Nick Heard

An Introduction to Bayesian Inference, Methods and Computation

Springer

An Introduction to Bayesian Inference, Methods and Computation

Nick Heard

An Introduction to Bayesian Inference, Methods and Computation

Springer

Nick Heard
Imperial College London
London, UK

ISBN 978-3-030-82807-3 ISBN 978-3-030-82808-0 (eBook)
https://doi.org/10.1007/978-3-030-82808-0

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2021, corrected publication 2022
This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

The aim of writing this text was to provide a fast, accessible introduction to Bayesian statistical inference. The content is directed at postgraduate students with a background in a numerate discipline, including some experience in basic probability theory and statistical estimation. The text accompanies a module of the same name, *Bayesian Methods and Computation*, which forms part of the online *Master of Machine Learning and Data Science* degree programme at Imperial College London.

Starting from an introduction to the fundamentals of subjective probability, the course quickly advances to modelling principles, computational approaches and then advanced modelling techniques. Whilst this rapid development necessitates a light treatment of some advanced theoretical concepts, the benefit is to fast track the reader to an exciting wealth of modelling possibilities whilst still providing a key grounding in the fundamental principles.

To make possible this rapid transition from basic principles to advanced modelling, the text makes extensive use of the probabilistic programming language *Stan*, which is the product of a worldwide initiative to make Bayesian inference on user-defined statistical models more accessible. Stan is written in C++, meaning it is computationally fast and can be run in parallel, but the interface is modular and simple. The future of applied Bayesian inference arguably relies on the broadening development of such software platforms.

Chapter 1 introduces the core ideas of Bayesian reasoning: Decision-making under uncertainty, specifying subjective probabilities and utility functions and identifying optimal decisions as those which maximise expected utility. Prediction and estimation, the two core tasks in statistical inference, are shown to be special cases of this broader decision-making framework. The application-driven reader may choose to skip this chapter, although philosophically it sets the foundation for everything that follows.

Chapter 2 presents representation theorems which justify the prior $\times$ likelihood formulation synonymous with Bayesian probability models. Simply believing that unknown variables are exchangeable, meaning probability beliefs are invariant to relabelling of the variables, is sufficient to guarantee that construction must hold. The prior distribution distinguishes Bayesian inference from frequentist statistical

methods, and several approaches to specifying prior distributions are discussed. The prior and likelihood construction leads naturally to consideration of the posterior distribution, including useful results on asymptotic consistency and normality which suggest large sample robustness to the choice of prior distribution.

Chapter 3 shows how graphical models can be used to specify dependencies in probability distributions. Graphical representations are most useful when the dependency structure is a primary target of inferential interest. Different types of graphical model are introduced, including belief networks and Markov networks, highlighting that the same graph structure can have different interpretations for different models.

Chapter 4 discusses parametric statistical models. Attention is focused on conjugate models, which present the most mathematically convenient parametric approximations of true, but possibly hard to specify, underlying beliefs. Although these models might appear relatively simplistic, later chapters will show how these basic models can form the basis of very flexible modelling frameworks.

Chapter 5 introduces the computational techniques which revolutionised the application of Bayesian statistical modelling, enabling routine performance of inferential tasks which had previously appeared infeasible. Relatively simple Markov chain Monte Carlo methods were at the heart of this development, and these are explained in some detail. A higher level description of Hamiltonian Monte Carlo methods is also provided, since these methods are becoming increasingly popular for performing simulation-based computational inference more efficiently. For high-dimensional inference problems, some useful analytic approximations are presented which sacrifice the theoretical accuracy of Monte Carlo methods for computational speed.

Chapter 6 discusses probabilistic programming languages specifically designed for easing some of the complexities of implementing Bayesian inferential methods. Particular attention is given to Stan, which has experienced rapid growth in deployment. Stan automates parallel Hamiltonian Monte Carlo sampling for statistical inference on any suitably specified Bayesian model on a continuous parameter space. In the subsequent chapters which introduce more advanced statistical models, Stan is used for demonstration wherever possible.

Chapter 7 is concerned with model checking. There are no expectations for subjective probability models to be correct, but it can still be useful to consider how well observed data appear to fit with an assumed model before making any further predictions using the same model assumptions; it may make sense to reconsider alternatives. Posterior predictive checking provides one framework for model checking in the Bayesian framework, and its application is easily demonstrated in Stan. For comparing rival models, Bayes factors are shown to be a well-calibrated statistic for quantifying evidence in favour for one model or the other, providing a vital Bayesian analogue to Neyman-Pearson likelihood ratios.

Chapter 8 presents the Bayesian linear model as the cornerstone of regression modelling. Extensions from the standard linear model to other basis functions such as polynomial and spline regression highlight the flexibility of this fundamental model structure. Further extensions to generalised linear models, such as logistic and Poisson regression, are demonstrated through implementation in Stan.

Chapter 9 characterises nonparametric models as more flexible parametric models with a potentially infinite number of parameters, distributing probability mass across larger function spaces. Dirichlet process and Polya tree models are presented as respective nonparametric models for discrete and continuous random probability measures. Partition models such as Bayesian histograms are also included in this class of models.

Chapter 10 covers nonparametric regression. Particular attention is given to Gaussian processes, which can be regarded as generalisations of the Bayes linear model. Spline models and partition models are also re-examined in this context.

Chapter 11 combines clustering and latent factor models. Both classes of model assume a latent underlying structure, which is either discrete or continuous, respectively. Finite and infinite mixture models are considered for clustering data into homogeneous groupings. Topic modelling of text and other unstructured data is considered as both a finite and infinite mixture problem. Finally, continuous latent factor models are presented as an extension of linear regression modelling, through the inclusion of unobserved covariates. Again, example Stan code is used to illustrate this class of models.

Throughout the text, there are exercises which should form an important component of following this course. Exercises which require access to a computer are indicated with a 💻 symbol; these become increasingly prevalent as the chapters progress, reflecting the transition within the text from laying fundamental principles to applied practice.

London, UK Nick Heard
June 2021

The original version of the book was revised. Copyright page text has been updated. The correction to the book is available at https://doi.org/10.1007/978-3-030-82808-0_12.

Contents

1 Uncertainty and Decisions 1
1.1 Subjective Uncertainty and Possibilities 1
1.1.1 Subjectivism 1
1.1.2 Subjective Uncertainty 2
1.1.3 Possible Outcomes and Events 2
1.2 Decisions: Actions, Outcomes, Consequences 3
1.2.1 Elements of a Decision Problem 3
1.2.2 Preferences on Actions 3
1.3 Subjective Probability 5
1.3.1 Standard Events 5
1.3.2 Equivalent Standard Events 6
1.3.3 Definition of Subjective Probability 6
1.3.4 Contrast with Frequentist Probability 7
1.3.5 Conditional Probability 7
1.3.6 Updating Beliefs: Bayes Theorem 8
1.4 Utility 8
1.4.1 Principle of Maximising Expected Utility 9
1.4.2 Utilities for Bounded Decision Problems 10
1.4.3 Utilities for Unbounded Decision Problems 10
1.4.4 Randomised Strategies 11
1.4.5 Conditional Probability as a Consequence of Coherence 11
1.5 Estimation and Prediction 12
1.5.1 Continuous Random Variables and Decision Spaces 12
1.5.2 Estimation and Loss Functions 12
1.5.3 Prediction 13

2 Prior and Likelihood Representation 15
2.1 Exchangeability and Infinite Exchangeability 15
2.2 De Finetti's Representation Theorem 16

2.3 Prior, Likelihood and Posterior . . . 18
2.3.1 Prior Elicitation . . . 18
2.3.2 Non-informative Priors . . . 18
2.3.3 Hyperpriors . . . 19
2.3.4 Mixture Priors . . . 19
2.3.5 Bayesian Paradigm for Prior to Posterior Reporting . . . 20
2.3.6 Asymptotic Consistency . . . 20
2.3.7 Asymptotic Normality . . . 21

3 Graphical Modelling and Hierarchical Models . . . 23
3.1 Graphs . . . 23
3.1.1 Specifying a Graph . . . 23
3.1.2 Neighbourhoods of Graph Nodes . . . 24
3.1.3 Paths, Cycles and Directed Acyclic Graphs . . . 25
3.1.4 Cliques and Separation . . . 25
3.2 Graphical Models . . . 26
3.2.1 Belief Networks . . . 26
3.2.2 Markov Networks . . . 27
3.2.3 Factor Graphs . . . 29
3.3 Hierarchical Models . . . 30

4 Parametric Models . . . 33
4.1 Parametric Modelling . . . 33
4.2 Conjugate Models . . . 34
4.3 Exponential Families . . . 36
4.4 Non-conjugate Models . . . 36
4.5 Posterior Summaries for Parametric Models . . . 37
4.5.1 Marginal Distributions . . . 37
4.5.2 Credible Regions . . . 38

5 Computational Inference . . . 39
5.1 Intractable Integrals in Bayesian Inference . . . 39
5.2 Monte Carlo Estimation . . . 40
5.2.1 Standard Error . . . 41
5.2.2 Estimation Under a Loss Function . . . 41
5.2.3 Importance Sampling . . . 42
5.2.4 Normalising Constant Estimation . . . 43
5.3 Markov Chain Monte Carlo . . . 44
5.3.1 Technical Requirements of Markov Chains in MCMC . . . 44
5.3.2 Gibbs Sampling . . . 46
5.3.3 Metropolis-Hastings Algorithm . . . 48
5.4 Hamiltonian Markov Chain Monte Carlo . . . 50
5.5 Analytic Approximations . . . 52
5.5.1 Normal Approximation . . . 52
5.5.2 Laplace Approximations . . . 53

5.5.3 Variational Inference . . . 55
5.6 Further Topics . . . 60

6 Bayesian Software Packages . . . 61
6.1 Illustrative Statistical Model . . . 61
6.2 Stan . . . 62
6.2.1 PyStan . . . 63
6.3 Other Software Libraries . . . 65
6.3.1 PyMC . . . 65
6.3.2 Edward . . . 65

7 Criticism and Model Choice . . . 67
7.1 Model Uncertainty . . . 68
7.2 Model Averaging . . . 68
7.3 Model Selection . . . 69
7.3.1 Selecting From a Set of Models . . . 69
7.3.2 Pairwise Comparisons: Bayes Factors . . . 70
7.3.3 Bayesian Information Criterion . . . 72
7.4 Posterior Predictive Checking . . . 73
7.4.1 Posterior Predictive p-Values . . . 73
7.4.2 Monte Carlo Estimation . . . 74
7.4.3 PPC with Stan . . . 74

8 Linear Models . . . 79
8.1 Parametric Regression . . . 79
8.2 Bayes Linear Model . . . 80
8.2.1 Conjugate Prior . . . 81
8.2.2 Reference Prior . . . 85
8.3 Generalisation of the Linear Model . . . 86
8.3.1 General Basis Functions . . . 86
8.4 Generalised Linear Models . . . 87
8.4.1 Poisson Regression . . . 88
8.4.2 Logistic regression . . . 90

9 Nonparametric Models . . . 93
9.1 Random Probability Measures . . . 94
9.2 Dirichlet Processes . . . 94
9.2.1 Discrete Base Measure . . . 97
9.3 Polya Trees . . . 98
9.3.1 Continuous Random Measures . . . 101
9.4 Partition Models . . . 102
9.4.1 Partition Models: Bayesian Histograms . . . 102
9.4.2 Bayesian Histograms with Equal Bin Widths . . . 104

10 Nonparametric Regression 107
10.1 Nonparametric Regression Modelling 107
10.2 Gaussian Processes 108
10.2.1 Normal Errors 109
10.2.2 Inference 110
10.3 Spline Models 113
10.3.1 Spline Regression with Equally Spaced Knots 114
10.4 Partition Regression Models 116
10.4.1 Changepoint Models 117
10.4.2 Classification and Regression Trees 119

11 Clustering and Latent Factor Models 121
11.1 Mixture Models 121
11.1.1 Finite Mixture Models 122
11.1.2 Dirichlet Process Mixture Models 126
11.2 Mixed-Membership Models 128
11.2.1 Latent Dirichlet Allocation 129
11.2.2 Hierarchical Dirichlet Processes 131
11.3 Latent Factor Models 133
11.3.1 Stan Implementation 134

Correction to: An Introduction to Bayesian Inference, Methods and Computation C1

Appendix A: Conjugate Parametric Models 137

Appendix B: Solutions to Exercises 141

Glossary 163

References 165

Index 167