Model misspecification and general Bayesian bootstraps

Simon Lyddon

St Peter's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2018

To Hannah

Acknowledgements

I would like to sincerely thank my supervisor, Chris Holmes, for his patience, guidance and support through my DPhil. His vast knowledge, creativity and passion for research has truly inspired me.

Thank you to my main collaborator, Stephen Walker, for sharing his time and expertise with me so generously during my research visits to the University of Texas at Austin, and in follow up meetings since.

Thank you to all of the members of the Department of Statistics at the University of Oxford that I have worked with over the past four years, making this time so interesting, enjoyable and enriching.

Thank you to my family and friends for all of their support. In particular, thank you to my parents for always being there for me. Finally thank you to my wife, Hannah, for all her love, support and encouragement along the way.

Model misspecification and general Bayesian bootstraps

Simon Lyddon
St Peter's College, University of Oxford

A thesis submitted for the degree of *Doctor of Philosophy*
Michaelmas 2018

Abstract

This thesis explores how a Bayesian should update their beliefs in the knowledge that any model available to them, however useful, is wrong. A number of connections are made between ideas in the model misspecification, Bayesian nonparametrics and bootstrapping literatures. These have led to the development of novel methods for scalable Bayesian inference under model misspecification for a family of general Bayesian parameters.

In Chapter 1 we introduce the core topics of this thesis, in particular the foundations of statistical modelling, model misspecification and general Bayesian updating.

In Chapter 2 we revisit the weighted likelihood bootstrap. This method can be interpreted as sampling exactly from a Bayesian nonparametric posterior distribution, making only weak assumptions about the data-generating mechanism. The loss-likelihood bootstrap is developed as a generalisation of this idea to a wider family of parameters. By making a connection with general Bayesian updating, we provide a means of calibrating the parametric form of the general Bayesian posterior by matching asymptotic posterior information.

In Chapter 3 we present a novel framework for Bayesian nonparametric learning from Bayesian parametric models. A mixture of Dirichlet processes is well suited for incorporating information from a misspecified model into a nonparametric analysis. The properties of this method are examined, and a computationally scalable algorithm for posterior sampling, called the posterior bootstrap, is described. Extensions to the correction of approximate posterior distributions, and to more general parameters of interest, are pursued.

In Chapter 4 we explore two extensions to the work of Chapters 2 and 3. In the first, we explore an application of our nonparametric learning framework to sparse regression problems and investigate the properties of nonparametric learning under a lasso objective function. In the second we consider how we should extend the posterior bootstrap to parametric statistical models containing random effects. Both of these extensions are based on the notion of uncertainty quantification by suitably randomising an objective function.

In Chapter 5 we conclude and discuss future work.

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Foundations of statistical modelling 3

1.2.1 Preliminaries 3

1.2.2 Statistical inference 6

1.3 Model misspecification 11

1.3.1 Misspecification in theory 11

1.3.2 Modelling in practice 14

1.3.3 Robustness 16

1.3.4 Model uncertainty and $\mathcal{M}$-open 18

1.3.5 Nonparametric approaches 20

1.4 Model-free methods 22

1.4.1 The general Bayesian approach 23

1.5 Contribution 27

2 General Bayesian Updating and the Loss-Likelihood Bootstrap **29**

2.1 Preamble 29

2.2 Abstract 30

2.3 Bayesian inference and model misspecification 30

2.4 Revisiting the weighted likelihood bootstrap 34

2.5 Calibrating general Bayesian posteriors by asymptotic information matching 38

2.6 Illustrations 41

2.6.1 Normal model with quadratic loss 41

2.6.2 Bayesian support vector machine for binary classification 43

2.7 Discussion 48

2.8 Supplementary material 50

2.8.1 Details of Theorem 1 50

2.8.1.1 Background 50

2.8.1.2 Conditions . . . 51
2.8.1.3 Main result . . . 53
2.8.2 General Bayesian posterior asymptotics . . . 54
2.9 Further discussion . . . 55
2.9.1 Bayesian bootstraps . . . 55
2.9.2 Alternatives to Fisher information matching . . . 57
2.9.3 Non-uniqueness of the loss . . . 58
2.9.4 Utility of general Bayesian uncertainty quantification . . . 58

3 Nonparametric learning from Bayesian models with randomized objective functions 60
3.1 Preamble . . . 60
3.2 Abstract . . . 61
3.3 Introduction . . . 61
3.4 Foundations of Nonparametric Learning . . . 63
3.4.1 Bayesian updating of misspecified models . . . 63
3.4.2 An NP prior using a MDP . . . 65
3.4.3 Monte Carlo conditional maximization . . . 66
3.4.4 A more general construction . . . 67
3.4.5 The Posterior Bootstrap algorithm . . . 68
3.4.6 Adaptive Nonparametric Learning: aNPL . . . 69
3.4.7 Properties of NPL . . . 70
3.5 Illustrations . . . 72
3.5.1 Exponential family, [S1] . . . 72
3.5.2 Updating approximate posteriors [S2]: Variational Bayes uncertainty correction . . . 72
3.5.3 Directly updating the prior: Bayesian Random Forests, using synthetic generated data . . . 75
3.5.4 Direct updating of utility functions [S3]: population median . . . 76
3.6 Discussion . . . 77
3.7 Supplementary Material . . . 78
3.7.1 Setting the MDP concentration parameter . . . 78
3.7.2 Posterior Bootstrap for adaptive Nonparametric Learning . . . 79
3.7.3 Bayesian Random Forests - further details . . . 80
3.8 Further discussion . . . 82
3.8.1 General form of nonparametric learning . . . 82

3.8.2 Graphical description of the exponential family case 83
3.8.3 Nonparametric learning for small samples 83
3.8.4 Recovering the prior under zero observations 84
3.8.5 NPL sensitivity analysis 85
3.8.6 Privacy-preserving learning by sharing pseudo-datasets 85

4 Posterior bootstraps for sparse regression and mixed models 88
4.1 Introduction 88
4.2 A posterior bootstrap for sparse regression 89
4.2.1 Background 89
4.2.2 The NPLasso 92
4.2.3 Illustrations 94
4.2.4 Discussion 97
4.3 A posterior bootstrap for mixed models 99
4.3.1 Introduction 99
4.3.2 Inference for mixed models 101
4.3.3 H-likelihood and the Extended Likelihood Principle 103
4.3.4 The weighted h-likelihood bootstrap 105
4.3.5 Normal means example 107
4.3.6 Salamander mating dataset 110
4.3.7 Discussion 113

5 Discussion 116
5.1 Summary 116
5.2 Future work 117
5.2.1 Methodology 117
5.2.2 Implementation 118
5.2.3 Applications 119
5.2.3.1 The general supra-Bayesian 119
5.2.3.2 Randomising objective functions 120

References 121

List of Figures

1.1 Model misspecification schematic diagram. 12
1.2 Model construction schematic diagram. 15

2.1 Left: Margin loss plot for hinge loss (solid) and smooth hinge loss (dashed). Middle and right: marginal density plots for α (middle) and β (right) for the general Bayesian posterior (black) and the loss-likelihood bootstrap posterior (grey), for a synthetic dataset with $n = 100$. 45
2.2 Synthetic example plots. Left: General Bayesian probability of class 1, as a function of the covariate z, for $n = 100, 200, 500, 1000$ (from dark to light). Middle: Box plot of $\widehat{w}$ as a function of sample size. Right: Box plot of misclassification rate, relative to a linear support vector machine, for $n = 100$ observations, for general Bayes (GB) and the loss-likelihood bootstrap (LLB) methods. 46
2.3 Statlog German Credit plots. Left and middle: marginal density plots for α (left) and β_1 (middle) for the general Bayesian posterior (black) and the loss-likelihood bootstrap (grey). Right: Box plot of misclassification rate minus the support vector machine misclassification rate for the general Bayesian (GB) and loss-likelihood bootstrap (LLB) methods. 47

3.1 Posterior 95% probability contour for a bivariate Gaussian, comparing VB-NPL with $c \in \{1, 10^2, 10^3, 10^4\}$ (red, orange, green, blue respectively) to the known Bayes posterior (grey dashed line) and a VB approximation (black dashed line). 73
3.2 Posterior contour plot for β_{22} vs β_{21}, for VB-NPL (green) and VB (blue), for $c \in \{1, 1000, 20000\}$. Scatter plot is a Bayesian logistic posterior sample (red) using a Polya-Gamma scheme. 74
3.3 Box plot of classification accuracy minus that of RF, for 13 UCI datasets. . 76

3.4 Posterior histogram for median, with $c = 0, 20, 80, 1000$ (left to right). Right-most: posterior expected loss as a function of observation x. Dotted line shows the loss at sample median. 77
3.5 Mean classification accuracy vs c for RF (black solid), RF-all (black dashed), and BRF (red) for varying c. over 100 repetitions, for the Bank dataset. Error bars for BRF and the grey lines for RF and RF-all represent one standard error from the mean. The BRF with $c = 0$ has a mean accuracy of 0.895 and a standard error of 0.0004. 82
3.6 Schematic diagram of the NPL method for the Exponential family example of Section 3.5.1. 84
3.7 Flow diagrams for random forests and Bayesian random forests. 85

4.1 Mean Squared Error as a function of the penalty parameter λ, for the synthetic dataset. Error bars represent ± 1 standard error from the mean. . . 95
4.2 Profile plots of NPLasso (top-left and top-right) and Bayesian lasso (bottom-left) coefficients, as a function of $\| \beta \|_1$, and posterior probability of covariate inclusion for NPLasso (bottom-right), as a function of λ, for the prostate cancer dataset. 97
4.3 Normal means posterior moment plots with dispersion parameters known. Left: WHLB vs Bayes scatter plots for the three weight distributions, with $(n_1, I) = (100, 100)$. Right: WHLB vs Bayes scatter plots for all data structures and Independent Dirichlet weights. The dotted line represents equality in posterior summaries. 109
4.4 Normal means posterior plots with dispersion parameters unknown, for the three weight distributions, for the saturated configuration $(n_1, I) = (100, 100)$. The dotted line represents equality in posterior summaries. . . . 109
4.5 Normal means posterior plots with dispersion parameters unknown, for the four balanced data structures, using the Independent Dirichlet weight distribution. The dotted line represents equality in posterior summaries. . . 110
4.6 Pairs plots for the posterior distributions for two fixed effects and two random effects, for the Salamander dataset, for the Bayesian posterior and the WHLB posterior. 113

List of Tables

4.1 Average projected time for generating 1 million samples, computed over 10 repetitions. All runtime standard errors are below 6% of the mean. The rightmost two columns display the ratio of projected runtimes of the Bayesian lasso relative to the two NPLasso methods. 96
4.2 WHLB weight configurations considered. $n := \sum_i n_i$. All sets of Dirichlet-distributed weights are independent. 107
4.3 Parameter estimates for the Salamander dataset using 4 different methods. For Bayesian methods the posterior mean is shown, with posterior standard deviation as a subscript. For frequentist methods the maximum likelihood estimate is shown, with the standard error as a subscript. 112

Chapter 1

Introduction

1.1 Motivation

Gaining insights into an unknown quantity via data analysis is a vital aspect of the scientific method. Arguably the most successful means of achieving this is via the expedient use of statistical models, which invoke a probabilistic representation of uncertainty surrounding the observations that is linked to the unknown quantity of interest. Often these statistical models make strong assumptions that may seem acceptable for small-data problems, whereas in the presence of larger, higher-dimensional datasets the same assumptions may become questionable and could materially impact the outcomes of the analysis. In this thesis, we investigate honest ways of reasoning with data and models by seeking to relax or remove these strong assumptions.

A statistical model provides a precise statement concerning how data is distributed in terms of a *known* set of parameters with *unknown* values. This statement determines the range of permissible probability distributions for the data that are considered, specifying dependencies between different components of the data and specific parameters, as well as possibly distributional assumptions for the parameters in the form of a prior. Inaccuracies in these assumptions are almost certainly present, which means that the model is 'wrong' (Box,

1976). Despite this, in many data analysis tasks a carefully constructed statistical model can deliver valuable insight into the phenomena of interest.

The consequences of accepting dubious modelling assumptions can be severe. Decisions that are made 'optimally' conditional on a misspecified model being true can lead to erroneous measurements of risk and subsequent suboptimal decision making. Real-world examples of this happening are abundant in biomedical research (Hanin, 2017; Limpert and Stahel, 2011) through to financial crises (Brigo et al., 2010). From a methodological standpoint, the violation of model assumptions brings into question the meaningfulness and utility of any subsequent inferential statements made. From a theoretical standpoint, optimal procedures conditional on the model being true may cease to be optimal.

Over the last 50 years, humanity's capacity to capture, store, transmit and compute with data has seen immense growth (Marx, 2013). This information revolution has profoundly changed the way we live our lives, and has already led to some leaps forward in our understanding of the world; Polson and Scott (2018) describe several examples in depth. This progress has heightened the profile of statistical learning across many scientific and technological domains. However, many of the statistical methods used to learn from large datasets are not fully fit for modern purposes.

In this thesis we explore how statistical methodology could better satisfy the demands of modern data-rich applications by not overly restricting the extent to which the data can speak for itself. The differences between statistical models and the actual underlying distribution can be significant but is rarely accounted or corrected for. Statistical frameworks that utilise the many positive qualities of parametric statistical modelling, while remaining honest about the fact that the underlying model is misspecified, are explored.

Additionally, algorithmic approaches to data analysis, such as deep learning (LeCun et al., 2015), have increased in popularity in recent years. Such methods commonly forego the rigour of the statistical modelling approach to data analysis and can be lacking in terms of uncertainty quantification as a result. We explore general Bayesian methodologies that

can provide uncertainty quantification for such algorithms, on targeted parameters of interest, without the need to construct a global probability model at all.

We take a pragmatic Bayesian approach to this research agenda, as the strong foundations, prescriptive inferential framework and coherent uncertainty propagation available under this approach are invaluable for complex inference tasks. However, there are many connections made with frequentist procedures along the way. In fact, we argue that our work blurs the boundaries between Bayesian and frequentist inference. We study both parametric and nonparametric methods, while considering both subjective and objective Bayesian viewpoints. We start by investigating simple independent and identically distributed data sources, before advancing to more complex hierarchical data structures. Throughout this work computational scalability is considered to be of paramount importance, as it must be if the methods are to be relevant to datasets of ever-increasing scale.

In the rest of this introduction we shall delve into the foundations of statistical modelling, review current approaches in the literature for dealing with model uncertainty and model misspecification, and introduce a recently developed general Bayesian theory for coherent and targeted belief updating in the absence of a full statistical model. We conclude by highlighting the scientific contribution of this thesis.

1.2 Foundations of statistical modelling

1.2.1 Preliminaries

Suppose we have access to a dataset $x_{1:n} = (x_1, \ldots, x_n) \in \mathcal{X}^n \subseteq \mathbb{R}^{p_n}$ comprised of n observations from an unknown joint distribution $F_{0,n} \in \mathfrak{F}^n$, where $\mathfrak{F}^n$ is the set of all distributions over $\mathcal{X}^n$. Suppose further that we are interested in some aspect of $F_{0,n}$, written as $\theta_{0,n} := \theta(F_{0,n})$ for some functional $\theta(\cdot)$. We are interested in the following problem that is fundamental to data analysis: how should we learn about $\theta_{0,n}$ given $x_{1:n}$?

To make progress on this problem it is necessary to make some further simplifying

assumptions. For example, if each x_i is a repetition of a single experiment, it may be reasonable to assume that the $x_i \in \mathcal{X} \subseteq \mathbb{R}^p$ are independent and identically distributed according to a distribution function F_0. To simplify notation, we can now write $F_0 \in \mathfrak{F}$, where $\mathfrak{F}$ is the set of all distributions over $\mathcal{X}$, and our parameter of interest is $\theta_0 := \theta(F_0)$.

Under this assumption, if an arbitrarily large dataset is available, then there is a non-controversial way of determining θ_0 from the data. Considering just the first n observations, we could construct an empirical distribution function $\hat{F}_n$ from our dataset,

$$\hat{F}_n(x \in A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i \in A}$$

where $\mathbb{1}_{\cdot}$ is the indicator function. Under regularity conditions (Van Der Vaart, 1998, Chapter 5, for example) the value of the empirical functional $\theta(\hat{F}_n)$ will almost surely converge to the 'true' value $\theta_0 = \theta(F_0)$, as the number of observations tends to infinity; i.e. as $n \to \infty$ we have

$$\theta(\hat{F}_n) \to \theta_0 \quad \text{in probability.}$$

An arbitrarily large dataset will not generally be available, and even if it was, computing with it would provide a practical constraint to taking this approach. So instead we must study the harder problem of making inferences about θ_0 given a finite dataset. The statistical modelling approach to the data analysis problem proceeds by making further assumptions about the underlying data-generating mechanism. This simplification can lead to inductions about θ_0 given $x_{1:n}$ that shall be referred to as statistical inferences.

Statistical modelling is not the only approach to the data analysis problem. Indeed, other less principled approaches can be highly effective for particular domains. Statistical modelling is particularly well-suited to *problems of science*, where the goal of the analysis is to probe the underlying data-generating mechanism, as opposed to *problems of technology*, where the goal of the analysis is to perform an action, such as prediction. This dichotomy has appeared on numerous occasions in the literature; Lehmann (1990) attributes it to a

taught course by J. Neyman. Healy (1978) argues that, philosophically, statistics should be considered a technology, or an applied science if this is more palatable, as opposed to the unrelenting search for the truth that is science. Breiman (2001b) is critical of the sole use of statistical modelling, even for problems of science:

> The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. [...] If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

However Breiman's preferred *algorithmic* approach is often found lacking in terms of uncertainty quantification and interpretability. Efron and Hastie (2016) points out that it is the role of the modern-day statistician to work on understanding properties of algorithms better.

Returning to the statistical modelling approach for the time being, a *statistical model* is simply defined as a subset of $\mathfrak{F}$. In this thesis we sometimes refer to a statistical model as simply a *model*. We say that a statistical model $\mathcal{F}$ is *parametric* if it admits a finite-dimensional parametrisation; i.e. for a suitably chosen dimension $p \in \mathbb{N}$ and parameter space $\Phi \subseteq \mathbb{R}^p$ we can write

$$\mathcal{F} \equiv \mathcal{F}_\Phi := \{F_\phi \in \mathfrak{F} : \phi \in \Phi\}. \tag{1.1}$$

Otherwise we say $\mathcal{F}$ is a *nonparametric statistical model*. Classically, it is often assumed that the model is defined such that $\phi = (\theta, \lambda)$, where λ is a nuisance parameter (Cox and Hinkley, 1979, p. 3), though this is not a necessary restriction.

The role of the statistician is to carefully construct a statistical model that allows for honest, useful and well-principled inferences to be made about the parameter of interest,

given data. Quoting Pace and Salvan (1997):

> The statistical model must be able to succinctly describe the aspects that are of primary interest and must also be sufficiently flexible to allow a realistic description of additional aspects.

Stronger assumptions about F_0 commonly lead to stronger inferences. However, if our assumptions are wrong then our model may misrepresent important aspects of F_0 and hence our inferences about θ_0 may not be useful at all. Classically the statistician would strive for parsimony; i.e. find an adequate statistical model that is as simple as possible. However, there is a great danger in making any modelling assumption that the statistician is not sure will hold. As discussed in Box (1980), this could distort the outcome of the analysis and potentially exclude information relevant to the problem at hand. For this reason, Lindley (2000) suggested to "always choose the largest model that your computational powers will tolerate". Before the computer age, simple parametric models were the only tractable option available for most statistical analyses. Such models are by construction less complex to understand and implement than nonparametric alternatives. However, we argue in this thesis that as computational resources are now available in abundance, we should be using more flexible models that allow the data to speak for themselves.

1.2.2 Statistical inference

Statistical inference refers to the process of making statements about an unknown parameter, given data and any other information deemed relevant, using probability. There is not a single agreed-upon procedure for statistical inference; one reason for this is that there is more than one interpretation of probability. In this regard, there are two main schools of thought:

- *The frequentist school*, where probability is defined as the limit of frequencies of a repeated experiment. The parameter ϕ indexing the distribution function in the

statistical model is an unknown constant; it does not have a probability distribution. Functions of the data called statistics are chosen for statistical inference based on their long-run frequency properties.

- *The Bayesian school*, where probability is axiomatically developed as a personalistic quantification of degrees of belief. The uncertainty in ϕ is represented precisely by a probability distribution called a prior, and statistical inferences are derived logically, via Bayes' rule, from the conditional distribution of ϕ given the data.

The Bayesian approach dates back to Bayes (1763). Savage (1954) showed that subjective probability and utility theory can be arrived at by an axiomatic treatment of preferences over decision-theoretic actions and consequences. Bayesian analysis provides a prescription for rational decision-making, given a person's prior beliefs, and data.

We define a *Bayesian statistical model* as a statistical model plus a prior distribution π over all unknown parameters. Given a Bayesian statistical model and a utility function, the Bayesian statistician has a single well-defined method for decision-making, including inference; inference as a decision problem is discussed in Bernardo and Smith (2006, p. 67). Prior uncertainty about the unknown ϕ is updated to a posterior belief distribution by Bayes' rule:

$$\pi(\phi \mid x) = \frac{\pi(\phi) f_\phi(x)}{\int_\Phi \pi(\phi') f_{\phi'}(x) d\phi'}$$

Decision making should follow by maximising expected utility under this probability measure. This single *method* for inference and decision is a valuable property of the Bayesian approach. However, the actual *computation* in practice is far from trivial; see for example Robert and Casella (2005).

The validity of the Bayesian method hinges on the assumed model being 'true'. This is problematic in that all models are wrong, and the construction of an *adequate* model (Box and Tiao, 1973, p. 6) can be a formidable task. Undertaking such an endeavour may seem an inefficient use of resources if the purpose of the analysis is a single decision

relating to one aspect of the statistical model. Goldstein (1981) argue that there is seldom adequate information to construct a full prior distribution over all unknowns. For scientific applications, some deem it more appropriate to adopt an objective Bayesian view (Jeffreys, 1939) reliant on non-informative priors. A detailed review of these ideas can be found in Kass and Wasserman (1994).

The frequentist swaps the Bayesian's controversy of a prior for imprecision over which method to use. Further, as Chatfield (1995) points out, somewhat sharply:

> Frequentists who object to the 'guess-work' involved in obtaining priors should perhaps reflect that they also must 'hazard a guess' at a model.

A further problem is that there can be multiple, possibly conflicting, approaches to the same inference question. The seminal work of R. A. Fisher did much to turn classical statistics from a range of methods for specific problems into a well-defined scientific discipline (Fisher, 1922; Fisher, 1925; Fisher, 1934). Introducing requirements for estimators, such as consistency, sufficiency and efficiency, provided a much-needed structure for problems of estimation. Arguably Fisher's greatest contribution was his development of the concept of *likelihood*. The likelihood function is simply the probability of the observed data given the unknown parameter, written as a function of the unknown parameter (defined up to a constant of proportionality),

$$L(\phi; x_{1:n}) \propto f(x_{1:n} \mid \phi).$$

Fisher proposed that the likelihood should represent the state of information of the unknown parameters in the statistical model, given the observed data. This idea caught on and is now a central tenet of statistical methodology. It was formalised and further motivated by Barnard (1949), and subsequently Birnbaum (1962) in the form of the likelihood principle, which follows from principles of sufficiency and conditionality. Berger and Wolpert (1988) argue that a Bayesian approach is "the most sensible and realistic method of implementation"

of the likelihood principle.

Fisher's method of maximum likelihood continues to have a profound influence on frequentist statistics, due to its good properties and intuitive nature. However, Fisher was critical of much of the research that formalised these ideas, such as those relating to hypothesis testing (Neyman and Pearson, 1933) and decision theory (Wald, 1949). He also thought the Bayesian approach to be "founded upon an error, and must be wholly rejected" (Fisher, 1990, p. 9). However he did believe in a logic of inductive inference, without a prior, via a so-called fiducial argument (Fisher, 1990). Although his approach was flawed, the underlying ideas have attracted much interest since. One example of this is the recent study of confidence distributions (Schweder and Hjort, 2016).

Frequentist methods are widely used in both scientific and technological applications. Prior to the computer revolution, Bayesian methods were in practice reliant on simplifying assumptions or approximations. This changed as computational resources started to grow, with ever larger and more structured datasets requiring the application of more complex statistical models. For the frequentist the development of new methods in this paradigm was a nontrivial challenge, whereas for the Bayesian approach, the method was known, once the model was written down. For this reason the Bayesian approach started to attract interest from practitioners faced with complex data problems. In terms of Bayesian computation, a significant development in the early 1990s was that of Markov chain Monte Carlo (MCMC) (Gelfand and Smith, 1990; Tierney, 1994), which provided an algorithm for sampling from Markov chains with a stationary distribution equal to a specified posterior distribution. Functionals of interest could then be estimated by Monte Carlo approximation. As Clifford states in the discussion of Smith and Roberts (1993),

> What the authors have done today is to announce that from now on we can compare our data with the model that we actually want to use rather than with a model which has some mathematically convenient form. This is surely a revolution.

However, MCMC algorithms can be expensive computationally, require tuning and suffer from other issues, such as providing serially correlated samples that only targeted the posterior asymptotically. Green et al. (2015) provides a thorough review of recent developments in Bayesian computation. Bardenet et al. (2017) reviews potential efficiency improvements for 'tall' data, i.e. when there are a large number of observations relative to the dimension of the parameter. Numerous attempts have been made to incorporate parallelism into posterior sampling algorithms without any resounding success (Scott et al., 2016; Neiswanger et al., 2014; Wang and Dunson, 2014). Approximate approaches such as variational Bayes (Blei et al., 2017), the integrated nested Laplace approximation (Rue et al., 2009) and approximate Bayesian computation (Marin et al., 2012) have proven to be popular when MCMC is infeasible.

There are many reasons to favour a Bayesian approach to statistical inference for complex data analysis problems. Modelling and decision-making are separated, meaning that the scientist can focus on constructing an adequate model, in the knowledge that, given the information available, decision-making conditional on the model will be optimal. By assigning a distribution to the unknowns, the Bayesian approach provides a comprehensive and coherent quantification of uncertainty. Prior information can be incorporated into an analysis, and a 'leak-proof' (Box and Tiao, 1973) belief updating mechanism, in the form of Bayes' rule, is available.

The foundational arguments in favour of Bayesian analysis, such as those put forward in Bernardo and Smith (2006), are idealistic and of questionable practical utility, given the impossibility of correctly specifying a full likelihood and prior. Furthermore, when considering issues such as model selection it is useful to step outside of the Bayesian framework to assess how well a model may fit the data. Hence, throughout this thesis we take a pragmatic Bayesian approach: we use the idealisms of Bayesian analysis to guide methodological decisions, but acknowledge that Bayesian inferences cannot be blindly trusted, as no model is perfect.

1.3 Model misspecification

Problems arise when we rely on statistical models whose underlying assumptions deviate from reality. Statistical procedures are often justified by some notion of optimality, under the assumption that the data is distributed according to a member of the statistical model. Furthermore, decision-theoretic actions may prove to be suboptimal and their associated risks misleading. As all models are wrong, what are the consequences of using a statistical model that is misspecified?

There has been much research into this topic, from theoretical studies of large sample properties of inference algorithms under model misspecification, to the study of robust methods that provide more stable outcomes to a statistical analysis. In this section we present a review of this literature.

1.3.1 Misspecification in theory

We shall focus on the simple situation where the dataset $x_{1:n} = (x_1, \ldots, x_n)$ is comprised of n independent and identically distributed observations from an unknown distribution F_0. Suppose further that we have a parametric model $\mathcal{F}$ as defined in (1.1) which we use to learn about the indexing parameter $\phi \in \Phi$.

We say that a statistical model is *well specified*, or *true* if there exists $\phi_0 \in \Phi$ such that $F_0 \equiv F_{\phi_0}$, i.e. $F_0 \in \mathcal{F}$. Otherwise we say that a statistical model is *misspecified*. In this section we consider the behaviour of maximum likelihood estimation and Bayesian inference under misspecification. Figure 1.1 provides a schematic depiction of this setting.

Let $\hat{\phi}_n \equiv \hat{\phi}(x_{1:n})$ denote the maximum likelihood estimate (MLE) for ϕ given $x_{1:n}$:

$$\hat{\phi}(x_{1:n}) := \arg\max_{\phi \in \Phi} \sum_{i=1}^{n} \log f_\phi(x_i).$$

If the model is well specified then under mild regularity conditions (see for example Van Der Vaart (1998), Chapter 5), the MLE has the follow properties:

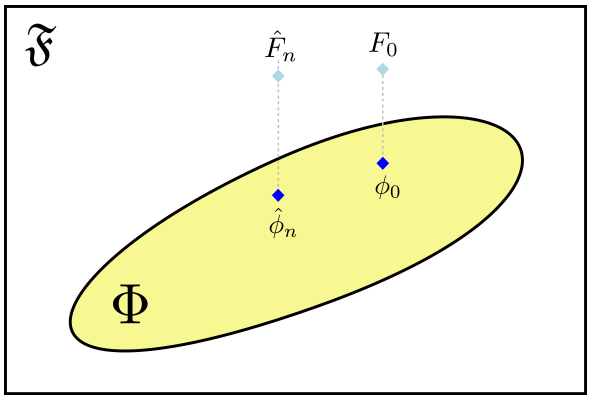


Figure 1.1: Model misspecification schematic diagram.

1. The MLE is consistent; i.e. $\hat{\phi}_n \to \phi_0$ in probability.

2. The MLE has an asymptotic normal distribution: $\sqrt{n}(\hat{\phi}_n - \phi_0) \to Z$, where $Z \sim \mathcal{N}(0, J^{-1})$, and J is the Fisher information matrix,

$$J = -\int \nabla^2_\phi \log f(x \mid \phi_0) dF_0(x). \tag{1.2}$$

Huber (1967) studies the behaviour of $\hat{\phi}_n$ without assuming that the model was well specified, showing under weak assumptions that

- the MLE converges in probability to the parameter value that minimises the Kullback-Leibler divergence (Kullback and Leibler, 1951) of F_ϕ to F_0:

$$\hat{\phi}_n \to \phi_0 := \arg\min_{\phi \in \Phi} d_{\mathrm{KL}}(f_0 \parallel f_\phi), \quad \text{in probability.}$$
$$d_{\mathrm{KL}}(f_0 \parallel f_\phi) := \int f_0(x) \log \frac{f_0(x)}{f_\phi(x)} dx$$

- the MLE has an asymptotic normal distribution with a sandwich covariance matrix: $\sqrt{n}(\hat{\phi}_n - \phi_0) \to Z'$, where $Z' \sim \mathcal{N}(0, J^{-1} I J^{-1})$, with J as in (1.2) and I is the covariance matrix of the score function,

$$I = \int \nabla_\phi \log f(x_{1:n} \mid \phi_0) \cdot \nabla_\phi \log f(x_{1:n} \mid \phi_0)^T f_0(dx).$$

As we would hope, the well-specified case is recovered as a particular case of the potentially misspecified theory. The limiting parameter value, ϕ_0, is sometimes called the *pseudo-true* parameter value, as it is recovered as $n \to \infty$ in the same way that the 'true' value is in the well-specified case. The sandwich covariance matrix plays an important role in robust statistics (Royall and Tsou, 2003; Cox, 1961); we shall come across it on a number of occasions in this thesis. Note that under misspecification the negative log likelihood can be viewed as a loss function for ϕ given $x_{1:n}$; the log likelihood has no special status relating the parameter to the data-generating mechanism.

Analysis of maximum likelihood estimation under misspecification in the literature is not restricted to independent and identically distributed observations. Gourieroux et al. (1984) summarises results for exponential families including covariates. Eicker (1967) studies regression models under heteroskedastic and dependent errors. White (1980) shows the sandwich covariance matrix remains consistent under heteroskedasticity. Consequences of misspecification for models with random effects have been the focus of recent research; see Jacqmin-Gadda et al. (2007) and Yu et al. (2018) for a discussion of linear and generalised linear mixed models, respectively.

In the Bayesian paradigm, the Bernstein-von Mises theorem states that under regularity conditions, a well-specified Bayesian posterior has an asymptotic normal distribution equivalent to that of the MLE (see Van Der Vaart (1998), Chapter 10). This notion of calibration is used as motivation for some Bayesian models (Welch and Peers, 1963). Berk (1966) studied the limiting behaviour of a misspecified Bayesian posterior. Under mild regularity conditions, the posterior distribution concentrates at the pseudo-true parameter value, as long as it is supported by the prior. The aforementioned calibration of Bayesian credible regions and maximum likelihood confidence regions is not guaranteed when the statistical model is misspecified. The asymptotic covariance matrix of the Bayesian posterior is the inverse Fisher information matrix, as opposed to the sandwich covariance matrix of the misspecified MLE (Kleijn and Vaart, 2012). Bunke and Milhaud (1998) extended these results to Bayes

estimates under a general loss function. The implications of using a Bayesian model under misspecification can be more severe than for maximum likelihood. Müller (2013) provides some formal results and examples of this.

Shalizi (2009) investigates the asymptotic properties of the misspecified Bayesian posterior when 'all hypotheses are wrong', in particular relaxing the independent and identically distributed assumption on F_0. O'Hagan (2013) questions the utility of Bayesian inference under misspecification. Grünwald and Ommen (2017) demonstrate that Bayesian inference can be inconsistent for misspecified linear models with a naturally optimal linear structure. Kleijn and Van Der Vaart (2006) and De Blasi and Walker (2013) explore the asymptotic properties of Bayesian nonparametric models.

1.3.2 Modelling in practice

For the results of the previous section to be of any utility, we need to have come up with a model in the first place. What procedure should we follow for constructing a suitable model for a particular problem?

This is a very contextual question. Models are chosen for a variety of reasons, including, but not limited to, previous usage for similar problems, suitability given knowledge of the field, desirable theoretical properties, recent methodological breakthroughs, desirable computational properties, flexibility and parsimony.

There is a significant literature on how we *should* go about this process, in particular several articles by G. E. P. Box (Box, 1976; Box, 1979; Box, 1980), also Box and Tiao (1973). In this work it is argued that scientific progress is made by iterating between conjecture and experiment. The initial conjecture should come from the scientific investigator. This conjecture is then put to the test by means of an appropriately designed experiment. Once the experiment has been carried out, the data collected from the experiment should be analysed. If the model is deemed adequate for answering the question of interest, the iterating can stop. Otherwise a follow-up model will be conjectured, and the iterations continue.

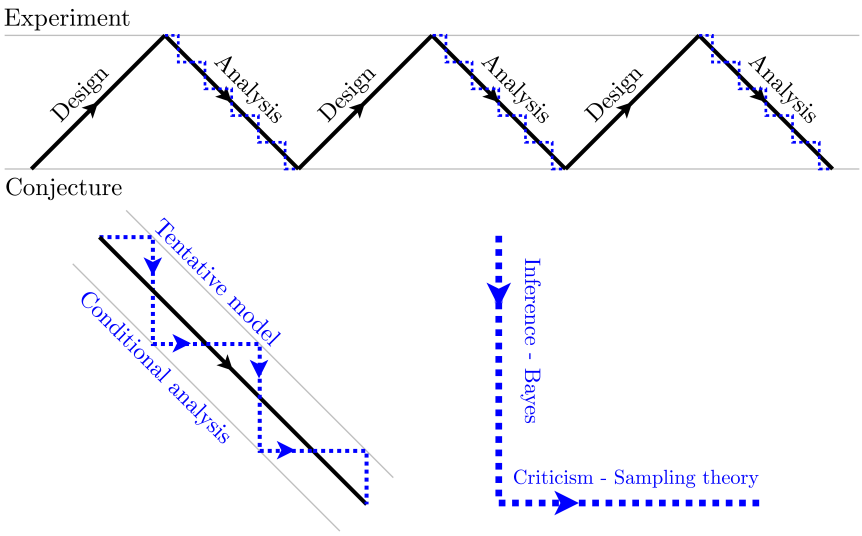


Figure 1.2: Model construction schematic diagram.

Each data analysis stage requires statistical modelling, and although we can carry forward previously-attained modelling knowledge, within each data analysis step there is a model building sub-iteration. A tentative model is proposed, then inferences are made, conditional on the model being true. Next, given these inferences, the model is criticised, leading to an updated model specification and the start of the next iteration. This iteration continues until an *adequate* model is settled on. A schematic of these model construction iterations can be found in Figure 1.2.

Model criticism can take many different forms. Classically this might involve the study of quantities such as fitted values, model parameters and residuals, often graphically (Anscombe and Tukey, 1963). If a Bayesian model is being used, a common approach is to use the posterior distribution to perform posterior predictive checks on diagnostic functions of the data (Gelman et al., 1996) potentially in comparison to some held-out data. Predictive distributions are a useful tool for model assessment as they are comparable across models; see for example Geisser and Eddy (1979). Box (1980) explains that both Bayesian and frequentist ideas play a role in model fitting:

> I believe that, subject to some overlap discussed later, sampling theory is needed for exploration and ultimate *criticism* of an entertained model in the light

> of current data, while Bayes' theory is needed for *estimation* of parameters conditional on the adequacy of the entertained model.

In practice scientists will have constraints that may conflict with this idealistic iterative approach. A scientifically appropriate model may not be computationally feasible. It may be that the modelling sub-iteration just does not converge satisfactorily. Finite resources may mean that not all aspects of the model's adequacy can be considered. Model construction will in practice deliver an imperfect model. As datasets grow in complexity, one would expect the extent of these imperfections to grow and the model criticism task to get harder.

For *problems of technology* much of the scientific approach presented above can still be applied, though the technologist's motives and preferences may differ from those of the scientist. As discussed by Bernardo and Smith (2006), in essence the scientist wants a good model for the data $p(x \mid \theta)$, whereas the technologist is more interested in $p(x)$, or some functional thereof. Although a good model of $p(x \mid \theta)$ should lead to a good predictive, it may well be that drastic simplifications to $p(x \mid \theta)$ may be good enough for the purposes of a single action. Overall, the scientist is likely to be more directly concerned about model misspecification than the technologist.

1.3.3 Robustness

How can we make informed choices about which models to consider during model construction? If we could find a model that provides stability in decision-making under small deviations of the data-generating mechanism away from the model class, this would give us a form of protection against model misspecification. As Huber (2009) points out, this does not come for free:

> And one should never forget that robustness is based on compromise, as was most clearly enunciated by Anscombe (1960) with his insurance metaphor: sacrifice some efficiency at the model, in order to insure against accidents caused by deviations from the model.

Within a Bayesian analysis there is no such thing as a robust method per se, as there is only one Bayesian method. There are, however, robust models. Box (1979) discussed model robustification as a means of obtaining a level of insensitivity to important potential deviations from assumptions. One common way of achieving this is to use a flat-tailed family of densities for the data, such as the student t distribution (West, 1984), as a means of downweighting outliers. Box (1980) argued that model elaboration, i.e. extending a model in a Bayesian fashion, is another sensible way to obtain targeted robustness.

As part of a comprehensive review of Bayesian robustness, Berger (1994) points out that robustness of not just the likelihood, but also the prior and any loss function associated with the analysis, should be considered. Indeed, much of the focus of Bayesian robustness is on the prior, as it is deemed by many to be the most questionable part of a Bayesian analysis. Commonly, robust Bayesian procedures start by determining a target functional, of interest to the scientist, related to the outcome of the analysis. Next, an underlying modelling assumption is perturbed, and the sensitivity of the value of the target is observed. Both global and local perturbations are considered; the global perturbations change a distributional assumption in a relevant neighbourhood of the assumed Bayesian model, whereas local perturbations take the form of functional derivatives. Computationally these approaches can be expensive and become less useful as the modelling complexity increases. As stated by Berger (1994),

> Statistical problems fall into several different categories. The most difficult are problems in which it is a challenge to perform any Bayesian analysis whatsoever. For such problems it will inherently be the case that formal Bayesian robustness cannot be investigated; at best, the informal "try a few models and priors" will be done.

Likelihood tempering, i.e. raising the likelihood to a power, has seen some popularity of late as a means of making a statistical model robust. From a classical standpoint, Royall and Tsou (2003) consider robustifying likelihoods in this way while conserving asymptotic

efficiency. Grünwald and Ommen (2017) describe a 'Safe Bayesian' procedure of adaptively tempering the likelihood, that can protect against heteroskedasticity. Walker and Hjort (2001) consider consistency of likelihood tempering in a Bayesian context. Miller and Dunson (2018) arrive at likelihood tempering by conditioning not on the data, but on the event that the model generates data close to the observed data. More recently, Wang et al. (2017) presented a reweighted likelihood method, where geometric weights for the likelihood contribution of each observation are computed as a means of obtaining model robustness.

The classical robust statistics literature studies robustness properties, such as efficiency, stability and breakdown, for classes of statistics such as M-estimators. For a detailed account of this rich literature, see Huber (2009).

1.3.4 Model uncertainty and $\mathcal{M}$-open

A number of statistical models will likely be considered during model construction. It is natural, then, to consider whether a superior model could be constructed by combining these sub-models in some way. In a Bayesian framework this can be achieved by Bayesian model averaging (Hoeting et al., 1999). Prior probabilities for each sub-model are assigned and propagated, given data, via Bayes' rule. Posterior model probabilities can then be used for decision making, effectively averaging across the submodels according to these posterior probabilities. Madigan and Raftery (1994) proved the predictive superiority of model-averaging over any of the sub-models, as measured by a logarithmic scoring rule.

Bayesian model averaging fully propagates and combines uncertainty from each sub-model, on the condition that one of the statistical models in the ensemble contains F_0. Bernardo and Smith (2006) call this the *$\mathcal{M}$-closed view* of model selection. This assumption is problematic, as Bernardo and Smith (2006, p.384) points out:

> Nature does *not* provide us with an exhaustive list of possible mechanisms and a guarantee that one of them is true. Instead, we ourselves choose the lists as part of the process of settling on a predictive specification that we hope will

prove “fit for purpose”.

Additionally, under mild regularity conditions, as the number of observations grows, the sub-model posterior probability mass concentrates fully on the model or models closest in Kullback-Leibler divergence to the true data-generating mechanism (Diebold, 1991). Thus the perceived practical benefit of contemplating a diverse set of sub-models is lost.

Bernardo and Smith (2006) also describe a more realistic view called $\mathcal{M}$-*open*, that acknowledges that no true model is available. In this case, Bayesian model averaging cannot formally be applied: what would the model probabilities mean in this case? In $\mathcal{M}$-open, no model has a special status connecting it to the true distribution, and uncertainty quantification loses some of its meaning.

Key et al. (1999) consider the problem of model selection in $\mathcal{M}$-open from a decision-theoretic standpoint. The model is chosen that maximises expected utility. Although this expectation is not available directly, it can be approximated by a cross-validation procedure. Walker and Gutierrez-Pena (1999) arrive at a similar method by assigning a Dirichlet process prior to the unknown distribution function. Clyde and Iversen (2013) considered $\mathcal{M}$-open model averaging as a decision problem by learning optimal model weights directly. Recently Li and Dunson (2016) suggested constructing model weights relative to a nonparametric reference model.

Many other statistical methods adhere to the $\mathcal{M}$-open view without this being formally asserted. These include Akaike’s information criterion for model selection (Akaike, 1981), the prequential approach to forecast assessment (Dawid, 1984), and a number of information-theoretic approaches such as those based on data compression via minimum description length codes (Grünwald, 1998).

For problems of technology in particular, the notion of *ensembling* a pool of weak sub-models or predictors to create a superior meta-model is popular and effective. Examples include bootstrap aggregating (bagging) (Breiman, 1996), boosting (Freund and Schapire, 1999) and stacking (Wolpert, 1992). Random forests is a popular classification and regression

algorithm that applies bagging to random decision trees. Fernández-Delgado et al. (2014) demonstrated its state of the art general predictive performance for black-box classification.

We take an $\mathcal{M}$-open view throughout this thesis. Statistical models, even if heavily misspecified, may still be useful tools for data analysis. Thus we seek uncertainty quantification for such models without assuming that the model is true. Throughout this thesis we find that the honesty of such an approach pays off in many ways.

1.3.5 Nonparametric approaches

As we have seen and discussed already, Bayesian parametric models make strong assumptions about the distribution of the data. As the number of observations becomes large, the imperfections of such models become clear to see. Quoting Hjort et al. (2010), Chapter 1:

> If a Bayesian has a prior distribution and suspects there is additional uncertainty, there are two possible actions. The first is to consider an alternative prior and then select one or the other after the data have been observed. The second action is to enlarge the prior before observing the data to cover the additional uncertainty. It is the latter action which is correct and coherent.

Moving towards using more adaptable models with broader support can provide reassurance that as the dataset grows, the model will still provide an adequate approximation to the distribution of the data.

Thus one way to mitigate the problems of model misspecification for large datasets is just to increase the range of distributions that the prior supports. This should mean that there is a better approximation to F_0 available in the model space, which will be located as the number of observations tends to infinity. Bayesian nonparametric priors support infinite-dimensional spaces of distribution functions, though to ensure tractability only support a restricted set of distributions. For example, perhaps the most common Bayesian nonparametric prior is the Dirichlet process (DP) (Ferguson, 1973). Draws of probability measures from the

Dirichlet process are almost surely discrete, but the smallest closed set of DP-probability 1 consists of all probability measures within the support of the DP's base measure. Bayesian nonparametric posteriors may lack smoothness for small sample sizes, however they have a strong asymptotic justification.

Bayesian nonparametric models are understandably more complicated, both theoretically and practically, compared to parametric Bayesian models. However, given an abundance of data a nonparametric analysis can provide a better approximation of the true data-generating mechanism. For this reason we find Bayesian nonparametric approaches suitable for a number of problems considered in this thesis.

An alternative to using a fully Bayesian nonparametric model is to construct a nonparametric likelihood. These are constructed from the data alone, and can sometimes then be used to update beliefs in a Bayesian analysis. Examples include empirical likelihood (Owen, 2001), synthetic likelihood (Wood, 2010) and the bootstrap likelihood (Davison et al., 1992). Boos and Monahan (1986) use a bootstrap estimate of the sampling density of a robust estimator in place of the likelihood in Bayes' rule. Lazar (2003) and Mengersen et al. (2013) are examples of Bayesian methods using empirical likelihood.

The nonparametric bootstrap (Efron, 1979) is a widely applicable frequentist method that provides a notion of uncertainty quantification by perturbing a dataset, by repeatedly sampling with replacement from that dataset. It is well suited to the modern computing paradigm, and has strong asymptotic motivations (Bickel and Freedman, 1981). There are many connections between Bayesian inference, likelihood methods and bootstrapping, notably the Bayesian bootstrap (Rubin, 1981) and the weighted likelihood bootstrap (Newton and Raftery, 1994). These are discussed in depth in Chapter 2.

1.4 Model-free methods

Many approaches to the data analysis problem outlined in Section 1.2.1 don't involve statistical models at all. We have discussed already the prevalence of algorithmic approaches for problems of technology. Algorithms are commonly assessed by their empirical performance, and in some cases can be lacking in interpretability; a so-called black box. Examples of such algorithms are deep learning (Goodfellow et al., 2016), random forests (Breiman, 2001a) and support vector machines (Vapnik, 2013). The outperformance of these algorithms compared to traditional modelling is vast in some domains. For example, in computer vision, deep convolutional neural networks have in recent years provided a leap forward in image classification accuracy (Krizhevsky et al., 2012).

Efron and Hastie (2016, p. 446) points out the importance of the role of the statistician in this new paradigm:

> A great amount of ingenuity and experimentation has gone into the development of modern prediction algorithms, with statisticians playing an important but not dominant role. [...] It is the job of statistical inference to connect "dangling algorithms" to the central core of well-understood methodology.

An often fruitful way of achieving this connection is to make an analogy with a well understood statistical method. Examples of this relating to the algorithms above include Polson (2017), Scornet et al. (2015) and Tipping (2001). A better understanding of the properties of such methods should lead to increased trust and subsequently a more widespread usage, particularly in application areas where interpretability is important. One example of this is the biomedical domain, where the social benefit of using state of the art algorithms is potentially great, but it is imperative for any applied method to be trustable (Pamies, 2018). Experts require interpretable and auditable explanatory tools to help build this trust. Making uncertainty quantification accessible for such methods would help in this regard, and this is a goal that we contribute towards in this thesis.

1.4.1 The general Bayesian approach

Given the expense of model construction, it's natural to ask if there is a work-around. How should a Bayesian update their beliefs about functionals of an unknown data-generating mechanism, without having to first construct a full probability model? In this section we shall introduce the general Bayesian approach of Bissiri et al. (2016), which provides prior-to-posterior belief updates for parameters defined by a loss function, under the assumption that the data-generating mechanism is unknown.

Suppose we are interested in learning about a *decision functional*, by which we mean a parameter that is defined as the minimiser of an expected loss, for some loss function $\ell : \Theta \times \mathcal{X} \to \mathbb{R}$,

$$\theta_0 = \theta(F_0)$$

$$\theta(F) = \arg\min_{\theta} \int \ell(\theta, x) dF(x)$$

We assume F_0 is unknown. Suppose we have prior beliefs $\pi(\theta)$ about the parameter θ, representing our subjective uncertainty in the true parameter value θ_0. This is different to a conventional Bayesian analysis where the prior must be specified for all unknown parameters indexing the family of distribution functions. The decision functional is often a more natural object over which to provide a prior distribution, as it is the ultimate quantity of interest in the analysis.

Suppose we have some data $x_{1:n}$ comprised of n independent draws from F_0. These samples clearly give us some information about $\theta(F_0)$, as discussed in Section 1.2.1, in the limit as $n \to \infty$ we could recover F_0 empirically and hence determine θ_0.

We shall consider the specific problem of updating the prior in the presence of data. A conventional Bayesian approach would be to first build a Bayesian statistical model for F, parametrised by ϕ, where ϕ is not necessarily the same as θ. A prior for ϕ would be constructed such that the marginal prior for θ is as specified. The posterior distribution for ϕ would be computed via Bayes' rule, from which uncertainty measures for θ could be derived.

This approach assumes that the prior for ϕ and the model are available and well specified. We have no guarantee about the optimality of making decisions under a misspecified model.

Surely, though, a prior to posterior belief update must exist, given data that has information about θ_0. Bissiri et al. (2016) argue this point, and show via a decision-theoretic formulation, with the posterior distribution as the 'action', that under a coherency condition this posterior belief update has a tractable functional form. We call this the general Bayesian update, and outline the main argument below.

We start by writing down a loss function for the optimal belief distribution, π^*, given the data and the prior. As π is a 'prior', it is argued that the loss function should be additive across the two information sources; i.e.

$$L(\pi^*; \pi, x) = L_{\text{data}}(\pi^*, x) + L_{\text{prior}}(\pi^*, \pi)$$

Next, a coherency condition is stipulated, such that updating the prior sequentially over several datasets will yield the same result as updating the prior once, given an augmented dataset. Bissiri and Walker (2012) showed that in this case L_{prior} must be the Kullback-Leibler divergence,

$$L_{\text{prior}}(\pi^*, \pi) = d_{\text{KL}}(\pi^* \parallel \pi) = \int \pi^*(d\theta) \log\{\pi^*(d\theta)/\pi(d\theta)\}.$$

For L_{data} a natural choice (Bissiri et al. (2016) argue the only choice) is to use the expected loss of the observed data under π^*:

$$L_{\text{data}}(\pi^*, x) = \int_{\Theta} \ell(\theta, x)\pi^*(d\theta)$$

The scale of either loss functions is of no consequence if considered in separation. However, when adding two loss functions together their scales determine the importance of one loss versus the other. Hence we add a loss scale w to the first term when writing down

our joint loss function:

$$\begin{aligned} L(\pi^*, \pi, x) &= \int_\Theta w\ell(\theta, x)\pi^*(\theta)d\theta + \int_\Theta \pi^*(\theta) \log\left[\frac{\pi^*(\theta)}{\pi(\theta)}\right] d\theta \\ &= \int_\Theta \pi^*(\theta) \log\left[\frac{\pi^*(\theta)}{\pi(\theta)\exp\{-w\ell(\theta, x)\}}\right] d\theta \end{aligned}$$

Under mild regularity conditions, this loss can be minimised explicitly, by interpreting L as a Kullback-Leibler divergence, up to a constant. We call this optimal belief distribution the general Bayesian posterior distribution $\pi(\theta \mid x)$:

$$\begin{aligned} \pi(\theta \mid x) &= \arg\min_{\pi^*} L(\pi^*, \pi, x) \\ &= \frac{\pi(\theta)\exp\{-w\,\ell(\theta, x)\}}{\int_\Theta \pi(\theta)\exp\{-w\,\ell(\theta, x)\}\,d\theta}. \end{aligned} \tag{1.3}$$

This provides us with a parametric functional form for the belief update of general parameters of interest, that bypasses the construction of a statistical model. Belief updating is targeted to the true quantity of interest to the practitioner. The framework is robust to model misspecification, as no statistical model is used. The practitioner can focus on eliciting a loss and prior, as opposed to spending time constructing a full statistical model. By adopting this framework, a quantification of uncertainty is available for general parameter types without further assumptions.

The argument of Bissiri et al. (2016) provides a formal justification for a number of statistical methods that are otherwise lacking in this regard. For example, Grünwald (1999) considered a class of 'safe' statistical methods that could be used for prediction give a loss function, arriving at an exponentiated utility function similar to (1.3), via the minimum description length principle (Grünwald, 1998). Zhang (2006) found that distributions of this form are optimal in terms of an information complexity measure. The study of probably approximately correct (PAC)-Bayes methods in machine learning has yielded probabilistic bounds for such posteriors (Shawe-Taylor and Williamson, 1997; Zhang, 2006). In the

econometrics literature, this type of posterior has been studied as a means of performing a Bayesian analysis for parameters defined by a moment condition (Chernozhukov and Hong, 2003).

The utility of general Bayesian posterior distributions has been demonstrated in many different settings; such a posterior is sometimes referred to as a *pseudo-posterior*, or *Gibbs posterior*. For example Jiang and Tanner (2008) consider their use in variable selection problems. Alquier et al. (2016) explores the properties of variational approximations to Gibbs posteriors. Grünwald (2011) uses this form of posterior for robust Bayesian learning. Jordan et al. (2018) demonstrate communication-efficient distributed learning covering this class of posteriors. Li et al. (2013) use Gibbs posteriors for a ranking problem. Syring and Martin (2017) use Gibbs posteriors to assess clinical significance without needing to specify a model.

Although the underlying motivation and applicability is promising, there are a number of open theoretical and methodological questions surrounding general Bayesian methods. As an example, we have introduced an arbitrary scaling to the loss in (1.3) - how should this be dealt with in practice? Another involves computation: the functional form of the posterior is suitable for posterior sampling methods such as those based on Markov chain Monte Carlo; however, these methods do not generally scale well to large datasets or parallel computer architectures. What can we do to improve on this? Regarding generalisation, much of the motivation above is applicable only for simple independent and identically distributed data structures. How should we extend the general Bayes methodology to non-independent and identically distributed settings? Regarding prior information, it may seem too weak an assumption that F_0 is simply treated as an unknown. How could we incorporate some structural prior information about F_0 into the general Bayesian framework?

The idea of general Bayesian updating is central to this thesis. It is an attractive way to approach uncertainty quantification for a range of modern statistical learning problems. In this thesis we contribute to a number of the open questions in this field.

1.5 Contribution

This thesis makes the following scientific contributions to the topics of statistical learning under model misspecification and general Bayesian methodology:

- In Chapter 2 we reinterpret the weighted likelihood bootstrap (Newton and Raftery, 1994) as a well-motivated, exact Bayesian nonparametric method that makes weak assumptions about the underlying data-generating mechanism. We generalise this method to a wider family of parameters, leading to a method that we call the loss-likelihood bootstrap. Asymptotic properties of these methods are studied. A method for calibrating the loss scale of the general Bayesian posterior via the loss-likelihood bootstrap is presented, by means of asymptotic information matching.

- In Chapter 3 we present a novel framework for Bayesian nonparametric learning using misspecified Bayesian parametric models. It is shown that the mixture of Dirichlet processes (Antoniak, 1974) is well suited for this task. The parametric model acts to regularise nonparametric learning. A concentration parameter representing faith in the parametric model enables a traversal between misspecified Bayesian posterior and a non-informative Bayesian bootstrap. It is shown that the resulting posterior is predictively optimal relative to a Bayesian approach under misspecification. The posterior bootstrap sampling algorithm performs scalable Bayesian inference. We extend this idea to apply to the correction of posterior approximations such as variational Bayes approximations, and a form of general Bayesian learning that incorporates a parametric model. We develop the notion of uncertainty quantification and Bayesian learning by the randomisation of objective functions, which enables inference and uncertainty quantification for machine learning algorithms that may not be based on a statistical model.

- In Chapter 4 we extend nonparametric learning in two directions. Firstly, we present nonparametric learning for sparse regression problems, via randomising a lasso objec-

tive function, as a principled means of quantifying uncertainty about sparse models given unknown data-generating mechanisms. This is advantageous over Bayesian shrinkage prior methods because the posterior samples are sparse, and has computational advantages over discrete mixture variable selection methods. Importantly, the method does not assume the data-generating mechanism is of a known parametric form. Secondly, we consider hierarchical data structures, where observations are grouped, with a conditionally independent and identically distributed structure. We present a posterior bootstrap for statistical models containing both fixed and random effects, based on an h-likelihood objective (Lee and Nelder, 1996). Statistical learning can be performed simultaneously, and at scale, on group-specific parameters, as well as across the distribution of group-level parameters. Both of the extensions are empirically validated on a number of examples.

This is an integrated thesis, consisting of publications and conventional thesis chapters. The content of Chapters 2 and 3 has been accepted for publication, the former in *Biometrika* (Lyddon et al., 2019) and the latter in *Advances in Neural Information Processing Systems 31, 2018* (Lyddon et al., 2018). These chapters reproduce each article in full, and additionally are preceded with a short preamble and succeeded by some additional discussion that did not appear in the original publication. Chapter 4 is written as a conventional thesis chapter. Chapter 5 concludes.

Chapter 2

General Bayesian Updating and the Loss-Likelihood Bootstrap

2.1 Preamble

In this chapter the initial goal was to develop a principled means of calibrating the general Bayesian posterior loss scale; the 'w' in (1.3). This was achieved by making connections between general Bayesian updating and Bayesian nonparametrics, and by reinterpreting Newton and Raftery (1994)'s weighted likelihood bootstrap as an exact Bayesian nonparametric method. This was joint work with Chris Holmes and Stephen Walker, and has been accepted for publication in *Biometrika*. A statement of the author's contribution, that has been agreed by all authors, is provided at the end of the chapter. The remaining content of the chapter, aside from Section 2.9, is precisely the accepted manuscript and its supplementary material. Section 2.9 provides some additional commentary on this work that is relevant to the thesis as a whole.

2.2 Abstract

In this paper we revisit the weighted likelihood bootstrap, a method that generates samples from an approximate Bayesian posterior of a parametric model. We show that the same method can be derived, without approximation, under a Bayesian nonparametric model with the parameter of interest defined as minimising an expected negative log-likelihood under an unknown sampling distribution. This interpretation enables us to extend the weighted likelihood bootstrap to posterior sampling for parameters minimizing an expected loss. We call this method the loss-likelihood bootstrap. We make a connection between this and general Bayesian updating, which is a way of updating prior belief distributions without needing to construct a global probability model, yet requires the calibration of two forms of loss function. The loss-likelihood bootstrap is used to calibrate the general Bayesian posterior by matching asymptotic Fisher information. We demonstrate the methodology on a number of examples.

2.3 Bayesian inference and model misspecification

Bayesian theory provides a comprehensive framework for quantitative reasoning about uncertainty which is built on axiomatic foundations and is well suited to many modern scientific applications. For the statistician, much of the validity of the Bayesian approach hinges on the condition that the statistical model for the data is well specified, in that it contains the underlying data-generating mechanism. This cannot be expected to hold in general, if at all, and the extent and impact of the misspecification can be hard to quantify.

In this article we investigate statistical methods that remain honest about the inability of the statistician to perfectly model the data. Suppose we observe a sample $x_{1:n} = (x_1, \ldots, x_n)$ with the $x_i \in \Omega \subseteq \mathbb{R}^p$ being independent draws from an unknown distribution F_0, which we assume admits a density, f_0. In a parametric Bayesian modelling paradigm, a family of densities $\mathbb{F}_\Theta = \{p(\cdot \mid \theta);\, \theta \in \Theta \subseteq \mathbb{R}^d\}$ is specified along with a prior belief

distribution $p(\theta)$ for the unknown parameter θ. Assuming that $f_0 \in \mathbb{F}_\Theta$, beliefs about the true parameter are updated by conditioning on the observed data, using Bayes' rule; $p(\theta \mid x_{1:n}) \propto p(\theta) \prod_{i=1}^n p(x_i \mid \theta)$.

We say that such a model is well specified if there exists a $\theta_0 \in \Theta$ supported by the prior such that $f_0(\cdot) = p(\cdot \mid \theta_0)$. If this is the case, then under mild regularity conditions the posterior will concentrate at θ_0 as the number of observations increases. Conversely, if $f_0 \notin \mathbb{F}_\Theta$, we say that the model is misspecified. In this case the posterior will, under mild regularity conditions, concentrate at the pseudo-true parameter value which minimizes the Kullback–Leibler divergence (Kullback and Leibler, 1951) to the true sampling distribution (Berk, 1966); i.e.

$$\theta_0 = \arg\min_{\theta \in \Theta} \int \log \left\{ \frac{f_0(x)}{p(x \mid \theta)} \right\} f_0(x) dx = \arg\max_{\theta \in \Theta} \int \log p(x \mid \theta) \, dF_0(x).$$

Müller (2013) showed that, under regularity conditions, the asymptotic frequentist risk associated with misspecified Bayesian estimators is inferior to that of an artificial posterior which is normally distributed, centred at the maximum likelihood estimator and with a certain covariance matrix. Additionally, the rate at which the Bayesian learns about the parameter of a misspecified model is not necessarily optimal, as it is in the well-specified case; see Zellner (1988).

It is therefore important that the statistician accounts for the inadequacy of the model when making inference about θ_0 and we shall study ways to do this. We show that the weighted likelihood bootstrap (Newton and Raftery, 1994), a method originally designed for approximate sampling from the Bayesian posterior of a well-specified parametric model, can also be viewed as generating exact posterior samples under a Bayesian nonparametric model which assumes much less about the structure of the data-generating mechanism than the parametric model does. This approach can then be extended to generate posterior samples for a much larger family of parameters than just those indexing parametric models.

More often than not, when a statistical model is used, it is with a specific action in mind, such as a prediction. For the Bayesian, the optimal action is chosen by maximising an expected utility or, equivalently, minimising an expected loss. To do this it is necessary to construct a full probability model for the data. This can be a prohibitively expensive exercise, if it is indeed at all possible. It is natural, then, to question whether this global modelling approach is necessary, or whether we could instead focus solely and directly on the functional that is of interest.

Bissiri et al. (2016) presented a general framework for updating targeted belief distributions of this kind. Instead of restricting themselves to parameters that index a family of distribution functions, the authors considered general parameters θ whose true value θ_0 minimizes an expected loss for some loss function $\ell : \Theta \times \Omega \to [0, \infty)$;

$$\theta_0 = \theta(F_0) = \arg\min_{\theta \in \Theta} \int \ell(\theta, x)\, dF_0(x). \tag{2.1}$$

Treating F_0 as unknown, Bissiri et al. (2016) use a decision-theoretic argument, relying on a coherency property, that leads to a unique functional form for the update of prior beliefs $p(\theta)$, given observations $x_{1:n}$. The resulting posterior distribution $p_{GB,w}$ is fully determined up to a loss scale $w > 0$, and given by

$$p_{GB,w}(\theta \mid x_{1:n}) \;\propto\; p(\theta) \exp\{-w\ell(\theta, x_{1:n})\}, \tag{2.2}$$

where the loss for multiple observations is defined additively, i.e. $\ell(\theta, x_{1:n}) = \sum_{i=1}^{n} \ell(\theta, x_i)$. We assume throughout this paper that this posterior distribution is proper; i.e. the loss and prior are provided such that the right-hand side of (2.2) is integrable.

We refer to this posterior distribution as the general Bayesian posterior, because it can be computed for a more general family of parameters than those simply indexing a family of distributions. Bayes' rule is recovered as a special case by choosing $w = 1$ and using the self-information loss, $\ell(\theta, x) = -\log p(x \mid \theta)$. We call the data-dependent component

$\exp\{-w\ell(\theta, x)\}$ the loss likelihood as it provides a prior-to-posterior belief update for the parameter θ, in analogy with the likelihood in the Bayesian setting. The loss scale w is a non-negative scalar controlling the learning rate about θ attributable to the observed data. If the learning rate is too large, the posterior will be too concentrated, exaggerating the extent of the information about θ coming from the data. Conversely if w is too small, the posterior will underplay the information in the data relative to the prior.

The theory presented in Bissiri et al. (2016) has a number of attractions compared to the Bayesian approach. It is built upon the assumption that the true underlying data-generating mechanism is unknown. It provides a principled means for performing targeted prior belief updates without the burden of having to construct a global probabilistic model. The prior specification is local to the parameter of interest. However, although a number of suggestions have appeared in the literature, the setting of w remains an open problem. We propose that a Bayesian bootstrap should be used for this purpose and we extend the weighted likelihood bootstrap's interpretation to cover the wider class of models built from loss functions, of the form given in (2.1). We refer to this method as the loss-likelihood bootstrap. The asymptotic structure of this bootstrap is studied, alongside that of the general Bayesian posterior. These theoretical results are used to determine a loss scale for the general Bayesian posterior by matching asymptotic posterior information. This provides us with a calibrated general Bayesian posterior with a number of desirable properties, such as the Bayes posterior being recovered if the model is well specified. The general Bayesian approach admits a subjective prior and provides a prior-to-posterior belief update, which in some settings will be preferable over the prior-free loss-likelihood bootstrap. As noted in Newton and Raftery (1994), the weighted likelihood bootstrap may provide a better approximation to the Bayesian posterior than a normal approximation, asymptotically, if the prior used is the square of the Jeffreys prior.

2.4 Revisiting the weighted likelihood bootstrap

The weighted likelihood bootstrap (Newton and Raftery, 1994) is a method for approximately sampling from a posterior distribution of a well-specified parametric statistical model. Samples are generated by computing randomly-weighted maximum likelihood estimates. The weights are drawn from a Dirichlet distribution, scaled by the number of observations. The weights perturb the contribution of each observation to the likelihood; see Algorithm 1 for details.

Algorithm 1. The Weighted Likelihood Bootstrap

For $j = 1$ to $j = B$:
Draw random weights $(g_{j1}, \ldots, g_{jn})$ with $n^{-1}(g_{j1}, \ldots, g_{jn}) \sim \text{Dirichlet}(1, \ldots, 1)$
Compute $\theta^{(j)} = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} p(x_i \mid \theta)^{g_{ji}}$
$= \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} g_{ji} \log p(x_i \mid \theta)$
Output $\left(\theta^{(1)}, \ldots, \theta^{(B)}\right)$

The method produces independent samples and is trivially parallelizable over $j = 1, \ldots, B$, which is advantageous over Markov chain Monte Carlo methods. However, the weighted likelihood bootstrap is not an exact method for sampling from the posterior of a parametric model and does not accommodate a prior. However, it is asymptotically first-order equivalent to a Bayesian posterior if the parametric model is well specified, with a higher order of asymptotic equivalence, under certain conditions, if the prior used is the square of the Jeffreys prior.

If interest is in the θ minimizing $-\int \log p(x \mid \theta)\, dF_0(x)$ then it is possible to arrive at a method functionally equivalent to the weighted likelihood bootstrap, but without assuming the parametric model is well-specified. Under misspecification, the data-generating distribution function is unknown and so it is appropriate to construct a prior on the sampling distribution function F, whose unknown true value is F_0. Uncertainty about θ is inherited from uncertainty about F. More precisely, a prior on F, say $P_{\mathcal{F}}$, induces a prior probability

P_Θ on Θ, via

$$\theta(F) = \arg\max_{\theta \in \Theta} \int \log p(x \mid \theta)\, dF(x); \tag{2.3}$$

the induced prior being $P_\Theta(\theta \in A) = P_{\mathcal{F}}[\{F : \theta(F) \in A\}]$.

The Bayesian nonparametric literature provides a number of methods for constructing priors over the space of distribution functions. The Dirichlet process prior (Ferguson, 1973) is a natural choice, as it is simple to use. The hyperparameter that determines the Dirichlet process prior is a finite measure α. Upon observing $x_{1:n}$ the posterior is also a Dirichlet process with unit mass added to the base measure at each observation; i.e.

$$\mathcal{L}\left(F \mid x_{1:n}\right) = \mathrm{DP}\left(\alpha + \sum_{i=1}^{n} \delta_{x_i}\right).$$

For a detailed account of the Dirichlet process, see Ghosal and Vaart (2017). Under regularity conditions the posterior for F will concentrate at F_0. In a noninformative setting, a small value of $\alpha(\Omega)$ is chosen. In the limit $\alpha(\Omega) \to 0$ the posterior distribution is supported only by the observations $x_{1:n}$, with Dirichlet-distributed probabilities for each state. This sampling procedure is commonly referred to as the Bayesian bootstrap (Rubin, 1981) as it has a limiting Bayesian interpretation and is closely associated to Efron's bootstrap (Efron, 1979). Posterior sampling under the Bayesian bootstrap is direct and fast and trivially parallelizable; see Algorithm 2 for further details.

Algorithm 2. The Bayesian Bootstrap

For $j = 1$ to $j = B$:
  Draw random distribution $F^{(j)} = \sum_{i=1}^{n} g_{ji}\, \delta_{x_i}$
    with $(g_{j1}, \ldots, g_{jn}) \sim \mathrm{Dirichlet}(1, \ldots, 1)$
  Compute $\theta^{(j)} = \theta\left(F^{(j)}\right)$ as in (2.3)
Output $\left(\theta^{(1)}, \ldots, \theta^{(B)}\right)$

If we apply the Bayesian bootstrap strategy to $\theta(F)$ as defined in (2.3), we recover the weighted likelihood bootstrap. Conceptually, however, the parametric and nonparametric

constructions are quite different. This can perhaps most easily be seen through the role that the random weights play. For the weighted likelihood bootstrap, the weights perturb the contribution of each observation to the likelihood, whereas for the Bayesian bootstrap the weights represent a posterior sample of the unknown distribution function.

The Bayesian bootstrap can produce posterior samples for a much larger family of parameters than just those that maximize expected log likelihoods, see for example Chamberlain and Imbens (2003). Of interest to us is the family of parameters that minimize an expected loss, as in (2.1). We call the method of posterior sampling for this family of parameters the loss-likelihood bootstrap; see Algorithm 3 for details. The reason for giving this name is that the method will prove important when we consider the problem of calibrating general Bayesian posterior distributions in Section 2.5.

Algorithm 3. The Loss-likelihood Bootstrap

For $j = 1$ to $j = B$:
  Draw random distribution $F^{(j)} = \sum_{i=1}^{n} g_{ji}\, \delta_{x_i}$
    with $(g_{j1}, \ldots, g_{jn}) \sim \text{Dirichlet}(1, \ldots, 1)$
  Compute $\theta^{(j)} = \arg\min_{\theta \in \Theta} \int \ell(\theta, x) dF^{(j)}(x)$
Output $\left(\theta^{(1)}, \ldots, \theta^{(B)}\right)$

Asymptotic properties of the weighted likelihood bootstrap have been studied in the 1991 University of Washington PhD thesis by M.A. Newton. It was shown that if the parametric model is well-specified then the distribution of samples under the weighted likelihood bootstrap is asymptotically first-order correct to a Bayesian posterior distribution. By this we mean that the probability laws $\mathcal{L}\{n^{1/2}(\theta - \hat{\theta}_n) \mid x_{1:n}\}$ and $\mathcal{L}\{n^{1/2}(\tilde{\theta}_n - \hat{\theta}_n) \mid x_{1:n}\}$ converge to the same limit as $n \to \infty$, almost surely with respect to $x_{1:n} \sim p(\cdot \mid \theta_0)$, where $\hat{\theta}_n$ is the maximum likelihood estimator, $\tilde{\theta}_n$ is a random draw from the weighted likelihood bootstrap, and θ represents the parameter under a Bayesian posterior.

The proof proceeds by determining the relevant asymptotic properties for both the weighted likelihood bootstrap under misspecification, and also for the loss-likelihood boot-

strap. The relevant generalized asymptotic theory is contained in the following theorem:

Theorem 1. *Let $\tilde{\theta}_n$ be a loss-likelihood bootstrap sample of a parameter defined in (2.1) with loss function ℓ, given n observations $x_{1:n}$, and let P_{LL} be its probability measure. Under regularity conditions, for any Borel set $A \subset \mathbb{R}^d$, as $n \to \infty$ we have*

$$P_{LL}\left\{n^{1/2}\left(\tilde{\theta}_n - \hat{\theta}_n\right) \in A \mid x_{1:n}\right\} \;\to\; P(z \in A),$$

a.s. $x_{1:\infty}$, where $z \sim N_d\{\,0,\; J(\theta_0)^{-1}I(\theta_0)J(\theta_0)^{-1}\,\}$, with

$$I(\theta) = \int_\Omega \nabla\ell(\theta, x)\nabla\ell(\theta, x)^{\mathrm{T}}\, dF_0(x) \quad \textit{and} \quad J(\theta) = \int_\Omega \nabla^2\ell(\theta, x)\, dF_0(x),$$

where ∇ is the gradient operator with respect to θ, and $\hat{\theta}_n = \arg\min_\theta n^{-1}\sum_{i=1}^n \ell(\theta, x_i)$.

Proof. The proof follows along the lines of the weighted likelihood bootstrap asymptotic normality proof in the 1991 University of Washington PhD thesis by M.A. Newton. Details can be found in the Supplementary Material. □

The asymptotic covariance matrix in Theorem 1 is a well-known quantity in the robust statistics literature; sometimes called the sandwich covariance matrix. It was shown in Huber (1967) to be the asymptotic covariance matrix for general, potentially misspecified, maximum likelihood estimators. This asymptotic distribution and that of the Bayesian posterior do not coincide if the model is misspecified, in general. Müller (2013) showed that the sandwich covariance matrix can lead to an improvement in frequentist risk over a misspecified Bayesian posterior. Others, such as Royall and Tsou (2003), have argued that the sandwich covariance matrix can be used to make misspecified likelihood functions robust.

2.5 Calibrating general Bayesian posteriors by asymptotic information matching

General Bayesian models admit a prior distribution and provide a belief update, for the same types of parameters as the loss-likelihood bootstrap, which does not admit a prior. However, before it can be used the general Bayesian loss scale, w in (2.2), must be calibrated. Under regularity conditions, the impact of a prior diminishes as the number of observations grows. This indicates a potential strategy for calibrating the general Bayes loss scale. We set w to ensure that asymptotically the general Bayesian posterior and the loss-likelihood bootstrap posterior contain the same amount of information.

We choose to measure asymptotic posterior information with the Fisher information number. As we shall see, it has a number of good properties and is simple to compute. We shall denote the Fisher information number as $K(p)$, which is defined as follows,

$$K(p) = \int \frac{|\nabla p(\theta)|^2}{p(\theta)} d\theta.$$

It takes the value zero for a flat posterior, and is positive otherwise, which is not true of differential entropy, one possible alternative. It is the sum of the marginal Fisher information for each dimension, which is a well-understood quantity in statistics that summarizes the amount of information in a sample about a parameter. Walker (2016) showed that the Fisher information number can be used to measure the information in a Bayesian experiment, in analogy with the work of Lindley (1956) who used differential entropy. Further, Holmes and Walker (2017) used the Fisher information number to calibrate a power likelihood temperature; see the discussion for further details. Ferentinos and Papaioannou (1981) investigates the properties of the Fisher information number as a measure of information.

The general Bayesian posterior of (2.2) has the following asymptotic normal distribution,

under regularity conditions:

$$n^{1/2}(\theta - \hat{\theta}_n) \to z' \quad \text{in distribution,} \quad \text{a.s. } F_0^\infty, \tag{2.4}$$

where $z' \sim N_d\{0, w^{-1}J(\theta_0)^{-1}\}$, and $\hat{\theta}_n$ is the empirical risk minimizer, i.e. $\nabla\ell(\hat{\theta}_n, x_{1:n}) = 0$. Regularity conditions and a proof of this result can be found in Chernozhukov and Hong (2003); see also the Supplementary Material.

Thus under regularity conditions, both the loss-likelihood bootstrap and general Bayesian posterior distributions are asymptotically normal, with the same centring and scaling but different covariance matrices. Asymptotically, the Fisher information number for a normal distribution is equal to the trace of the Fisher information matrix, so this is the quantity we use to match the general Bayesian posterior to the loss-likelihood bootstrap. We would not expect the general Bayesian and loss-likelihood bootstrap posteriors to match in general; we are simply using the notion that asymptotic information should match to calibrate the general Bayesian loss scale. Once the loss scale is determined the full general Bayesian posterior can be used without approximation.

The following lemma determines the loss scale required to match the Fisher information number of the general Bayesian posterior to the loss-likelihood bootstrap.

Lemma 1. *The value of the loss scale w which equates the Fisher information number of the asymptotic distributions, p_{LL} and $p_{GB,w}$, is*

$$w = \frac{\mathrm{tr}\left\{J(\theta_0)I(\theta_0)^{-1}J(\theta_0)^{\mathrm{T}}\right\}}{\mathrm{tr}\left\{J(\theta_0)\right\}}. \tag{2.5}$$

Proof. For $z \sim N_d(0, \Sigma)$, if $p(z)$ is the density of z, then the Fisher information number is $K(p) = \mathrm{tr}(\Sigma^{-1})$. Applying this result for the general Bayesian asymptotic posterior, (2.4), we have $K(p_{GB,w}) = \mathrm{tr}\{wJ(\theta_0)\}$. Similarly, for the loss-likelihood bootstrap asymptotic posterior we have $K(p_{LL}) = \mathrm{tr}\{J(\theta_0)I(\theta_0)^{-1}J(\theta_0)\}$. Equating these expressions gives the required result. □

In practice θ_0 is unknown so the empirical risk minimizer $\hat{\theta}_n$ can be used as a strongly consistent estimator of θ_0. Also, as F_0 is unknown, matrices I and J can be estimated empirically.

$$\widehat{w} = \frac{\mathrm{tr}\left\{ J_n(\hat{\theta}_n) I_n(\hat{\theta}_n)^{-1} J_n(\hat{\theta}_n)^{\mathrm{T}} \right\}}{\mathrm{tr}\left\{ J_n(\hat{\theta}_n) \right\}}, \tag{2.6}$$

where

$$J_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla^2 \ell(\theta, x_i)\,, \qquad I_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \{ \nabla \ell(\theta, x_i) \nabla \ell(\theta, x_i)^{\mathrm{T}} \}\,.$$

This methodology can be applied to a parameter defined by an arbitrary loss function. For the case where a self-information loss is used, and the data is generated from the associated sampling distribution for some value of θ_0, we still recover Bayes' theorem. This is shown in the following lemma; for which regularity conditions can be found in the Supplementary Material.

Lemma 2. *If* $f_0(x) = \exp\{ -w_0\, \ell(\theta_0, x) \}$ *for some* $\theta_0 \in \Theta$, $w_0 > 0$, *then* $w = w_0$.

Proof. We recall the result that for a regular density f_{θ_0},

$$-\int_{\Omega} \nabla^2 \log f_{\theta_0}(x)\, dF_0(x) \;=\; \int_{\Omega} \nabla \log f_{\theta_0}(x) \nabla \log f_{\theta_0}(x)^{\mathrm{T}}\, dF_0(x),$$

where F_0 has density $f_0 = f_{\theta_0}$. This immediately gives us $J(\theta_0) = w_0 I(\theta_0)$. Plugging this into (2.5) gives us $w = w_0$. □

2.6 Illustrations

2.6.1 Normal model with quadratic loss

Suppose we observe independent data from a p-dimensional multivariate normal distribution, $x_i \sim N_p(\theta_0, \Sigma_0),\ i = 1, \ldots, n$, and we have a loss function that is of quadratic form,

$$\ell(\theta, x) = \frac{1}{2}(x - \theta)^{\mathrm{T}} \Sigma_1^{-1} (x - \theta).$$

Such a quadratic loss is non-trivial, as it could be used to estimate multiple integrals of the type $\int h(x)\, dF_0(x)$, where in the loss function above we would take $x \to h(x)$.

We assume throughout that Σ_0 and Σ_1 are strictly positive-definite matrices. The parameter of interest is the population mean θ_0, regardless of Σ_0 and Σ_1. The loss covariance Σ_1 is free to be set by the practitioner; this changes the general Bayesian posterior but does not impact the loss-likelihood bootstrap posterior.

The loss-likelihood bootstrap amounts to repeatedly drawing Dirichlet weights $g_{1:n} \sim \mathrm{Dirichlet}(1, \ldots, 1)$ and computing the weighted mean $\tilde{\theta}_n = \sum_i g_i x_i$. Using standard properties of the Dirichlet distribution it can be shown that this posterior is centred at the sample mean with a covariance matrix given by

$$\mathrm{var}\left(\tilde{\theta}_n \mid x_{1:n}\right) = x_{1:n}^{\mathrm{T}} \left\{ \frac{1}{n(n+1)} I_{n\times n} - \frac{1}{n^2(n+1)} 1_{n\times n} \right\} x_{1:n},$$

where $I_{n\times n}$ is the n-dimensional identity matrix and $1_{n\times n}$ is the $n \times n$ matrix with each element equal to 1. As we know F_0 we can compute the expectation of these quantities,

$$E\left\{E\left(\tilde{\theta}_n \mid x_{1:n}\right)\right\} = \theta_0, \qquad E\left\{\mathrm{var}\left(\tilde{\theta}_n \mid x_{1:n}\right)\right\} = \frac{n-1}{n(n+1)}\Sigma_0.$$

Now let us consider the general Bayesian approach. If $\Sigma_0 = \Sigma_1$ then the loss function is equal, up to a constant, to the negative log likelihood associated with the data-generating

mechanism. We find from Lemma 1 that we should set $w = 1$. For this special case, general Bayesian updating coincides with Bayesian updating under a normal location model.

Given our knowledge of the distribution of the data, we can determine the matrices I and J using standard properties of the normal distribution. In particular, we have $I(\theta_0) = \Sigma_1^{-1}\Sigma_0\Sigma_1^{-1}$ and $J(\theta_0) = \Sigma_1^{-1}$. Using Lemma 1 we get the following expression for the loss scale that calibrates the asymptotic posterior Fisher information number,

$$w = \frac{\mathrm{tr}(\Sigma_0^{-1})}{\mathrm{tr}(\Sigma_1^{-1})}. \tag{2.7}$$

Suppose that the loss function is the standard quadratic loss ($\Sigma_1 = I_p$), and each dimension of x is independent, with $\Sigma_0 = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, the $p \times p$ diagonal matrix with diagonal elements $(\sigma_1^2, \ldots, \sigma_p^2)$. Plugging these expressions into (2.7) we obtain $w = p^{-1}\sum_{i=1}^{p}\sigma_i^{-2}$, i.e. the average precision. If $\max_i \sigma_i^2 < 1$, then the data is under-dispersed relative to the loss, if we view the loss as a negative log likelihood. This calibration enables the loss likelihood to better match the likelihood of the well-specified model. Specifically, we will obtain a $w > 1$, thus reducing the variance associated with our loss likelihood. Similarly if $\min_i \sigma_i^2 > 1$ then we obtain $w < 1$. Our loss likelihood will not always match the likelihood under the well-specified model, as the data's covariance structure has many more degrees of freedom than the scalar w available to calibrate the general Bayesian posterior. Reassuringly, Lemma 2 tells us that if $\Sigma_1 = w_0\Sigma_0$ we do obtain correct calibration; i.e. $w = w_0$.

More generally, given matrices Σ_0 and Σ_1 we obtain the following general Bayesian posterior:

$$p_{GB}(\theta \mid x_{1:n}) \;\propto\; p(\theta)\,\exp\left[-\frac{1}{2}(\theta-\bar{x})^{\mathrm{T}}\left\{\frac{n\,\mathrm{tr}(\Sigma_0^{-1})}{\mathrm{tr}(\Sigma_1^{-1})}\Sigma_1^{-1}\right\}(\theta-\bar{x})\right],$$

where $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$. For univariate data this setting of w ensures that the loss likelihood is equal to the likelihood, and thus the general Bayesian update will coincide with the Bayesian update.

In practice the distribution of the data will be unknown. In this case we still have $J(\theta) = \Sigma_1^{-1}$ for all θ, and $I(\theta) = \int \Sigma_1^{-1}(x-\theta)(x-\theta)^{\mathrm{T}}\Sigma_1^{-1} dF_0(x)$. As θ_0 is the population mean, we can see that $I(\theta_0) = \Sigma_1^{-1}\,\mathrm{C}\,\Sigma_1^{-1}$, where C is the population covariance matrix for x. So we can rewrite w as $w = \mathrm{tr}(\mathrm{C}^{-1})/\mathrm{tr}(\Sigma_1^{-1})$. If $p = 1$, an unbiased estimator $\widehat{w^{-1}}$ of w^{-1} is the sample variance divided by the loss variance. In this case, $\widehat{w} = 1/\widehat{w^{-1}}$ may be a biased but consistent estimator of w. The plug-in estimator of (2.6) is available for $p > 1$.

2.6.2 Bayesian support vector machine for binary classification

Consider the problem of binary classification where we observe $\{x_i = (y_i, z_i) \in \{-1, 1\} \otimes \mathbb{R}^p \;;\; i = 1, \ldots, n\}$, with each $x_i \sim F_0$, where z_i denotes the covariates of an observation belonging to class y_i. We would like to predict future classes y^* given their covariates z^*. Specifically, the objective is to learn about the optimal linear classification rule that minimizes the expected loss under ϕ, a margin-based loss function,

$$(\alpha_0, \beta_0) = \arg\min_{\alpha,\beta} \int \phi\left\{y\left(\alpha + \beta^{\mathrm{T}} z\right)\right\}\, dF_0(y, z).$$

In a general Bayesian framework, we can compute a posterior distribution for the parameter of interest $\theta = (\alpha, \beta)$ given just the loss function, prior beliefs for θ and a loss scale. Our work in this paper provides a means for determining the loss scale. The loss-likelihood bootstrap also provides a posterior sample without requiring a loss scale, however it does not admit a prior.

Often the loss of interest is the $0-1$ loss, $\phi_{0-1}(\xi) = \mathbb{1}(\xi \leq 0)$, whose non-convexity leads to a number of computational problems when optimizing under this loss. A popular approach is to use a convex surrogate loss in place of ϕ_{0-1}. Bartlett et al. (2006) explore this idea formally, and provides some conditions for a surrogate loss to be Bayes-risk consistent to the $0-1$ loss.

A popular classification method in the machine learning literature is the support vector

machine (Cortes and Vapnik, 1995). Applied to linear classifiers as considered above, this method amounts to penalized optimization of a convex surrogate loss:

$$(\widehat{\alpha}, \widehat{\beta}) = \arg\min_{\alpha,\beta} \sum_{i=1}^{n} \phi_0 \left\{ y_i \left(\alpha + \beta^{\mathrm{T}} z_i \right) \right\} + \frac{\lambda}{2} \parallel \beta \parallel^2 .$$

Here $\lambda \geq 0$ is a hyperparameter and $\phi_0(\xi) = \max(0, 1 - \xi)$ is often referred to as the hinge loss. The optimization can be solved efficiently as a convex quadratic programming problem. The output is a linear classification rule that performs well empirically on a wide range of classification problems. The method doesn't provide any uncertainty quantification about the optimal linear boundary. We consider how we can use the ideas developed in this paper to provide uncertainty quantification about the optimal linear discriminant.

The problem with using the hinge loss in our framework is that we use first and second derivatives of the loss in the calculation of w. The hinge loss can easily be smoothed however; see for example Zhang (2004). We construct a smoothed hinge loss ϕ_2 that coincides with the first three derivatives of ϕ_0 outside of $(0, 1)$ and is a monotonic polynomial in between:

$$\phi_2(\xi) = \begin{cases} \frac{1}{2} - \xi, & \text{if } \xi < 0 \\ \xi^6 - 3\xi^5 + \frac{5}{2}\xi^4 - \xi + \frac{1}{2}, & \text{if } \xi \in [0, 1] \\ 0 & \text{otherwise.} \end{cases}$$

A comparison of this loss function to the standard hinge loss can be found in Fig. 2.1.

We can sample from the loss-likelihood bootstrap or the calibrated general Bayes posterior of the parameter that minimises the expectation of $\ell(\alpha, \beta, x) = \phi_2 \left\{ y \left(\alpha + \beta^{\mathrm{T}} z \right) \right\}$. We propose the following prediction routine for a new observation $\widehat{y}$ given the covariates of the new observation z^*. For each element of a posterior sample (α_i, β_i) we compute the class prediction $\widehat{y}_i = \mathrm{sign}(\alpha_i + \beta_i^{\mathrm{T}} z^*)$. We then predict the modal class across all of these predictions.

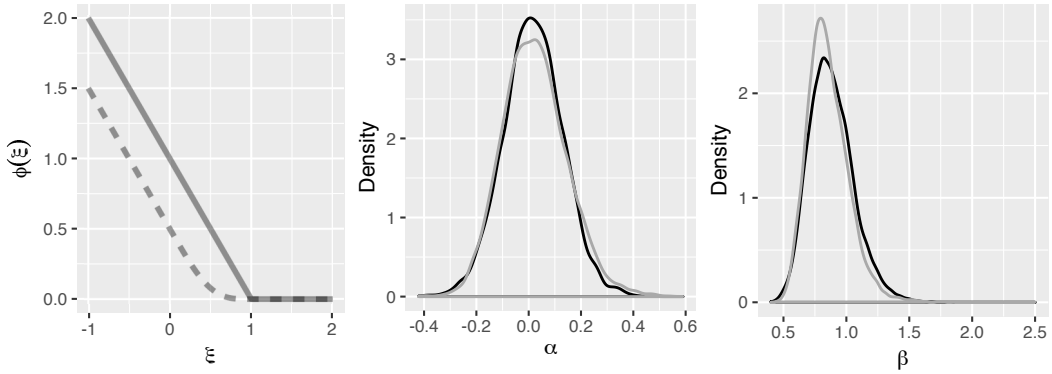


Figure 2.1: Left: Margin loss plot for hinge loss (solid) and smooth hinge loss (dashed). Middle and right: marginal density plots for α (middle) and β (right) for the general Bayesian posterior (black) and the loss-likelihood bootstrap posterior (grey), for a synthetic dataset with $n = 100$.

We test these classification methods on synthetic datasets from the following distribution:

$$\mathbb{P}(y = 1) = \frac{1}{2} = \mathbb{P}(y = -1), \quad \text{with} \quad p(z \mid y) = N(z \mid y, 1).$$

A linear classification boundary attains the Bayes risk under the $0 - 1$ loss. It is of the form $\widehat{y} = 2\,\mathbb{1}(z > 0) - 1$, which implies that $\alpha = 0$ and $\beta > 0$. Our classification posterior concentrates here for large sample sizes.

We generated synthetic datasets and assigned independent $N(0, 10^2)$ priors to α and β. For the general Bayesian posterior we computed a loss scale using the plug-in estimator $\widehat{w}$ of w specified in (2.6), under the smooth hinge loss ϕ_2. We generated general Bayesian posterior samples using the Hamiltonian Monte Carlo routine implemented in probabilistic programming language Stan (Carpenter et al., 2017). We also generated samples from the loss-likelihood bootstrap posterior. The marginal density plots for α and β, are shown in Fig. 2.1, given a synthetic dataset containing 100 observations. The general Bayesian posterior distribution matches well the loss-likelihood bootstrap, though the general posterior is calibrated to match asymptotic posterior information, as opposed to coverage.

The left pane of Fig. 2.2 shows the general Bayesian posterior probability that $\alpha + \beta z > 0$

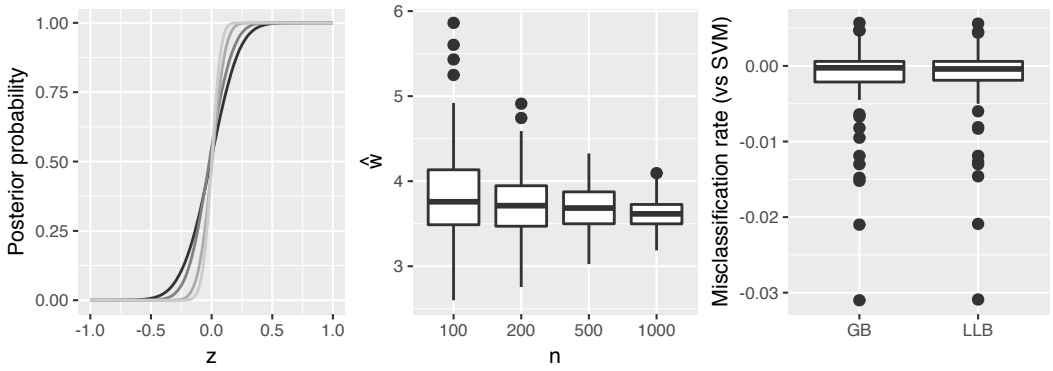


Figure 2.2: Synthetic example plots. Left: General Bayesian probability of class 1, as a function of the covariate z, for $n = 100, 200, 500, 1000$ (from dark to light). Middle: Box plot of $\widehat{w}$ as a function of sample size. Right: Box plot of misclassification rate, relative to a linear support vector machine, for $n = 100$ observations, for general Bayes (GB) and the loss-likelihood bootstrap (LLB) methods.

given z, as a function of z, averaged over 100 repetitions, for various sample sizes. These probability curves steepen as the number of observations increase, at $z = 0$, showing that the posterior mass does concentrate at the optimal classification boundary. The loss scale estimate $\widehat{w}$ for various sample sizes, and the misclassification error under a training dataset of 100 samples and a test dataset of $10\,000$ samples, relative to the misclassification error of a linear support vector machine, are also displayed in Fig. 2.2. These are based on 100 randomly generated synthetic datasets. We used the e1071 support vector machine implementation in R. Posterior samples of size 1000 were used for prediction; for the general Bayesian posterior these were collected after discarding 1000 samples as burn-in. The plug-in loss scale $\widehat{w}$ exhibits higher variance for small datasets. Additionally our plug-in estimator $\widehat{w}$ exhibits some bias which diminishes with the number of observations. The performance of our classification routines is very similar to a linear support vector machine.

As a more challenging example, we took the Statlog German Credit dataset from the UCI Machine Learning Depository (http://archive.ics.uci.edu/ml), preprocessed following Fernández-Delgado et al. (2014). It contains 24 covariates across 1000 customers, where each customer belongs to one of two classes relating to the lending experience of the

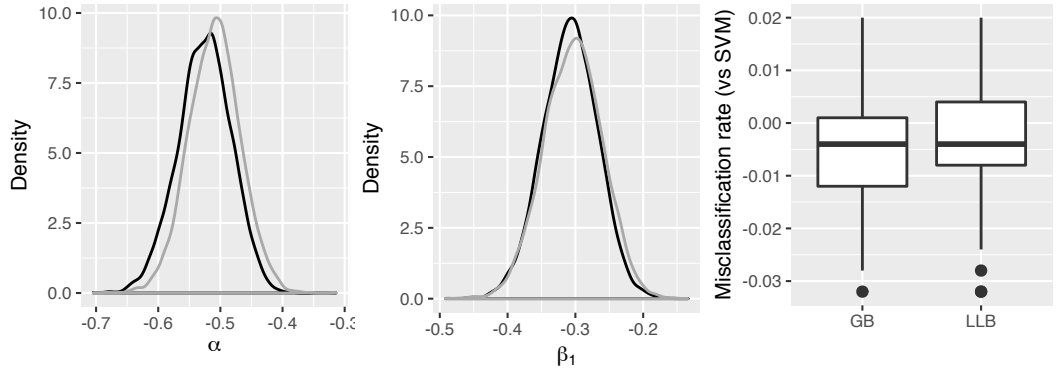


Figure 2.3: Statlog German Credit plots. Left and middle: marginal density plots for α (left) and β_1 (middle) for the general Bayesian posterior (black) and the loss-likelihood bootstrap (grey). Right: Box plot of misclassification rate minus the support vector machine misclassification rate for the general Bayesian (GB) and loss-likelihood bootstrap (LLB) methods.

customer.

The following test was repeated 100 times: we randomly split our dataset 75:25 to obtain a training and a test dataset. For the general Bayesian method, we use an independent $N(0, 10^2)$ prior on each covariate dimension and the intercept. We computed $\widehat{w}$ and generated a posterior sample of size $B = 1000$ in the same way as in the synthetic dataset example. The test dataset misclassification rate was recorded in comparison to that of a linear support vector machine.

The classification performance of our general Bayesian and loss-likelihood bootstrap procedures is very similar to that of the support vector machine method that does not provide uncertainty quantification. Posterior marginals for α and the first component β_1 of β are displayed in Fig. 2.3 for a single run. They seem to be well aligned again. The loss-likelihood bootstrap is well suited to parallel computation; to generate $100,000$ samples across a 48 CPU server took 3 minutes, compared to 137 minutes on a single CPU, a speed up of 45 times.

Both of our methods can easily be adapted to cost-sensitive classification problems, by use of an asymmetric hinge loss (Scott, 2012).

2.7 Discussion

In this paper we have presented two methods for generating samples from posterior distributions targeting a parameter defined by a loss function. Both methods make minimal assumptions about the data-generating mechanism.

The loss-likelihood bootstrap is a prior-free Bayesian nonparametric method. It does not require any calibration. Computationally, the method centres around optimization. It is well-suited for scalable Bayesian inference: independent samples can be generated in parallel with none of the practical inconveniences of MCMC approaches to standard Bayesian posterior sampling.

General Bayesian updating provides a decision-theoretic posterior under a coherency condition and a proper prior. It does require the calibration of data information relative to the prior information, for which we have provided a new method using the loss-likelihood bootstrap.

The prior-free nature of the loss-likelihood bootstrap may be unappealing to the subjective Bayesian. Previous attempts to incorporate a prior in similar settings include importance weighting (Newton and Raftery, 1994) and marginally specified priors (Kessler et al., 2015); both suffer from a curse of dimensionality. The loss-likelihood bootstrap doesn't satisfy the coherency condition of Bissiri et al. (2016), either. Suppose we have two datasets from different distributions, with a common parameter of interest. The general Bayesian posterior can be computed, using a loss-likelihood with a different loss scale for each dataset. However, using the loss-likelihood bootstrap in this setting is problematic as it does not provide a belief update.

Our calibration argument relies on the matching of asymptotic posterior information, which is well motivated as both methods target the same parameter. In Bissiri et al. (2016) a number of possible alternative means of calibration are discussed, including subjective calibration, unit information loss matching and hierarchical treatment via a loss function on w. Although these ideas may suit particular applications, they are less well motivated

in general, compared to the proposal in this paper. Syring and Martin (2018) provide an iterative method for setting w by ensuring calibration of a single user-specified posterior credible region. However, this method is computationally demanding and switches the issue of setting the loss scale to one of choosing a single credible region to calibrate.

For the special case of a potentially misspecified self-information loss, Holmes and Walker (2017) set the loss scale by matching the expected information gain from a single observation, for two experiments. In the first experiment the self-information loss is well specified in relation to the distribution of the observations. In this case the Bayesian update is optimal, so $w = 1$. In the second experiment the data comes from an unknown distribution so a general Bayesian update with self-information loss is used. For self-information losses w acts as a tempering parameter on the likelihood. The Fisher information distance is used to measure the prior-to-posterior information gain. The data is used to estimate these quantities.

The approach of Holmes and Walker (2017) focuses on prior information, whereas our method calibrates asymptotic posterior information to a bootstrap. One important quality of our method is that it recovers the parametric Bayesian learning rate when the model is correct up to an arbitrary tempering, i.e. any $w_0 > 0$ in Lemma 2, whereas Holmes and Walker (2017) will only recover the correct rate if the correct log likelihood is used; i.e $w_0 = 1$. However, Holmes and Walker (2017) is not an option in the general loss case as we do not have a benchmark experiment for which w is known. The criterion of Fisher information matching with the loss-likelihood bootstrap is applicable to calibrating posterior distributions based on loss functions, under minimal assumptions about the underlying data generating mechanism. This includes self information loss, $\ell(\theta, x) = -\log p(x \mid \theta)$, as a particular case.

Acknowledgements

The authors are grateful for the comments and suggestions of 6 referees which have led to a substantial improvement to the paper. Lyddon is funded by the Engineering and Physical Sciences Research Council. Holmes gratefully acknowledges support for this research from the Medical Research Council, the Engineering and Physical Sciences Research Council, The Alan Turing Institute, Health Data Research UK, and the Li Ka Shing foundation. Walker is partially supported by the U.S. National Science Foundation.

2.8 Supplementary material

2.8.1 Details of Theorem 1

2.8.1.1 Background

This theorem is a generalization of a result in the 1991 University of Washington PhD thesis by M.A. Newton that proved asymptotic normality of the weighted likelihood bootstrap. The generalization we provide covers the misspecified case, as well as a parameter defined by a generalized loss function. For the well-specified case the weighted likelihood bootstrap distribution is first-order correct to the Bayes posterior, which was part of the original justification for using the weighted likelihood bootstrap as an approximate Bayesian method.

The entire proof is in full as Chapter 3 of Newton's thesis. To avoid much repetition, we shall just highlight here the differences required for our generalized proof to work, both in terms of regularity conditions and the argument of the proof itself.

In terms of notation we follow the paper where we can. As a recap, let $x_{1:\infty}$ be an independent sequence of observations from distribution function F_0, where $x_i \in \Omega \subseteq \mathbb{R}^p$. We have a loss function $\ell(\theta, x)$, and care about a parameter $\theta_0 \in \Theta$, where

$$\theta_0 = \arg\min_{\theta \in \Theta} \int \ell(\theta, x) dF_0(x). \tag{2.8}$$

Let $\{(g_{n1}, \ldots, g_{nn})\}_{n=1}^{\infty}$ denote the $\mathrm{Dirichlet}(1, \ldots, 1)$ weights used by the weighted bootstrap, which are independent of F_0.

Let ∇ be the gradient operator with respect to θ. We call $S_n(\theta) = n^{-1} \sum_{i=1}^{n} \nabla \ell(\theta, x_i)$ the score function, in analogy with the log likelihood case in well-specified statistical theory. Similarly we will call J_n the sample (generalized) information matrix.

$$J_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla^2 \ell(\theta, x_i)$$

We add a tilde to the score and information functions to define the weighted versions:

$$\tilde{S}_n(\theta) = \sum_{i=1}^{n} g_{ni} \, \nabla \ell(\theta, x_i),$$

$$\tilde{J}_n(\theta) = \sum_{i=1}^{n} g_{ni} \, \nabla^2 \ell(\theta, x_i), \qquad \tilde{I}_n(\theta) = \sum_{i=1}^{n} g_{ni} \, \nabla \ell(\theta, x_i) \nabla \ell(\theta, x_i)^T.$$

Let $\hat{\theta}_n$ be the empirical risk minimizer statistic for $x_{1:n}$,

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} n^{-1} \sum_{i=1}^{n} \ell(\theta, x_i).$$

This takes the place of the maximum likelihood estimator in our generalized theory - the two estimators are the same when the loss function used is a negative log likelihood.

2.8.1.2 Conditions

We require the following set of regularity conditions relating to the loss function ℓ and the distribution of the observations F_0.

Condition 1 (Loss function). *The loss function $\ell : \theta \times \Omega \to \mathbb{R}$ is a measurable bounded from below function, with $\int \ell(\theta, x) \, dF_0(x) < \infty$ for all $\theta \in \Theta$, where Θ is a compact and convex subset of a d-dimensional Euclidean space.*

Condition 2 (Identifiability). *There exists a unique minimizing parameter value* $\theta_0 = \arg\min_{\theta\in\Theta} \int \ell(\theta, x)\, dF_0(x)$*, and further for all* $\delta > 0$ *there exists an* $\epsilon > 0$ *such that*

$$\liminf_n P\left\{ \sup_{|\theta-\theta_0|\geq\delta} \frac{1}{n}\sum_{i=1}^{n} [\ell(\theta, x_i) - \ell(\theta_0, x_i)] \geq \epsilon \right\} = 1.$$

Condition 3 (Smoothness of loss). *There is an open ball* B *containing* θ_0 *such that all first, second and third partial derivatives of* $\ell(\theta, x)$ *with respect to* $\theta \in B$ *exist and are continuous for almost all* x*. Furthermore, there exist measurable functions* G_j*,* G_{jk}*,* G_{jkl} *and* H_{jkl} *such that for* $\theta \in B$ *we have*

$$\begin{aligned}
\left|\frac{\partial\ell(\theta,x)}{\partial\theta_j}\right| &\leq G_j(x) && \text{with} \quad \int G_j(x)\, dF_0(x) < \infty,\\
\left|\frac{\partial^2\ell(\theta,x)}{\partial\theta_j\partial\theta_k}\right| &\leq G_{jk}(x) && \text{with} \quad \int G_{jk}(x)\, dF_0(x) < \infty,\\
\left|\frac{\partial^3\ell(\theta,x)}{\partial\theta_j\partial\theta_k\partial\theta_l}\right| &\leq G_{jkl}(x) && \text{with} \quad \int G_{jkl}(x)\, dF_0(x) < \infty,\\
\left|\frac{\partial\ell(\theta,x)}{\partial\theta_j}\frac{\partial^2\ell(\theta,x)}{\partial\theta_k\partial\theta_l}\right| &\leq H_{jkl}(x) && \text{with} \quad \int H_{jkl}(x)\, dF_0(x) < \infty.
\end{aligned}$$

Condition 4 (Positive definiteness of information matrices). *The matrices* $I(\theta)$ *and* $J(\theta)$ *are positive definite for* $\theta \in B$ *with all elements finite, where*

$$\begin{aligned}
I(\theta) &= \int_\Omega \nabla\ell(\theta, x)\nabla\ell(\theta, x)^T\, dF_0(x),\\
J(\theta) &= \int_\Omega \nabla^2\ell(\theta, x)\, dF_0(x).
\end{aligned}$$

Condition 1 is a reasonably standard condition on the loss function, and is the natural generalization of conditions on the log likelihood to the generalized loss function case. Condition 2 ensures identifiability for the loss case and is inspired by Chernozhukov and Hong (2003), in which asymptotic results for a general Bayesian posterior are developed (we will discuss this further in Section 2.8.2). Much of the proof revolves around Taylor expansions of the weighted score or information matrix, for which smoothness conditions

provided by Condition 3 are required. A first derivative condition has been included that is not required in the well-specified case, to ensure boundedness. This is covered in the original theory by a condition on the density and the information equality $I(\theta_0) = J(\theta_0)$ that holds for regular models. Additionally, positive definiteness of both I and J is required for the generalized loss function case, whereas in the well-specified theory the information equality is again used in place of this.

2.8.1.3 Main result

Theorem 1. *Let $\tilde{\theta}_n$ be a loss-likelihood bootstrap sample of a parameter defined in (2.8) with loss function ℓ, given n observations $x_{1:n}$, and let P_{LL} be its probability measure. Under conditions 1-4, for any Borel set $A \subset \mathbb{R}^d$, as $n \to \infty$ we have*

$$P_{LL}\left\{n^{1/2}\left(\tilde{\theta}_n - \hat{\theta}_n\right) \in A \mid x_{1:n}\right\} \;\to\; P(z \in A),$$

a.s. $x_{1:\infty}$, where $z \sim N_d\{\, 0,\; J(\theta_0)^{-1} I(\theta_0) J(\theta_0)^{-1} \,\}$.

Outline of proof. The proof follows the same structure as the well-specified version which can be found in the 1991 PhD thesis of M.A. Newton (Chapter 3, up to Theorem 7). First a number of properties of the weighted score and weighted sample information are shown to be similar to their unweighted versions. The weighted score function $\tilde{S}_n(\theta_0)$ converges to 0 in conditional probability, and the weighted observed information converges uniformly in an open ball containing θ_0. These results follow by replacing log likelihood with negative loss, if the generalized conditions 1-3 are satisfied. The inclusion of a first derivative condition (the first equation of condition 3) ensures that differentiating under the integral sign is valid and along with conditions 1-2 ensures that $\int \nabla \ell(\theta, x) dF_0(x) = 0$.

The remaining results required to prove consistency of the loss-likelihood bootstrap use a Taylor series argument where our generalized condition 3 replaces the original log likelihood ones. The weighted sample information evaluated at $\hat{\theta}_n$ is shown to converge in

conditional probability to the (expected) information, and is positive definite with probability converging to 1 (with n). Here the only difference is the distinction we must make in the misspecified case between I and J, as in general $I(\theta_0) \neq J(\theta_0)$. Condition 3 enables us to just work with J here. The positive definite result requires our additional condition 4 which ensures that both I and J are positive definite. Existence and uniqueness follows as in the well-specified case using the ideas of Foutz (1977), who applies the inverse function theorem to obtain consistency and uniqueness.

For the asymptotic normality result, we do a Taylor expansion of the weighted score function about $\hat{\theta}_n$:

$$\tilde{S}_n(\hat{\theta}_n) = (\, \tilde{J}_n(\hat{\theta}_n) - R_n \,)(\tilde{\theta}_n - \hat{\theta}_n)$$

where R_n is a remainder term. First we use the smoothness assumptions of condition 3 to show that $\tilde{J}_n(\hat{\theta}_n) - R_n$ converges to $J(\theta_0)$ and is invertible with high probability for large n. Next the Cramér-Wold device is used to show that $n^{1/2}\tilde{S}_n(\hat{\theta}_n) \to N_d\{0, I(\theta_0)\}$ in conditional distribution. The main result then follows by an application of Slutsky's theorem. □

2.8.2 General Bayesian posterior asymptotics

We state a result from Chernozhukov and Hong (2003) which provides us with, under regularity conditions, the asymptotic normality result for a general Bayesian posterior distribution. We require conditions 1-4, and two further conditions. These extra conditions are taken from Lemma 2 of Chernozhukov and Hong (2003), who also provide a discussion of the conditions used for the interested reader.

Condition 5 (Asymptotic normality of score function). *There is a positive definite matrix* $\Lambda(\theta_0)$ *with finite elements such that*

$$\Lambda(\theta_0)^{-1/2} n^{-1/2} \nabla \ell(\theta_0, x) \to N(0, I)$$

in distribution.

Condition 6 (Information matrix condition). *For some $\delta > 0$ and each $\epsilon > 0$ we have*

$$\limsup_{n\to\infty} P\left\{ \sup_{|\theta-\theta_0|<\delta} \left| n^{-1}\nabla^2 \ell(\theta, x_{1:n}) - \int \nabla^2 \ell(\theta, x) dF_0(x) \right| > \epsilon \right\} = 0.$$

This provides us with a convergence result for our general Bayesian posterior distribution.

Theorem 2. *Under conditions 1-6 we have*

$$\int | \, p(\, t \mid x_{1:n} \,) - p_\infty(\, t \,) \, | \, dt \to 0$$

in probability, where

$$t = n^{1/2}(\theta - \hat{\theta}_n)$$

and

$$p_\infty(t) = [\det J(\theta_0)]^{1/2} \, [2\pi]^{-d/2} \, \exp\left\{ -\frac{1}{2} t^T J(\theta_0) t \right\}$$

Proof. See Chernozhukov and Hong (2003), Theorem 1 and Lemma 2. □

Finally, conditions 1-6 are sufficient for Lemma 2 of the paper.

2.9 Further discussion

2.9.1 Bayesian bootstraps

Rubin (1981) presented the Bayesian bootstrap as a means of criticising the nonparametric bootstrap;

> Because both methods of drawing inferences are based on somewhat peculiar model assumptions and the resulting inferences are generally sensitive to these assumptions, neither method should be applied without some consideration of

the reasonableness of these assumptions. In this sense, neither method is a true bootstrap procedure yielding inferences unaided by external assumptions.

The Bayesian bootstrap, like any model, is misspecified; indeed for small datasets the assumptions underpinning it may seem unreasonable. Additionally, it is incoherent in that the support of the posterior over F changes as the dataset grows. Some may be uncomfortable with the use of an improper prior. However, it does have very broad support, which is highly advantageous relative to parametric statistical modelling, when an abundance of data is available. Asymptotic properties of the Bayesian bootstrap have been studied, including Lo (1987) who explored the first order equivalence of the Bayesian bootstrap and nonparametric bootstrap for a number of target functionals. Gasparini (1995) characterised the Bayesian bootstrap distribution of the population mean in terms of a multivariate B-spline.

Different weight distributions for the Bayesian bootstrap have been investigated. Weng (1989) showed that Gamma$(4, 1)$ weights achieve second-order agreement between Bayesian bootstrap credible interval and bootstrap confidence intervals. Praestgaard and Wellner (1993) extended the nonparametric bootstrap to exchangeable observation weights, and provided conditions for the asymptotic validity of the nonparametric bootstrap to be conserved. Under the standard nonparametric bootstrap, a weight w_i is assigned to observation x_i, with $nw_i \sim$ Multinomial$(n; n^{-1})$. Suppose instead that $w_i = W_i / \sum_j W_j$ with the W_i independent and identically distributed according to some known distribution. Then, under regularity conditions, for this *IID-weighted bootstrap* to satisfy the same central limit theorem as the nonparametric bootstrap, it is required that

$$\operatorname{var} W_1 = (\mathbb{E}\, W_1)^2.$$

This holds for the Bayesian bootstrap, as $W_i \sim \text{Exp}(1)$.

Newton and Raftery (1994) is highly cited but received a mixed reception, and the weighted likelihood bootstrap procedure has been scarcely used as it was intended in

applied work. Clyde and Lee (2001) and Chapter 1 of Hjort et al. (2010) both mention a connection between the Bayesian bootstrap and the weighted likelihood bootstrap but this is not developed further. Fushiki (2010) used the Bayesian bootstrap in favour of the nonparametric bootstrap for problems where the non-zero probability of a weight being zero leads to degeneracy problems. For example, in least squares computation when the number of observations is similar to the number of covariates, there may not be a unique least squares estimate for a bootstrap sample. Recently, Newton et al. (2018) has explored the potential for using the Bayesian bootstrap in machine learning applications.

2.9.2 Alternatives to Fisher information matching

The Fisher information number is most commonly used when the distribution of interest is only known up to an incalculable normalising constant. For example, in Holmes and Walker (2017) this property is used to give a closed-form expression for a power likelihood coefficient. However, in this chapter, this is not what is important, as our information measure is derived from the asymptotic normal distribution of the posterior. If they so wished, the practitioner could choose to calibrate the general Bayesian posterior distribution by another information measure of their choosing. For example if one were to choose differential entropy as the information metric,

$$K_{\text{ent}}(p) = \int p(\theta) \log p(\theta) d\theta$$

then w could be selected by a ratio of determinants as opposed to using (2.5),

$$\hat{w}_{\text{ent}} = \left(\frac{\det J(\theta_0)}{\det I(\theta_0)} \right)^{1/p}.$$

In Section 2.5 we point out a number of reasons in favour of using the Fisher information number.

2.9.3 Non-uniqueness of the loss

The normal model example of Section 2.6.1 highlights an aspect of general Bayesian updating that may seem undesirable at first sight. If our parameter of interest is the population mean, $\mathbb{E}x$, then any loss function of the form

$$\ell(\theta, x) = \frac{1}{2}(x-\theta)^T\Sigma_1^{-1}(x-\theta),$$

for any positive definite matrix Σ_1, targets that parameter. It is troubling that different choices of Σ_1 can give quite different general Bayesian posterior distributions. However, the loss provided here is doing more than just defining a parameter: it must encapsulate the individual's preferences. We don't have the freedom of choosing Σ_1, this will be determined as part of the loss elicitation, if the loss is truly quadratic.

2.9.4 Utility of general Bayesian uncertainty quantification

In Section 2.6.2 a general Bayesian support vector machine was constructed. The general Bayesian posterior samples encompass the uncertainty in the optimal linear decision rule, under a smooth hinge loss. This is different to a Bayesian approach, where posterior samples represent uncertainty in a parameter that determines the distribution of the data. Knowledge of this leads to an optimal prediction rule. The value of the risk is lost in the general Bayesian approach; the general Bayesian posterior parameter is already the optimal classification rule. At first this may seem less useful than the Bayesian posterior, but actually the Bayesian posterior only has this interpretation if the model is well specified.

The story for the loss-likelihood bootstrap is different, as we have posterior samples for $F \mid x_{1:n}$. Conditional on F, we can now just compute the optimal prediction given that F, but also the risk of a given prediction rule. This highlights one disadvantage of bypassing model construction in favour of targeted belief updating. The targeted approach means we lose information about phenomena that are not captured in the loss itself.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

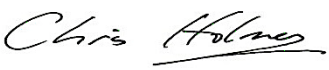
Title of Paper	General Bayesian Updating and the Loss-Likelihood Bootstrap
Publication Status	Accepted for publication.
Publication Details	Lyddon, S. P., Holmes, C. C., & Walker, S. G. (2018). General Bayesian Updating and the Loss-Likelihood Bootstrap. In *Biometrika (to appear)*.

Student Confirmation

Student Name:	Simon Lyddon		
Contribution to the Paper	Formulated the scientific idea. Undertook the literature review. Developed novel methodology. Developed theoretical results. Generated and interpreted material for illustrations. Wrote manuscript (redrafted by co-authors).		
Signature		Date	08/01/2019

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor Chris Holmes		
Supervisor comments The paper was largely conceived and developed by the candidate with guidance from the other authors. The candidate's description of their contribution is accurate.			
Signature		Date	08/01/2019

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 3

Nonparametric learning from Bayesian models with randomized objective functions

3.1 Preamble

Although all statistical models are wrong, some are useful (Box, 1976) and despite misspecification can provide useful insights into aspects of the data-generating mechanism. In this chapter we incorporate misspecified Bayesian parametric models into a Bayesian nonparametric learning algorithm and study the properties of this approach. This was joint work with Stephen Walker and Chris Holmes, has been presented at *Neural Information Processing Systems 31, 2018* and will be published in the conference proceedings. A statement of this author's contribution that was agreed by all authors is provided at the end of the chapter. The remaining content of the chapter, aside from Section 3.8, is exactly the accepted manuscript and supplementary material of this publication. In Section 3.8 we provide some additional comments and perspective that are of relevance to the thesis as a whole.

3.2 Abstract

Bayesian learning is built on an assumption that the model space contains a true reflection of the data generating mechanism. This assumption is problematic, particularly in complex data environments. Here we present a Bayesian nonparametric approach to learning that makes use of statistical models, but does not assume that the model is true. Our approach has provably better properties than using a parametric model and admits a Monte Carlo sampling scheme that can afford massive scalability on modern computer architectures. The model-based aspect of learning is particularly attractive for regularizing nonparametric inference when the sample size is small, and also for correcting approximate approaches such as variational Bayes (VB). We demonstrate the approach on a number of examples including VB classifiers and Bayesian random forests.

3.3 Introduction

Bayesian updating provides a principled and coherent approach to inference for probabilistic models (Robert, 2007), but is predicated on the model class being true. That is, for an observation x and a generative model $F_\theta(x)$ parametrized by a finite-dimensional parameter $\theta \in \Theta$, then for some parameter value $\theta_0 \in \Theta$ it is that $x \sim F_{\theta_0}(x)$. In reality, however, all models are false. If the data is simple and small, and the model space is sufficiently rich, then the consequences of model misspecification may not be severe. However, data is increasingly being captured at scale, both in terms of the number of observations as well as the diversity of data modalities. This poses a risk in conditioning on an assumption that the model is true.

In this paper we discuss a scalable approach to Bayesian nonparametric learning (NPL) from models without the assumption that $x \sim F_{\theta_0}(x)$. To do this we use a nonparametric prior that is centered on a model but does not assume the model to be true. A concentration parameter, c, in the nonparametric prior quantifies trust in the baseline model and this is

subsequently reflected in the nonparametric update, through the relative influence given to the model-based inference for θ. In particular, $c \to \infty$ recovers the standard model-based Bayesian update while $c = 0$ leads to a Bayesian bootstrap estimator for the object of interest.

Our methodology can be applied in a number of situations, including:

[S1] Model misspecification: where we have used a parametric Bayesian model and we are concerned that the model may be misspecified.

[S2] Approximate posteriors: where for expediency we have used an approximate posterior, such as in variational Bayes (VB), and we wish to account for the approximation.

[S3] Direct updating from utility-functions: where the sole purpose of the modelling task is to perform some action or take a decision under a well-specified utility function.

Our work builds upon previous ideas including Newton and Raftery (1994) who introduced the weighted likelihood bootstrap (WLB) as a way of generating approximate samples from the posterior of a well-specified Bayesian model. Lyddon et al. (2019) highlighted that the WLB in fact provides an exact representation of uncertainty for the model parameters that minimize the Kullback-Leibler (KL) divergence, $d_{\mathrm{KL}}(F_0, F_\theta)$, between the unknown data-generating distribution and the model likelihood $f_\theta(x)$, and hence is well motivated regardless of model validity. These approaches, however, do not allow for the inclusion of prior knowledge and do not provide a Bayesian update as we do here.

A major underlying theme behind our paper, and indeed an open field for future research, is the idea of obtaining targeted posterior samples via the maximization of a suitably randomized objective function. The WLB randomizes the log-likelihood function, effectively providing samples which are randomized maximum likelihood estimates, whereas we randomize a more general objective function under a Bayesian nonparametric (NP) posterior. The randomization takes into account knowledge captured through the choice of a model and parametric prior.

3.4 Foundations of Nonparametric Learning

We begin with the simplest scenario, namely [S1], concerning a possibly misspecified model before moving on to more complicated situations. It is interesting to note that all of what follows can also be considered from a viewpoint of NP regularization, using a parametric model to centre a Bayesian NP analysis in a way that induces stability and parametric structure to the problem.

3.4.1 Bayesian updating of misspecified models

Suppose we have a parametric statistical model, $\mathcal{F}_\Theta = \{f_\theta(\cdot);\ \theta \in \Theta)\}$, where for each $\theta \in \Theta \subseteq \mathbb{R}^p$, $f_\theta : \mathcal{X} \to \mathbb{R}$ is a probability density. The conventional approach to Bayesian learning involves updating a prior distribution to a posterior through Bayes' theorem. This approach is well studied and well understood (Bernardo and Smith, 2006), but formally assumes that the model space contains the true data-generating mechanism. We will derive a posterior update under weaker assumptions.

Suppose that $\mathcal{F}_\Theta$ has been selected for the purpose of a prediction, or a decision, or some other modelling task. Consider the thought experiment where the modeller somehow gains access to Nature's true sampling distribution for the data, $F_0(x)$, which does not necessarily belong to $\mathcal{F}_\Theta$. How should they then update their model?

With access to F_0 the modeller can simply request an infinite training set, $x_{1:\infty} \overset{iid}{\sim} F_0$, and then update to the posterior $\pi(\theta|x_{1:\infty})$. Under an infinite sample size all uncertainty is removed and for regular models the posterior concentrates at a point mass at θ_0, the parameter value maximizing the expected log-likelihood, assuming that the prior has support there; i.e.

$$\theta_0 = \arg\max_{\theta\in\Theta} \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \log f_\theta(x_i) = \arg\max_{\theta\in\Theta} \int_{\mathcal{X}} \log f_\theta(x)\, dF_0(x).$$

It is straightforward to see that θ_0 minimizes the KL divergence from the true data-

generating mechanism to a density in $\mathcal{F}_\Theta$

$$\theta_0 = \arg\max_{\theta\in\Theta} \int_{\mathcal{X}} \log f_\theta(x) dF_0(x) = \arg\min_{\theta\in\Theta} \int_{\mathcal{X}} \log \frac{f_0(x)}{f_\theta(x)} dF_0(x). \tag{3.1}$$

This is true regardless of whether F_0 is in the model space of $\mathcal{F}_\Theta$ and is well-motivated as the target of statistical model fitting (Akaike, 1981; Burnham and Anderson, 2002; Walker, 2013; Bissiri et al., 2016).

Uncertainty in this unknown value θ_0 flows directly from uncertainty in F_0. Of course F_0 is unknown, but being "Bayesian" we can place a prior on it, $\pi(F)$, for $F \in \mathcal{F}$, that should reflect our honest uncertainty about F_0. Typically the prior should have broad support unless we have special knowledge to hand, which is a problem with a parametric modelling approach that only supports a family of distribution functions indexed by a finite-dimensional parameter. The Bayesian NP literature, however, provides a range of priors for this sole purpose (Hjort et al., 2010). Once a prior for F is chosen, the correct way to propagate uncertainty about θ comes naturally from the posterior distribution for the law $\mathcal{L}[\theta(F)|x_{1:n}]$, via $\mathcal{L}[F|x_{1:n}]$, where $\theta(F) = \arg\max_{\theta\in\Theta} \int \log f_\theta(x) dF(x)$. The posterior for the parameter is then captured in the marginal by treating F as a latent auxiliary probability measure,

$$\tilde{\pi}(\theta \mid x_{1:n}) = \int_{\mathcal{F}} \pi(\theta, dF \mid x_{1:n}) = \int_{\mathcal{F}} \pi(\theta \mid F)\pi(dF \mid x_{1:n}), \tag{3.2}$$

where $\pi(\theta|F)$ assigns probability 1 to $\theta = \theta(F)$. We use $\tilde{\pi}$ to denote the NP update to distinguish it from the conventional Bayesian posterior $\pi(\theta|x_{1:n}) \propto \pi(\theta) \prod_{i=1}^n f_\theta(x_i)$, noting that in general the nonparametric posterior $\tilde{\pi}(\theta \mid x_{1:n})$ will be different to the standard Bayesian update as they are conditioning on different states of prior knowledge. In particular, as stated above, $\pi(\theta|x_{1:n})$ assumes artificially that $F_0 \in \mathcal{F}_\Theta$.

3.4.2 An NP prior using a MDP

For our purposes, the mixture of Dirichlet processes (MDP) (Antoniak, 1974) is a convenient vehicle for specifying prior beliefs $\pi(F)$ centered on parametric models.[1] The MDP prior can be written as

$$[F \mid \theta] \sim \text{DP}(c, f_\theta(\cdot)); \qquad \theta \sim \pi(\theta). \tag{3.3}$$

This is a mixture of standard Dirichlet processes with mixing distribution or hyper-prior $\pi(\theta)$, and concentration parameter c. We write this as $F \sim \text{MDP}(\pi(\theta), c, f_\theta(\cdot))$.

The MDP provides a practical, simple posterior update. From the conjugacy property of the DP applied to (3.3), we have the conditional posterior update given data $x_{1:n}$, as

$$[F \mid \theta, x_{1:n}] \sim \text{DP}\left(c+n,\ \frac{c}{c+n} f_\theta(\cdot) + \frac{1}{c+n}\sum_{i=1}^{n} \delta_{x_i}(\cdot)\right) \tag{3.4}$$

where δ_x denotes the Dirac measure at x. The concentration parameter c is an effective sample size, governing the trust we have in $f_\theta(x)$. The marginal posterior distribution for $\mathcal{L}[F|x_{1:n}]$ can be written as

$$\pi(dF \mid x_{1:n}) = \int_\Theta \pi(dF \mid \theta, x_{1:n})\, \pi(\theta \mid x_{1:n})\, d\theta, \tag{3.5}$$

i.e.

$$[F \mid x_{1:n}] \sim \text{MDP}\left(\pi(\theta \mid x_{1:n}),\ c+n,\ \frac{c}{c+n} f_\theta(\cdot) + \frac{1}{c+n}\sum_{i=1}^{n} \delta_{x_i}(\cdot)\right). \tag{3.6}$$

The mixing distribution $\pi(\theta|x_{1:n})$ coincides with the parametric Bayesian posterior, $\pi(\theta|x_{1:n})$, assuming there are no ties in the data (Antoniak, 1974), although as noted above it does not follow that the NP marginal $\tilde{\pi}(\theta|x_{1:n})$ is equivalent to the parametric Bayesian posterior $\pi(\theta|x_{1:n})$.

[1] The MDP should not to be confused with the Dirichlet process mixture (DPM) model (Lo, 1984).

We can see from the form of the conditional MDP (3.4) that the sampling distribution of the centering model, $f_\theta(x)$, regularizes the influence of the empirical data $\sum_{i=1}^n \delta_{x_i}(\cdot)$. The resulting NP posterior (3.5) combines the information from the posterior distribution of the centering model $\pi(\theta|x_{1:n})$ with the information in the empirical distribution of the data. This leads to a simple and highly parallelizable Monte Carlo sampling scheme as shown below.

3.4.3 Monte Carlo conditional maximization

The marginal in (3.2) facilitates a Monte Carlo estimator for functionals of interest under the posterior, which we write as $G = \int_\Theta g(\theta)\tilde{\pi}(\theta|x_{1:n})d\theta$. This is achieved by sampling $\pi(\theta, dF|x_{1:n})$ jointly from the posterior,

$$\begin{aligned} \int_\Theta g(\theta)\tilde{\pi}(\theta \mid x_{1:n})d\theta &\approx \frac{1}{B}\sum_{i=1}^{B} g(\theta^{(i)}) \\ \theta^{(i)} = \theta(F^{(i)}) &= \arg\max_{\theta\in\Theta} \int_{\mathcal{X}} \log f_\theta(x) dF^{(i)}(x) \qquad (3.7) \\ F^{(i)} &\sim \pi(dF \mid x_{1:n}). \qquad (3.8) \end{aligned}$$

This involves an independent Monte Carlo draw (3.8) from the MDP marginal followed by a conditional maximization of an objective function (3.7) to obtain each $\theta^{(i)}$. This Monte Carlo conditional maximization (MCCM) sampler is highly amenable to fast implementation on distributed computer architectures; given the parametric posterior samples, each NP posterior sample, $F^{(i)}$, can be computed independently and in parallel from (3.8).

We can see from (3.6) that the parametric posterior samples are not required if $c = 0$. If $c > 0$ it may be computationally intensive to generate samples from the parametric posterior. However, as we will see next, we do not need to sample from this posterior directly. This makes the approach particularly attractive to fast, tractable approximations for $\pi(\theta|x_{1:n})$, such as a variational Bayes (VB) posterior approximation. The NP update corrects for the approximation in a computationally efficient manner, leading to a posterior distribution with

optimal properties as shown below.

3.4.4 A more general construction

So far we have assumed, hypothetically, that:

(i) the modeller is interested in learning about the MLE under an infinite sample size, $\theta_0 = \arg\max_\theta \int \log f_\theta(x) dF_0(x)$, rather than $\alpha_0 = \arg\max_\alpha \int u(x,\alpha) dF_0(x)$ more generally, for a utility function $u(x,\alpha)$.

(ii) the parametric mixing distribution $\pi(\theta|x_{1:n})$ of the MDP posterior in (3.6) is constructed from the same centering model that defines the target parameter, $\theta_0 = \arg\max_\theta \int \log f_\theta(x) dF_0(x)$.

Both of these assumptions can be relaxed. For the latter case, it is valid to use a tractable parametric mixing distribution $\pi(\gamma|x_{1:n})$ and baseline model f_γ, while still learning about θ_0 in (3.1) through the marginal $\tilde{\pi}(\theta|x_{1:n})$ as in (3.2) obtained via $\theta(F)$ and

$$[F \mid x_{1:n}] \sim \text{MDP}\left(\pi(\gamma \mid x_{1:n}),\ c+n,\ \frac{c}{c+n} f_\gamma(\cdot) + \frac{1}{c+n}\sum_i \delta_{x_i}\right). \tag{3.9}$$

For (i), we can use the mapping $\alpha(F) = \arg\max_\alpha \int u(x,\alpha) dF(x)$ to derive the NPL posterior on actions or parameters maximizing some expected utility under a model-centered MDP posterior. This can be written as $\tilde{\pi}(\alpha|x_{1:n}) = \int \pi(\alpha|F)\pi(dF|x_{1:n})$, where $\pi(\alpha|F)$ assigns probability 1 to $\alpha = \alpha(F)$.

This highlights a major theme of the paper: the idea of obtaining posterior samples via maximization of a suitably randomized objective function. In generality the target is $\alpha_0 = \arg\max_\alpha \int u(x,\alpha) dF_0(x)$, obtained by maximizing an objective function, and the randomization arises from the uncertainty in F_0 through $\pi(F|x_{1:n})$ that takes into account the information, and any misspecification, associated with a parametric centering model.

Algorithm 4. The Posterior Bootstrap

Data: Dataset $x_{1:n} = (x_1, \ldots, x_n)$.
Parameter of interest $\alpha_0 = \alpha(F_0) = \arg\max_\alpha \int u(x, \alpha) dF_0(x)$.
Mixing posterior $\pi(\gamma|x_{1:n})$, concentration parameter c, centering model $f_\gamma(x)$.
Number of centering model samples T.
begin
  for $i = 1, \ldots, B$ **do**
    Draw centering model parameter $\gamma^{(i)} \sim \pi(\gamma|x_{1:n})$;
    Draw posterior pseudo-sample $x^{(i)}_{(n+1):(n+T)} \overset{iid}{\sim} f_{\gamma^{(i)}}$;
    Generate weights
    $(w_1^{(i)}, \ldots, w_n^{(i)}, w_{n+1}^{(i)}, \ldots, w_{n+T}^{(i)}) \sim \text{Dirichlet}(1, \ldots, 1, c/T, \ldots, c/T)$;
    Compute parameter update

$$\tilde{\alpha}^{(i)} = \arg\max_\alpha \left\{ \sum_{j=1}^{n} w_j^{(i)} u(x_j, \alpha) + \sum_{j=1}^{T} w_{n+j}^{(i)} u(x_{n+j}^{(i)}, \alpha) \right\};$$

  Return NP posterior sample $\{\tilde{\alpha}^{(i)}\}_{i=1}^{B}$.

3.4.5 The Posterior Bootstrap algorithm

We will use the general construction of Section 3.4.4 to describe a sampling algorithm. We assume we have access to samples from the posterior parametric mixing distribution, $\pi(\gamma|x_{1:n})$, in the MDP. In the case of model misspecification, [S1], if the data contains no ties, this is simply the parametric Bayesian posterior under $\{f_\gamma(x), \pi(\gamma)\}$, for which there is a large literature of computational methods available for sampling - see for example Robert and Casella (2005). If there are ties then we refer the reader to Antoniak (1974) or note that we can simply break ties by adding a new pseudo-variable, such as $x^* \sim N(0, \epsilon^2)$ for small ϵ.

The sampling algorithm, found in Algorithm 4, is a mixture of Bayesian posterior bootstraps. After a sample $\gamma^{(i)}$ is drawn from the mixing posterior, $\pi(\gamma|x_{1:n})$, a posterior pseudo-sample is generated, $x^{(i)}_{(n+1):(n+T)} \overset{iid}{\sim} f_{\gamma^{(i)}}(x)$, and added to the dataset, which is then randomly weighted. The parameter under this implicit distribution function is then computed as the solution of an optimization problem.

Note for the special case of correcting model misspecification [S1], we have $\gamma \equiv \theta$, $f_\gamma(\cdot) \equiv f_\theta(\cdot)$, $\pi(\gamma|x_{1:n}) \equiv \pi(\theta|x_{1:n})$, $\alpha \equiv \theta$, $u(x, \alpha) \equiv \log f_\theta(x)$, so that the posterior

sample is given by

$$\tilde{\theta}^{(i)} = \arg\max_{\theta} \left\{ \sum_{j=1}^{n} w_j^{(i)} \log f_\theta(x_j) + \sum_{j=1}^{T} w_{n+j}^{(i)} \log f_\theta(x_{n+j}^{(i)}) \right\}.$$

where $w^{(i)} \sim \text{Dirichlet}(\cdot)$ following Algorithm 4 and $x^{(i)}_{(n+1):(n+T)}$ are T synthetic observations drawn from the parametric sampling distribution under $\theta^{(i)}$ which itself is drawn from $\pi(\theta|x_{1:n})$. We leave the concentration parameter c to be set subjectively by the practitioner, representing faith in the parametric model. Some further guidance to the setting of c can be found in Section 3.7.1 of the Supplementary Material.

3.4.6 Adaptive Nonparametric Learning: aNPL

Instead of the Dirichlet distribution approximation to the Dirichlet process, we propose an alternative stick-breaking procedure that has some desirable properties. This procedure entails following the usual DP stick-breaking construction (Sethuraman, 1994) for the model component of the MDP posterior, by repeatedly drawing Beta$(1, c)$-distributed stick breaks, but terminating when the unaccounted for probability measure $\prod_j (1 - v_j)$, multiplied by the average mass assigned to the model, $c/(c + n)$, drops below some threshold ϵ set by the user. This adaptive nonparametric learning (aNPL) algorithm is written out in full in Section 3.7.2 of the Supplementary Material.

One advantage of this approach is that a number of theoretical results then hold, as for large enough n, under this adaptive scheme the parametric model is in effect 'switched off', and essentially the MDP with $c = 0$ is used to generate posterior samples. This is an interesting notion in itself. For small samples, we prefer the regularization that our model provides, though as n grows the average probability mass assigned to the model decays like $(c + n)^{-1}$, as seen in (3.4). In the adaptive version, we agree a hard threshold at which point we discard the model entirely and allow the data to speak for itself. We set this point at a level such that we are a priori comfortable that there is enough information in the sample

alone with which to quantify uncertainty in our parameter of interest. For example, $\epsilon = 10^{-4}$ and $c = 1$ only utilizes the centering model for $n < 10,000$. Further, we could use this idea to set c: this quantity is determined if a tolerance level, ϵ, and a threshold $n_{\max}$ over which the parametric model would be discarded, are provided by the practitioner.

3.4.7 Properties of NPL

Bayesian nonparametric learning has a number of important properties that we shall now describe.

Honesty about correctness of model. Uncertainty in the data-generating mechanism is quantified via a NP update that takes into account the model likelihood, prior, and concentration parameter c. Uncertainty about model parameters flows from uncertainty in the data-generating mechanism.

Incorporation of prior information. The prior for θ is naturally incorporated as a mixing distribution for the MDP. This is in contrast to a number of Bayesian methods with similar computational properties but that do not admit a prior (Newton and Raftery, 1994; Chamberlain and Imbens, 2003).

Parallelized bootstrap computation. As shown in Section 3.4.5, NPL is trivially parallelizable through a Bayesian posterior bootstrap and can be coupled with misspecified models or approximate posteriors to deliver highly scalable and exact inference.

Consistency. Under mild regularity, all posterior mass concentrates in any neighbourhood of θ_0 as defined in (3.1), as the number of observations tends to infinity. This follows from an analogous property of the DP (see, for example Hjort et al. (2010)).

Standard Bayesian inference is recovered as $c \to \infty$. This follows from the property of the DP that it converges to the prior degenerate at the base probability distribution in the limit of $c \to \infty$.

Non-informative learning with $c = 0$. If no model or prior is available, setting $c = 0$ recovers the WLB. This has an exact interpretation as an objective NP posterior (Lyddon

et al., 2019), where the asymptotic properties of the misspecified WLB were studied. Müller (2013) demonstrated the suboptimality of a misspecified Bayesian posterior, asymptotically, relative to an asymptotic normal distribution with the same centering but a sandwich covariance matrix (Huber, 1967). We will see next that for large samples the misspecified Bayesian posterior distribution is predictively suboptimal as well.

A superior asymptotic uncertainty quantification to Bayesian updating. A natural way to compare posterior distributions is by measuring their predictive risk, defined as the expected KL divergence of the posterior predictive to F_0. We consider only the situation where there is an absence of strong prior information, following Shimodaira (2000) and Fushiki (2005).

We say predictive π_1 asymptotically dominates π_2 up to $o(n^{-k})$ if for all distributions q there exists a non-negative and possibly positive real-valued functional $K(q(\cdot))$ such that:

$$\mathbb{E}_{x_{1:n}\sim q}\, d_{\mathrm{KL}}(q(\cdot), \pi_2(\cdot \mid x_{1:n})) - \mathbb{E}_{x_{1:n}\sim q}\, d_{\mathrm{KL}}(q(\cdot), \pi_1(\cdot \mid x_{1:n})) \;=\; K(q(\cdot)) + o(n^{-k}).$$

We have the following theorem about the asymptotic properties of the MDP with $c = 0$. This result holds for aNPL, as the model component is ignored for suitably large n.

Theorem 3. *The posterior predictive of the MDP with $c = 0$ asymptotically dominates the standard Bayesian posterior predictive up to $o(n^{-1})$.*

Proof. In Fushiki (2005) the bootstrap predictive is shown to asymptotically dominate the standard Bayesian predictive up to $o(n^{-1})$. In Theorem 1 of Fushiki (2010), the predictive of the MDP with $c = 0$ and the bootstrap predictive are shown to be equal up to $o_p(n^{-3/2})$. A Taylor expansion argument shows that the predictive risk of the MDP with $c = 0$ has the same asymptotic expansion up to $o(n^{-1})$ as that of the bootstrap. Thus Theorem 2 of Fushiki (2005) can be proven with the predictive of the MDP with $c = 0$ in place of the bootstrap predictive. Thus the predictive of the MDP with $c = 0$ must also dominate the standard Bayesian predictive up to $o(n^{-1})$. □

3.5 Illustrations

3.5.1 Exponential family, [S1]

Suppose the centering model is an exponential family with parameter θ and sufficient statistic $s(x)$,

$$\mathcal{F}_\Theta = \left\{ f_\theta(x) = g(x) \exp\left\{\theta^T s(x) - K(\theta)\right\};\ \theta \in \Theta \right\}.$$

Under assumed regularity, by differentiating under the integral sign of (3.1) we find that our parameter of interest must satisfy $\mathbb{E}_{F_0} s(x) = \nabla_\theta K(\theta_0)$. For a particular F drawn from the posterior bootstrap, the expected sufficient statistic is

$$\nabla_\theta K(\tilde{\theta}) = \lim_{T\to\infty} \left\{ \sum_{j=1}^{n} w_j s(x_j) + \sum_{j=n+1}^{n+T} w_j s(x_j) \right\}.$$

with $\tilde{\theta}$ the NP posterior parameter value, weights $w_{1:(n+T)}$ arising from the Dirichlet distribution as set out in Algorithm 4, and $x_j \sim f_\theta(\cdot)$ for $j = n+1, \ldots, n+T$, with θ drawn from the parametric posterior. This provides a simple geometric interpretation of our method, as convex combinations of (randomly-weighted) empirical sufficient statistics and model sufficient statistics from the parametric posterior. The distribution of the random weights is governed by c and n only. Our method generates stochastic maps from misspecified posterior samples to corrected NP posterior samples, by incorporating information in the data over and above that captured by the model.

3.5.2 Updating approximate posteriors [S2]: Variational Bayes uncertainty correction

Variational approximations to Bayesian posteriors are a popular tool for obtaining fast, scalable but approximate Bayesian posterior distributions (Bishop, 2006; Blei et al., 2017). The approximate nature of the variational update can be accounted for using our approach.

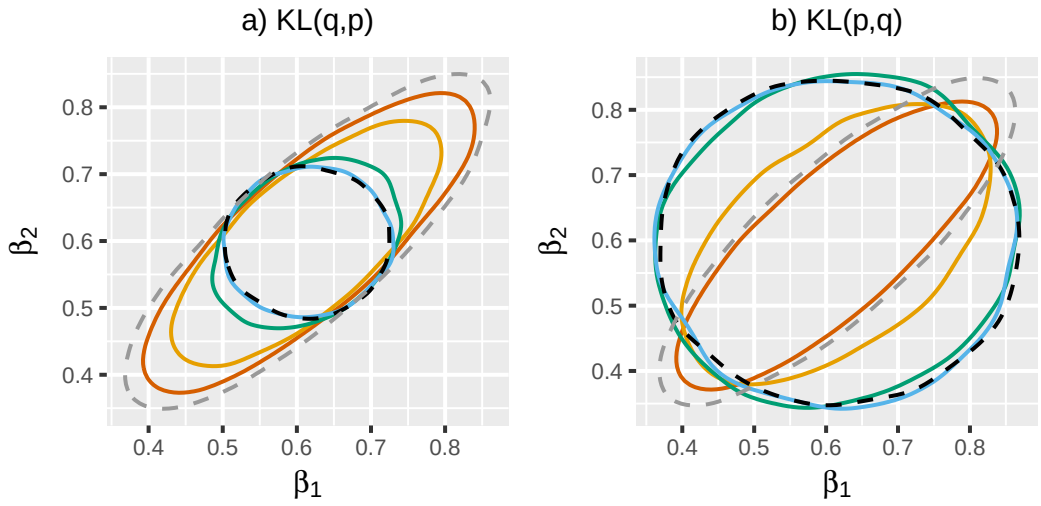


Figure 3.1: Posterior 95% probability contour for a bivariate Gaussian, comparing VB-NPL with $c \in \{1, 10^2, 10^3, 10^4\}$ (red, orange, green, blue respectively) to the known Bayes posterior (grey dashed line) and a VB approximation (black dashed line).

Figure 3.1 shows a mean-field normal approximation q to a correlated normal posterior p, an example similar to one from Bishop (2006), Section 10.1. We generated 100 observations from a bivariate normal distribution, centered at $(\frac{1}{2}, \frac{1}{2})$, with dimension-wise variances both equal to 1 and correlation equal to 0.9, and independent normal priors on each dimension, both centered at 0 with variance 10^2. Each posterior contour plotted is based on $10,000$ posterior samples.

By applying the posterior bootstrap with a VB posterior (VB-NPL) in place of the Bayesian posterior, we recover the correct covariance structure for decreasing prior concentration c. If instead of $d_{\mathrm{KL}}(q, p)$ we use $d_{\mathrm{KL}}(p, q)$ as the objective, as it is for expectation propagation, the model posterior uncertainty may be overestimated, but is still corrected by the posterior bootstrap.

We demonstrate this in practice through a VB logistic regression model fit to the Statlog German Credit dataset, containing 1000 observations and 25 covariates (including intercept), from the UCI ML repository (Dheeru and Karra Taniskidou, 2017), preprocessing via Fernández-Delgado et al. (2014). An independent normal prior with variance 100 was assigned to each covariate, and 1000 posterior samples were generated for each method.

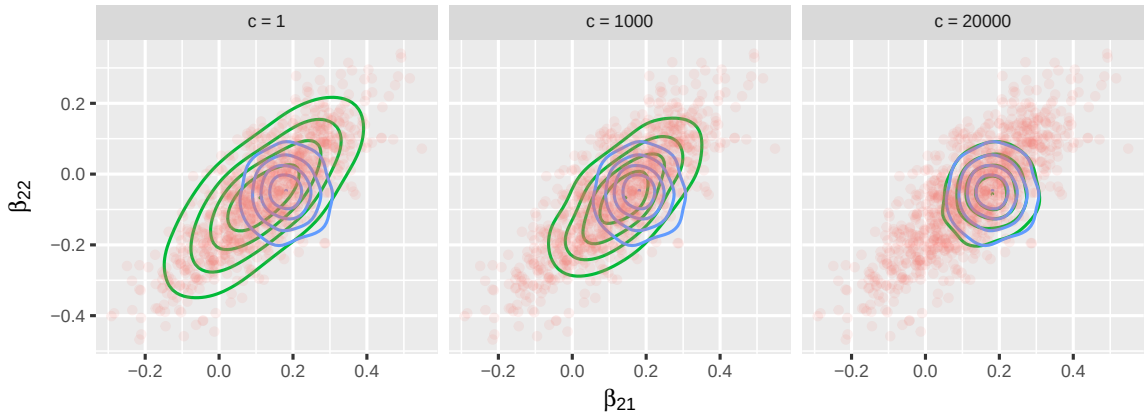


Figure 3.2: Posterior contour plot for β_{22} vs β_{21}, for VB-NPL (green) and VB (blue), for $c \in \{1, 1000, 20000\}$. Scatter plot is a Bayesian logistic posterior sample (red) using a Polya-Gamma scheme.

We obtain a mean-field VB sample using automatic differentiation variational inference (ADVI) in Stan (Kucukelbir et al., 2015). When generating synthetic samples for the posterior bootstrap, both features and classes are needed. Classes are generated, given features, according to the probability specified by the logistic distribution. In this example (and the example in Section 3.5.3) we repeatedly re-use the features of the dataset for our pseudo-samples. In Fig. 3.2 we show that the NP update effectively corrects the VB approximation for small values of c. Of course, local variational methods can provide more accurate posterior approximations to Bayesian logistic posteriors (Jaakkola and Jordan, 1997), though these too are approximations, that NP updating can correct.

Comparison with Bayesian logistic regression. The conventional Bayesian logistic regression assumes the true log-odds of each class is linear in the predictors, and would use MCMC for inference (Polson et al., 2013). The MCMC samples, shown as points in Fig. 3.2, are a good match to the NPL update but MCMC requires a user-defined burn-in and convergence checking. The runtime to generate 1 million samples by MCMC (discarding an equivalent burn-in), was 33 minutes, compared to 21 seconds with NPL, using an m5.24xlarge AWS instance with 96 vCPUs; a speed-up of 95 times. Additionally, NPL has provably better predictive properties, as detailed in Section 3.4.7.

3.5.3 Directly updating the prior: Bayesian Random Forests, using synthetic generated data

Random forests (RF) (Breiman, 2001a) is an ensemble learning method that is widely used and has demonstrated excellent general performance (Fernández-Delgado et al., 2014). We construct a Bayesian RF (BRF), via NPL with decision trees, under a prior mixing distribution (a variant of [S1]). This enables the incorporation of prior information, via synthetic data generated from a prior prediction function, in a principled way that is not available to RF. The step-like generative likelihood function arising from the tree partition structure does not reflect our beliefs about the true sampling distribution; the trees are just a convenient compression of the data. Because of this we simply update the prior in the MDP by specifying $\pi(\gamma|x_{1:n}) = \pi(\gamma)$. Details of our implementation of BRF can be found in Section 3.7.3 of the Supplementary Material.

To demonstrate the ability of BRF to incorporate prior information, we conducted the following experiment. For 13 binary classification datasets from the UCI ML repository (Dheeru and Karra Taniskidou, 2017), we constructed a prior, training and test dataset of equal size. We measured test dataset predictive accuracy for three methods (relative to an RF trained on the training dataset only): BRF (c=0) (a non-informative BRF with $c = 0$, trained on the training dataset only), BRF (c>0) (a BRF trained on the training dataset, incorporating prior pseudo-samples from a non-informative BRF trained on the prior dataset, setting c equal to the number of observations in the prior dataset), and RF-all (an RF trained on the combined training and prior datasets). See Fig. 3.3 for boxplots of the test accuracy over 100 repetitions.

As detailed in Section 3 of the Supplementary Material, for small c we find that BRF and RF have similar performance, but as c increases, more weight is given to the externally-trained component and we find that BRF outperforms RF. The best performance of our BRF tends to occur when c is set equal to the number of samples in the external training dataset, in line with our intuition of the role of c as an effective sample size. Overall, the

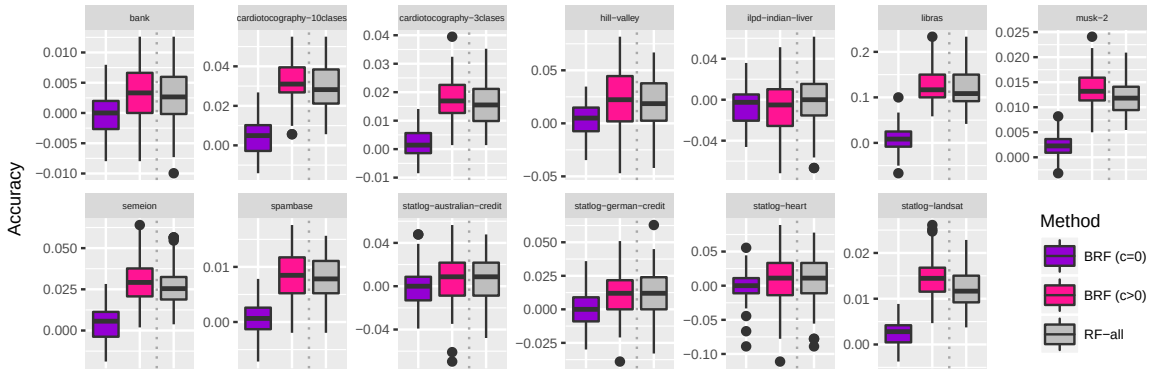


Figure 3.3: Box plot of classification accuracy minus that of RF, for 13 UCI datasets.

BRF accuracy is better than that of RF, and close to that of RF-all. BRF may have privacy benefits over RF-all as it only requires synthetic samples; both the original data and the model can be kept private.

3.5.4 Direct updating of utility functions [S3]: population median

We demonstrate direct inference for a functional of interest using the population median, under a misspecified univariate Gaussian model, as an example, where the parameter of interest is $\alpha_0 = \arg\min_\alpha \int |\alpha - x| dF_0(x)$, and an MDP prior centered at a $\mathcal{N}(\theta, 1)$ with prior $\pi(\theta) = \mathcal{N}(0, 10^2)$ and data generated from a skew-normal distribution. We use the posterior bootstrap to generate posterior samples that combine the prior model information with that from the data. Figure 3.4 presents histograms of posterior medians given a sample of 20 observations from a skew-normal distribution with mean 0, variance 1 and median approximately -0.2. The left-most histogram is sharply peaked at the sample median but does not have support outside of $(x_{\min}, x_{\max})$. As c grows smoothness from the parametric model is introduced to a point where the normal location parameter is used.

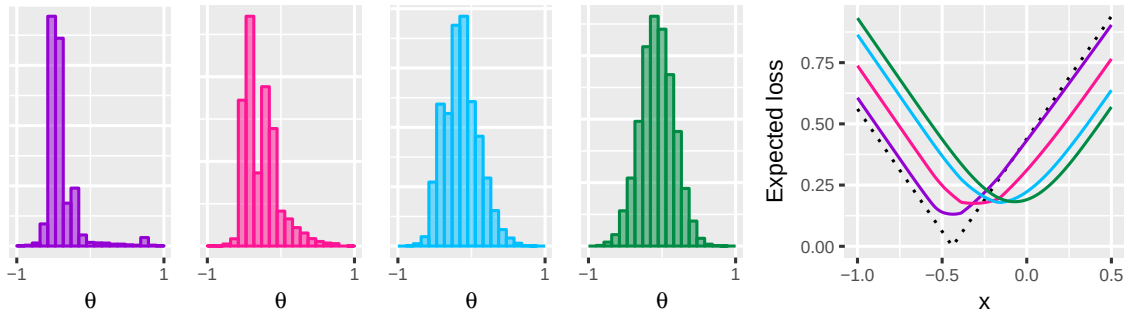


Figure 3.4: Posterior histogram for median, with $c = 0, 20, 80, 1000$ (left to right). Rightmost: posterior expected loss as a function of observation x. Dotted line shows the loss at sample median.

3.6 Discussion

We have introduced a new approach for scalable Bayesian nonparametric learning, NPL, for parametric models that facilitates prior regularization via a baseline model, and corrects for model misspecification by incorporating an empirical component that has greater influence as the number of observations grows. A concentration parameter c encodes subjective beliefs on the validity of the model; $c = \infty$ recovers Bayesian updating under the baseline model, and $c = 0$ ignores the model entirely. Under regularity conditions, asymptotically, our method closely matches parametric Bayesian updating if the posited model were indeed true, and will provide an asymptotically superior predictive if the model is misspecified. The NP posterior predictive mixes over the parametric model space as opposed to targeting F_0 directly, though this may aid interpretation compared to fully nonparametric approaches. Additionally, our construction admits a trivially parallelizable sampler once the parametric posterior samples have been generated (or if $c = 0$).

Our approach can be seen to blur the boundaries between Bayesian and frequentist inference. Conventionally, the Bayesian approach conditions on data and treats the unknown parameter of interest as if it was a random variable with some prior on a known model class. The frequentist approach treats the object of inference as a fixed but unknown constant and characterizes uncertainty through the finite sample variability of an estimator targeting this

value. Here we randomize an objective function (an estimator) according to a Bayesian nonparametric prior on the sampling distribution, leading to a quantification of subjective beliefs on the value that would be returned by the estimator under an infinite sample size.

At the heart of our approach is the notion of Bayesian updating via randomized objective functions through the posterior bootstrap. The posterior bootstrap acts on an augmented dataset, comprised of data and posterior pseudo-samples, under which randomized maximum likelihood estimators provide a well-motivated quantification of uncertainty while assuming little about the data-generating mechanism. The precursor to this is the weighted likelihood bootstrap, which utilized a simpler form of randomization that ignored prior information. Our approach provides scope for quantifying uncertainty for more general machine learning models by randomizing their objective functions suitably.

Acknowledgements

SL is funded by the EPSRC OxWaSP CDT, through EP/L016710/1. CH gratefully acknowledges support for this research from the MRC, The Alan Turing Institute, and the Li Ka Shing foundation.

3.7 Supplementary Material

3.7.1 Setting the MDP concentration parameter

The MDP concentration parameter c represents the subjective faith we have in the model: setting $c = 0$ means we discard the baseline model completely; whereas for $c = \infty$ we obtain the Bayesian posterior, equivalent to knowing that the model is true.

One way that the concentration parameter may be specified in practice is via the prior uncertainty of a functional. For example, we could set c using a priori variance of the population mean. Using standard properties of the Dirichlet process, it is straightforward to

show that under an MDP$(\pi(\theta), c, f_\theta(\cdot))$, the variance in the mean functional is given by

$$\text{var}\,\mu(\theta) \;+\; \frac{1}{1+c}\,\mathbb{E}\,\sigma^2(\theta)$$

where $\mu(\theta)$ is the mean of $x|\theta$ under the model, and $\sigma^2(\theta)$ is the variance of $x|\theta$, and the variance and expectation in the expression above are both with respect to the prior $\theta \sim \pi$. Thus, if we can elicit a prior variance over the population mean, this will lead directly to a specific setting for c.

Another option is the Bayesian learning of c via a hyper-prior; see Ghosal and Vaart (2017), Section 4.5 for details. In practice it may be preferable to consider a number of different values of c; once the parametric posterior samples are generated the posterior for each value of c could be computed in parallel.

3.7.2 Posterior Bootstrap for adaptive Nonparametric Learning

We present the posterior bootstrap for adaptive Nonparametric Learning (aNPL), as discussed in Section 2.6 of the paper. This algorithm uses the stick-breaking construction of the Dirichlet process, as opposed to the Dirichlet distribution approximation of Section 2.5. The stick-breaking threshold, at which point the stick-breaking process ceases, is a function of the number of observations n directly, meaning that for large enough n the parametric model will be ignored, switching off the pseudo-samples from the parametric posterior predictive.

Algorithm 5. The Posterior Bootstrap for adaptive Nonparametric Learning

Data: Dataset $x_{1:n} = (x_1, \ldots, x_n)$.
Parameter of interest $\alpha_0 = \alpha(F_0) = \arg\max_\alpha \int u(x, \alpha) dF_0(x)$.
Mixing posterior $\pi(\gamma|x_{1:n})$, concentration parameter c, centering model $f_\gamma(x)$.
aNPL stick-breaking tolerance ϵ.
begin
  for $i = 1, \ldots, B$ **do**
    Draw centering model parameter $\gamma^{(i)} \sim \pi(\gamma|x_{1:n})$;
    Draw model vs data weight: $s^{(i)} \sim \text{Beta}(c, n)$;
    Set data weights $w^{(i)}_{1:n} = (1 - s^{(i)})v^{(i)}_{1:n}$, with $v^{(i)}_{1:n} \sim \text{Dirichlet}(1, \ldots, 1)$;
    Set $v_{\text{rem}} = c/(c + n)$, $T = 0$;
    while $v_{rem} \geq \epsilon$ **do**
      $T \to T + 1$;
      Draw posterior pseudo-sample $x^{(i)}_{n+T} \sim f_{\gamma^{(i)}}$;
      Stick-break $w^{(i)}_{n+T} = s^{(i)} \prod_{j=1}^{T-1}(1 - v^{(i)}_{n+j})v^{(i)}_{n+T}$ with
       $v^{(i)}_{n+T} \sim \text{Beta}(1, c)$;
      Update $v_{\text{rem}} = c/(c + n) \prod_{j=1}^{T}(1 - v^{(i)}_{n+j})$;
    Compute parameter update

$$\tilde{\alpha}^{(i)} = \arg\max_\alpha \left\{ \sum_{j=1}^{n} w^{(i)}_j u(x_j, \alpha) + \sum_{j=1}^{T} w^{(i)}_{n+j} u(x^{(i)}_{n+j}, \alpha) \right\};$$

  Return NP posterior sample $\{\tilde{\alpha}^{(i)}\}_{i=1}^{B}$.

3.7.3 Bayesian Random Forests - further details

We base our Bayesian Random Forests (BRF) implementation on the Scikit-learn `RandomForestClassifier` class (Pedregosa et al., 2011), with an alternative fitting method. Instead of applying bootstrap aggregation as Random Forests (RF) does, for each tree we construct an augmented dataset, containing the training data and prior pseudo-data. We then fit the decision tree to the augmented dataset, suitably weighted.

If the concentration parameter c is set to zero then no prior data is necessary; we just need to generate sample weights $w_{1:n} \sim \text{Dirichlet}(1, \ldots, 1)$ for observations $x_{1:n}$ and fit a tree to the weighted training dataset. When $c > 0$ we use a BRF with $c = 0$, trained on the prior data, to generate the prior pseudo-data. For each tree in the BRF we generate

a pseudo-dataset before training a decision tree on the augmented dataset containing the prior pseudo-dataset and the training dataset. If T pseudo-observations are generated, we generate a vector of Dirichlet$(1, \ldots, 1, c/T, \ldots, c/T)$ weights, relating to observations $x_{1:(n+T)}$, where $x_{1:n}$ is the training dataset and $x_{(n+1):(n+T)}$ is the prior pseudo-data.

When training RF decision trees, terminal node splits are made until the leaves only contain training data from a single class. This is computationally problematic for our BRF method, as after augmenting our internal training data with a large pseudo-dataset of samples, trees may need to grow ever deeper until leaf purity is attained. To avert this issue, we threshold the proportion of sample weight required to be present across samples at a node before a split can take place. This can be done via the `min_weight_fraction_leaf` argument of the Scikit-learn `RandomForestClassifier` class. In our testing we set a weight proportion threshold of $0.5(n + c)^{-1}$.

Internal, external and test datasets were obtained by equally sized class-stratified splits. Each forest contained 100 trees, and 10,000 external pseudo-samples were generated for the BRFs with $c > 0$. Predictions were made in our BRF by majority vote across the forest of trees.

A plot of the mean classification accuracy as a function of c is given for the Bank dataset in Figure 3.5. As discussed in Section 3.3, for small c the external data is given little weight and our method performs similarly to an RF trained only on the internal dataset. As c grows, the accuracy improves, peaking around where c is equal to the number of external samples. This peak performance is roughly the same as that attained by an RF trained on both the external and internal data.

Note that when setting $c = 0$ our method is equivalent to Bayesian bootstrapping random decision trees. Similarly to our construction Taddy et al. (2015) uses the Bayesian bootstrap as an underlying model of the data-generating mechanism, viewing the randomly weighted trees generated as a posterior sample over a functional of the unknown data-generating mechanism. Previously, a number of attempts have been made in the literature to construct

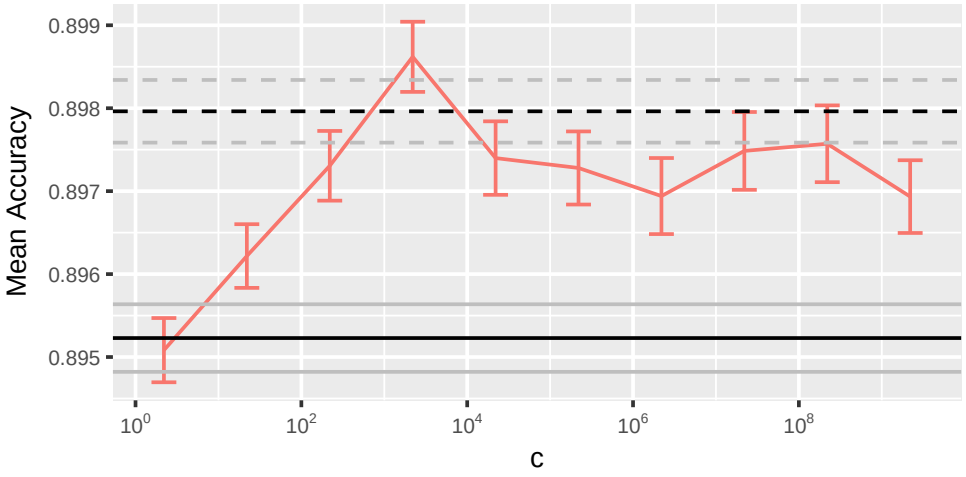


Figure 3.5: Mean classification accuracy vs c for RF (black solid), RF-all (black dashed), and BRF (red) for varying c. over 100 repetitions, for the Bank dataset. Error bars for BRF and the grey lines for RF and RF-all represent one standard error from the mean. The BRF with $c = 0$ has a mean accuracy of 0.895 and a standard error of 0.0004.

Bayesian models for decision trees (Chipman et al., 1998; Denison et al., 1998) but the associated MCMC sampling routines tend to mix poorly. Our method, whilst remaining honest that our trees are poor generative models, is very similar to RF in nature and performance. However, it has the additional benefit of enabling the user to incorporate prior information via a prediction function, in a principled manner.

3.8 Further discussion

We highlight some further methodological and practical considerations that were not included in the publication.

3.8.1 General form of nonparametric learning

The presentation of NPL in this paper is restricted to using an MDP prior for the unknown data-generating mechanism. The incorporation of Bayesian parametric model information is achieved very naturally for this choice of prior, as described in Section 3.4.2. It should be emphasised, however, that the idea of nonparametric learning from Bayesian models

is much more general and could apply to a range of different nonparametric priors. The exploration of NPL under Bayesian nonparametric priors other than the MDP would be interesting future work.

3.8.2 Graphical description of the exponential family case

Figure 3.6 provides a schematic description of the posterior bootstrap algorithm for the exponential family model, as described in Section 3.5.1. A posterior sample $\theta^{(b)}_{\text{model}}$ is drawn from the parametric model posterior, and mapped to its associated sufficient statistic. The observed data is mapped to its empirical sufficient statistic. These sufficient statistics are randomised according to their Dirichlet-distributed weights, and then a convex combination of these sufficient statistics is computed. The combining weight $v^{(b)}$ follows a Beta(n, c) distribution. This resultant randomised sufficient statistic is then mapped back to the parameter space to return one posterior bootstrap sample.

3.8.3 Nonparametric learning for small samples

Although we have some assurance from the asymptotics about the performance of NPL versus misspecified Bayes for large samples, we do not have any quantification of the quality of the NPL posterior, relative to, say, a well-specified Bayesian approach, for finite samples. Finite sample results are often a challenge for nonparametric procedures. We can see, directly, that for very small samples, NPL may exhibit some strange behaviour. As an example, suppose that $c = 0$, there is one observation x_1, and the parameter is population mean. The NP posterior is then a point mass at x_1, so there is no uncertainty! Also, note that in Figure 3.1, the Bayesian posterior under 100 observations is more dispersed than the NPL posterior with $c = 1$; this is also true for $c = 0$. This may seem counter-intuitive. As is not uncommonly the case, some care should be taken when applying nonparametric methods to very small samples.

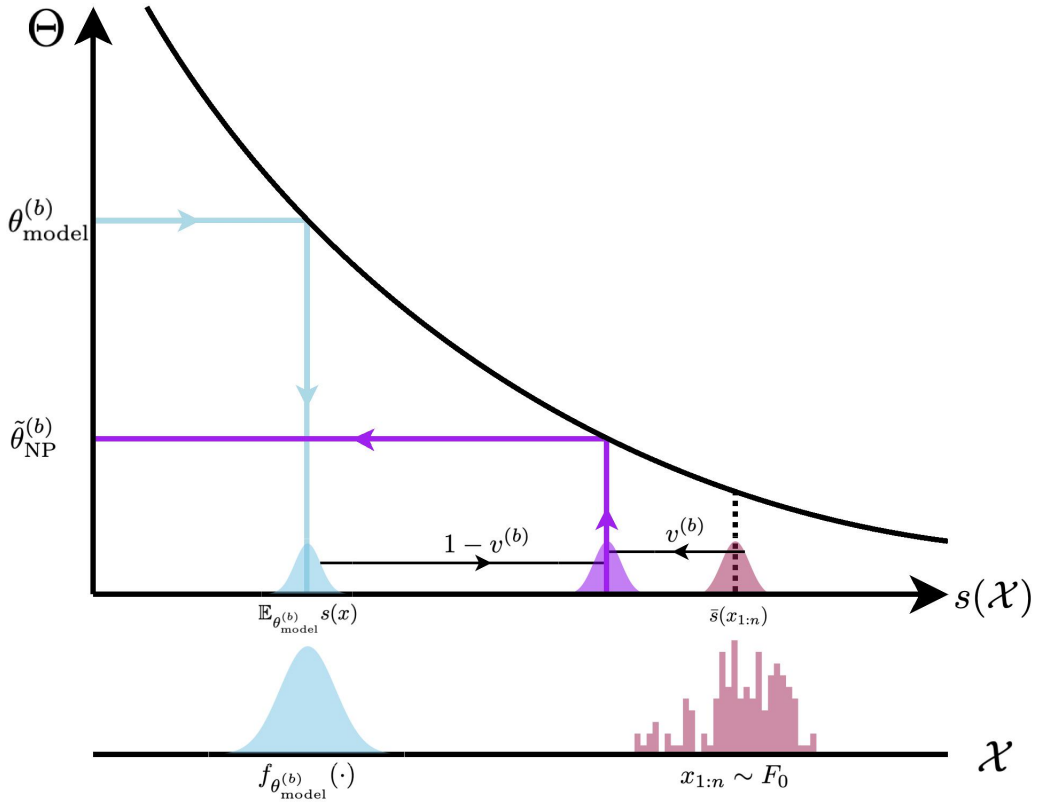


Figure 3.6: Schematic diagram of the NPL method for the Exponential family example of Section 3.5.1.

3.8.4 Recovering the prior under zero observations

NPL has the property that the Bayesian posterior is recovered if we have full faith in the model, i.e. $c = \infty$. It follows that, given zero data, we recover the prior if $c = \infty$. It might seem desirable that our method would recover the prior if there are no observations for all values of $c > 0$, but this is not the case for the MDP. When we don't have full faith in the Bayesian model, this means we also don't have full faith in the prior. So although our average NP prior sample, conditional on drawing θ from the prior for the parametric model, is equal to θ, the random weights mean there is additional variability about the NP prior $\tilde{\theta}(\theta)$. It would be interesting future work to consider if there is a way of recovering the prior for all values of c under an alternative Bayesian nonparametric formulation. In this case, c would represent faith in the statistical model alone, and not the prior. Practically, one way to recover the prior is to use equal deterministic weights for the prior pseudo-samples, as

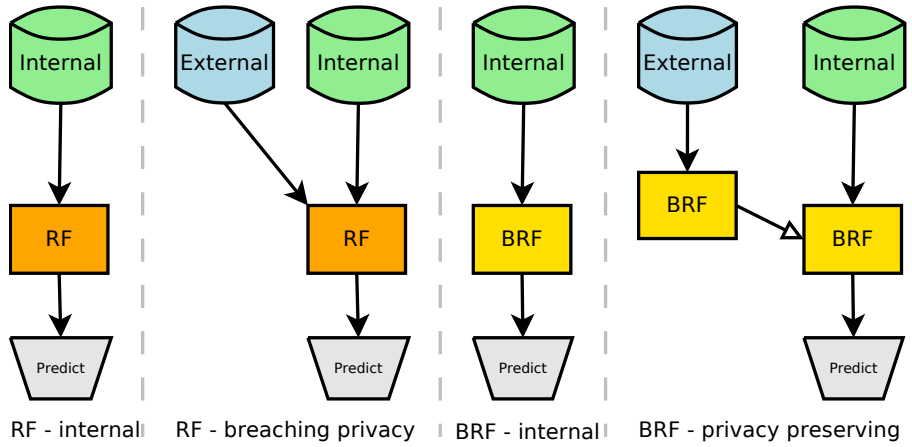


Figure 3.7: Flow diagrams for random forests and Bayesian random forests.

opposed to the random weights that are consistent with the MDP formulation.

3.8.5 NPL sensitivity analysis

The setting of the concentration-type parameters, such as c, is a common issue in applied work for a number of Bayesian nonparametric models. We discussed a variety of possible methods for setting c in Section 3.7.1. Additionally, as highlighted by an anonymous reviewer, NPL can be used to perform a misspecification sensitivity analysis on a parametric posterior. The plotting of posterior quantities of interest as a function of c gives the practitioner a measure of the variability of this quantity between model-based and nonparametric learning. Conveniently, the computation of a sensitivity curve over a grid of values of the concentration parameter could be computed in parallel.

3.8.6 Privacy-preserving learning by sharing pseudo-datasets

An interesting line of future work is the application of NPL to the problem of information sharing under privacy constraints. Suppose we are in a situation whereby Analyst 1 and Analyst 2, each with a separate dataset, wish to share information, but are bound by a privacy restriction preventing direct access of both datasets by either analyst. If Analyst 2 is willing to share privacy-preserving pseudo-data with Analyst 1, then nonparametric learning could

be used by Analyst 1 to incorporate information external to Analyst 1's dataset into their analysis, without breaching the privacy constraint. The external pseudo-data would take the place of the model-based samples of the posterior bootstrap, and c would represent the faith that Analyst 1 has in Analyst 2's pseudo-data.

The Bayesian random forests example in Section 3.5.3 can be viewed in this way. Figure 3.7 shows four flow diagrams for random forests and Bayesian random forests that demonstrate this. Going from left to right, the first flow diagram shows the privacy-preserving setting available to the RF algorithm: Analyst 1 just uses their internal dataset as training data, and cannot use any external data. The next diagram shows a RF algorithm using both the external and internal dataset, which provides better predictions, as it has more training data, but breaches the privacy constraint. The third diagram represents a BRF algorithm using one dataset; it was shown in Section 3.5.3 that this has similar performance to that of RF. The right-most diagram shows how pseudo-data, generated from the output of a BRF trained on the external data, could be incorporated into a subsequent BRF. It was shown in Section 3.5.3 that doing this leads to a performance improvement over BRF trained only on the internal dataset.

The further investigation and formalisation of NPL-based privacy-preserving procedures is left as future work.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

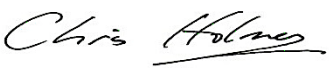
Title of Paper	Nonparametric learning from Bayesian models with randomized objective functions
Publication Status	Published.
Publication Details	Lyddon, S. P., Walker, S. G., & Holmes, C. C. (2018). Nonparametric learning from Bayesian models with randomized objective functions. In *Advances in Neural Information Processing Systems 31, 2018.*

Student Confirmation

Student Name:	Simon Lyddon		
Contribution to the Paper	Formulated the scientific idea. Undertook the literature review. Development of novel methodology. Developed theoretical results. Generated and interpreted material for illustrations. Wrote manuscript (redrafted by co-authors).		
Signature		Date	08/01/2019

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor Chris Holmes		
Supervisor comments The paper was largely conceived and developed by the candidate with guidance from the other authors. The candidate's description of their contribution is accurate.			
Signature		Date	08/01/2019

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 4

Posterior bootstraps for sparse regression and mixed models

4.1 Introduction

In Chapters 2 and 3 a framework for nonparametric learning with misspecified models has been developed, and extended to approximate posterior distributions and inference based on a loss function. In this chapter we investigate two further extensions to this framework.

First we consider the nonparametric (NP) learning framework in the context of sparse regression. Common Bayesian approaches to sparse regression are either computationally infeasible in high dimensions, or can involve strong prior assumptions that make their associated uncertainty quantification of questionable utility. We argue that these issues can be avoided by the use of NP learning. Splitting nonparametric data modelling from sparsity-inducing regression provides a more honest uncertainty quantification.

Second, we study the extension of nonparametric learning to parametric statistical models that contain random effects. In many modern applications observations cannot be assumed to be independent and identically distributed, as has been central to our NP learning development so far. We propose to extend our methodology to mixed models by

randomising an objective function based on an extended likelihood. The application of bootstraps to such data structures is far from straightforward (Davison and Hinkley, 1997, p. 100), so during this investigation we discuss a number of issues that we are faced with in practice.

The key message we take from the previous chapters is that computationally-efficient general Bayesian uncertainty quantification can be generated by suitably randomising an objective function. Up until now we had implicitly assumed that our loss function is additive with respect to observations, i.e.

$$\ell(\theta, x_{1:n}) = \sum_{i=1}^{n} \ell(\theta, x_i).$$

and that the data is comprised of n independent draws from a distribution F_0. In this chapter we seek ways of relaxing these assumptions.

The chapter is joint work with Stephen Walker and Chris Holmes, who have participated in discussions on its development. The investigations and write up of this chapter are solely this author's work.

4.2 A posterior bootstrap for sparse regression

4.2.1 Background

In recent years, datasets have not just grown in terms of the number of observations, but also in the number of measurements made per observation. In many cases this has been driven by technological advances that have made it cheaper and easier to automate data capture. For example, a human genome can now be sequenced in under a day, for around $1000, producing around 200 gigabytes of raw data (Turnbull et al., 2018). In comparison, the first human genome sequencing finished in 2003, having taken 13 years and costing $2.7bn (Venter et al., 2001).

When faced with high-dimensional data, sparsity-inducing statistical methods have a number of appealing properties. Low-dimensional summaries are usually easier to interpret than high-dimensional ones, can be faster to compute and often lead to better and more stable predictive performance. Hastie et al. (2009) describes the compelling 'bet on sparsity' principle, which states, informally, that if the 'truth' is sparse, then sparse methods are suitable for efficient estimation; however, if the truth is dense, then no method will be able to recover the truth.

Ridge regression was introduced by Hoerl and Kennard (1970) as a practical improvement on least squares that involves applying regression parameter shrinkage via ℓ_2 regularisation. This can improve predictive performance via a bias-variance tradeoff, but still results in dense parameter estimates. Traditional variable selection procedures such as best subset selection (Hastie et al., 2009, page 57) provide interpretable low-dimensional prediction algorithms that can improve predictive performance. However, these methods suffer from a curse of dimensionality, making them infeasible in high dimensions. Tibshirani (1996) introduced the *lasso* as a means of performing shrinkage and variable selection for the linear model by adding in an ℓ_1 norm penalisation to the least squares objective. This norm induces sparsity, while still being calculable in high dimensions, due to the convexity of the ℓ_1 penalty. Computationally-efficient algorithms have been developed for the lasso (Efron et al., 2004; Friedman et al., 2007) and have been made widely available (Friedman et al., 2010). This has led to the adoption of the method as a default approach for high-dimensional linear regression. The method has been extended beyond the linear model, for example to generalized linear models (Park and Hastie, 2007). The basic idea of using ℓ_1 regularisation to efficiently induce sparsity has been applied in a number of different fields, from graphical model edge selection (Meinshausen and Bühlmann, 2006) to sparse principal components analysis (Witten et al., 2009).

Generally considered as an adaptive prediction algorithm rather than an inference algorithm, the lasso is found to be lacking in terms of uncertainty quantification. From a

frequentist perspective, there has recently been some progress in post-selection inference (Berk et al., 2013) for methods such as the lasso, for example in constructing valid p-values and confidence intervals after a subset of the variables have been selected (Lockhart et al., 2014). From a Bayesian perspective, the lasso parameter estimates can be interpreted as the posterior mode under a normal linear model with double Exponential prior. The Bayesian lasso (Park and Casella, 2008) involves sampling from this posterior, instead of searching for its maximum. Although this is a preferable approach for a Bayesian, it is still likely to be problematic: even if we assume the statistical model is adequate, does this prior really represent our beliefs about the 'true' parameter value? Is the model true? If we are interested in sparse solutions, what use do we have for non-sparse posterior samples? Further, the underlying Gibbs sampler of the Bayesian lasso scales poorly with the dimension of the data.

If a Bayesian does believe that the data is generated by a normal linear model with a sparse parameter vector $\beta = (\beta_1, \ldots, \beta_p)$, then their prior for β should assign positive prior probability mass to $\beta_i = 0$ for some or all of $i = 1, \ldots, p$. A number of Bayesian variable selection methods take this approach, including the spike and slab prior (Mitchell and Beauchamp, 1988), where the prior is a mixture of a point mass at $\beta_i = 0$ and a continuous probability density. The problem with this approach is a computational one, as the solution set of sparse β vectors grows combinatorially with p.

The linear model is wrong, however, so a more reasonable approach is to interpret the lasso as the solution to an optimisation given a sparsity-inducing penalised least squares objective function. In this section we consider the application of the nonparametric learning approach of Chapter 3, to quantify uncertainty in the target parameter by randomising an objective function.

4.2.2 The NPLasso

For uncertainty quantification of sparse linear predictors, we present nonparametric learning for the least absolute shrinkage and selection operator, or the *NPLasso* for short.

Suppose there is a dataset $\mathcal{D}_n$ made up of n observations, each comprised of a vector of covariates and a real-valued response,

$$\mathcal{D}_n = \{z_i = (x_i, y_i) \in \mathbb{R}^p \otimes \mathbb{R},\ i = 1, \ldots, n\}.$$

Suppose for simplicity that the z_i's are IID draws from an unknown distribution F_0,

$$z_i \sim F_0(z_i) = F_0(x_i)F_0(y_i \mid x_i); \quad i = 1, \ldots, n.$$

We are interested in learning a sparse linear prediction model for y^* as a linear function of x^*, where $z^* = (x^*, y^*)$ is another independent draw from F_0. The lasso of Tibshirani (1996) amounts to the minimisation of the following objective function,

$$\ell_{\lambda_n}(\mathcal{D}_n, \beta) = \sum_{i=1}^{n}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|,$$

where $\beta_0 \in \mathbb{R}$, $\beta \in \mathbb{R}^p$ are the parameters and $\lambda_n \geq 0$ is a penalty hyperparameter.

The parameters of interest, denoted $(\beta_0^0, \beta^0) \in \mathbb{R} \otimes \mathbb{R}^p$, are the minimisers of the expected loss,

$$\begin{aligned}\beta_0^0, \beta^0 &= \arg\min_{\beta_0,\beta} \int \ell_{\lambda_n}(\mathcal{D}_n, \beta) dF_0(\mathcal{D}_n) \\ &= \arg\min_{\beta_0,\beta} \left\{ \int \left(y - \beta_0 - x^T\beta\right)^2 dF_0(x, y) + \lambda_n^* \parallel \beta \parallel_1 \right\},\end{aligned}$$

with $\lambda_n^* = \lambda_n/n$. Note that if λ_n^* does not depend on n then we have a loss function that is additive over observations. In this case, the generalisation of NP-learning to utility functions of Section 3.4.4 can be applied directly. If this is not the case, we still have a well-defined

objective function, for any n, and so can proceed by randomising this objective function in accordance with the posterior bootstrap algorithm of Section 3.4.5.

As the lasso is often used as a default procedure, we suggest that the concentration parameter of NP-learning, c, should be set to zero as a default. Our method then simplifies to repeatedly, and independently, drawing Dirichlet-distributed weights for each observation, and then computing sparse parameter estimates by optimisation of the weighted objective. This can be done efficiently using weighted versions of the algorithms implementing the lasso, such as least angle regression (Efron et al., 2004) and coordinate descent (Friedman et al., 2007). These methods accommodate weighted samples as a simple extension. Further, these routines compute the whole parameter path as a function of the penalty λ_n. Thus, we could consider the parameter of interest to be the function that maps a value of λ_n to a parameter vector.

If we have at our disposal a useful Bayesian parametric model for (x, y, β_0, β), then we can include this information into our nonparametric learning routine, in the way that we described in Section 3.4.4, by generating and randomly weighting pseudo-data from the posterior predictive of the parametric Bayesian model. One candidate for this would be a conjugate Bayes linear model, which could be sampled from efficiently. However, our method could accommodate any Bayesian model that can be sampled from.

Posterior samples retain the sparsity of lasso estimates, which is not true of other shrinkage-based Bayesian variable selection routines. For example, the probability that a parameter is non-zero can be computed; this is not possible under the Bayesian lasso. Further, NPLasso samples can be generated quickly and in parallel, not suffering from the computational issues that MCMC-based sampling routines face in high dimensions.

Central to our method is the separation of the nonparametric modelling of the data-generating mechanism, and the determination of an optimal sparse prediction model. Most Bayesian methods for variable selection intertwine these two aspects, leaving the user with a heavily misspecified model which may bring into question the usefulness of the associated

uncertainty quantification. Instead, by reasoning honestly about the data-generating mechanism and the objective function, our method provides a more reasonable quantification of the uncertainties present.

4.2.3 Illustrations

The performance of the NPLasso is now demonstrated for linear regression on a synthetic dataset, defined as follows. For each of n observations, the covariates are independent draws from a p-dimensional multivariate Gaussian, centred at zero with unit variance and pairwise correlation $\rho = 1/2$. The response y_i is generated according to a linear model

$$y_i = x_i^T \beta_0 + \epsilon_i, \quad i = 1, \ldots, n,$$

with β_0 fixed, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ independently.

Given a dataset, B posterior samples are generated under the NPLasso, the Bayesian lasso, as well as two other common shrinkage priors: the horseshoe prior (Carvalho et al., 2010) and the normal-gamma prior of Griffin and Brown (2010). For all methods apart from the NPLasso, the posterior samples are collected after the first B samples are discarded as burn-in. Performance is measured by the squared error of the posterior $\bar{\beta}$ mean for β, relative to its true value, β_0,

$$\text{SE}(\bar{\beta}, \beta_0) = \| \bar{\beta} - \beta_0 \|^2 .$$

For the lasso and NPLasso we record the squared error as a function of λ_n, as well as the squared error when λ_n is determined by ten-fold cross-validation for the lasso. This whole procedure is repeated for 100 independently generated datasets. The lasso implementation from the `glmnet` R package (Friedman et al., 2010), which uses the coordinate descent algorithm, is used for the lasso and within the NPLasso. Posterior sampling for the Bayesian lasso, horseshoe and normal Gamma priors was achieved using the implementation in the

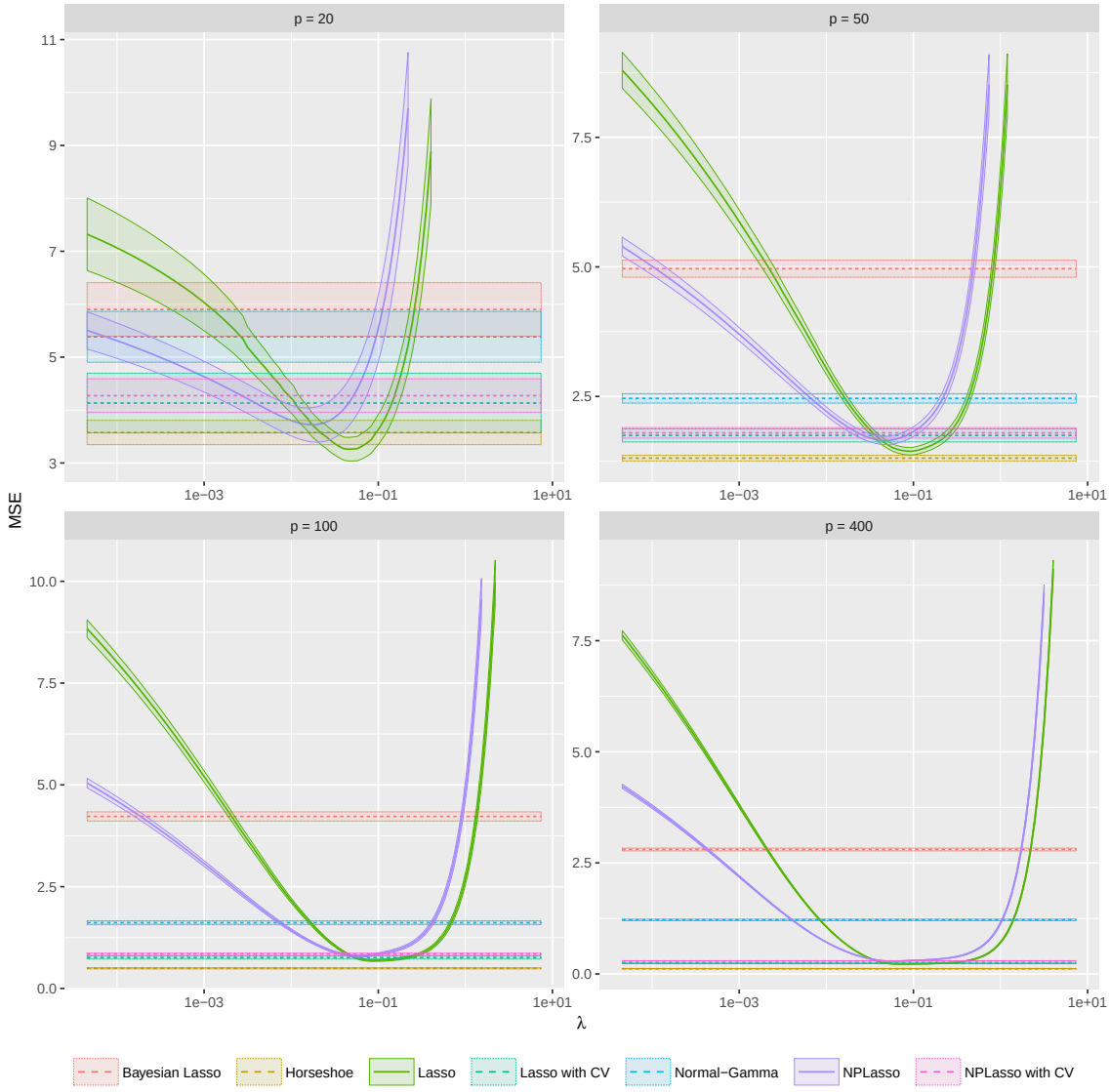


Figure 4.1: Mean Squared Error as a function of the penalty parameter λ, for the synthetic dataset. Error bars represent ± 1 standard error from the mean.

`monomvn` R package (Gramacy and Pantaleo, 2010).

Figure 4.1 shows the mean squared error for $p \in \{20, 50, 100, 400\}$, $n = p + 5$, $\sigma^2 = 1$ and $\beta_0 = (1, 2, \ldots, 10, 0, \ldots, 0)$. A similar set up was used in Carvalho et al. (2009). The posterior means are estimated from $B = 1000$ samples, except for the $p = 400$ case where $B = 100$ is used, due to the slow running times of the Bayesian lasso, horseshoe and normal-gamma prior methods. The NPLasso outperforms the Bayesian lasso and the normal-gamma prior method in terms of mean squared error, for all values of p. The performance of

p	Bayesian lasso	NPLasso$_1$	NPLasso$_{96}$	NPLasso$_1$ vs. Bayesian lasso	NPLasso$_{96}$ vs. Bayesian lasso
20	17 mins	128 minutes	99 seconds	$0.14\times$	$10.5\times$
50	6.1 hours	145 minutes	113 seconds	$2.5\times$	$196\times$
100	3.4 days	154 minutes	150 seconds	$32\times$	$1,975\times$
200	52 days	6.2 hours	4.7 minutes	$201\times$	$16,069\times$
300	264 days	12 hours	9.5 minutes	$530\times$	$39,820\times$
400	2.6 years	25 hours	20 minutes	$904\times$	$69,710\times$

Table 4.1: Average projected time for generating 1 million samples, computed over 10 repetitions. All runtime standard errors are below 6% of the mean. The right-most two columns display the ratio of projected runtimes of the Bayesian lasso relative to the two NPLasso methods.

the NPLasso, lasso and horseshoe prior are very similar, with the horseshoe prior method slightly outperforming the NPLasso and lasso. The cross-validation solution for λ_n seems to work well for the NPLasso, getting close to the global minimiser.

Table 4.1 provides a comparison of average runtimes between our methods for the synthetic example. Runtimes were computed on a `m5d.24xlarge` AWS EC2 instance containing 96 CPU cores. The runtimes shown represent the sampling rate for generating 1 million posterior samples. This number of samples was actually generated under the parallel NPLasso$_{96}$ method that exploited the 96 CPUs available, though for the serial NPLasso$_1$ the runtime for 1000 posterior samples was scaled to calculate this rate. For the Bayesian lasso, we generated 1000 samples for $p \in \{20, 50, 100\}$ and only 100 samples for $p \in \{200, 300, 400\}$. The Bayesian lasso runtime scales poorly with dimension, whereas the NPLasso scales much better, and can be further accelerated by applying the computation in parallel. The NPLasso functions are coded in R; additional performance gains are to be expected if the method was implemented in a lower-level programming language.

The prostate cancer dataset was considered in Tibshirani (1996), and is a canonical choice for linear model variable selection; see for example Hastie et al. (2009). The response consists of 97 measurements of the log prostate specific antigen, `lpsa`, with 8 covariates provided for each observation. A linear model is used to predict the response from the covariates. Figure 4.2 shows the coefficient profile plots for the Prostate cancer data under the

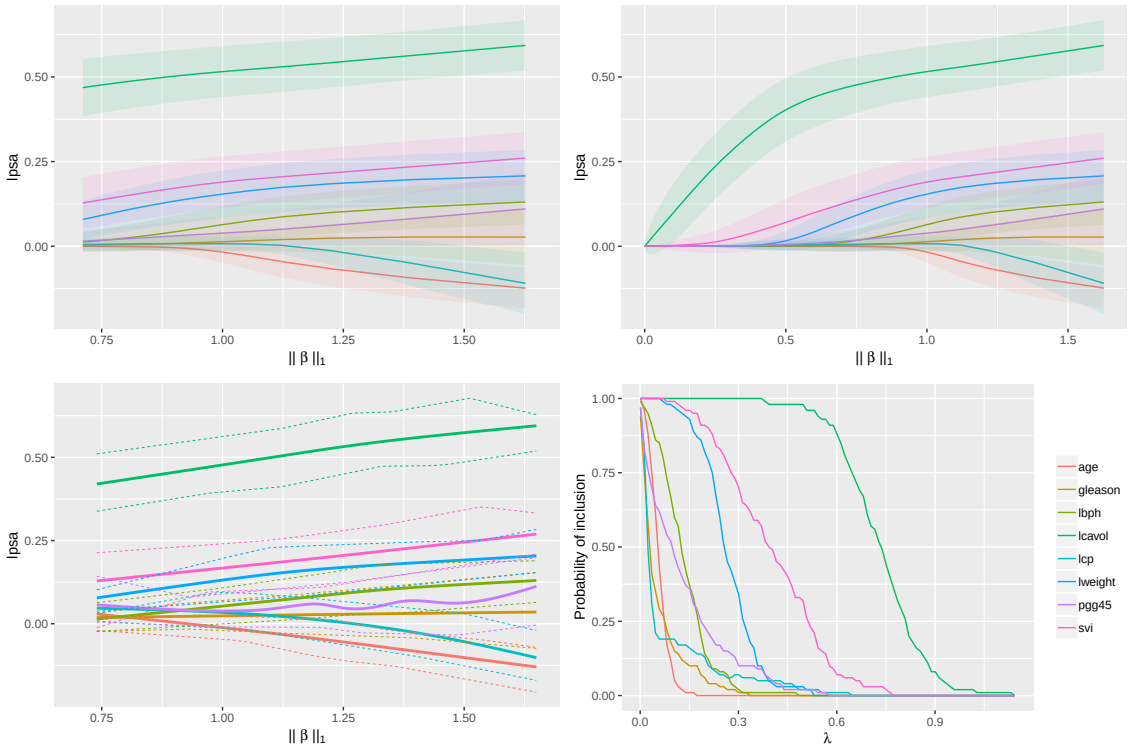


Figure 4.2: Profile plots of NPLasso (top-left and top-right) and Bayesian lasso (bottom-left) coefficients, as a function of $\| \beta \|_1$, and posterior probability of covariate inclusion for NPLasso (bottom-right), as a function of λ, for the prostate cancer dataset.

NPLasso and the Bayesian lasso. The conditional posterior mean and standard deviations are hard to distinguish when put on the same scale. Note that the NPLasso provides a posterior over the parameter vector as a function of λ, whereas the Bayesian lasso provides a set of posterior samples over (λ, β), jointly. Figure 4.2 also shows the posterior probability that each covariate takes a non-zero value, as a function of the penalty λ. This quantity is not directly available for the lasso or any of the Bayesian shrinkage prior-based methods.

4.2.4 Discussion

The NPLasso provides a direct and honest means of quantifying uncertainty about sparse parameter models for linear regression, and could easily be extended to the many other applications of ℓ_1 penalisation. It is computationally scalable and can incorporate information from a parametric model, or approximate posterior if available. Bayesian approaches to

sparse signal regression, such as the Bayesian lasso, are based on questionable Bayesian modelling assumptions that in turn lead to uncertainty quantification of questionable utility. Additionally, MCMC-based posterior sampling scales poorly with the dimensionality of the dataset.

In terms of further work, the NPLasso as presented can either be considered as providing a path of parameter estimates as a function of the penalty λ, or for a specific value of λ chosen by the practitioner. It should be possible to use the posterior samples and Dirichlet weights to estimate an appropriate value of λ directly. If there was positive probability that one of the observation weights would take the value zero, then we could estimate the error using the observations associated with the zero weights. This is basically the idea underlying the .632 estimator for the nonparametric bootstrap (Efron, 1983). Something similar could be done with the smooth Dirichlet weights by mapping smaller *training* weights to larger *test* weights, and vice versa.

It would be interesting to study the performance of the method, relative to the Bayesian approaches, when the linear model assumptions are violated. By targeting the objective directly, does the NPLasso provide more robust inferences? If not, the investigation of a robust version of the NPLasso may be of interest.

By setting $c = 0$ the NPLasso amounts to a Bayesian bootstrap for the lasso objective. Nonparametric bootstraps applied to the lasso have been studied previously. Both Bach (2008) and Chatterjee and Lahiri (2011) use the bootstrap in conjunction with the lasso to study consistency properties of the lasso. This is not of direct interest to us, but the theoretical asymptotic superiority of the NPLasso relative to the Bayesian lasso would be interesting future work.

4.3 A posterior bootstrap for mixed models

4.3.1 Introduction

In many problems in applied statistics the data-generating mechanism will exhibit some hierarchical structure. In the simplest case this could mean that there are groups of conditionally independent observations, or, more generally, that observations share some group-level dependence structure. Ignoring such structure can lead to suboptimal statistical analyses, where the information contained in the dataset is not fully utilised.

Statistical models for grouped data are commonly constructed by hierarchically modelling within-group and between-group effects. This enables the sharing of information across groups, and allows for more insightful inferences about the phenomena of interest. In this section we study mixed effects models, where it is assumed that the data belongs to a statistical model parametrised by three distinct types of parameter:

- fixed effect parameters, whose values *are not* assumed to have been drawn from a probability distribution,
- random effect parameters, whose values *are* assumed to have been drawn from a probability distribution,
- dispersion parameters, that parametrise the probability distribution from which the random effects are drawn, and any additional variation not determined by the fixed and random effects.

Two canonical examples of mixed models are the linear mixed model and the generalised linear mixed model, which we shall now introduce.

Example 1 (The linear mixed model). *Let y be an n-dimensional response vector, $X \in \mathbb{R}^{n\times p}$ be called the fixed effect model matrix, $Z \in \mathbb{R}^{n\times q}$ the random effect model matrix, $\beta \in \mathbb{R}^p$ the fixed effect parameter vector, $v \in \mathbb{R}^q$ the random effect vector, and $\epsilon \in \mathbb{R}^n$ the*

noise term. The normal linear mixed model (LMM) is defined as

$$y = X\beta + Zv + \epsilon$$

with v *and* ϵ *independent, with* $v \sim \mathcal{N}(0, D)$ *and* $\epsilon \sim \mathcal{N}(0, \Sigma)$.

Example 2 (The generalised linear mixed model). *Let* y *be an* n*-dimensional response vector,* $X \in \mathbb{R}^{n \times p}$ *the fixed effect model matrix,* $Z \in \mathbb{R}^{n \times q}$ *the random effect model matrix,* $\beta \in \mathbb{R}^p$ *the fixed effect parameter vector,* $v \in \mathbb{R}^q$ *the random effect vector,* $\psi \in \mathbb{R}^n$ *the canonical parameter,* g *the link function and* $\phi > 0$ *the dispersion parameter. The generalised linear mixed model (GLMM) is defined as follows:*

- *The distribution* y *given* v *belongs to a generalised linear model (McCullagh and Nelder, 1989) with linear predictor* $g(\mu) = \eta = X\beta + Zv$*, with* μ *the conditional mean; that is*

$$f(y \mid u, \psi, \phi) = C(y, \phi) \exp\left\{ \left(y^T \psi - b(\psi) \right) / \phi \right\}$$

- $v \sim \mathcal{N}(0, D)$

In both examples, we call β the fixed effects and v the random effects. The covariance matrices D and Σ are dispersion parameters for the linear mixed model; for the generalised linear mixed model it is the same with Σ replaced by ϕ. In practice the matrices are commonly given a lower-dimensional parametrisation, i.e. $D = D(\theta)$, $\Sigma = \Sigma(\theta)$ for some $\theta \in \mathbb{R}^d$. We shall discuss these models further later in the chapter.

The posterior bootstrap algorithm developed in Chapter 3 cannot be applied to data with a hierarchical structure; it is only formally applicable to data generated as independent draws from an unknown distribution function. In the rest of this section we consider how to generalise the ideas surrounding the posterior bootstrap of Chapter 3 to hierarchical data

structures containing random effects. This is done by extending the notion of uncertainty quantification by randomising an objective function.

4.3.2 Inference for mixed models

First, we provide some background by briefly reviewing the common methods for statistical inference for a mixed model.

The Bayesian approach is a natural fit for mixed models, in essence because it is clear how probability distributions over unknown parameters should be treated. It should be noted, however, that there is a distinction in interpretation between a subjective prior over a fixed effect, and a physically modelled but unobserved distribution over a random effect.

Suppose we have a generalised linear mixed model as described in Example 2, where θ represents all unknown dispersion parameters. The joint probability distribution can be factorised as follows:

$$f(y, \beta, v, \theta) = f(y \mid \beta, v, \theta) f(v \mid \theta) \pi(\beta, \theta),$$

where $\pi(\beta, \theta)$ represents the joint prior over the fixed effects and dispersion parameters. The posterior distribution $\pi(\beta, v, \theta \mid y)$ is proportional to $f(y, \beta, v, \theta)$, though the constant of proportionality is a potentially high-dimensional integral. Direct estimation of this constant can be avoided via MCMC methods, which can be used to generate a set of samples from a Markov chain whose equilibrium distribution is the joint posterior over all parameters. The Bayesian approach fully quantifies the uncertainty across the parameters and future observations, conditional on the model being true. Leaving aside model misspecification and the appropriateness of the prior, which was discussed in Chapter 1, the main problem with taking a Bayesian approach is that the MCMC-based sampling methods are computationally burdensome and have some undesirable properties. These algorithms often scale poorly with the size of the dataset, can be challenging to tune adequately so that they mix well over

the parameter space, and are commonly serial algorithms that produce correlated posterior samples, only whose stationary distribution is of interest.

For problems of moderate size, the computational issues of MCMC are reasonably well understood, and a number of probabilistic programming packages are available to smooth over some of these computational issues and provide a fairly reliable set of posterior samples. Popular examples include BUGS (Lunn et al., 2000) and Stan (Carpenter et al., 2017). Both admit a graphical model as an input and return a set of posterior samples as the output. For high-dimensional problems it is common to resort to an approximation, such as a variational approximation, or to use a more tractable conjugate model that may not perfectly reflect the practitioner's subjective beliefs. These approaches bring into question the validity and usefulness of the uncertainty quantification provided by the Bayesian approach.

In the frequentist school, a number of methods have been proposed for statistical inference with mixed models. These methods conflict in some respects, and use different approaches for the different types of unknown, which is in contrast with the completeness of the Bayesian approach. For the fixed effects and dispersion parameters the method of maximum likelihood can be applied to the marginal likelihood,

$$\hat{\beta}, \hat{\theta} = \arg\max_{\beta,\theta} \int f(y, v \mid \beta, \theta) dv.$$

This integral can quickly become problematic, as the number of random effects grows. Additionally, joint maximisation over fixed and dispersion parameters can yield dispersion parameter estimates that some consider undesirable, as their values may be sensitive to the fixed effect estimates. Restricted maximum likelihood (REML) overcomes this problems. Originally constructed for the linear mixed model (Patterson and Thompson, 1971) and later generalised (Schall, 1991), REML is maximum likelihood inference for a linear function of the response vector chosen so that it is invariant to the values of the fixed effects. Such estimates have better robustness properties, but, as is also true under maximum likelihood,

can still provide degenerate estimates of variance parameters. Further, as they have been eliminated, REML does nothing to provide fixed effect estimates. A common approach is to plug in the REML dispersion parameter estimates and maximise with respect to the fixed effects. Estimation of the random effects is often referred to as *prediction*, given the different status of the unobserved random effects relative to the fixed effects. Decision-theoretic estimates are popular, such as best linear unbiased prediction (BLUP); Robinson (1991) provides a number of motivations for this.

The complexities of the above approaches, as well as concerns about the goodness of the underlying model specification, have led to the development of approximate methods that don't rely on a full model specification. Examples include quasi-likelihood (Breslow and Clayton, 1993) and estimating equations (Liang and Zeger, 1986), as well as maximum h-likelihood estimation, which we shall discuss next.

4.3.3 H-likelihood and the Extended Likelihood Principle

Another approximate means of avoiding maximising the marginal likelihood is to maximise the *h-likelihood*, which is the joint probability density of the observations and the random effects, written as a function of the fixed and random effects,

$$f(y, v \mid \beta, \theta) = f(y \mid v, \beta, \theta) f(v \mid \theta)$$

It is not entirely clear what the 'h' stands for here, though the obvious candidates are 'hierarchical', or 'Henderson' who originally studied this approach for the linear mixed model (Henderson, 1953). Lee and Nelder (1996) explored this idea for the hierarchical GLM, which is similar to the GLMM but with a conjugate random effect distribution that undergoes a transformation before entering into the linear predictor. The authors showed that this method provides fixed effect estimates that coincide with maximum marginal likelihood estimates as the sample size tends to infinity, and random effect predictions that

are asymptotically best unbiased, however under restrictive regularity conditions where the number of random effects may not grow with the number of observations.

Although the method of Lee and Nelder (1996) lacked a formal justification, over time a more concrete motivation for maximum h-likelihood estimation (MHLE) was developed. This was achieved by extending the likelihood principle (Birnbaum, 1962) to models including random effects. Bjørnstad (1996) did this by suitably generalising the sufficiency and conditionality principles, and showing that the extended likelihood principle followed from these two. In the following we assume that all dispersion parameters are known; Lee and Nelder (1996) suggest prior estimation of these via an adjusted profile h-likelihood.

Following the argument of Lee et al. (2006), likelihood-based inference for fixed effects models involves making a connection between data generation and parameter inference via Fisher's likelihood. That is, the probability of the observed data, written as a function of the unknown parameters, contains all of the information in the sample relevant to inference for the unknown parameters. If our statistical model includes random effects, this argument does not hold, but there is an extension to it. For mixed models we again relate the concepts of data generation and parameter inference, but each of these is now a two-stage process. Data-generation involves first generating a random effect, and then generating data given that random effect. For parameter inference, the likelihood principle tells us that inferences about β should be made using the marginal likelihood,

$$L(\beta; y) \equiv f_\beta(y) = \int f_\beta(y, v) dv.$$

Further, it is argued that if we knew β, we should make inferences about v using the conditional likelihood,

$$L(\beta, v; v \mid y) \equiv f_\beta(v \mid y).$$

The extended likelihood for the fixed and random effects is defined as the product of these

likelihoods, and the connection with the data-generating mechanism is clear:

$$L(\beta, v; y, v) \equiv f_\beta(y, v) = f_\beta(y \mid v) f(v).$$

This definition of extended likelihood has appeared elsewhere in the literature, see for example Berger and Wolpert (1988). As is the case with the likelihood principle, the extended likelihood principle does not prescribe how inferences should be made. Under misspecification, we have argued in Chapter 2 that the log likelihood is an appropriate objective when observations are independent and identically distributed. Our argument now is that, in the same way, the log h-likelihood is the appropriate objective for models containing unobservables. We extend the posterior bootstrap to cover models with random effects by randomising the log h-likelihood objective.

Lee and Nelder (1996) received a mixed reception. The method performed poorly on a number of canonical examples, including logistic regression, particularly in settings where there were a small number of samples per group. Further, it lacked any firm foundational justification. In later work by the same authors (Lee and Nelder, 2005; Lee et al., 2007) it was proposed that the h-likelihood should only be used for predicting random effects, and that fixed effects should be estimated by maximising the marginal likelihood.

4.3.4 The weighted h-likelihood bootstrap

Suppose we have observed I groups of conditionally independent observations, where group i contains n_i observations $y = (y_{i1}, \ldots, y_{in_i})$, for $i = 1, \ldots, I$. Under a GLMM with fixed effects β, random effects $v = (v_1, \ldots, v_I)$ and dispersion parameters θ, the log h-likelihood can be written as follows:

$$\log f(y, v \mid \beta, \theta) = \sum_{i=1}^{I} \log f(v_i \mid \theta) \; + \; \sum_{i=1}^{I} \sum_{j=1}^{n_i} \log f(y_{ij} \mid v_i, \beta, \theta).$$

Following the weighted likelihood bootstrap of Newton and Raftery (1994), we assign random weights to each term of the log h-likelihood. This can be done in a number of ways, however. We could just assign weights to the observation-level terms of the likelihood factorisation, $f(y_{ij} \mid \beta, v, \theta)$, as is done by the weighted likelihood bootstrap, and determine group weights by adding up the observation-level weights, suitably scaled. Alternatively we could assign weights to the group-level *unobservable* terms $f(v_i \mid \theta)$ as well as the *observable* terms $f(y_{ij} \mid v_i, \beta, \theta)$. There are two natural ways that this could be achieved. We could assign independent random weights to the group-level and observation-level terms, or we could apply the i^{th} group weight to the $f(v_i \mid \theta)$ term and to all of the observation-level terms in group i, and then apply additional observation-level weights that are multiplied by the group-level weights. We shall investigate all of these weight configurations empirically. We can write down the weighted h-likelihood bootstrap (WHLB) objective covering all of these possibilities as follows:

$$\tilde{v}, \tilde{\beta} = \arg\max_{v,\beta} \left\{ \sum_{i=1}^{I} w_i^* \log f(v_i \mid \hat{\theta}) + \sum_{i=1}^{I} \sum_{j=1}^{n_i} w_{ij} \log f(y_{ij} \mid v_i, \beta, \hat{\theta}) \right\}. \tag{4.1}$$

The three weight configurations, by which we mean the distributions on the weights

$$w = \{w_{ij}; j = 1, \ldots, n_i, i = 1, \ldots, I\}, \qquad w^* = \{w_i; i = 1, \ldots, I\},$$

are specified in Table 4.2. Dirichlet weights are used in keeping with the NP learning weight configuration of Chapter 3. These are suitably scaled to match the log likelihood, in expectation over the weights. The use of unnormalised weights that do not require scaling, such as $\text{Exp}(1)$ weights, was investigated but made little difference to the properties of the WHLB.

Note that in (4.1) we have replaced the unknown dispersion parameters θ with a plug-in estimate $\hat{\theta}$. This is in keeping with the frequentist notion of h-likelihood, though is not in keeping with Bayesian reasoning. In our LMM example we also consider maximum

	Observation weights	Group weights
Independent Dirichlet	$n^{-1}w \sim \text{Dirichlet}(1, \ldots, 1)$	$I^{-1}w^* \sim \text{Dirichlet}(1, \ldots, 1)$
Observation Dirichlet	$n^{-1}w \sim \text{Dirichlet}(1, \ldots, 1)$	$nI^{-1}w_i^* = \sum_{j=1}^{n_i} w_{ij}$
Product Dirichlet	$w_{ij} = nw_i^* v_{ij}$ $v \sim \text{Dirichlet}(1, \ldots, 1)$	$I^{-1}w^* \sim \text{Dirichlet}(1, \ldots, 1)$

Table 4.2: WHLB weight configurations considered. $n := \sum_i n_i$. All sets of Dirichlet-distributed weights are independent.

h-likelihood dispersion parameters as well, i.e.

$$\tilde{v}, \tilde{\beta}, \tilde{\theta} = \arg\max_{v,\beta,\theta} \left\{ \sum_{i=1}^{I} w_i^* \log f(v_i \mid \theta) + \sum_{i=1}^{I} \sum_{j=1}^{n_i} w_{i,j} \log f(y_{i,j} \mid v_i, \beta, \theta) \right\}.$$

To investigate the properties of the weighted h-likelihood bootstrap further, we study the method's performance for a simple linear mixed model using synthetically-generated data, followed by a generalised linear mixed model for a real dataset.

4.3.5 Normal means example

Consider the following simple normal means model with a random effect:

$$y_{ij} = \beta + v_i + \epsilon_{ij}; \quad j = 1, \ldots, n_i, \; i = 1, \ldots, I; \tag{4.2}$$

$$v_i \sim \mathcal{N}(0, \sigma_v^2); \quad i = 1, \ldots, I; \tag{4.3}$$

$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma_y^2); \quad j = 1, \ldots, n_i, \; i = 1, \ldots, I. \tag{4.4}$$

We shall first study the case where the dispersion parameters, σ_v^2 and σ_y^2 are known. There is no tractable description of the distribution of the posterior parameters in general, so we proceed by means of simulation. We compare empirically the WHLB posterior distribution to the Bayesian posterior, computed using Stan, for a number of different balanced hierarchical data structures. The following experiment was conducted 100 times. A dataset was generated according to equations (4.2, 4.3, 4.4), with the true $\beta = 1$, $\sigma_y^2 = 1$ and $\sigma_v \in \{1, 10\}$. A posterior sample of size 1000 was generated from both the WHLB posterior

and the well-specified Bayesian posterior. The posterior mean and standard deviation were calculated for all unknowns. This experiment was conducted on balanced datasets, i.e. $n_1 = \ldots = n_I$, with $(n_1, I) \in \{(1, 100), (10, 100), (100, 10), (100, 100)\}$. Several further configurations were investigated but these four provide a concise comparison of the properties of the WHLB. Flat priors were applied to all unknowns in the Bayesian model. We considered all three weight configurations specified in Table 4.2.

In the left pane of Figure 4.3 we see that when the dispersion parameters are known, the *Independent Dirichlet* and *Product Dirichlet* weight configurations seem to provide a reasonable match to the well-specified Bayesian posterior. However, the *Observation Dirichlet* configuration, that generates just one set of weights for the observations, produces WHLB posterior samples that are reasonably well centred but much less dispersed than under the Bayesian posterior. Looking just at the Independent Dirichlet weights in the right pane of Figure 4.3, the posterior means match well but there are large discrepancies in the posterior variances when there are few observations per group, or few groups. The findings are similar for the Product Dirichlet weights, though with slightly less similarity with Bayes for the mean of the first random effect, mean(v_1).

Figure 4.4 shows the same plot as the left pane of Figure 4.3, but for the case where the dispersion parameters are also solved for as part of the WHLB maximisation, i.e.

$$\tilde{v}, \tilde{\beta}, \tilde{\sigma}_v, \tilde{\sigma}_y = \underset{v,\beta,\sigma_v,\sigma_y}{\arg\max} \left\{ \sum_{i=1}^{I} w_i^* \log f(v_i \mid \sigma_v) + \sum_{i=1}^{I} \sum_{j=1}^{n_i} w_{i,j} \log f(y_{i,j} \mid v_i, \beta, \sigma_y) \right\}.$$

This does not seem to perform too badly for the Independent Dirichlet weights, but there are some significant differences under the other weight distributions. The Observation Dirichlet weights underplay the dispersion of the fixed and random effects, and σ_v, compared to Bayes. The standard deviation of σ_y under the Product Dirichlet weights is roughly double the value attained by the Bayesian posterior.

Figure 4.5 provides a comparison of the posterior means and standard deviations, under

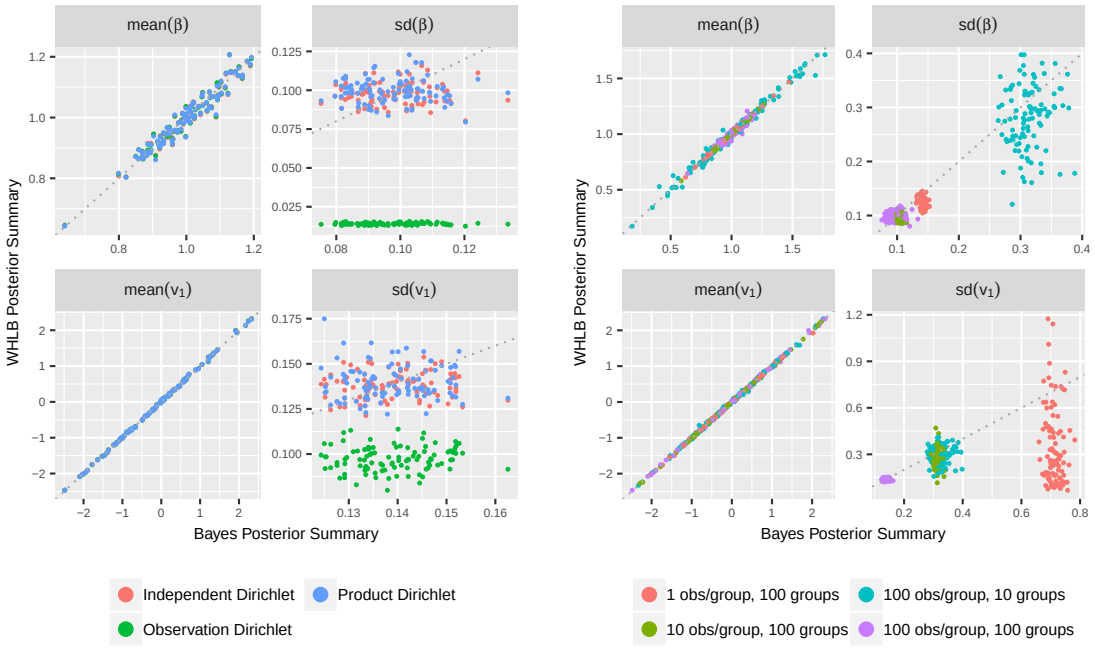


Figure 4.3: Normal means posterior moment plots with dispersion parameters known. Left: WHLB vs Bayes scatter plots for the three weight distributions, with $(n_1, I) = (100, 100)$. Right: WHLB vs Bayes scatter plots for all data structures and Independent Dirichlet weights. The dotted line represents equality in posterior summaries.

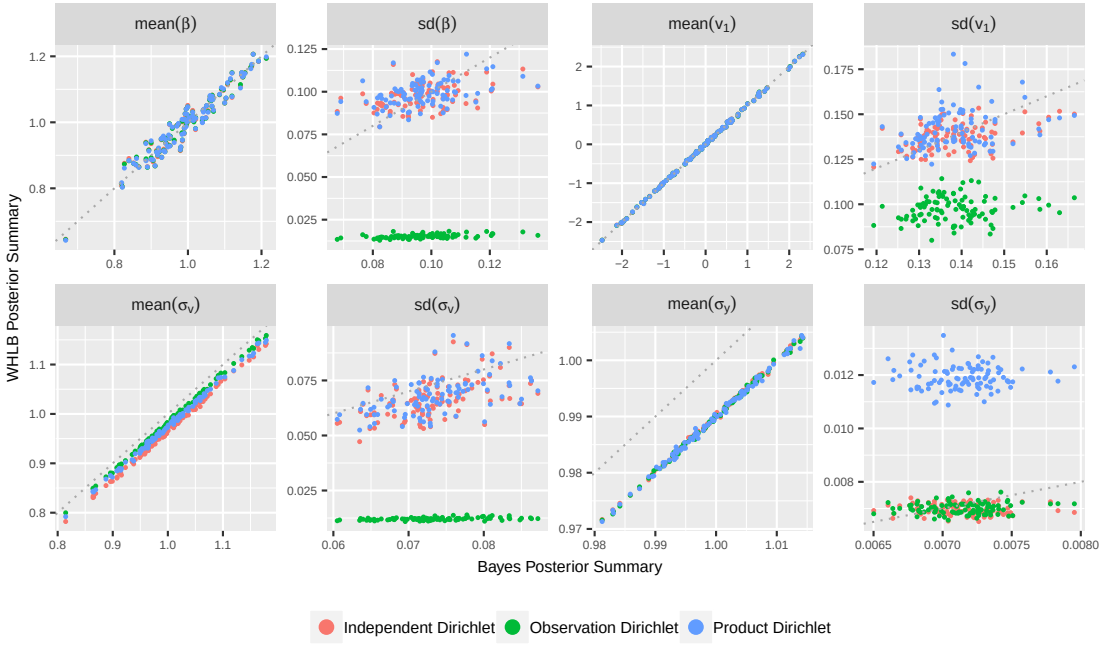


Figure 4.4: Normal means posterior plots with dispersion parameters unknown, for the three weight distributions, for the saturated configuration $(n_1, I) = (100, 100)$. The dotted line represents equality in posterior summaries.

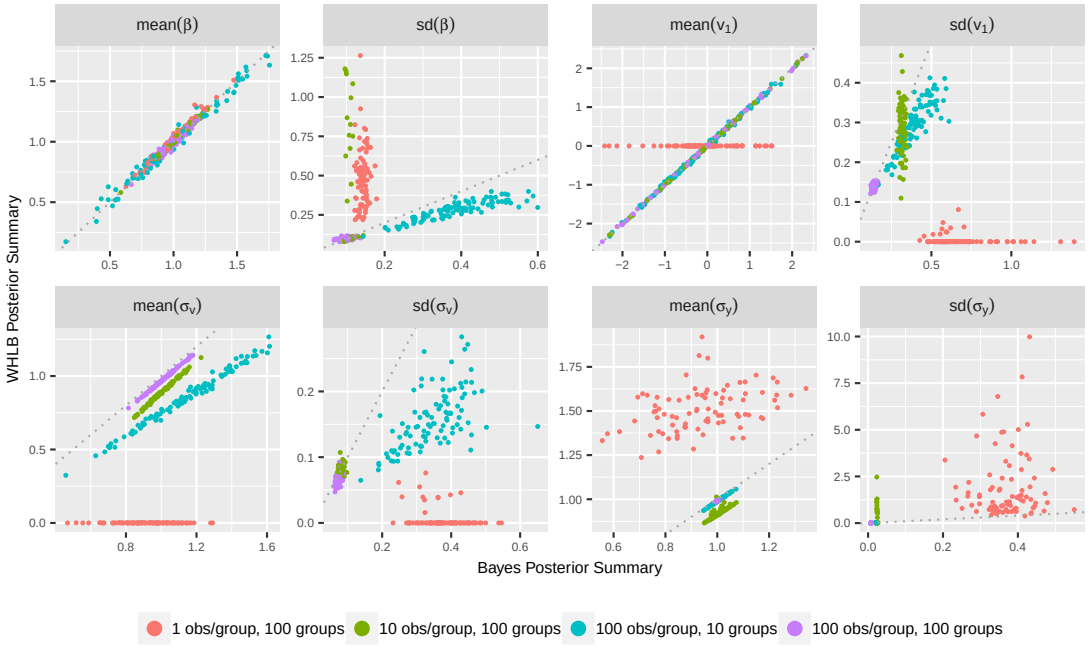


Figure 4.5: Normal means posterior plots with dispersion parameters unknown, for the four balanced data structures, using the Independent Dirichlet weight distribution. The dotted line represents equality in posterior summaries.

the *Independent Dirichlet* weight configuration, for the four data structures considered. There are meaningful differences with the Bayesian posterior in all but the saturated case where $(n_1, I) = (100, 100)$. This is in contrast with what is observed when the dispersion parameters are known, where, aside from the one observation per group case, the two posteriors are reasonably comparable.

4.3.6 Salamander mating dataset

The Salamander dataset, originally in McCullagh and Nelder (1989), consists of a binary response corresponding to success of mating between 360 specified pairs of salamanders. The data combines the results from three experiments, two of which were done with the same salamanders in summer and autumn, and a third in the same autumn but with a different set of salamanders. For each experiment 20 male and 20 female salamanders, from either the 'Whiteside' and 'Rough Butt' populations, were paired on six occasions with a salamander

of the opposite sex, from either population.

Lee and Nelder (1996) proposed a logistic generalised linear mixed model with two random effects which are crossed; one for each salamander of each sex. Additionally, fixed effects are included; specifically an intercept, an indicator variable for each of the male and female population, plus an interaction term. The binary response for the experiment involving the i^{th} female and the j^{th} male in the k^{th} experiment is denoted as $y_{ijk} \in \{0, 1\}$. The model can be written as

$$\begin{aligned} \log\{p_{ijk}/(1-p_{ijk})\} &= x_{ijk}^T\beta + v_{ik}^f + v_{jk}^m \\ v_{ik}^f &\sim \mathcal{N}(0, \sigma_f^2) \\ v_{jk}^m &\sim \mathcal{N}(0, \sigma_m^2) \end{aligned}$$

for $i, j = 1, \ldots, 20$ and $k = 1, 2, 3$, where $p_{ijk} := \mathbb{P}(y_{ijk} = 1 \mid v_{ik}^f, v_{jk}^m)$, with $v^f = \{v_{ik}^f; i = 1, \ldots, 20; k = 1, 2, 3\}$ and $v^m = \{v_{jk}^m; j = 1, \ldots, 20; k = 1, 2, 3\}$ mutually independent. We use the following WHLB objective:

$$\begin{aligned} \tilde{v}^f, \tilde{v}^m, \tilde{\beta} = \underset{v^f, v^m, \beta}{\arg\max} \Bigg\{ &\sum_{k=1}^{3}\sum_{i=1}^{20} w_{ik}^* \log f(v_i^f \mid \hat{\sigma}_f^2) + \sum_{k=1}^{3}\sum_{j=1}^{20} w_{jk}^\star \log f(v_j^m \mid \hat{\sigma}_m^2) \\ &+ \sum_{k=1}^{3}\sum_{l=1}^{120} w_{lk} \log f(y_{i[lk]j[lk]k} \mid v_{i[lk]k}^f, v_{j[lk]k}^m, \beta) \Bigg\} \end{aligned} \tag{4.5}$$

where $i[lk], j[lk]$ denotes the index of the male, female in the l^{th} pairing of the k^{th} experiment, respectively. The weights in (4.5) are distributed as follows:

$$\begin{aligned} 1/360\ (w_{1,1}, \ldots, w_{120,3}) &\sim \text{Dirichlet}(1, \ldots, 1), \\ 1/60\ (w_{1,1}^*, \ldots, w_{20,3}^*) &\sim \text{Dirichlet}(1, \ldots, 1), \\ 1/60\ (w_{1,1}^\star, \ldots, w_{20,3}^\star) &\sim \text{Dirichlet}(1, \ldots, 1). \end{aligned}$$

We consider only this weight configuration due to its superior performance in the normal

Method	Bayes	GLMM	H-likelihood	WHLB
β_0	$1.041_{0.468}$	$1.014_{0.509}$	$0.824_{0.360}$	$1.028_{0.325}$
β_1	$-3.089_{0.622}$	$-2.817_{0.562}$	$-2.411_{0.485}$	$-3.042_{0.446}$
β_2	$-0.715_{0.528}$	$-0.733_{0.561}$	$-0.568_{0.407}$	$-0.767_{0.392}$
β_3	$3.824_{0.652}$	$3.406_{0.572}$	$2.976_{0.518}$	$3.790_{0.483}$
σ_f	$1.306_{0.292}$	$1.131_{0.385}$	$1.131_{0.385}$	$1.306_{0.292}$
σ_m	$1.283_{0.277}$	$0.871_{0.418}$	$0.871_{0.418}$	$1.283_{0.277}$
v_1^f	$0.700_{0.900}$	$0.208_{1.134}$	$0.544_{0.775}$	$0.690_{0.653}$
v_1^m	$1.352_{0.977}$	$0.675_{1.298}$	$0.768_{0.698}$	$1.449_{0.857}$

Table 4.3: Parameter estimates for the Salamander dataset using 4 different methods. For Bayesian methods the posterior mean is shown, with posterior standard deviation as a subscript. For frequentist methods the maximum likelihood estimate is shown, with the standard error as a subscript.

means example.

The estimates for the parameter values under a Bayesian model with flat priors on all unknowns, the fitting routine of the `glmm` package in R based on maximising the marginal likelihood, h-likelihood and the WHLB are shown in Table 4.3. The Bayesian posterior is sampled from using Stan. The dispersion parameters for the h-likelihood were plugged in from the `glmm` fitting routine. The dispersion parameters for the WHLB were taken from the Bayes posterior. If we estimate the dispersion parameters by maximising weighted h-likelihood jointly with the fixed and random effects, we find these parameters are given negligible value. The fixed effects seem to be partly robust to this, but the random effects are not. Using the WHLB with the Bayes posterior mean dispersion parameters plugged in, all parameters compared are within one standard deviation of the Bayes posterior mean. This holds true if the `glmm` dispersion parameters are plugged in to the WHLB instead. So for this example it seems that the WHLB matches Bayes reasonably well, and although the WHLB needs the dispersion parameters to be provided separately, it does not seem overly sensitive to the values of the dispersion parameters used. A pairs plot for two fixed effects and a male and female random effect can be found in Figure 4.6, showing much similarity in the two posteriors. The WHLB posterior is slightly less dispersed than the

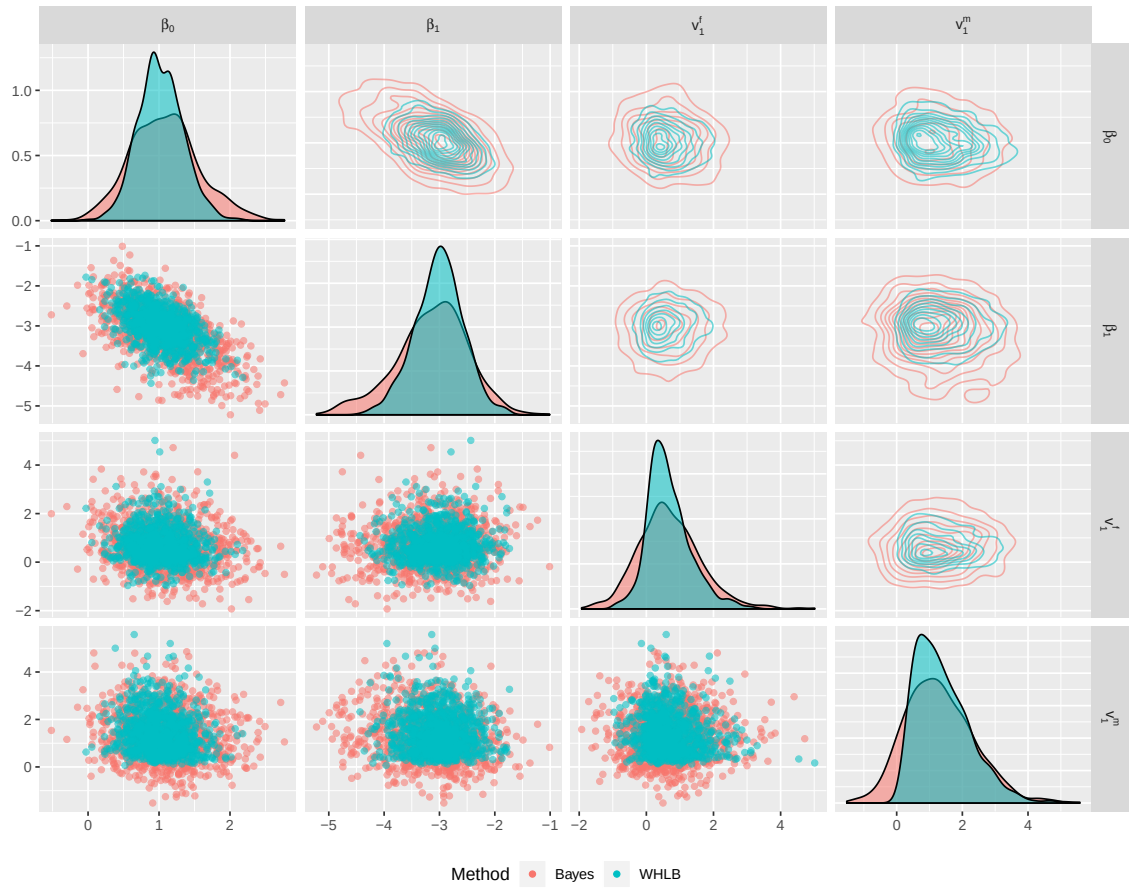


Figure 4.6: Pairs plots for the posterior distributions for two fixed effects and two random effects, for the Salamander dataset, for the Bayesian posterior and the WHLB posterior.

Bayesian posterior. In some respects this is unsurprising, given that the WHLB ignores any uncertainty in the dispersion parameters.

4.3.7 Discussion

The weighted h-likelihood bootstrap provides a novel means of generating general Bayesian posterior samples for statistical models with mixed effects, in a computationally efficient manner. However, the framework does inherit some of the less desirable traits of frequentist methods for fitting mixed models. Notably, the method seems to work best when the dispersion parameters are known, or plugged in, and as Bayesians we would ideally provide a posterior over these parameters as well. It would be interesting to develop a general

Bayesian bootstrap procedure for determining the dispersion parameters in future work.

The method can be seen as an extension to the notion of uncertainty quantification via randomising an objective function. However, the Bayesian nonparametric interpretation of the method is less clear than in the IID case. One problem here is the dual role that the 'unobservables' play, both as a type of observation, and as a more traditional type of parameter.

The WHLB method can be viewed as applying random weights to each term of a likelihood factorisation. Although this lends itself to further extension to inference for a general graphical model, different parametrisations can lead to different posterior distributions. It would be important in the future to develop some guidance in this regard, as well as further foundational motivation for any particular randomisation.

In this exploration we have said little about the problem of misspecification in mixed models. This topic is understandably more complex when unobserved random effects are involved, though has been studied in the literature in a frequentist setting. For example, recently, Yu et al. (2018) provide consistency results for fixed effect MHLEs under identifiability constraints. There seems to be have been little progress in terms of provable properties in Bayesian analyses of misspecified mixed models.

In Chapter 2 we extended the weighted likelihood bootstrap from log likelihoods to general loss functions, for IID data. Analogously, we can extend the weighted h-likelihood bootstrap to general Bayesian inference under a hierarchical loss function containing three types of argument: observation, unobservable random effects and any other fixed parameters. Suppose we have a hierarchical loss of the form

$$\ell_h(y, v, \beta) = \ell_y(y, v) + \ell_v(v, \beta)$$

where ℓ_y denotes the term for a dataset y equal to the sum of individual observation losses,

and the ℓ_v is equal to the sum of individual group losses,

$$\ell_y(y, v) = \sum_{i,j} \ell(y_{i,j}, v_i), \qquad \ell_v(v, \beta) = \sum_i \ell_{v_i}(v_i, \beta).$$

This loss is non-additive and the observations assumed non-IID. The WHLB could be applied by assigning random weights to each term in the loss, in analogy with (4.1).

As an example of the utility of this general approach, suppose we have grouped data, and are interested in both the group-level medians, and also the hierarchical median of the group-level medians. We argue that an appropriate loss in this case would be setting both ℓ_y and ℓ_v equal to the ℓ_1 loss. An appropriate scaling of one of the losses may be made, to represent the relative weight of the two objectives. Further investigations into the practical utility of such hierarchical losses should be undertaken in future work.

The use of the h-likelihood as an objective function can be motivated via the extended likelihood principle, though this principle itself is not accepted by all. To be able to argue that the h-likelihood is an appropriate objective function for general Bayesian inference under misspecification would provide an alternative justification for this concept.

Chapter 5

Discussion

5.1 Summary

In this thesis we have developed a framework for statistical learning for a general family of parameters that can include information from misspecified parametric models. The loss-likelihood bootstrap, the posterior bootstrap, the NPLasso and the weighted h-likelihood bootstrap are all examples of *general Bayesian bootstraps*, in that they are closely related to the Bayesian bootstrap (Rubin, 1981), but are general Bayesian methods in that they do not assume that we know the true model class. It is arguable that these general Bayesian bootstraps are more suitable to statistical learning under potential misspecification than if we used the parametric form of the general Bayesian posterior that is presented in Bissiri et al. (2016), and Section 1.4.1 of this thesis. The reasons in favour of general Bayesian bootstraps are that no calibration is required and they admit computationally scalable Bayesian inference routines, while still targeting the same broad family of parameters defined by a loss function. Also, the general Bayesian bootstrap applies a nonparametric prior to the distribution function of the data, which allows for a much richer set of inferences to be made than is possible under the approach of Bissiri et al. (2016). One advantage of the approach of Bissiri et al. (2016) is that it admits a targeted prior, defined only on the

parameter of interest, as opposed to over all unknowns in a statistical model. This may be valuable when a full probability model and associated prior information is expensive to build, but using the non-informative MDP with $c = 0$ is not considered appropriate.

In the rest of this chapter we highlight some thoughts about future work on this topic, split up into methodological, computational and applied aspects.

5.2 Future work

5.2.1 Methodology

In Section 3.6 the blurring of the boundaries between Bayesian and frequentist inference was discussed. The Bayesian bootstrap plays an important role here, due to its limiting Bayesian interpretation and its closeness to the nonparametric bootstrap. A frequentist should be comfortable to switch their multinomial nonparametric bootstrap weights for Dirichlet distributed ones; Rubin (1981) initially pointed out the closeness of these two methods in terms of first and second moments. The argument in favour of indifference about switching to Dirichlet weights was strengthened by Praestgaard and Wellner (1993) who showed that the Bayesian bootstrap has the same asymptotic properties as the nonparametric bootstrap. If this is accepted, then many frequentist (bootstrap) and noninformative Bayesian procedures coincide. If a subjective Bayesian wishes to incorporate prior information, there is an optional extension of this framework to cater for this. In a sense, the admission that all models are wrong helps bring the frequentist and Bayesian schools closer together.

One advantage of the general Bayesian approach is that it admits a prior belief distribution about the parameter of interest. In many settings the target parameter is a simpler and more natural object from which to elicit a prior, rather than a parameter indexing a potentially vast model class. The NP learning approach does admit a prior, but via a predictive, or pseudo-samples, not just for a targeted parameter defined by a loss. The extension of the posterior bootstrap to targeted priors would be valuable future work. This challenge has been

faced many times in the literature, including recently by Bornn et al. (2016), also Kessler et al. (2015). Both of these approaches require a nonparametric density estimation step, which would ideally be avoided, due to its poor properties in moderate and high dimensional problems. As pointed out in Section 2.9.4, we should expect such posterior uncertainty to be of a targeted nature, and we may lose some peripheral information that is not captured by the loss function.

Much of our work can be viewed as the melding of information from misspecified models with information from data. When we have no data, all we have to make an inference is the misspecified model and prior. As the amount of data grows, it is attractive to consider the discarding of the model, when the quantity of data allows it to speak for itself. For NP-learning with an MDP, as investigated in Chapter 3, the traversal between these two regimes is determined by the MDP. It is unclear if this approach is appropriate under very small sample sizes, for example if moving from zero to one observation. The Bayesian update is much smoother than the NP learning update in this case. Should we consider an alternative means of traversing between parametric model and data that is smoother, in a Bayesian parametric sense, than NP learning under an MDP? This would be an interesting topic for future work.

5.2.2 Implementation

From a computational perspective, there are several pieces of future work that could have a valuable impact. Mainly, a foundationally-grounded general posterior bootstrap recipe, for a general graphical model, would be a widely applicable procedure. In Chapter 4 we have made some progress on extending our NP learning framework to incorporate unobservables, but as discussed in Section 4.3.7, there is still much work to do in this regard. A Bayesian nonparametric interpretation of our methodology, and a notion of asymptotic optimality, as provided for the independent and identically distributed case in Chapter 3, would put this work on a much firmer footing.

If such a methodology was developed, the embedding of an NP learning diagnostic into a popular probabilistic programming language such as Stan (Carpenter et al., 2017) would provide practitioners with information about their sensitivity to misspecification, and more generally raise awareness in the problems associated with using convenient but misspecified models. One option for this would be the sensitivity analysis tool that was discussed in Section 3.8.5. There is scope to do similar for approximate posteriors such as variational Bayes, as well as exact posterior sampling routines such as MCMC.

For some applied problems the dataset is so large that it cannot fit on a single machine. Instead the data, and hence the computation, is distributed across a number of worker machines. This issue was not addressed in our NP learning development in this thesis, but is an important problem. The extension of nonparametric learning to distributed data settings would be a valuable avenue of future work. The bag of little bootstraps (Kleiner et al., 2014) would be one idea for scaling up the posterior bootstrap in this regard. Topics such as communication-efficient statistical learning (Jordan et al., 2018) as well as distributed optimisation routines may also be useful avenues of enquiry.

5.2.3 Applications

5.2.3.1 The general supra-Bayesian

In the process of devising ways to combine prior information with data using bootstraps, one idea we considered was to treat this as a problem of combining beliefs from a number of experts. This problem has been studied in some depth in the literature; see Genest and Zidek (1986) for a review. Perhaps the best-motivated approach is the 'supraBayesian': logarithmic pooling of belief distributions ensures the operations of combining the belief distributions, and updating them given data, is commutative. It would be interesting to consider extensions of NP learning to the setting where we have M experts, each with a belief distribution, and/or a predictive, and the goal is to optimally combine the belief distributions. As discussed in Section 3.8.6, one means of achieving this is via pseudo data,

which could also allow for privacy-preserving group learning routines.

5.2.3.2 Randomising objective functions

There is scope for NP learning to be applied to a range of machine learning applications. Although often excellent predictively, these algorithms are often hard to interpret and may provide little uncertainty quantification. It would be interesting to see what practical benefits there are from applying NP learning to deep learning. Active learning problems seem to fit naturally into the general Bayesian framework, due to their often targeted nature. Extending general Bayesian updating to such problems would be a valuable extension, and could lead to the application of such ideas to reinforcement learning problems.

References

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics* 16(1):3–14.

Alquier, P., Ridgway, J., and Chopin, N. (2016). On the properties of variational approximations of Gibbs posteriors. *Journal of Machine Learning Research* 17:1–41.

Anscombe, F. J. and Tukey, J. W. (1963). The Examination and Analysis of Residuals. *Technometrics* 5(2):141–160.

Antoniak, C. E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics* 2(6):1152–1174.

Bach, F. R. (2008). Bolasso: Model Consistent Lasso Estimation through the Bootstrap. *International Conference on Machine Learning*.

Bardenet, R., Doucet, A., and Holmes, C. C. (2017). On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research* 18:1–43.

Barnard, G. A. (1949). Statistical Inference. *Journal of the Royal Statistical Society. Series B (Methodological)* 11(2):115–149.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. (2006). Convexity, Classification, and Risk Bounds. *Journal of the American Statistical Association* 101(473):138–156.

Bayes, T. (1763). An Essay towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society of London* 53:370–418.

Berger, J. O. (1994). An overview of robust Bayesian analysis. *Test* 3(1):5–124.

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle: A review, generalizations, and statistical implications*. Institute of Mathematical Statistics.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics* 41(2):802–837.

Berk, R. H. (1966). Limiting Behavior of Posterior Distributions when the Model is Incorrect. *The Annals of Mathematical Statistics* 37(1):51–58.

Bernardo, J. M. and Smith, A. F. M. (2006). *Bayesian Theory*. Chichester: Wiley.

Bickel, P. J. and Freedman, D. A. (1981). Some Asymptotic Theory for the Bootstrap. *The Annals of Statistics* 9(6):1196–1217.

Birnbaum, A. (1962). On the Foundations of Statistical Inference. *The Annals of Mathematical Statistics* 32(2):414–435.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer.

Bissiri, P. G., Holmes, C. C., and 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.

Bissiri, P. G. and Walker, S. G. (2012). Converting information into probability measures with the Kullback-Leibler divergence. *Annals of the Institute of Statistical Mathematics* 64(6):1139–1160.

Bjørnstad, J. F. (1996). On the Generalization of the Likelihood Function and the Likelihood Principle. *Journal of the American Statistical Association* 91(434):791–806.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association* 112(518):859–877.

Boos, D. D. and Monahan, J. F. (1986). Bootstrap Methods Using Prior Information. *Biometrika* 73(1):77–83.

Bornn, L., Shephard, N., and Solgi, R. (2016). Moment conditions and Bayesian nonparametrics. arXiv: 1507.08645.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Wiley Classics Library. Wiley.

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

Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building. *Army Research Office Workshop on Robustness in Statistics*:1–36.

Box, G. E. P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Journal of the Royal Statistical Society. Series B (Methodological)*.

Breiman, L. (1996). Bagging predictors. *Machine Learning* 24(2):123–140.

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

Breiman, L. (2001b). Statistical Modeling: The Two Cultures. *Statistical Science* 16(3):199–231.

Breslow, N. E. and Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* 88(421):9–25.

Brigo, D., Pallavicini, A., and Torresetti, R. (2010). *Credit Models and the Crisis: A Journey into CDOs, Copulas, Correlations and Dynamic Models*. John Wiley & Sons.

Bunke, O. and Milhaud, X. (1998). Asymptotic Behavior of Bayes Estimates under Possibly Incorrect Models. *The Annals of Statistics* 26(2):617–644.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag New York.

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

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* 97(2):465–480.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling Sparsity via the Horseshoe. *International Conference on Artificial Intelligence and Statistics* 5:73–80.

Chamberlain, G. and Imbens, G. W. (2003). Nonparametric Applications of Bayesian Inference. *Journal of Business & Economic Statistics* 21(1):12–18.

Chatfield, C. (1995). Model Uncertainty , Data Mining and Statistical Inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 158(3):419–466.

Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping lasso estimators. *Journal of the American Statistical Association* 106(494):608–625.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics* 115(2):293–346.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). Bayesian CART Model Search. *Journal of the American Statistical Association* 93(443):935–948.

Clyde, M. A. and Iversen, E. S. (2013). Bayesian model averaging in the M-open framework. *Bayesian Theory and Applications*. Ed. by P. Damien, P. Dellaportas, N. G. Polson, and D. A. Stephens. Oxford University Press, 484–498.

Clyde, M. A. and Lee, H. K. H. (2001). Bagging and the Bayesian bootstrap. *Artificial Intelligence and Statistics* 2001:169–174.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning* 20(3):273–297.

Cox, D. R. (1961). Tests of Separate Families of Hypotheses. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley, Calif.: University of California Press, pp. 105–123.

Cox, D. R. and Hinkley, D. V. (1979). *Theoretical Statistics*. Taylor & Francis.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Davison, A. C., Hinkley, D. V., and Worton, B. J. (1992). Bootstrap likelihoods. *Biometrika* 79:113–130.

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.

De Blasi, P. and Walker, S. G. (2013). Bayesian asymptotics with misspecified models. *Statistica Sinica* 23:169–187.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). A Bayesian CART Algorithm. *Biometrika* 85(2):363–377.

Dheeru, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository.

Diebold, F. X. (1991). On Bayesian Forecast Combination Procedures. *Economic Structural Change: Analysis and Forecasting*. Springer. Chap. 15, 225–232.

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

Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. *Journal of the American Statistical Association* 78(382):316–331.

Efron, B. and Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge University Press.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics* 32(2):407–499.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability* 1:59–82.

Ferentinos, K. and Papaioannou, T. (1981). New parametric measures of information. *Information and Control* 51(3):193–208.

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics* 1(2):209–230.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research* 15:3133–3181.

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society A* 222:309–368.

Fisher, R. A. (1925). Theory of Statistical Estimation. *Proceedings of the Cambridge Philosophical Society* 22:700–725.

Fisher, R. A. (1934). Two New Properties of Mathematical Likelihood. *Philosophical Transactions of the Royal Society A* 144:285–307.

Fisher, R. A. (1990). *Statistical Methods, Experimental Design, and Scientific Inference*. Oxford science publications. Oxford University Press.

Foutz, R. V. (1977). On the Unique Consistent Solution to the Likelihood Equations. *Journal of the American Statistical Association* 72(357):147.

Freund, Y. and Schapire, R. E. (1999). A brief introduction to boosting. *IJCAI International Joint Conference on Artificial Intelligence* 2(5):1401–1406.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics* 1(2):302–332.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33(1).

Fushiki, T. (2005). Bootstrap prediction and Bayesian prediction under misspecified models. *Bernoulli* 11(4):747–758.

Fushiki, T. (2010). Bayesian bootstrap prediction. *Journal of Statistical Planning and Inference* 140(1):65–74.

Gasparini, M. (1995). Exact Multivariate Bayesian Bootstrap Distributions of Moments. *The Annals of Statistics* 23(3):762–768.

Geisser, S. and Eddy, W. F. (1979). A Predictive Approach to Model Selection. *Journal of the American Statistical Association* 74(365):153–160.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* 85(410):398–409.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica* 6:733–807.

Genest, C. and Zidek, J. V. (1986). Combining Probability Distributions: A Critique and an Annotated Bibliography. *Statistical Science* 1(1):114–148.

Ghosal, S. and Vaart, A. van der (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Goldstein, M. (1981). Revising Previsions : A Geometric Interpretation. *Journal of the Royal Statistical Society. Series B (Methodological)* 43(2):105–130.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Adaptive computation and machine learning. MIT Press.

Gourieroux, C., Monfort, A., and Trognon, A. (1984). Pseudo Maximum Likelihood Methods: Theory. *Econometrica* 52(3):681–700.

Gramacy, R. B. and Pantaleo, E. (2010). Shrinkage Regression for Multivariate Inference with Missing Data, and an Application to Portfolio Balancing. *Bayesian Analysis* 5(2):237–262.

Green, P. J., Latuszy, K., Pereyra, M., and Robert, C. P. (2015). Bayesian computation: a perspective on the current state, and sampling backwards and forwards. *Statistics and Computing* 25(4):835–862.

Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* 5(1):171–188.

Grünwald, P. (1998). The Minimum Description Length Principle and Reasoning under Uncertainty. PhD thesis, pp. 1–302.

Grünwald, P. (1999). Viewing all models as "probabilistic". *Proceedings of the twelfth annual conference on Computational learning theory*, pp. 171–182.

Grünwald, P. (2011). Safe Learning: bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity. *Journal of Machine Learning Research: Workshop and Conference Proceedings* 19:397–419.

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

Hanin, L. (2017). Why statistical inference from clinical trials is likely to generate false and irreproducible results. *BMC Medical Research Methodology* 17(1):1–12.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Vol. 1. Springer Series in Statistics. Springer New York, pp. 337–387.

Healy, M. J. R. (1978). Is Statistics a Science? *Journal of the Royal Statistical Society. Series A (General)* 141(3):385–393.

Henderson, C. R. (1953). Estimation of Variance and Covariance Components. *Biometrics* 9(2):226–252.

Hjort, N. L., Holmes, C. C., Müller, P., and Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression : Biased Estimation for Nonorthogonal Problems. *Technometrics* 12(1):55–67.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science* 14(4):382–417.

Holmes, C. C. and Walker, S. G. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika* 104(2):497–503.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, Calif.: University of California Press, pp. 221–233.

Huber, P. J. (2009). *Robust Statistics*. Wiley.

Jaakkola, T. S. and Jordan, M. I. (1997). A variational approach to Bayesian logistic regression models and their extensions. *International Workshop on Artificial Intelligence and Statistics*.

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J. M., and Thiébaut, R. (2007). Robustness of the linear mixed model to misspecified error distribution. *Computational Statistics and Data Analysis* 51(10):5142–5154.

Jeffreys, H. (1939). *The Theory of Probability*. Oxford Classic Texts in the Physical Sciences. Oxford University Press.

Jiang, W. and Tanner, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics* 36(5):2207–2231.

Jordan, M. I., Lee, J. D., and Yang, Y. (2018). Communication-Efficient Distributed Statistical Inference. *Journal of the American Statistical Association*:1–14.

Kass, R. E. and Wasserman, L. (1994). Formal Rules for Selecting Prior Distributions: A Review and Annotated Bibliography:1–81.

Kessler, D. C., Hoff, P. D., and Dunson, D. B. (2015). Marginally specified priors for non-parametric Bayesian estimation. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 77(1):35–58.

Key, J. T., Pericchi, L. R., and Smith, A. F. M. (1999). Bayesian Model Choice: What and Why? *Bayesian Statistics 6*. Ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith:343–370.

Kleijn, B. J. K. and Vaart, A. W. van der (2012). The Bernstein-Von-Mises theorem under misspecification. *Electronic Journal of Statistics* 6:354–381.

Kleijn, B. J. K. and Van Der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics* 34(2):837–877.

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 76:795–816. eprint: `1112.5016`.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems 25*, pp. 1097–1105.

Kucukelbir, A., Ranganath, R., Gelman, A., and Blei, D. M. (2015). Automatic Variational Inference in Stan. *Advances in Neural Information Processing Systems 28*. Vol. 1, pp. 1–9.

Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics* 22(1):79–86.

Lazar, N. A. (2003). Bayesian empirical likelihood. *Biometrika* 90(2):319–326.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521(7553):436–444.

Lee, Y. and Nelder, J. A. (1996). Hierarchical Generalized Linear Models. *Journal of the Royal Statistical Society. Series B (Methodological)* 58(4):619–678.

Lee, Y., Nelder, J. A., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. CRC Press.

Lee, Y. and Nelder, J. A. (2005). Likelihood for random-effect models. *SORT* 29(2):141–164.

Lee, Y., Nelder, J. A., and Noh, M. (2007). H-likelihood: Problems and solutions. *Statistics and Computing* 17(1):49–55.

Lehmann, E. L. (1990). Model Specification: The Views of Fisher and Neyman, and Later Developments. *Statistical Science* 5(2):160–168.

Li, C., Jiang, W., and Tanner, M. A. (2013). General oracle inequalities for Gibbs posterior with application to ranking. *Journal of Machine Learning Research: Workshop and Conference Proceedings* 30:1–10.

Li, M. and Dunson, D. B. (2016). A framework for probabilistic inferences from imperfect models. arXiv: `1611.01241v1`.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika* 73(1):13–22.

Limpert, E. and Stahel, W. A. (2011). Problems with using the normal distribution - and ways to improve quality and efficiency of data analysis. *PLoS ONE* 6(7).

Lindley, D. V. (1956). On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics* 27(4):986–1005.

Lindley, D. V. (2000). The philosophy of statistics. *Journal of the Royal Statistical Society. Series D (The Statistician)* 49(3):293–337.

Lo, A. Y. (1984). On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. *The Annals of Statistics* 12(1):351–357.

Lo, A. Y. (1987). A large sample study of the Bayesian Bootstrap. *The Annals of Statistics* 15(1):360–375.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics* 42(2):413–468.

Lunn, D. J., Thomas, A., Best, N. G., and Spiegelhalter, D. (2000). WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* 10:325–337.

Lyddon, S. P., Holmes, C. C., and Walker, S. G. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika* 106(2):465–478.

Lyddon, S. P., Walker, S. G., and Holmes, C. C. (2018). Nonparametric learning from Bayesian models with randomized objective functions. *Advances in Neural Information Processing Systems 31*, pp. 2071–2081.

Madigan, D. and Raftery, A. E. (1994). Model Selection and Accounting in Graphical Models for Model Uncertainty Using Occam's Window. *Journal of the American Statistical Association* 89(428):1535–1546.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing* 22:1167–1180.

Marx, V. (2013). The big challenges of big data. *Nature* 498(7453):255–260.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, p. 532.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* 34(3):1436–1462.

Mengersen, K. L., Pudlo, P., and Robert, C. P. (2013). Bayesian computation via empirical likelihood. *Proceedings of the National Academy of Sciences of the United States of America* 110(4):1321–1326.

Miller, J. W. and Dunson, D. B. (2018). Robust Bayesian Inference via Coarsening. *Journal of the American Statistical Association*.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association* 83(401):1023–1032.

Müller, U. K. (2013). Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix. *Econometrica* 81(5):1805–1849.

Neiswanger, W., Wang, C., and Xing, E. (2014). Asymptotically Exact, Embarrassingly Parallel MCMC. *Uncertainty in Artificial Intelligence (UAI)*, pp. 623–632.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)* 56(1):3–48.

Newton, M., Polson, N. G., and Xu, J. (2018). Weighted Bayesian Bootstrap for Scalable Bayes. *Technical report*:1–16. arXiv: 1803.04559.

Neyman, J. and Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 231(694-706):289–337.

O'Hagan, A. (2013). Bayesian inference with misspecified models: Inference about what? *Journal of Statistical Planning and Inference* 143:1643–1648.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC.

Pace, L. and Salvan, A. (1997). *Principles of Statistical Inference: From a Neo-Fisherian Perspective*. Advanced Series on Statistical Science & Applied Probability. World Scientific Publishing Company.

Pamies, P. (2018). Towards trustable machine learning. *Nature Biomedical Engineering* 2(10):709–710.

Park, M. Y. and Hastie, T. (2007). L_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 69(4):659–677.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* 103(482):681–686.

Patterson, H. D. and Thompson, R. (1971). Recovery of Inter-Block Information when Block Sizes are Unequal. *Biometrika* 58(3):545–554.

Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Polson, N. G. (2017). Deep Learning: A Bayesian Perspective. *Bayesian Analysis* 12(4):1275–1304.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables. *Journal of the American Statistical Association* 108(504):1339–1349.

Polson, N. and Scott, J. (2018). *AIQ: How People and Machines Are Smarter Together*. St. Martin's Press, Inc.

Praestgaard, J. and Wellner, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *The Annals of Probability* 21(4):2053–2086.

Robert, C. P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer New York.

Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer New York.

Robinson, G. K. (1991). The BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science* 6(1):15–51.

Royall, R. and Tsou, T. S. (2003). Interpreting statistical evidence by using imperfect models: Robust adjusted likelihood functions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 65(2):391–404.

Rubin, D. B. (1981). The Bayesian Bootstrap. *The Annals of Statistics* 9(1):130–134.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 71(2):319–392.

Savage, L. J. (1954). *The Foundations of Statistics*. Wiley.

Schall, R. (1991). Estimation in generalized linear models with random effeects. *Biometrika* 78(4):719–727.

Schweder, T. and Hjort, N. L. (2016). *Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Scornet, E., Biau, G., and Vert, J. P. (2015). Consistency of random forests. *The Annals of Statistics* 43(4):1716–1741.

Scott, C. (2012). Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics* 6:958–992.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. (2016). Bayes and big data: The consensus monte carlo algorithm. *International Journal of Management Science and Engineering Management* 11.

Sethuraman, J. (1994). A constructive definition of Dirichlet Process Priors. *Statistica Sinica* 4:639–650.

Shalizi, C. R. (2009). Dynamics of bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics* 3:1039–1074.

Shawe-Taylor, J. and Williamson, R. (1997). A PAC analysis of a Bayesian estimator. *Proceedings of the Tenth Annual Conference on Computational Learning Theory*.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* 90(2):227–244.

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian Computation Via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods Author. *Journal of the Royal Statistical Society. Series B (Methodological)* 55(1):3–23.

Syring, N. and Martin, R. (2018). Calibrating general posterior credible regions. *Biometrika (to appear)*. arXiv: 1509.00922.

Syring, N. and Martin, R. (2017). Gibbs posterior inference on the minimum clinically important difference. *Journal of Statistical Planning and Inference* 187:67–77.

Taddy, M., Chen, C.-S., Yu, J., and Wyle, M. (2015). Bayesian and Empirical Bayesian Forests. *Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015.* Vol. 37.

Tibshirani, R. (1996). Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 58(1):267–288.

Tierney, L. (1994). Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics* 22(4):1701–1762.

Tipping, M. (2001). Sparse Bayesian Learning and the Relevance Vector Mach. *Journal of Machine Learning Research* 1:211–244.

Turnbull, C. et al. (2018). The 100 000 Genomes Project: Bringing whole genome sequencing to the NHS. *BMJ (Online)* 361(April):1–7.

Van Der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Vapnik, V. (2013). *The Nature of Statistical Learning Theory*. Information Science and Statistics. Springer New York.

Venter, J. C. et al. (2001). The Sequence of the Human Genome. *Science* 291(5507):1304–1351.

Wald, A. (1949). Statistical Decision Functions. *The Annals of Mathematical Statistics* 20(2):165–205.

Walker, S. G. (2013). Bayesian inference with misspecified models. *Journal of Statistical Planning and Inference* 143(10):1621–1633.

Walker, S. G. (2016). Bayesian information in an experiment and the Fisher information distance. *Statistics and Probability Letters* 112:5–9.

Walker, S. G. and Gutierrez-Pena, E. (1999). Robustifying Bayesian Procedures. *Bayesian Statistics 6*, 685–710.

Walker, S. G. and Hjort, N. L. (2001). On Bayesian Consistency. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 63(4):811–821.

Wang, X. and Dunson, D. B. (2014). Parallelizing MCMC via Weierstrass Sampler:1–35. arXiv: `1312.4605v2`.

Wang, Y., Kucukelbir, A., and Blei, D. M. (2017). Robust Probabilistic Modeling with Bayesian Data Reweighting. *Proceedings of the 34th International Conference on Machine Learning*, pp. 1–12.

Welch, B. L. and Peers, H. W. (1963). On Formulae for Confidence Points Based on Integrals of Weighted Likelihoods. *Journal of the Royal Statistical Society. Series B (Methodological)* 25(2):318–329.

Weng, C.-S. (1989). On a Second-Order Asymptotic Property of the Bayesian Bootstrap Mean. *The Annals of Statistics* 17(2):705–710.

West, M. (1984). Outlier Models and Prior Distributions in Bayesian Linear Regression. *Journal of the Royal Statistical Society. Series B (Methodological)* 46(3):431–439.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* 48(4):817–838.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10(3):515–534.

Wolpert, D. H. (1992). Stacked Generalization. *Neural networks* 5(2):241–259.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* 466(7310):1102–1104.

Yu, D., Zhang, X., and Yau, K. K. W. (2018). Asymptotic properties and information criteria for misspecified generalized linear mixed models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 80(4):817–836.

Zellner, A. (1988). Optimal Information Processing and Bayes's Theorem. *The American Statistician* 42(4):278–280.

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. *Proceedings of the 21st International Conference on Machine Learning*.

Zhang, T. (2006). From epsilon-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *The Annals of Statistics* 34(5):2180–2210.