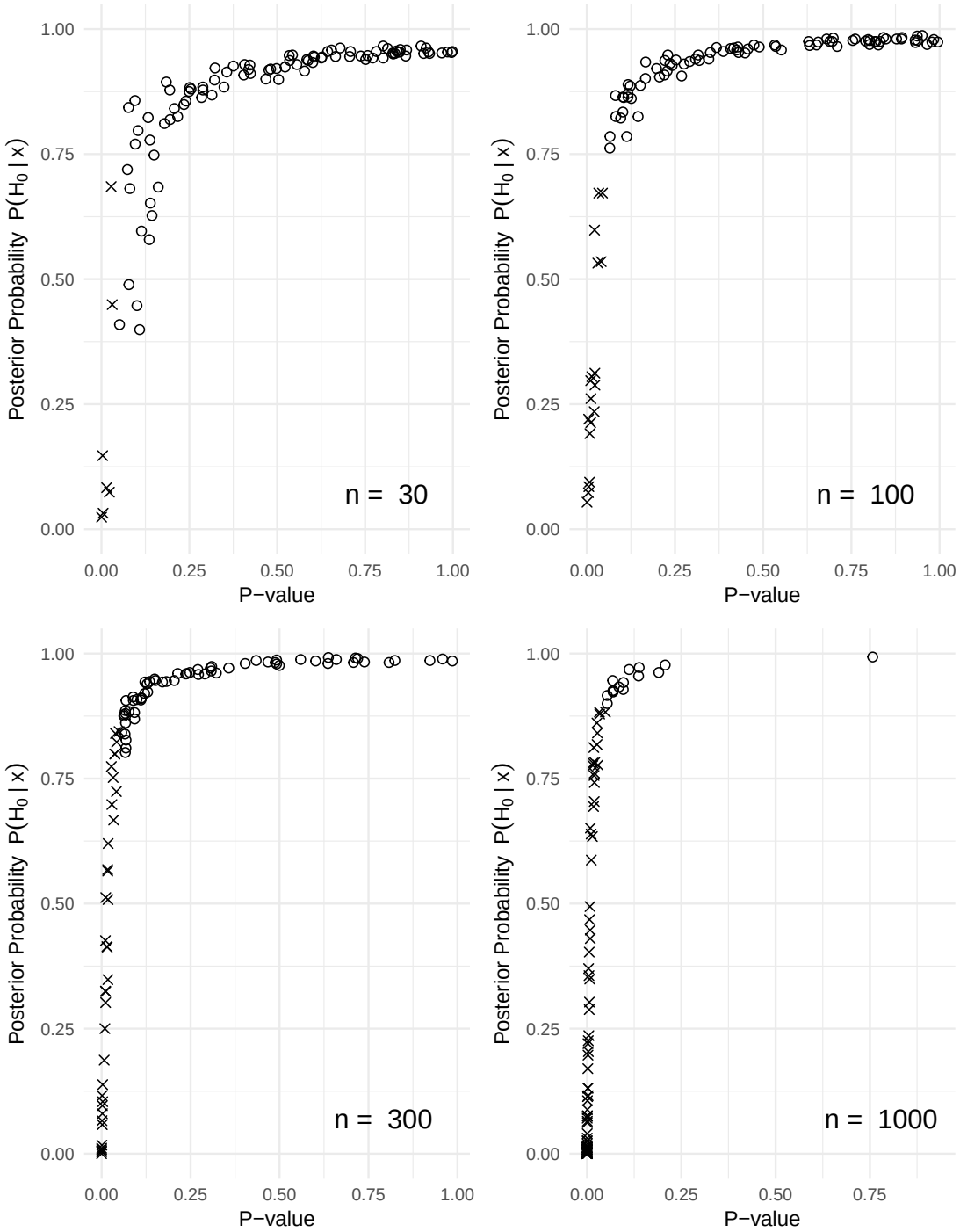


Figure 3.4: Observed p-value vs. resampling posterior probability of H_0 for data generated under the alternative $\mathcal{N}(0.1, 1)$ across 100 out of 400 random seeds. X denotes tests with $p < 0.05$ where classical testing rejects H_0, and O denotes tests with $p > 0.05$.

As an additional point of interest, for the experiment with data generated under the alternative hypothesis, we compare the results of our predictive resampling procedure with those obtained from the BIC-based approximation, $p(H_0 \mid y_{1:n}) = 1/(1 + E \cdot n^{-1/2})$. As seen in Figure 3.5, the two methods produce relatively similar posterior probabilities, with predictive resampling closely tracing the smooth (logistic-shaped) curve implied by the BIC function. This may imply that predictive resampling is unnecessarily elaborate for a simple problem like point-null hypothesis testing, but the close agreement also offers encouraging validation. Despite its greater generality and flexibility, the predictive approach encodes and propagates uncertainty in a way that aligns with the penalized likelihood structure of BIC, while directly targeting interpretable model probabilities.

3.3.2 Multi-level testing

For a more detailed demonstration, we consider the well-known example of the "hot hand" in basketball — i.e., whether players are subject to swings in "momentum" that affect their accuracy (Gilovich et al., 1985; Hsiao et al., 2005; Kass & Raftery, 1995).

For a particular player, we observe data (n_i, k_i) for $i = 1, 2, \ldots, g$ games, where n is the number of shots attempted and k is the number scored. If the player's true underlying shooting percentage is stable, then the variation across individual games should be explained by binomial distributions with the same success probability. If not, the game-level shooting percentages should be sufficiently independent from each other. A simple hypothesis test that captures this difference is

$$H_0 : k_i \sim \text{Binomial}(n_i, p) \text{ with a common percentage } p$$
$$H_1 : k_i \sim \text{Binomial}(n_i, p_i) \text{ with independent } p_i \sim B(a, b)$$

where the null model is binomial for all games and the alternative model is beta-binomial.

The null model likelihood is

$$\mathcal{L}_0(1, 1 \mid \{n_i, k_i\}_{i=1}^g) = \prod_{i=1}^{g} \binom{n_i}{k_i} \frac{B(k_i + 1, n_i - k_i + 1)}{B(1, 1)}$$

with a uniform prior on the beta distribution.

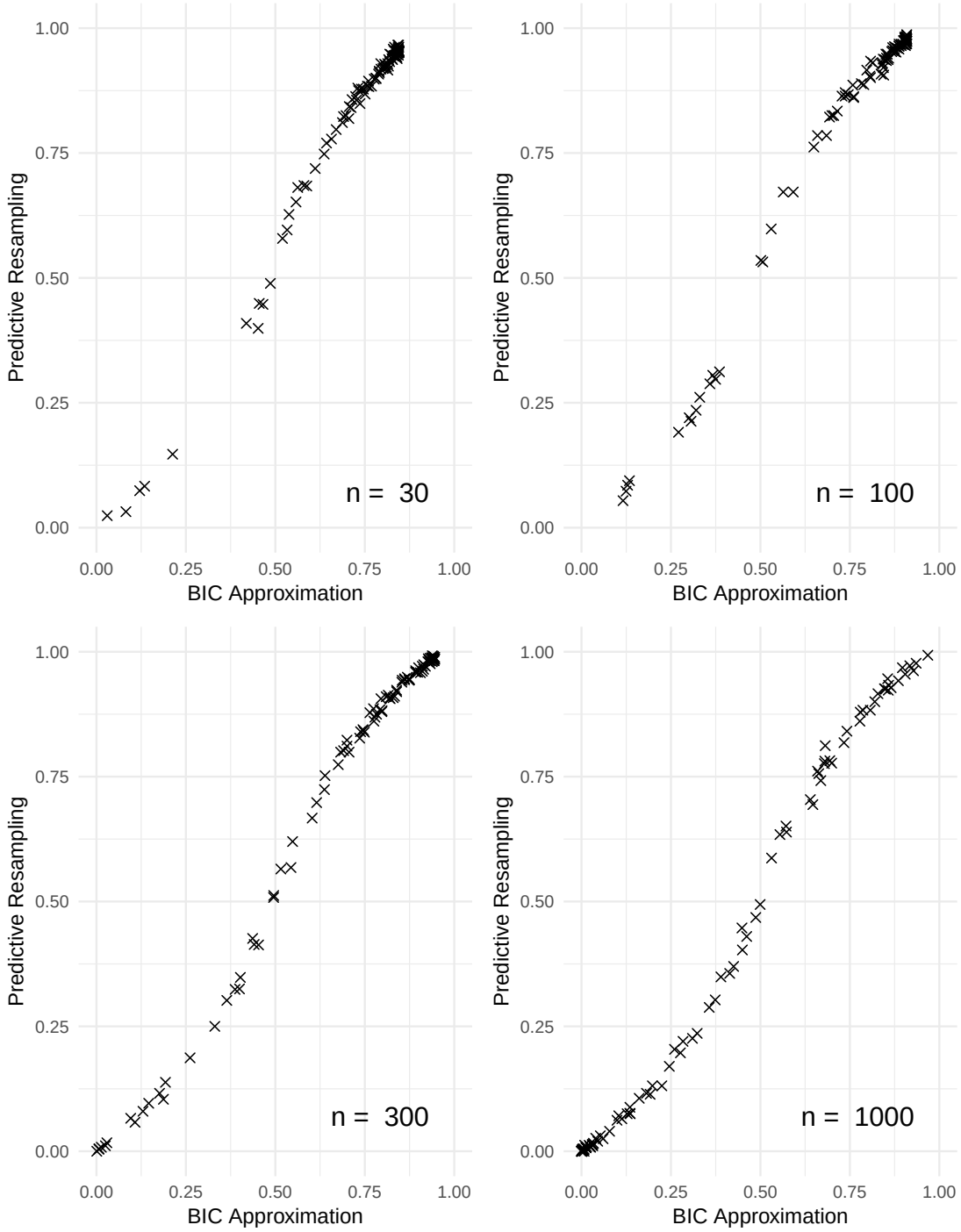


Figure 3.5: Approximate $\exp(-\frac{1}{2}\text{BIC})$ probability vs. resampling posterior probability of H_0 for data generated under the alternative $\mathcal{N}(0.1, 1)$ across 100 out of 400 random seeds.

For the alternative model likelihood, we have

$$\mathcal{L}_1(a, b \mid \{n_i, k_i\}_{i=1}^g) = \prod_{i=1}^{g} \binom{n_i}{k_i} \frac{B(k_i + a, n_i - k_i + b)}{B(a, b)}$$

where the best values of $\hat{a}$ and $\hat{b}$ can be fitted with numerical optimization.

We maximize the log-likelihood for each model and compare the BIC. The dimension penalty is one for the null model (the choice of prior distribution) and two for the alternative model (a, b), while the sample size is the number of games. For the preferred model, we simulate a new game by drawing n_{g+1} from $(n_1, \ldots, n_g)$ and then $k_{g+1} \sim \text{BetaBin}(n_{g+1}, 1, 1)$ in the null case, or $k_{g+1} \sim \text{BetaBin}(n_{g+1}, \hat{a}, \hat{b})$ in the alternative case.

To demonstrate, we generate data for $g = 20$ games with $k_i \sim B(n_i, p_i)$, where $n_i \sim \text{Unif}[5, 15]$ and $p_i \sim B(\alpha, \alpha)$. We vary the value of $\alpha \in \{0.5, 1, 1.5, 2, 2.5\}$ and resample up to $G = 200$ games across 100 trials. Figure 3.6 shows the acceptance proportion of H_0 with respect to the α used for data generation. As expected, we accept H_0 in most cases when the initial data are generated under $\text{BetaBin}(n_i, 1, 1)$, corresponding to the original null hypothesis, and vice versa.

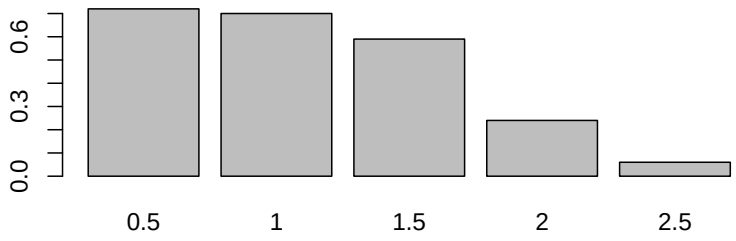


Figure 3.6: Null hypothesis acceptance probability vs. α parameter used to sample data from beta-binomial distribution. H_0 is accepted in 70% of trials when the data are generated under $B(1, 1)$, decreasing to 6% of trials under $B(2.5, 2.5)$.

Chapter 4

Bayesian prediction without parameter uncertainty

4.1 Introduction

The goal of prediction is to quantify uncertainty over future observables, not to recover posterior distributions over latent parameters. However, even when prediction is the primary objective, Bayesian methods typically approach uncertainty entirely through the lens of parameters — requiring that we specify a prior distribution, update it via a likelihood, and marginalize over the resulting posterior weights before making statements about observables. This can be a cumbersome and misdirected detour, particularly in Bayesian deep learning, where prior distributions lack clear interpretation and posterior updates are generally intractable.

In this chapter, building upon recent developments in Bayesian predictive inference (Fong et al., 2024; Fortini & Petrone, 2025), we argue that a model is fundamentally defined by how it sequentially resolves predictive uncertainty. In particular, a Bayesian model is one whose future predictions form a c.i.d. sequence (Berti et al., 2004). Under this view, if our goal is to retrieve Bayesian predictive uncertainty, a prior-likelihood update is no longer strictly necessary — we can adopt a plug-in approach, where we specify and directly optimize a flexible predictive distribution, provided that the resulting model produces c.i.d. observations.

This shifts the core task of Bayesian learning: rather than computing a parametric posterior distribution, often via imprecise approximation or sampling, we simply need to

assess whether a model's predictions satisfy the c.i.d. property. To this end, we introduce a new diagnostic based on predictive consistency under one-step sequential imputation, and demonstrate its application in a supervised learning setting. We additionally build on the related work of Mlodozeniec et al. (2024), who test whether a one-dimensional predictive rule is "implicitly Bayesian" by evaluating the exchangeability of new predictions with observed data, as quantified by the variance of the log-joint probability. In our demonstration, this idea is extended to supervised learning and implemented as an alternative diagnostic.

Taken together, our findings suggest a complementary perspective for practitioners and researchers interested in Bayesian uncertainty quantification. The traditional prior-posterior workflow, while foundational, is not the only path to producing coherent predictive distributions. We suggest that there is also value in designing and optimizing flexible models with the explicit goal of producing coherent, c.i.d. predictions. This reframing introduces its own set of modeling choices and diagnostic trade-offs, but allows a broader array of procedures, including plug-in or frequentist methods, to inherit the key virtues of Bayesian reasoning. By focusing on predictive coherence as a central criterion, we open the door to new modeling strategies that are both practically effective and theoretically grounded.

The remainder of this chapter is organized as follows. Section 4.2 provides a brief background on related work in generalized Bayesian inference and model uncertainty. In Section 4.3, we apply the predictive representation of Bayesian inference to argue that Bayesian uncertainty only requires the construction of a c.i.d. predictive model. Section 4.4 introduces novel metrics to evaluate predictive coherence, based on consistent prediction under sequential imputation and the log-joint variance method of Mlodozeniec et al. (2024). In Section 4.5, we apply these metrics to supervised learning examples, demonstrating that plug-in models can produce coherent Bayesian conditional predictions. Section 4.6 provides some concluding remarks. Code to replicate all experiments is available at `https://github.com/vshirvaikar/MPPrediction`.

4.2 Related work

In this section, we provide a brief overview of relevant literature on non-standard Bayesian updates and model uncertainty, particularly in the context of deep learning.

4.2.1 Generalized Bayesian inference

The standard Bayesian pipeline provides a number of desirable properties, such as coherence, consistency, and asymptotic optimality (Bernardo & Smith, 2004). However, these guarantees typically depend on a set of critical assumptions. Knoblauch et al. (2022) highlight three such assumptions that are often fragile in practice: model specification, prior specification, and computational feasibility.

First, Bayesian inference typically assumes an $\mathcal{M}$-closed setting (Bernardo & Smith, 2004), where the true data-generating process lies within the specified model class — in other words, there exists some value of the random parameter $\theta \in \Theta$ for which the likelihood accurately describes the data. Second, this parameter requires an associated prior distribution $\pi(\theta)$ that meaningfully encodes initial uncertainty (MacKay, 1992a). Third, it must be computationally feasible to update this prior distribution based on observed data to yield a posterior distribution. In many Bayesian machine learning use cases, one or more of these are violated; in the case of Bayesian neural networks (BNNs), all three are likely violated, as we discuss further in Section 4.2.2. As a result, several "generalized" approaches have emerged that aim to leverage the useful properties of Bayesian inference, but without strict reliance on the standard prior-posterior pipeline.

Power posteriors (Friel & Pettitt, 2008; Lartillot & Philippe, 2006) introduce a "temperature" parameter $\lambda > 0$ that controls the influence of the likelihood relative to the prior distribution. The resulting posterior distribution takes the form

$$\pi_\lambda(\theta \mid \mathcal{D}) \propto p(\mathcal{D} \mid \theta)^\lambda \, \pi(\theta).$$

When $\lambda = 1$, this reduces to the standard Bayesian posterior. Smaller values of λ down-weight the likelihood, while larger values of λ downweight the prior distribution, providing

a flexible mechanism for robustness and control of either model or prior misspecification (Grünwald, 2012; Grünwald & Ommen, 2017).

Gibbs posteriors (Zhang, 1999, 2006) replace the log-likelihood with a general loss function L, yielding the update

$$\pi_L(\theta \mid \mathcal{D}) \propto \exp\left(-\lambda \sum_{i=1}^{n} L(\theta, x_i, y_i)\right) \pi(\theta)$$

where λ is again a scaling parameter. This framework allows Bayesian-like updates even when no likelihood is specified, enabling coherent inference in misspecified or loss-driven contexts (Bissiri et al., 2016; Kallioinen et al., 2023). When $L(\theta, x, y) = -\log p(y \mid x, \theta)$, the Gibbs posterior also reduces to the standard Bayesian posterior distribution.

Other approaches in this space include PAC-Bayes, which provides high-probability performance guarantees by balancing empirical fit and model complexity (McAllester, 1999); and optimization-based generalized variational inference, with a flexible objective that trades off data fit and prior regularization (Knoblauch et al., 2022).

4.2.2 Bayesian neural networks

Bayesian neural networks (BNNs) are a widely popular approach for uncertainty quantification in deep learning (MacKay, 1992b; Neal, 1996). In principle, they apply full Bayesian inference to neural networks, placing a prior distribution over the weights and biases of the network θ, which is then updated based on observed data. This is believed to improve calibration for the resulting posterior predictive by capturing epistemic uncertainty in the modeling process (Izmailov et al., 2021; Snoek et al., 2015).

In practice, the extent to which BNNs satisfy the underlying assumptions of Bayesian inference is questionable. Full posterior inference is generally intractable, so approximation schemes are typically applied to update θ (Blundell et al., 2015; Welling & Teh, 2011). Specifying or engineering a meaningful prior distribution over θ is also difficult (Nalisnick & Smyth, 2018; Sharma et al., 2023), since a choice made for computational ease or flexibility in parameter space may have unintended downstream properties in predictive space (Fortuin et al., 2021; Noci et al., 2024). Ultimately, this tends to result in model

updates which diverge substantially from the "true" Bayesian posterior distribution (Coker et al., 2022; Foong et al., 2020; Trippe & Turner, 2018).

Several BNN-specific efforts have therefore also moved away from strict adherence to standard Bayesian updating. One line of work introduces partial stochasticity, applying uncertainty only to a subset of weights or layers to balance expressiveness and tractability (Daxberger et al., 2021; Izmailov et al., 2020; Kristiadi et al., 2020; Sharma et al., 2023). Another explores power posteriors, referred to in the related literature as "cold posteriors" (Wenzel et al., 2020), though their theoretical justification remains debated (Izmailov et al., 2021; McLatchie et al., 2024). A third direction has attempted to learn the posterior predictive distribution directly, typically using variational techniques (Farquhar et al., 2020; Rudner et al., 2020; Sun et al., 2019).

4.2.3 Model uncertainty in deep learning

The challenges associated with BNNs have led to an active line of deep learning research focused on more directly encoding model uncertainty. The standard approach is deep ensembling (Lakshminarayanan et al., 2017), which trains multiple models with different initializations and aggregates their predictions at test time. Despite its relative simplicity, deep ensembling has been shown to provide strong empirical performance (Abdar et al., 2021; Wilson & Izmailov, 2020). Recent work on MixupMP (Wu & Williamson, 2024) extends deep ensembling through the predictive lens of martingale posterior distributions (Fong et al., 2024), using data augmentation to train ensemble members.

Other approaches also aim to capture model uncertainty directly in predictive space. These include epistemic neural networks (Osband et al., 2021), which place a joint distribution over the predictions themselves; and evidential deep learning (Sensoy et al., 2018), which targets the parameters of a higher-order distribution (e.g. a Dirichlet over class probabilities) to simultaneously represent aleatoric and epistemic uncertainty.

4.3 Rethinking Bayesian prediction

We now consider a general supervised learning setting where the primary interest is prediction, conditional on an observed sample of size n. More specifically, we observe training data $\mathcal{D}_{\text{obs}} = \{(x_i, y_i)\}_{i=1}^n$ and wish to return a predictive density $p(y \mid x^*, \mathcal{D}_{\text{obs}})$ at a new, unseen observation x^*. This setting arises across many real-world applications, from image classification to medical diagnosis.

The standard Bayesian approach to this problem is well-documented: specify a likelihood $p(y \mid x^*; \theta)$, then mix it over the parametric posterior distribution to compute the posterior predictive

$$p(y \mid x^*, \mathcal{D}_{\text{obs}}) = \int \underbrace{p(y \mid x^*; \theta)}_{\text{aleatoric}} \underbrace{\pi(\theta \mid \mathcal{D}_{\text{obs}})}_{\text{epistemic}} \, d\theta. \tag{4.1}$$

The parametric posterior distribution captures *epistemic uncertainty* in the underlying θ, which is reducible as $n \to \infty$ and more data is observed, while the likelihood captures *aleatoric uncertainty* inherent in the data, which is irreducible even when the true parameter is known (Bickford Smith et al., 2024).

This approach is coherent and optimal in a decision-theoretic sense, but as previously discussed, it requires a well-specified model with the notion of a random parameter θ, the specification of an associated prior distribution $\pi(\theta)$, and the ability to compute the resulting prior-posterior update. The validity of these constructions has increasingly been questioned, especially in the Bayesian deep learning literature.

It is therefore worth stepping back to consider when and why we want to accept the complexity of a full Bayesian analysis. Bayesian updates are especially valuable in sequential settings where the model must be updated iteratively as new data arrives, such as active learning, continual learning, or Bayesian optimization (Houlsby et al., 2011; Snoek et al., 2012). They are also important when we care about a principled decomposition of uncertainty, such as distinguishing between epistemic and aleatoric components as outlined above, or designing adaptive learning rates based on model confidence.

However, our original setting of interest is one-step-ahead prediction, with a fixed dataset and no intention to update the model. We only care about the total predictive

uncertainty for a new input, and do not need to determine how it decomposes into model-based (epistemic) and data-based (aleatoric) components. In this context, there is nothing fundamental about the parameter θ in Equation 4.1 — it solely serves to index the sampling density, and must ultimately be marginalized out so that the final predictive can be expressed entirely in terms of observable data. A full Bayesian posterior distribution may therefore be nothing but a poorly specified and computationally expensive detour, especially when applying a complex model such as a BNN.

A more direct alternative is to adopt a plug-in approach: specify a predictive model $g(y \mid x; \alpha)$, possibly over an extended model space, and optimize $\hat{\alpha}$ using a flexible machine learning technique. If g is sufficiently expressive, the resulting predictive distribution $g(y \mid x^*; \hat{\alpha})$ can yield calibrated uncertainty estimates and prediction intervals without a prior-posterior calculation. While there have been efforts to directly target the posterior predictive using BNNs, these typically rely on variational inference and still assume an underlying Bayesian structure with a posterior update.

In the absence of a typical Bayesian model, how can we retain guarantees of coherent or well-calibrated prediction (Dawid, 1982)? As discussed in Section 1.3, Doob (1949) demonstrates the consistency of Bayesian updating, where new observations sequentially resolve uncertainty about the parameter. The key condition underlying this result is that as we impute data, the posterior mean of the quantity of interest forms a martingale. This requirement can be reinterpreted as a predictive coherence condition, which corresponds to the future data being c.i.d. given the observed data, as defined in Equation 1.7.

This c.i.d. property can be assessed operationally. Consider the setup of Equation 1.3, where we generate a one-step predictive update from our model, add it to the training set, and repeat, with the eventual goal of recursively imputing the complete missing data. A core property of a well-calibrated Bayesian model is that it should preserve a self-consistent amount of uncertainty when updated with its own predictions. That is, averaging over different possible imputations, the updated predictive distribution at each step should remain stable. If the sequential prediction intervals become wider, it means the updating scheme is losing information from previous data; if they become narrower, it

suggests the model has "hallucinated" certainty without any new data (Falck et al., 2024).

If the c.i.d. condition holds, we argue that under the predictive representation, our approach is producing meaningful Bayesian uncertainty estimates. This is true even if its underlying functionality actually relies on nothing more than a plug-in approach, such as maximum likelihood estimation. In other words, the fundamental essence of Bayesian inference is not subjective prior specification or a prior-posterior update — it is the coherent sequential resolution of predictive uncertainty, and this can still be achieved by a model that does not appear "Bayesian" in the traditional sense.

Explained another way, recall that the predictive framework views missing data as the source of uncertainty, where any identifiable target quantity would be known given the complete data. Let the complete data $\mathcal{D} = \{\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{mis}}\}$ be composed of the observed data $\mathcal{D}_{\text{obs}} = \{(x_i, y_i)\}_{i=1}^{n}$ and the missing (or future) data $\mathcal{D}_{\text{mis}} = \{(x_i, y_i)\}_{i=n+1}^{\infty}$. When the target quantity is the latent parameter θ, this is represented as

$$p(\theta \mid \mathcal{D}_{\text{obs}}) = \int p(\theta, \mathcal{D}_{\text{mis}} \mid \mathcal{D}_{\text{obs}})\, d\mathcal{D}_{\text{mis}} = \int p(\theta \mid \mathcal{D})\, p(\mathcal{D}_{\text{mis}} \mid \mathcal{D}_{\text{obs}})\, d\mathcal{D}_{\text{mis}}. \quad (4.2)$$

If our one-step predictive update forms a martingale for the posterior mean of θ, this representation yields the standard parametric posterior distribution.

Our goal, however, is the posterior predictive distribution from Equation 4.1. We can connect these two frameworks by substituting Equation 4.2 directly into the posterior predictive equation as

$$\begin{aligned} p(y \mid x^*, \mathcal{D}_{\text{obs}}) &= \int p(y \mid x^*; \theta)\, p(\theta \mid \mathcal{D}_{\text{obs}})\, d\theta \\ &= \int p(y \mid x^*; \theta) \left[\int p(\theta \mid \mathcal{D})\, p(\mathcal{D}_{\text{mis}} \mid \mathcal{D}_{\text{obs}})\, d\mathcal{D}_{\text{mis}} \right] d\theta. \end{aligned}$$

By changing the order of integration, this becomes

$$p(y \mid x^*, \mathcal{D}_{\text{obs}}) = \int \left[\int p(y \mid x^*; \theta)\, p(\theta \mid \mathcal{D})\, d\theta \right] p(\mathcal{D}_{\text{mis}} \mid \mathcal{D}_{\text{obs}})\, d\mathcal{D}_{\text{mis}}.$$

The inner integral is simply the predictive distribution given the complete data $\mathcal{D}$, which we can write as $p(y \mid x^*, \mathcal{D})$. This is because $p(\theta \mid \mathcal{D})$ becomes a point mass at the true parameter determined by the complete data. The expression then simplifies to

$$p(y \mid x^*, \mathcal{D}_{\text{obs}}) = \int p(y \mid x^*, \mathcal{D})\, p(\mathcal{D}_{\text{mis}} \mid \mathcal{D}_{\text{obs}})\, d\mathcal{D}_{\text{mis}}.$$

If our one-step predictive update is c.i.d., which corresponds to the martingale condition, then this representation again yields the same uncertainty as the typical parametric approach. Our key insight is that this means *we can directly interpret the update as our actual predictive uncertainty.* By constructing a c.i.d. sequence, we ensure that the uncertainty in our initial update is the same as the uncertainty we would get from imputing and marginalizing over the missing data; paradoxically, but conveniently, this means that we no longer have to perform the complete update.

4.4 Evaluating predictive coherence

The insights of the previous section motivate a new perspective: to operationally assess whether a model is performing "Bayesian" inference, we simply need to evaluate whether it produces c.i.d. observations under sequential updating. In other words, if a predictive resampling procedure generates a c.i.d. sequence, then the model is predictively resolving uncertainty in a manner consistent with Bayesian updating — even if this does not involve a prior distribution, a likelihood, or any explicit posterior computation.

How do we conduct this evaluation? Let the observed data $\mathcal{D}_{\text{obs}} = (\mathbf{X}, \mathbf{y}_1)$ consist of a covariate design matrix and outcome vector. To implement sequential updates for supervised learning, we apply the "block resampling" strategy from Section 2.4.2, where the observed covariates are treated as fixed, and the target for each update is the imputation of a new outcome vector. In other words, we fit the model to $\mathcal{D}_{\text{obs}}$, then sample a new $n \times 1$ outcome vector $\mathbf{Y}_2$ from the model's predictive distribution conditional on $\mathbf{X}$. We append the realized $(\mathbf{X}, \mathbf{y}_2)$ to the data and repeat to generate a sequence of imputations. If the model is performing a valid Bayesian update, then this sequence, $(\mathbf{Y}_k)_{k\geq 2} = \{\mathbf{Y}_2, \mathbf{Y}_3, \ldots\}$, should be c.i.d. given $\mathbf{y}_1$, with the covariates $\mathbf{X}$ fixed throughout.

Under this construction, we present a novel metric based on the consistency of predictions under one-step sequential updating, followed by a metric that extends the exchangeability test of Mlodozeniec et al. (2024) to the supervised learning setting.

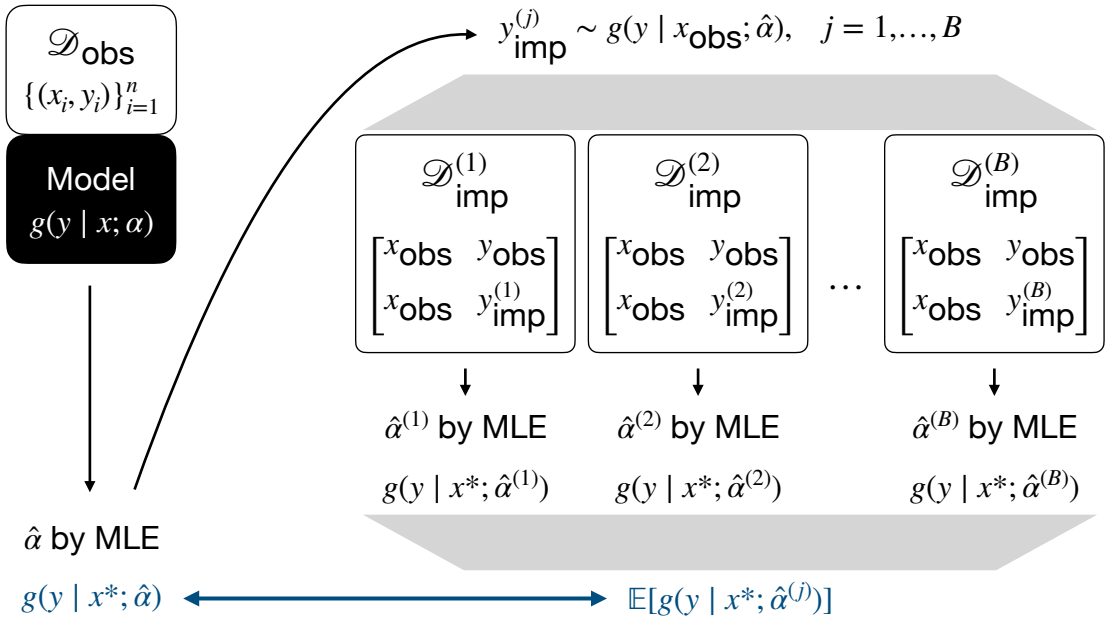


Figure 4.1: Predictive coherence check via consistency under one-step imputation. A coherent model should yield similar predictive distributions at a new test point x^* before (left) and after (right) updating the model with synthetic outcomes imputed at x_{obs}.

4.4.1 Consistency under sequential updating

In this test, we evaluate whether the average predictive distribution at a new input x^* remains stable under a one-step model update. In other words, we evaluate how the model's predictions evolve when its own outputs are used to update it. The central intuition is simple: conditionally on the observed data, the model learns a representation of predictive uncertainty over future observations. If this uncertainty is c.i.d., then it should remain consistent in expectation as those hypothetical outcomes are realized. If not, it suggests that new information is being introduced through the sequential update, and the model predictions have incorrectly represented the underlying uncertainty.

Formally, let $g(y \mid x^*, \hat{\alpha}_1)$ denote the predictive density at a test point x^* under optimized model parameters $\hat{\alpha}_1$ fit to the observed dataset $(\mathbf{X}, \mathbf{y}_1)$. We assess whether this density remains stable after the model is updated on synthetic data drawn from itself. Across B trials, we sample responses $\mathbf{Y}_2^{(b)} \sim p(\cdot \mid \mathbf{X}, \hat{\alpha}_1)$, augment the training set with realizations $(\mathbf{X}, \mathbf{y}_2)$, refit the model to obtain updated parameters $\hat{\alpha}_2^{(b)}$, and compute the new predictive density as $g(y \mid x^*, \hat{\alpha}_2^{(b)})$.

Averaging over the trials gives us an equal-weighted mixture density formed from the

B post-update predictive densities. If the model is predictively coherent, the original predictive density should closely match this mixture density, or

$$g(y \mid x^*, \hat{\alpha}_1) \approx \frac{1}{B} \sum_{b=1}^{B} g(y \mid x^*, \hat{\alpha}_2^{(b)}).$$

No new information is introduced through the resampled datasets, only uncertainty already encoded by the initial model. If the model's predictive uncertainty is c.i.d., we should therefore expect it to retain a consistent level of uncertainty through a one-step update. A shift in the predictive density would indicate that the model is either over- or under-confident, and is effectively "hallucinating" or introducing artificial variability into its predictions (Falck et al., 2024). This diagnostic is illustrated in Figure 4.1.

4.4.2 Exchangeability via log-joint variance

In this test, we assess the conditional exchangeability of future predictions by measuring the variance of the log-joint likelihood across sequentially resampled blocks, with respect to the initial model. In other words, we alternate between generating data from a model and using that data to progressively re-optimize the model parameters. We then measure whether each of those imputed data blocks would have been equally likely under the original model. If so, it suggests that future observations are c.i.d., and the model is preserving consistent uncertainty through updates.

This extends the test proposed by Mlodozeniec et al. (2024), originally developed for one-dimensional sequences, to our block resampling setting. Specifically, they measure

$$m_{\text{var}}(s_1, \ldots, s_n) = \mathbb{E}_{X_1,\ldots,X_n} \left[\text{Var}_\pi \left(\sum_{i=1}^{n} \log s_i(X_{\pi(i)} \mid X_{\pi(1)}, \ldots, X_{\pi(i-1)}) \right) \right]$$

for predictive densities $(s_1, s_2, \ldots, s_n)$ over permutations π (Mlodozeniec et al., 2024, Equation 5). If future predictions are exchangeable, then this should be invariant to reordering and the variance should be near zero.

To adapt this to our setting, we again consider one-step sequential updates. If a model produces c.i.d. predictions, then the likelihood of future blocks of data should remain consistent even after updating the model on intermediate blocks. To test this, we first optimize the model on the observed data $(\mathbf{X}, \mathbf{y}_1)$ to obtain initial parameters $\hat{\alpha}_1$. We

then iterate K times for each of B independent trials. At each iteration, we sample a new outcome random vector $\mathbf{Y}^{(b)}_{\text{imp},k+1} \sim p(\cdot \mid \mathbf{X}, \hat{\alpha}^{(b)}_k)$ and append its realization $\mathbf{y}^{(b)}_{\text{imp},k+1}$ to the data alongside the replicated design matrix. We refit the model on all data available so far to obtain the updated parameters $\hat{\alpha}^{(b)}_{k+1}$.

This process ultimately generates $B \times K$ blocks $\mathbf{y}^{(b)}_{\text{imp},k}$. The log-likelihood of each imputed block, evaluated under the original model (with parameters $\hat{\alpha}_1$), is

$$\log p\left(\mathbf{y}^{(b)}_{\text{imp},k} \mid \mathbf{X}, \hat{\alpha}_1\right), \quad b = 1, \ldots, B \text{ and } k = 2, \ldots, K.$$

If the model is predictively coherent and produces exchangeable blocks, then within each of the B independent chains, the empirical variance of these K log-likelihoods should be close to zero, or

$$\text{Var}_k\left[\log p\left(\mathbf{y}^{(b)}_{\text{imp},k} \mid \mathbf{X}, \hat{\alpha}_1\right)\right] \approx 0, \quad b = 1, \ldots, B \text{ and } k = 2, \ldots, K. \tag{4.3}$$

A large variance suggests that some imputed updates lead to future data which is much more or less likely under the original model, indicating a breakdown in the c.i.d. condition. This would suggest that the model's sequential updates are introducing mis-calibrated predictive behavior.

4.5 Illustrations

In this section, we demonstrate the predictive coherence diagnostics on a regression example. Our demonstration is limited to a simple linear model; while we expect analogous predictive representations to hold in deep learning contexts, a full treatment is beyond the current scope and we leave it to future work. For $n \in \{20, 50\}$ data points with $p = 5$ covariates (including an intercept), we generate synthetic data according to

$$y_i = x_i^\top \beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2),$$

where the (heteroskedastic) conditional variance is defined as $\sigma_i^2 = \text{softplus}(x_i^\top \gamma)$. We sample the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p} \sim \text{Uniform}(-1, 1)$, with the first column fixed as 1 to serve as the intercept. The true regression weights $\beta \in \mathbb{R}^p$ and variance weights $\gamma \in \mathbb{R}^p$

are also drawn independently from Uniform$(-1, 1)$, with the intercept coefficient γ_0 fixed to 1 to ensure nondegenerate variance.

For baseline comparison, we fit a homoskedastic Bayesian linear regression (BLR) model with a normal-inverse-gamma prior distribution over the regression weights and noise variance. The posterior distribution over the parameters is conjugate, and the posterior predictive at a new point x^* is a Student's t-distribution

$$y^* \sim t_{2a}\left(x^{*\top}\mu,\ \frac{b}{a}\left(1 + x^{*\top}\Lambda^{-1}x^*\right)\right).$$

where μ and Λ are the mean and precision matrix for the weights, and (a, b) are shape and scale parameters for the precision. We use a diffuse prior with $\mu_0 = \mathbf{0}$, $\Lambda_0 = 10^{-3}I_p$, and $(a, b) = (2, 0.5)$.

To directly target this predictive distribution via maximum likelihood, we fit a parametric t-distribution to the data using the following flexible form. The predictive mean and variance are modeled as

$$\mu(x) = x^\top\beta, \qquad \sigma^2(x) = \text{softplus}(x^\top\gamma),$$

with shared degrees of freedom ν. The parameters $\theta = (\beta, \gamma, \nu)$ are estimated by minimizing the negative log-likelihood, with an optional ℓ_2 (ridge) regularization penalty controlled by λ according to the loss function

$$\mathcal{L}(\theta) = -\sum_{i=1}^{n} \log t_\nu\left(y_i \mid \mu(x_i), \sqrt{\sigma^2(x_i)}\right) + \lambda\left(\|\beta\|_2^2 + \|\gamma\|_2^2\right).$$

Prediction at a new input x^* then uses the fitted parameters θ to evaluate

$$y^* \sim t_\nu\left(x^{*\top}\beta,\ \sqrt{\text{softplus}(x^{*\top}\gamma)}\right).$$

This model bypasses the need for a prior-likelihood decomposition by directly fitting the predictive distribution, while maintaining a comparable structure to the Bayesian posterior predictive.

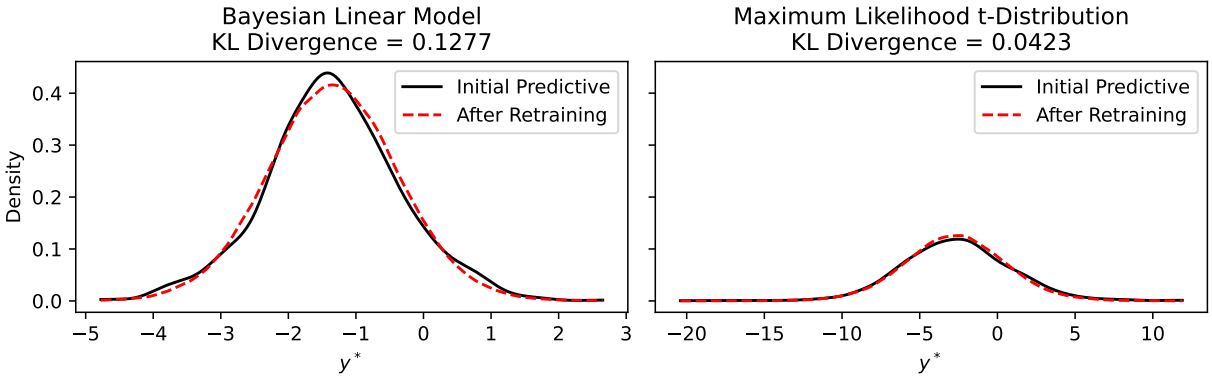


Figure 4.2: Predictive consistency at a fixed test input x^* with $n = 20$ observed data points, comparing Bayesian linear regression (left) with a directly maximized t-distribution (right). Solid black lines show the initial predictive density, while dotted red lines show the mixture of predictions across $B = 100$ trials after retraining with synthetic outcomes.

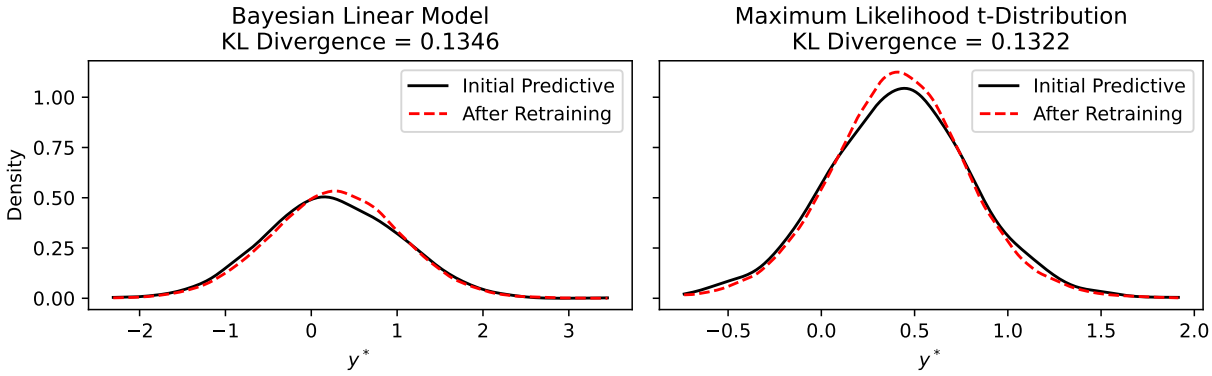


Figure 4.3: Predictive consistency at a fixed test input x^* with $n = 50$ observed data points, comparing Bayesian linear regression (left) with a directly maximized t-distribution (right). Solid black lines show the initial predictive density, while dotted red lines show the mixture of predictions across $B = 100$ trials after retraining with synthetic outcomes.

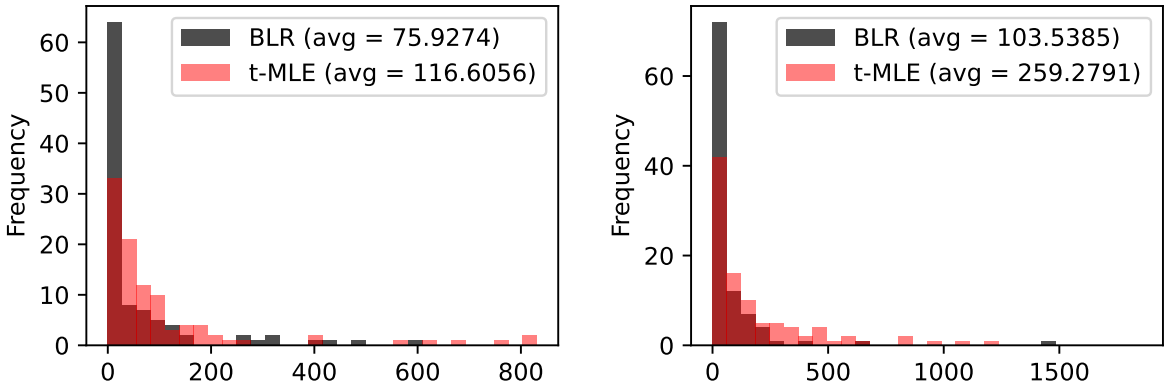
4.5.1 Consistency under sequential updating results

For the first test, we visualize sequential consistency by plotting kernel density estimates of the predictive density at a fixed input x^*. We compare the initial predictive density $g(y \mid x^*, \hat{\alpha}_1)$ to the equal-weighted mixture of updated predictive densities, $\frac{1}{B}\sum_{b=1}^{B} g(y \mid x^*, \hat{\alpha}_2^{(b)})$, obtained after retraining on synthetic data imputed at $\mathbf{X}$ across $B = 100$ trials. The Kullback-Leibler (KL) divergence between the initial and mixture models is estimated using 10^5 Monte Carlo samples to quantify the degree of predictive drift. Results show that the maximum likelihood-based model retains a stable predictive distribution, with a KL divergence less than or equal to the BLR baseline. This suggests that direct optimization can preserve uncertainty without an explicit posterior update.

Notably, compared to BLR, the MLE-based predictive distribution is very diffuse in the $n = 20$ setting (Figure 4.2), but more concentrated for $n = 50$ observed data points (Figure 4.3). This behavior can be interpreted through the lens of (possibly implicit) prior specification. Under BLR, the spread of posterior predictive uncertainty is largely determined by prior dispersion. Based on the observed results, the specific MLE approach implemented here behaves in the manner of a Bayesian model whose prior distribution becomes increasingly concentrated as the sample size grows. From this perspective, different specifications of the MLE model, provided they satisfy predictive coherence, can be understood as Bayesian under different implicit priors.

4.5.2 Exchangeability via log-joint variance results

For the second test, we assess conditional exchangeability by evaluating the average variance of the log-likelihood for synthetic outcomes. We alternate between simulating a new block of data and retraining the model on the updated dataset for $K = 10$ forward steps, and compute the log-likelihood of each new block under the original (pre-update) model. We compute the empirical variance of the K log-likelihoods, then repeat this across $B = 100$ trials and average the results. Results show that the MLE-based model has a larger variance than the Bayesian baseline, suggesting that this plug-in model may struggle to generate conditionally exchangeable updates as resampling progresses.



(a) Histogram for $n = 20$ experiment

(b) Histogram for $n = 50$ experiment

Figure 4.4: Average variance of joint log-likelihood for $K = 10$ resampled blocks under the original model, comparing Bayesian linear regression (black) with a directly maximized t-distribution (red) over $B = 100$ trials.

4.6 Conclusion

Bayesian inference has been interpreted in many different ways — for example, as a framework for updating beliefs in light of new data, or more broadly as the use of probability as the language of uncertainty. In this work, we emphasize a predictive interpretation of Bayes, under which any learning rule is evaluated based on the sequence of predictive distributions it produces. For a learning rule to provide coherent, unbiased uncertainty quantification within a bootstrap scheme, the essential requirement is that its predictions are identically distributed conditional on the observed data, a condition often mathematically formalized through the use of martingales.

In practice, this predictive perspective opens the door for methods that are not traditionally Bayesian to be interpreted through a Bayesian lens, as long as they yield c.i.d. predictions. We introduce two diagnostics to assess this property and apply them to examples from supervised learning. The key direction for future work will be the development of more expressive parametric models that can capture a wider class of distributions via direct optimization. Our t-distribution regression example demonstrated only limited success, highlighting the need for more flexible predictive representations before these concepts can be reliably extended to more complex settings.

Chapter 5

Predictive resampling with conditionally identically distributed parametric models

5.1 Introduction

In this chapter, we examine resampling schemes for predictive inference. While martingale posterior-style approaches provide a compelling alternative to prior-likelihood calculation, it is crucial to establish the necessary properties for a bootstrap-based updating rule to yield meaningful uncertainty quantification.

We first review existing predictive constructions, including plug-in parametric updates and c.i.d. sequences. We then introduce a resampling procedure based on c.i.d. parametric models that enables valid inference from a one-step c.i.d. update, thereby avoiding a full resampling procedure. The approach is illustrated with applications from nonparametric curve fitting and time-to-event survival analysis.

5.2 Review of existing approaches

Recall from Section 1.3 that the predictive view of Bayesian uncertainty quantification directly targets the density of unobserved data, via the sequential factorization of Equation 1.2. This can be implemented using the Bayesian bootstrap (Equation 1.4). However, a frequent objection to established bootstrap methods, both classical and Bayesian, is that resampling from discrete point masses yields a discrete predictive distribution, lacking smoothness and failing to capture continuous uncertainty.

In contrast, a key advantage of the martingale posterior framework is that it enables the use of a broader class of predictive models, allowing modern machine learning methods to be leveraged for Bayesian inference as a form of generalized bootstrap (Fong & Yiu, 2024a). The one-step-ahead approach allows for fine control over the procedure, with the primary requirement being that the predictive updates form a c.i.d. sequence, as specified in Equation 1.7. Recent work from Battiston and Cappello (2025) on asymptotically c.i.d. sequences may provide a pathway towards relaxing this condition even further.

Building on work by Hahn et al. (2020), Fong et al. (2024) propose a recursive c.i.d. update using bivariate copulas. A copula is a function that couples multivariate distribution functions to their marginal distributions. In this case, the copula construction is applied to update the predictive distribution,

$$p_{i+1}(y) = c_{i+1}\{P_i(y), P_i(y_{i+1})\},$$

where the copula is defined as

$$c_{i+1}\{P_i(y), P_i(y_{i+1})\} = \frac{\int f_\theta(y)\, f_\theta(y_{i+1})\, \pi(\theta \mid y_{1:i})\, d\theta}{p_i(y)\, p(y_{i+1})}.$$

This provides a flexible mechanism which can be tailored for various settings (univariate density estimation, multivariate density estimation, regression, etc.) to yield nonparametric c.i.d. predictive distributions.

An alternative approach from Holmes and Walker (2023) builds upon the parametric bootstrap of Efron (2012). The parametric bootstrap replaces the Bayesian bootstrap's empirical distribution with a plug-in density estimator $f(\cdot \mid \hat{\theta})$, where $\hat{\theta}$ is some functional of the observed data, such as the MLE. This allows the bootstrap procedure to incorporate model structure and parametric assumptions, potentially yielding more efficient resampling when the model is well-specified.

Holmes and Walker (2023) show that if $\hat{\theta}$ is an unbiased estimator of the true parameter θ, then predictive resampling can similarly be conducted with a plug-in parametric predictive density $Y_{n+1} \sim f(\cdot \mid \hat{\theta}_n)$, where data is sequentially drawn from the model fit to the current MLE. The parameter is updated via stochastic gradient descent according to

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \frac{1}{n+1} \nabla_\theta \log p(y_{n+1} \mid \hat{\theta}_n),$$

where the score function $\nabla_\theta \log p$ has mean zero under the model, ensuring that the sequence $(\hat{\theta}_{m \geq n})$ forms a martingale. As $n \to \infty$, the limiting distribution P_∞ therefore exists, and can be interpreted as a sample from a "parametric martingale" posterior.

Though we do not discuss them in detail here, additional efforts have sought to further optimize the resampling process for particular settings. For example, Fong and Yiu (2024b) derive a quantile-based predictive update that allows resampling to be performed entirely via Uniform$[0, 1]$ draws. This yields a quantile regression procedure that is substantially faster than existing approaches in dependent Bayesian nonparametrics.

5.3 Basic constructions for conditionally identically distributed sequences

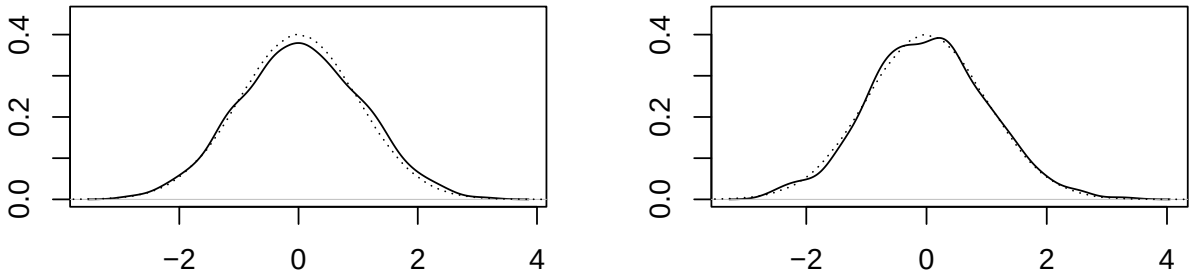
In this section, we examine a class of closed-form c.i.d. constructions, where we directly derive a sequence of predictions that are identically distributed given the initial data. While this approach is not feasible for every setting, it reinforces the key insight from Section 4.3. In particular, we argue that the c.i.d. condition is both necessary and sufficient for coherent predictive uncertainty, so if we can specify a c.i.d. model to impute the first missing observation, then carrying out the full predictive resampling process is no longer necessary. This is because the initial predictive distribution fully captures the Monte Carlo uncertainty that would otherwise be approximated through resampling.

5.3.1 Univariate location parameter

As a basic example of a c.i.d. sequence, consider the autoregressive update defined by

$$\begin{aligned} Y_m \mid y_{1:m-1} &\sim \mathcal{N}(\theta_m, \sigma_m^2) \\ \theta_{m+1} &= (1-a)\theta_m + a y_m \\ \sigma_{m+1}^2 &= (1-a^2)\sigma_m^2 \end{aligned}$$

where the hyperparameter $a \in (0, 1)$ is a fixed value that determines the level of autocorrelation between subsequent samples. To initialize the process, we let $\theta_1 = 0$ and $\sigma_1^2 = 1$, meaning that the first observation is randomly drawn according to $Y_1 \sim \mathcal{N}(0, 1)$. Based on its realization, the mean and variance are recursively updated to generate the next



(a) Density with autocorrelation $a = 0.3$ (b) Density with autocorrelation $a = 0.7$

Figure 5.1: Initial distribution $y_1 \sim \mathcal{N}(0, 1)$ (dotted) compared to kernel density of final samples y_M across trials (solid) for univariate resampling.

observation, and so on. This means that at any step m, the parameters θ_m and σ_m^2 are deterministic given the history $y_{1:m-1}$, while the next observation Y_m is a random variable.

To verify that this sequence is c.i.d., we confirm that the expected predictive density for the next step equals the current density,

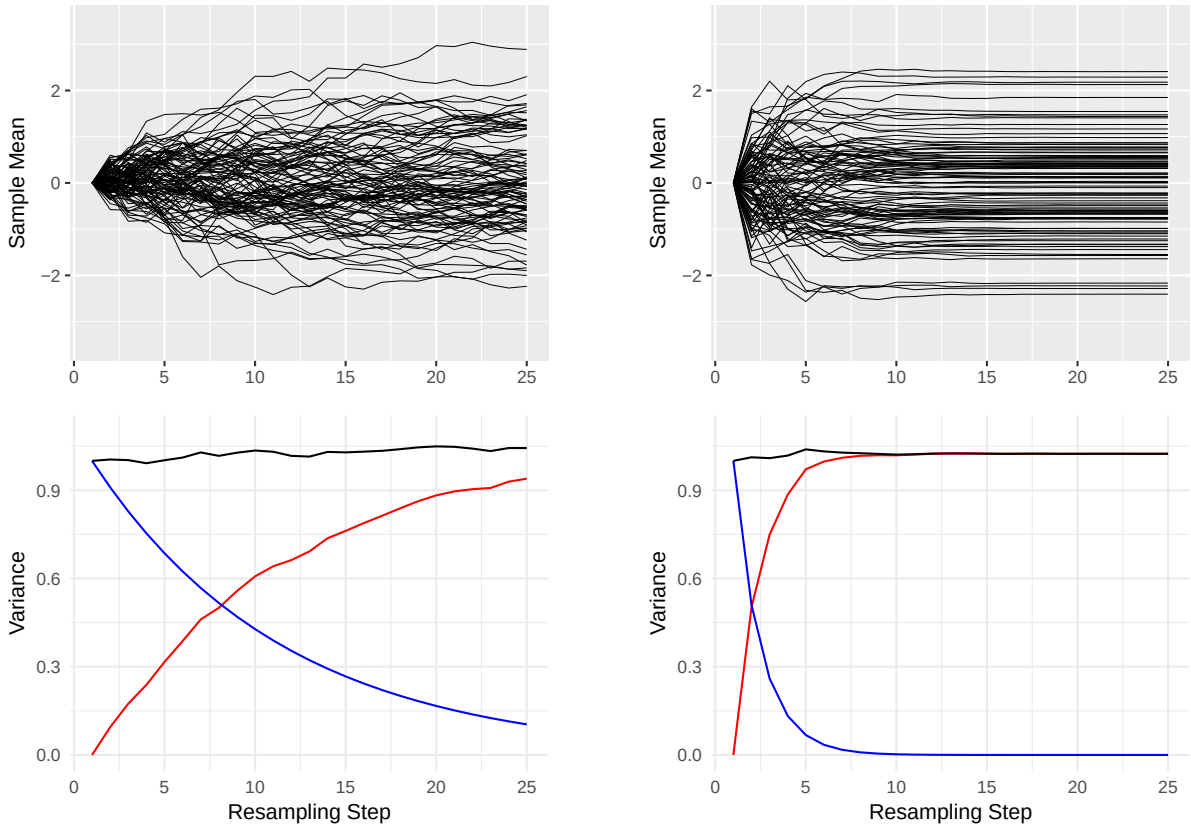
$$\mathbb{E}\left[\mathcal{N}(y \mid \theta_{m+1}, \sigma_{m+1}^2) \mid y_{1:m-1}\right] = \mathcal{N}(y \mid \theta_m, \sigma_m^2).$$

This equality holds because, conditional on $y_{1:m-1}$, the expected parameter updates preserve the moments of the predictive distribution. The expected mean is $\mathbb{E}[\theta_{m+1}] = (1 - a)\theta_m + a\theta_m = \theta_m$, and the total variance is $\mathbb{E}[\sigma_{m+1}^2] + \mathrm{Var}(\theta_{m+1}) = (1 - a^2)\sigma_m^2 + a^2\sigma_m^2 = \sigma_m^2$.

To demonstrate this in simulation, we draw samples up to $M = 100$ for 1000 trials and compare the distribution of the original $Y_1 \sim \mathcal{N}(0, 1)$ to the distribution of the final observed values at the end of each trial. Figure 5.1 shows that the two kernel densities are roughly equivalent, for autocorrelation parameters $a = 0.3$ and 0.7.

As the resampling progresses, the model variance σ_i^2 decreases while the empirical variance in θ_i across the trials increases. These effects combine to maintain a constant total variance, consistent with the c.i.d. nature of the sequence. Figure 5.2 illustrates this behavior over the first 25 resampling steps, showing sample mean trajectories from 100 trials (top) and variance decomposition (bottom), for $a = 0.3$ and 0.7.

In other words, our total uncertainty remains fixed, but the nature of the uncertainty shifts from epistemic (model-based, reducible with additional data) to aleatoric (data-based, irreducible once outcomes are observed) as more data is observed. The rate of this transition is governed by the autocorrelation parameter a, as shown in Figure 5.2.



(a) Uncertainty decomposition with $a = 0.3$ (b) Uncertainty decomposition with $a = 0.7$

Figure 5.2: Univariate resampling trajectories (top) and decomposition of variance (bottom). The model variance σ_i^2 (blue) decreases and the data variance of θ_i across trials (red) increases with resampling step m, resulting in constant overall variance (black).

This helps to demonstrate why the construction of a c.i.d. parametric model eliminates the need for actual imputation of the infinite data — the initial model variance exactly captures the eventual data variance that would be realized across the Monte Carlo trials.

5.3.2 Simple linear model

We now consider a simple linear model with no intercept. Let $\mathbf{X} \sim \mathcal{N}(0, 1)$ be a fixed $n \times p$ data matrix and $\mathbf{y}_{obs} = \mathbf{y}_1 = \mathbf{X}\beta_{sim} + \epsilon$ be a vector of responses, generated using some random coefficients β_{sim}. As in Sections 2.4.2 and 4.3, the c.i.d. update condition relies on a "block resampling" procedure where the full design matrix $\mathbf{X}$ is copied at each step, and the target for imputation is a complete vector of outcomes.

Algorithm 2 Conditionally identically distributed linear model

$\mathbf{X} \leftarrow \mathcal{N}(0,1)$
$\mathbf{H} \leftarrow \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ (hat matrix)
$\beta_{sim} \leftarrow \text{Unif}[1,5]$
$\mathbf{y}_{obs} \leftarrow \mathbf{X}\beta_{sim} + \epsilon$
$s_1 = 1, a \in (0,1)$ (hyperparameters)
for $m = 2, ..., 1000$ **do**
    $\mathbf{y}_{m+1} \leftarrow \mathbf{X}\beta_m + s_m\mathbf{Hz}$
    $\beta_{m+1} \leftarrow (1-a)\beta_m + a(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}_{m+1}$
    $s_{m+1} \leftarrow s_m\sqrt{1-a^2}$
end for

Specifically, our c.i.d. model for the outcomes is defined by

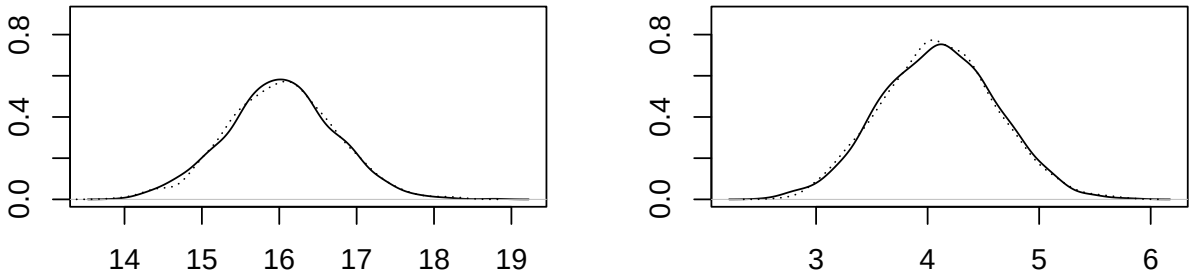
$$\mathbf{y}_{m+1} = \mathbf{X}\beta_m + s_m\mathbf{Hz}$$

$$\beta_{m+1} = (1-a)\beta_m + a(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}_{m+1}$$

$$s_{m+1} = s_m\sqrt{1-a^2}$$

where β_1 is fitted by an ordinary least squares regression of $\mathbf{y}_{obs}$ on $\mathbf{X}$ and we initialize $s_1 = 1$. The hyperparameter $a \in (0,1)$ again determines the level of autocorrelation; the hat matrix is denoted as $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ and $\mathbf{z} \sim \mathcal{N}(0,1)$ represents a vector of standard normals. This procedure is fully detailed in Algorithm 2.

To simulate, we let $n = 20$ and $p = 7$ and follow the procedure in Algorithm 2, with the regression coefficients drawn from $\beta_{sim} \sim \text{Uniform}[1,5]$. We resample up to $M = 100$ for 1000 trials with $a = 0.7$. Note that n is used here to indicate the length of the observed outcome vectors, whereas M is used to index the resampling procedure. We empirically verify the c.i.d. condition by comparing the distributions of $y_{2,i}$ (the first resampled observation at index i) and $y_{M,i}$ (the final resampled observation at index i) at two randomly chosen indices (specifically, $i \in \{3, 17\}$ out of the 20 individuals). Figure 5.3 shows that these kernel densities are roughly equivalent. In other words, across the 1000 trials, the first and final imputed predictions for each individual have the same distribution, centered in each case on the initial predicted value $(\mathbf{X}\beta_1)_i$.



(a) Densities for observation $i = 3$ (b) Densities for observation $i = 17$

Figure 5.3: Kernel density of initial $y_{2,i}$ (dotted) and final $y_{M,i}$ samples at specific indices i for linear model resampling.

5.4 Illustrations

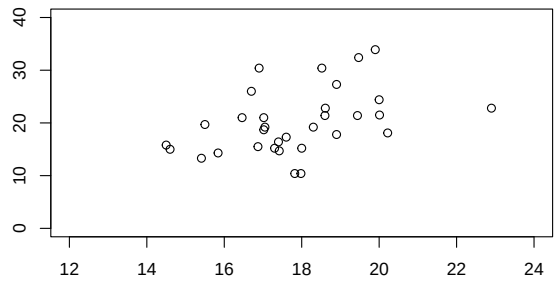
In this section, we demonstrate predictive resampling with parametric c.i.d. models on example problems from nonparametric curve fitting and time-to-event survival analysis. Code for all illustrations can be found at `https://github.com/vshirvaikar/MPPrediction`.

5.4.1 Nonparametric curve fitting

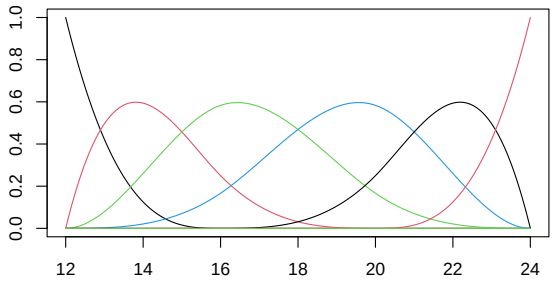
To explore predictive uncertainty under different resampling schemes, we consider basis splines or B-splines, a versatile tool in nonparametric curve fitting. B-splines, or basis splines, are piecewise-defined polynomials that facilitate flexible modeling of curves without imposing strict parametric assumptions.

We demonstrate our procedure on a univariate real-world data example from the R dataset `mtcars` example, plotted in Figure 5.4a. The original outcome y is miles per gallon for $n = 32$ different vehicle models, with $p = 10$ predictive covariates (horsepower, weight, cylinders, etc.), of which we focus on x_6 (quarter-mile time) in this example. Our basis functions are pre-specified using cubic curves (order 3) with 2 internal knots or breakpoints, as seen in Figure 5.4b. Applied spline-based modeling often employs adaptive strategies where knots are inserted or removed, or the order of the basis functions is modified, based on the features of the data, but we simplify this step in order to focus on changes in the model coefficients as additional data points are imputed.

For our simulations, we follow the previously described "block resampling" procedure, where the B-spline regression is refitted at each step to impute a new outcome vector for the fixed design matrix. This is repeated up to a total of 8 blocks across 25 trials,

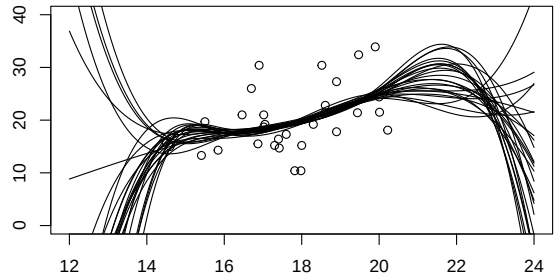


(a) Observed data for covariate x_6 (quarter-mile time) and y (miles per gallon).

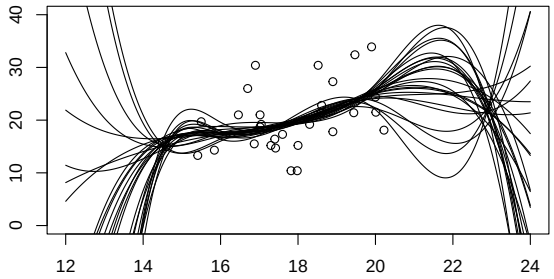


(b) B-spline basis functions of order 3 (cubic) with 2 internal knots on domain [12, 24].

Figure 5.4: Observations and basis functions for B-spline modeling of `mtcars` data



(a) Splines with "blockstrapped" data points (imputed by c.i.d. full-population repetition.)



(b) Splines with uniform random data points sampled from the observed domain.

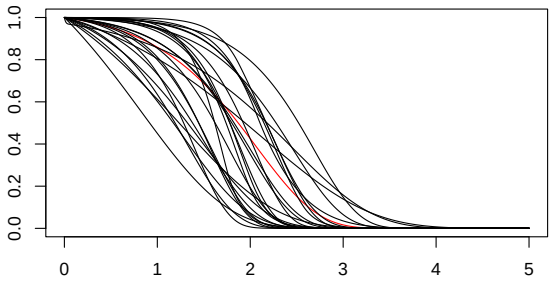
Figure 5.5: Comparison of predictive uncertainty under different resampling schemes

with the results shown in Figure 5.5a. The Monte Carlo predictive uncertainty is tightly bounded within the domain of the observed data, and much wider outside this region.

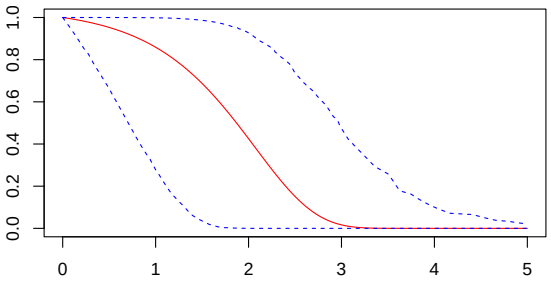
To compare, we consider a resampling scheme where new covariate values are uniformly drawn from the observed domain of $[14.5, 22.9]$. At each iteration, we generate $X_{n+1} \sim \text{Unif}[14.5, 22.9]$, predict Y_{n+1} using the fitted B-spline regression, append $(x, y)_{n+1}$ to the observed data, and refit the regression. This is repeated up to the same total sample size of $N = 32 * 8 = 256$. Figure 5.5b displays the results, with substantially wider predictive uncertainty outside the observed domain.

5.4.2 Time-to-event survival analysis

We next consider hazard regression, a method for analyzing time-to-event data, which flexibly estimates the conditional hazard function by using splines to capture the instantaneous risk of an event occurring at any given time (Kooperberg et al., 1995). In this setup, we define $f(\cdot|\mathbf{x})$ as the conditional density of survival time T given covariates $\mathbf{x}$,



(a) Predicted survival function at new observation $\mathbf{x}^*$ for baseline model (red) and 25 predictive resampling trials (black).



(b) Predicted survival function at $\mathbf{x}^*$ for baseline model (red) with 95% error bounds from 1,000 resampling trials (blue).

Figure 5.6: Survival function with error bounds for time-to-event analysis.

with corresponding conditional distribution function

$$F(t|\mathbf{x}) = \int_0^t f(u|\mathbf{x})du$$

The hazard function is then

$$h(t|\mathbf{x}) = \frac{f(t|\mathbf{x})}{1 - F(t|\mathbf{x})}$$

and polynomial splines are used to fit a linear model for the conditional log-hazard function, with coefficients selected by maximum likelihood estimation.

We simulate $n = 50$ observed data points with five-dimensional covariate vectors $\mathbf{x}_{1:n}$. The true survival times $T_{1:n}$ are drawn from a Weibull distribution dependent on the covariates, and the individual censoring times $C_{1:n}$ are drawn from a uniform distribution. The hazard regression model only receives $(\mathbf{x}, y, \delta)_{1:n}$ where each $y_i = \min\{t_i, c_i\}$ is the observed failure time and each $\delta_i = \mathbb{1}\{t_i < c_i\}$ is a censoring indicator that takes the value of 1 if the event was observed and 0 otherwise.

For predictive resampling, we iteratively bootstrap $\mathbf{x}_{n+1}$ from the observed $\mathbf{x}_{1:n}$; sample a new survival time T_{n+1} using the fitted hazard regression model, with an upper bound of 120 years; add $(\mathbf{x}, y, \delta)_{n+1}$ to the observed data, with $y_{n+1} = t_{n+1}$ and $\delta_{n+1} = 1$; and refit the hazard regression model. This is repeated up to a total sample size of $N = n + 500$.

To visualize predictive uncertainty, we generate a new observation $\mathbf{x}^*$ for prediction. Figure 5.6a displays the predicted survival function $1 - F(t|\mathbf{x}^*)$ under the baseline model in red, and the survival functions from 25 predictive resampling trials in black. Figure 5.6b displays the 2.5% and 97.5% error bounds across 1000 predictive resampling trials in blue, providing an uncertainty measure around the initial estimate.

Chapter 6

Targeting relative risk heterogeneity with causal forests

6.1 Introduction

In clinical settings, we are often interested in exploring the evidence for heterogeneous treatment effects (HTE), or determining whether specific subgroups of the population respond differently to a treatment under investigation. Over the past few decades, this question has received significant attention in the causal inference literature, as it represents a key step toward personalized medicine. Identifying HTEs is also essential for transporting and generalizing trial findings to populations that have been under-represented in clinical studies (Lipkovich et al., 2024; Sechidis et al., 2024).

More specifically, HTE discovery calls for nonparametric methods which can evaluate the dataset as a whole to identify subgroups of interest (Watson & Holmes, 2020). Classical subgroup analysis, where potentially relevant covariates are pre-specified in a clinical trial protocol, may fail to detect strong but unexpected heterogeneity, and additionally raises concerns related to multiple testing (Cook et al., 2004). For this reason, forest-based methods have become especially popular, building upon the seminal work of Breiman (2001a) to flexibly model high-dimensional interactions between covariates.

In particular, causal forests (Athey & Imbens, 2015; Athey et al., 2019; Wager & Athey, 2018) are a state-of-the-art approach for HTE estimation in real-world clinical trial analysis (Athey & Wager, 2019; Basu et al., 2018; Raghavan et al., 2022). While a standard decision tree recursively partitions the input data to maximize the variability in

an outcome, a causal decision tree instead maximizes the variability in treatment effect by separating the treatment and control samples within each node. To adjust for confounding and directly estimate HTEs, causal forests orthogonalize the outcome and treatment propensity with respect to the covariates, leveraging ideas from the double/debiased machine learning literature (Chernozhukov et al., 2018).

However, the node-splitting criterion in causal forests targets heterogeneity in the absolute Risk Difference (RD) between subgroups. In certain contexts, a preferable approach is to target heterogeneity in the relative Risk Ratio (RR). The RR has been found to extrapolate more effectively across populations (Deeks, 2002; Furukawa et al., 2002), meaning that heterogeneity in the RR may be more indicative of a true difference in treatment efficacy (Spiegelman & VanderWeele, 2017; Sun et al., 2014).

In addition, the RD over-emphasizes individuals with a high baseline risk level (Kent et al., 2010). As a result, HTE estimates based on RD could be biased towards high-risk individuals, and spuriously identify prognostic covariates that are related to baseline risk, rather than predictive covariates that indicate treatment heterogeneity. Targeting the RR can therefore improve statistical power, especially in scenarios with large variation in individual baseline risk, by weighting all individuals equally regardless of risk level.

We therefore aim to adjust the structure of causal forests such that the RR can be chosen as the quantity of interest. We present a novel method that uses exhaustive generalized linear model (GLM) comparison as the basis for the splitting rule within the forest. By fitting a GLM with an interaction term between the treatment and every possible candidate split in succession, we identify the split that induces the most significant heterogeneity in the treatment effect. Changing the link function of the GLM then allows for quantities other than the RD, such as the RR, to be targeted. Our approach is implemented as an update to the `grf` software package in `R` (Tibshirani et al., 2023), available at `https://github.com/vshirvaikar/rrcf`.

Previous partitioning approaches such as RECPAM (Ciampi et al., 1988) have implemented an exhaustive model-based search, but never in the causal setting, or with a focus on RR. We additionally note the recent body of work on "model-based forests", which

identify the optimal split by fitting a model within each node of a decision tree (Seibold et al., 2016; Zeileis et al., 2008). However, the mechanism and motivation of this approach are fundamentally different: model-based forests fit a single model within each node in order to simultaneously estimate prognostic and predictive effects, while we fit several models within each node for the purpose of identifying the optimal split.

The remainder of this chapter is organized as follows. We present the problem setting in Section 6.2, and motivate the importance of relative risk in Section 6.3. In Section 6.4, we review existing forest-based methods for HTE estimation, and in Section 6.5, we present our proposed methodology and demonstrate how it fits into the causal forest framework. Section 6.6 validates our approach on simulated data, and Section 6.7 concludes.

6.2 Problem setting

Consider a clinical trial with n subjects, each with covariates X_i and a binary treatment assignment indicator $W_i \in \{0, 1\}$ for $i = 1, \ldots, n$. In the current work, we focus on the binary outcome setting with $Y_i \in \{0, 1\}$. To estimate the causal effect of the treatment on the outcome, we adopt the potential outcomes framework of Rubin (2005). Under this framework, each subject has potential outcomes $Y_i^{(1)}$, representing the outcome if the subject were treated ($W_i = 1$), and $Y_i^{(0)}$, representing the outcome if the subject were untreated ($W_i = 0$). The fundamental challenge in causal inference is that for each subject, we can only observe one of these outcomes depending on treatment assignment,

$$Y_i = W_i Y_i^{(1)} + (1 - W_i) Y_i^{(0)}.$$

To estimate causal effects from observed data, we must therefore introduce certain key assumptions. The Stable Unit Treatment Value Assumption states that there is no hidden variation in the treatment W, and no interference between the treatment outcomes for different subjects. Ultimately, this implies that each subject's outcome depends only on their own treatment status $\left[Y_i = Y_i^{(W_i)}\right]$. We additionally assume ignorability or unconfoundedness, meaning that treatment assignment is independent of potential outcomes given the observed covariates $\left[(Y_i^{(1)}, Y_i^{(0)}) \perp\!\!\!\perp W_i \mid X_i\right]$. Finally, we assume

positivity or overlap, where each subject has a nonzero probability of receiving either treatment $[0 < \mathbb{P}(W_i = 1 \mid X_i) < 1 \ \ \forall X_i]$.

Together, these assumptions allow us to identify various causal estimands for our treatment effect. Note that in the binary setting, expected outcomes can be interpreted as probabilities $\mathbb{E}\left[Y_i^{(W_i)}\right] = \mathbb{P}\left[Y_i^{(W_i)} = 1\right]$. Two common choices for outcome measure are the RD, which measures the absolute change in event probability due to treatment,

$$\tau_{RD} = \mathbb{E}\left[Y_i^{(1)} - Y_i^{(0)}\right] = \mathbb{P}\left[Y_i^{(1)} = 1\right] - \mathbb{P}\left[Y_i^{(0)} = 1\right],$$

and the RR, which measures the relative change in event probability between the treated and untreated groups,

$$\tau_{RR} = \frac{\mathbb{E}\left[Y_i^{(1)}\right]}{\mathbb{E}\left[Y_i^{(0)}\right]} = \frac{\mathbb{P}\left[Y_i^{(1)} = 1\right]}{\mathbb{P}\left[Y_i^{(0)} = 1\right]}.$$

Though we will not focus on them here, other common measures used in the binary setting include the Odds Ratio, which compares the odds of the event under treatment versus control and frequently appears in case-control studies; and the Number Needed to Treat, which contextualizes the RD using patient counts and is widely used in medical practice (Altman, 1999; Cook & Sackett, 1995).

The RD is also referred to as the average treatment effect (ATE) across the population. However, this does not capture variation in treatment response across different subgroups of the study population, motivating the need for HTE investigation (Rothman, 2012). A typical target quantity for HTE estimation is then the expected RD for a subject with a given set of covariates

$$\tau_{RD}(x) = \mathbb{E}[Y_i^{(1)} - Y_i^{(0)} | X_i = x]$$

or, equivalently, the conditional average treatment effect (CATE).

6.3 The importance of relative risk

In many practical applications, there are compelling reasons to prefer HTE estimation on a relative rather than an absolute scale, or to evaluate both in parallel. Instead of the

CATE, the relevant estimand becomes the expected RR given covariates

$$\tau_{RR}(x) = \frac{\mathbb{E}[Y_i^{(1)}|X_i = x]}{\mathbb{E}[Y_i^{(0)}|X_i = x]},$$

which has been referred to as the conditional Risk Ratio (Wang et al., 2016).

Relative effect measures such as the RR are often preferred due to their perceived stability across settings. Multiple meta-analyses of clinical trials have found that RR reductions extrapolate most reliably across different time periods and populations (Deeks, 2002; Furukawa et al., 2002; Schmid et al., 1998). As a result, it has been argued that heterogeneity in the RR is more indicative of true underlying effect modification than heterogeneity in the RD or OR (Spiegelman & VanderWeele, 2017; Sun et al., 2014). We note that certain contrasting empirical settings have been highlighted, in which RD or OR estimates are likely more stable (Doi et al., 2022; Poole et al., 2015).

The RR also possesses several desirable theoretical properties. Under a range of plausible causal models, Huitfeldt et al. (2022) and Colnet et al. (2023) show that the RR remains stable between patient groups, provided that the standard RR is used if the treatment is beneficial, and the Survival Ratio (the RR reversed to count null events) is used if the treatment is harmful. This formulation is equivalent to the Switch Relative Risk of Van Der Laan et al. (2007), and addresses long-standing concerns about the asymmetry of the RR first discussed in Sheps (1958). Stensrud and Smith (2023) find that the RR is stable when exposure is partially unobserved, while Piccininni and Stensrud (2025) find that the RR is stable when immune individuals are excluded from a study.

The RR may present advantages in terms of explanation and interpretability. Relative effects often better align with how patients and practitioners perceive risk (Schechtman, 2002; Simon, 2001). For example, expressions such as "five times higher" can be more intuitively meaningful than statements like "a 4% increase," particularly when the outcome of interest is rare. In applied settings, it is generally recommended that relative and absolute effect measures are reported in parallel, in order to provide context and support accurate interpretation (Colnet et al., 2023; Noordzij et al., 2017).

Finally, RR-based estimands can be more appropriate in settings with substantial variation in baseline risk. For instance, Watson and Holmes (2020) present a malaria

case study from the AQUAMAT clinical trial where patients' predicted baseline mortality risk varied from less than 1% to greater than 80%. In such cases, estimates based on differences in RD are often dominated by covariates associated with overall mortality. Individuals with a high pre-treatment risk may be "overrepresented" in the final results, since a risk reduction from 40% to 10% (for example) is considered ten times as important as a reduction from 4% to 1%. While this weighting is appropriate when estimating a population-average effect, it may dilute statistical power for HTE detection by discounting individuals with a low pre-treatment risk level. Estimates based on RR would weight these reductions equally, potentially leading to a more balanced detection of subgroup differences across the covariate distribution.

Together, these motivations help to explain the widespread use of the RR in clinical research. A survey of 100 clinical trials published in the *New England Journal of Medicine* between 2018 and 2020 found that the RR was employed approximately twice as often as the RD to report subgroup effects (Andersen, 2021). Despite this, state-of-the-art HTE estimation methods, such as causal forests, are still primarily designed to target heterogeneity on an absolute scale. Given the prominent role of the RR in medical decision-making, we aim to develop an HTE estimation method that specifically operates on a relative scale. In the following sections, we review existing forest-based approaches to HTE estimation, then describe our modifications for targeting relative treatment effects.

6.4 Review of existing approaches

Tree-based methods non-parametrically partition the covariate space to adaptively model complex interactions, making them particularly well-suited for detecting HTEs, especially in high-dimensional settings. Classical decision trees and random forests (Breiman, 2001a; Breiman et al., 1984) are widely used in supervised learning, where the partitioning process aims to optimize a variance-based loss function (for regression) or a Gini impurity criterion (for classification) with respect to the outcome.

However, HTE estimation requires a different objective: rather than identifying covariates which explain variability in the outcome itself, the goal is to identify covariates

which explain variability in the treatment effect, or in other words, *differences in the differences* between the treated and untreated groups. Various modifications to recursive partitioning have been developed to achieve this; we discuss two key frameworks below.

6.4.1 Causal forests

Causal forests (Athey & Imbens, 2015; Wager & Athey, 2018) take a direct approach, specifying the ATE as the quantity of interest within a random forest framework. Suppose that we are evaluating whether to split on variable Z at value z at a given node in a decision tree. For some target quantity ξ, we would calculate the reduction in variance from the parent node to the resulting children, choosing the split that maximizes

$$\Delta V = \text{Var}(\xi_i) - p_L \cdot \text{Var}(\xi_i \mid Z_i < z) - p_R \cdot \text{Var}(\xi_i \mid Z_i > z)$$

where $p_L = \mathbb{P}(Z_i < z)$ and $p_R = \mathbb{P}(Z_i > z)$ denote the proportions of observations in the left and right child nodes, respectively. In a classical regression tree, the quantity of interest is the outcome $\xi_i = Y_i$, but in a causal tree, it is instead the treatment effect τ_{RD}.

However, individual-level treatment effects $\tau_{RD}(X_i)$ cannot be directly observed, so causal forests instead rely on proxy estimates using average group-level differences. Intuitively, this can be understood as maintaining a separation between the treated and control samples in each node, enabling the tree structure to capture HTEs through local differences in group means. Wager and Athey (2018) describe this as a form of data-driven population stratification, where each leaf "acts as though [it] had come from a randomized experiment" restricted to individuals within that particular subgroup.

To ensure valid inference and prevent overfitting in the resulting tree, the key detail is a sample-splitting procedure known as honesty, where the data is randomly partitioned into two subsamples: one to construct the tree (i.e., determine splits), and another to then estimate treatment effects within the terminal leaves. The major theoretical results of Athey et al. (2019), including consistency and asymptotic normality, rely on honesty along with several additional structural constraints, as specified below.

Specification 1 (Athey et al., 2019)**.** *All trees are symmetric, in that their output is invariant to permuting the indices of training examples; make balanced splits, in the sense*

that every split puts at least a fraction ω of the observations in the parent node into each child, for some $\omega > 0$; and are randomized in such a way that, at every split, the probability that the tree splits on the j-th feature is bounded from below by some $\pi > 0$. The forest is honest and built via subsampling with subsample size s satisfying $s/n \to 0$ and $s \to \infty$.

Together, these conditions ensure the resulting forest is an ensemble of learners that are both sufficiently diverse (due to randomization and subsampling) and stable (due to honesty and balanced splits), allowing for valid pointwise confidence intervals.

To improve the empirical performance of causal forests, Athey et al. (2019) incorporate local centering through the double/debiased machine learning (DML) framework of Chernozhukov et al. (2018). This approach yields smoother and more robust estimates by separately regressing out the influence of the baseline covariates on both the treatment and the outcome. In practice, this is implemented by first fitting regression forests on $Y \sim X$ and $W \sim X$ to estimate the baseline risk $\widehat{Y} = \mathbb{E}[Y|X]$ and propensity score $\widehat{W} = \mathbb{E}[W|X]$. The main causal forest algorithm is then applied to the residualized outcome $\widetilde{Y} = Y - \widehat{Y}$ and residualized treatment $\widetilde{W} = W - \widehat{W}$.

An important theoretical insight underlying DML is the concept of Neyman orthogonality, which ensures that the final treatment effect estimate is insensitive to small misspecifications in the initial regression models. In the context of causal forests, the baseline risk and propensity score forests are not of direct inferential interest, so they are considered nuisance functions, collectively denoted by $\eta = \{\widehat{Y}, \widehat{W}\}$. For a moment function $\psi(W, Y, X; \tau, \eta)$ which defines some condition for estimating the treatment effect τ, Neyman orthogonality refers to the additional requirement that

$$\frac{\partial}{\partial \eta} \mathbb{E}[\psi(W, Y, X; \tau, \eta)]\Big|_{\eta=\eta_0} = 0,$$

meaning that small perturbations in η around the true value η_0 do not affect the first-order expectation of the score function. In other words, the estimator for τ is doubly robust: it remains consistent even if one of the nuisance functions is misspecified, provided the other is accurately estimated. This property underpins the stability of orthogonalized causal forest estimates in high-dimensional or flexible settings.

In recent years, causal forests have gained widespread adoption across the biomedical and social sciences (Athey & Wager, 2019; Davis & Heller, 2017; Raghavan et al., 2022). The current state-of-the-art implementation is the generalized random forests (`grf`) package in `R` (Tibshirani et al., 2023), which extends forest-based inference to a broad class of statistical problems, including HTE estimation. Another popular extension is Bayesian causal forests (Hahn et al., 2020), which leverages the benefits of Bayesian regularization and shrinkage.

6.4.2 Model-based forests

Meanwhile, model-based forests (Seibold et al., 2016; Zeileis et al., 2008) explicitly specify a parametric model with the outcome as a function of the treatment and covariates

$$Y_i = \mu(X_i) + W_i\tau(X_i) + \epsilon_i.$$

In this setting, $\mu(X)$ represents the prognostic effect of baseline covariates that directly impact the outcome, while $\tau(X)$ is the predictive effect of covariates that influence treatment efficacy. Any given covariate can be both prognostic and predictive.

For this model, define an objective function $\Psi((Y, X), \theta)$, such as the negative log-likelihood, with respect to parameters $\theta = (\mu, \tau)$. The model minimizes this function,

$$\arg\min_{\theta} \sum_{i=1}^{n} \Psi((y_i, x_i), \theta),$$

or equivalently solves the score equation

$$\sum_{i=1}^{n} \frac{\partial \Psi((y_i, x_i), \theta)}{\partial \theta} = \sum_{i=1}^{n} \psi((y_i, x_i), \theta) = 0$$

where ψ is the score function, the gradient of the objective function with respect to the parameters (Seibold et al., 2016).

The key insight of model-based forests is that covariates associated with heterogeneity induce instabilities in the parameter estimates, and that this can be measured directly via score functions, which quantify how much each individual's data influences the estimation of μ and τ. Specifically, define the partial score functions as

$$\psi_\mu((y, x), \theta) = \frac{\partial \Psi((y, x), \theta)}{\partial \mu} \quad \text{and} \quad \psi_\tau((y, x), \theta) = \frac{\partial \Psi((y, x), \theta)}{\partial \tau}.$$

If the true baseline effect μ is constant across individuals, then ψ_μ should be independent of any covariates; similarly, if the true treatment effect τ is homogeneous, then ψ_τ should be independent of all partitioning variables.

To detect covariates associated with heterogeneity, two hypothesis tests of independence can therefore be conducted for each candidate split variable Z_j (where the splitting variables Z and model variables X are often the same, but are not required to be)

$$H^0_{\mu,j} : \psi_\mu((y,x),\hat{\theta}) \perp Z_j, \quad j = 1,\ldots,J$$
$$H^0_{\tau,j} : \psi_\tau((y,x),\hat{\theta}) \perp Z_j, \quad j = 1,\ldots,J.$$

Out of these $2 \times J$ tests, the partitioning variable with the smallest p-value (subject to a certain threshold) is selected, indicating the strongest dependence between the covariate and changes in the prognostic or predictive effect.

Model-based forests build upon older approaches, such as RECPAM (Ciampi et al., 1988) and GUIDE (Loh, 2002), which conduct model comparison within each node of a tree to identify the optimal splitting variable. However, these approaches have typically been applied to general regression or classification, rather than causal HTE estimation. In recent years, model-based forests have been extended to several other specific applications, including individual-level modeling (Seibold et al., 2018) and observational data integration (Dandl et al., 2024).

6.5 Relative risk causal forests

Forest-based methods are fundamentally characterized by the rule that determines whether and where to split the data at each node. In standard causal forests, this rule targets heterogeneity in the absolute risk because it aims to maximize variation in $\tau_{RD}(X_i)$. To instead target heterogeneity in the relative risk, we modify this splitting criterion to focus on variation in $\tau_{RR}(X_i)$. We begin by proposing a general alternative node-splitting procedure based on exhaustive GLM comparison, and then specialize it to the relative risk setting. Our implementation modifies the open-source `grf` package in `R`, and is available at `https://github.com/vshirvaikar/rrcf`.

6.5.1 Forest construction

Recall that at any given parent node, we observe a set of outcomes Y_i, covariates X_i, and binary treatment indicators $W_i \in \{0, 1\}$. To evaluate a candidate split on covariate Z at threshold z, we first define a binary split indicator $S_i = \mathbb{1}\{Z_i > z\}$ which denotes whether observation i would fall to the left or right of the proposed split. We then fit the GLM

$$Y_i \sim X_i + W_i + S_i + W_i \cdot S_i. \tag{6.1}$$

Here, the X_i terms adjust for baseline covariate effects, W_i controls for the average treatment effect throughout the parent node, and S_i controls for the main effect of the candidate split. The interaction term $W_i \cdot S_i$ then captures the difference in treatment effects between the two sides of the split — in other words, the degree of treatment effect heterogeneity induced by the partition. We repeat this procedure for each candidate split, and select the one with the smallest p-value (or most extreme test statistic) on the interaction coefficient. While this approach is computationally intensive, it is easily parallelizable, and designed specifically for long-term clinical trial analysis where runtime is not a primary concern.

The motivation for GLM-based splitting is that we can target different treatment effect estimands by toggling the assumed response distribution and link function. For example, if we use linear regression (Gaussian distribution with an identity link), the estimated effect corresponds to an absolute risk difference. If we use logistic regression (binomial distribution with a logit link), the estimated effect represents a log-odds ratio.

To target the relative risk ratio, we require a model with a log link function, as this directly models multiplicative effects on the outcome scale. For a binary outcome, this leads to two primary candidates: log-binomial regression (with a binomial likelihood) and Poisson regression (with a Poisson likelihood). Log-binomial regression is theoretically more appropriate for binary data, as it ensures that fitted probabilities remain within the unit interval. However, log-binomial models suffer from convergence issues, especially in small samples or near the boundaries.

Meanwhile, Poisson regression is typically used for count data, but enjoys stable estimation due to its canonical link function. In comparative studies, it has been shown to perform similarly to the log-binomial approach, providing a valid approximation for binary outcomes when the focus is on estimating relative risks (Chen et al., 2018; Petersen & Deddens, 2008). We therefore adopt Poisson regression, which specifically models the response as

$$Y_i \sim \text{Poisson}(\lambda_i), \text{ with } \log(\lambda_i) = \beta_0 + \sum_k \beta_k x_{ik}.$$

In this setting, λ_i represents the expected value of Y_i and can be interpreted as an intensity parameter, analogous to a hazard function in survival analysis, capturing the instantaneous rate at which the binary outcome transitions from 0 to 1.

We incorporate additional adjustments to mirror the orthogonalization strategy used in standard causal forests. In classical DML, both the outcome and treatment variables are residualized to remove variation explained by baseline covariates (Chernozhukov et al., 2018). However, our outcome variable must remain non-negative and integer-valued for Poisson regression, so we cannot use $\widetilde{Y} = Y - \widehat{Y}$ in Equation 6.1. Instead, we still estimate a baseline risk model of the form $Y \sim X$, but then use the linear predictor

$$\hat{\nu} = \log(\widehat{Y}) = \log(\mathbb{E}[Y \mid X])$$

to replace X in Equation 6.1, yielding a univariate adjustment for baseline outcome risk under the log link. While the functional form of this model is flexible, we use Poisson regression and extract the fitted linear predictor as $\hat{\nu} = X\hat{\beta}$. In the resulting GLM, $\hat{\nu}$ typically has a coefficient close to one, reflecting its role in capturing the background log-risk. This substitution also provides a computational benefit: it reduces the dimensionality of the repeated GLM fits from $p + 3$ covariates to exactly four.

In the case of randomized controlled trial (RCT) data, where treatment is unconfounded by design, the adjusted model becomes

$$Y_i \sim \hat{\nu}_i + W_i + S_i + W_i \cdot S_i.$$

In observational or mixed designs, however, additional adjustment is required to mitigate confounding. Following the application of DML used in causal forests, we estimate the

propensity score via a regression forest on $W \sim X$, yielding $\widehat{W} = \mathbb{E}[W \mid X]$, and replace the treatment indicator with the residualized treatment $\widetilde{W} = W - \widehat{W}$. The final model used for evaluating candidate splits in such settings is then

$$Y_i \sim \hat{\nu}_i + \widetilde{W}_i + S_i + \widetilde{W}_i \cdot S_i.$$

We implement our modified splitting criterion within the `grf` package in `R`, which relies on a `C++` backend for computational efficiency and scalability via the `Rcpp` framework (Eddelbuettel & François, 2011). The GLM fits are performed in `C++` using iteratively reweighted least squares (IRLS), a standard algorithm for maximum likelihood estimation in GLMs. IRLS solves a sequence of weighted least squares problems, where both the weights and the working response are updated based on the current parameter estimates. In our implementation, we use Householder QR decomposition at each step for improved numerical stability. Convergence is assessed via the L_2 norm of successive coefficient updates, and we terminate early if convergence stalls or fails. Upon convergence, we extract the final coefficient vector and its estimated covariance matrix to compute standard errors, and select the candidate split with the largest absolute t-statistic for the treatment-split interaction term. If no candidate split yields a converged model, we stop splitting. This commonly occurs in small subsamples or when outcomes are highly imbalanced, mirroring standard stopping behavior in classical random forests.

Because our method is implemented within the existing `grf` framework, it inherits the key structural properties described in Specification 1, including honesty and subsampling. We also retain the randomized selection of candidate splitting variables at each node, where the number of features considered is drawn from $\min\{\max\{\text{Poisson}(m), 1\}, k\}$ with tuning parameter $m > 0$, ensuring that every feature has a strictly positive probability of being selected (Denil et al., 2014). The GLM-based splitting rule is symmetric (invariant to the ordering of training observations) and balanced (ensuring that each split places a nonzero fraction of samples from the parent node into each child). As a result, the theoretical guarantees established by Athey and Wager (2019) under Specification 1, including consistency and asymptotic normality, continue to apply to our method.

6.5.2 Treatment effect estimation

With a trained forest in hand, we now consider how to estimate the treatment effect at a new test point x. Classical random forests frame prediction as an ensemble procedure: for each tree, the test point is passed down to a terminal leaf, the average outcome of training points in that leaf is computed, and the leaf-level means are aggregated across trees. This yields the estimate

$$\hat{\mu}(x) = \frac{1}{B} \sum_{b=1}^{B} \left(\frac{\sum_{X_i \in L_b(x)} Y_i}{\#\{X_i \in L_b(x)\}} \right),$$

where $L_b(x)$ denotes the set of training samples that fall into the same leaf as x in tree b, and B is the total number of trees.

However, computational implementations of causal forests, including the `grf` package, recast this aggregation step as an adaptive nearest-neighbor procedure (Tibshirani et al., 2023). Rather than computing tree-level predictions, these methods directly assign a weight to each training point based on how frequently it co-occurs in a leaf with x across the forest. Let the resulting forest weights be denoted by $\alpha_i(x)$, defined as

$$\alpha_i(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{1}(X_i \in L_b(x))}{\#\{X_i \in L_b(x)\}},$$

where $\sum_{i=1}^{n} \alpha_i(x) = 1$. This formulation defines a kernel centered at x, allowing the forest to be interpreted as a locally weighted estimator over training examples. For causal applications, this enables smoother and more stable estimates of treatment effects, particularly when leaf sizes or class proportions vary across trees.

The adaptive weighting setup enables efficient computation of final estimates using pre-computed forest-wide statistics. In causal forests, this takes the form of a two-stage least squares estimate, motivated by the equivalence between causal and instrumental forests where the treatment assignment vector serves as the instrument (Athey et al., 2019). However, this cannot be directly extended to the relative risk setting, as the estimand is multiplicative rather than additive. Instead, we use the forest weights $\alpha_i(x)$ to compute weighted averages of outcomes within each treatment group. Specifically, for

each test point x, we estimate the conditional risk under treatment and control as

$$\hat{\mu}_1(x) = \frac{\sum_{W_i=1} \alpha_i(x) Y_i}{\sum_{W_i=1} \alpha_i(x)}, \quad \hat{\mu}_0(x) = \frac{\sum_{W_i=0} \alpha_i(x) Y_i}{\sum_{W_i=0} \alpha_i(x)}$$

The absolute risk estimate returned by a standard causal forest is then equivalent to

$$\hat{\tau}_{RD}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

and so we can define the relative risk estimate as

$$\hat{\tau}_{RR}(x) = \frac{\hat{\mu}_1(x)}{\hat{\mu}_0(x)}$$

6.5.3 Omnibus testing

Our final step is to evaluate whether the forest's individual-level relative treatment effect predictions $\hat{\tau}_{RR}(X_i)$ are well-calibrated and statistically significant. To this end, we develop an omnibus test for the detection of overall heterogeneity, extending the calibration test proposed by Tibshirani et al. (2023) to the multiplicative setting.

The `grf` package assesses the overall quality of a forest via a simple linear fit on held-out data. Let $\widetilde{Y}$ and $\widetilde{W}$ denote the residualized outcome and treatment vectors as before, and let $\hat{\tau}_{RD}(X_i)$ represent the individual predicted absolute treatment effects. Define $\bar{\tau}_{RD}(X) = \mathbb{E}[\hat{\tau}_{RD}(X_i)]$ as the mean prediction across the test set. The `grf` calibration test fits the linear model

$$\widetilde{Y}_i \sim \alpha\, \widetilde{W}_i\, \bar{\tau}_{RD}(X) + \beta\, \widetilde{W}_i\, \left(\hat{\tau}_{RD}(X_i) - \bar{\tau}_{RD}(X)\right), \tag{6.2}$$

where α indicates whether the mean forest prediction is centered and β indicates whether the individual HTE estimates are well-calibrated. A "correct" forest would have $\alpha = \beta = 1$, and Equation 6.2 would reduce to $\widetilde{Y}_i = \widetilde{W}_i \hat{\tau}_{RD}(X_i)$. The p-value on β is then used as an omnibus test for absolute heterogeneity: a small p-value indicates that the individual predictions contain significant explanatory signal beyond the global average.

In the relative risk setting, we analogously predict $\hat{\tau}_{RR}(X_i)$ for each individual in a held-out test set. We begin with the baseline model $Y \sim X + W$, and evaluate the added contribution of an interaction term $W \cdot \log(\hat{\tau}_{RR}(X))$. This yields the model

$$Y_i \sim X_i + W_i + W_i \cdot \log(\hat{\tau}_{RR}(X_i)), \tag{6.3}$$

where the final term becomes exactly $\hat{\tau}_{RR}(X_i)$ for the treatment group and zero for the control group under the log link. The p-value on the final term therefore serves as an omnibus test for relative heterogeneity in the relative treatment effect: it measures whether the individual-level predictions provide significant improvement over a model with only main effects for the covariates and treatment.

6.6 Simulation study

We conduct a simulation study to assess the effectiveness of the relative risk causal forest in detecting HTEs, and to compare its statistical power against the `grf` baseline.

6.6.1 Data generation

For a meaningful simulation study with a known true treatment effect, it is critical to simulate data from a correctly specified data-generating process. While this is relatively straightforward in the case of additive treatment effects, greater care is required when modeling potential outcomes on a relative scale. Following Lin et al. (2024), we adopt the frugal parameterization proposed by Evans and Didelez (2024) to define and simulate from marginal structural models. This framework allows for precise control over the HTE structure by separately specifying the key components of the joint distribution. Specifically, the joint distribution of covariates, treatment, and potential outcomes is decomposed into three distinct components: (1) the marginal causal quantity of interest, i.e., the HTE function; (2) the joint distribution of the treatment and covariates; and (3) the dependence between the outcome and covariates conditional upon the treatment.

We simulate seven covariates in total. X_1, X_2, and X_3 are prognostic for the outcome Y but do not modify the treatment effect, while C_1, C_2, C_3, and C_4 are predictive effect modifiers. These variables are a mix of discrete and continuous as specified below.

$$X_1 \sim N(0,1) \qquad X_2 \sim \text{Gamma}(0.1 + 0.2X_1, 1) \qquad X_3 \sim \text{Beta}(0.1 + 0.1X_1, 1)$$

$$C_1 \sim \text{Bernoulli}(0.5) \qquad C_2 \sim \text{Bernoulli}(\text{expit}(-2 + C_1))$$

$$C_3 \sim N(0.1C_1C_2, 1) \qquad C_4 \sim t_{20}(0.1C_1, 0.1)$$

A pair-copula construction is used to flexibly encode the dependency between the non-modifier covariates and the outcome. Specifically, we employ a Gaussian copula to model the $X_1 - Y$ dependency with correlation depending on C_2; a Clayton copula to capture higher-tail dependency for $X_2 - Y$; and a Gumbel copula to introduce lower-tail dependency for $X_3 - Y$. We simulate these complex distributions to better reflect real-world scenarios, where data often deviates from simple Gaussian and linear models.

We consider both an RCT setting, where treatment is assigned independently of covariates, and an observational setting, where treatment depends on a subset of the covariates. In the RCT setup, treatment is simulated as $W \sim \text{Bernoulli}(0.5)$. In the observational setup, we simulate $W \sim \text{Bernoulli}(\text{expit}(-1 + 2X_1 + 2X_3 - C_2))$,

Potential outcomes are drawn from the marginal $Y(w) \sim \text{Bernoulli}(\mu_y)$, where

$$\log(\mu_y) = -2 + 0.3C_1 + 0.4\sin(C_4) + W\left\{-0.2 + \rho\left(C_1 + C_2 + \mathbb{I}(C_3 > 0) + C_4^2\right)\right\}$$

This results in a CATE given by

$$\tau_{RR}(\boldsymbol{c}) = \exp\left\{-0.2 + \rho\left(C_1 + C_2 + \mathbb{I}(C_3 > 0) + C_4^2\right)\right\}$$

where ρ controls the degree of heterogeneity. When $\rho = 0$, there is no heterogeneity and the treatment effect is homogeneous at a factor of $\exp(-0.2) \approx 0.819$. We vary ρ across $\{0, 0.25, 0.5, 0.75\}$, resulting in a progressively stronger signal.

6.6.2 Results

We compare the methods across 100 random seeds, for a range of sample sizes and heterogeneity levels, with forests comprised of 500 trees. We first report statistical power, defined as the proportion of simulations in which the omnibus test yields a p-value below 0.05, with each forest trained on 80% of the given sample size and tested on the other 20%. We also report the average variable importance assigned to the true predictive covariates (C_1 to C_4) in the forest, as a measure of whether each method successfully identifies the relevant sources of heterogeneity. Variable importance is computed using the built-in function in the `grf` package, which weights variables based on how frequently

they are used for splits, adjusted for their depth in the tree. Complete numerical results are provided in Table A.1, located in the appendix.

In the RCT setting, Figure 6.1 displays the power of each method, while Figure 6.2 shows the average importance assigned to the true predictive variables. As expected, performance improves with increasing sample size and heterogeneity level. Across both metrics, the relative risk approach consistently outperforms the baseline causal forest. In non-null simulations, the relative risk causal forest achieves an average increase in power of 5.2%, and assigns 5.5% more importance to the true effect modifiers.

In the observational setting, Figure 6.3 displays power, while Figure 6.4 compares key variable importance. The relative risk approach again uniformly outperforms the baseline causal forest in both metrics. The difference in performance is larger than in the RCT experiment, with an average power increase of 10.3% in non-null simulations, and 9.5% more importance assigned to the true effect modifiers. These results suggest that the relative risk forest is able to effectively adjust for differences in treatment propensity, making it a promising candidate for observational data applications.

6.7 Conclusion

In this study, we present an adaptation of causal forests that specifically targets heterogeneity in the relative risk. The RR is an important clinical measure for several reasons, including its ability to generalize across populations. HTE discovery methods that only target the absolute risk may overlook critical sources of heterogeneity on the RR scale, especially in settings with large variation in individual baseline risk. The modification is based on an alternative splitting rule for the data in a causal forest that uses exhaustive GLM comparison. Specifying Poisson regression as the GLM of choice allows the RR to be targeted as the quantity of interest. Validation on simulated data suggests that the RR adjustment can improve the power of causal forests to identify heterogeneity.

Next steps for this project will focus on robust uncertainty quantification for our relative risk predictions. HTE discovery faces the fundamental challenge that true individual treatment effects are never observed, due to the missing counterfactual data, and so we

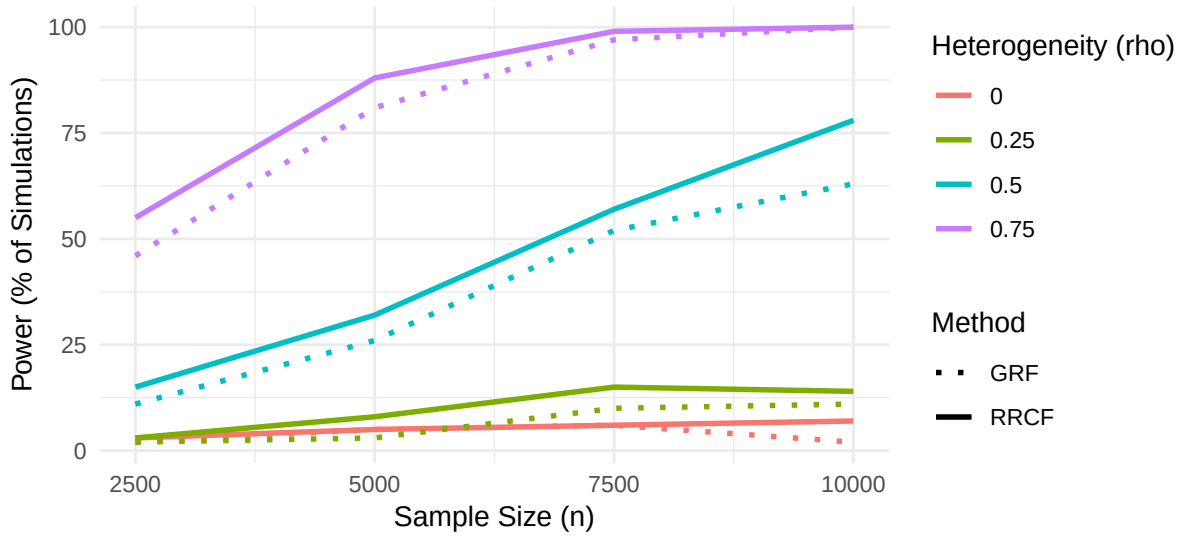


Figure 6.1: Power (proportion of trials where omnibus test p-value on additional $W_i \log(\hat{\tau}_{RR}(X_i))$ term was significant) across 100 RCT simulations, as a function of sample size n and heterogeneity level ρ.

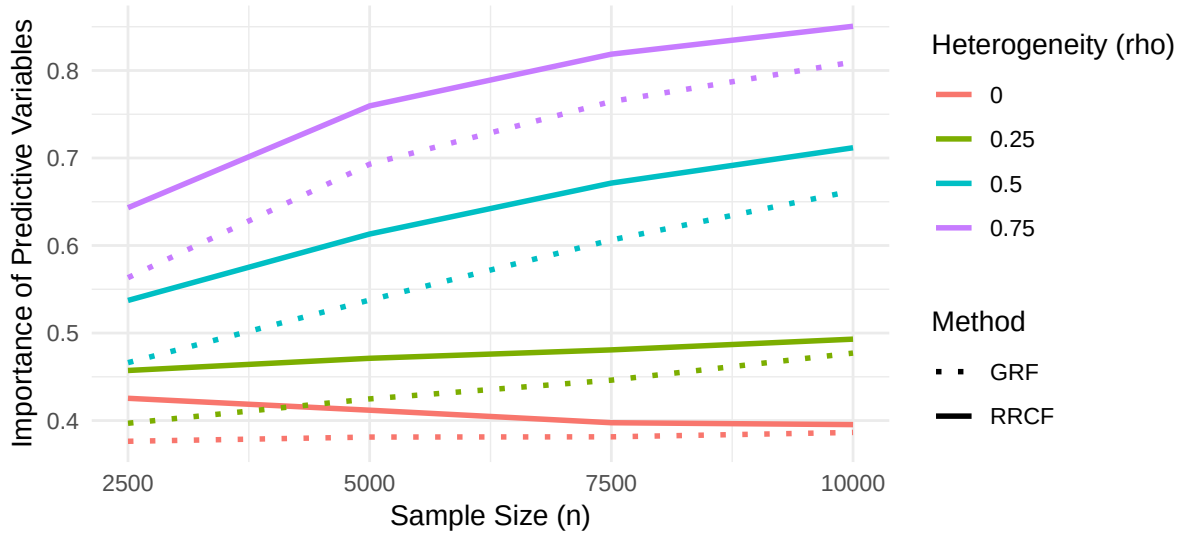


Figure 6.2: Average variable importance assigned to true predictive covariates (C_1 to C_4) across 100 RCT simulations, as a function of sample size n and heterogeneity level ρ.

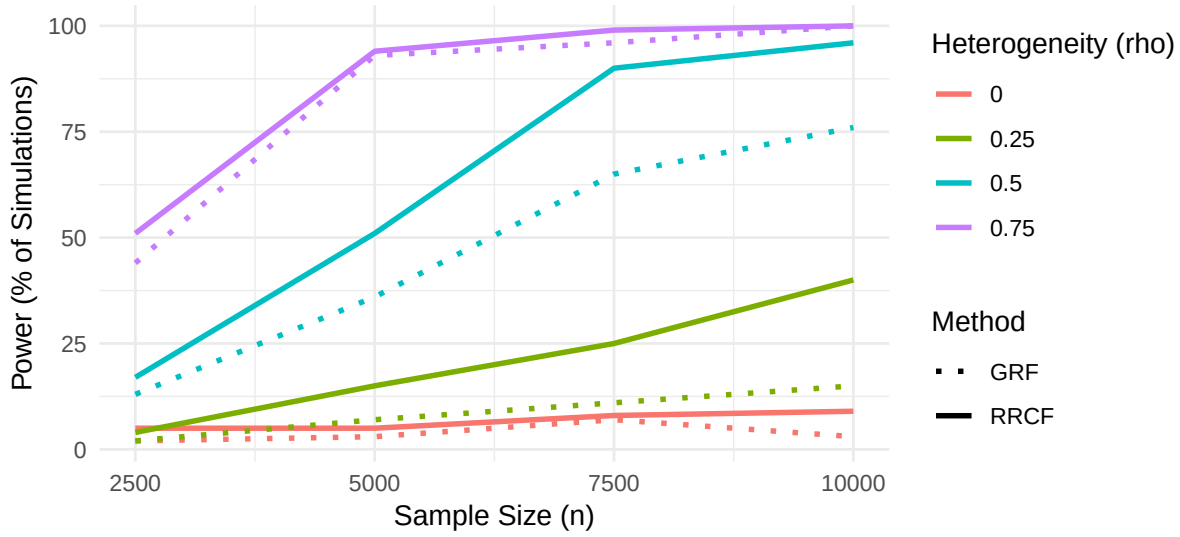


Figure 6.3: Power (proportion of trials where omnibus test p-value on additional $W_i \log(\hat{\tau}_{RR}(X_i))$ term was significant) across 100 observational simulations, as a function of sample size n and heterogeneity level ρ.

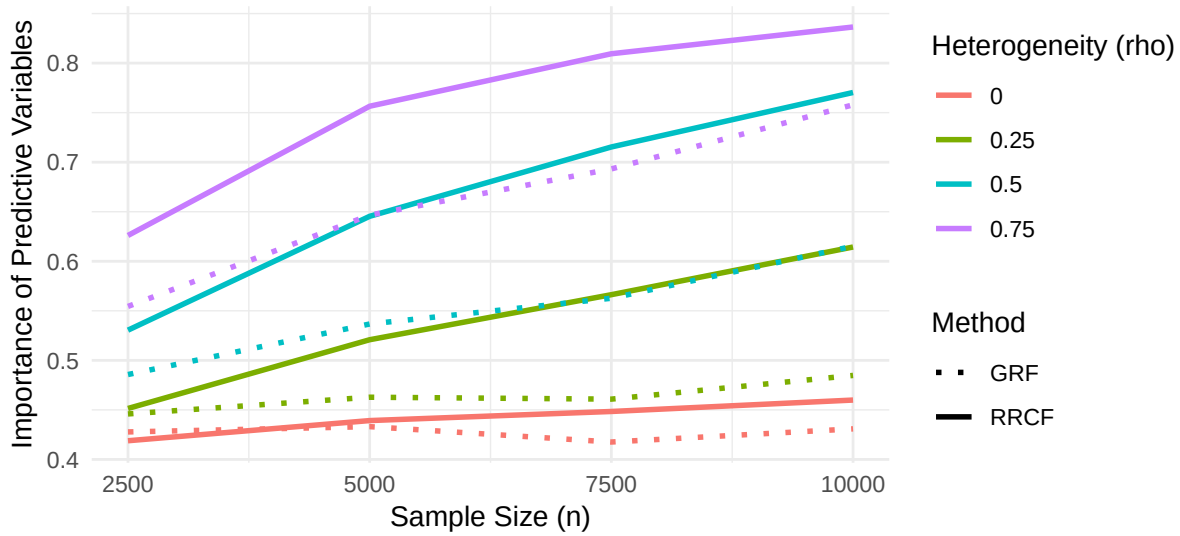


Figure 6.4: Average variable importance assigned to true predictive covariates (C_1 to C_4) across 100 observational simulations, as a function of sample size n and heterogeneity level ρ.

cannot rely on standard methods such as cross-validation to assess variability in our estimates. Instead, causal forest implementations typically turn to the "bootstrap of little bags", using the empirical variance across subsamples to approximate standard errors and confidence intervals. We plan to apply this same approach to our relative risk forest, and test its coverage and calibration in simulation, with the ultimate goal of providing transparent measures of uncertainty.

Chapter 7

Exploring relative risk heterogeneity in the LEADER clinical trial

7.1 Introduction

In this chapter, we apply relative risk causal forests to the analysis of real-world data from the Liraglutide Effect and Action in Diabetes: Evaluation of Cardiovascular Outcome Results (LEADER) clinical trial. We first screen endpoints based on variation in baseline risk to identify the most promising candidates for further analysis. The relative risk approach uncovers significant heterogeneity in one particular outcome, indicating composite heart failure or death. To translate the forest's estimates into clinically actionable insights, we then conduct a downstream analysis of the identified subgroups, exploring variable differences and importances to pinpoint key drivers of heterogeneity.

7.2 Trial details

LEADER was initiated in 2010 to evaluate the benefit of liraglutide, a glucagon-like peptide 1 analogue, in the treatment of patients with type 2 diabetes (Marso et al., 2016). 9,340 patients underwent randomization with a median follow-up time of 3.8 years. The primary outcome (MACE, or major adverse cardiovascular events) occurred in significantly fewer patients in the treatment group — 13.0% compared to 14.9%, with a hazard ratio of 0.87 (95% confidence interval from 0.78 to 0.97).

From the LEADER dataset, 70 covariates were identified as potentially relevant based on guidance from collaborators at Novo Nordisk. A complete list of these covariates is

provided below. Only patients with no missing data across all covariates were used, resulting in a final analysis set of 8,750 observations, with a very similar outcome distribution to the complete data.

- **6 demographic fields:** Age, Diabetes Duration, Sex, Race, Current Smoker, Previous Smoker
- **11 baseline vital signs:** Waist Circumference, Body Mass Index, Pulse, Systolic Blood Pressure, Diastolic Blood Pressure, Hemoglobin A1C (HbA1C), High-Density Lipoprotein (HDL) Cholesterol, Low-Density Lipoprotein (LDL) Cholesterol, Total Cholesterol, Triglycerides, Serum Creatinine
- **14 lab measurements:** Alanine Aminotransferase, Amylase, Bilirubin, Calcium, Estimated Glomerular Filtration Rate (eGFR) by the Chronic Kidney Disease-Epidemiology Collaboration (CKD-EPI) Formula, eGFR by the Modification of Diet in Renal Disease (MDRD) Formula, Potassium, Triacylglycerol Lipase, Sodium, Hematocrit, Hemoglobin, Platelets, Erythrocytes, Leukocytes
- **19 medical history flags:** Antihypertensive Therapy, Myocardial Infarction, Stroke, Stroke Sensitivity Analysis, Revascularization, Carotid Stenosis on Angiography, Coronary Heart Disease, Ischaemic Heart Disease, Chronic Heart Failure, Chronic Kidney Failure, Microalbuminuria and Proteinuria, Hypertension and Left Ventricular Hypertrophy, Left Ventricular Systolic and Diastolic Dysfunction, Ankle/Brachial Index, Cardiovascular High Risk, Cardiovascular Medium Risk, Diabetic Retinopathy, Diabetic Nephropathy
- **20 concomitant medication flags:** Insulin, Metformin, Sulfonylureas, Alpha Glucosinade Inhibitors, Thiazolidinediones, Glinides, Vitamin K Antagonists, Platelet Inhibitors, Other Antihypertensives, Thiazides, Thiazide-like Diuretics, Loop Diuretics, Aldosterone Antagonists, Beta-blockers, Calcium Channel Blockers, Angiotensin-Converting Enzyme Inhibitors, Angiotensin Receptor Blockers, Statins, Other Lipid Lowering Drugs, Ezetimibe

Outcome	Event Description	Baseline Risk Variance
PRMACETM	Expanded MACE*	14.4%
HFDTHEVT	Composite heart failure or death	12.9%
MICROTM	Microvascular event	9.3%
NEPHROTM	Secondary nephropathy event	7.8%
MACEEVTM	MACE*	7.7%
FRMASATM	MACE prior to 15th visit*	7.6%
ALDTHTM	All-cause death	7.2%
PPEVENT	MACE without pause over 120 days*	5.2%
OTR30EVT	MACE within 30 days of completion*	4.9%
EXMCHFTM	Heart failure requiring hospitalization*	4.5%

Table 7.1: Baseline risk variances $\text{Var}(\hat{Y}_i)$ for the top ten outcomes from LEADER, using predicted risk from a logistic regression model. MACE refers to major adverse cardiovascular events; outcomes marked with an asterisk (*) require confirmation from an event adjudication committee.

7.3 Results

7.3.1 Baseline risk screening

Including the primary MACE endpoint, the LEADER dataset contains records for 30 total primary and secondary outcomes. Recall that HTE estimation based on relative risk can improve power when baseline risk varies substantially across individuals, since estimates based on absolute risk may overweight high-risk patients with large event probabilities. This suggests that we can identify cases where the relative risk approach could be beneficial by searching for outcomes with large variability in baseline risk.

To operationalize this idea, we conduct a screening procedure across all available outcomes $Y^{(1)}, ..., Y^{(30)}$. We fit baseline logistic regression models $Y^{(j)} \sim X$ for $j = 1, ..., 30$ using the observed covariates, compute individual risk estimates $\hat{Y}_i^{(j)}$, and calculate the empirical variance of the predicted risk $\text{Var}(\hat{Y}_i^{(j)})$ across the population. Table 7.1 lists the ten outcomes with the highest baseline risk variance. The top five outcomes were selected for further analysis, with the primary MACE outcome ranking fifth. For reference, the baseline risk of MACE across the complete population ranges from 1.2% to 62.4% with a mean of 13.9%.

Outcome	Event Description	Omnibus p-value	
		GRF	RRCF
PRMACETM	Expanded MACE*	0.724	0.082
HFDTHEVT	Composite heart failure or death	0.921	**0.027**
MICROTM	Microvascular event	0.904	0.512
NEPHROTM	Secondary nephropathy event	0.238	0.072
MACEEVTM	MACE*	0.163	0.908

Table 7.2: Cross-validated omnibus test results for the top five outcomes from LEADER, comparing p-values for absolute and relative risk causal forests. MACE refers to major adverse cardiovascular events; outcomes marked with an asterisk (*) require confirmation from an event adjudication committee. The bolded p-value (composite heart failure or death, for the relative risk causal forest) indicates the only significant finding at $\alpha = 0.05$.

7.3.2 Omnibus testing

For each of the top five outcomes, we perform five-fold cross-validation using forests composed of 2,000 trees. In each fold, 80% of the data are used to train the forest, and relative risk coefficients $\hat{\tau}_{RR}$ are predicted for the remaining 20%. The out-of-fold predictions are concatenated to yield a vector of estimates across the full dataset of 8,750 observations, and the omnibus test from Equation 6.3 is applied.

Table 7.2 summarizes the results: for the `grf` absolute risk forest, none of the five p-values indicate useful findings, but for the relative risk approach, two are suggestive at $\alpha = 0.1$, and one outcome is significant at $\alpha = 0.05$. While this does not include a multiple testing correction, guidance from collaborators at Novo Nordisk indicates the result is still of exploratory clinical interest. We focus on this outcome (HFDTHEVT, indicating composite heart failure or death) for subsequent analysis.

7.3.3 Covariate analysis

To translate our HTE findings into clinically relevant insights, a key question is which covariates or interactions drive variation in treatment effectiveness, as captured by the fitted relative risk estimates $\hat{\tau}_{RR}(X_i)$ from the causal forest. For the composite heart failure or death outcome, we focus on two subgroups: (1) the decile of patients with the greatest predicted benefit (those with $\hat{\tau}_{RR}(X_i) < 0.815$), which we refer to as the *high-benefit group*; and (2) the subset of patients for whom treatment is predicted to be

	Subgroup Mean		
Covariate	**Reference**	**High-Benefit**	**Percent Difference**
Systolic blood pressure	135	147	+8.35%
Total cholesterol	4.37	5.13	+17.3%
LDL cholesterol	2.30	2.90	+26.1%
Sex (Male=1)	0.66	0.39	-40.9%
Hemoglobin	8.53	8.02	-6.02%
	Reference	**Low-Benefit**	
Total cholesterol	4.37	3.76	-13.9%
Platelets	269	222	-17.6%
LDL cholesterol	2.30	1.89	-17.8%
Sex (Male=1)	0.66	0.85	+28.5%
Systolic blood pressure	135	128	-5.61%

Table 7.3: Covariates with the five most significant differences, measured by p-value, from the reference population to the high-benefit (top) and low-benefit (bottom) subgroups. All p-values are significant at $\alpha = 10^{-25}$.

harmful (those with $\hat{\tau}_{RR}(X_i) > 1$), comprising 7.6% of the population, which we refer to as the *low-benefit group*. The remaining 82.4% of individuals not in either subgroup are designated as the reference population.

To characterize each subgroup, we perform two-sample t-tests comparing each covariate to the reference population. Table 7.3 lists the five covariates with the most significant differences (indicated by p-value) between each subgroup and the reference. Overall, the high-benefit group appears less healthy, with elevated blood pressure and cholesterol, while the low-benefit group is comparatively healthier. However, the relationship is not explained by overall health alone. Figure 7.1 plots baseline risk estimates $\hat{Y}_i$ from the initial logistic regression against relative treatment effects $\hat{\tau}_{RR}(X_i)$ from the relative risk causal forest. There is a slight negative trend, but the correlation is weak ($\rho = -0.053$), suggesting that more complex interactions underlie treatment heterogeneity.

In addition, simulation results suggest that relative risk causal forests more effectively split on predictive covariates which drive treatment effect heterogeneity. To explore this, we compare variable importance between the absolute and relative risk approaches. Table 7.4 displays the five covariates with the largest gains and losses in importance under the relative risk model, averaged across five training folds.

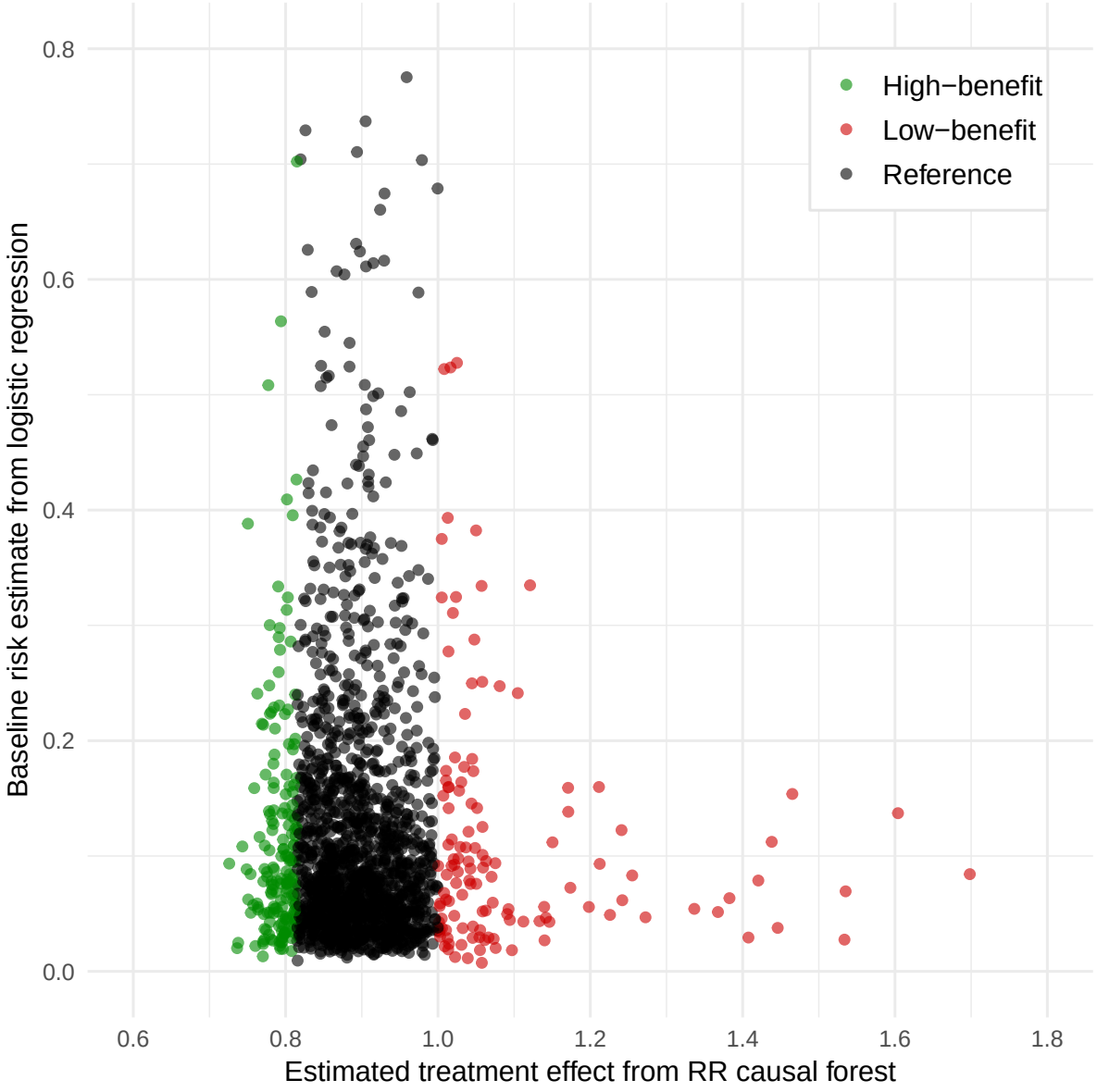


Figure 7.1: Baseline risk estimates $\hat{Y}_i$ from logistic regression against relative treatment effects $\hat{\tau}_{RR}(X_i)$ from the causal forest. Points are color-coded by subgroup.

Covariate	Importance		Difference
	GRF	RRCF	
White blood cells	3.39%	14.58%	+11.19%
Platelets	3.17%	8.01%	+4.84%
Amylase	2.27%	3.60%	+1.33%
Potassium	2.16%	2.86%	+0.70%
Sodium	0.87%	1.55%	+0.68%
eGFR by CKD-EPI	4.52%	1.87%	-2.65%
Hematocrit	4.51%	2.52%	-1.99%
Systolic blood pressure	5.62%	3.79%	-1.83%
eGFR by MDRD	3.83%	2.15%	-1.68%
Age	4.47%	2.81%	-1.66%

Table 7.4: Covariates with the five largest gains (top) and five largest losses (bottom) in variable importance between absolute and relative risk forests, averaged across five training folds. eGFR refers to estimated glomerular filtration rate, a marker of kidney function, which is estimated according to two different formulas in the provided data.

White blood cell and platelet counts show the greatest increases in importance, suggesting that blood-related biomarkers, particularly those linked to immune or inflammatory activity, play a central role in driving heterogeneity on the relative scale. However, these effects may be underrepresented in absolute risk models, possibly because their influence is stronger among individuals with lower baseline risk. Conversely, two of the five largest decreases in importance involve estimated glomerular filtration rate (eGFR), a marker of kidney function. While Marso et al. (2016) identified eGFR as a potential effect modifier, our findings suggest that its role may be overstated when heterogeneity is analyzed in absolute terms, and less relevant under a relative risk framework.

Chapter 8

Discussion and future work

In this thesis, we have explored a range of questions concerning the principled selection and specification of statistical models, through a predictive inference lens that treats missing data as the fundamental source of uncertainty. In particular, we have motivated probabilistic model uncertainty by comparing candidate models to impute data through one-step-ahead predictive updates; characterized the distribution of these updates to understand when and how models are propagating uncertainty in a consistent manner; and developed ensembling strategies to adjust causal random forests toward clinically relevant quantities of interest, with an application to real-world clinical trial data.

Methodologically, these contributions offer a framework for evaluating and understanding models entirely through the predictions they generate. Modern machine learning excels at precisely the task of rapid, high-quality prediction, and we aim to leverage this as a tool that enables principled inference and uncertainty quantification, rather than viewing the "two cultures" as orthogonal or mutually exclusive. As a further advantage, the predictive perspective may provide insights that help bridge the long-standing divide between frequentist and Bayesian approaches, by directly targeting probability distributions on quantities of interest without the need for explicit prior specification.

As outlined in previous chapters, a key direction for future work is computational improvements to make these methods more scalable and stable. This includes more efficient frameworks for recursive model updating and optimization, as well as more expressive model classes that can capture uncertainty beyond finite or closed-form settings. In conjunction with this work, a deeper understanding of the asymptotic behavior of

predictive resampling paths would also provide both theoretical and practical benefits, allowing for early stopping with provable guarantees.

In the longer term, another goal is to extend predictive resampling to a broader range of model classes and data structures. A central challenge in this direction is identifying the appropriate level of abstraction at which to sample. For example, consider graphical models, originally referred to as Bayesian networks, which represent variables and their conditional dependencies in a dataset using nodes and (optionally directed) edges. In fact, graphical models were the primary focus of early model-based MCMC approaches (Madigan et al., 1995), but direct resampling over all possible graph structures is computationally infeasible due to their super-exponential growth and highly irregular posterior landscapes.

Subsequent work has found that sampling over constrained model spaces, such as topological orderings (Friedman & Koller, 2003) or graph partitions (Kuipers & Moffa, 2017), can significantly improve uncertainty quantification, by leveraging structural assumptions to reduce the problem complexity. This suggests that predictive resampling may become feasible in more complex settings by thinking carefully about the structure and representation of the underlying models.

Appendix A

Supplementary figures and tables

A.1 Density estimation (Chapter 2)

This set of "jellyfish plots" displays empirical convergence for the density estimation illustrations in Section 2.4.1. Unlike MCMC, each individual resampling path ultimately follows a fixed trajectory, so visual inspection is typically sufficient to determine when the final empirical results have stabilized.

For the first example with two components, we track the ongoing model choice between $G = 1$ and $G = 2$ components as data points are imputed across 100 random trials. We add some jitter to each resampling trajectory to improve the visibility of the diagram. In Figures A.1 and A.2, for data with $\sigma = 0.6$ and $\sigma = 0.9$ respectively, we see that $N = n + 200$ additional observations are sufficient for the choice of model to converge.

For the second example with three components, we now plot candidate models with $G = \{1, \ldots, 9\}$ components across 100 trials. We again add some jitter to each trajectory for visibility. In Figure A.3, models with up to 9 components are explored, and that $N = n + 600$ additional observations are sufficient for $\mathcal{M}_{\widehat{k}(N)}$ to approximate $\mathcal{M}_{k(\infty)}$. In Figure A.4, model space is only explored up to $G = 4$ components, reflecting the principle that observing more initial data should tighten our final uncertainty estimate.

Finally, for the real-world galaxies dataset, we again plot candidate models with $G = \{1, \ldots, 9\}$ components across 100 trials, with some jitter for visibility. Figure A.5 shows that $N = n + 1500$ additional observations are sufficient for convergence.

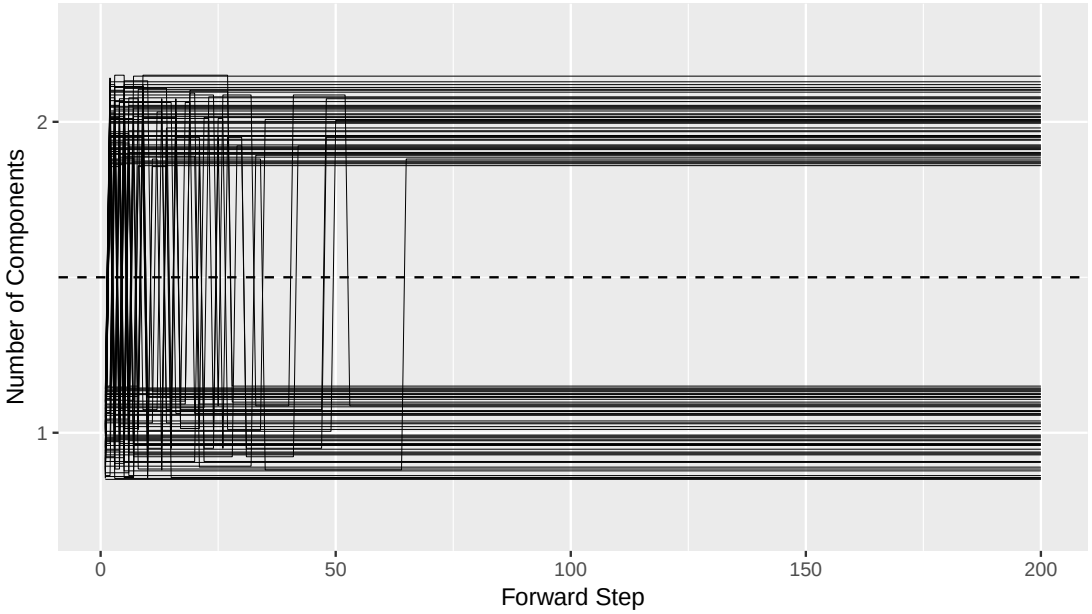


Figure A.1: Sample trajectory diagram for density estimation in two-component GMM with $\sigma = 0.6$, showing that model choice converges after $N = n+200$ imputed observations.

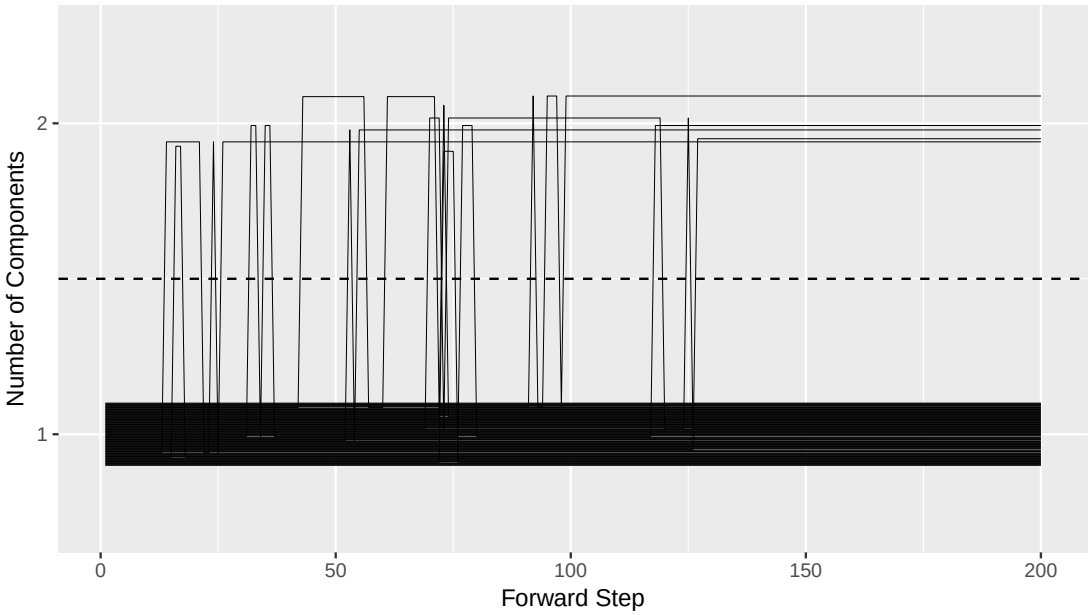


Figure A.2: Sample trajectory diagram for density estimation in two-component GMM with $\sigma = 0.9$, showing that model choice converges after $N = n+200$ imputed observations.

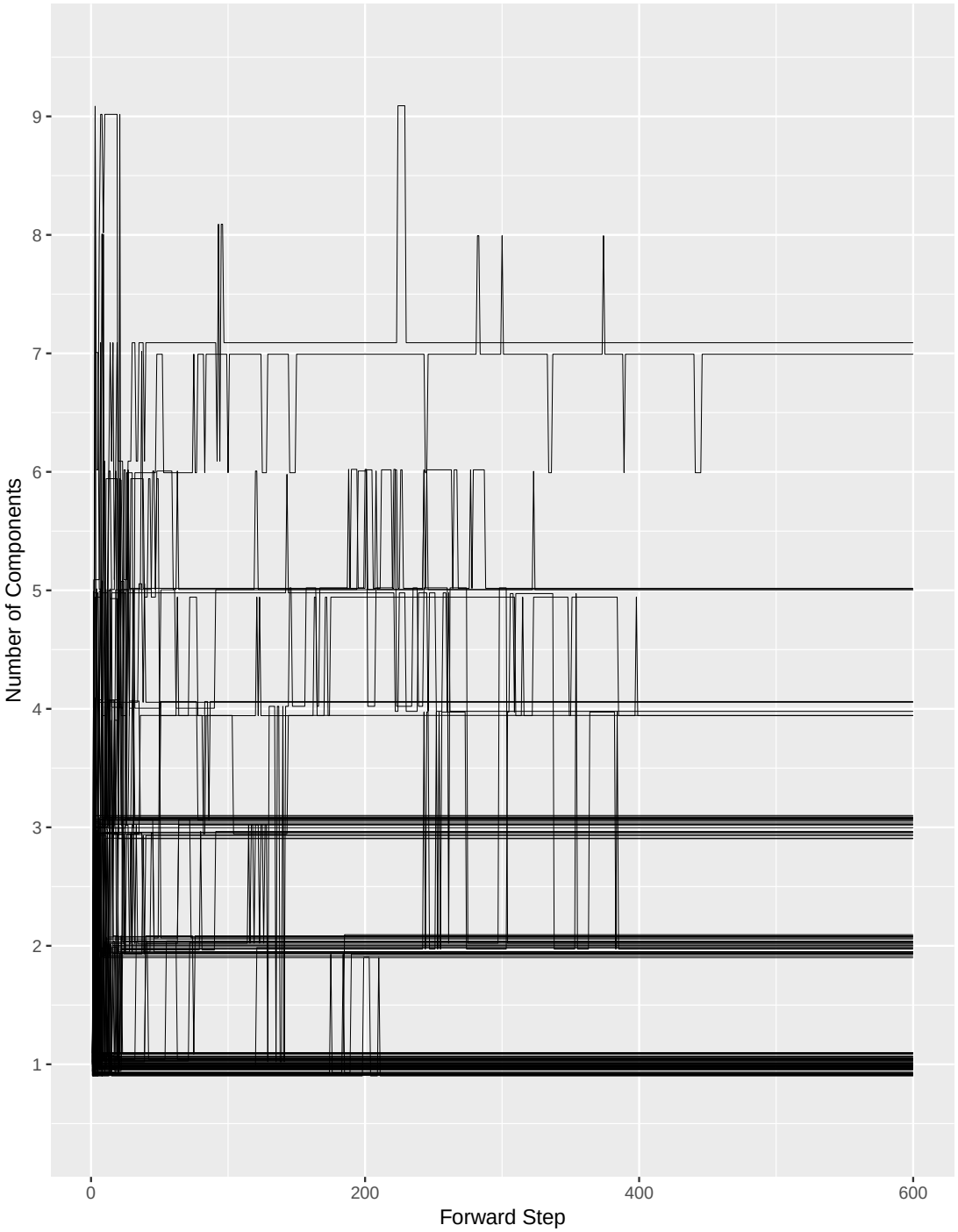


Figure A.3: Sample trajectory diagram for density estimation in three-component GMM with $n = 20$ initial data points, showing that model choice converges after $N = n + 600$ imputed observations.

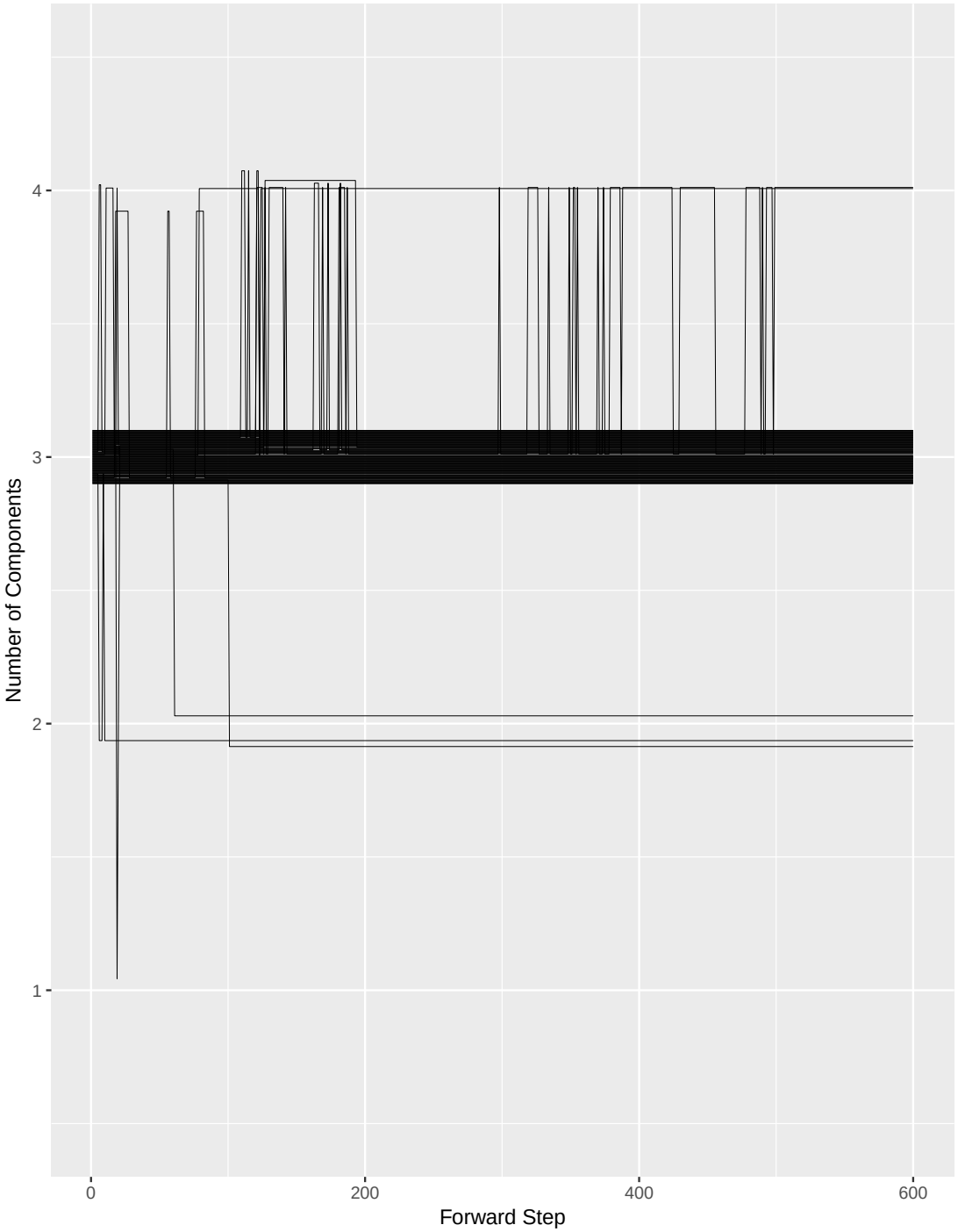


Figure A.4: Sample trajectory diagram for density estimation in three-component GMM with $n = 50$ initial data points, showing that model choice converges after $N = n + 600$ imputed observations.

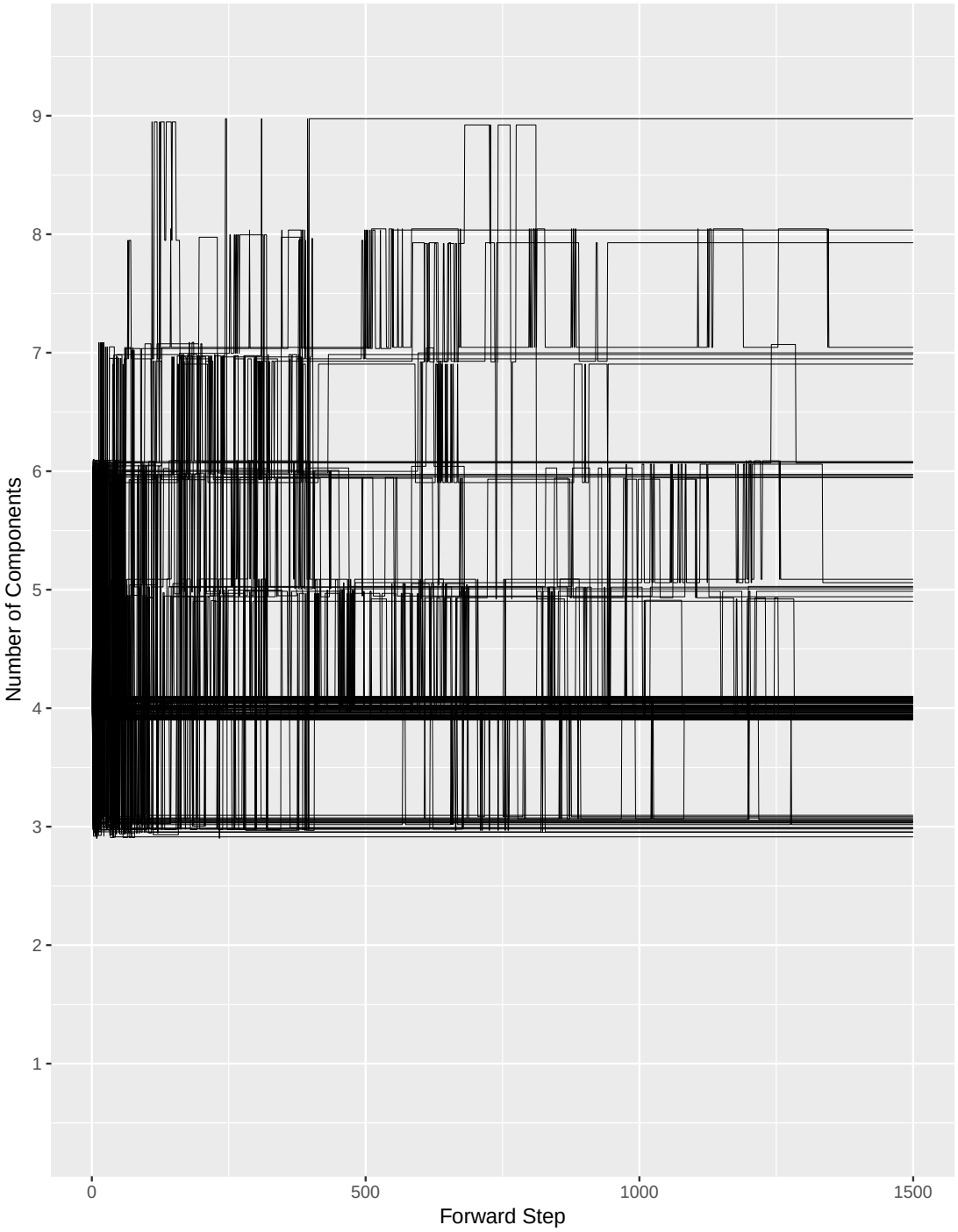


Figure A.5: Sample trajectory diagram for density estimation with galaxies dataset, showing that model choice converges after $N = n + 1500$ imputed observations.

A.2 Two-sided hypothesis testing (Chapter 3)

The first set of diagrams here demonstrates empirical convergence for the two-sided hypothesis testing demonstration in Section 3.3. For $n = \{30, 100, 300, 1000\}$ observations simulated from the "true alternative" $\mathcal{N}(0.1, 1)$, we see in Figures A.6, A.7, A.8, and A.9 respectively that $N = n + 20n$ additional observations are sufficient for $\mathcal{M}_{\widehat{k}(N)}$ to closely approximate the final $\mathcal{M}_{k(\infty)}$ over 400 random trials. We note that the scale of the y-axis (the range of final model means) decreases as the initial observed n increases, reflecting the intuition that uncertainty should be reduced with the observation of additional data.

The second set of diagrams compares resampling with e-values. In Figure A.10, we plot the (log-scaled) e-values for each seed on the horizontal axis and the resampling posterior probabilities of H_0 on the vertical axis, for data generated under H_0. This corresponds to Figure 3.3 for p-values. The X marks on the plots indicate tests with $e > 10$ where Jeffreys' rule of thumb finds strong evidence against H_0, and the O marks indicate $e < 10$. We note that the e-values are invariant to sample size, with a constant Type I error rate regardless of n, but that the resampling probabilities tend towards 1 as the sample size grows. Figure A.11 is the corresponding diagram to Figure 3.4 for p-values, with data generated under H_1. The e-values build evidence against the null as n increases, with the points gradually migrating towards the bottom and right as we observe more data.

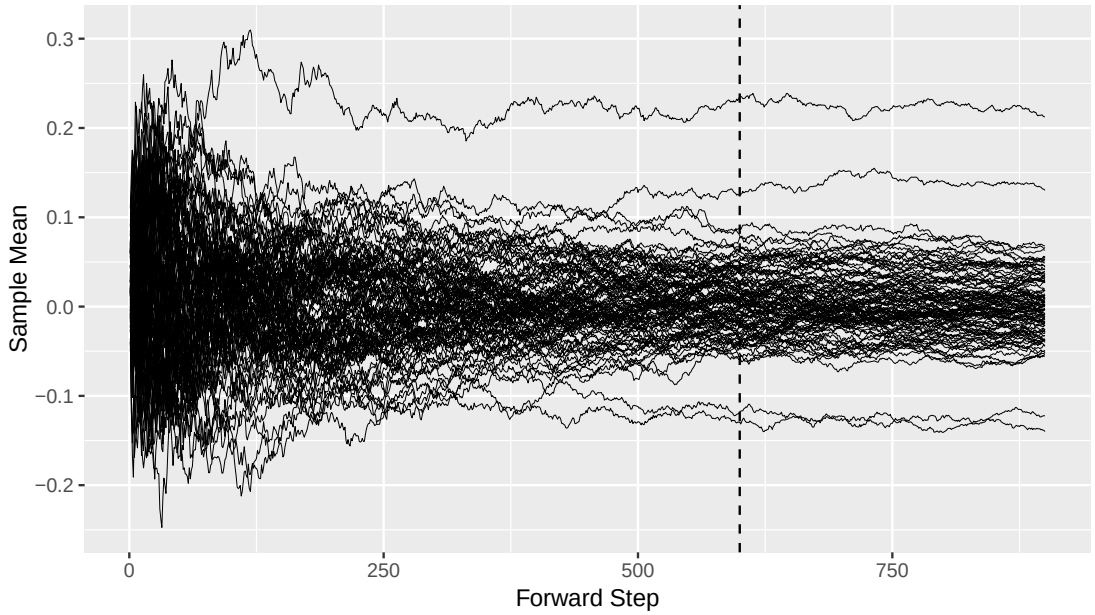


Figure A.6: Sample trajectory diagram for predictive resampling with $m = 30$ observations simulated from $\mathcal{N}(0.1, 1)$, showing that model choice converges after $M = m+20m$ imputed observations, indicated by the dashed vertical line.

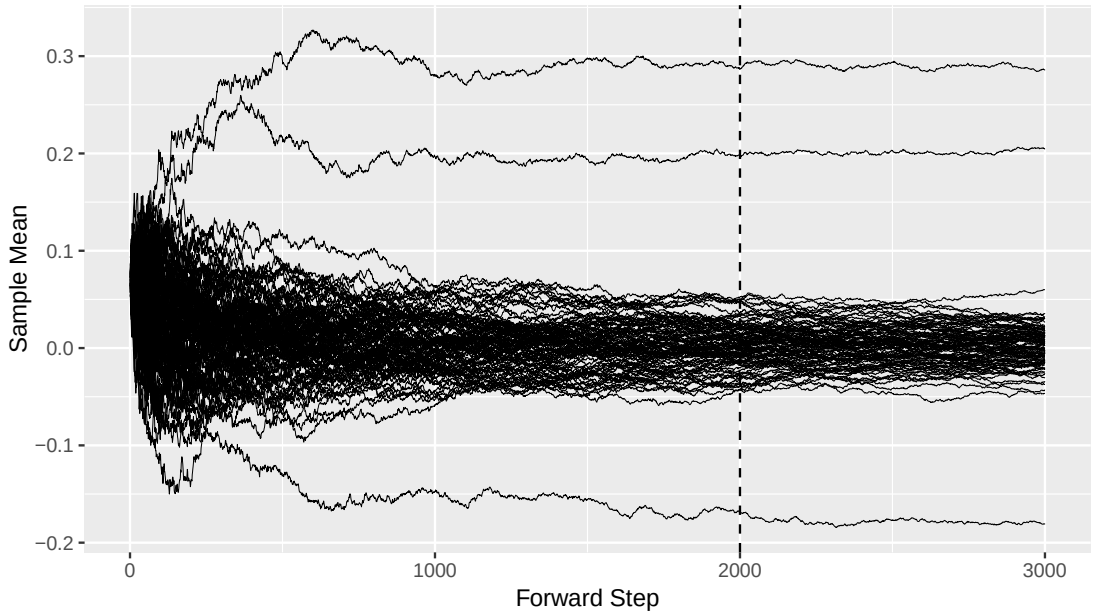


Figure A.7: Sample trajectory diagram for predictive resampling with $m = 100$ observations simulated from $\mathcal{N}(0.1, 1)$, showing that model choice converges after $M = m + 20m$ imputed observations, indicated by the dashed vertical line.

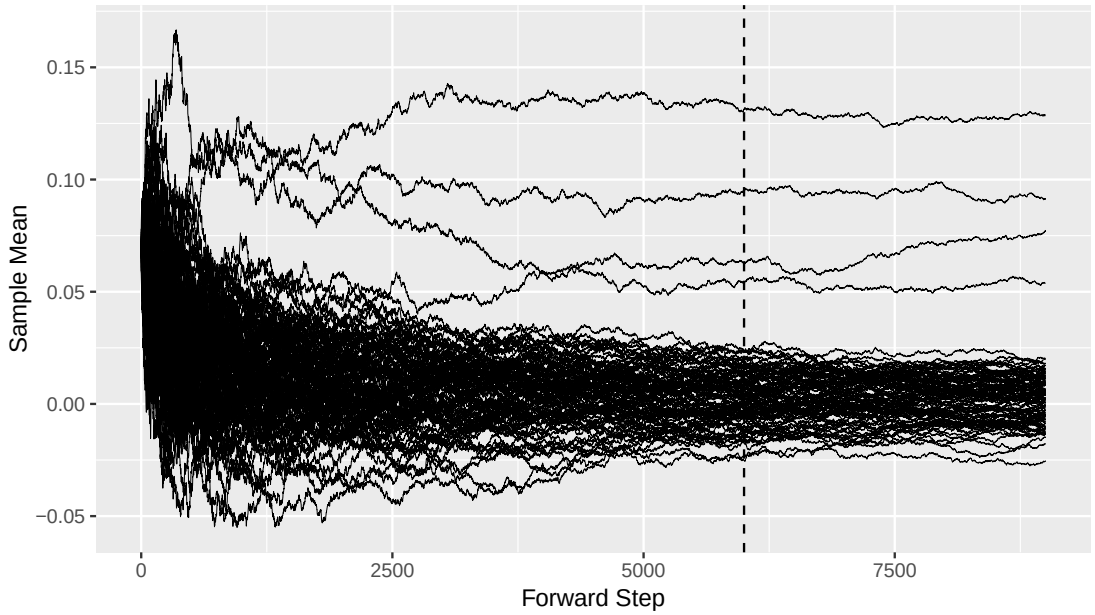


Figure A.8: Sample trajectory diagram for predictive resampling with $m = 300$ observations simulated from $\mathcal{N}(0.1, 1)$, showing that model choice converges after $M = m + 20m$ imputed observations, indicated by the dashed vertical line.

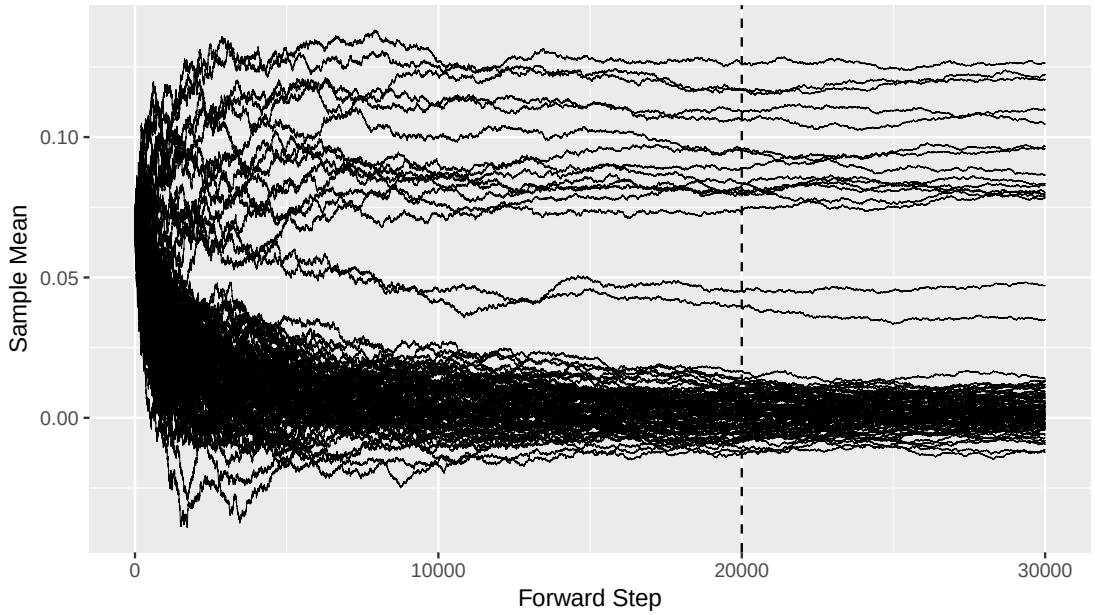


Figure A.9: Sample trajectory diagram for predictive resampling with $m = 1000$ observations simulated from $\mathcal{N}(0.1, 1)$, showing that model choice converges after $M = m + 20m$ imputed observations, indicated by the dashed vertical line.

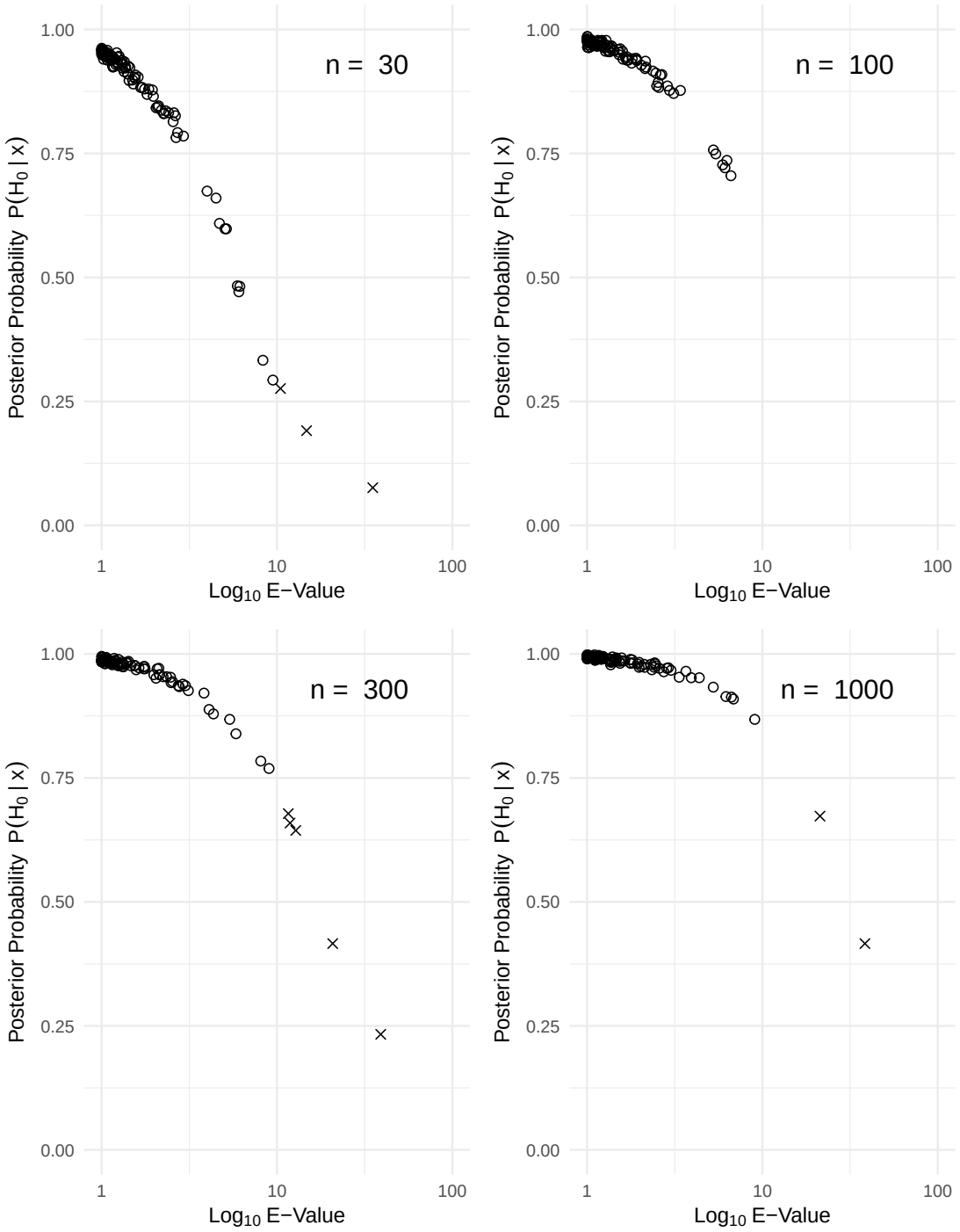


Figure A.10: Observed e-value (log-scale) vs. resampling posterior probability of H_0 for data generated under the null $\mathcal{N}(0, 1)$ across 100 out of 400 random seeds. X denotes tests with $e > 10$ where Jeffreys' rule of thumb finds strong evidence against H_0, and O denotes tests with $e < 10$.

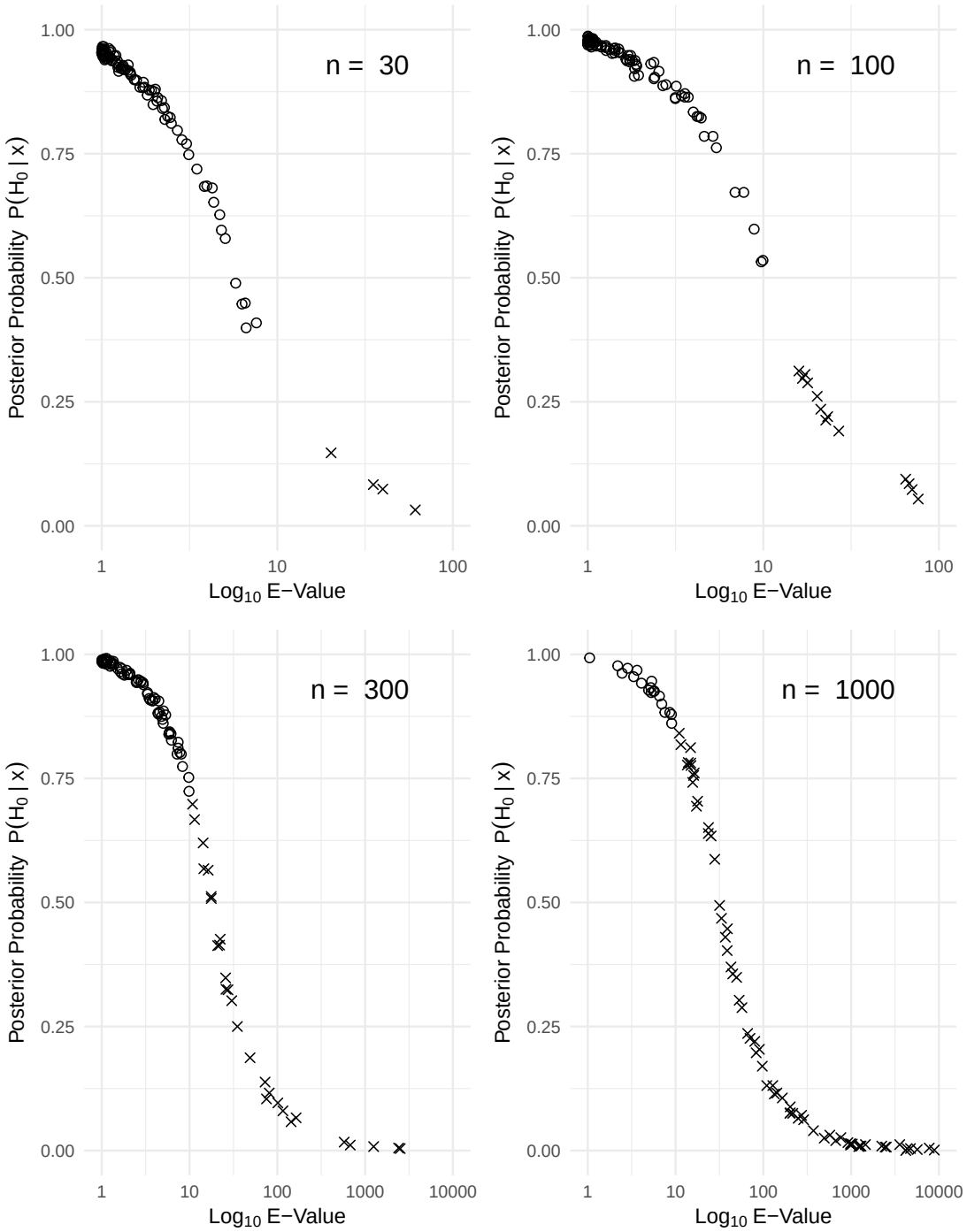


Figure A.11: Observed e-value (log-scale) vs. resampling posterior probability of H_0 for data generated under the alternative $\mathcal{N}(0.1, 1)$ across 100 out of 400 random seeds. X denotes tests with $e > 10$ where Jeffreys' rule of thumb finds strong evidence against H_0, and O denotes tests with $e < 10$.

A.3 Causal forest simulation (Chapter 6)

The table below displays complete numerical metrics for the simulation study in Section 6.6, including both the randomized-treatment and observational settings.

		RCT data				**Observational data**			
		Power (%)		**VI (%)**		**Power (%)**		**VI (%)**	
HTE ρ	n	**RD**	**RR**	**RD**	**RR**	**RD**	**RR**	**RD**	**RR**
0	**2,500**	2	3	37.6	42.5	2	5	42.8	41.9
	5,000	5	5	38.1	41.2	3	5	43.3	43.9
	7,500	6	6	38.1	39.7	7	8	41.8	44.8
	10,000	2	7	38.6	39.5	3	9	43.1	46.0
0.25	**2,500**	2	3	39.7	45.7	2	4	44.6	45.1
	5,000	3	8	42.5	47.1	7	15	46.3	52.1
	7,500	10	15	44.6	48.1	11	25	46.1	56.6
	10,000	11	14	47.7	49.3	15	40	48.5	61.4
0.5	**2,500**	11	15	46.6	53.7	13	17	48.6	53.0
	5,000	26	32	53.8	61.3	36	51	53.7	64.5
	7,500	52	57	60.6	67.1	65	90	56.3	71.5
	10,000	63	78	66.3	71.2	76	96	61.5	77.0
0.75	**2,500**	46	55	56.3	64.3	44	51	55.4	62.6
	5,000	81	88	69.3	76.0	93	94	64.7	75.6
	7,500	97	99	76.5	81.9	96	99	69.3	80.9
	10,000	100	100	81.0	85.1	100	100	75.8	83.6

Table A.1: Complete numerical results for simulation experiment, including power (proportion of tests with omnibus p-value below 0.05) and variable importance (average weight assigned to true predictive covariates). Results are compared for `grf` forest based on absolute Risk Difference (**RD**) and `rrcf` forest based on relative Risk Ratio (**RR**).

Bibliography

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., Makarenkov, V., & Nahavandi, S. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, *76*, 243–297.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Altman, D. G. (1999). *Practical statistics for medical research*. Chapman & Hall/CRC.

Andersen, L. W. (2021). Absolute vs. relative effects—implications for subgroup analyses. *Trials*, *22*(1), 50.

Athey, S., & Imbens, G. (2015). Recursive Partitioning for Heterogeneous Causal Effects [arXiv:1504.01132].

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

Athey, S., & Wager, S. (2019). Estimating Treatment Effects with Causal Forests: An Application [arxiv:1902.07409].

Balakrishnan, S., Wainwright, M. J., & Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, *45*(1), 77–120.

Barbieri, M. M., & Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, *32*(3).

Basu, S., Raghavan, S., Wexler, D. J., & Berkowitz, S. A. (2018). Characteristics Associated With Decreased or Increased Mortality Risk From Glycemic Therapy Among Patients With Type 2 Diabetes and High Cardiovascular Risk: Machine Learning Analysis of the ACCORD Trial. *Diabetes Care*, *41*(3), 604–612.

Battiston, M., & Cappello, L. (2025). Bayesian predictive inference beyond martingales [arXiv:2507.21874].

Bayarri, M. J., & Berger, J. (2000). *P* Values for Composite Null Models. *Journal of the American Statistical Association*, *95*(452), 1127–1142.

Bayarri, M. J., & Berger, J. (2004). The Interplay of Bayesian and Frequentist Analysis. *Statistical Science*, *19*(1).

Bayes, T. (1764). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418.

Berger, J. (2003). Could Fisher, Jeffreys and Neyman Have Agreed on Testing? *Statistical Science*, *18*(1), 1–12.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *1*(3).

Berger, J., & Pericchi, L. R. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association*, *91*(433), 109–122.

Bernardo, J. M., & Smith, A. F. M. (2004). *Bayesian theory*. John Wiley; Sons, Inc.

Berti, P., Dreassi, E., Leisen, F., Pratelli, L., & Rigo, P. (2023). A Probabilistic View on Predictive Constructions for Bayesian Learning. *Statistical Science*.

Berti, P., Dreassi, E., Pratelli, L., & Rigo, P. (2021). A class of models for Bayesian predictive inference. *Bernoulli*, *27*(1).

Berti, P., Pratelli, L., & Rigo, P. (2004). Limit theorems for a class of identically distributed random variables. *The Annals of Probability*, *32*(3).

Bickford Smith, F., Kossen, J., Trollope, E., van der Wilk, M., Foster, A., & Rainforth, T. (2024). Rethinking Aleatoric and Epistemic Uncertainty [arXiv:2412.20892].

Bissiri, P. G., Holmes, C. C., & Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *78*(5), 1103–1130.

Blundell, C., Cornebise, J., Kavukcuoglu, K., & Wierstra, D. (2015). Weight Uncertainty in Neural Network. *Proceedings of the 32nd International Conference on Machine Learning*, 1613–1622.

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L. (2001b). Statistical Modeling: The Two Cultures. *Statistical Science*, *16*(3), 199–215.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification And Regression Trees* (1st ed.). Routledge.

Cai, D., Campbell, T., & Broderick, T. (2020). Finite mixture models do not reliably learn the number of components.

Chen, W., Qian, L., Shi, J., & Franklin, M. (2018). Comparing performance between log-binomial and robust Poisson regression models for estimating risk ratios under model misspecification. *BMC Medical Research Methodology*, *18*(1), 63.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.

Ciampi, A., Hogg, S. A., McKinney, S., & Thiffault, J. (1988). RECPAM: A computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. I. Methods and program features. *Computer Methods and Programs in Biomedicine*, *26*(3), 239–256.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*.

Coker, B., Bruinsma, W. P., Burt, D. R., Pan, W., & Doshi-Velez, F. (2022). Wide Mean-Field Bayesian Neural Networks Ignore the Data. *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 5276–5333.

Colnet, B., Josse, J., Varoquaux, G., & Scornet, E. (2023). Risk ratio, odds ratio, risk difference... Which causal measure is easier to generalize? [arxiv:2303.16008].

Cook, D., Gebski, V. J., & Keech, A. C. (2004). Subgroup analysis in clinical trials. *Medical Journal of Australia*, *180*(6), 289–291.

Cook, R., & Sackett, D. L. (1995). The number needed to treat: A clinically useful measure of treatment effect. *BMJ*, *310*(6977), 452–454.

Cox, D. R. (1958). Some Problems Connected with Statistical Inference. *The Annals of Mathematical Statistics*, *29*(2), 357–372.

Dandl, S., Bender, A., & Hothorn, T. (2024). Heterogeneous treatment effect estimation for observational data using model-based forests. *Statistical Methods in Medical Research*, *33*(3), 392–413.

Davis, J. M., & Heller, S. B. (2017). Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs. *The American Economic Review*, *107*(5), 546–550.

Dawid, A. P. (1982). The Well-Calibrated Bayesian. *Journal of the American Statistical Association*, *77*(379), 605–610.

Dawid, A. P. (1984). Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach. *Journal of the Royal Statistical Society. Series A (General)*, *147*(2), 278–292.

Daxberger, E., Nalisnick, E., Allingham, J. U., Antoran, J., & Hernandez-Lobato, J. M. (2021). Bayesian Deep Learning via Subnetwork Inference. *Proceedings of the 38th International Conference on Machine Learning*, 2510–2521.

de Finetti, B. (1937). La prevision : Ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, *7*, 1–68.

Deeks, J. J. (2002). Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*, *21*(11), 1575–1600.

Denil, M., Matheson, D., & Freitas, N. D. (2014). Narrowing the Gap: Random Forests In Theory and In Practice. *Proceedings of the 31st International Conference on Machine Learning*, 665–673.

Doi, S. A., Furuya-Kanamori, L., Xu, C., Lin, L., Chivese, T., & Thalib, L. (2022). Controversy and Debate: Questionable utility of the relative risk in clinical research: A call for change to practice. *Journal of Clinical Epidemiology*, *142*, 271–279.

Doob, J. (1949). Application of the theory of martingales. *Actes du Colloque International Le Calcul des Probabilités et ses applications*, 23–27.

Doob, J. (1953). *Stochastic processes*. Wiley.

Draper, D. (1995). Assessment and Propagation of Model Uncertainty. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *57*(1), 45–70.

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18.

Edwards, A. (1997). What Did Fisher Mean by "Inverse Probability" in 1912 – 1922? *Statistical Science*, *12*(3), 177–184.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1).

Efron, B. (2012). Bayesian inference and the parametric bootstrap. *The Annals of Applied Statistics*, *6*(4), 1971–1997.

Efroymson, M. (1960). Multiple regression analysis. *Mathematical Methods for Digital Computers*.

Escobar, M. D., & West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, *90*(430), 577–588.

Evans, R. J., & Didelez, V. (2024). Parameterizing and simulating from causal models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *86*(3), 535–568.

Falck, F., Wang, Z., & Holmes, C. (2024). Is in-context learning in large language models bayesian? a martingale perspective. *Proceedings of the 41st International Conference on Machine Learning*, *235*, 12784–12805.

Farquhar, S., Smith, L., & Gal, Y. (2020). Liberty or Depth: Deep Bayesian Neural Nets Do Not Need Complex Weight Posterior Approximations.

Fisher, A., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, *20*, 177.

Fisher, R. (1925). *Statistical Methods for Research Workers*. Oliver; Boyd.

Fong, E., & Holmes, C. C. (2020). On the marginal likelihood and cross-validation. *Biometrika*, *107*(2), 489–496.

Fong, E., Holmes, C., & Walker, S. G. (2024). Martingale posterior distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(5), 1357–1391.

Fong, E., & Yiu, A. (2024a). Asymptotics for parametric martingale posteriors [arXiv:2410.17692].

Fong, E., & Yiu, A. (2024b). Bayesian Quantile Estimation and Regression with Martingale Posteriors [arXiv:2406.03358].

Foong, A., Burt, D., Li, Y., & Turner, R. (2020). On the Expressiveness of Approximate Inference in Bayesian Neural Networks. *Advances in Neural Information Processing Systems*, *33*, 15897–15908.

Fortini, S., & Petrone, S. (2012). Predictive construction of priors in Bayesian nonparametrics. *Brazilian Journal of Probability and Statistics*, *26*(4).

Fortini, S., & Petrone, S. (2020). Quasi-Bayes Properties of a Procedure for Sequential Learning in Mixture Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(4), 1087–1114.

Fortini, S., & Petrone, S. (2025). Exchangeability, Prediction and Predictive Modeling in Bayesian Statistics. *Statistical Science*, *40*(1).

Fortuin, V., Garriga-Alonso, A., Ober, S. W., Wenzel, F., Rätsch, G., Turner, R. E., van der Wilk, M., & Aitchison, L. (2021). Bayesian Neural Network Priors Revisited. *Proceedings of the International Conference on Machine Learning*.

Fraley, C., & Raftery, A. E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, *97*(458), 611–631.

Fraser, D. (1961). The Fiducial Method and Invariance. *Biometrika*, *48*(3/4), 261–280.

Friedman, N., & Koller, D. (2003). Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. *Machine Learning*, *50*(1), 95–125.

Friel, N., & Pettitt, A. N. (2008). Marginal Likelihood Estimation via Power Posteriors. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *70*(3), 589–607.

Furukawa, T. A., Guyatt, G. H., & Griffith, L. E. (2002). Can we individualize the 'number needed to treat'? An empirical study of summary effect measures in meta-analyses. *International Journal of Epidemiology*, *31*(1), 72–76.

Geisser, S. (1975). The Predictive Sample Reuse Method with Applications. *Journal of the American Statistical Association*, *70*(350), 320–328.

Geisser, S. (1982). Aspects of the Predictive and Estimative Approaches in the Determination of Probabilities. *Biometrics*, *38*, 75–85.

Geisser, S. (1993). *Predictive inference*. Chapman & Hall/CRC.

Gelman, A., & Loken, E. (2019). The garden of forking paths : Why multiple comparisons can be a problem , even when there is no " fishing expedition " or " p-hacking " and the research hypothesis was posited ahead of time.

Gelman, A., & Stern, H. (2006). The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant. *The American Statistician*, *60*(4), 328–331.

George, E. I., & McCulloch, R. E. (1993). Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*, *88*(423), 881–889.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*(3), 295–314.

Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Good, I. J. (1952). Rational Decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *14*(1), 107–114.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711–732.

Grünwald, P. (2012). The Safe Bayesian. In N. H. Bshouty, G. Stoltz, N. Vayatis, & T. Zeugmann (Eds.), *Algorithmic Learning Theory* (pp. 169–183).

Grünwald, P., De Heide, R., & Koolen, W. (2024). Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *86*(5), 1091–1128.

Grünwald, P., & Ommen, T. v. (2017). Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis*, *12*(4), 1069–1103.

Hahn, P. R., Martin, R., & Walker, S. G. (2018). On Recursive Bayesian Predictive Distributions. *Journal of the American Statistical Association*, *113*(523), 1085–1093.

Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis*, *15*(3).

Hannig, J., Iyer, H., Lai, R. C. S., & Lee, T. C. M. (2016). Generalized Fiducial Inference: A Review and New Results. *Journal of the American Statistical Association*, *111*(515), 1346–1361.

Harder, J. A. (2020). The Multiverse of Methods: Extending the Multiverse Analysis to Address Data-Collection Decisions. *Perspectives on Psychological Science*, *15*(5), 1158–1177.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, *14*(4), 382–401.

Holmes, C. C., & Walker, S. G. (2023). Statistical inference with exchangeability and martingales. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *381*(2247), 20220143.

Houlsby, N., Huszár, F., Ghahramani, Z., & Lengyel, M. (2011). Bayesian Active Learning for Classification and Preference Learning [arXiv:1112.5745].

Hsiao, C. K., Lee, M.-h., & Kass, R. E. (2005). Bayesian tests of extra-Binomial variability. *Statistics in Medicine*, *24*(1), 49–64.

Huitfeldt, A., Fox, M. P., Murray, E. J., Hróbjartsson, A., & Daniel, R. M. (2022). Shall we count the living or the dead? [arxiv:2106.06316].

Izmailov, P., Maddox, W. J., Kirichenko, P., Garipov, T., Vetrov, D., & Wilson, A. G. (2020). Subspace Inference for Bayesian Deep Learning. *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, 1169–1179.

Izmailov, P., Vikram, S., Hoffman, M. D., & Wilson, A. G. (2021). What Are Bayesian Neural Network Posteriors Really Like? [arXiv:2104.14421].

Jeffreys, H. (1961). *Theory of probability*.

Johnson, V. E., & Rossell, D. (2010). On the Use of Non-Local Prior Densities in Bayesian Hypothesis Tests. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *72*(2), 143–170.

Kallioinen, N., Paananen, T., Bürkner, P.-C., & Vehtari, A. (2023). Detecting and diagnosing prior and likelihood sensitivity with power-scaling. *Statistics and Computing*, *34*(1), 57.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kent, D. M., Rothwell, P. M., Ioannidis, J. P., Altman, D. G., & Hayward, R. A. (2010). Assessing and reporting heterogeneity in treatment effects in clinical trials: A proposal. *Trials*, *11*(1), 85.

Knoblauch, J., Jewson, J., & Damoulas, T. (2022). An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference. *Journal of Machine Learning Research*, *23*(132), 1–109.

Kooperberg, C., Stone, C. J., & Truong, Y. K. (1995). Hazard Regression. *Journal of the American Statistical Association*, *90*(429), 78–94.

Kristiadi, A., Hein, M., & Hennig, P. (2020). Being Bayesian, Even Just a Bit, Fixes Overconfidence in ReLU Networks. *Proceedings of the 37th International Conference on Machine Learning*, 5436–5446.

Kuipers, J., & Moffa, G. (2017). Partition MCMC for inference on acyclic digraphs. *Journal of the American Statistical Association*, *112*(517), 282–299.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *Advances in Neural Information Processing Systems*, *30*.

Laplace, P. S. (1774). Memoire sur la probabilite des causes par les evenemens. *Memoires de mathematique et de physique presentes i I'Academie royale des sciences, par divers savans, & lts dans ses assemblee*, *6*, 621–656.

Lartillot, N., & Philippe, H. (2006). Computing Bayes Factors Using Thermodynamic Integration. *Systematic Biology*, *55*(2), 195–207.

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*. Wiley.

Lin, X., Tarp, J. M., & Evans, R. J. (2024). Data fusion for efficiency gain in ATE estimation: A practical review with simulations [arxiv:2407.01186].

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*(1-2), 187–192.

Lindley, D. V. (1958). Fiducial Distributions and Bayes' Theorem. *Journal of the Royal Statistical Society. Series B (Methodological)*, *20*(1), 102–107.

Lipkovich, I., Svensson, D., Ratitch, B., & Dmitrienko, A. (2024). Modern approaches for evaluating treatment effect heterogeneity from clinical trials and observational data. *Statistics in Medicine*, *43*(22), 4388–4436.

Loh, W.-Y. (2002). Regression Trees With Unbiased Variable Selection and Interaction Detection. *Statistica Sinica*, *12*, 361–386.

MacKay, D. J. C. (1992a). Bayesian Interpolation. In *Maximum Entropy and Bayesian Methods* (pp. 39–66). Springer.

MacKay, D. J. C. (1992b). A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, *4*(3), 448–472.

Madigan, D., York, J., & Allard, D. (1995). Bayesian Graphical Models for Discrete Data. *International Statistical Review / Revue Internationale de Statistique*, *63*(2), 215–232.

Marso, S. P., Daniels, G. H., Brown-Frandsen, K., Kristensen, P., Mann, J. F., Nauck, M. A., Nissen, S. E., Pocock, S., Poulter, N. R., Ravn, L. S., Steinberg, W. M., Stockner, M., Zinman, B., Bergenstal, R. M., & Buse, J. B. (2016). Liraglutide and

Cardiovascular Outcomes in Type 2 Diabetes. *New England Journal of Medicine*, *375*(4), 311–322.

Matheson, J. E., & Winkler, R. L. (1976). Scoring Rules for Continuous Probability Distributions. *Management Science*, *22*(10), 1087–1096.

McAllester, D. A. (1999). PAC-Bayesian model averaging. *Proceedings of the twelfth annual conference on Computational learning theory*, 164–170.

McLatchie, Y., Fong, E., Frazier, D. T., & Knoblauch, J. (2024). Predictive performance of power posteriors [arXiv:2408.08806].

Meng, X.-L. (1994). Posterior Predictive p-Values. *The Annals of Statistics*, *22*(3).

Mlodozeniec, B., Krueger, D., & Turner, R. (2024). Implicitly Bayesian Prediction Rules in Deep Learning. *Proceedings of the 6th Symposium on Advances in Approximate Bayesian Inference*, 79–110.

Nalisnick, E., & Smyth, P. (2018). Learning Priors for Invariance. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, 366–375.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks* (Vol. 118). Springer.

Neyman, J., & Pearson, E. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London, Series A, Containing Papers of a Mathematical or Physical Character*, *281*, 289–337.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301.

Noci, L., Roth, K., Bachmann, G., Nowozin, S., & Hofmann, T. (2024). Disentangling the roles of curation, data-augmentation and the prior in the cold posterior effect.

Proceedings of the 35th International Conference on Neural Information Processing Systems, 12738–12748.

Noordzij, M., Van Diepen, M., Caskey, F. C., & Jager, K. J. (2017). Relative risk versus absolute risk: One cannot be interpreted without the other. *Nephrology Dialysis Transplantation*, *32*(suppl_2), ii13–ii18.

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 99–138.

Osband, I., Wen, Z., Asghari, S. M., Dwaracherla, V., Ibrahimi, M., Lu, X., & Van Roy, B. (2021). Epistemic Neural Networks. *Advances in Neural Information Processing Systems*.

Petersen, M. R., & Deddens, J. A. (2008). A comparison of two methods for estimating prevalence ratios. *BMC Medical Research Methodology*, *8*(1), 9.

Piccininni, M., & Stensrud, M. J. (2025). Immune-selection stability is a neglected property of the causal risk ratio. *American Journal of Epidemiology*, kwaf086.

Plummer, M., Stukalov, A., & Denwood, M. (2023). Rjags: Bayesian Graphical Models using MCMC.

Poole, C., Shrier, I., & VanderWeele, T. J. (2015). Is the Risk Difference Really a More Heterogeneous Measure? *Epidemiology (Cambridge, Mass.)*, *26*(5), 714–718.

Raghavan, S., Josey, K., & Ghosh, D. (2022). Generalizability of heterogeneous treatment effects based on causal forests applied to two randomized clinical trials of intensive glycemic control. *Annals of Epidemiology*.

Richardson, S., & Green, P. J. (1997). On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *59*(4), 731–792.

Riha, A. E., Siccha, N., Oulasvirta, A., & Vehtari, A. (2024). Supporting Bayesian modelling workflows with iterative filtering for multiverse analysis [arXiv:2404.01688].

Roberts, H. V. (1965). Probabilistic Prediction. *Journal of the American Statistical Association*, *60*(309), 50–62.

Rodríguez, C. E., Mena, R. H., & Walker, S. G. (2025). Martingale Posterior Inference for Finite Mixture Models and Clustering. *Journal of Computational and Graphical Statistics*, 1–10.

Roeder, K. (1990). Density Estimation with Confidence Sets Exemplified by Superclusters and Voids in the Galaxies. *Journal of the American Statistical Association*, *85*(411), 617–624.

Ross, G. J., & Markwick, D. (2019). Dirichletprocess: An R Package for Fitting Complex Bayesian Nonparametric Models.

Rothman, K. J. (2012). *Epidemiology: An introduction* (2nd ed). Oxford University Press.

Rubin, D. B. (1981). The Bayesian Bootstrap. *The Annals of Statistics*, *9*(1), 130–134.

Rubin, D. B. (2005). Causal Inference Using Potential Outcomes: Design, Modeling, Decisions. *Journal of the American Statistical Association*, *100*(469), 322–331.

Rudner, T., Chen, Z., & Gal, Y. (2020). Rethinking Function-Space Variational Inference in Bayesian Neural Networks. *Advances in Approximate Bayesian Inference*.

Savage, L. J. (1976). On Rereading R. A. Fisher. *The Annals of Statistics*, *4*(3), 441–500.

Schechtman, E. (2002). Odds Ratio, Relative Risk, Absolute Risk Reduction, and the Number Needed to Treat—Which of These Should We Use? *Value in Health*, *5*(5), 431–436.

Schmid, C. H., Lau, J., McIntosh, M. W., & Cappelleri, J. C. (1998). An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. *Statistics in Medicine*, *17*(17), 1923–1942.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2).

Scott, J. G., & Berger, J. O. (2010). Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem. *The Annals of Statistics*, *38*(5), 2587–2619.

Scrucca, L., Fraley, C., Murphy, B. T., & Raftery, A. E. (2023). *Model-Based Clustering, Classification, and Density Estimation Using mclust in R.*

Sechidis, K., Sun, S., Chen, Y., Lu, J., Zang, C., Baillie, M., Ohlssen, D., Vandemeulebroecke, M., Hemmings, R., Ruberg, S., & Bornkamp, B. (2024). WATCH: A Workflow to Assess Treatment Effect Heterogeneity in Drug Development for Clinical Trial Sponsors [arXiv:2405.00859].

Seibold, H., Zeileis, A., & Hothorn, T. (2016). Model-Based Recursive Partitioning for Subgroup Analyses. *The International Journal of Biostatistics*, *12*(1), 45–63.

Seibold, H., Zeileis, A., & Hothorn, T. (2018). Individual treatment effect prediction for amyotrophic lateral sclerosis patients. *Statistical Methods in Medical Research*, *27*(10), 3104–3125.

Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential Deep Learning to Quantify Classification Uncertainty. *Advances in Neural Information Processing Systems*.

Shafer, G. (2021). Testing by Betting: A Strategy for Statistical and Scientific Communication. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *184*(2), 407–431.

Shao, J. (1993). Linear Model Selection by Cross-validation. *Journal of the American Statistical Association*, *88*(422), 486–494.

Sharma, M., Farquhar, S., Nalisnick, E., & Rainforth, T. (2023). Do Bayesian Neural Networks Need To Be Fully Stochastic? *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, 7694–7722.

Sheps, M. C. (1958). Shall We Count the Living or the Dead? *New England Journal of Medicine*, *259*(25), 1210–1214.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3).

Simon, S. D. (2001). Understanding the Odds Ratio and the Relative Risk. *Journal of Andrology*, *22*(4), 533–536.

Sivula, T., Magnusson, M., Matamoros, A. A., & Vehtari, A. (2023). Uncertainty in Bayesian Leave-One-Out Cross-Validation Based Model Comparison.

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems*.

Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M., Prabhat, M., & Adams, R. (2015). Scalable Bayesian Optimization Using Deep Neural Networks. *Proceedings of the 32nd International Conference on Machine Learning*, 2171–2180.

Spiegelman, D., & VanderWeele, T. J. (2017). Evaluating Public Health Interventions: 6. Modeling Ratios or Differences? Let the Data Tell Us. *American Journal of Public Health*, *107*(7), 1087–1091.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*(5), 702–712.

Stensrud, M. J., & Smith, L. (2023). Identification of Vaccine Effects When Exposure Status Is Unknown. *Epidemiology*, *34*(2), 216–224.

Stigler, S. M. (1982). Thomas Bayes's Bayesian Inference. *Journal of the Royal Statistical Society. Series A (General)*, *145*(2), 250.

Stigler, S. M. (1986). Laplace's 1774 Memoir on Inverse Probability. *Statistical Science*, *1*(3), 359–363.

Sun, S., Zhang, G., Shi, J., & Grosse, R. (2019). Functional Variational Bayesian Neural Networks. *Proceedings of the 36th International Conference on Machine Learning*.

Sun, X., Ioannidis, J. P. A., Agoritsas, T., Alba, A. C., & Guyatt, G. (2014). How to Use a Subgroup Analysis: Users' Guide to the Medical Literature. *JAMA*, *311*(4), 405.

Tibshirani, J., Athey, S., Sverdrup, E., & Wager, S. (2023). Grf: Generalized Random Forests.

Trippe, B., & Turner, R. (2018). Overpruning in Variational Bayesian Neural Networks [arXiv:1801.06230].

Van Der Laan, M. J., Hubbard, A., & Jewell, N. P. (2007). Estimation of Treatment Effects in Randomized Trials With Non-Compliance and a Dichotomous Outcome. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *69*(3), 463–482.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432.

Vehtari, A., & Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, *14*(10), 2439–2468.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*.

Vovk, V., & Wang, R. (2021). E-values: Calibration, combination, and applications. *The Annals of Statistics*, *49*(3).

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Wang, X., Jin, Y., & Yin, L. (2016). Measuring and estimating treatment effect on dichotomous outcome of a population. *Statistical Methods in Medical Research*, *25*(5), 1779–1790.

Watson, J. A., & Holmes, C. C. (2020). Graphing and reporting heterogeneous treatment effects through reference classes. *Trials*, *21*(1), 386.

Welling, M., & Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 681–688.

Wenzel, F., Roth, K., Veeling, B., Swiatkowski, J., Tran, L., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., & Nowozin, S. (2020). How Good is the Bayes Posterior in Deep Neural Networks Really? *Proceedings of the 37th International Conference on Machine Learning*, 10248–10259.

Wilson, A. G., & Izmailov, P. (2020). Bayesian Deep Learning and a Probabilistic Perspective of Generalization. *Advances in Neural Information Processing Systems*, *33*, 4697–4708.

Wu, L., & Williamson, S. A. (2024). Posterior Uncertainty Quantification in Neural Networks using Data Augmentation. *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, 3376–3384.

Xie, M.-g., & Singh, K. (2013). Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review. *International Statistical Review*, *81*(1), 3–39.

Xu, L., & Jordan, M. I. (1996). On Convergence Properties of the EM Algorithm for Gaussian Mixtures. *Neural Computation*, *8*(1), 129–151.

Yang. (2007). Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, *35*(6), 2450–2473.

Yang, Xia, E., Ho, N., & Jordan, M. I. (2019). Posterior Distribution for the Number of Clusters in Dirichlet Process Mixture Models.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Analysis*, *13*(3).

Zabell, S. (1989). R. A. Fisher on the History of Inverse Probability. *Statistical Science*, *4*(3), 247–256.

Zabell, S. (1992). R. A. Fisher and Fiducial Argument. *Statistical Science*, *7*(3), 369–387.

Zabell, S. (2022). Fisher, Bayes, and Predictive Inference. *Mathematics*, *10*(10), 1634.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514.

Zhang, J. L. (2014). Comparative investigation of three Bayesian p values. *Computational Statistics & Data Analysis*, *79*, 277–291.

Zhang, T. (1999). Theoretical analysis of a class of randomized regularization methods. *Proceedings of the twelfth annual conference on Computational learning theory*, 156–163.

Zhang, T. (2006). Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, *52*(4), 1307–1321.