The θ-augmented Model for Bayesian Semiparametric Inference on Functional Parameters

Vivian Meng, Department of Mathematics and Statistics
McGill University, Montreal

March 2022

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of

Doctor of Philosophy in Mathematics and Statistics

Abstract

In this thesis we develop the θ-augmented Bayes (TAB) method, which refers to Bayesian inference via the θ-augmented model, as a way of Bayesian semiparametric inference on functional parameters. It is a method for inferring functionals of the sampling distribution for the observable, without requiring the target parameter to be a part of the likelihood specification. We first define the θ-augmented model which augments a "proposal" nonparametric model with the target parameter θ by 1) partitioning the proposal model space according to the contours of θ and 2) modifying the corresponding probability measure via a re-weighting function, in the fashion of change of measure/importance sampling. This allows us to control the marginal prior distribution for θ while maintaining a fully nonparametric model space. We show asymptotic consistency of TAB posterior inference for functionals defined via estimating equations when the proposal model is nonparametric and weakly consistent for the data-generating mechanism. This thesis also documents some recommendations with regards to algorithms for sampling from the TAB posterior, and suitable proposal models. In general, the θ-augmented model is most useful when the proposal model is taken to be the Dirichlet process (DP), due to an ease of implementation for most functional parameters. Given suitable hyperparameters for the DP proposal model, the TAB posterior approximately extends the Bayesian bootstrap (BB) with a subjective prior. This behaviour is particularly useful as the BB exhibits good asymptotic properties. Through simulation, we show that the TAB posterior has good Frequentist properties in small sample inference, and performs well among competitors.

Abrégé

Dans cette thèse, nous développons la méthode θ- augmentée de Bayes (TAB), qui réfère à l'inférence bayésienne à travers la méthode θ-augmentée comme moyen d'inférence semi-paramétrique bayésienne. Cette méthode est utilisée pour l'inférence de fonctions des distributions d'échantillonnage pour les observables sans restreindre le paramètre d'intérêt aux specifications fondées sur la vraisemblance. D'abord, nous définissons la méthode θ-augmentée qui augmente un modèle non-paramétrique proposé dont le paramètre d'intérêt est θ en 1) partitionnant l'espace du modèle proposé selon les contours de θ et en 2) modifiant la mesure de probabilité correspondante par l'intermédiaire d'une fonction de masse d'une manière analogue à l'échantillonnage préférentiel. Cela nous permet de contrôler la loi marginale a priori suivie par θ tout en maintenant un espace totalement non-paramétrique. Nous démontrons la convergence asymptotique de la méthode TAB a posteriori pour les fonctions définies à travers des équations d'estimation dans le cas où le modèle proposé est non-paramétrique et faiblement convergent pour le mécanisme générateur des données. Cette thèse documente certaines recommandations concernant les algorithmes d'échantillonage de la distribution TAB a posteriori.

En général, le modèle θ-augmenté est le plus utile quand le modele proposé est un processus de Dirichlet (DP), vue la simplicité de l'implémentation pour la plupart des paramètres fonctionnels. Étant donnée des hyperparamètres convenables pour le modèle DP proposé, la méthode a posteriori TAB est approximativement une extension du bootstrap bayésien (BB) avec une distribution a priori subjective. Ce fait est particulièrement utile vu que le BB possède de bonnes propriétés asymptotiques. Par le biais de simulations, nous montrons que la TAB a posteriori possède de bonnes propriétés fréquentistes dans le cas où l’échantillon est petit et performe bien parmi ses concurrents.

Acknowledgements

I would like to acknowledge Professor David Stephens for his invaluable advice and vision throughout the development of this thesis. His guidance and patience helped me persevere through the difficult parts of my research. Many thanks to my family for their support during this time.

Contribution to Original Knowledge

This thesis contains several original scholarly contributions, primarily with regards to the construction of the θ-augmented measure in Chapter 3 and methods with which to carry-out the θ-augmented Bayesian inference (Chapters 4 and 5). Chapter 6 provides a first comparison of θ-augmented Bayesian inference with existing methods in Bayesian semiparametrics.

Contribution of Author

This thesis is written in its entirety by the named author, Vivian Meng. The author is the sole person responsible for the literature review, the development of the TAB methodology, the development of any theoretical proofs contained herein, the conception of computer schemes for implementation of the methodology, and the construction of all simulation studies. The interpretations of main findings in this thesis, as presented in the discussion and conclusion sections, are attributed entirely to the author.

Contents

List of Figures

List of Tables

Chapter 1

Introduction: the case for Bayesian semiparametric inference based on functional parameters of the distribution for observables

The Bayesian paradigm of statistical analysis has yielded an important class of tools toward the coherent characterization of uncertainty associated with observable events. Through the axioms of subjective probability and theorems regarding representation, one's uncertainty toward an observable quantity is transformed by available data. Statements regarding uncertainties are made via probabilities which, even without a true grasp of the data-generating mechanism, provides an accounting system for quantifying subjective beliefs and acting under uncertainty. This is an important difference between the Frequentist and Bayesian paradigm; for the Frequentist, inference statements are ascertained based on asymptotic properties associated with an estimator without bridging for how one should coherently act upon limited information conditional on the data at hand. An introduction to the Bayesian paradigm and the derivation of coherent actions can be found in Bernardo and Smith (1994).

One of the most important theorems for describing uncertainty regarding observables is de Finetti's representation theorem. This theorem applies to a sequence of random observables that is infinitely exchangeable, and makes an explicit connection as to how the array of possible models for the observable influences one's assessment of marginal probability regarding the observable. More specifically, infinite exchangeability implies that a subjective probability regarding the observed random variable can be represented as a weighted average over a collection of models for the observable, such that, had we known which model to place all of our confidence in, the data would be independent and identically distributed accordingly. The weighting assigned to the space of models for the observable is referred to as the prior distribution.

Thus the first step in many Bayesian procedures is to identify an appropriate prior over the space of models for the observable, one which complies with our genuine subjective belief regarding the observables. To be coherent, it is necessary to ensure that the prior distribution encapsulates one's true judgments regarding the observed data, so that statements of posterior probabilities are genuine and meaningful for the analyst.

Often, in a pure inferential problem, some low-dimensional parameter of the distribution for the observable will be the target of inference. The target parameter is often of scientific importance that we may hold subjective belief over it *a priori*. To be coherent, it is necessary to ensure that the prior distribution we assume is compatible with one's prior belief regarding the target parameter, which is sometimes difficult to do due to the structure of the distribution for observables we work with.

In Bayesian analysis the space of models for the observable is typically parametric, which makes the specification of a marginal prior on the target parameter particularly easy. However, a parametric model for the observable is subject to misspecification, which may lead to a lack of consistency of the Bayesian posterior.

As a way to avoid model misspecification, many authors advocate the use of nonparametric models (Hjort et al., 2010) for analysis of real data. Nonparametric Bayesian models are well studied and their properties well understood, with key results and models summarized by Ghosal and van der Vaart (2017), Ghosh and Ramamoorthi (2003), among others. Together, the low-dimensional target parameter and the nonparametric model make up a "semiparametric" inference problem.

There are in general, two ways to structure a semiparametric problem. In one approach, the target parameter is a general functional of the distribution for the observable, whereas in the alternative approach a nonparametric model is parameterized by the partitioned vector (θ, η) containing the target θ and the remaining nonparametric part η; see Tsiatis (2007) for elaboration. In the latter case, it is relatively straightforward to control the prior distribution such that the subjective belief regarding the target parameter is respected; see examples in Bayesian semiparametric regression (Section 4.2 of Müller et al. (2015) and Section 23.4 of Gelman et al. (2013)). Semiparametric models with a partitioned structure tend to arise from the use of conditional models as building blocks, with the caveat that these conditional models may require unrealistic assumptions.

If the conditional models are known to be untenable, posterior parametric inference may not be asymptotically consistent, and probabilities calculated under the given model would not be appropriate as personal probabilities for decision making for a subjective Bayesian. For example, to find the line-of-best-fit through the data (X, Y), we may define the problem via a partitioned nonparametric model, where

$$\begin{aligned}
&Y = \beta_0 + \beta_1 X + \epsilon \\
&\epsilon \sim f_\epsilon(\cdot|\eta_1) \\
&\mathbb{E}[\epsilon|X] = 0 \\
&X \sim f_x(\cdot|\eta_2)
\end{aligned}$$

so that the semiparametric model is defined by $(\beta_0, \beta_1, \eta_1, \eta_2)$. We may also consider the problem with the target parameter identified by a functional of the distribution F_{XY} for the observable (X, Y), i.e.

$$(\beta_0, \beta_1) = \underset{(b_0, b_1)\in\Theta}{\arg\min} \int (y - (b_0 + b_1 x))^2 \mathrm{d}F_{XY}(x, y). \tag{1.1}$$

Formulation of the problem by partitioning is in fact more restrictive, as it specifies a linear form for the conditional mean $\mathbb{E}[Y|X]$; even though this restriction is unlikely to be true, it has an easier interpretation as the "effect" of X. Whereas, definition as a general functional via Eqn. (1.1) is valid regardless of the structure of model F_{XY}, though the parameter is interpreted as the minimizer of an objective function. If we do not believe that the data follows a conditional linear model then the posterior probabilities based on a conditional linear model formulation are not subjective probabilities for the statistician nor theoretically optimal. As yet another example, in the context of Bayesian semiparametric estimation of a mean when data is missing at random, Ray and van der Vaart (2020) introduced a semiparametric likelihood based on propensity scores using both conditional and nonparametric likelihoods as building blocks, with the target of inference given by a transformation of the parameters for the conditional likelihood function. As this particular model asserts explicit assumptions regarding the missing data mechanism in the conditional likelihood, it is subject to misspecification.

Even though inference based on functional parameters of nonparametric models has an important advantage in Bayesian inference against likelihood misspecification, very few methods exist in the current literature to facilitate this task. Existing methods are either pseudo-Bayesian or lacking in ways to control the marginal prior for the target parameter. In the present thesis we introduce the θ-augmented Bayesian (TAB) method, which refers to Bayesian inference via the θ-augmented (TA) model, as a way of performing Bayesian semiparametric inference on functional parameters. The TA model gives us an opportunity to specify the required subjective prior for a functional parameter while maintaining a fully nonparametric model space with minimal assumptions. In the next chapter we will review existing methods in Bayesian semiparametric inference of general functional parameters and identify shortcomings of these existing methods in providing coherent Bayesian semiparametric inference for functional parameters.

1.1 Specific examples of functional parameters

Prior to presenting a literature review, we wish to present some common functional parameters that will serve as running examples throughout the thesis. Let there be a distribution function F_X, the most general interpretation of a functional parameter θ is simply a map from F_X to $\mathbb{R}^d$. We give two simple examples below.

Example 1.1. (The mean of a distribution $\mu(\cdot)$)

For random variable X supported on $\mathbb{R}$ with distribution function F_X

$$\mu(F_X) = \int_{\mathbb{R}} x \mathrm{d}F_X(x).$$

Example 1.2. (The variance of a distribution $\sigma^2(\cdot)$)

For a random variable X supported on $\mathbb{R}$ with distribution function F_X

$$\sigma^2(F_X) = \int_{\mathbb{R}} x^2 \mathrm{d}F_X(x) - \left(\int_{\mathbb{R}} x \mathrm{d}F_X(x)\right)^2.$$

Functional parameters may also be defined as the solution to an estimating equation. The estimating equations themselves may have resulted from conditional modelling, or from optimization of an objective function, etc. We provide some examples below as a non-exhaustive list.

Example 1.3. (Simple least squares regression parameter $\beta(\cdot)$) For $Y \in \mathbb{R}$, $X \in \mathbb{R}^m$, with joint distribution F_{XY} the least squares regression coefficient

$$\begin{aligned}\beta(F_{XY}) &:= \left\{b \in \mathbb{R}^m : \int (x)^\top (y - x^\top b) \mathrm{d}F_{XY}(x, y) = 0\right\} \\ &= \underset{b \in \mathbb{R}^m}{\arg\min} \int (y - x^\top b)^2 \mathrm{d}F_{XY}(x, y).\end{aligned}$$

Example 1.4. (Parameter of the closest logistic regression model $\psi(\cdot)$)

For $C \in 0, 1$, $X \in \mathbb{R}^m$, with joint distribution function F_{XC}, the parameter identifying the closest logistic regression model to F_{XC} is given by

$$\psi(F_{XC}) := \left\{t \in \mathbb{R}^m : \int x^\top (c - \tau(x^\top t)) \mathrm{d}F_{XC}(x, c) = 0\right\},$$

with $\tau(z) = \dfrac{1}{1 + \exp(-z)}$, the logistic function. The functional above is derived based on maximizing the expectation log-likelihood function of the logistic regression model with respect to F_{XC}.

Example 1.5. (Parameter of nonlinear least squares regression $\theta_{\text{NLS}}(\cdot)$)

For $Y \in \mathbb{R}$, $X \in \mathbb{R}^m$, with joint distribution F_{XY}. Let $f(x, t)$ be a nonlinear function, and define

$$\theta_{\text{NLS}}(F_{XY}) := \underset{t \in \mathbb{R}^m}{\arg\min} \int (y - f(x,t))^2 \, \mathrm{d}F_{XY}(x, y),$$

which can also be expressed as an estimating equation if the minimum is unique and occurs where the gradient of the objective function with respect to t is 0.

In general, any Frequentist estimator that can be expressed as a functional of the empirical distribution may be cast as a functional parameter for the purpose of Bayesian semiparametric estimation by substituting a nonparametric distribution F for the observable in place of the empirical distribution. For further examples and use cases, please refer to references in empirical likelihood methods (e.g. Owen (2001)).

Chapter 2

Review of semiparametric Bayesian and pseudo-Bayesian methods for inferring functional parameters

We have identified three relevant methods in the field of nonparametric Bayesian estimation of functional parameters, which are the Bayesian bootstrap (BB) (Rubin, 1981), Bayesian empirical likelihood (BEL) (Lazar, 2003), and finally, the general Bayes (GB) method (Bissiri et al., 2016). These methods will be discussed in chronological order of their first appearance in the literature.

2.1 The Bayesian bootstrap

The Bayesian bootstrap refers to the posterior inference conditional on the data whereby the sampling distribution for the observable is assumed to be multinomial and supported only on the observed data values, while the parameter of the multinomial likelihood, a vector of probabilities summing to one, is randomly distributed according to the Dirichlet distribution with hyperparameters all equal to 1. It was first proposed by Rubin (1981). Rubin obtained the BB via coupling the multinomial likelihood with a Dirichlet prior with common parameter α which tends to 0. The BB posterior can also be constructed via the Dirichlet process prior, by letting the prior concentration parameter tend to 0. In the BB posterior, the probability vector for the multinomial distribution is distributed according to a Dirichlet$(1, \ldots, 1)$ distribution. Every point $(w_1, \ldots, w_n)$ in the $(n-1)$ probability simplex maps to a model for the observable with the probability mass function $F_X(x) = \sum_{i=1}^{n} w_i \delta_{X_i}(x)$, where X_i are the observed data points. The method is nonparametric in the sense that the support of the observable grows with the number of data points.

Since the target functional parameter is a map θ from the space of distribution functions for the observable to $\mathbb{R}^d$, a posterior probability distribution over the space of distribution functions for the observable induces a distribution for the target parameter. While exact expressions of the BB posterior for functional parameters are often difficult to derive, good approximations may be obtained by repeatedly sampling F_X from the posterior Bayesian bootstrap, then transforming the sampled F_X to $\theta(F_X)$. In most cases, including those listed in Section 1.1, a functional parameter is mapped via an integral transform with respect to the distribution for the observable. The BB posterior for the observable, being supported on the space of discrete distributions, lends itself to efficient computations of the required integrals in the process of mapping F_X to $\theta(F_X)$.

One advantage of the Bayesian bootstrap is its well-behaved asymptotics. Asymptotic properties of the BB has been discussed in Lo (1987), Chatterjee and Bose (2005) for parameters defined via estimating equations, and Cheng and Huang (2010) for general semiparametric estimation in the case where the nonparametric model admits partition into a parametric target parameter and a nonparametric part. For these functionals, the BB is asymptotically consistent and shows distributional convergence to the Gaussian distribution. Interval estimation via smooth estimating equation are asymptotically also known to be consistent at the nominal level; for a definition of interval consistency, see Section 23.2 of van der Vaart (2000).

Despite being well behaved asymptotically, the method has several shortcomings. Firstly, elements of F_X contained in the BB posterior, being discrete and only supported on the data points, often draws criticism for lacking in realism as a model for continuous observables. That F_X is supported on the data points also results in the "convex hull" limitation, where the posterior parametric inference is bound by some convex hull condition given by the support, for example, if the observed data ranges between some $(X_{(1)}, X_{(n)})$ then the posterior distribution of the mean of F_X can never be outside the observed data range; see Owen (2001) for elaboration on the convex hull condition in the context of empirical likelihood which also uses models supported on the observed data points. Furthermore, the prior distribution which leads to the BB is improper. One noteworthy peculiarity with the use of an improper prior is that the prior parameter space may not be the same as that of the posterior. In this case, the limiting Dirichlet prior is supported on the vertices of the $(n-1)$ probability simplex, whereas in the posterior the Dirichlet distribution is supported on the interior of the $(n-1)$ probability simplex. Lastly, the method does not provide a way of controlling the marginal prior for a functional parameter when prior information exists, which could render the method incoherent for those following a strict Bayesian paradigm.

2.2 Bayesian empirical likelihood

The idea of conducting Bayesian inference with the profile empirical likelihood (PEL) function as a substitute for the sampling distribution was first described by Lazar (2003). The PEL is the maximum of the likelihood function over the set of models supported on the observed data points with the same value for the target parameter of interest. It was first described in the Frequentist literature by Owen (1990). The PEL is nonparametric in the sense that the space of models to profile over increases in dimensionality with increasing number of data points.

Using the notation of $F(x; \tilde{w})$ to denote the distribution function of a model parameterized by a weight vector $\tilde{w} := (w_1, \ldots, w_n)$ with w_i assigned to data point x_i, the profile empirical likelihood function, $R_n(t)$, is

$$R_n(t) = \max_{\tilde{w} \in \mathcal{S}^{n-1}} \left\{ \prod_{i=1}^{n} w_i : \theta(F(x;\tilde{w})) = t \right\}.$$

Bayesian (profile) empirical likelihood posterior distribution is the conditional distribution

$$\pi_{\text{BEL}}(\theta | X_1, \ldots, X_n) \propto R_n(t) p_\theta(t),$$

where $p_\theta(t)$ is the statistician's subjective prior distribution regarding the target parameter.

In the case that definition of the target of inference involves other nuisance parameters, e.g. in linear regression one may consider the intercept to be a nuisance, BEL inference for the target parameter may be difficult to obtain when the number of nuisance parameters is large. One approach is to marginalize $\pi_{\text{BEL}}(\theta|X_1, \ldots, X_n)$ over the nuisance parameters after obtaining a joint posterior for the target and nuisance. A second approach is to first profile out any nuisance parameters in the PEL function and then couple it with a marginal prior for the target parameter to obtain direct marginal inference. Either of these approaches will be computationally challenging. In the first approach we merely trade the computation burden of profiling out the nuisance parameter from the PEL function for the burden of sampling from a high-dimensional space and having to specify a high dimensional prior. Nevertheless, both approaches to BEL inference in the presence of nuisance parameters yield asymptotically correct inference, due to the asymptotic properties of the PEL function (Owen, 1990; Qin and Lawless, 1994). Specifics on the BEL method regarding asymptotic consistency and Gaussian tuning is found in Lazar (2003), Yang et al. (2012), and Zhao et al. (2020) for various use cases.

Despite the noteworthy feature of having good asymptotic performance without the pitfall of model misspecification, there are several downsides to BEL inference. The PEL function cannot be defined prior to seeing the data; as such, it does not constitute a true likelihood function/sampling distribution. Hence BEL inference is typically considered pseudo-Bayesian even by proponents of the method (Lazar, 2003; Yang et al., 2012; Zhao et al., 2020). Lazar (2003) argues that despite being pseudo-Bayesian the method is well justified based on an idea of Monahan and Boos (1992). However, for the Bayesian purist, the inferential statements resulting from pseudo-Bayesian methods are not subjective probabilities and thus incompatible with the theory of Bayesian decision making, and not subject to advantages of the Bayesian paradigm in small sample inference.

Besides not being a proper Bayesian distribution, another disadvantage of the Bayesian profile empirical likelihood method has to do with the convex hull condition (Section 10.4 of Owen (2001)) which can lead to significant under-coverage of interval estimates in small samples, similar to the Bayesian bootstrap. The pseudo-posterior distribution for a multivariate θ can be tricky to sample from due to the irregular shaped domain arising from the convex hull condition, which has prompted the development of a Hamiltonian MCMC algorithm by Chaudhuri et al. (2017) which solves the problem of sampling at the cost of complexity of the algorithm.

2.3 The general Bayes method of Bissiri et al. (2016)

The general Bayes (GB) method of Bissiri et al. (2016) is a method applicable to a functional parameter that is defined as the minimizer of an expected loss function, such that the "true value" θ_0 is

$$\theta_0 := \arg\min_{t\in\Theta} \int l(x,t)\mathrm{d}F_0(x),$$

where $F_0(x)$ is the distribution function of the true data generating mechanism. The general Bayes method proposes the conditional density function

$$\pi_{\mathrm{GB}}(\theta|x_1,\ldots,x_n) \propto \exp\left(-w\sum_{i=1}^{n} l(x_i,\theta)\right) p_\theta(\theta)$$

as the optimal choice after seeing the data, either according to a given "coherence property," or decision theoretic criteria. The scaling constant w is arbitrary; Bissiri et al. offered some suggestions for how to choose w in their original paper, though not one strategy was singled out above others as being more appropriate. Several of the strategies require choosing w based on the observed data. Curiously, the coherence property outlined on p. 1104 of Bissiri et al. (2016) requires that the GB posterior be invariant to the order in which the data is received, and be amenable to sequential update. Yet if we follow the suggestions for selecting w based on data, then the loss function weighting changes at each step of the sequential update, thus violating the coherence property.

The Bayes estimator generated under this method is asymptotically justified due to

$$\lim_{n\to\infty}\left\{\sum_{i=1}^{n} l(x_i,t)\right\} = \int l(x,t)\mathrm{d}F_0(x),$$

such that the mode of the posterior distribution should converge to the truth at θ_0. However, it is not known if the interval estimates are asymptotically consistent - but it is easy to see that interval consistency will depend on having the correct w.

Another marketed advantage of the general Bayes update is that it avoids the pitfall of model misspecification due to not having been derived through de Finetti's representation theorem. If $\pi_{\mathrm{GB}}(\theta|x_1,\ldots,x_n)$ is indeed a genuine conditional subjective probability, then, Bayes theorem for conditioning should also hold, such that

$$\pi_{\mathrm{GB}}(\theta|x_1,\ldots,x_n) = \frac{P(x_1,\ldots,x_n|\theta)\times p_\theta(\theta)}{P(x_1,\ldots,x_n)}.$$

The entity $P(x_1,\ldots,x_n)$ must exist, and when the data are exchangeable, must also be subject to de Finetti's representation theorem, which then partially informs $P(x_1,\ldots,x_n|\theta)$ considering that θ is a functional of the distribution for observables. Based on the above manipulation of probabilities, we found that in many situations, specifying the loss function for GB inference uniquely identifies the sampling model, and therefore the method is not actually model-free. For example, using the squared loss in GB inference leads to an inherent belief that the data came from a parametric Gaussian sampling model; see Appendix A.1 and A.2. One can examine the implications placed on the model space. In the case that one judges the model space as unrealistic, the GB posterior loses the interpretation as a subjective belief distribution and the advantages of subjective Bayesian paradigm in small sample inference.

However the general Bayes method is also purported as an optimal choice based on the minimization of an *empirical* risk function, $R(\nu; \hat{F}_n)$, which, as given by Bissiri et al. (2016), is

$$R(\nu; \hat{F}_n) \equiv \int \sum_{i=1}^{n} l(\theta, x_i)\nu(\mathrm{d}\theta) + d_{KL}(\nu||\pi) = \int \left[n \int l(\theta, x)\nu(\mathrm{d}\theta) + d_{KL}(\nu||\pi) \right] \hat{F}_n(\mathrm{d}x), \tag{2.1}$$

with $\hat{F}_n$ denoting the empirical data distribution, and π denoting one's subjective prior distribution for θ. The concept of minimizing empirical risk is popular in the machine learning discipline (Vapnik, 1992). Eqn. (2.1) was also the subject of Jiang and Tanner (2008) through which they developed the same conditional distribution as π_{GB} from the perspective of best model for classification in data mining. However, as a treatment based on decision theory, it seems to lack certain elements. For a review of Bayesian and Frequentist decision theory, see Berger (1985). Here we note that a Frequentist decision under the principle of risk minimization (or equally, maximizing the expected utility) would consider a choice optimal if $R(\nu; F_0)$ is available. Alternatively, a Frequentist choosing to employ the minimax principle would choose

$$\arg\min_{\nu}\{\max_{F \in A} R(\nu, F)\},$$

where A is the set of all distributions with $x_1, \ldots, x_n$ in its support, and may be problematic if l is unbounded, and if not, regions with extreme values of l would dominate the selection process.

From the perspective of minimizing Bayes risk, we would have to model uncertainty in F through $Q(F|\text{Data})$, perhaps nonparametrically if we are not willing to assume a parametric likelihood, and proceed to minimizing the Bayes risk

$$\int R(\nu; F)Q(\mathrm{d}F|\text{Data}),$$

such that the optimal choice is

$$\hat{\nu} = \arg\min_{\nu} \int R(\nu; F)Q(\mathrm{d}F|\text{Data}).$$

The general Bayes solution will be chosen as $\hat{\nu}$ only if the our posterior $Q(F|\text{Data})$ over the model space is concentrated at $\hat{F}_n$ almost surely- which is certainly unrealistic. Further, note that regardless of what the form of $Q(F|\text{Data})$ is, there is most likely a discrepancy between $\hat{\nu}$ and the induced distribution of θ based $Q(F|\text{Data})$. If the target functional of interest is indeed the one defined through $\theta(F) = \arg\min_{t\in\Theta} \int l(x,t)\mathrm{d}F(x)$ then it would be incoherent to choose anything other than the probability distribution induced by $Q(\theta(F)|\text{Data})$ to represent our belief and there will be not a lot of importance in identifying $\hat{\nu}$ through minimization of Bayes risk.

Chapter 3

Theory of θ-augmented Bayesian inference

Since functional parameters are defined via various mappings from the probability distribution for the observable to $\mathbb{R}^d$, uncertainties regarding functional parameters are induced by our uncertainties with regards to the probability distribution for the observable. This thesis presents what we term the θ-augmented Bayesian method as a coherent way to account for induced uncertainties, via a θ-augmented probability measure over random measures for the observable.

The theory for θ-augmentation has two prerequisites, that being

1. infinite exchangeability of the observed data, hence applicability of de Finetti's representation theorem to exert existence of a prior distribution over a model space, and,

2. that all models for the observable in the prior model space be dominated by the same dominating measure *almost surely*.

Without loss of generality, let the distribution function for the observable be denoted by F_X. One begins by asserting the existence of a prior distribution over distribution functions for the observables, which is denoted by $\Pi(F_X)$. When we judge the observed quantities $x_1, \ldots, x_n$ as being infinitely exchangeable, we may apply de Finetti's representation theorem, the general form of which is found on p. 177 of Bernardo and Smith (1994), or p. 83 of Ghosh and Ramamoorthi (2003).

The de Finetti theorem states that, loosely speaking, letting $M(\mathbb{R})$ be the space of all distribution functions on $\mathbb{R}$, and $X_1, X_2, \ldots$ an infinitely exchangeable sequence of realizations of real-valued random quantities with probability measure P, there exists a measure $\Pi(F_X)$ over the space $M(\mathbb{R})$ such that the joint distribution function of $\tilde{x}_n := (x_1, \ldots, x_n)$ has the form

$$P(\tilde{x}_n) = \int_{M(\mathbb{R})} \prod_{i=1}^{n} F_X(x_i) \mathrm{d}\Pi(F_X).$$

The space $M(\mathbb{R})$ mentioned by the representation theorem is quite general and may include all nonparametric models as well as parametric ones. We assume the probability model $\mathcal{P}_\Pi := (M(\mathbb{R}), \Sigma, \Pi)$, where Σ is the σ-algebra over the space of all random measures for X. The subscript of $\mathcal{P}_\Pi$ shows our notation for the corresponding measure of this probability model explicitly. Without loss of generality, suppose that $\theta : M(\mathbb{R}) \to \mathbb{R}$, i.e. θ is a measurable function which maps elements of $M(\mathbb{R})$ to the real line. Existence of a prior distribution $\Pi(F_X)$ induces a distribution function for θ.

Letting F_θ^Π denote the distribution function of θ induced by Π, we have

$$F_\theta^\Pi(t) = \int_{M(\mathbb{R})} \mathbb{I}[\theta(F_X) \leq t] \mathrm{d}\Pi(F_X).$$

In the case that the distribution function of θ is absolutely continuous with respect to the Lebesgue measure, a density function exists s.t.

$$F_\theta^\Pi(t) = \int_{-\infty}^{t} q_\theta^\Pi(u) \mathrm{d}u,$$

where q_θ^Π denotes the density function of θ induced by Π.

Let us denote the set of observed data as $\tilde{x}_n := (x_1, \ldots, x_n)$. Let $F_{\theta|\tilde{x}_n}^\Pi$ denote the distribution function of θ conditional on $\tilde{x}_n$ induced by the measure Π conditional on $\tilde{x}_n$, that is,

$$F_{\theta|\tilde{x}_n}^\Pi(t|\tilde{x}_n) = \int_{M(\mathbb{R})} \mathbb{I}[\theta(F_X) \leq t] \mathrm{d}\Pi(F_X|\tilde{x}_n).$$

In the case that that θ is continuous, an induced density function $q^{\Pi}_{\theta|\tilde{x}_n}$ exists s.t.

$$F^{\Pi}_{\theta|\tilde{x}_n}(t|\tilde{x}_n) = \int_{-\infty}^{t} q^{\Pi}_{\theta|\tilde{x}_n}(u|\tilde{x}_n)\mathrm{d}u.$$

As for the second prerequisite of the TAB method, we require applicability of Bayes rule for finding (a version of) the conditional distribution $\Pi(F_X|\tilde{x}_n)$, that is

$$\Pi(F_X \in A|\tilde{x}_n) = \frac{\int_A \prod_{i=1}^n F_X(x_i)\mathrm{d}\Pi(F_X)}{\int \prod_{i=1}^n F_X(x_i)\mathrm{d}\Pi(F_X)}. \tag{3.1}$$

Equation (3.1) differs slightly from the way Bayes rule is written in the books by Ghosal and van der Vaart (2017) and Ghosh and Ramamoorthi (2003) because these books introduced it in the setting of parametric inference. According to Section 1.3 of Ghosal and van der Vaart (2017), the key is to determine if every distribution F_X is dominated by some σ-finite measure *almost surely* Π. This is necessarily the case if Π is a prior supported on the subset of parametric models. However, many nonparametric Bayesian models are not dominated in this way, leading to the absence of Bayes rule when developing the posterior distributions for these Bayesian nonparametric methods, with the Dirichlet process prior being a simple example.

Luckily, if we restrict our attention to kernel mixture models with kernels that are dominated by either the Lebesgue measure or a version of the counting measure, Bayes rule applies. Specifically, let

$$f(x) = \int \mathcal{K}(x|\eta)\mathrm{d}F_H(\eta),$$

where η parameterizes the mixture kernel $\mathcal{K}$. The observation model $f(x)$ depends on the measure F_H. Regardless of F_H being parametric or nonparametric, $f(x)$ will be dominated by the Lebesgue measure (or counting measure) if the kernel $\mathcal{K}$ is dominated; see Section 5.3 of Ghosh and Ramamoorthi (2003).

3.1 Defining the θ-augmented probability measure and probability model

The underlying idea of the θ-augmented semiparametric method is to gently modify a nice proposal measure Π according to the contours of the model space as given by θ. Without loss of generality, we assume that $\theta : M(\mathbb{R}) \to \mathbb{R}$. Let us further assume some $m : \mathbb{R} \to \mathbb{R}^+$ such that the composite function $m \circ \theta : M(\mathbb{R}) \to \mathbb{R}^+$ is a Σ-measurable positive integrable function, that is,

$$m(\theta(F_X)) \geq 0,$$

$$\int_{M(\mathbb{R})} m(\theta(F_X)) \mathrm{d}\Pi(F_X) < \infty.$$

We define a new probability measure $\Pi^\star$, the θ-augmented measure, to be,

$$\begin{aligned}\Pi^\star(A) &= \frac{\int_{M(\mathbb{R})} \mathbb{I}[F_X \in A] m(\theta(F_X)) \mathrm{d}\Pi(F_X)}{\int_{M(\mathbb{R})} m(\theta(F_X)) \mathrm{d}\Pi(F_X)} \\ &= \frac{\int_{M(\mathbb{R})} \mathbb{I}[F_X \in A] m(\theta(F_X)) \mathrm{d}\Pi(F_X)}{Z_{\Pi^\star}},\end{aligned} \tag{3.2}$$

where

$$Z_{\Pi^\star} := \int_{M(\mathbb{R})} m(\theta(F_X)) \mathrm{d}\Pi(F_X),$$

for any measurable event $A \in \Sigma$. This is conceptually similar to slicing, or “partitioning”, of the model space based on $\theta(F_X) = t$, and for every model in this subspace, adjusting its probability content by a factor $m(t)$. Furthermore, the expectation of any Σ-measurable function $g(F_X)$ with regards to $\Pi^\star$ is

$$\mathbb{E}_{\Pi^\star}[g(F_X)] = \frac{\int_{M(\mathbb{R})} g(F_X) m(\theta(F_X)) \mathrm{d}\Pi(F_X)}{Z_{\Pi^\star}}.$$

We can verify that $\Pi^\star$ is a valid measure on $(M(\mathbb{R}), \Sigma)$. The measure of any event $A \in \Sigma$ under the unnormalized measure $\Pi^\star \cdot Z_{\Pi^\star}$ is equal to the expectation of the measurable non-negative function $m(\theta(F_X)) \times \mathbb{I}[F_X \in A]$ taken under Π. The countable additivity of the unnormalized $\Pi^\star \cdot Z_{\Pi^\star}$ can easily be verified. Therefore $\Pi^\star$ is a probability measure.

While we see Eqn. (3.2) as a recipe for constructing semiparametric models out of nonparametric ones via parameter augmentation, the corresponding change of measure can also be regarded as generalizing the usual approach to importance sampling, with $\Pi^\star$ playing the role of the target distribution and Π the biasing distribution. Typically, importance sampling is not performed with nonparametric distributions. Our construction of $\Pi^\star$ in Eqn. (3.2) ensures that the Radon–Nikodym derivative of the target distribution $\Pi^\star$ with respect to Π depends only on the functional parameter θ; in this case that derivative is $(m \circ \theta)/Z_{\Pi^\star}$.

The measure $\Pi^\star$ as given by Eqn. (3.2) is a distribution over random measures with parameters m and Π, and it serves as the prior in our Bayesian analysis. As a probability model for F_X, we define the θ-augmented probability model, denoted as

$$\mathrm{TA}(m, \mathcal{P}_\Pi),$$

to be the probability model which inherits the sample space and σ-algebra of $\mathcal{P}_\Pi$, and is equipped with a measure $\Pi^\star$ which is a modification of the measure Π of $\mathcal{P}_\Pi$ by m according to Eqn. (3.2). We will refer to the parameter m as the *weighting function*, and refer to $\mathcal{P}_\Pi$ as the *proposal model*, and Π as the *proposal measure*. As we do not wish to overload the symbol $\Pi^\star$ with subscripts or superscripts, we will state it as the probability measure $\Pi^\star$ of a $\mathrm{TA}(m, \mathcal{P}_\Pi)$ model, as often as needed, so that the parameters m and Π associated with a $\Pi^\star$ are understood clearly. The domain of the weighting function m will be indicated explicitly in its specification, so that we remove any ambiguity with regards to the construction of $\Pi^\star$.

Suppose that θ is absolutely continuous, and we aim to specify a θ-augmented model such that the induced marginal density of θ is equal to p_θ. To achieve this, we may choose

$$m = \frac{p_\theta}{q_\theta^\Pi},$$

where p_θ/q_θ^Π denotes pointwise division of p_θ by q_θ^Π. As a convention in this thesis, if f and g are two functions with the same domain, we shall use fg or $f \cdot g$ to denote the pointwise product of these functions, and $\frac{f}{g}$ or f/g to denote pointwise quotient.

To verify, denote with $F_\theta^{\Pi^\star}$ the distribution function of θ induced by the measure $\Pi^\star$ of the TA$(m = p_\theta/q_\theta^\Pi, \mathcal{P}_\Pi)$ model, we have that

$$\begin{aligned} F_\theta^{\Pi^\star}(t) := \Pi^\star(\theta \leq t) &= \frac{\int_{M(\mathbb{R})} \mathbb{I}[\theta(F_X) \leq t] \frac{p_\theta(\theta(F_X))}{q_\theta^\Pi(\theta(F_X))} \mathrm{d}\Pi(F_X)}{Z_{\Pi^\star}} \qquad (3.3)\\ &\propto \int_{-\infty}^{\infty} \mathbb{I}[u \leq t] \frac{p_\theta(u)}{q_\theta^\Pi(u)} \cdot q_\theta^\Pi(u) \mathrm{d}u \qquad (3.4)\\ &= \int_{-\infty}^{t} p_\theta(u) \mathrm{d}u, \end{aligned}$$

i.e. the density function corresponding to $F_\theta^{\Pi^\star}(t)$ is exactly p_θ. The simplification from Eqn. (3.3) to (3.4) is due to

$$\mathbb{I}[\theta(F_X) \leq t] \frac{p_\theta(\theta(F_X))}{q_\theta^\Pi(\theta(F_X))}$$

depending only on the value of $\theta(F_X)$.

As for the induced distribution of $\theta|\tilde{x}_n$ when $F_X \sim$ TA$(m = p_\theta/q_\theta^\Pi, \Pi)$, because we restrict ourselves to Π where every F_X in the space of random measures is dominated Π *almost surely*, Bayes rule applies. We apply Bayes rule to obtain $F_{\theta|\tilde{x}_n}^{\Pi^\star}$, the distribution function of $\theta|\tilde{x}_n$ induced by $\Pi^\star$, as

$$\begin{aligned} F_{\theta|\tilde{x}_n}^{\Pi^\star}(t|\tilde{x}_n) := \Pi^\star(\theta \leq t|\tilde{x}_n) &\propto \int_{M(\mathbb{R})} \mathbb{I}[\theta(F_X) \leq t] \cdot \prod_{i=1}^{n} F_X(x_i) \mathrm{d}\Pi^\star(F_X) \qquad (3.5)\\ &\propto \int_{M(\mathbb{R})} \mathbb{I}[\theta(F_X) \leq t] \cdot \frac{p_\theta(\theta(F_X))}{q_\theta^\Pi(\theta(F_X))} \cdot \prod_{i=1}^{n} F_X(x_i) \mathrm{d}\Pi(F_X) \qquad (3.6)\\ &\propto \int_{M(\mathbb{R})} \mathbb{I}[\theta(F_X) \leq t] \cdot \frac{p_\theta(\theta(F_X))}{q_\theta^\Pi(\theta(F_X))} \cdot \mathrm{d}\Pi(F_X|\tilde{x}_n) \qquad (3.7)\\ &\propto \int_{-\infty}^{\infty} \mathbb{I}[u \leq t] \frac{p_\theta(u)}{q_\theta^\Pi(u)} \cdot q_{\theta|\tilde{x}_n}^\Pi(u|\tilde{x}_n) \mathrm{d}u \qquad (3.8)\\ &= \int_{-\infty}^{t} \frac{q_{\theta|\tilde{x}_n}^\Pi(u|\tilde{x}_n)}{q_\theta^\Pi(u)} \cdot p_\theta(u) \mathrm{d}u. \qquad (3.9) \end{aligned}$$

Bayes rule is important here in order to deduce the form of the posterior distribution of θ under $\Pi^\star$ i.e. Eqn. (3.5). Had it not applied, the form of $F_{\theta|\tilde{x}_n}^{\Pi^\star}$ would have to be found through other means similar to other non-dominated models in Bayesian nonparametrics, and the weighting function p_θ/q_θ^Π may not have appeared in $F_{\theta|\tilde{x}_n}^{\Pi^\star}$ as elegantly as it did in Eqn. (3.9).

There are several observations to be made based on Eqn. (3.5)-(3.9). Firstly, let $q^{\Pi^\star}_{\theta|\tilde{x}_n}$ denote the target posterior density function corresponding to $F^{\Pi^\star}_{\theta|\tilde{x}_n}$, we have that

$$q^{\Pi^\star}_{\theta|\tilde{x}_n} \propto \frac{q^{\Pi}_{\theta|\tilde{x}_n}}{q^{\Pi}_{\theta}} \cdot p_\theta.$$

This expression shows clearly that, after seeing the data, the subjective prior distribution p_θ is modified by $q^{\Pi}_{\theta|\tilde{x}_n}/q^{\Pi}_{\theta}$, which we term the *effective likelihood*. Secondly, for $F_X \sim \text{TA}(m = p_\theta/q^{\Pi}_{\theta}, \Pi)$, the conditional random variable $F_X|\tilde{x}_n$ is again distributed according to a θ-augmented model but with the proposal measure updated to $\Pi(\cdot|\tilde{x}_n)$.

In the case that the distribution of θ induced by Π is discrete, it is easy to show that, given a subjective marginal probability mass function (PMF) P_θ for θ, and an induced PMF $\mathrm{Q}^{\Pi}_{\theta}$ given the proposal measure Π, one would specify

$$m = \frac{\mathrm{P}_\theta}{\mathrm{Q}^{\Pi}_{\theta}},$$

as the weighting function in the θ-augmented model, which results in

$$\begin{aligned} \Pi^\star(\theta = t) &= \mathrm{P}_\theta(t), \\ \Pi^\star(\theta = t|\tilde{x}_n) &\propto \frac{\mathrm{P}_\theta(t)}{\mathrm{Q}^{\Pi}_{\theta}(t)} \mathrm{Q}^{\Pi}_{\theta|\tilde{x}_n}(t|\tilde{x}_n), \end{aligned}$$

for t in the support of $\mathrm{Q}^{\Pi}_{\theta}$.

3.1.1 θ-augmented Bayesian inference

We take the θ-augmented Bayesian (TAB) inference (or method) as referring to inference with a $\text{TA}(m = p_\theta/q^{\Pi}_{\theta}, \mathcal{P}_\Pi)$ model as the nonparametric prior, unless otherwise specified. The probability measure $\Pi^\star$ associated with a $\text{TA}(m = p_\theta/q^{\Pi}_{\theta}, \mathcal{P}_\Pi)$ model is defined according to Eqn. (3.2) while making the appropriate substitutions for the parameters m and $\mathcal{P}_\Pi$. In the weighting function $p_\theta/q^{\Pi}_{\theta}$, the numerator is specified by the statistician as their selected model, while the denominator is completely determined by the proposal measure Π. Any posterior distribution obtained via the TAB method will be termed TAB posterior.

To show how the model relates to observables, we may write

$$X \sim F_X$$
$$F_X \sim \text{TA}(m = p_\theta / q_\theta^\Pi, \mathcal{P}_\Pi).$$

To conduct TAB inference, we are therefore required to

1. define the target parameter as a functional of the distribution for the observable,

2. select a proposal model $\mathcal{P}_\Pi$ that results in dominated models for the observable Π *almost surely*. Suitable proposal models for semiparametric inference will be discussed in Section 5,

3. ensure p_θ and $\mathcal{P}_\Pi$ are compatible by checking that the resulting weighting function p_θ/q_θ^Π is integrable– an assumption in the definition of a θ-augmented measure,

4. select an algorithm to approximate the conditional distribution of $\theta|\tilde{x}_n$ under the given TA$(m = p_\theta/q_\theta^\Pi, \mathcal{P}_\Pi)$ model, to be discussed in Section 4.

3.2 Multivariate θ and marginal inference

The method of θ-augmentation is not limited to $\theta \in \mathbb{R}^1$. In the case that the target parameter is $\theta \in \mathbb{R}^d$, we require a weighting function m which takes $(\theta_1(F_X), \ldots, \theta_d(F_X))$ as the argument. In specific, as a simple extension to Eqn. (3.3), taking the θ-augmented model with weighting function

$$m(\theta_1, \cdots, \theta_d) = \frac{p_\theta(\theta_1, \ldots, \theta_d)}{q_\theta^\Pi(\theta_1, \ldots, \theta_d)}, \tag{3.10}$$

as the prior will lead to the prior marginal density of θ being exactly $p_\theta(\theta_1, \ldots, \theta_d)$ as required. However, actual implementation utilizing Eqn. (3.10) will be challenging, due to high dimensionality of the induced joint density function $q_\theta^\Pi(\theta_1, \ldots, \theta_d)$.

As a compromise, if we limit ourselves to specifying our subjective prior in only one of the dimensions, say θ_1, then the implementation will be a lot more feasible. That is, suppose

$$
\begin{aligned}
m(\theta_1, \cdots, \theta_d) &= \frac{p_{\theta_1}(\theta_1) \times q^{\Pi}_{(-\theta_1)|\theta_1}(\theta_2, \ldots, \theta_d)}{q^{\Pi}_{\theta}(\theta_1, \ldots, \theta_d)} \\
&= \frac{p_{\theta_1}(\theta_1)}{q^{\Pi}_{\theta_1}(\theta_1)},
\end{aligned} \tag{3.11}
$$

where $q^{\Pi}_{(-\theta_1)|\theta_1}$ is the conditional distribution of the remaining parameters given θ_1 induced by Π. Then, marginal prior of θ_1 under the corresponding θ-augmented model will be exactly p_{θ_1} as required, but for some other dimension θ_j, marginal prior under this θ-augmented model will be

$$
\begin{aligned}
\Pi^{\star}(\theta_j \leq t_j) &= \int\int \mathbb{1}[\theta_j \leq t_j] \frac{p_{\theta_1}(\theta_1)}{q^{\Pi}_{\theta_1}(\theta_1)} q^{\Pi}_{\theta_1, \theta_j}(\theta_1, \theta_j) \mathrm{d}\theta_1 \mathrm{d}\theta_j \\
&= \int_{-\infty}^{t_j} \left[\int_{-\infty}^{\infty} q^{\Pi}_{\theta_j|\theta_1}(u|\theta_1) p_{\theta_1}(\theta_1) \mathrm{d}\theta_1 \right] \mathrm{d}u,
\end{aligned}
$$

which shows that construction θ-augmented models through Eqn. (3.11) impacts the marginal prior (and posterior) of untargeted dimensions of θ. In other words, θ_1 will be informative of θ_j if $q^{\Pi}_{\theta_j|\theta_1}(\theta_j|\theta_1)$ depends on θ_1 under the proposal measure Π. The above method for inferring multivariate θ preserves coherence between the joint and marginal distributions. Although coherent, we relinquish our control over all individual margins except the one that we presume to be the most important.

As yet another an alternative, at the sacrifice of coherence in joint inference, one could conduct multiple TAB inference via the collection of models $\{\text{TA}(m_j, \mathcal{P}_{\Pi}); j = 1 \ldots, d\}$, where $\text{TA}(m_j, \mathcal{P}_{\Pi})$ fulfills the prior marginal p_{θ_j} for the j-th dimensions of θ with m_j specified in the fashion of Eqn. (3.11). It is clear that in this approach, each marginal inference is derived through a different probability model so no joint inference is available. This method of inferring multiple θ_j has the advantage of being computationally simple. Ultimately, the choice of how to construct weighting function m given the options above would depend on the application and one's priorities regarding various aspects of one's subjective prior regarding the target parameter vector.

3.3 Asymptotics

Despite subjective probability being the currency for accounting in the Bayesian world, Bayesians should also be concerned with recovering truth, and asymptotic consistency, and worried about misspecification. Though the true data-generating mechanism is unknown to us, we believe that the subjective nature of Bayesian inference does not preclude the existence, and one's quest, of appropriate models that are asymptotic consistent.

In terms of estimator consistency, a Bayes estimator $\hat{\theta}_{B,n}$, which is a function of the posterior distribution, is said to be consistent for the truth θ_0 if

$$\lim_{n\to\infty} \int \mathbb{1}\left[|\hat{\theta}_{B,n} - \theta_0| > \epsilon\right] \mathrm{d}F_0 = 0.$$

Whereas, asymptotic consistency of a Bayes procedure for a particular parameter typically refers to a type of convergence in probability. Ghosal (1997) defines consistency for a Bayesian posterior $P(\cdot|\tilde{x}_n)$ for a parameter as, for every neighbourhood U of θ_0

$$\lim_{n\to\infty} P(U|\tilde{x}_n) = 1 \quad a.s. \quad \mathbb{P}_{F_0},$$

which also appears in Hjort et al. (2010) page 52, in the complement U^c,

$$\lim_{n\to\infty} P(U^c|\tilde{x}_n) = 0 \quad a.s. \quad \mathbb{P}_{F_0}. \tag{3.12}$$

It is easy to see that asymptotic consistency of the TAB method is dependent on the particular proposal model employed. Suppose that the proposal model $\mathcal{P}_\Pi$ equipped with measure Π satisfies consistency per Eqn. (3.12), that is, for every neighbourhood U of θ_0,

$$\lim_{n\to\infty} \Pi(\theta \in U^c|\tilde{x}_n) = 0 \quad a.s. \quad \mathbb{P}_{F_0}.$$

If the particular weighting function p_θ/q_θ^Π we use to define a θ-augmented model is bounded by a number M, then clearly the induced distribution of θ under the TAB posterior $\Pi^\star(\cdot|\tilde{x})$ is consistent for θ_0 as

$$\begin{aligned}
&\lim_{n\to\infty} \Pi^\star(\theta \in U^c|\tilde{x}_n) \\
&= \lim_{n\to\infty} \int \mathbb{1}\left[\theta(F_X) \in U^c\right] \cdot p_\theta(\theta(F_X))/q_\theta^\Pi(\theta(F_X)) \mathrm{d}\Pi(F_X|\tilde{x}_n) \\
&\leq \lim_{n\to\infty} M \int \mathbb{1}\left[\theta(F_X) \in U^c\right] \mathrm{d}\Pi(F_X|\tilde{x}_n) = 0
\end{aligned}$$

It is possible to ensure that $p_\theta/q^\Pi_\theta < M$ because, as we need to ensure integrability of the weighting function in a θ-augmented model, in practice we will almost always choose a proposal measure where the induced q^Π_θ is more heavy-tailed than p_θ, satisfying the boundedness condition.

For conciseness, let the weighting function p_θ/q^Π_θ be denoted again by m. The form of TAB posterior, which is proportional to $q^\Pi_{\theta|\tilde{x}_n} \cdot m(\theta)$, has clear parallels in standard Bayesian inference with $q^\Pi_{\theta|\tilde{x}_n}(\theta|\tilde{x}_n)$ taking the place of the likelihood function; see p. 287 of Bernardo and Smith (1994) for a development of asymptotics in standard Bayesian inference. Suppose that $q^\Pi_{\theta|\tilde{x}_n}$ has a unique maximum $t_{q,n}$, and a Hessian matrix at $t_{q,n}$ of Σ_n^{-1}. We may expand $\log q^\Pi_{\theta|\tilde{x}_n}$ about its maximum $t_{q,n}$ to obtain

$$\log q^\Pi_{\theta|\tilde{x}_n}(\theta) = \log q^\Pi_{\theta|\tilde{x}_n}(t_{q,n}) - \frac{1}{2}(\theta - t_{q,n})^\top(\Sigma_n^{-1})(\theta - t_{q,n}) + R_n,$$

with R_n being the remainder term. Suppose that $m(\theta)$ has a unique maximum t_0, and a Hessian matrix at t_0 of H_0. An expansion of $\log m$ about its maximum t_0 gives

$$\log m(\theta) = \log m(t_0) - \frac{1}{2}(\theta - t_0)^\top H_0(\theta - t_0) + R_0,$$

where R_0 is some remainder term. We assume regularity conditions that make R_n and R_0 negligible for large n. Then, the TAB posterior for θ (up to proportionality constant) may be approximated by

$$\exp\left(-\frac{1}{2}(\theta - t_{q,n})^\top(\Sigma_n^{-1})(\theta - t_{q,n}) - \frac{1}{2}(\theta - t_0)^\top H_0(\theta - t_0)\right)$$

for large n. Therefore, asymptotically, $q^{\Pi\star}_{\theta|\tilde{x}_n}(\theta)$ is approximately Normal(t_n, H_n), where

$$H_n = H_0 + \Sigma_n^{-1}$$

and

$$t_n = H_n^{-1}\left(H_0 t_0 + \Sigma_n^{-1} t_{q,n}\right).$$

If we have parametric consistency under the proposal model, then the curvature of $\log q^\Pi_{\theta|\tilde{x}_n}$ about its maximum as given by Σ_n^{-1} must increase with sample size. As $n \to \infty$, H_0

will be negligible compared to Σ_n^{-1}, which warrants an approximation of $q^{\Pi^\star}_{\theta|\tilde{x}_n}$ based on Normal$(t_{q,n}, \Sigma_n^{-1})$.

It is clear that asymptotic behaviour of the TAB posterior for the parameter θ depends on the behaviour of $q^{\Pi}_{\theta|\tilde{x}_n}$. In the next section we will examine the asymptotic behaviour of functionals identified via estimating equations when the proposal model corresponding is a weakly consistent Bayesian nonparametric model.

3.3.1 Parametric consistency of TAB posterior for parameters defined via estimating equations when the proposal model is weakly consistent

Conditions for asymptotic consistency of Bayesian nonparametric models can be found throughout standard references in Bayesian nonparametrics, e.g. Ghosal and van der Vaart (2017) and Ghosh and Ramamoorthi (2003). In this section we identify the conditions for parametric consistency of TAB posterior for functional parameters defined via estimating equations, while assuming weak consistency of the nonparametric proposal model. We assume that the data-generating distribution has a density f_0, and the nonparametric proposal probability model with the measure Π, is supported on random measures admitting density functions f. Generalization of the results to probability models supported on random distribution functions is straightforward.

We have that, for functional parameters defined via an estimating equations $g(x, t)$,

$$\theta(f) := \left\{ t \in \Theta \quad s.t. \int g(x,t) f(x) \mathrm{d}x = 0 \right\},$$

boundedness of $g(x, t)$, continuity of $g(x, t)$ in x and t, integrability of $\int g(x,t)f(x)\mathrm{d}x$ for all f in the support of Π, integrability of $\int g(x,t)f_0(x)\mathrm{d}x$ at any given t, and weak consistency of the proposal model together imply parametric consistency under the proposal model.

This stems from the definition of weak consistency (Definition 1 of Ghosal et al. (1999)). Given a probability measure Π with sample space $\mathcal{F}$, let $\phi_i, i = 1, \ldots, k$ be bounded continuous functions on $\mathbb{R}$. Let

$$U = \left\{ f \in \mathcal{F} : \left| \int \phi_i(x) f(\mathrm{d}x) - \int \phi_i(x) f_0(\mathrm{d}x) \right| < \epsilon, i = 1, 2, \ldots, k \right\} \tag{3.13}$$

be a weak neighbourhood of f_0. A probability model with measure Π is said to be weakly consistent for f_0 if with $P_{f_0^-}$ probability 1

$$\Pi(U | X_1, X_2, \ldots, X_n) \to 1$$

for *all* weak neighbourhoods U of f_0. The boundedness condition for ϕ_i's that define a weak neighbourhood may not be satisfied by $g(x, t)$ when the domain of the function is not compact at any given t. In this case, a workaround is to truncate x to be within some bounds – this is generally possible to do when the observables are natural/physical phenomena. It is therefore assumed that $g(x, t)$ at any given t can be used to define weak neighbourhoods in the development of the subsequent proof.

For estimating equations that lead to explicit solutions for the parameter, in the form of $\theta(f) = \int k(x) f(\mathrm{d}x)$ for some general function $k(x)$, parametric consistency is immediately apparent. For example, when $\theta(f) := \int x f(\mathrm{d}x)$, weak consistency of a model for f_0 implies, for all $\epsilon > 0$,

$$\lim_{n \to \infty} \Pi \left(\left\{ f \in \mathcal{F} : \left| \int x f(\mathrm{d}x) - \int x f_0(\mathrm{d}x) \right| < \epsilon \right\} | X_1, X_2, \ldots, X_n \right) = 1,$$

due to

$$\left\{ f \in \mathcal{F} : \left| \int x f(\mathrm{d}x) - \int x f_0(\mathrm{d}x) \right| < \epsilon \right\}$$

being a weak neighbourhood per definition given in Eqn. (3.13). Similar arguments apply to show parametric consistency for estimating the k-th moment, variance/covariance, coefficients for linear regression or, in general, simple functions of moments when Slutsky's theorem applies.

Now consider the case that the parameter θ is defined as the solution to some estimating function. Without loss of generality assume $\theta \in \mathbb{R}$. Let

$$h_0(t) := \int g(x, t) f_0(x) \mathrm{d}x$$

$$s.t. \quad h_0(\theta_0) = 0,$$

i.e. the estimating function identifies the true parameter θ_0 under the true data generating mechanism f_0. We assume that θ_0 is unique, and sgn $h_0(\theta_0 + \delta) \neq$ sgn $h_0(\theta_0 - \delta)$, i.e. the function h_0 has opposite signs at the two sides of θ_0. By the continuity of $g(x,t)$ in t for any x, the function h_0 is continuous. Then there exist for ϵ small enough

$$\begin{aligned}
C_a(\epsilon) &:= \{t \in \Theta : h_0(t) = \epsilon\} \\
C_b(\epsilon) &:= \{t \in \Theta : h_0(t) = -\epsilon\} \\
t_a(\epsilon) &:= \arg\min_{t \in C_a} |t - \theta_0| \\
t_b(\epsilon) &:= \arg\min_{t \in C_b} |t - \theta_0|,
\end{aligned}$$

i.e. t_a, t_b are the closest points to θ_0 with $|h_0(t)|$ equal to ϵ. Due to continuity of $h_0(t)$, $\epsilon_1 < \epsilon_2 \implies |t_a(\epsilon_1) - \theta_0| \leq |t_a(\epsilon_2) - \theta_0|$, similarly for t_b, so that these points can only get closer to θ_0 when we decrease the deviation ϵ.

Define

$$U_{t_a,t_b}(\epsilon) := \\ \left\{ f \in \mathcal{F} : \left| \int g(x, t_a(\epsilon))(f(x) - f_0(x))\mathrm{d}x \right| < \epsilon^c \text{ AND } \left| \int g(x, t_b(\epsilon))(f(x) - f_0(x))\mathrm{d}x \right| < \epsilon^c \right\},$$

which are weak neighbourhoods of f_0 indexed by ϵ, for arbitrary c. We will set c according to ϵ, such that when $\epsilon < 1$, $c > 1$, whereas when $\epsilon \geq 1$, $c < 1$.

We have that

$$\begin{aligned}
&\epsilon < 1, c > 1, \text{ and } f \in U_{t_a,t_b}(\epsilon) \text{ OR } \epsilon > 1, c < 1, \text{ and } f \in U_{t_a,t_b}(\epsilon) \\
\Longrightarrow &\int g(x, t_a(\epsilon)) f(x)\mathrm{d}x \in (\epsilon - \epsilon^c, \epsilon + \epsilon^c) > 0 \text{ and} \\
&\quad \int g(x, t_b(\epsilon)) f(x)\mathrm{d}x \in (-\epsilon - \epsilon^c, -\epsilon + \epsilon^c) < 0 \\
\Longrightarrow &\,\theta(f) \in (\min(t_a(\epsilon), t_b(\epsilon)), \max(t_a(\epsilon), t_b(\epsilon))) \\
\Longrightarrow &\,|\theta(f) - \theta_0| < \max(|t_a(\epsilon) - \theta_0|, |t_b(\epsilon) - \theta_0|),
\end{aligned}$$

the second to last implication is due to continuity of $g(x,t)$ hence that of $\int g(x,t)f(x)\mathrm{d}x$. Define

$$e(\epsilon) = \max(|t_a(\epsilon) - \theta_0|, |t_b(\epsilon) - \theta_0|),$$

Hence

$$U_{t_a,t_b}(\epsilon) \subseteq \{f \in \mathcal{F} : |\theta(f) - \theta_0| < e(\epsilon)\},$$

i.e. $U_{t_a,t_b}(\epsilon)$ is a subset of the right-hand-side expression. Recall that both $t_a(\epsilon)$ and $t_b(\epsilon)$ $\to \theta_0$ as $\epsilon \to 0$ (monotonically), such that $e(\epsilon) \to 0$ monotonically as $\epsilon \to 0$. Then, given weak consistency of the proposal model $\mathcal{P}_\Pi$ with measure Π, $\forall \epsilon > 0$, we have,

$$\lim_{n\to\infty} \Pi\left(U_{t_a,t_b}(\epsilon) \mid X_1, X_2, \ldots, X_n\right) = 1$$
$$\implies \lim \Pi(|\theta(f) - \theta_0)| < e(\epsilon)|X_1, X_2, \ldots, X_n) = 1.$$

Hence, posterior TAB parametric inference for a parameter defined via estimating equation $g(x,t)$ that is continuous in t and integrable will be consistent if the proposal nonparametric model is at least weakly consistent for the true data-generating distribution f_0. The result above easily extends to the case where the data-generating distribution is discrete or a mixed distribution, by considering the definition of weak neighbourhoods of distribution functions which appears on p. 81 of Ghosh and Ramamoorthi (2003).

3.3.2 Parametric consistency of TAB posterior for parameters defined via estimating equations when the proposal model is the Dirichlet process

In the case that $\mathcal{P}_\Pi$ is the Dirichlet process, we have that for functional parameters defined via an estimating equations $g(x,t)$, that is,

$$\theta(F) := \left\{ t \in \Theta \quad s.t. \int g(x,t)\mathrm{d}F(x) = 0 \right\},$$

parametric consistency does not require a boundedness condition on $g(x,t)$. We have that continuity of $g(x,t)$ in t, integrability of $\int g(x,t)\mathrm{d}F(x)$ for all F in the support of Π, and integrability of $\int g(x,t)\mathrm{d}F_0(x)$ at any given t together imply the parametric consistency of the Dirichlet process proposal model.

To see this, the proof contained in Section 3.3.1 can be used, with the only modification being the definition of the neighbourhood $U_{t_a,t_b}(\epsilon)$ as

$$\begin{aligned}&U_{t_a,t_b}(\epsilon) := \\ &\left\{F \in \mathcal{F} : \left|\int g(x,t_a(\epsilon))\mathrm{d}(F(x)-F_0(x))\right| < \epsilon^c \text{ AND}\right. \\ &\qquad\qquad \left.\left|\int g(x,t_b(\epsilon))\mathrm{d}(F(x)-F_0(x))\right| < \epsilon^c\right\}.\end{aligned}$$

When $\mathcal{P}_\Pi$ is the Dirichlet process, Proposition 4.3 of Ghosal and van der Vaart (2017) implies that

$$1 = \lim_{n\to\infty} \Pi\left(\left|\int g(x,t_a(\epsilon))\mathrm{d}(F(x)-F_0(x))\right| \le \epsilon^c \mid X_1 \ldots, X_n\right) \le \lim_{n\to\infty} \Pi\left(U_{t_a,t_b}(\epsilon) \mid X_1,\ldots,X_n\right),$$

$\forall \epsilon > 0$ at arbitrary c, as long as $g(x,t)$ is a integrable function at any fixed t.

Chapter 4

Algorithms for sampling from the posterior θ-augmented probability model for Bayesian inference

If we take $F_X \sim \text{TA}(m = p_\theta/q_\theta^\Pi, \mathcal{P}_\Pi)$, then the TAB posterior for θ has the density function

$$q_{\theta|\tilde{x}_n}^{\Pi^\star}(\theta|\tilde{x}_n) \propto \frac{p_\theta(\theta)}{q_\theta^\Pi(\theta)} \cdot q_{\theta|\tilde{x}_n}^\Pi(\theta|\tilde{x}_n) \tag{4.1}$$

when the induced distribution for θ is absolutely continuous. If the induced distribution for θ is discrete, we would use the model $F_X \sim \text{TA}(m = \text{P}_\theta/\text{Q}_\theta^\Pi, \mathcal{P}_\Pi)$, which results in the TAB posterior

$$\Pi^\star(\theta = t|\tilde{x}_n) \propto \frac{\text{P}_\theta(t)}{\text{Q}_\theta^\Pi(t)} \text{Q}_{\theta|\tilde{x}_n}^\Pi(t|\tilde{x}_n),$$

We note the similarities between these two versions. The following discussion on sampling focuses on the continuous version of the posterior distribution for θ but easily generalizes to the discrete version.

We discuss two types of sampling strategies, direct sampling and Markov chain Monte Carlo (MCMC) for drawing samples from $q_{\theta|\tilde{x}_n}^{\Pi^\star}$ when our prior is the $\text{TA}(m = p_\theta/q_\theta^\Pi, \mathcal{P}_\Pi)$ model. In most scenarios, closed-form expressions for q_θ^Π and $q_{\theta|\tilde{x}_n}^\Pi$ are not known. While we can easily sample from these densities, the samples only provide us with estimates of these functions, denoted as $\hat{q}^{\Pi}_{\theta}$ and $\hat{q}^{\Pi}_{\theta|\tilde{x}_n}$, via common density estimation methods. Regardless, we describe both direct sampling and MCMC algorithms in the following sections, and provide some insights with regards to our ability to approximate posteriors given a TA($m = p_\theta/q^{\Pi}_{\theta}, \mathcal{P}_{\Pi}$) prior. Throughout this section, inference will be based on using the Normal-Inverse-Gamma (NIG) prior with normal likelihoods as $\mathcal{P}_{\Pi}$, which gives us the benefit that exact expressions for $q^{\Pi}_{\theta|\tilde{x}_n}$ and q^{Π}_{θ} are known for some functional parameters. We show that, under natural and agreeable limitations to the choice of $\mathcal{P}_{\Pi}$, differences in posterior TAB inference due to the use of $\hat{q}^{\Pi}_{\theta}$ and $\hat{q}^{\Pi}_{\theta|\tilde{x}_n}$ in place of exact expressions are negligible.

4.1 Direct sampling

The marginal posterior density of θ under the TA($m = p_\theta/q^{\Pi}_{\theta}, \mathcal{P}_{\Pi}$) model is proportional to a multiplication of three terms, shown in Eqn. (4.1). If we have the expressions for these three components, direct sampling of $\Pi^{\star}$ will be possible based on either a grid approximation or the inverse cumulative distribution function (CDF) method. Grid approximation is the subject of the following study as it is quick to perform in low-dimensional space, and works well given enough grid points. Under the scheme of grid approximation, we will study the conditions that affect our ability to produce good approximations to TAB posterior densities based on a comparison of $q^{\Pi}_{\theta|\tilde{x}_n}/q^{\Pi}_{\theta}$, the effective likelihood, and $\hat{q}^{\Pi}_{\theta|\tilde{x}_n}/\hat{q}^{\Pi}_{\theta}$, the estimated effective likelihood, as this allows us to make comparisons without specifying p_θ.

Suppose that we are interested in estimating the mean parameter and the proposal model is one which assumes a normal sampling distribution and an NIG prior. In this case exact expressions for q^{Π}_{θ} and $q^{\Pi}_{\theta|\tilde{x}_n}$ are known to be the density functions of t-distributions. The density functions $\hat{q}^{\Pi}_{\theta}$ and $\hat{q}^{\Pi}_{\theta|\tilde{x}_n}$ are estimated separately, via the R package **ks** (Duong, 2020) with 1×10^6 samples each. Samples over a fine grid with sampling probability proportional to an (approximate) effective likelihood function were obtained to visualize this function. Several hyperparameter settings for the proposal model $\mathcal{P}_\Pi$ were tested to identify the conditions under which the use of estimated $\hat{q}_\theta^\Pi$ and $\hat{q}_{\theta|\tilde{x}_n}^\Pi$ led to good approximation of the exact effective likelihoods. In this study, we used the same dataset of 10 observations drawn from a Gaussian distribution, and the resulting effective likelihoods for the mean parameter are shown in Figures 4.1 and 4.2.

Figures 4.1 and 4.2 show, with each row corresponding to a particular hyperparameter setting for the TAB proposal model, the estimated and exact versions of the effective likelihood (right-hand-side panels), along with the corresponding $q_{\theta|\tilde{x}_n}^\Pi$ and q_θ^Π (right-hand-side panels). When the proposal prior q_θ^Π had heavier tails than the proposal posterior $q_{\theta|\tilde{x}_n}^\Pi$, the approximate effective likelihood accurately tracked the exact expression. When $q_{\theta|\tilde{x}_n}^\Pi$ had fatter tails than q_θ^Π, or was situated in the tail of q_θ^Π, then the approximate version of the effective likelihood was quite inaccurate; this is to be expected since the approximate effective likelihood is calculated based on a *division* by the estimated $\hat{q}_\theta^\Pi$ therefore inherently unstable for small values of $\hat{q}_\theta^\Pi$ which are estimated less accurately to begin with. This result can be translated to the approximation of the posterior density function of θ under the TA$(m = p_\theta/q_\theta^\Pi, \mathcal{P}_\Pi)$ model; that is, due to the same limitations in density estimation in tail regions, we require q_θ^Π to be heavier tailed than $q_{\theta|\tilde{x}_n}^\Pi \cdot p_\theta$ in order for posterior inference based on a TA model with an estimated weighting function to be a good approximation to inference based on an exact weighting function.

Note that, in the third row of Figure 4.2, the effective likelihood has a peculiar ("M") shape. We are reminded by this peculiar example to keep in mind the scale of q_θ^Π when interpreting an effective likelihood. In this case, the proposal q_θ^Π is extremely precise, and as a result of the integrability condition of weighting function p_θ/q_θ^Π, this particular proposal model can only be appropriate if our subjective prior p_θ is even more precise. Coupled with a p_θ of high precision, bi-modality of the effective likelihood does not translate to the TAB posterior inference.

In the case of joint inference of multivariate θ the situation is similar. The R library **ks** provides density estimation of a multivariate random variable for up to 6 dimensions.

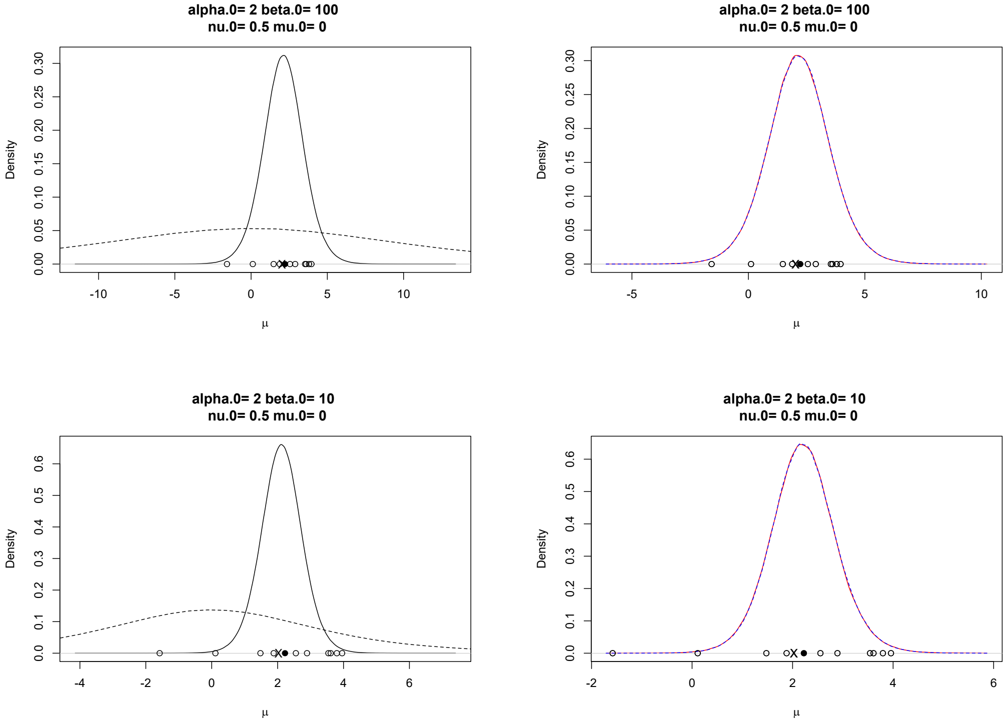
However, to estimate multivariate density well enough via grid approximation requires increasingly more samples as the number of dimensions increases. We conducted a simulation, again, using the Gaussian likelihood and a NIG prior as the proposal model. The target of inference was the vector of mean and variance parameters. Figure 4.3a shows the estimated contours of the effective likelihood superimposed on the exact contours for the same data set as in Figure 4.1. The estimated contours of the likelihood function was generated based on 10^6 samples from $\hat{q}_{\theta}^{\Pi}$ and 10^7 samples from $\hat{q}_{\theta|\tilde{x}_n}^{\Pi}$ via grid approximation over a $(8 \times 10^2) \times (8 \times 10^2)$ rectangular grid between $\mu \in [-5, 10]$ and $\sigma^2 \in [10^{-5}, 80]$. Inaccuracies in the estimated contours of the effective likelihood can be visually detected in the region where $q_{\theta|\tilde{x}_n}^{\Pi}$ was high yet q_{θ}^{Π} was low (bottom centre of Figure 4.3a). In comparison, in Figure 4.3b, when the q_{θ}^{Π} was more spread-out than the$q_{\theta|\tilde{x}_n}^{\Pi}$ in every direction, the estimated contour of the effective likelihood function matched the exact contours very well with just 10^6 samples from each of $\hat{q}_{\theta}^{\Pi}$ and $\hat{q}_{\theta|\tilde{x}_n}^{\Pi}$.

We note that the 95th percent contour of the effective likelihood and that of $q_{\theta|\tilde{x}_n}^{\Pi}$ occupied only a moderate portion of the the region $(-5, 10) \times (0, 80)$ in the parameter space, which suggests that sampling of the TAB posterior $q_{\theta|\tilde{x}_n}^{\Pi^\star}$ via grid approximation is likely inefficient in high-dimensional space. An alternative approach in this case is to approximate the TAB posterior $q_{\theta|\tilde{x}_n}^{\Pi^\star}$ via Markov chain Monte Carlo sampling; see Section 4.2.

In summary, our results show that direct sampling via grid approximation from the posterior $q_{\theta|\tilde{x}_n}^{\Pi^\star}$ or the effective likelihood works well when the distribution q_{θ}^{Π} induced by a proposal measure Π is well dispersed compared to $q_{\theta|\tilde{x}_n}^{\Pi^\star}$ and $q_{\theta|\tilde{x}_n}^{\Pi}$.

4.2 Markov chain Monte Carlo sampling

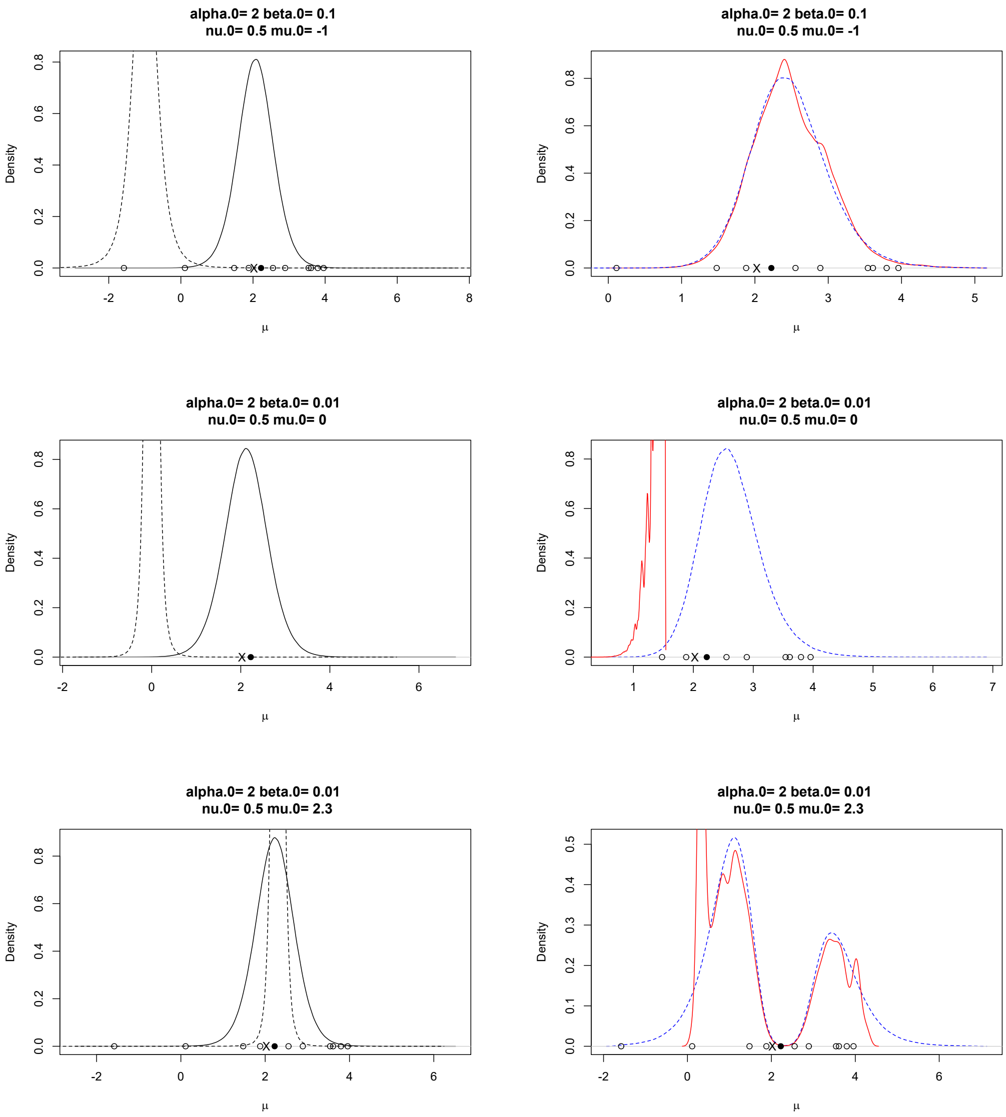
Again, assume that $F_X \sim \text{TA}(m = p_\theta / q_{\theta}^{\Pi}, \mathcal{P}_{\Pi})$ *a priori*. A simple Metropolis-Hastings (MH) algorithm to sample from the TAB posterior proposes moves in the space of F_X based on the conditional measure $\Pi(\cdot|\tilde{x}_n)$ of a given proposal model. In the case that independent sampling of F_X from $\Pi(\cdot|\tilde{x}_n)$ is possible, the acceptance ratio of a move from F_X to F_X' is



Prior and posterior

Effective lIkelihood

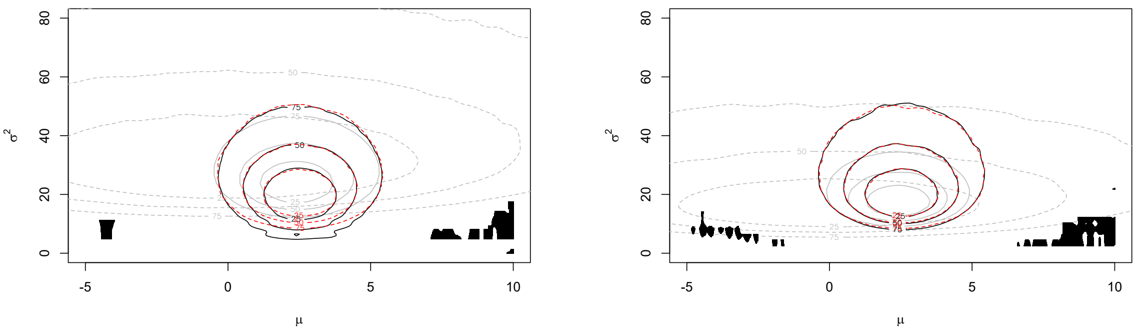
Figure 4.1: Approximating the effective likelihood under a NIG model. Estimated effective likelihoods were accurate as long as $q^{\Pi}_{\theta|\tilde{x}_n}$ was not in the tail region of q^{Π}_{θ}. Each row in the figure corresponds to a particular specification for $\mathcal{P}_{\Pi}$, with parameters given above each subplot. The left-hand-side plot within a row show q^{Π}_{θ} (dotted lines) and $q^{\Pi}_{\theta|\tilde{x}_n}$ (solid lines) of a proposal model. The right-hand-side plot within a row show the effective likelihoods obtained two ways, either estimated (red line) or exact (blue line).



Prior and Posterior

Effective Likelihood

Figure 4.2: Approximating the effective likelihood under a NIG model. Estimated effective likelihoods were inaccurate whenever $q^{\Pi}_{\theta|\tilde{x}_n}$ was in the tail region of the q^{Π}_{θ}. Each row in the figure corresponds to a particular specification for $\mathcal{P}_{\Pi}$, with parameters given above each subplot. The left-hand-side plot within a row show q^{Π}_{θ} (dotted lines) and $q^{\Pi}_{\theta|\tilde{x}_n}$ (solid lines) of a proposal model. The right-hand-side plot within a row show the effective likelihoods obtained two ways, either estimated (red line) or exact (blue line).



(a) proposal posterior in tail region of prior

(b) proposal prior more dispersed

Figure 4.3: Contours of estimated likelihood functions (black solid lines) by grid approximation, superimposed over the exact likelihood functions (red solid lines). Grid points with invalid estimates of a likelihood function are blacked out in the figure. Grey solid lines show the contours of estimated $\hat{q}^{\Pi}_{\theta|\tilde{x}_n}$, and grey dotted lines show that of estimated $\hat{q}^{\Pi}_{\theta}$. The left-hand-side panel shows a situation where grid approximation of the likelihood function was inaccurate in a low-density region of q^{Π}_{θ}. The right-hand-side panel shows a situation where grid approximation of the likelihood function worked well in general, when q^{Π}_{θ} is more dispersed than $q^{\Pi}_{\theta|\tilde{x}_n}$. The estimated effective likelihoods were generated based on a minimum of 10^6 samples from each of $q^{\Pi}_{\theta|\tilde{x}_n}$ and q^{Π}_{θ} .

$$\begin{aligned} A(F'_X, F_X) &= \min\left(1, \frac{\Pi^\star(F'_X|\tilde{x}_n)}{\Pi^\star(F_X|\tilde{x}_n)} \cdot \frac{\Pi(F_X|\tilde{x}_n)}{\Pi(F'_X|\tilde{x}_n)}\right) \\ &= \min\left(1, \frac{p_\theta(\theta(F'_X))/q^\Pi_\theta(\theta(F'_X))\Pi(F'_X|\tilde{x}_n)}{p_\theta(\theta(F_X))/q^\Pi_\theta(\theta(F_X))\Pi(F_X|\tilde{x}_n)} \cdot \frac{\Pi(F_X|\tilde{x}_n)}{\Pi(F'_X|\tilde{x}_n)}\right) \\ &= \min\left(1, \frac{p_\theta(\theta(F'_X))/q^\Pi_\theta(\theta(F'_X))}{p_\theta(\theta(F_X))/q^\Pi_\theta(\theta(F_X))}\right). \end{aligned}$$

In our algorithm the MCMC proposal distribution will, in most cases, be wider than the MCMC target distribution as the weighting function p_θ/q^Π_θ is necessarily integrable. However, the algorithm may be inefficient if the MCMC proposal distribution is too dispersed, or if $q^{\Pi^\star}{}_{\theta|\tilde{x}_n}$ is concentrated in a region of $q^\Pi_{\theta|\tilde{x}_n}$ with low density.

In most situations, a substitution of $\hat{q}^\Pi_\theta$ for q^Π_θ in the acceptance ratio is necessary. The accuracy with which $p_\theta/\hat{q}^\Pi_\theta$ tracks p_θ/q^Π_θ is low in the tail regions of q^Π_θ, due to a lack of accuracy in density estimation for tail regions. We can general avoid problems if the proposal measure Π is selected such that $q^\Pi_{\theta|\tilde{x}_n}$ is not located near the tail of q^Π_θ – a condition that can be checked visually for low-dimensional θ.

Figure 4.4 shows the contours of various target posteriors estimated with 10^6 MCMC samples, with marginal priors and posteriors of TAB inference (p_θ and $q^{\Pi^\star}_{\theta|\tilde{x}_n}$) shown in blue, and marginal priors and posteriors of the proposal measure (q^Π_θ and $q^\Pi_{\theta|\tilde{x}_n}$) shown in grey. As expected, we observed that sampling efficiency depends how well $q^\Pi_{\theta|\tilde{x}_n}$ resembles $q^{\Pi^\star}_{\theta|\tilde{x}_n}$, but regardless of efficiency, the MCMC approximation of the target posteriors were quite accurate in all cases tested.

It is possible to use other MCMC proposal distributions, but the advantage of using exactly $\Pi(\cdot|\tilde{x}_n)$ as the MCMC proposal is that the acceptance ratio involves only values of θ, thus we need not keep track of the full F_X when running the Markov chain.

To run any MCMC we require a proper probability density as the target distribution. Therefore, when implementing the algorithm it is necessary to assume a particular p_θ, and check that it satisfies the integrability condition of p_θ/q^Π_θ. The integrability condition may be difficult to verify when θ is of high dimensionality. One workaround is to match partial subjective information for a single dimension of θ, for which the weighting func-

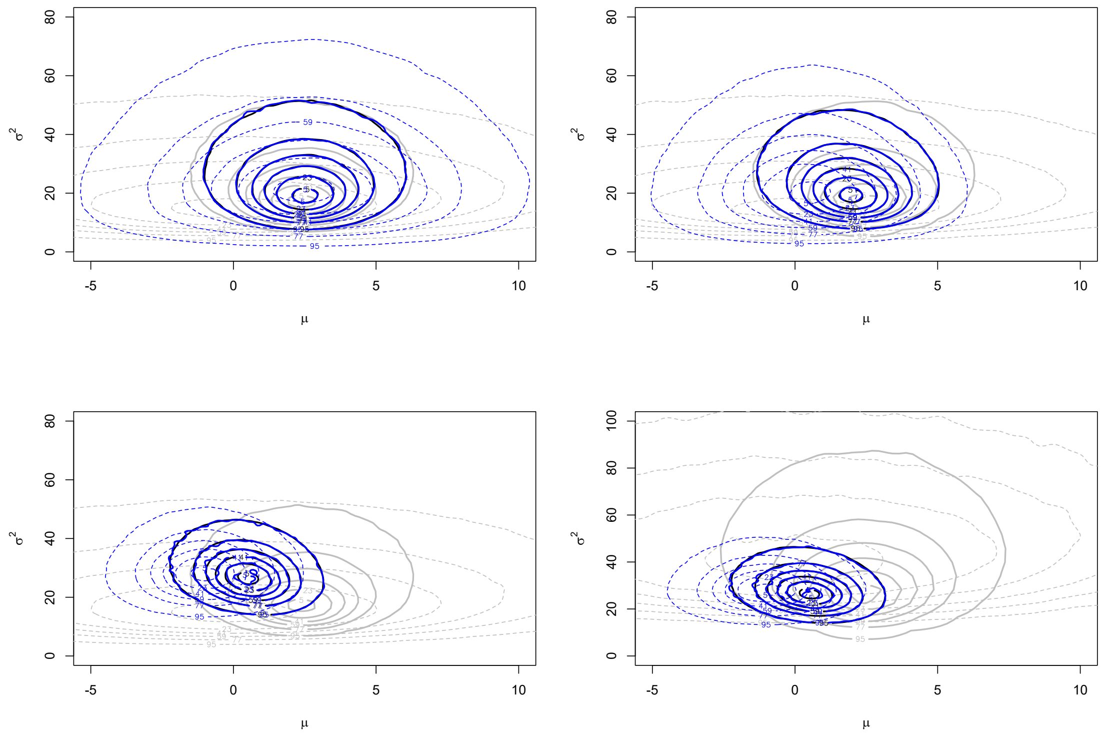


Figure 4.4: Two-dimensional joint inference of $\theta = (\mu, \sigma^2)$ based on MCMC approximation (10^6 samples) to $q^{\Pi\star}_{\theta|\tilde{x}_n}$. The panels represent inference under θ-augmented models that differ in p_θ but share the same $\mathcal{P}_\Pi$. In each panel, p_θ is shown with blue dotted lines, and approximated $q^{\Pi\star}_{\theta|\tilde{x}_n}$ is shown with blue solid lines. The exact contours of the target posterior distributions are shown with black solid lines. Contours relating to $\mathcal{P}_\Pi$ are shown in grey (dotted= prior, solid= posterior). With the same number of MCMC iterations for approximating the target posterior, the estimated target posteriors (blue solid lines) appear slightly smoother in the top row of figures, due to the proposal posteriors being a lot more dispersed than the target posteriors.

tion is $m = p_{\theta_j}/q^{\Pi}_{\theta_j}$, as per Eqn. (3.11). This partial specification allows us to check the integrability condition based on marginal distributions in 1 or 2 dimensions which can usually be done visually.

Having an effective likelihood function is useful for situations when we want to try out various subjective priors without having to perform another MCMC. Although it may be possible to obtain an approximate effective likelihood based on a division of the estimated density of the TAB posterior by the exact p_θ used in the MCMC, estimation of the effective likelihood function in the tail region of the TAB posterior is challenging, as seen in Figure 4.5. Even though an estimate of the effective likelihood function can be regarded as a by-product of having performed a density estimation for the TAB posterior, we are likely better off obtaining the effective likelihood directly by the calculation of $q^{\Pi}_{\theta|\tilde{x}_n}/q^{\Pi}_{\theta}$.

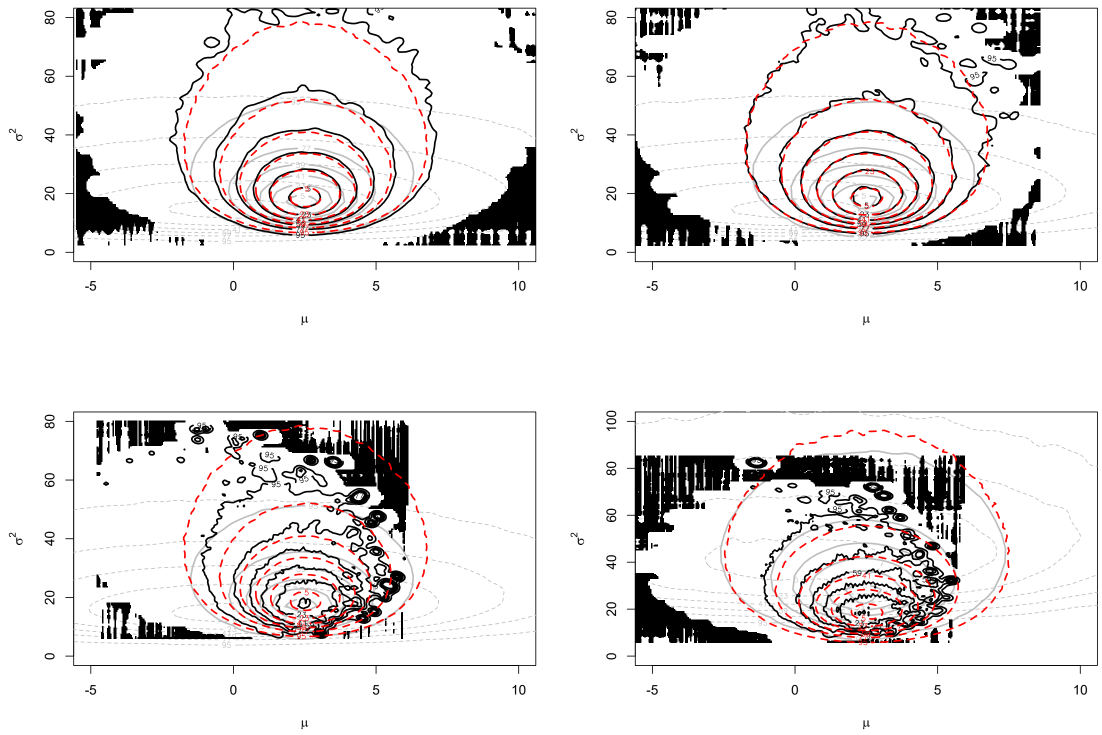


Figure 4.5: Contours of effective likelihoods estimated based on MCMC approximations. The posterior $q^{\Pi\star}_{\theta|\tilde{x}_n}$ was approximated with 10^6 MCMC samples. Subplots present results from different θ-augmented models, corresponding to respective panels of Figure 4.4. Contours relating to proposal model are shown in grey (dotted= q^{Π}_{θ}, solid= $q^{\Pi}_{\theta|\tilde{x}_n}$). Contours of effective likelihoods obtained via $\hat{q}^{\Pi\star}_{\theta|\tilde{x}_n}/p_\theta$ are shown in black, while those of exact effective likelihoods are shown in red. Visualization of the contours is done via grid sampling. Locations where $\hat{q}^{\Pi\star}_{\theta|\tilde{x}_n}/p_\theta$ resulted in invalid values are blacked out in the plots. Estimation of the effective likelihood based on a division of MCMC estimated target posterior distribution by the subjective prior appear rather unstable.

Chapter 5

Proposal models for θ-augmented Bayesian semiparametric inference

Although a θ-augmented model may be specified with either a parametric or nonparametric proposal model, we focus on nonparametric proposal models in this thesis. Employing a nonparametric proposal model in the specification of a θ-augmented model effectively produces semiparametric Bayesian inference.

The theory of θ-augmentation limits compatible proposal models to ones that are dominated. One well-studied class of dominated nonparametric models is the Dirichlet process mixture (DPM) model. The DPM can be represented as an infinite mixture of kernels. It is typically implemented with Gaussian kernels for ease of computation. In the limit with kernel variance approaching 0, we recover the Dirichlet process model. Thus the DP may be represented as an infinite mixture of kernels, but with kernels that are Dirac measures. One needs to be cautious as, in general, the DP is not dominated and violates the prerequisite of TAB inference. A simple workaround is to instead consider the data as measured to finite precision which results in a dominated DP with a discrete base distribution. As we will demonstrate, the practical difference induced by rounding of observed data is minimal in finite sample inference. Furthermore, limitations in computer architecture leads to most mathematical calculations being performed with finite precision arithmetic in practice.

Depending on the type of proposal model, the difficulty of calculating $\theta(F_X)$ will vary. In general, calculation of $\theta(F_X)$ will be simple when F_X is discrete. If we take $\mathcal{P}_\Pi$ to be a DP, discreteness of F_X is guaranteed, and the sampling algorithm will be simpler to implement than that of DPM. In this section, we aim to compare the use of DP versus DPM as the proposal model for inferring various functional parameters. We will discuss differences in the resulting TAB posterior and in implementation difficulty.

Without loss of generality, we describe the candidate proposal models for an observable $X \in \mathbb{R}$ in Model 5.1 and Model 5.2. These models can be extended to an observable vector in higher dimensions while maintaining the same overall structure.

Model 5.1. (Dirichlet process mixture model with absolutely continuous kernels)

Assume $\mathcal{K}(x|\eta)$ is an absolutely continuous density function for x,

$$\begin{aligned} f_X(x) &= \int \mathcal{K}(x|\eta)\mathrm{d}F_H(\eta) \\ F_H &\sim \text{DP}(\phi, G_0). \end{aligned}$$

Model 5.2. (DP with discrete base distribution)

For a fixed bin-width h,

$$\begin{aligned} F_X &\sim \text{DP}(\phi, G_0') \\ G_0' &= \text{Discretized version of continuous distribution } G_0, \\ &\quad \text{with mass assigned to points } \{ih | i \in \mathbb{Z}\} \text{ for fixed bin width } h \text{ s.t.} \\ G_0'(X = ih) &= \int \mathbb{1}\left[x \in \left(ih - \frac{h}{2}, ih + \frac{h}{2}\right)\right] \mathrm{d}G_0(x). \end{aligned}$$

5.1 Sampling of random measures from the proposal model

Both DP and DPM models rely on the Dirichlet process at some level. In order to sample from the prior Dirichlet process, we may use the stick-breaking construction (Sethuraman, 1994). Sampling from the prior $\mathcal{P}_\Pi$, regardless of it being DP or DPM, is relatively straightforward as no data is involved. In practice one terminates the stick-breaking process (SBP) after a sufficient number of components have been realized, resulting in an approximation of a random measure from $\mathcal{P}_{\Pi}$; see Ishwaran and James (2001) for some algorithms. To obtain good approximations, the level of truncation of the stick-breaking process should depend on the precision ϕ. When $\phi < 1$ we can typically terminate the SBP fairly quickly.

Under a DP prior, posterior distribution of F_X is again a Dirichlet process, i.e. Model 5.2, but updated with a new base distribution,

$$F_X|\tilde{x}_n \sim \text{DP}\left((\phi+n), \left(\frac{\phi}{\phi+n}G_0' + \frac{n}{\phi+n}\hat{F}_n\right)\right),$$

where $\hat{F}_n$ denotes the empirical distribution of the observed data and n denotes sample size. We may draw approximately from the posterior DP via the truncated SBP. While this is conceptually simple, there are some ways to make the algorithm efficient, as the posterior precision parameter $(\phi + n)$ can be large. For an efficient algorithm to sample approximately from the posterior DP, we may extend the results of Lemma 3 of Paisley and Jordan (2016), which also appears in Paisley et al. (2010); see Appendix A.3.

Sampling of $F_X|\tilde{x}_n$ under a DPM prior, however, is more computationally involved than under a DP prior. Dirichlet process mixture models, especially with Gaussian kernels, are well studied in the literature. Section 23.3 of Gelman et al. (2013) gives an extensive account of the DPM with Gaussian kernels, and a blocked Gibbs Sampler (p. 553) for generating approximate samples from the posterior DPM. The kernels are continuous densities supported on the real line, hence membership of a data point in each mixture component is not certain. A blocked Gibbs sampler alternates between the sampling of cluster membership given data and kernel parameters, and subsequently the sampling of kernel parameters given cluster membership and data. A slice sampler (Kalli et al., 2011) can be used to generate exact samples from a posterior DPM. However, mixing problems can be extreme under the fringe case where kernels of the DPM resemble degenerate distributions with practically no chance of clustering. This is a highly unrealistic model for density estimation purposes, but will be explored here to present the continuum of inferences possible between DPM and DP. Label-switching moves can alleviate the problem of poor mixing in this extreme scenario, or alternatively, an approximate algorithm built on the Polya urn representation of DP; see Appendix A.4.

5.1.1 Estimating the mean and variance

Calculating the mean functional is an example where the integral of the estimating equation presented in Example 1.1 has a simple closed form expression for F_X sampled from either a DP or DPM; the same goes for moments in general. Every random measure sampled from a DP or DPM model can be expressed as an infinite weighted sum of kernels, denoted with K_i of unspecified form. Therefore

$$\begin{aligned}\mu(F_X) &= \int x \mathrm{d}F_X(x) \\ &= \int x \sum_{i=1}^{\infty} w_i K_i(x) \mathrm{d}x \\ &= \sum_{i=1}^{\infty} w_i \int x K_i(x) \mathrm{d}x \\ &= \sum_{i=1}^{\infty} w_i \mathbb{E}_{K_i}[X],\end{aligned}$$

when the above integral is finite (Fubini's Theorem). The expected value of kernel K_i can be obtained easily for common kernels because the first moment of a common distribution is typically well known. The variance of F_X can be derived similarly due it being $\sigma^2(F_X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Example

We performed a simulation study to highlight the effects of varying proposal model specification on the posterior inference for mean and variance parameters. We obtained marginal posterior inference for the mean and variance functionals separately. When inferring the mean, we used the model

$$\text{TA}(m = p_\mu / q_\mu^\Pi, \mathcal{P}_\Pi)$$

as the Bayesian prior. Whereas, for inferring variance, we used the model

$$\mathrm{TA}(m = p_\sigma^2 / q_{\sigma^2}^{\Pi}, \mathcal{P}_\Pi)$$

as the Bayesian prior. The data was distributed according to a skewed distribution, detailed in Model 5.3. The density function for the observable is shown in Figure 5.1. We drew 20 samples from Model 5.3. The data points are represented as open circles in Figure 5.3.

Model 5.3. (Data-generating mechanism of a skewed random observable, Section 5.1.1)

$$\begin{aligned}
T_1 &\sim \text{Normal}(\mu = 5, \sigma^2 = 5) \\
T_2 &\sim \text{Normal}(\mu = -1, \sigma^2 = 1) \\
L &\sim \text{Bernoulli}(0.3) \\
X &= LT_1 + (1 - L)T_2.
\end{aligned}$$

Model 5.4. (DPM proposal model for estimating mean and variance, Section 5.1.1)

$$\begin{aligned}
&f_X(x) = \int N(x|\mu, \sigma^2)\mathrm{d}F_H(\mu, \sigma^2) \\
&F_H \sim \text{DP}(\phi, G_0) \\
&G_0 = \text{NIG}(\mu_0, \lambda_0, \alpha_0, \beta_0),
\end{aligned}$$

where $N(x|\mu, \sigma^2)$ denotes the Gaussian density function.

Model 5.5. (DP proposal model for estimating mean and variance, Section 5.1.1)

$$\begin{aligned}
F_X \sim& \text{DP}(\phi, G_0') \\
G_0' =& \text{ Discretized version of continuous distribution } G_0\text{,} \\
& \quad \text{with mass assigned to points } \{ih | i \in \mathbb{Z}\} \text{ for fixed bin-width } h \text{ s.t.} \\
& \quad G_0'(X = ih) = \int \mathbb{1}\left[x \in \left(ih - \frac{h}{2}, ih + \frac{h}{2}\right)\right] \mathrm{d}G_0(x) \\
G_0 =& \text{ Normal}(\mu_0, \sigma_0^2).
\end{aligned}$$

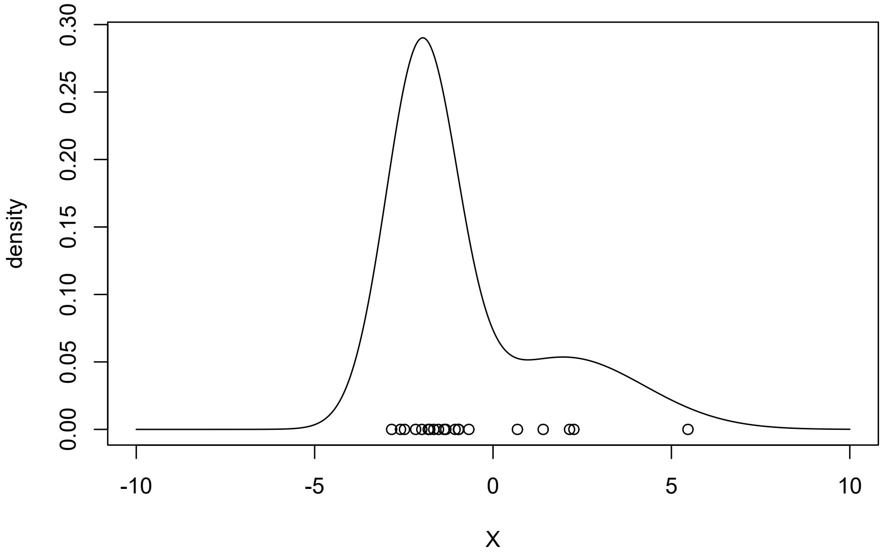


Figure 5.1: Density function of Model 5.3, the data-generating distribution for the example in Section 5.1.1. The dataset that was used in Section 5.1.1 is shown as open circles on the x-axis.

Table 5.1: Proposal model hyperparameters for the example in Section 5.1.1. The hyperparameters are tested in combination for estimation of mean and variance parameters. The proposal DPM and DP models are described in detail in Model 5.4 and Model 5.5.

DPM Model 5.4	**DP Model 5.5**
precision $\phi \in \{0.3, 0.75, 1.5\}$	precision $\phi \in \{0.3, 0.75, 1.5\}$
base measure $G_0 \in \{G_{0,1}, G_{0,2}, G_{0,3}\}$ $G_{0,1} =$ $\text{NIG}(\mu_0 = 0, \lambda_0 = 1.25 \times 10^{-4}, \alpha_0 = 3, \beta_0 = 1 \times 10^{-1})$ $G_{0,2} =$ $\text{NIG}(\mu_0 = 1, \lambda_0 = 1.25 \times 10^{-8}, \alpha_0 = 3, \beta_0 = 1 \times 10^{-5})$ $G_{0,3} =$ $\text{NIG}(\mu_0 = 1, \lambda_0 = 5 \times 10^{-8}, \alpha_0 = 3, \beta_0 = 1 \times 10^{-5})$	base measure prior to discretization $G_0 \in \{G_{0,1}, G_{0,2}\}$ $G_{0,1} = \text{Normal}(\mu_0 = 0, \sigma_0^2 = 20^2)$ $G_{0,2} = \text{Normal}(\mu_0 = 1, \sigma_0^2 = 10^2)$
	discretization precision $h \in \{h_1, h_2\}$ $h_1 = 1 \times 10^{-1}$ $h_2 = 1 \times 10^{-4}$

Two types of proposal models were employed, one of which is a DPM model (Model 5.4), and the other a DP, (Model 5.5). Several values of hyperparameters for the proposal models were used in combination to examine the effect of hyperparameter specification on posterior TAB inference. The selected hyperparameter values are detailed in Table 5.1. Examples of random measures we drew from posterior DPMs with varying hyperparameter specifications are shown in Figure 5.2.

The results are presented in terms of effective likelihoods, $q^{\Pi}_{\mu|\tilde{x}_n}/q^{\Pi}_{\mu}$ and $q^{\Pi}_{\sigma^2|\tilde{x}_n}/q^{\Pi}_{\sigma^2}$, which were estimated via kernel density estimation using the R package `ks`. The results based on using DPM as $\mathcal{P}_{\Pi}$ are summarized in Figures 5.3 and 5.5. The results from using DP as $\mathcal{P}_{\Pi}$ are summarized in Figures 5.4 and 5.6. We also provide the BB posterior in these Figures as a point of reference.

Algorithmically, the DP was much easier to implement and sample from. We found

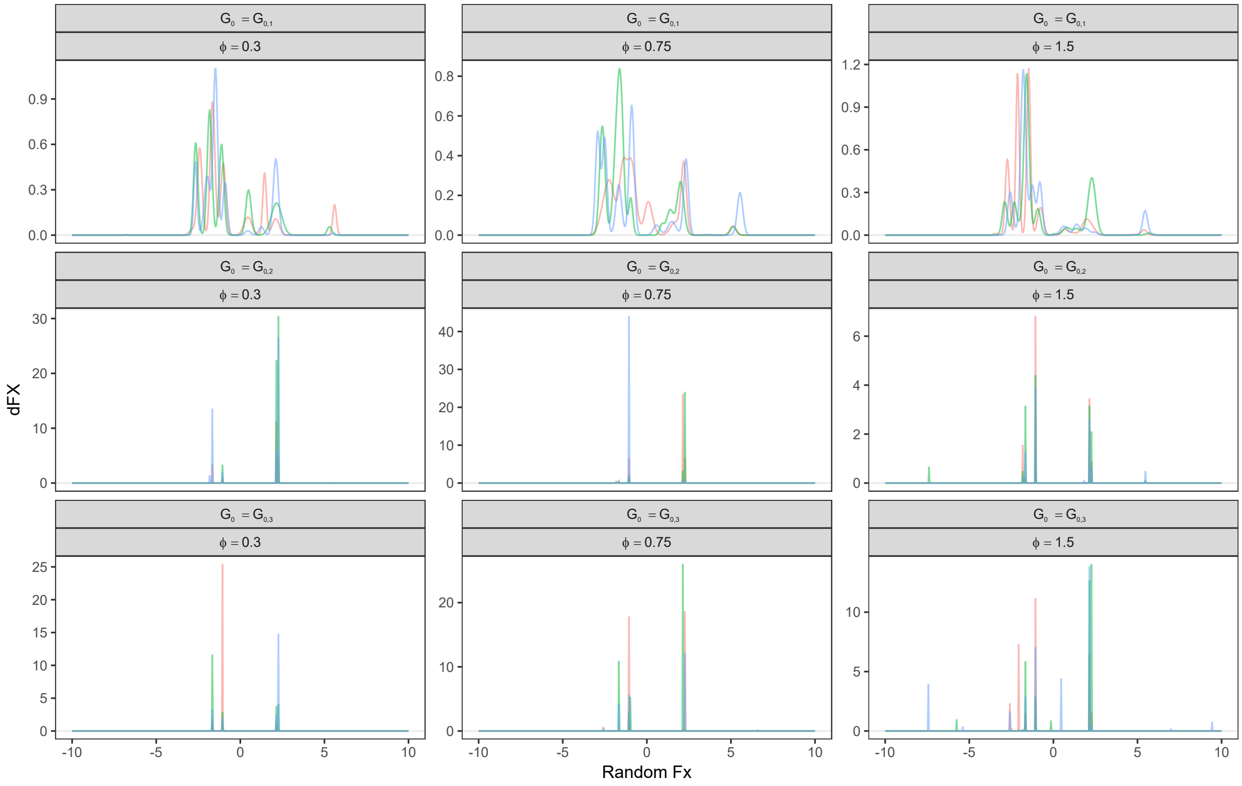


Figure 5.2: Random draws of F_X conditional on the data given Model 5.4, a DPM of Gaussian kernels, Section 5.1.1. The subplots correspond to various hyperparameter specifications for the same model, which are given above each subplot. Within a subplot, the samples of F_X from the same posterior DPM are differentiated by colour. Detailed hyperparameter specification is found in Table 5.1.

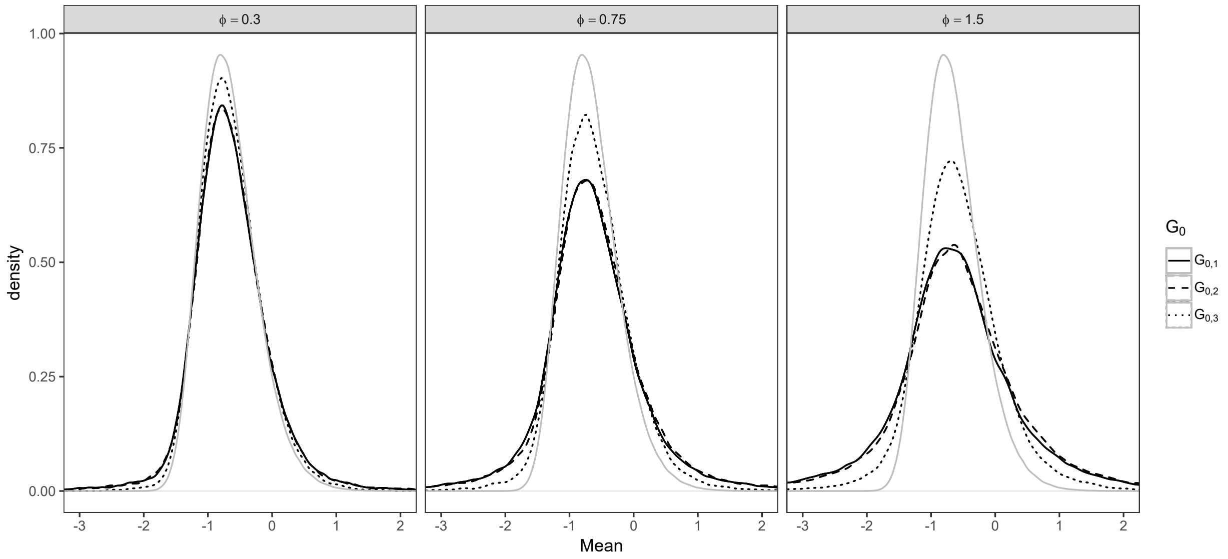


Figure 5.3: TAB effective likelihood curves for inferring the mean with $\mathcal{P}_\Pi$ being Model 5.4, a DPM of Gaussian kernels, Section 5.1.1. Black curves represent effective likelihoods obtained under various hyperparameter specifications, the line types of which correspond to different base distributions. The three panels are arranged in the order of increasing DPM precision parameter from left to right. The Bayesian bootstrap posterior is shown in grey in each panel for comparison. Note that a decreased ϕ resulted in an increased resemblance of the effective likelihood to the BB posterior. Within each panel, the curve corresponding to the model with $G_{0,3}$ as the base distribution is the most similar to the BB posterior, due to the induced variance of $q^{\Pi}_{\mu|\tilde{x}_n}$ being the smallest. Detailed hyperparameter specification is found in Table 5.1.

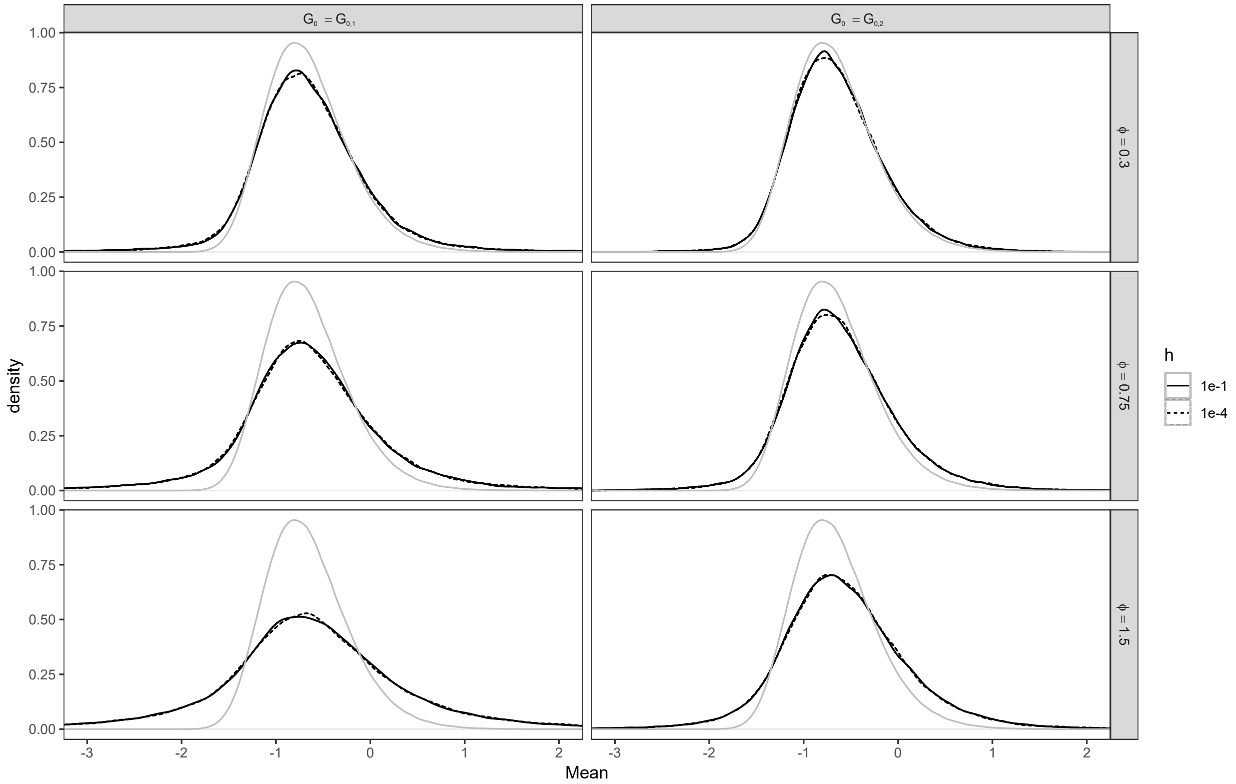


Figure 5.4: TAB effective likelihood curves for inferring the mean with $\mathcal{P}_\Pi$ being Model 5.5, a DP with discrete base distribution, Section 5.1.1. Black curves represent the effective likelihoods obtained under various hyperparameter specifications. The rows of subplots are arranged such that the precision parameter ϕ increases from top to bottom, while the columns of subplots represent different base distributions. Bandwidth of discretization, as controlled by parameter h, is indicated by the line type, which seems to have little effect on finite sample parametric inference. The Bayesian bootstrap posterior is shown in grey in each panel for comparison. Details regarding hyperparameter specification is given in Table 5.1. Since $G_{0,2}$ has a smaller variance than $G_{0,1}$, it seems that decreasing the spread of the base distribution increased the resemblance of TAB effective likelihood to the Bayesian bootstrap posterior. Decreasing ϕ also led to increased resemblance of the effective likelihoods to the BB.

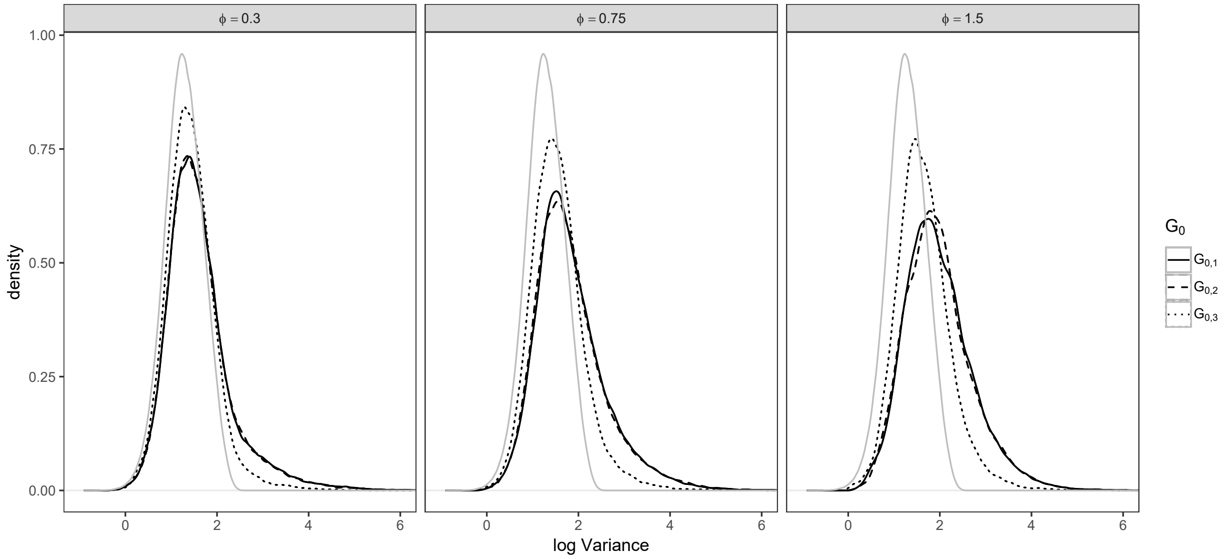


Figure 5.5: TAB effective likelihood curves for inferring the log of variance with $\mathcal{P}_{\Pi}$ being Model 5.4, a DPM of Gaussian kernels, Section 5.1.1. Black curves represent effective likelihoods obtained under various hyperparameter specifications, the line types of which correspond to different base distributions. The three panels are arranged in the order of increasing DPM precision parameter from left to right. The Bayesian bootstrap posterior is shown in grey in each panel for comparison. Note that a decreased ϕ resulted in an increased resemblance of the effective likelihood to the BB posterior. Within each panel, the curve corresponding to the model with $G_{0,3}$ as the base distribution is the most similar to the BB posterior, due to the induced variance of $q^{\Pi}_{\log\sigma^2|\tilde{x}_n}$ being the smallest. Detailed hyperparameter specification is found in Table 5.1.

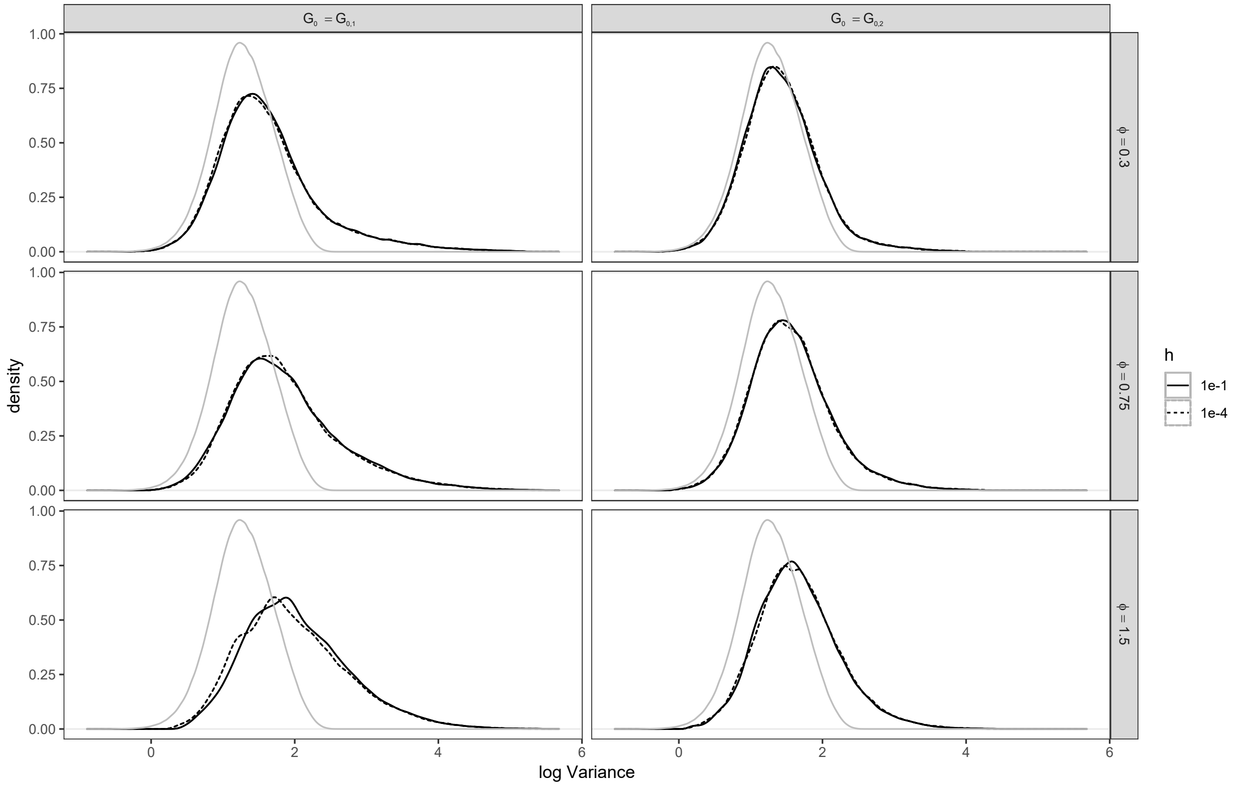


Figure 5.6: TAB effective likelihood curves for inferring the log of variance with $\mathcal{P}_\Pi$ being Model 5.5, a DP with discrete base distribution, Section 5.1.1. Black curves represent the effective likelihoods obtained under various hyperparameter specifications. The rows of subplots are arranged such that the precision parameter ϕ increases from top to bottom, while the columns of subplots represent different base distributions. Bandwidth of discretization, as controlled by parameter h, is indicated by the line type, which seems to have little effect on finite sample parametric inference. The Bayesian bootstrap posterior is shown in grey in each panel for comparison. Details regarding hyperparameter specification is given in Table 5.1. Decreasing ϕ and decreasing the spread of the base distribution both increased the resemblance of TAB effective likelihood to the Bayesian bootstrap posterior.

the level of precision, h, of the observable had little effect on the posterior inference. We found that the effective likelihoods became more and more similar to the Bayesian bootstrap when the value of ϕ was reduced, regardless if we used DP or DPM as the proposal model. Reduction in the kernel variance of DPM proposal models also tuned the effective likelihoods toward the BB. As the BB is known to have good properties, selecting ϕ as small as computationally feasible may be a good starting point for TAB inference unless we suspect a severe under-coverage of BB credibility intervals.

5.1.2 Linear least squares

Without loss of generality, let $X = (1, X_1)^\top$, $Y \in \mathbb{R}$. We adopt the functional defined in Example 1.3 for estimating the coefficients of linear regression, $\beta := (\beta_0, \beta_1)^\top$.

For any random measure F_{XY} in the sample space of infinite kernel mixture models based on the DP, β can be found by solving

$$\begin{aligned} 0 =& \mathbb{E}[XY - XX^\top \beta] \\ =& \mathbb{E}[XY] - \mathbb{E}[XX^\top]\beta, \end{aligned}$$

which leads to

$$\beta = \mathbb{E}[XX^\top]^{-1}\mathbb{E}[XY] = \left(\sum_{j=1}^{\infty} w_j \mathbb{E}_j[XX^\top]\right)^{-1} \left(\sum_{i=1}^{\infty} w_i \mathbb{E}_i[XY]\right). \tag{5.1}$$

The solution requires calculation of cross moments of the random variables under the chosen multivariate kernel, which may be nontrivial. When $\mathcal{P}_\Pi$ is a DP, calculation of cross moments is simple, since the kernel is a product of Dirac measures $\delta_{X_k}(x)\delta_{Y_k}(y)$ for the k-th mixture component.

When the multivariate kernels are continuous in every dimension, one can potentially simplify the calculation in Eqn. (5.1) if the kernel are jointly independent in X and Y, that is, $\mathcal{K}(x, y|\eta) = \mathcal{K}_X(x|\eta_X)\mathcal{K}_Y(y|\eta_Y)$. For a Dirichlet process mixture of multivariate Gaussian kernels with diagonal covariance matrices, its asymptotic behaviour and the conditions for consistency were studied by Wu and Ghosal (2010). However, we were unable to find results on the asymptotic behaviour of other types of jointly independent multivariate kernels. In general, if jointly independent multivariate Gaussian kernels suffice for the application, $\beta(F_{XY})$ will be simple to calculate, as all cross moments under this type of multivariate kernel will be 0.

Example

We performed a simulation study to highlight the effects of varying proposal model specification on posterior inference for regression coefficients. The functional of interest was $\beta = \{\beta_0, \beta_1\}$ as defined by Example 1.3. We focused on marginal inference of β_0 and β_1 separately. When inferring β_0, we used the model

$$\text{TA}(m = p_{\beta_0}/q^{\Pi}_{\beta_0}, \mathcal{P}_{\Pi})$$

as the Bayesian prior. When inferring β_1, we used the model

$$\text{TA}(m = p_{\beta_1}/q^{\Pi}_{\beta_1}, \mathcal{P}_{\Pi})$$

as the Bayesian prior.

The data (X, Y) was generated according to Model 5.6. A contour plot of the data-generating density function of (X, Y) is given in Figure 5.7. The DPM model shown in Model 5.7 and DP model shown in Model 5.8 were used as $\mathcal{P}_{\Pi}$. Several values of hyperparameters for the proposal models were tested in combination to examine the effect of hyperparameter specification on posterior TAB inference. The results of our simulation are presented in terms of effective likelihoods. This allows us to discuss the marginal prior-to-posterior update mechanism without selecting a specific subjective prior for the regression coefficients.

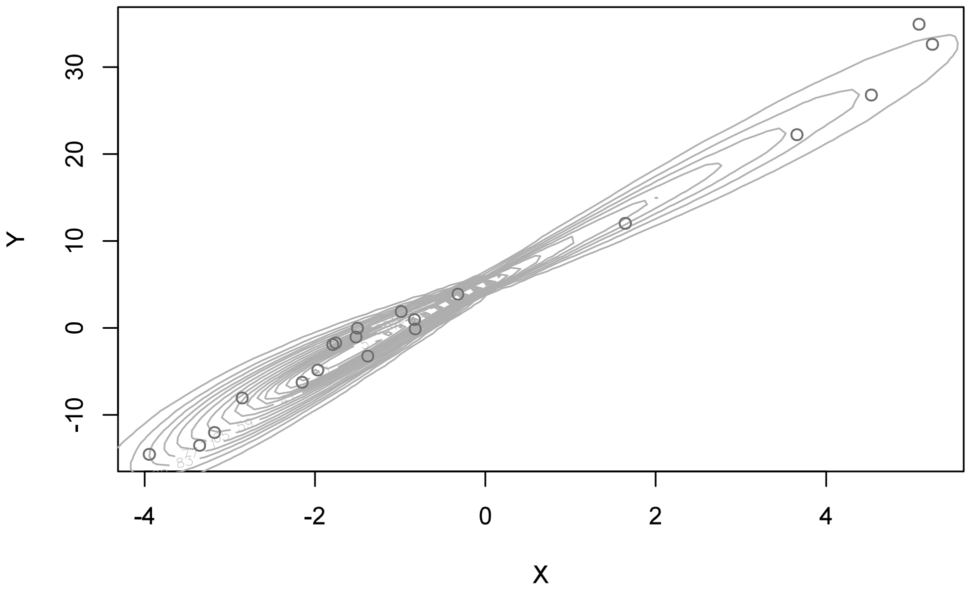


Figure 5.7: Scatter plot of the data used in Section 5.1.2, along with contours of the data-generating distribution given by Model 5.6. Data points in the dataset are shown as open circles.

Model 5.6. (Data-generating mechanism for observable (X, Y) in Section 5.1.2)

$$\begin{aligned}
T_1 &\sim \text{Normal}(\mu = 2, \sigma^2 = 5)\\
T_2 &\sim \text{Normal}(\mu = -1, \sigma^2 = 1)\\
L &\sim \text{Bernoulli}(0.3)\\
X &= LT_1 + (1 - L)T_2\\
Y &= 5 + 5X + \epsilon\\
\epsilon | X &\sim \text{Normal}(\mu = 0, \sigma^2 = X^{4/3}).
\end{aligned}$$

Model 5.7. (DPM proposal for estimating linear regression parameters in Section 5.1.2)

$$\begin{aligned}
f_{XY}(x, y) = &\int N(x|\mu_x, \sigma_x^2)N(y|\mu_y, \sigma_y^2)\mathrm{d}F_H(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2)\\
F_H \sim &\text{DP}(\phi, G_0)\\
G_0 &= G_{0X} \times G_{0Y}\\
G_{0X} &= \text{NIG}(\mu_{0,X}, \lambda_X, \alpha_X, \beta_X)\\
G_{0Y} &= \text{NIG}(\mu_{0,Y}, \lambda_Y, \alpha_Y, \beta_Y),
\end{aligned}$$

with kernel parameters (μ_x, σ_x^2) sampled from G_{0X}, and (μ_y, σ_y^2) sampled from G_{0Y}. $N(\cdot|\mu, \sigma^2)$ denotes the Gaussian density function.

Model 5.8. (DP proposal for estimating linear regression parameters in Section 5.1.2)

$$\begin{aligned}
F_{XY}(x, y) \sim &\text{DP}(\phi, G_0')\\
G_0' &= \text{Discretized version of the distribution } G_{0X} \times G_{0Y},\\
&\quad \text{with mass assigned to points } \{(ih, jh)|i, j \in \mathbb{Z}\} \text{ for fixed bin width } h \text{ s.t.}\\
&\quad G_0'(X = ih, Y = jh) = \int\int \mathbb{1}\left[x \in \left(ih - \frac{h}{2}, ih + \frac{h}{2}\right)\right] \times\\
&\qquad\qquad \mathbb{1}\left[y \in \left(jh - \frac{h}{2}, jh + \frac{h}{2}\right)\right] \mathrm{d}G_{0X}(x)\mathrm{d}G_{0Y}(y)\\
G_{0X} &= \text{Normal}(\mu_X, \sigma_X^2)\\
G_{0Y} &= \text{Normal}(\mu_Y, \sigma_Y^2)\\
h &= \text{a constant}
\end{aligned}$$

Table 5.2: Proposal model hyperparameters for the example in Section 5.1.2, which were tested in combination for estimation of linear regression parameters. The proposal DPM and DP models are described in detail in Model 5.7 and Model 5.8.

DPM Model 5.7	**DP Model 5.8**
DPM precision $\phi \in \{0.3, 1.5\}$	DP precision $\phi \in \{0.3, 1.5\}$
DPM base measure for the sampling of kernel parameters $G_{0X} \in \{G_{0X,1}, G_{0X,2}\}$ $G_{0Y} \in \{G_{0Y,1}, G_{0Y,2}\}$ $G_{0X,1} =$ $\text{NIG}(\mu_{0,X} = \bar{x}_n, \lambda_X = 1 \times 10^{-3}, \alpha_X = 3, \beta_X = 1 \times 10^{-3})$ $G_{0X,2} =$ $\text{NIG}(\mu_{0,X} = \bar{x}_n, \lambda_X = 4 \times 10^{-3}, \alpha_X = 2, \beta_X = 1 \times 10^{-4})$ $G_{0Y,1} =$ $\text{NIG}(\mu_{0,Y} = \bar{y}_n, \lambda_Y = 2 \times 10^{-2}, \alpha_Y = 3, \beta_Y = 15)$ $G_{0Y,2} =$ $\text{NIG}(\mu_{0,Y} = \bar{y}_n, \lambda_Y = 1.5 \times 10^{-2}, \alpha_Y = 3, \beta_Y = 5)$	DP base measure prior to discretization $G_{0X} \in \{G_{0X,1}, G_{0X,2}\}$ $G_{0Y} \in \{G_{0Y,1}, G_{0Y,2}\}$ $G_{0X,1} = \text{Normal}(\mu_X = 0, \sigma_X^2 = 10^2)$ $G_{0X,2} = \text{Normal}(\mu_X = 0, \sigma_X^2 = 0.5^2)$ $G_{0Y,1} = \text{Normal}(\mu_Y = 1, \sigma_Y^2 = 10^2)$ $G_{0Y,2} = \text{Normal}(\mu_Y = 1, \sigma_Y^2 = 20^2)$
	$h = 1 \times 10^-4$

We were interested in studying the conditions that lead to the effective likelihood, $q^{\Pi}_{\theta|\tilde{x}_n}/q^{\Pi}_{\theta}$, of TAB inference being approximately the same as the BB. Since we knew that DP and DPM posterior distributions would be approximately equal to the Bayesian bootstrap with ϕ set close to 0, ϕ was chosen to be 0.3 and 1.5 for the simulation. As for the base distributions, we thought of these as primarily influencing the flatness of $q^{\Pi}_{\beta_0}$ and $q^{\Pi}_{\beta_1}$. Although, given that ϕ cannot be 0, we had to make sure that when a base distribution led to $q^{\Pi}_{\beta_0}$ and $q^{\Pi}_{\beta_1}$ being flat, it did not cause $q^{\Pi}_{\beta_0|\tilde{x}_n}$ and $q^{\Pi}_{\beta_1|\tilde{x}_n}$ to become excessively wide. The discovery of suitable hyperparameters took some trial and error due to the numerator and denominator of an effective likelihood being affected simultaneously by changes in the hyperparameter specification. The selected hyperparameter values are detailed in

Table 5.2. The results for marginal parametric TAB inference based on these proposal models and hyperparameter values are shown in Figures 5.8 - 5.11.

We saw in the results that, as expected, decreasing ϕ towards 0 had the effect of tuning the effective likelihoods toward Bayesian Bootstrap posteriors. When ϕ was increased, the effective likelihoods became more spread out, reflecting a decrease in the informativeness of data, as expected. The effect of base distribution specification on the effective likelihood was less straight forward.

We detail our observations of the effects of changing base distribution specification on the effective likelihood. For the DP proposal, when we set the variance of base distribution G_{0X} to be much smaller than the sample variance, the proposal prior for β_1 became extremely spread out, which led to the effective likelihood of β_1 being approximately the Bayesian bootstrap posterior. The choice of a base distribution G_{0X} with variance smaller than the sample variance seemed counter intuitive, but it was necessary because a smaller variance of X translated to a larger covariate effect, when the spread of G_{0Y} was kept the same. The base distribution G_{0Y} also had an effect on the effective likelihood, in that increasing the variance of G_{0Y} led to a widening of the induced proposal prior for β_0 and β_1 when G_{0X} was kept the same. An appropriate specification for approximating the Bayesian bootstrap was found with the variance of G_{0Y} being approximately equivalent to or slightly larger than that of the observed values, and the variance of G_{0X} artificially smaller than that of the observed x values, while matching the means of these base distributions to the sample means.

Hyperparameter specification for the DPM proposal model worked in a similar fashion as for the DP. However, due to a location scale kernel being used, there were many more parameters in the base distribution of a DPM model, making hyperparameter tuning and selection comparatively more difficult. The complexity of the algorithm for implementing a DPM model for the 2-dimensional observable was not much more than that of a DPM for a 1-dimensional observable. We note that all combinations of base distributions used for the DPM in this study resulted in some clustering of the data, which made the resulting density estimation model more plausible. However, this likely was the reason why effective likelihoods of DPM proposals (Figures 5.8 and 5.10) did not track the Bayesian bootstrap very closely. Regardless, TAB posteriors based on a DPM proposal model that leads to clustering of data would still be considered genuine if we belief clustering of data to be likely *a priori*.

We also tested an additional scenario where the posterior DPM proposal model resulted in no clustering of the data, with kernels approaching point-mass. In this case, shown in Figure 5.12, the effective likelihoods track the Bayesian bootstrap posteriors very closely.

5.1.3 Logistic regression

When $\mathcal{P}_\Pi$ is a DPM of continuous kernels, the calculation of nonlinear functionals of random distributions sampled from this model tends to be difficult. As an example, we take the target of inference to be the coefficients of logistic regression. Suppose the observable random variables are $C \in \{0, 1\}$ and $X = (1, X_1)^\top$, with target parameter $\psi = (\psi_0, \psi_1)^\top$ defined according to Example 1.4. We denote a distribution function for the observable (X, C) with F_{XC}.

The observable C is a Bernoulli random variable, which does not have a density. The application of DP and DPM to model this type of data is somewhat unorthodox. Computationally, we do not encounter too much difficulty. For a DP proposal model, we can posit a base distribution $G_0 = G_{0X} \times G_{0C}$ where G_{0C} is supported on $\{0, 1\}$. When the precision parameter $\phi \to 0$, the limiting posterior inference based on a DP prior is the Bayesian bootstrap, which is commonly used as posterior inference for functional parameters regardless if the observed variable is discrete or continuous.

As for suitable specifications of a DPM proposal model, one possibility with regards to the kernel is

$$\mathcal{K}(x, c) = \mathcal{K}_X(x|\eta) \times \left[\theta^c (1-\theta)^{(1-c)}\right],$$

which has a jointly independent structure. The mapping from F_{XC} to $\psi(F_{XC})$ requires

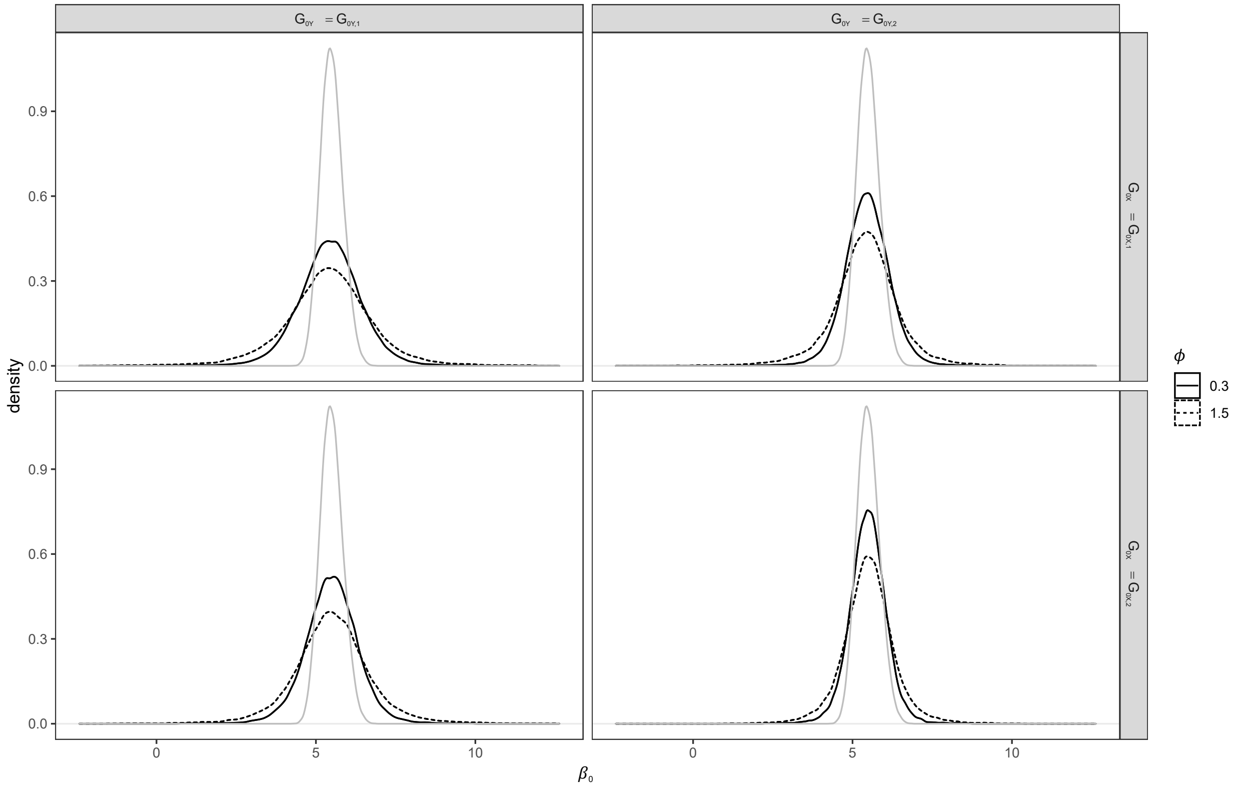


Figure 5.8: TAB effective likelihood curves for inferring regression parameter β_0 with $\mathcal{P}_{\Pi}$ being Model 5.7, a DPM of jointly independent Gaussian kernels. Black curves correspond to the effective likelihoods under various hyperparameter specifications. The figure contains several panels, arranged into rows and columns, with panels in the same column sharing the same specification for the parameter G_{0Y}, and panels in the same row sharing the same specification for the parameter G_{0X}. The line type of an effective likelihood curve indicates the value of parameter ϕ. The Bayesian bootstrap (in grey) is overlaid for comparison. Detailed hyperparameter specifications are given in Table 5.2. Out of the four panels, the effective likelihoods shown in the bottom-right panel are most similar to the BB. This is due to the prior variance observable X under parameter $G_{0,X2}$ being smaller than that of $G_{0,X1}$. The prior variance of observable Y under parameter $G_{0,Y2}$ is smaller than that of $G_{0,Y1}$.

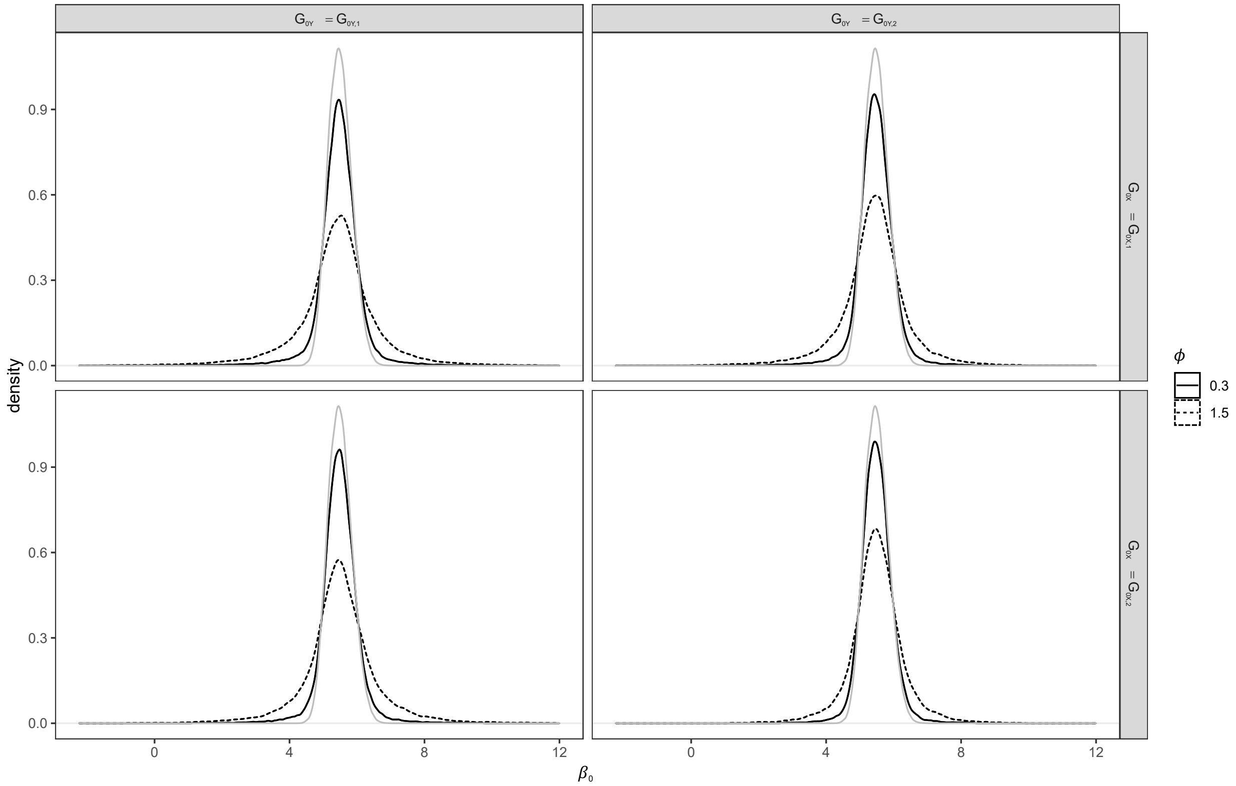


Figure 5.9: Marginal inference for the β_0 parameter of linear regression, Section 5.1.2. TAB effective likelihood curve for inferring the regression parameter β_0 with the DP (Model 5.8) as $\mathcal{P}_\Pi$. Black curve correspond to effective likelihoods under varying sets of hyperparameters. The panels are arranged into rows and columns with panels in the same column sharing the same specification for the parameter G_{0Y}, and panels in the same row sharing the same specification for the parameter G_{0X}. The line type of an effective likelihood curve indicates the value of parameter ϕ. The Bayesian bootstrap (in grey) is overlaid for comparison. Detailed hyperparameter specifications are given in Table 5.2. Out of the four panels, the effective likelihoods shown in the bottom-right panel are the most similar to the BB. This is due to the prior variance observable X under parameter $G_{0,X2}$ being smaller than that of $G_{0,X1}$. The prior variance of observable Y under parameter $G_{0,Y2}$ is smaller than that of $G_{0,Y1}$.

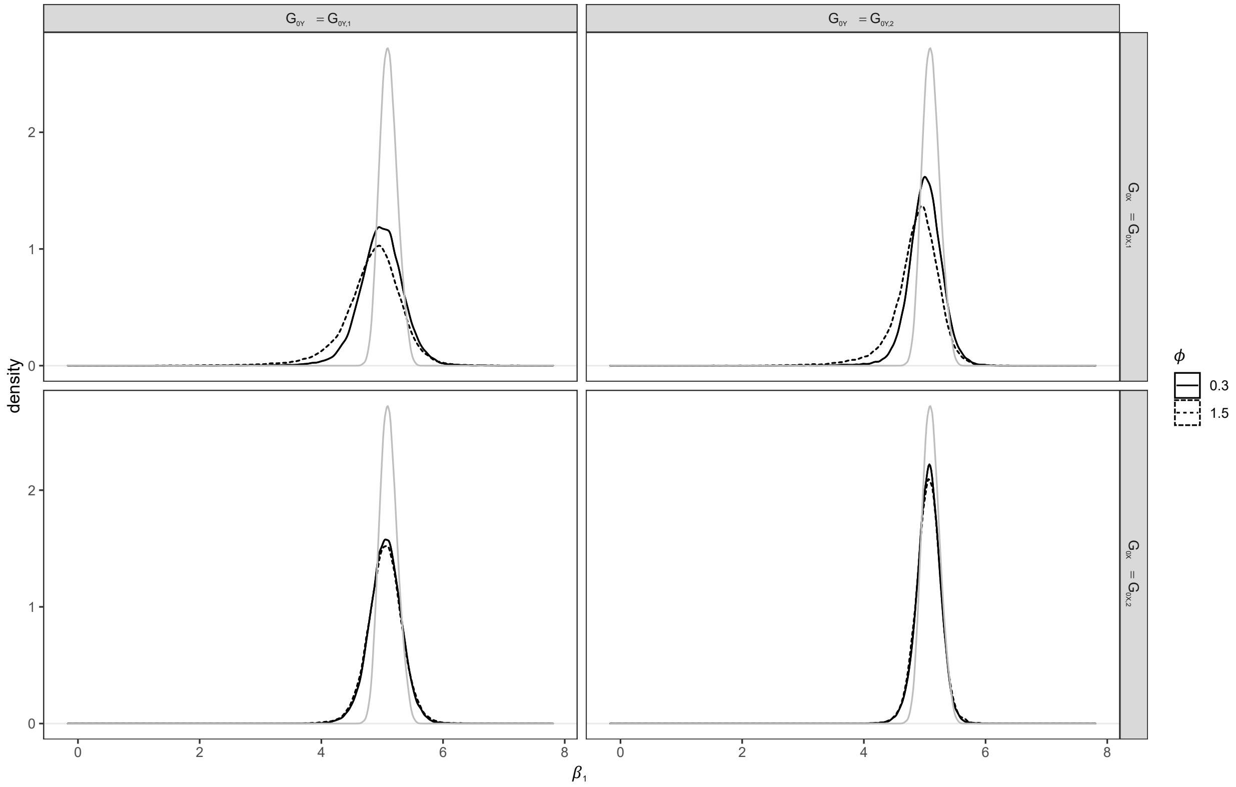


Figure 5.10: TAB effective likelihood curves for inferring regression parameter β_1 with the DPM (Model 5.7) as $\mathcal{P}_\Pi$. Black curves correspond to effective likelihoods under varying sets of hyperparameters. The panels are arranged into rows and columns with panels in the same column sharing the same specification for the parameter G_{0Y}, and panels in the same row sharing the same specification for the parameter G_{0X}. The line type of an effective likelihood curve indicates the value of parameter ϕ. The Bayesian bootstrap (in grey) is overlaid for comparison. Detailed hyperparameter specifications are given in Table 5.2. Out of the four panels, the effective likelihoods shown in the bottom-right panel are most similar to the BB. This is due to the prior variance observable X under parameter $G_{0,X2}$ being smaller than that of $G_{0,X1}$. The prior variance of observable Y under parameter $G_{0,Y2}$ is smaller than that of $G_{0,Y1}$.

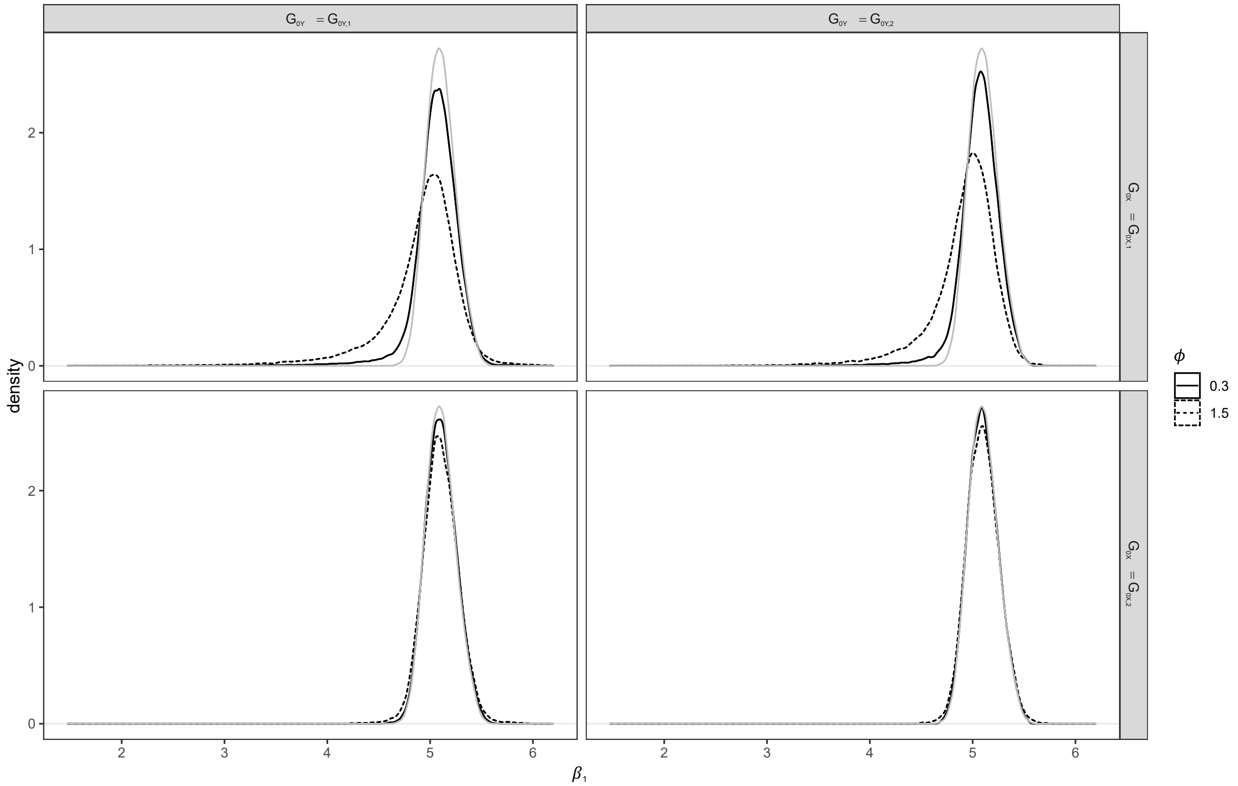


Figure 5.11: Marginal inference for the β_1 parameter of linear regression, Section 5.1.2. TAB effective likelihood curve for inferring the regression parameter β_0 with the DP (Model 5.8) as $\mathcal{P}_{\Pi}$. Black curve correspond to effective likelihoods under varying sets of hyperparameters. The panels are arranged into rows and columns with panels in the same column sharing the same specification for the parameter G_{0Y}, and panels in the same row sharing the same specification for the parameter G_{0X}. The line type of an effective likelihood curve indicates the value of parameter ϕ. The Bayesian bootstrap (in grey) is overlaid for comparison. Detailed hyperparameter specifications are given in Table 5.2. Out of the four panels, the effective likelihoods shown in the bottom-right panel are most similar to the BB. This is due to the prior variance observable X under parameter $G_{0,X2}$ being smaller than that of $G_{0,X1}$. The prior variance of observable Y under parameter $G_{0,Y2}$ is smaller than that of $G_{0,Y1}$.

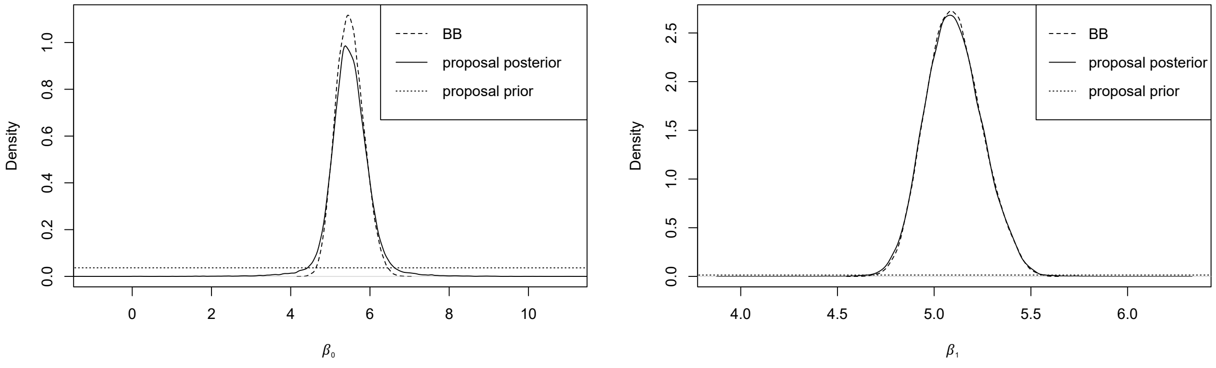


Figure 5.12: Plots of TAB effective likelihoods based on a specification of the DPM proposal (Model 5.7) which induced no clustering of data points. The left-hand-side panel shows induced distributions of β_0 and the right-hand-side shows that of β_1. The line type of a curve corresponds to a particular type of distribution (proposal prior, proposal posterior, and the BB posterior) as indicated by the plot legend. Base distribution parameters are $\phi = 0.3$, $G_{0X} = \text{NIG}(\mu_{0,X} = \bar{x}_n, \lambda_X = 1.2 \times 10^{-5}, \alpha_X = 20, \beta_X = 1 \times 10^{-5})$ and $G_{0Y} = \text{NIG}(\mu_{0,Y} = \bar{y}_n, \lambda_Y = 3.3 \times 10^{-9}, \alpha_Y = 15, \beta_Y = 1 \times 10^{-5})$. In each subplot, we observe that the TAB proposal prior is essentially flat over the region of interest, which means the proposal posterior shown in each panel may also be interpreted as the TAB effective likelihood.

that we solve the following estimating equation,

$$\begin{aligned}
0 &= \mathbb{E}[X^\top (C - \tau(X^\top \psi))] \\
&= \sum_{j=1}^{\infty} w_j \mathbb{E}_j[X^\top C] - \sum_{i=1}^{\infty} w_i \mathbb{E}_i[X^\top \tau(X^\top \psi)],
\end{aligned}$$

where $\tau(z) = \dfrac{1}{1+\exp(-z)}$. We found the integral $\mathbb{E}_i[X^\top \tau(X^\top \psi)]$ particularly challenging to calculate for the Gaussian kernel, even with the aid of symbolic math software. One could approximate the integral via numerical integration. However, the required computation time does not seem realistic as $\mathbb{E}_i[X^\top \tau(X^\top \psi)]$ needs to be calculated repeatedly in a Monte Carlo approximation of the posterior.

As a workaround, the calculation of $\mathbb{E}_i[X^\top \tau(X^\top \psi)]$ is typically feasible when the integration is performed with respect to an uniform distribution. Exact expressions for the definite integrals of estimating functions with respect to a uniform distribution may be found with the help of computer programs for symbolic mathematics.

In the context of modelling (X, C), we propose as the DPM (partial) kernel the uniform distribution

$$\mathcal{K}_U(x|\mu, h) = \frac{1}{h}\mathbb{I}\left[x \in \left[\mu - \frac{1}{2}h, \mu + \frac{1}{2}h\right]\right].$$

The complete kernel is jointly independent in x and c, that is,

$$\mathcal{K}(x, c) = \mathcal{K}_U(x|\mu, h) \times \left[\theta^c (1-\theta)^{(1-c)}\right],$$

and the kernel parameters are (μ, h, θ). Based on this partially uniform kernel, the estimating equation for logistic regression can be simplified, and the exact expressions of the relevant integrals are shown in Appendix A.6. We may further simplify the sampling algorithm with the following DPM base distribution:

$$\begin{aligned}
G_0 &= G_{0\mu} \times G_{0h} \times G_{0\theta} \\
G_{0\theta} &= \text{Beta}(\alpha_\theta, \beta_\theta) \\
G_{0\mu} &= \sum_{j=1}^{J} p_j \frac{1}{(b_j - a_j)} \mathbb{I}[\mu \in (a_j, b_j)] \\
G_{0h} &= \text{Gamma}(\alpha_h, \beta_h).
\end{aligned}$$

Sampling from this DPM model can be performed with a blocked Gibbs algorithm, with the conditional distributions given in Appendix A.5.

Example

We performed a simulation study to highlight the effects of varying proposal model specification on posterior inference for logistic regression. The functional of interest was $\psi = \{\psi_0, \psi_1\}$ as defined by Example 1.4. We focused on marginal inference of ψ_0 and ψ_1 separately. When inferring ψ_0, we used the model

$$\text{TA}(m = p_{\psi_0}/q^{\Pi}_{\psi_0}, \mathcal{P}_{\Pi})$$

as the Bayesian prior. When inferring ψ_1, we used the model

$$\text{TA}(m = p_{\psi_1}/q^{\Pi}_{\psi_1}, \mathcal{P}_{\Pi})$$

as the Bayesian prior.

In this example, we generated the data according Model 5.9. Twenty data points were drawn from the data-generating mechanism. Figure 5.13 shows a scatter plot of the data, overlaid with a curve showing the true $\text{Pr}(C = 1|X)$, and a curve showing the best fitting logistic model for $\text{Pr}(C = 1|X)$ as given by maximum likelihood.

Model 5.9. (Data-generating model for the logistic regression example, Section 5.1.3)

$$\begin{aligned} X &\sim \text{Normal}(\mu = 5, \sigma^2 = 25) \\ p_c(X) &= 1/(1 + \exp(-(5 - X))) \\ C|X &\sim \text{Bernoulli}(p_c(X)). \end{aligned}$$

The DPM and DP models we chose as $\mathcal{P}_{\Pi}$ for conducting TAB inference are detailed, respectively, in Model 5.10 and Model 5.11. Specifications for the hyperparameters of the DP proposal, i.e. Model 5.11, is given in Table 5.3. Specifications for the hyperparameters of the DPM proposal is given directly in Model 5.10. The DPM proposal was only studied under one set of hyperparameter values as an attempt to demonstrate the feasibility of our

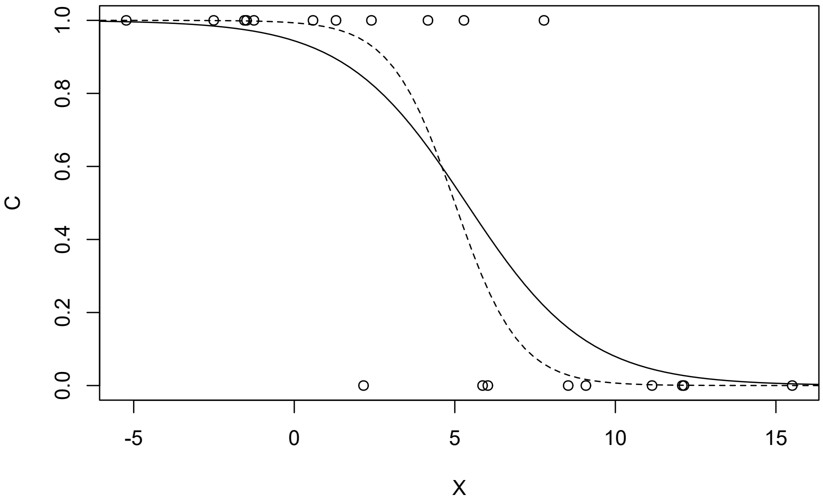


Figure 5.13: Logistic regression in Section 5.1.3. A scatter plot of the data (n=20) used in the example, with each data point represented by an open circle. The best fitting logistic model $p_c(X)$ for this particular sample is shown in solid line, and the true model is shown in dotted line.

method. Under the chosen hyperparameters, our DPM proposal introduced a moderate amount of clustering in the data points, which provided some contrast to our DP proposal model. Please refer to Figure 5.18 for examples of random draws of F_X from the posterior DPM. The results for TAB inference are again presented in terms of effective likelihoods, and are summarized in Figures 5.14- 5.16.

Model 5.10. (DPM proposal for estimating logistic regression parameters in Section 5.1.3)

$$f(x,c) = \int \frac{1}{h}\mathbb{I}\left[x \in \left[\mu - \frac{1}{2}h, \mu + \frac{1}{2}h\right]\right]\left[\theta^{c}(1-\theta)^{(1-c)}\right] \mathrm{d}F_H(\mu,h,\theta)$$

$$F_H \sim \text{DP}(\phi, G_0)$$

$$\phi = 0.5$$

$$G_0 = G_{0\mu} \times G_{0h} \times G_{0\theta}$$

$$G_{0\theta} = \text{Beta}(\alpha_\theta = 1 \times 10^{-5}, \beta_\theta = 1 \times 10^{-5})$$

$$G_{0\mu} = \sum_{j=1}^{3} p_j \frac{1}{(b_j - a_j)}\mathbb{I}[\mu \in (a_j, b_j)]$$

$$(a_1, a_2, a_3) = (-20, -10, -6)$$

$$(b_1, b_2, b_3) = (50, 30, 16)$$

$$(p_1, p_2, p_3) = (0.001, 0.01, 0.989)$$

$$G_{0h} = \text{Gamma}(\alpha_h = 10, \beta_h = 10).$$

Model 5.11. (DP proposal for estimating logistic regression parameters in Section 5.1.3)

$$\begin{aligned}
F_{XC}(x,c) \sim & \text{DP}(\phi, G_0') \\
G_0' = & \text{a discrete distribution with mass assigned to} \\
& \text{points } \{(ih,c)|i \in \mathbb{Z}, c \in \{0,1\}\} \text{ for fixed bin width } h \text{ s.t.} \\
& G_0'(X=ih, C=c|h) = \\
& \int \mathbb{1}\left[x \in \left(ih - \frac{h}{2}, ih + \frac{h}{2}\right)\right] \mathrm{d}G_{0X}(x) \times p_c^{\mathbb{1}[c=1]}(1-p_c)^{\mathbb{1}[c=0]} \\
G_{0X} = & \text{ Normal}(\mu, \sigma^2) \\
G_{0C} = & \text{ Bernoulli}(p_c) \\
h = & \text{ a constant.}
\end{aligned}$$

As with previous examples, the effective likelihoods based on a DP proposal model with small ϕ approximated BB posteriors closely. We were able to make the denominators of the effective likelihoods, i.e. $q^{\Pi}_{\psi_0}$ and $q^{\Pi}_{\psi_1}$, relatively flat by choosing a DP base distribution G_{0X} with small variance, since X was the covariate in the logistic regression. We kept the mean of G_{0X} the same as the sample mean throughout the simulation, which helped maintain the resemblance of the effective likelihoods to BB posteriors. The same base distribution G_{0C} was used throughout this study, as it seemed to have little effect on the effective likelihoods under the chosen structure for the DP proposal model. More investigation is needed to understand the effect of G_{0C} on the coefficients of logistic regression.

The posterior DPM proposal, Model 5.10, introduced a fair amount of clustering in the data (Figures 5.17 and 5.18). For F_{XC} sampled from this DPM proposal mode, calculation of $\psi(F_{XC})$ was quite tricky numerically, as different root finding/optimization algorithms often led to different solutions, suggesting that the estimating equation for logistic regression gave rise to a highly complex surface. The effective likelihoods for ψ_0 and ψ_1 under the DPM proposal model are shown in Figure 5.16.The effective likelihoods of TAB inference based on using DPM Model 5.10 as $\mathcal{P}_{\Pi}$ also tracked BB posteriors closely in the

Table 5.3: Specifications of the hyperparameters in DP proposal, Model 5.11, in Section 5.1.3; the parameters are tested in combination.

DP Model 5.11
DP precision $\phi \in \{0.3, 1.5\}$
DP base measure $G_{0C} = \text{Bernoulli}(0.5)$ $G_{0X} \in \{G_{0X,1}, G_{0X,2}\}$ $G_{0X,1} = \text{Normal}(\mu = \bar{x}_n, \sigma^2 = 10^2)$ $G_{0X,2} = \text{Normal}(\mu = \bar{x}_n, \sigma^2 = 2.5^2)$
Bandwidth of discretization $h = 1 \times 10^{-4}$

marginal parametric inference for both ψ_0 and ψ_1. The difference in parametric TAB inference between the use of DP versus DPM as $\mathcal{P}_\Pi$ was minor, despite a moderate amount of clustering in the data under the posterior DPM.

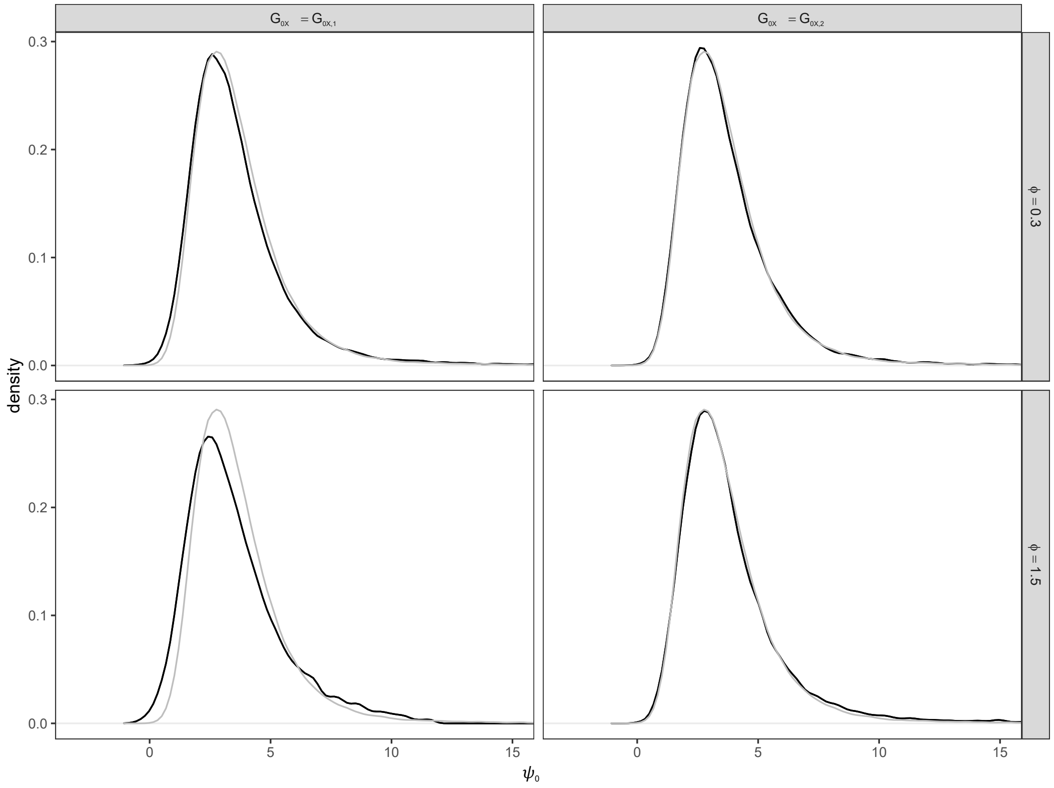


Figure 5.14: Marginal inference for logistic regression parameter ψ_0 in Section 5.1.3 using DP Model 5.11 as the TAB proposal model. Black curves represent TAB effective likelihoods. The panels are arranged into rows and columns with panels in the same column sharing the same specification for the parameter G_{0X}, and panels in the same row share the same specification for the parameter ϕ. The Bayesian bootstrap posterior is shown in grey in each subplot for comparison. Detailed hyperparameter specifications are given in Table 5.3. We note that $G_{0,X2}$ has a smaller variance than $G_{0,X1}$, which shows that decreasing the spread of the base distribution increased the resemblance of TAB effective likelihood to the Bayesian bootstrap posterior. Decreasing ϕ also led to increased resemblance of the effective likelihoods to the BB.

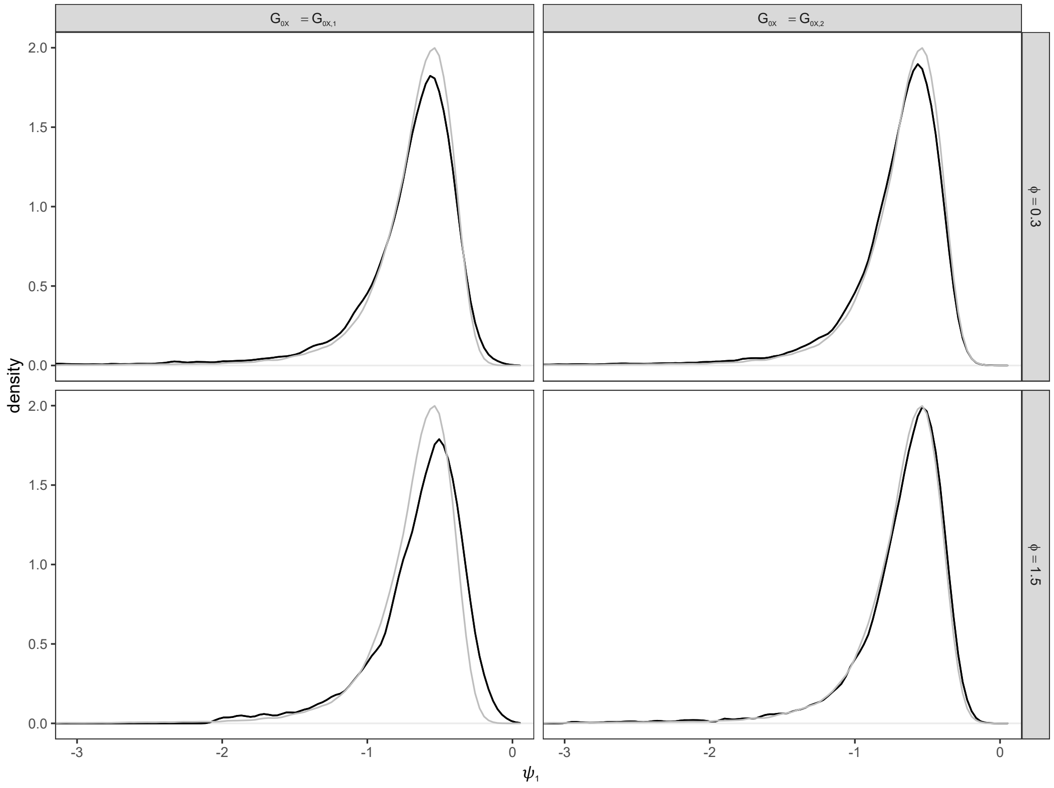


Figure 5.15: Marginal inference for logistic regression parameter ψ_1 in Section 5.1.3 using DP Model 5.11 as the TAB proposal model. Black curves represent TAB effective likelihoods. The panels are arranged into rows and columns with panels in the same column sharing the same specification for the parameter G_{0X}, and panels in the same row share the same specification for the parameter ϕ. The Bayesian bootstrap posterior is shown in grey in each subplot for comparison. Detailed hyperparameter specifications are given in Table 5.3. We note that $G_{0,X2}$ has a smaller variance than $G_{0,X1}$, which shows that decreasing the spread of the base distribution increased the resemblance of TAB effective likelihood to the Bayesian bootstrap posterior. Decreasing ϕ also led to increased resemblance of the effective likelihoods to the BB.

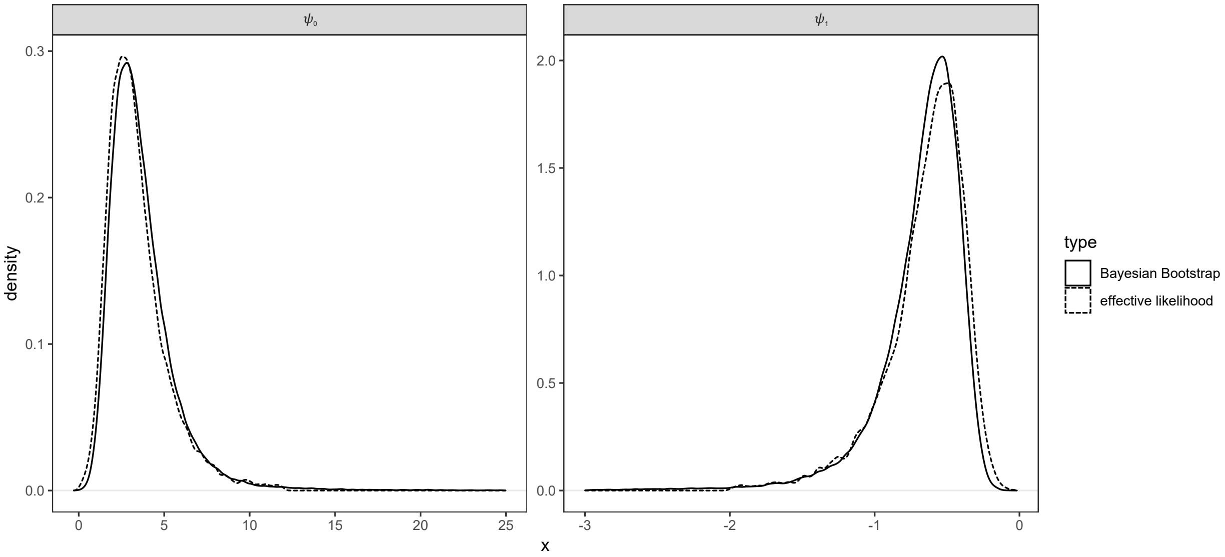


Figure 5.16: Marginal TAB inference for logistic regression parameter ψ_0 (left-hand-side panel) and ψ_1 (right-hand-side panel) in Section 5.1.3 using DPM Model 5.10 as the proposal model. TAB effective likelihoods are shown with dotted lines while the Bayesian bootstrap posteriors are shown with solid lines.

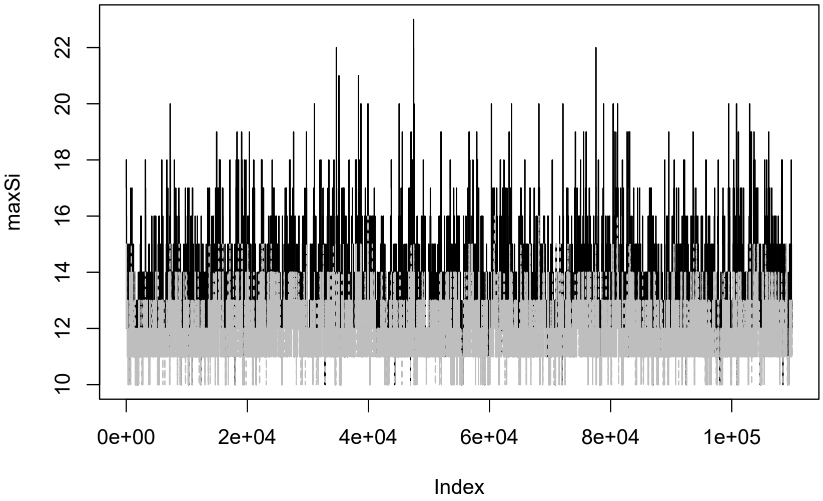


Figure 5.17: Fitting of DPM Model 5.10 in Section 5.1.3 to the example data: plot of maximum occupied kernel index and numbers of latent kernels over MCMC iterations. Maximum occupied kernel index is shown in black solid line and the total number of latent clusters is shown in grey dotted line.

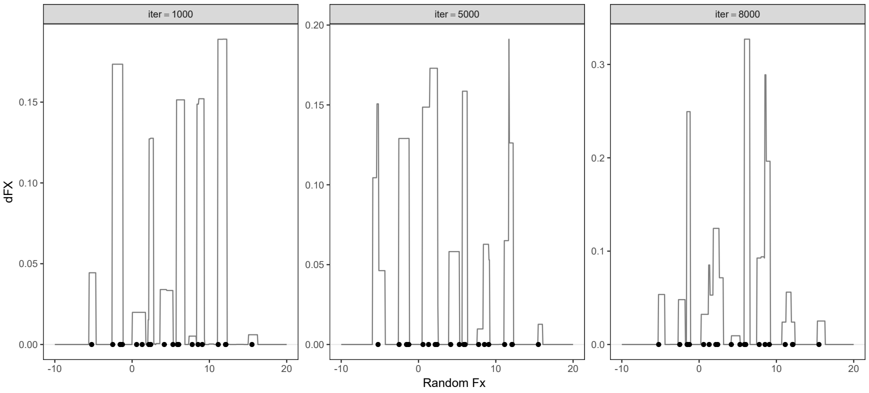


Figure 5.18: Fitting of DPM Model 5.10 in Section 5.1.3 to the example data: random measures for the observed X drawn from the posterior DPM. The observed X values are overlaid as black dots on the x-axis. Each panel shows a random measure for X obtained at the MCMC iteration indicated above the panel.

Chapter 6

Comparison of semiparametric Bayesian methods via simulations

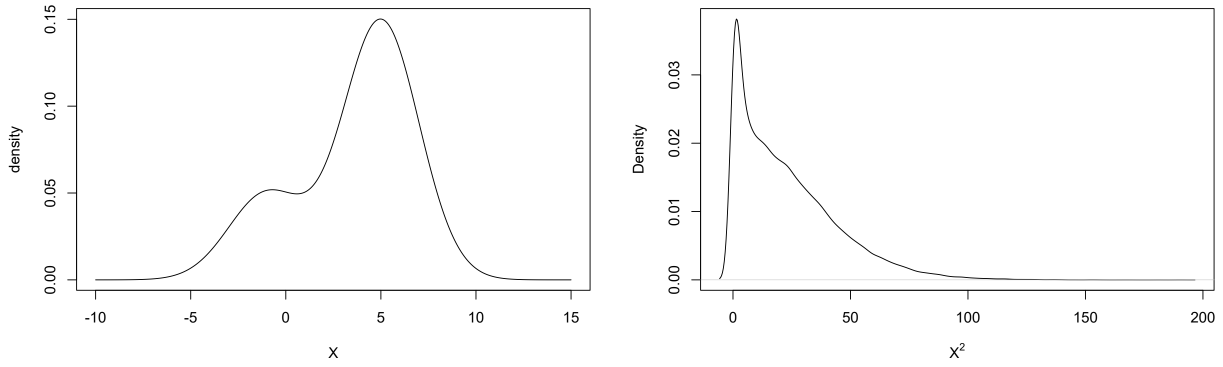
The philosophical gain of the θ-augmented Bayesian method over its competitors is clear. It is fully Bayesian, in contrast with existing methods which are at best pseudo-Bayesian. This section explores the practical differences between existing methods and the TAB method. The evaluation of competing methods is performed under a Frequentist framework.

Four methods – the θ-augmented Bayesian method, Bayesian bootstrap, general Bayes method, and Bayesian empirical likelihood inference, are compared. Prior to the simulations we have identified several weaknesses of the competing methods. The Bayesian bootstrap does not allow one to incorporate a subjective prior, and as such, actual performance should be worse than competing methods that incorporate a correctly specified prior. The GB method sometimes assumes implicit parametric likelihood models, such that, when the random observable being modelled deviates from this implicit model, the estimation is likely inefficient. The Bayesian empirical likelihood posterior can be computationally challenging to implement when the number of nuisance parameters is large. Both BB and BEL posteriors are bound by the convex hull condition which suggests difficulty with coverage of interval estimates in small sample inference.

Given the considerations above, we chose to highlight the strength of TAB via two examples. The first example is a problem of estimating the mean and variance parameters, the second example is a problem of estimating the mean of a population when some of the observations are missing at random. The variety of problems includes scenarios where all estimators perform quite well practically, typically when the estimating equation involves a summary variable which is more or less unimodal with little skew, and scenarios where the performance of competing methods is quite varied.

The results for TAB inference were obtained based on a TA model with Dirichlet process as $\mathcal{P}_{\Pi}$. We chose the DP due to the sampling algorithm being simple to implement, and the ease with which we calculate functionals based on random measures drawn from the DP. We believe these qualities will positively influences the adoption rate of TAB inference in future applications. Secondly, in Chapter 5 we saw that inference via a DP proposal model did not differ significantly from inference via a DPM proposal model when there was minimal clustering.

Under a Frequentist framework, we evaluated each method via estimates of the following metrics. For joint inference, we considered the expected size and coverage probability of credible regions. For marginal inference, we considered the expected length and coverage probability of credible intervals, bias of the posterior mean, and expected quadratic risk for predicting the truth. Depending on the method being evaluated, samples from a posterior distribution were drawn either by direct sampling or MCMC. Upon obtaining samples from a posterior distribution, the joint credible regions was identified via the `kde` function from the R package `ks` (Duong, 2020). Marginally, credible intervals were chosen to be highest posterior density (HPD) intervals, and calculated via the R package `HDInterval` (Meredith and Kruschke, 2020). The bias of posterior mean was defined as $\mathbb{E}_{F_0}[(\hat{\theta}_n - \theta_0)]$, with expectation taken over the true data-generating mechanism F_0, where $\hat{\theta}_n$ denotes the mean of a posterior distribution conditional on a sample of size n. The expected quadratic risk was defined as $\mathbb{E}_{F_0}[\int(\theta - \theta_0)^2 f_\theta(\mathrm{d}\theta|\tilde{x}_n)]$. The expected quadratic risk measured how well random conditional posterior distributions captured



(a) X

(b) X^2

Figure 6.1: Comparison study of Section 6.1: marginal density functions of random observable X and X^2 generated by Model 6.1. The density function of X is shown on the left-hand-side, and that of X^2 is shown on the right-hand-side.

the true value. Estimation of the above performance metrics were based on 300 to 1,000 resampling events.

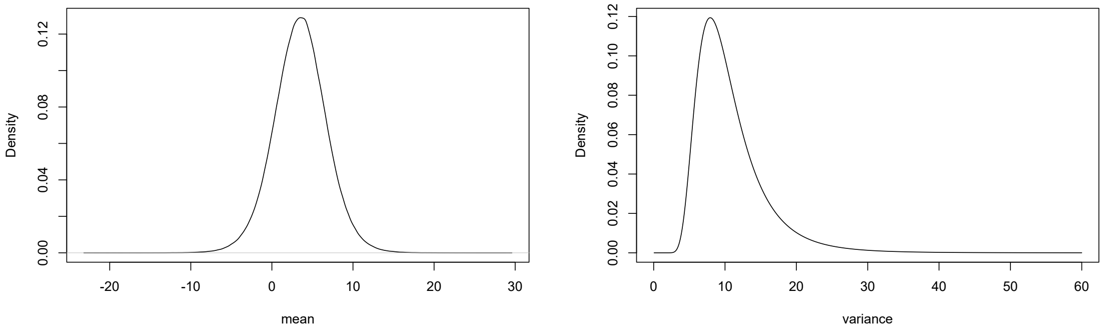
6.1 Estimating mean and variance parameters

In this example we generated the data from a slightly skewed distribution, as given by Model 6.1, the density function of which is shown in Figure 6.1.

Model 6.1. (Data-generating mechanism in Section 6.1)

$$\begin{aligned} &X = -6Z + T \\ &T \sim \text{Normal}(\mu = 5, \sigma^2 = 4) \\ &Z \sim \text{Bernoulli}(0.25). \end{aligned}$$

For methods requiring a subjective prior distribution, the same prior was utilized. We wished to emulate Bayesian inference with good prior information, and therefore chose the subjective prior, as specified in Model 6.2, to be distributed according to the



(a) μ

(b) σ^2

Figure 6.2: Comparison study of Section 6.1: marginal density functions of subjective prior according to Model 6.2. The marginal subjective prior for μ is shown on the left, and for σ^2 on the right.

Normal-Inverse-Gamma distribution with the marginal prior mean of variance parameter matching the variance of data-generating distribution, and the marginal prior mean of mean parameter matching the mean of data-generating distribution.

Model 6.2. (Joint subjective prior $p_{\mu,\sigma^2}(\mu, \sigma^2)$ in Section 6.1)

$$1/\sigma^2 \sim \text{Gamma}(\alpha = 6.623, \beta = 60.442)$$
$$\mu|\sigma^2 \sim \text{Normal}(\mu_0 = 3.5, \sigma_0^2 = \sigma^2).$$

The marginal densities of the above prior distribution are shown in Figure 6.2. In the simulation study, for each test dataset, θ-augmented Bayesian posterior inference was obtained via MCMC according to the method outlined in Section 4.2 with 4×10^5 runs, using the proposal model:

Model 6.3. (Proposal model for implementing θ-augmented method in Section 6.1)

$$\begin{aligned}
F_X \sim & \text{DP}(\phi, G_0') \\
G_0' = & \text{ Discretized version of continous distribution } G_0, \\
& \quad \text{with mass assigned to points } \{ih | i \in \mathbb{Z}\} \text{ for fixed bin width } h \text{ s.t.} \\
& \quad G_0'(X = ih) = \int \mathbb{1}\left[x \in \left(ih - \frac{h}{2}, ih + \frac{h}{2}\right)\right] \mathrm{d}G_0(x) \\
G_0 = & \text{ Normal}(\mu_0 = 0, \sigma_0^2 = 100) \\
\phi = & \ 0.5 \\
h = & \ 1 \times 10^{-5}.
\end{aligned}$$

Let q^{Π}_{μ,σ^2} denote the parametric joint distribution induced by the TAB proposal prior (Model 6.3). Noting p_{μ,σ^2} to be the density function of the actual subjective prior (Model 6.2). TAB inference was performed based on the following model:

Model 6.4. (Bayesian Prior for obtaining TAB inference in Section 6.1)

$$\begin{aligned}
X &\sim F_X \\
F_X &\sim \text{TA}(m = p_{\mu,\sigma^2}/q^{\Pi}_{\mu,\sigma^2}, \mathcal{P}_{\Pi} = \text{Model 6.3}).
\end{aligned}$$

Marginal inference on μ and σ^2 was obtained by marginalization of the joint TAB posterior.

Figure 6.3 shows the subjective prior p_{μ,σ^2} as being more concentrated than q^{Π}_{μ,σ^2}. Hence we believe that substitution of estimated $\hat{q}^{\Pi}_{\mu,\sigma^2}$ for q^{Π}_{μ,σ^2} in the MCMC provided a good approximation to the exact TAB posterior. Each MCMC chain for approximating a TAB posterior distribution was obtained with 4×10^5 iterations.

As for general Bayes inference for the mean and variance parameter, we note that there does not exist any distribution-free M-estimator for the variance parameter directly. Therefore we chose to estimate the mean and second moment via the 2-dimensional loss

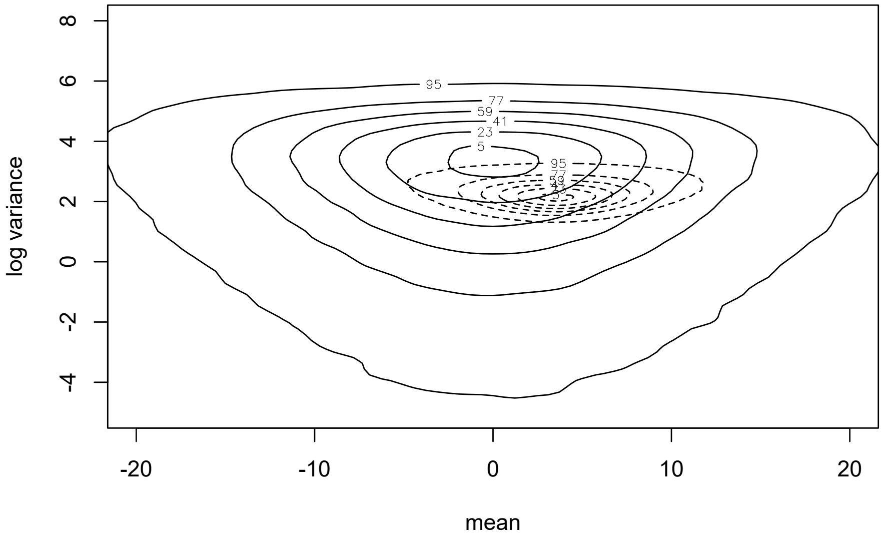


Figure 6.3: Comparison study of Section 6.1: Contour plots comparing q^{Π}_{μ,σ^2}, in solid lines, versus the p_{μ,σ^2}, in dotted lines. In this plot, distributions are parameterized in terms of mean and log variance.

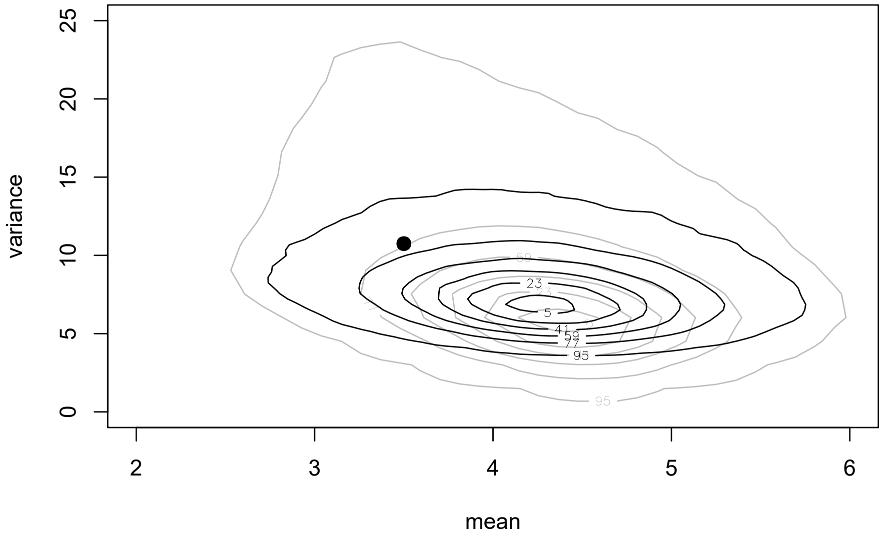


Figure 6.4: Comparison study of Section 6.1: contour plot of a TAB target posterior, in black, and that of the corresponding TAB proposal posterior, in grey, based on a particular dataset with 20 samples. The black dot marks the true mean and variance. The contours are shown under (μ, σ^2) parameterization because the posterior densities do not concentrate around 0 for σ^2 to necessitate a conversion to log variance scale.

functions $l(x, m_1, m_2) := (l_1(x, m_1), l_2(x, m_2))$, where

$$l_1(x, m_1) = (x - m_1)^2$$
$$l_2(x, m_2) = (x^2 - m_2)^2;$$

the first coordinate l_1 is a loss function targeting the first moment, whereas l_2 is a loss function targeting the second moment (denoted as $\mu_{(2)}$). The subjective prior $p(\mu, \sigma^2)$ was reparameterized and a joint density in the first two moments was obtained via the change of variable formula. The tuning parameter, w, of GB loss function was chosen based on each dataset; specifically it was set to the inverse of sample variance covariance matrix of (x, x^2). This choice was made based on asymptotic tuning and similar in spirits as Section 3.2 of Bissiri et al. (2016). Each MCMC for approximating a GB posterior distribution was obtained with a multivariate normal proposal distribution with mean of $(\bar{x}_n, \frac{1}{n}\sum_{i=1}^{n} x_i^2)$ and variance of w/n, and ran for 1×10^5 iterations.

Lastly, Bayesian empirical likelihood inference was obtained with the profile empirical likelihood function R_n defined as

$$R_n(\mu, \sigma^2) := \max_{\tilde{w} \in \mathcal{S}^{n-1}} \left\{ \prod_{i=1}^{n} w_i : \sum_{i=1}^{n} (x_i - \mu) = 0 \text{ and } \sum_{i=1}^{n} \frac{1}{n}(x_i - \mu)^2 - \sigma^2 = 0 \right\},$$

and BEL posterior being proportional to $R_n(\mu, \sigma^2)p(\mu, \sigma^2)$. The PEL function R_n was obtained via the iterative least squares technique described in Section 3.14 of Owen (2001). Due to a lack of explicit formula for R_n, finding good proposal distribution for the MCMC approximation of a BEL posterior requires trial-and-error. On a first pass, we chose the MCMC proposal distribution based on the posterior NIG distribution that would have resulted from a misspecified normal likelihood with a conjugate NIG prior equivalent to our subjective prior, while scaling the shape and rate parameters by 0.25 to ensure that the proposal distribution was disperse enough. This method of setting the proposal distribution worked well most of the time; we show a plot comparing the default MCMC proposal for BEL inference to the target BEL posterior of a particular dataset in Figure 6.5. Due to the low dimensionality of the target parameter, it was feasible to visually check for

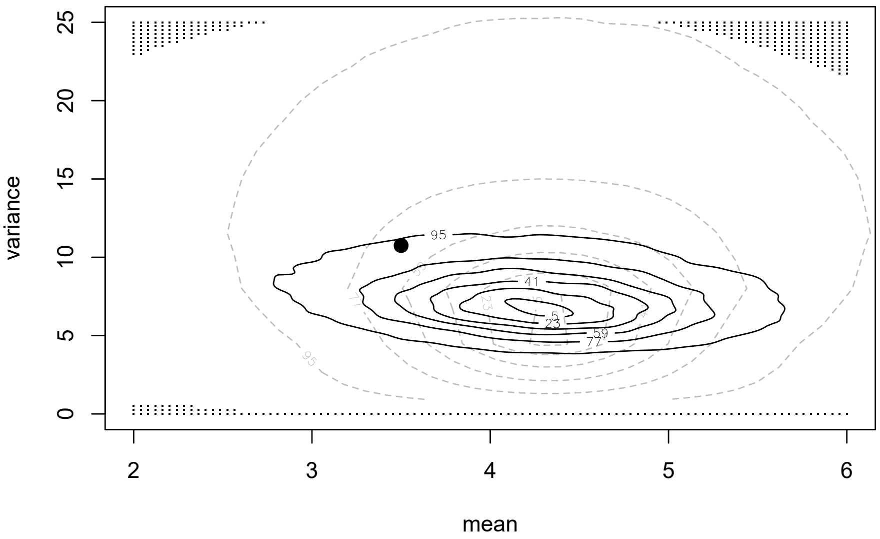


Figure 6.5: Comparison study of Section 6.1: contour plot of a BEL posterior distribution (solid lines) and that of the corresponding BEL MCMC proposal distribution (in grey dotted lines) based on a particular dataset. Square dots mark the locations (on a grid) outside of valid parameter space as given by the convex hull condition. The figure shows the contours under (μ, σ^2) parametrization since the BEL posterior in question is not concentrated around 0 for σ^2 to necessitate a conversion to log variance scale.

the quality of the MCMC proposal distribution. When the aforementioned MCMC proposal did not work well the proposal was adjusted individually until it was wider than the BEL posterior. Due to the time intensive nature of the iterative least squares algorithm for finding R_n at given values of (μ, σ^2), each MCMC chain contained just 1×10^4.

Performance of the four competing methods were explored at two sample sizes ($n = 20, n = 50$). At each sample size level, 300 resampling events were performed, and the performance metrics were calculated based on the average from these resampling events. The chosen number of resampling events is not large and therefore some stochastic errors in the results are expected. Results from the simulation study is summarized in Table 6.1 for joint inference and Table 6.2, Table 6.3 for marginal inference.

Table 6.1: Comparison study of Section 6.1: results for estimating μ and σ^2 jointly. Performance metrics were computed based on 300 resampling events at each sample size n.

	Method	Estimated coverage probability of 95% CR	Average size of 95% CR
$n = 20$	TAB	0.980	30.129
	BB	0.803	30.009
	BEL	0.837	25.755
	GB	0.783	25.946
$n = 50$	TAB	0.96	14.571
	BB	0.880	14.009
	BEL	0.893	13.330
	GB	0.87	13.476

Table 6.2: Comparison study of Section 6.1: results for marginal inference of μ. Performance metrics were computed based on 300 resampling events at each sample size n.

	Method	Estimated coverage probability of 95% CI	Average size of 95% CI	Estimated bias	Average quadratic risk
$n = 20$	TAB	0.947	2.611	-0.051	0.917
	BB	0.930	2.695	-0.095	1.027
	BEL	0.947	2.592	-0.084	0.914
	GB	0.933	2.633	-0.064	0.934
$n = 50$	TAB	0.937	1.726	0.039	0.398
	BB	0.923	1.759	0.009	0.422
	BEL	0.943	1.788	0.015	0.423
	GB	0.937	1.768	0.053	0.408

Table 6.3: Comparison study of Section 6.1: results for marginal inference of σ^2. Performance metrics were computed based on 300 resampling events at each sample size n.

	Method	Estimated coverage probability of 95% CI	Average size of 95% CI	Estimated bias	Average quadratic risk
$n = 20$	TAB	1.000	9.597	-0.607	10.537
	BB	0.830	9.857	-1.010	16.709
	BEL	0.877	8.316	-0.863	9.924
	GB	0.817	8.370	-1.596	10.551
$n = 50$	TAB	0.98	7.206	-0.425	6.324
	BB	0.883	7.094	-0.558	7.587
	BEL	0.907	6.380	-0.501	5.700
	GB	0.857	6.776	-1.083	6.462

We note that small sample Frequentist properties of Bayesian procedures are dependent on the prior distribution, that had the prior been exceptionally strong then all of results (except those for the BB) would be concentrated at the correct value, and the 95% credibility region would always cover the true value. However, we had assumed a subjective prior distribution with a moderate amount of uncertainty (Figure 6.2), and as such the coverage of the methods under comparison was less than certain.

In joint inference, only the TAB method achieved a coverage level close to nominal at the sample sizes tested. Average size of the 95% credible region for TAB posterior was similar to competing methods while having achieved much better coverage. For marginal inference on the variance parameter, the estimated coverage probability of TAB method was closest to nominal among competitors, while the average length of the credible interval stayed small. The estimated bias and expected quadratic risk were also small relative to that of competing methods. Poor performance of GB estimator for variance parameter was likely due to the skewness of X^2 (see Figure 6.1) conflicting with the implications of the loss function on the likelihood model as being Gaussian. Similarly, posteriors for BB and BEL are related to the empirical distribution, which was likely unrepresentative of the variance of the population given the skewness of the distribution of X^2. Interestingly, the TAB method may have accounted for this with the weighting by $p_{\mu,\sigma^2}/q^{\Pi}_{\mu,\sigma^2}$, with the induced proposal prior q^{Π}_{μ,σ^2} capturing the behaviour of variance parameter when random measures for the observable put large weights on a small number of support points.

In marginal inference of the mean, all methods performed well. The results in Table 6.2 did not highlight one method above others as being superior if we factor in the stochastic error from the relatively low number of resampling events.

6.2 Estimating the mean with data missing at random

Let us consider the problem of estimating the mean of some random variable Y when some observations are missing at random (MAR), while an auxiliary variable X is observed always. Let C be the indicator for whether variable Y is observed. When Y is said to be missing at random, it means that $C \perp Y|X$, or equivalently, that $\mathbb{P}[C|X,Y] = \mathbb{P}[C|X]$. The observed data is

$$(c_i y_i, c_i, x_i), \quad i = 1, \ldots, n.$$

Many of the methodologies for dealing with missing data can be found in Molenberghs et al. (2014).

To aid the view of parameters as functionals of the distribution for observables, we let $\mathbb{E}_{(F)}[\cdot]$ denote integral transform of measurable functions with respect to F, a distribution for the observable; we hope this notation makes clear that F itself is an argument of the integral transform. Under a Frequentist semiparametric setting, the target parameter $\mathbb{E}_0[Y]$ can be identified via several candidate estimating equations. Letting $\mathbb{P}_0[C|X]$ be the true conditional probability of Y being observed given X, and $\mathbb{E}_0[Y|X]$ be the true conditional mean of Y given X.

We consider the following estimating functions,

$$g_{\text{IPW}}(m_y, \mathbb{P}_0[C=1|X], F_0) := \mathbb{E}_{(F_0)}\left[\frac{CY}{\mathbb{P}_0[C=1|X]}\right] - m_y, \tag{6.1}$$

$$g_{\text{IPW2}}(m_y, \mathbb{P}_0[C=1|X], F_0) := \mathbb{E}_{(F_0)}\left[\frac{C}{\mathbb{P}_0[C=1|X]}(Y - m_y)\right], \tag{6.2}$$

$$g_{\text{AIPW}}(m_y, \mathbb{P}_0[C=1|X], \mathbb{E}_0[Y|X], F_0) := \mathbb{E}_{(F_0)}\left[\frac{CY}{\mathbb{P}_0[C=1|X]} - \frac{C - \mathbb{P}_0[C=1|X]}{\mathbb{P}_0[C=1|X]}\mathbb{E}_0[Y|X] - m_y\right]. \tag{6.3}$$

When set to equal to 0, the above estimating equations all identify $\mathbb{E}_0[Y]$ if $\mathbb{E}_0[Y|X]$ and $\mathbb{P}_0[C=1|X]$ of the data-generating mechanism are known.

In most applications, $\mathbb{P}_0[C=1|X]$ and $\mathbb{E}_0[Y|X]$, and F_0 are unknown. A more useful construction of the estimating equations is obtained when we substitute proxy models for $\mathbb{E}_0[Y|X]$ and $\mathbb{P}_0[C=1|X]$. We let

$$\hat{p}(x|\psi_0, \psi_1) = [1 + \exp\{-(\psi_0 + \psi_1 x)\}]^{-1}$$

be a class of proxy models of $\mathbb{P}_0[C=1|X]$, and

$$\hat{\zeta}(x|\beta_0, \beta_1) = \beta_0 + \beta_1 x$$

be a class of proxy models of $\mathbb{E}_0[Y|X]$. We can further define "closest" proxy models by metrics given in terms of estimating equations, for example

$$g_{\text{logistic}}(\psi_0, \psi_1, F_0) := \mathbb{E}_{(F_0)}[(1 \quad X)^\top (C - (1 + \exp(-\psi_0 + \psi_1 X))^{-1})]$$
$$g_{\text{LM}}(\beta_0, \beta_1, F_0) := \mathbb{E}_{(F_0)}[(1 \quad X)^\top (Y - (\beta_0 + \beta_1 X))],$$

such that $\{(\hat{\psi}_0, \hat{\psi}_1) \in \mathbb{R}^2 : g_{\text{logistic}}(\psi_0, \psi_1, F_0) = 0\}$ further specifies $\hat{p}(x|\hat{\psi}_0, \hat{\psi}_1)$ to be the closest logistic regression model to $\mathbb{P}_0[C=1|X]$. Setting $\{(\hat{\beta}_0, \hat{\beta}_1) \in \mathbb{R}^2 : g_{\text{LM}}(\beta_0, \beta_1, F_0) = 0\}$ specifies $\hat{\zeta}(x|\hat{\beta}_0, \hat{\beta})$ as the closest linear model to $\mathbb{E}_0[Y|X]$.

When approximating models are used in Eqn. (6.1) - (6.3) instead of the data-generating model, we have the following functionals of $\hat{F}_n$ which are asymptotically consistent for $\mathbb{E}_0[Y]$ when either $\hat{p}$ or $\hat{\zeta}$ is correct:

$$\mu_{\text{IPW}}(\hat{F}_n) := \left\{ m_y \in \mathbb{R}; \begin{pmatrix} g_{\text{IPW}}\left(m_y, \hat{p}(\cdot|\psi_0, \psi_1), \hat{F}_n\right) \\ g_{\text{logistic}}(\psi_0, \psi_1, \hat{F}_n) \end{pmatrix} = 0^{3\times 1} \right\}, \tag{6.4}$$

$$\mu_{\text{IPW2}}(\hat{F}_n) := \left\{ m_y \in \mathbb{R}; \begin{pmatrix} g_{\text{IPW2}}\left(m_y, \hat{p}(\cdot|\psi_0, \psi_1), \hat{F}_n\right) \\ g_{\text{logistic}}(\psi_0, \psi_1, \hat{F}_n) \end{pmatrix} = 0^{3\times 1} \right\}, \tag{6.5}$$

and

$$\mu_{\text{AIPW}}(\hat{F}_n) := \left\{ m_y \in \mathbb{R}; \begin{pmatrix} g_{\text{AIPW}}\left(m_y, \hat{p}(\cdot|\psi_0, \psi_1), \hat{\zeta}(\cdot|\beta_0, \beta_1), \hat{F}_n\right) \\ g_{\text{logistic}}(\psi_0, \psi_1, \hat{F}_n) \\ g_{\text{LM}}(\beta_0, \beta_1, \hat{F}_n) \end{pmatrix} = 0^{5\times 1} \right\}. \tag{6.6}$$

These functionals of $\hat{F}_n$ in fact correspond to famous Frequentist estimators of $\mathbb{E}_0[Y]$ (Robins et al., 1994; Horvitz and Thompson, 1952). Although Eqn. (6.4) - (6.6) are stated with the arguments being $\hat{F}_n$, they indicate the form of the functionals μ_{IPW}, μ_{IPW2}, and μ_{AIPW} for any F that is a distribution for the observable.

In a Bayesian formulation, we may obtain consistent Bayesian inference with μ_{IPW}, μ_{IPW2}, and μ_{AIPW} when our prior distribution over the space of random measures for the observables is weakly consistent, based on Section 3.3 of this thesis. In this case, the induced posterior distributions for μ_{IPW} and μ_{IPW2} will be asymptotically consistent for the mean of the data-generating distribution when the form of $\hat{p}$ is correct, whereas μ_{AIPW} will be consistent if either $\hat{p}$ or $\hat{\zeta}$ is correctly specified. Bayesian/Frequentist inference for the mean of the data-generating distribution works well in practice with any of the above functionals, with μ_{AIPW} being very efficient when the true model has conditional mean $\mathbb{E}[Y|X] = \beta_0 + \beta_1 X$.

To begin our simulation study, let us consider the following data generating distribution with the true $\mathbb{E}_0[Y] = 6$,

Model 6.5. (Data-generating distribution in Section 6.2)

$$\begin{aligned} &X \sim \text{Normal}(\mu = 10, \sigma^2 = 100) \\ &e \sim \text{Normal}(\mu = 0, \sigma^2 = 4) \\ &Y|X, e = 1 + 0.5X + e \\ &C|X \sim \text{Bernoulli}\left(p = \{1 + \exp\left(1 - 0.1X\right)\}^{-1}\right) \end{aligned}$$

In this data-generating model, $C|X$ follows a logistic regression model, and $Y|X$ follows a linear regression model, therefore the functionals μ_{IPW}, μ_{IPW2} and μ_{AIPW} are all asymptotically consistent, and μ_{AIPW} is of optimal efficiency. Here, we provide yet another functional, μ_{AIPW2}, defined as

$$\mu_{\text{AIPW2}}(F) := \left\{ m_y \in \mathbb{R}; \begin{pmatrix} g_{\text{AIPW}}\left(m_y, \hat{p}(\cdot|\psi_0, \psi_1), \hat{\zeta}(\cdot|\beta_0, \beta_1), F\right) \\ g_{\text{logistic}}(\psi_0, \psi_1, F) \\ g^{\star}_{\text{LM}}(\beta_1, F) \end{pmatrix} = 0^{4\times 1} \right\},$$

which uses an incorrect class of proxy regression model $\mathbb{E}[Y|X] = \beta_1 X$, defined via estimating equation

$$g^{\star}_{\text{LM}}(\beta_1, F) := \mathbb{E}_{(F)}[(1 \quad X)^{\top}(Y - (\beta_1 X))].$$

The estimator μ_{AIPW2} can be thought of as the intermediate between assuming no regression relationship between X and Y, i.e. μ_{IPW}, and assuming a fully correct relationship between X and Y, i.e. μ_{AIPW}. All four functionals, μ_{IPW}, μ_{IPW2}, μ_{AIPW}, and μ_{AIPW2}, are examined in the simulation study to follow.

Given the data-generating mechanism above, we also present the population distribution of $Y/\mathbb{P}_0(C = 1|X)$ and $(Y - \mathbb{E}_0[Y|X])/\mathbb{P}_0(C = 1|X)$ in Figure 6.6. Note that the density function of $Y/\mathbb{P}_0(C = 1|X)$ is highly skewed, with extremely long left tail.

We assumed the same subjective prior distribution for all functionals under examination, for any method requiring a subjective prior distribution. This marginal prior is given in Model 6.6. We assumed different Bayesian priors for modelling the observables depending on the functional of interest. The Bayesian prior for inferring μ_{IPW} is given in Model 6.7. The Bayesian priors for inferring the other functionals all had the same form as Model 6.7, but with the weighting function modified accordingly, e.g. $m = p_{\mu_{\text{IPW2}}}/q^{\Pi}_{\mu_{\text{IPW2}}}$ if inferring μ_{IPW2}, and so forth. The proposal model we used to construct our TA model is given in Model 6.8.

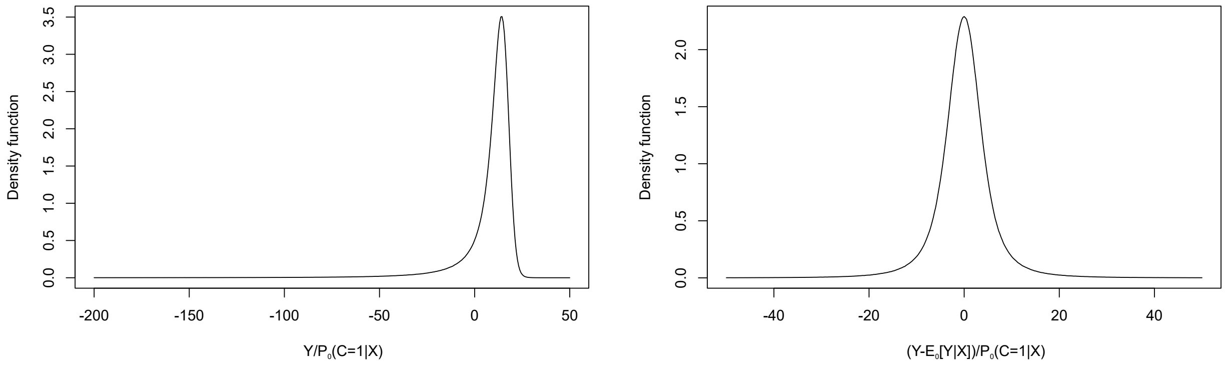


Figure 6.6: Comparison study of Section 6.2: density functions of $Y/\mathbb{P}_0(C=1|X)$ and $(Y-\mathbb{E}_0[Y|X]/\mathbb{P}_0(C=1|X))$ according to the MAR mechanism under Model 6.5.

Model 6.6. (Subjective prior for the unconditional mean of Y in Section 6.2)

$$p_{\mu_{\text{IPW}}} = p_{\mu_{\text{IPW2}}} = p_{\mu_{\text{AIPW}}} = p_{\mu_{\text{AIPW2}}} =$$
$$\text{Normal}(\mu = 6, \sigma^2 = 9).$$

Model 6.7. (Bayesian prior for TAB inference on μ_{IPW} in Section 6.2)

$$\begin{aligned}(X, C, CY) &\sim F_{X,C,CY}\\ F_{X,C,CY} &\sim \text{TA}(m = p_{\mu_{\text{IPW}}}/q^{\Pi}_{\mu_{\text{IPW}}}, \mathcal{P}_{\Pi} = \text{Model 6.8}).\end{aligned}$$

Model 6.8. (TAB proposal model in Section 6.2)

$$
\begin{aligned}
F_{X,C,CY} &\sim \text{DP}(\phi, G_0') \\
\phi &= 0.5 \\
h &= 1 \times 10^{-4} \\
G_0' &= \text{ Discretized version of } G_{0X} \times G_{0C} \times G_{0CY}(\cdot|C), \\
&\quad \text{with mass assigned to points } \{(ih, c, jh)|i, j \in \mathbb{Z}\} \text{ for fixed bin width } h \text{ s.t.} \\
&\quad G_0'(X = ih, C = c, CY = jh|h) = G_{0C}(c) \times \\
&\qquad \int\int \mathbb{1}\left[x \in \left(ih - \frac{h}{2}, ih + \frac{h}{2}\right)\right] \times \\
&\qquad \mathbb{1}\left[cy \in \left(jh - \frac{h}{2}, jh + \frac{h}{2}\right)\right] \mathrm{d}G_{0X}(x)\mathrm{d}G_{0CY}(y|c) \\
G_{0X} &= \text{Normal}(\mu = 10, \sigma^2 = 1) \\
G_{0C} &= \text{Bernoulli}(0.2) \\
G_{0CY}(\cdot|C = 1) &= \text{Normal}(\mu = 0, \sigma^2 = 50^2) \\
G_{0CY}(\cdot|C = 0) &= 0 \text{ with probability } 1.
\end{aligned}
$$

The hyperparameters of the TAB proposal model were selected based on an examination of prior-to-posterior update mechanism. As ϕ tends to 0 the TAB proposal posterior becomes more similar to the Bayesian bootstrap, while the TAB proposal prior becomes harder to sample from. Therefore we chose ϕ of 0.5, which has worked well in past experience. Several considerations influenced our choice of a base distribution for the TAB proposal model. We noted that the functionals of interest involve weighting either Y or the residuals of linear regression by the inverse of $\hat{p}$. As such, these functionals are extremely sensitive to extreme values of X in the support of the distribution for observables. We chose a TAB base distribution G_{0X} with a small variance as to not place too much weight in the extreme values of the observable X. Secondly, we decided that $G_{0CY}(\cdot|C, X)$ should be sufficiently wide to reflect our relative lack of information in Y *a priori*. G_{0C} curiously played a role in terms of how much our DP base distribution contributed to the posterior distribution of the functionals– when $G_{0C}(C = 1)$ is small, the proposal posterior will be more data driven. Hence we chose G_{0C} to be a Bernoulli(0.2) distribution. Given the chosen DP proposal model, Figure 6.7 show a comparison of the induced prior to posterior distributions for the various functionals under consideration, conditional on a randomly chosen set of data points. Note that the induced prior for μ_{IPW} has a probability density function that is increasing to the left. Therefore under small sample sizes, we expected the TAB posterior of μ_{IPW} to assign more weight to larger values of μ_{IPW} as compared to the BB posterior. In the simulation, each TAB posterior distribution was obtained via an MCMC chain with 1×10^5 runs. The performance of TAB inference over repeated sampling is summarized in Table 6.5.

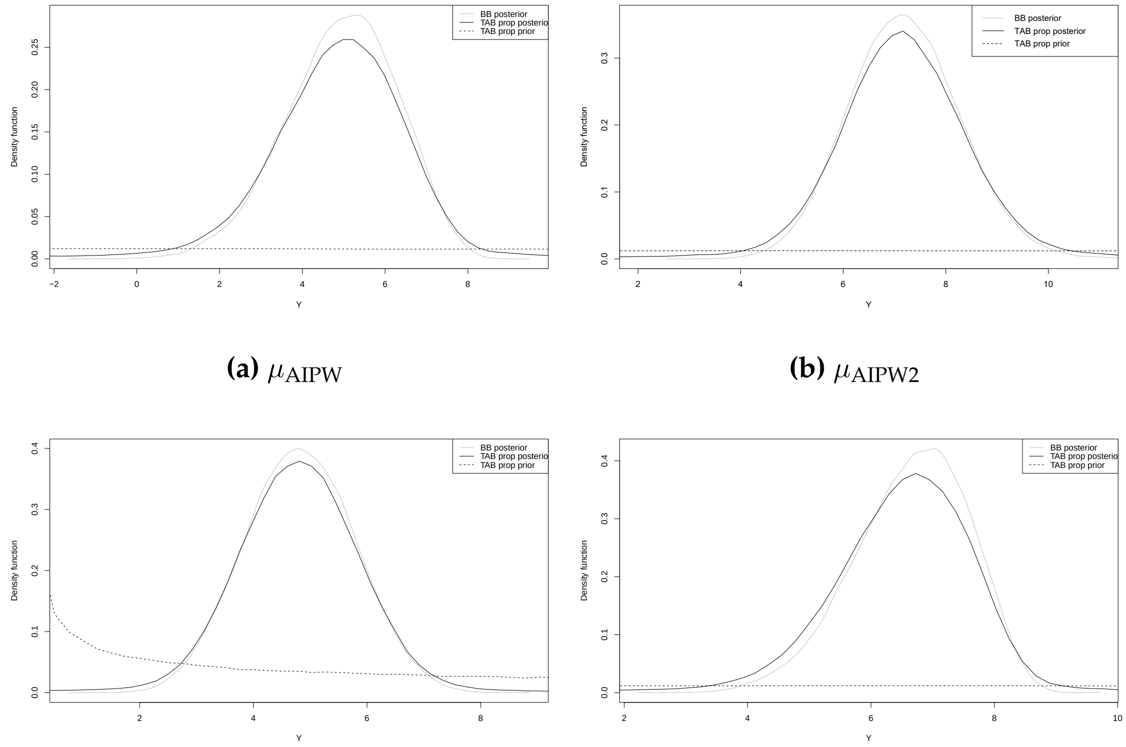
Bayesian bootstrap was implemented by direct sampling in the simulation study. We approximated each BB posterior with 5×10^4 samples. The performance of BB inference is shown in Table 6.4. As for GB inference, we had to cast the functionals as M-estimators to conform to the GB framework. While the original estimating equations defined these functionals via solving for roots, we casted them as the $\arg\min$ of loss functions by noting, for example,

$$\mu_{\text{IPW},GB}(\hat{F}_n) = \arg\min_{t \in \mathbb{R}} \sum_i \left(\frac{(CY)_i}{\hat{p}(X_i, \hat{\psi}_0(\hat{F}_n), \hat{\psi}_1(\hat{F}_n))} - t \right)^2 .$$

Since the tuning parameter of a GB loss function is arbitrary chosen, we chose it based on asymptotic arguments. Using $\mu_{\text{IPW},GB}$ as an example, as $n \to \infty$, the terms $\frac{(CY)_i}{\hat{p}(X_i \hat{\psi}_0(\hat{F}_n), \hat{\psi}_1(\hat{F}_n))}$ become approximately independent. Therefore we let the tuning parameter for $\mu_{\text{IPW},GB}$ be 0.5 times the inverse of the sample variance of $\frac{(CY)_i}{\hat{p}(X_i \hat{\psi}_0(\hat{F}_n), \hat{\psi}_1(\hat{F}_n))}$. The resulting GB posteriors for the target functionals all had a Gaussian form, therefore we were able to sample from them directly and efficiently. The performance of GB inference is summarized in Table 6.6.

It was rather challenging to obtain the BEL inference. The dimensionality of estimating equations for the target functionals was between $\mathbb{R}^3$ to $\mathbb{R}^5$, which meant that in the worst case we had four nuisance parameters. As we had specified the subjective prior marginally for the mean of Y, we had to profile over any nuisance parameters in the set of $\{\beta_0, \beta_1, \psi_0, \psi_1\}$ in the PEL prior to running the MCMC, to provide an apples-to-apples comparison of the BEL to its competitors. Unfortunately, profiling of the PEL function over nuisance parameters is rather computationally difficult and time consuming. As a compromise, the BEL posterior distribution was obtained for only two datasets. These results are shown, along with inference obtained under competing methods for the same data, in Figures 6.8 and 6.9.

Results from the simulation study show that, regardless of method chosen, inference based on μ_{AIPW} and μ_{AIPW2} exhibited good Frequentist properties. The GB method underperformed when it came to inference based on either μ_{IPW} or μ_{IPW2}, having much longer interval estimates on average than its competitors, and subsequently much higher than nominal coverage probability. This behaviour was anticipated, given that the skewness in the distribution of $Y/\mathbb{P}_0(C = 1|X)$ (Figure 6.6) under the data-generating mechanism conflicted with an assumption of Gaussian likelihood that is implicated by the GB loss function. As for the Bayesian bootstrap, its coverage probabilities were lower than nominal in general, and were much lower than nominal for the functionals μ_{IPW} and μ_{IPW2}; this observation conforms with the deficiency of not utilizing subjective prior information. In contrast, the TAB method performed fairly well in all metrics for each of the functional parameters we investigated. Due to the few number of examples we obtained for BEL, we can only suspect that its performance is likely somewhere in between TAB and GB, judging from on the two examples given in Figures 6.8 and 6.9. However, the computational difficulty of obtaining BEL posteriors in the presence of nuisance parameters is a big disadvantage of BEL over its competitors.



(a) μ_{AIPW}

(b) μ_{AIPW2}

(c) μ_{IPW}

(d) μ_{IPW2}

Figure 6.7: Comparison study of Section 6.2: induced TAB proposal prior/posterior distributions on the target functionals for an example dataset with 20 samples generated under MAR mechanism under Model 6.5. Each panel represents the results for the functional given by corresponding subplot caption. Black solid lines represent proposal posteriors, black dotted lines represent proposal priors, while grey solid lines show Bayesian bootstrap posteriors, provided for comparison purposes.

Table 6.4: Comparison study of Section 6.2: results from Bayesian bootstrap marginal inference of the mean of Y with data missing at random. Performance metrics were computed based on 1000 resampling events at each sample size n.

	Method	Estimated coverage probability of 95% CI	Average size of 95% CI	Estimated bias	Average quadratic risk
$n = 20$	μ_{AIPW}	0.918	5.111	-0.006	3.949
	μ_{AIPW2}	0.926	5.121	0.013	3.816
	μ_{IPW}	0.915	4.950	0.072	4.218
	μ_{IPW2}	0.809	4.789	0.314	4.952
$n = 50$	μ_{AIPW}	0.920	3.223	-0.016	1.500
	μ_{AIPW2}	0.924	3.280	-0.012	1.540
	μ_{IPW}	0.908	3.279	0.017	1.792
	μ_{IPW2}	0.868	3.511	-0.012	2.132

Table 6.5: Comparison study of Section 6.2: results from θ-augmented Bayes marginal inference of the mean of Y with data missing at random. Performance metrics were computed based on 1000 resampling events at each sample size n.

	Method	Estimated coverage probability of 95% CI	Average size of 95% CI	Estimated bias	Average quadratic risk
$n = 20$	μ_{AIPW}	0.976	4.951	0.017	2.837
	μ_{AIPW2}	0.979	4.986	-0.019	2.842
	μ_{IPW}	0.972	4.678	0.216	2.647
	μ_{IPW2}	0.933	4.869	0.276	3.420
$n = 50$	μ_{AIPW}	0.962	3.352	-0.017	1.402
	μ_{AIPW2}	0.959	3.391	-0.027	1.431
	μ_{IPW}	0.951	3.274	0.095	1.399
	μ_{IPW2}	0.959	3.588	0.109	1.776

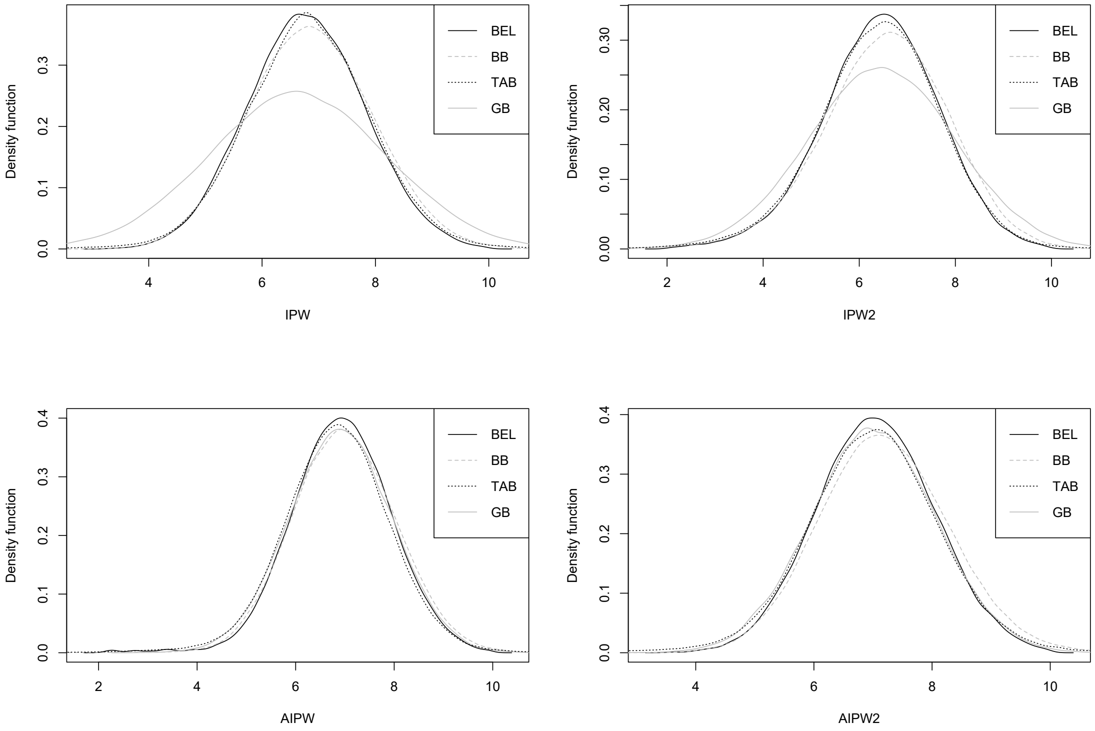


Figure 6.8: Comparison study of Section 6.2: example of posterior inference for various functionals estimating the mean of Y with data missing at random. Each panel represents the conditional posterior distribution for the functional indicated below the plot. Four methods of inference, Bayesian empirical likelihood, Bayesian bootstrap, θ-augmented Bayes, and general Bayes, are presented for each target functional. All posterior inference is based on data set no.1 with 20 samples drawn from the data generating distribution of Model 6.5.

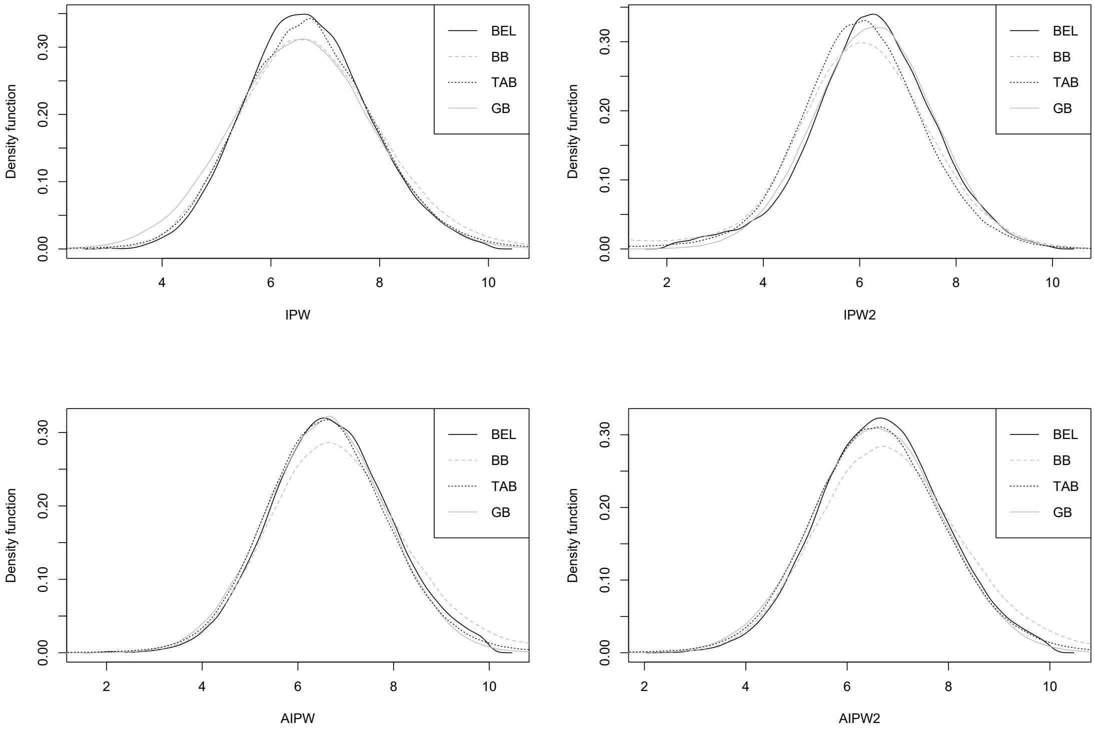


Figure 6.9: Comparison study of Section 6.2: example of posterior inference for various functionals estimating the mean of Y with data missing at random. Each panel represents the conditional posterior distribution for the functional indicated below the plot. Four methods of inference, Bayesian empirical likelihood, Bayesian bootstrap, θ-augmented Bayes, and general Bayes, are presented for each target functional. All posterior inference is based on data set no.2 with 20 samples drawn from the data generating distribution of Model 6.5.

Table 6.6: Comparison study of Section 6.2: results from general Bayes marginal inference of the mean of Y with data missing at random. Performance metrics were computed based on 1000 resampling events at each sample size n.

	Method	Estimated coverage probability of 95% CI	Average size of 95% CI	Estimated bias	Average quadratic risk
$n = 20$	μ_{AIPW}	0.935	4.516	0.030	2.688
	μ_{AIPW2}	0.957	4.784	-0.016	2.743
	μ_{IPW}	0.996	6.048	0.087	3.337
	μ_{IPW2}	0.996	6.212	0.259	3.833
$n = 50$	μ_{AIPW}	0.938	3.119	-0.011	1.307
	μ_{AIPW2}	0.947	3.249	-0.023	1.363
	μ_{IPW}	0.995	4.235	0.050	1.809
	μ_{IPW2}	0.987	4.256	0.112	2.047

Chapter 7

Discussion

The theory of Bayesian inference based on θ-augmented models builds on de Finetti's representation theorem in a simple manner. As a Bayesian prior for modelling a sequence of exchangeable observables $(X_1, \ldots, X_n)$, θ-augmented Bayesian inference proceeds by specifying

$$\begin{aligned} X &\sim F_X \\ F_X &\sim \text{TA}\left(m = \frac{p_\theta}{q_\theta^\Pi}, \mathcal{P}_\Pi\right), \end{aligned}$$

which guarantees the marginal prior density function of θ under this model to be exactly p_θ. The construction of the TA model with a θ-augmented measure provides a way to adjust any dominated proposal model $\mathcal{P}_\Pi$, via a simple weighting function, to achieve the desired marginal prior distribution for a functional parameter that is not part of the parametrization of the likelihood function of the model space. The weighting which happens when defining a θ-augmented measure can be regarded as a change of measure while indexing the proposal model space via subspaces generated by the functional.

Posterior TAB inference relates to the proposal model in an extremely elegant way, where

$$q_{\theta|\tilde{x}_n}^{\Pi^\star} \propto q_{\theta|\tilde{x}_n}^{\Pi} \frac{p_\theta}{q_\theta^\Pi},$$

with $q_{\theta|\tilde{x}_n}^{\Pi}$ denoting the induced density function of $\theta|\tilde{x}_n$ under the proposal model, and

$q^{\Pi\star}_{\theta|\tilde{x}_n}$ denoting the induced density function of $\theta|\tilde{x}_n$ under the TA model. In practice, the induced functions q^{Π}_{θ} and $q^{\Pi}_{\theta|\tilde{x}_n}$ will have to be estimated. However we showed in Chapter 4 that, even when q^{Π}_{θ} and $q^{\Pi}_{\theta|\tilde{x}_n}$ are estimated, our proposed sampling schemes can produce samples that approximate the posterior densities well as long as $q^{\Pi}_{\theta|\tilde{x}_n}$ does not have too much density in the tail regions of q^{Π}_{θ}.

The method provides an alternative to Bayesian conditional models for semiparametric inference. When conditional models are subject to misspecification, inference based on functional parameters of nonparametric models has the advantage that probability statements are not generated by taking the misspecified model to be true. Instead, the TAB method differentiates between a model for the observable and a "proxy" conditional model, so that posterior probabilities are generated under a well behaved model for the observable, while the proxy model helps identify the functional of interest. The trade-off being that the interpretation of what the target parameter represents is less straightforward. For example, when the target is inspired by some conditional mean model with the conditional mean given by $f(Y|X,t)$, a parameter defined as

$$\theta(F) = \arg\min_{t\in\Theta}\{\mathbb{E}_F[(f(Y|X,t) - Y)^2]\},$$

may be interpreted as identifying the "closest" proxy model $\hat{F}$ with the smallest expected squared distance between Y and $f(Y|X,t)$, for which the conditional mean satisfies some parametric form, and $\hat{F}(X) = F(X)$. Sometimes, the interpretation can be much more complicated as is the case with estimating the mean with data missing at random, Section 6.2.

In Chapter 5, we outlined two candidate TAB proposal models for semiparametric inference, which are the Dirichlet process model (with a discrete base measure) and the Dirichlet process mixture model. While the DPM has a better level of realism as a model for continuous observables, it is computationally more cumbersome. In many situations, the mapping from density model F_X to the target functional may be difficult to compute. This problem is sometimes alleviated by choosing a uniform mixing kernel for the DPM, as we did in the example of inferring the coefficients for logistic regression, but in general it is not certain if adopting a uniform kernel can help us in the calculation of every conceivable functional parameter. Furthermore, under the DPM, parametric consistency of a functional parameter defined via an estimating equation is only guaranteed when either the estimating equation is bounded or the support of the observed random variable can be perceived as bounded. Although the DP model is not very realistic as a model for continuous observables, it has proven to induce good parametric TAB inference for various target parameters. The robustness of DP as a proposal model for a while range of problems is not surprising, given that a huge amount of dimensionality reduction occurs when a nonparametric random measures for the observable is collapsed down to a low-dimensional target parameter. As yet another benefit of using a DP as the proposal model, parametric consistency of a functional parameter defined via an estimating equation is guaranteed as long as the estimating equation is integrable.

Since the behaviour of TAB posterior for a target parameter depends on the chosen proposal model, we must make an appropriate choice for the exact specification of $\mathcal{P}_{\Pi}$. We are reminded of the view that commitment to a particular model is equivalent to commitment to the corresponding prior-to-posterior update mechanism (Chapter 1, Hjort et al. (2010)). One can potentially compare proposal models based on effective likelihoods, that is, $q^{\Pi}_{\theta|\tilde{x}_n}/q^{\Pi}_{\theta}$, as a way to compare prior-to-posterior mechanisms in a standard way, concentrating on how marginal subjective prior p_θ is modified by data. Also relating to the selection of $\mathcal{P}_{\Pi}$, since the TAB weighting function m must be integrable, if we commit to a particular proposal model we inherently believe that our subjective prior p_θ is more concentrated than the q^{Π}_{θ} that is induced by the proposal model. Model selection should be restricted to the collection of proposal models with q^{Π}_{θ} wider than our subjective prior p_θ.

In Chapter 5, we observed that the TAB effective likelihood based on the class of infinite kernel mixture models, with carefully chosen hyperparameters, can be made to approximate the Bayesian bootstrap posterior. Let the Bayesian bootstrap posterior of a functional parameter be denoted as $BB(\theta)$. When the effective likelihood of TAB inference is approximately proportional to the BB, the TAB posterior for the target parameter is approximately proportional to $BB(\theta)p_\theta(\theta)$, which should perform well in small sample and asymptotically as the BB is known to behave well in general.

In a sense, the Bayesian bootstrap represents, within the general class of infinite kernel mixture models, the limiting prior-to-posterior update mechanism when the data is at its most informative. This behavior is seen throughout the examples in Chapter 5, where the effective likelihood functions of various DPM/DP proposal models tend towards the BB when we place more confidence in the data. When setting the hyperparameter of the DPM proposal model, we can make the data more informative by committing to a mechanism with less clustering, and smaller variance of the kernels, and, at the same time, reducing the DP precision parameter. By doing so we tune the effective likelihood of the proposal model towards the BB. Conversely, as the BB is neither compatible with subjective prior distribution nor equipped with a proper Bayesian prior, the TA model may also be viewed as a way to extend the Bayesian bootstrap to equip it with a subjective prior. Given this aim, we wonder whether a full TAB analysis is necessary over the simplification of approximating the TAB posterior with $BB(\theta)p(\theta)$. Although we save computation time by approximating a TAB posterior with $BB(\theta)p(\theta)$, the BB is bound by the convex-hull condition, which may be disadvantageous when we suspect the data to be unrepresentative of the population for certain functionals.

Interestingly, in some situations, specification of the hyperparameters of a proposal model may depend on the target of inference if we require the effective likelihood to approximate the BB. As an example, suppose we observe $(X_i, Y_i), i = 1 \ldots, n$, taking the DP model to be the TAB proposal model, increasing the variance of the DP base distribution for X results in an increased variance of the induced prior for the mean of X. However, when the same X is used as a covariate for regression purposes, increasing the variance of the base distribution for X translates to a reduced range in the regression coefficients. Hence to flatten the induced prior density function q_θ^Π when θ is a regression coefficient, one would decrease the variance of the base distribution for covariate X. We may find ourselves choosing different DP base distribution for the same observable random variable depending on the parameter of interest, if our goal is to produce a TAB posterior that is approximately proportional to $BB(\theta)p(\theta)$.

For the purist, as the TAB method is rooted in de Finetti representation theorem, it would feel incoherent to change the model specification when estimating different parameters if the observables have not changed. If theoretical coherence is desired, we could specify the DP base distribution to adhere to our prior views regarding the observable, which usually produces very sensible results, especially if the base distribution is not too different from the empirical distribution for the data. Though given how well the BB performs under Frequentist metrics we may end up sacrificing good Frequentist behavior for the gain of coherence.

Compared to existing (pseudo-)Bayesian semiparametric methods, TAB performs well in small sample inference under the Frequentist metrics mentioned in Chapter 6. We saw the TAB method excel in providing joint inference, while others faired poorly. In estimation of variance parameter, which was the only setting not involving a "location" parameter in our comparison study, TAB was the only method to have achieved close to nominal coverage probability. The BB showed significantly lower than nominal coverage probability in joint inference, which was likely due to the "convex hull condition" being detrimental in small sample inference. Generally speaking, when the BB posterior interval estimates are too short, using the TAB posterior for inference can improve coverage, bias, etc., if quality prior information is present. As for the GB method, we note that it exhibited several downsides. Firstly, implementation of GB inference requires fixing an arbitrary tuning parameter. Secondly, results from Chapter 6 show that the performance of GB posterior could be worse than other methods that involve data-driven likelihoods when the shape of the data is incompatible with the shape of the loss function. This is expected as the GB method did not stem from an attempt to model the observables. BEL posterior can be difficult and time-consuming to implement when there is a large number of nuisance parameters, either due to performing iterative least squares optimization repeatedly to profile out the nuisance parameters in the PEL function and/or requiring sampling in a high dimensional space with irregular domain. In comparison, in TAB inference, subjective prior belief regarding nuisance parameters need not enter the calculations at all, and the MCMC algorithm proposed in Section 4.2 tends to work efficiently when the TAB target posterior is not too different from $q^{\Pi}_{\theta|\tilde{x}_n}$. We note that the simulations in Chapter 6 were only performed at a few sample sizes, and with a low number (between 300-1000) of repeated sampling events. As such, the measures of performance which we reported are only estimates and are subject to stochastic error.

An issue with asymptotics arises in the non-standard use of DPM for modeling a joint vector of observables where one part of the observable is continuous and another part is discrete, as was the case in Section 5.1.3 for logistic regression. The DPM proposal model we used in Section 5.1.3 was devised with the goal of examining the effects of clustering in data points on parametric inference. The particular DPM proposed in Section 5.1.3 is by no means a well-studied nonparametric model for mixed observables, and therefore, further investigation regarding asymptotic consistency is required.

Chapter 8

Conclusion

In this thesis, a fully Bayesian method of semiparametric inference for functional parameters was developed. The TAB model is most useful when coupled with a nonparametric proposal model, to provide a fully Bayesian semiparametric inference. The proposal nonparametric distribution should be well-behaved. In several important use cases, we have shown that the asymptotic consistency of the proposal model for the data-generating distribution guarantees the TAB posterior for the induced functional to be asymptotically consistent.

TAB inference is easy to implement, and directly targets a low-dimensional parameter even when a large number of nuisance parameters are involved in defining the target, requiring only marginal prior for valid statistical inference. On the flip side, technical difficulties currently exist for the joint estimation of a parameter vector of high dimensionality. The denominator, q_θ^Π, of a TAB weighting function is an induced probability density or mass function which most of the time has no closed-form expression and must be estimated. There exists `R` library `ks` which provides computer functions for density estimation of a parameter in $\mathbb{R}^d, d \leq 6$. For the moment it seems that joint inference of a target parameter vector in 7-dimensions or more will require that we program our own density estimation algorithm, possibly via Bayesian mixtures of Gaussian kernels (via a DP), which is a significant task in itself that will likely prevent a wide-scale adoption of TAB inference for high-dimensional parameter vectors. Further research is needed to tackle this problem either from the algorithmic or theoretical front.

We note that the TAB method may be used to translate many existing Frequentist tools into a Bayesian setting, provided that the Frequentist procedure/estimator is expressed as a functional of the empirical data distribution. With the TAB method at our disposal we may begin tackling a wide range of important problems that may not have had a Bayesian solution previously. This broadens the scope of the Bayesian paradigm significantly as Frequentist estimators need not be motivated by a likelihood function. Although we are accustomed to tackling Bayesian inference by firstly analyzing the structure of the sampling distribution/likelihood function, writing it down explicitly as involving the parameter of interest, this mindset can produce less than desirable posterior inference when the likelihood function is known to be misspecified or unavailable. However, it could be that the problem has a simple and well accepted Frequentist solution. For example, the problem of confounding in treatment effects is typically addressed by inverse probability of treatment (IPT) weighting under a Frequentist framework. Application of TAB to this type of problem removes the need for correct structuring of the likelihood function, and the need for specifying a prior distribution for nuisance parameters. Further work is needed to quantify any additional advantages with regards to performance of the TAB method compared with likelihood-based Bayesian methods. The examples included in this thesis are far from comprehensive, and work remains to apply the TAB method to real datasets and under settings where appropriate likelihood functions do not exist. Some of the applications we look forward to include causal inference, survey sampling, longitudinal studies, etc.

While a great variety of functional parameters are compatible with the theory of TAB inference, the number of compatible proposal models is limited based on algorithmic practicality. For the moment, the Dirichlet process model with a discrete base distribution is the only proposal model for which mapping from the distribution for the observable to any general functional parameter can be performed with ease. The limitation to discrete base distribution is necessary so that the DP model is dominated, which is a requirement of TAB inference. Discretization of base distribution is justifiable by the practical limitation that all real world measurements are made with finite precision, as are most computer arithmetic schemes, and is likely not an issue with small sample inference. For better understanding of the theory, it may be interesting to characterize the tolerable precision of measurement at various sample sizes, based on both asymptotic consistency and practical significance.

Even though the Dirichlet process mixture model with Gaussian kernels has been a favourite model for density estimation purposes and well studied, it sees limited use as the proposal model in TAB, due to the difficulty in obtaining closed-form expressions for general functionals of the Gaussian distribution. More research is needed to find fast and reliable ways of approximating any general functional of the Gaussian distribution, so that we may utilize the DPM with Gaussian kernels as the TAB proposal model in a wide range of applications.

We saw in Chapter 5 that TAB inference could be sensitive to hyperparameter specification of the proposal model. In order to generate a particular prior-to-posterior mechanism, often the specification of hyperparameters will depend on the functional of interest. Our investigation into the effects of hyperparameter specification on posterior inference was rather brief; more work can be done to gain better understanding of how to specify the proposal model for various functionals of interest.

This thesis contains some insight regarding the asymptotic behaviour of the TAB posterior for functionals defined via continuous estimating equations when a weakly consistent nonparametric model or the DP is used as the TAB proposal model $\mathcal{P}_{\Pi}$. Work on asymptotics remains for other types of functionals and proposal distributions should these be involved in future applications.

We have, in the examples throughout this thesis, assessed the small sample performance of the TAB posterior. We focused on small sample inference, where a prior distribution has a big role in the posterior inference, as an attempt to showcase the advantages of Bayesian inference. Future work with regards to the evaluation and quantification of large sample performance and computation efficiency of the TAB method is still needed to have a better understanding of the practical merits and demerits of the TAB method.

Last but not least, by requiring that the TAB proposal models be necessarily dominated, we are limited in the class of nonparametric models that can be used to describe the observable random variable. A re-examination of the theory of TAB inference may be interesting to see how this requirement may be relaxed to make more nonparametric methods available as the starting point of Bayesian semiparametric inference for functional parameters.

Bibliography

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Science & Business Media.

Bernardo, J. M. and A. F. Smith (1994). *Bayesian Theory*. John Wiley & Sons.

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

Chatterjee, S. and A. Bose (2005). Generalized bootstrap for estimating equations. *The Annals of Statistics 33*(1), 414–436.

Chaudhuri, S., D. Mondal, and T. Yin (2017). Hamiltonian Monte Carlo sampling in Bayesian empirical likelihood computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(1), 293–320.

Cheng, G. and J. Z. Huang (2010). Bootstrap consistency for general semiparametric m-estimation. *The Annals of Statistics 38*(5), 2884–2915.

Duong, T. (2020). *ks: Kernel Smoothing*. R package version 1.11.7.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis, Third Edition*. CRC press.

Ghosal, S. (1997). A review of consistency and convergence of posterior distribution. In *Varanashi Symposium in Bayesian Inference, Banaras Hindu University*.

Ghosal, S., J. K. Ghosh, and R. Ramamoorthi (1999). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics 27*(1), 143–158.

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

Ghosh, J. K. and R. Ramamoorthi (2003). *Bayesian Nonparametrics*. Springer Science & Business Media.

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

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*(260), 663–685.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*(453), 161–173.

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

Kalli, M., J. E. Griffin, and S. G. Walker (2011). Slice sampling mixture models. *Statistics and Computing 21*(1), 93–105.

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

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

Meredith, M. and J. Kruschke (2020). *HDInterval: Highest (Posterior) Density Intervals*. R package version 0.2.2.

Molenberghs, G., G. Fitzmaurice, M. G. Kenward, A. Tsiatis, and G. Verbeke (2014). *Handbook of Missing Data Methodology*. CRC Press.

Monahan, J. F. and D. D. Boos (1992). Proper likelihoods for Bayesian analysis. *Biometrika 79*(2), 271–278.

Müller, P., F. A. Quintana, A. Jara, and T. Hanson (2015). *Bayesian Nonparametric Data Analysis*. Springer.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics 18*(1), 90–120.

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

Paisley, J. W. and M. I. Jordan (2016). A constructive definition of the beta process. *arXiv preprint arXiv:1604.00685*.

Paisley, J. W., A. K. Zaas, C. W. Woods, G. S. Ginsburg, and L. Carin (2010). A stick-breaking construction of the beta process. In *International Conference on Machine Learning*.

Qin, J. and J. Lawless (1994). Empirical likelihood and general estimating equations. *the Annals of Statistics*, 300–325.

Ray, K. and A. van der Vaart (2020). Semiparametric Bayesian causal inference. *The Annals of Statistics 48*(5), 2999–3020.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association 89*(427), 846–866.

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics 9*(1), 130–134.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica 4*(2), 639–650.

Tsiatis, A. (2007). *Semiparametric Theory and Missing Data*. Springer Science & Business Media.

van der Vaart, A. (2000). *Asymptotic Statistics*, Volume 3. Cambridge University Press.

Vapnik, V. (1992). Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pp. 831–838.

Wu, Y. and S. Ghosal (2010). The L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation. *Journal of Multivariate Analysis 101*(10), 2411–2419.

Yang, Y., X. He, et al. (2012). Bayesian empirical likelihood for quantile regression. *The Annals of Statistics 40*(2), 1102–1131.

Zhao, P., M. Ghosh, J. Rao, and C. Wu (2020). Bayesian empirical likelihood inference with complex survey data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(1), 155–174.

Appendix A

A.1 Implications of general Bayes loss function $l(\theta, x)$ on the likelihood model when it matches the kernel of a location-shift family

Suppose the general Bayes posterior, $\pi_{GB}(\theta|x)$, is considered to be a genuine subjective probability. Suppose we acknowledge that x is variable and hold subjective belief $\pi(x)$.

If the data is regarded as infinitely exchangeable, then de Finetti's representation theorem applies, to provide a structure for $\pi(x)$. Let the representation be

$$\pi(x_1, \dots, x_n) = \int \prod_{i=1}^{n} f(x_i; \theta, \eta) \mathrm{d}\pi(\theta, \eta).$$

Note that when θ is a general functional parameter and η fully specifies the sampling likelihood, i.e. $f(x; \eta, \theta) = f(x; \eta)$, the above representation theorem is still coherent. The joint prior in this case extends from the prior over η, and is

$$\pi(\theta \in A, \eta \in B) := \int \mathbb{1}[\theta(f(\cdot; \eta)) \in A] \mathbb{1}[\eta \in B] \mathrm{d}\pi(\eta).$$

The joint distribution $\pi(\theta, x)$ exists due to θ being a functional $f(x; \eta)$. The conditional distribution by definition is $\pi(\theta|x) := \pi(\theta, x)/\pi(x)$.

Suppose the general Bayes posterior, $\pi_{GB}(\theta|x) \propto \exp(-l(\theta, x)) \times \pi(\theta)$, is interpreted as one's genuine conditional probability, then,

$$\pi(x|\theta) = \pi_{GB}(\theta|x) \cdot \pi(x)/\pi(\theta) = \frac{\exp(-l(\theta, x)) \times \pi(\theta)}{\int \exp(-l(\theta, x))\pi(\theta)\mathrm{d}\theta} \times \frac{\pi(x)}{\pi(\theta)} = \frac{\exp(-l(\theta, x))}{\int \exp(-l(\theta, x))\pi(\theta)\mathrm{d}\theta} \times \pi(x), \tag{A.1}$$

When there are multiple samples, $(x_1, \ldots, x_n)$, by the coherence property in Bissiri et al. (2016),

$$\pi(x_1, \ldots, x_n|\theta) = \frac{\exp(-\sum_{i=1}^n l(\theta, x_i))}{\int \exp(-\sum_{i=1}^n l(\theta, x_i))\pi(\theta)\mathrm{d}\theta}\pi(x_1, \ldots, x_n) \tag{A.2}$$

On the other hand, as we can only hold one genuine subjective belief, we equate the usual marginal Bayesian posterior and the GB posterior for θ. In the case of $\pi(\theta, \eta)$ giving rise to dominated $f(x; \theta, \eta)$, we have that

$$\begin{aligned}\pi(\theta|x) &= \frac{\int f(x; \theta, \eta)\pi(\eta, \theta)\mathrm{d}\eta}{\pi(x)} \\ &= \frac{\left[\int f(x; \theta, \eta)\pi(\eta|\theta)\mathrm{d}\eta\right]\pi(\theta)}{\pi(x)} \\ &= \frac{\exp(-l(\theta, x)) \times \pi(\theta)}{\int \exp(-l(\theta, x))\pi(\theta)\mathrm{d}\theta},\end{aligned}$$

dividing both sides by $\pi(\theta)$ and multiplying by $\pi(x)$, we have that

$$\mathbb{E}_{\eta|\theta}[f(x; \theta, \eta)] := \int f(x; \theta, \eta)\pi(\eta|\theta)\mathrm{d}\eta = \exp(-l(\theta, x)) \times \frac{\pi(x)}{\int \exp(-l(\theta, x))\pi(\theta)\mathrm{d}\theta}.$$

Next, we integrate $\mathbb{E}_{\eta|\theta}[f(x; \theta, \eta)]$ by $\mathrm{d}x$, and notice that $f(x; \theta, \eta)$ is a distribution in x,

$$\begin{aligned}\int \mathbb{E}_{\eta|\theta}[f(x; \theta, \eta)]\mathrm{d}x &= \mathbb{E}_{\eta|\theta}\left[\int f(x; \theta, \eta)\mathrm{d}x\right] \\ &= 1 \\ &= \int \exp(-l(\theta, x)) \times \frac{\pi(x)}{\int \exp(-l(\theta, x))\pi(\theta)\mathrm{d}\theta}\mathrm{d}x,\end{aligned}$$

passing integration into the expectation by Tonelli's theorem. (Tonelli's theorem requires that the measure of x and the measure of (θ, η) be σ-finite, and that $f(x; \theta, \eta)$ be a measurable non-negative function from the product space. The measure $\pi(\theta, \eta)$ over the sample space of (θ, η)) is σ-finite due to being a probability measure.)

And if $\exp(-l(\theta, x))$ is proportional to a distribution of x (call this $h(x; \theta)$) with normalizing constant $c(\theta)$, and denoting $z(x) := \int \exp(-l(\theta, x))\pi(\theta)\mathrm{d}\theta$

$$1 = \int \frac{\pi(x)}{z(x)} \times c(\theta) \times h(x; \theta)\mathrm{d}x.$$

If h is a location shift family (e.g. Gaussian, or Laplace), then $c(\theta) = c$, i.e does not depend on θ. Then,

$$\frac{1}{c} = \mathbb{E}_{h(x;\theta)}\left[\frac{\pi(X)}{z(X)}\right], \forall\theta,$$

and we seek some function of x which has the same expected value regardless of the location of the distribution (while the shape stays the same). This implies that $\frac{\pi(x)}{z(x)} = \frac{1}{c}$. Substituting into Equation (A.1), we have that $\pi(x|\theta) = \exp(-l(\theta, x))/c$.

Extending the argument to arbitrary sample size n, we start with the fact that

$$\pi(\theta|x_1, \cdots, x_n) = \frac{\mathbb{E}_{\eta|\theta}[\prod_{i=1}^n f(x_i;\theta,\eta)]\pi(\theta)}{\pi(x_1, \cdots, x_n)} = \frac{\exp(-\sum_{i=1}^n l(\theta, x_i)) \times \pi(\theta)}{\int \exp(-\sum_{i=1}^n l(\theta, x_i))\pi(\theta)\mathrm{d}\theta},$$

and proceed to dividing both sides by $\pi(\theta)$ and multiplying by $\pi(x_1, \cdots, x_n)$, followed by the multiple integration of $\mathbb{E}_{\eta|\theta}[\prod_{i=1}^n f(x_i;\theta,\eta)]$ in x_1 to x_n, which leads to

$$\frac{\pi(x_1, \ldots, x_n)}{\int \exp(-\sum_{i=1}^n l(\theta, x_i))\pi(\theta)\mathrm{d}\theta} = \frac{1}{c^n}.$$

Noting Eq. (A.2), we have that

$$\pi(x_1, \ldots, x_n|\theta) = \prod_{i=1}^n (\exp(-l(\theta, x_i)))/c^n.$$

It seems that, when the GB loss function $\exp(-l(\theta, x))$ corresponds to some location-shift family in x, the loss function uniquely indicates the likelihood model. The use of GB update did not yield a model-free inference if the GB posterior distribution were taken as a genuine subjective probability, under the assumptions that the samples $(x_1, \ldots, x_n)$ are exchangeable and $f(x;\theta,\eta)$ is dominated $\pi(\theta,\eta)$ *almost surely*.

A.2 Investigating the form of likelihood models that follow coherence property

The goal of general Bayesian inference had been to provide a genuine Bayesian posterior while bypassing having to identify a particular likelihood function in order to avoid model misspecification. The proposition is useful, but it was not explicitly stated if the GB inference is to be taken as a "genuine" posterior belief in the traditional Bayesian sense.

To be a genuine posterior belief distribution, it is clear that one must necessarily quantify their belief distribution regarding the observable data via probabilities and via de Finetti's representation theorem or other assumptions on the model for the observable, as according to the theory of subjective belief, see Bernardo and Smith (1994).

While the likelihood function for any GB inference is unspecified, it is possible that the GB procedure leads to genuine belief distributions in the classical sense by corresponding to an extremely broad and flexible class of likelihood functions without committing to one any of them in particular.

Suppose that there exist a prior distribution of the parameters, $\pi(\theta, \eta)$, where $\pi(\theta)$ has some density, but η may exist in some general space equipped with a σ-algebra and a measure to better accommodate the situation that η may be infinite dimensional. The conditional probability formula still applies, where measure $\pi(\theta, \eta) = \pi(\eta|\theta)\pi(\theta)$, where $\pi(\theta)$ is a density function, but $\pi(\eta|\theta)$ is a general measure. We let the likelihood $f(x; \theta, \eta)$, $(\theta, \eta) \sim \pi(\theta, \eta)$, be dominated *almost surely*.

After one sample,

$$\begin{aligned}\pi(\theta|x_1) &\propto \int f(x_1; \theta, \eta)\pi(\theta, \eta)\mathrm{d}\eta \propto \mathbb{E}_{\eta|\theta}\left[f(x_1; \theta, \eta)\right]\pi(\theta) := \bar{f}(x_1; \theta)\pi(\theta) \\ &= \pi_{GB}(\theta|x_1) \propto \exp(-l(x_1, \theta))\pi(\theta),\end{aligned}$$

where $\mathbb{E}_{\eta|\theta}$ denotes the integral with respect to the conditional measure $\pi(\eta|\theta)$.

By Bissiri's coherence property requirement, we deduce that receiving a second piece of data,

$$\pi(\theta|x_1, x_2) \propto \exp(-l(x_1, \theta))\exp(-l(x_2, \theta))\pi(\theta) \propto \frac{\bar{f}(x_2; \theta)\left[\bar{f}(x_1; \theta)\pi(\theta)\right]}{\int \bar{f}(x_2; \theta)\left[\bar{f}(x_1; \theta)\pi(\theta)\right]\mathrm{d}\theta}. \tag{A.3}$$

Moreover, because f is assumed to produce genuine posterior belief in the classical sense, then by exchangeability,

$$\pi(\theta|x_1, x_2) = \frac{\mathbb{E}_{\eta|\theta}\left[\prod_{i=1}^{2} f(x_i; \theta, \eta)\right]\pi(\theta)}{\int \mathbb{E}_{\eta|\theta}\left[\prod_{i=1}^{2} f(x_i; \theta, \eta)\right]\pi(\theta)\mathrm{d}\theta}, \tag{A.4}$$

with $\mathbb{E}_{\eta|\theta}$ indicating an expectation with regards to the conditional prior distribution $\pi(\eta|\theta)$.

If we assume that the likelihood is factorizable, with the form being

$$f(x;\theta,\eta) = g_x \cdot g_\theta \cdot g_\eta \cdot g_{x\theta} \cdot g_{x\eta} \cdot g_{\theta\eta} \cdot g_{x\theta\eta},$$

where the subscripts denote the variables/parameters that appear in that term. In both Equation (A.3) and (A.4), factors of $f(x;\theta,\eta)$ that depend on only x will be appear in both numerator and denominator, whereas terms that depend on θ in Equation (A.3) and (A.4) will equate, so that the same posterior is obtained via GB sequential update and via full likelihood update.

Just to make things clear, we are interested in equating the terms that depend on θ in

$$\mathbb{E}\left[\prod_{i=1}^{2} f(x_i;\theta,\eta)\right] = g_{x_1}g_{x_2}g_\theta^2 g_{x_1\theta}g_{x_2\theta}\underbrace{\int g_\eta^2 g_{x_1\eta}g_{x_2\eta}g_{\theta\eta}^2 g_{x_1\theta\eta}g_{x_2\theta\eta}\pi(\eta|\theta)\mathrm{d}\eta}_{I_A},$$

and

$$\prod_{i=1}^{2}\mathbb{E}\left[f(x_i;\theta,\eta)\right] = g_{x_1}g_{x_2}g_\theta^2 g_{x_1\theta}g_{x_2\theta}\underbrace{\int g_\eta g_{x_1\eta}g_{\theta\eta}g_{x_1\theta\eta}\pi(\eta|\theta)\mathrm{d}\eta}_{I_B} \times \int g_\eta g_{x_2\eta}g_{\theta\eta}g_{x_2\theta\eta}\pi(\eta|\theta)\mathrm{d}\eta.$$

Case 1: $\pi(\eta|\theta) \neq \pi(\eta)$

Suppose $\pi(\eta|\theta) \neq \pi(\eta)$, then the integrals I_A and I_B are both functions of θ, Except in the fringe case where $\pi(\eta|\theta)$ cancels out the other terms involving θ in I_A and/or I_B. Letting the arbitrary $x_1 = x_2 = x$, $\mathbb{E}_{\eta|\theta}\left[(g_\eta g_{x\eta}g_{\theta\eta}g_{x\theta\eta})^2\right] = \mathbb{E}_{\eta|\theta}\left[g_\eta g_{x\eta}g_{\theta\eta}g_{x\theta\eta}\right]^2$ which implies that the variance will be 0, $\implies \pi(\eta|\theta) = \mathbb{1}[\eta = \eta(\theta)]$, that η is a mapping of θ. Either way, the likelihood function is defined by θ only and therefore f is fully parametrized by θ and the GB loss function fully identifies the likelihood.

Case 2: $\pi(\eta|\theta) = \pi(\eta)$**, but** $g_{\theta\eta} \neq 1$ **or** $g_{x\theta\eta} \neq 1$

if $g_{\theta\eta} \neq 1$ or $g_{x\theta\eta} \neq 1$ then both I_A and I_B will still be functions of θ and we will still require $\mathbb{E}_{\eta|\theta}\left[(g_\eta g_{x\eta}g_{\theta\eta}g_{x\theta\eta})^2\right] = \mathbb{E}_{\eta|\theta}\left[g_\eta g_{x\eta}g_{\theta\eta}g_{x\theta\eta}\right]^2$. To have this, either $\pi(\eta)$ is a point mass distribution, which means we have the likelihood fully parametrized by θ and GB loss function identifying the likelihood fully, or the integrand is constant which implies $g_{\theta\eta} = 1$ and $g_{x\theta\eta} = 1$, leading to Case 3 below.

Case 3: $\pi(\eta|\theta) = \pi(\eta)$**,** $g_{\theta\eta} = 1$ **and** $g_{x\theta\eta} = 1$

In this case, I_A and I_B are no longer functions of θ, only functions of x. There seem to be no contradictions with our assumptions. In this case, likelihood models compatible with the coherence property of Bissiri et al. have the form

$$f(x; \theta\eta) = PL_1(x; \theta) \times PL_2(x; \eta),$$

which indicates that the likelihood model must be strictly factorizable. Furthermore, the exponentiated negative GB loss function $\exp(-l(\theta, x))$ identifies the partial likelihood $PL_1(x; \theta)$.

The conclusion is that, if the GB posterior were to be taken as a genuine belief distribution following the axioms of subjective probability (Bernardo and Smith, 1994), then it corresponds (at the minimum) to the class of likelihood models that are strictly factorizable and the partial likelihood involving θ in particular must be fully identified by the GB loss function.

A.3 Extension of Paisley and Jordan (2016) for efficient sampling from a posterior DP

Under a DP prior, posterior distribution of F_X is again a Dirichlet process, but updated with a new base distribution,

$$F_X \sim \text{DP}\left((\phi + n), \left(\frac{\phi}{\phi + n} G_0' + \frac{n}{\phi + n} \hat{F}_n\right)\right), \tag{A.5}$$

where $\hat{F}_n$ denotes the empirical distribution of the observed data and n denotes sample size. We can also view the posterior DP as having a base distribution that is a mixture of G_0' and δ_{x_k}, $k = 1, \ldots, n$, i.e. Dirac measures at the observed data values.

To obtain an efficient algorithm for drawing approximately from this posterior DP, we first note that, Lemma 3 of Paisley and Jordan (2016) states that, if

$$\begin{aligned} V_i &\sim \text{Beta}(1, a+b), \quad i = 1, \ldots \\ Y_i &\sim \text{Bernoulli}\left(\frac{a}{a+b}\right), \quad i = 1, \ldots \\ \pi &= \sum_{i=1}^{\infty} \left\{ V_i \prod_{j=1}^{i-1} (1 - V_j) \mathbb{I}[Y_i = 1] \right\} \end{aligned}$$

then π has a Beta(a, b) distribution. We take the random variables V_i, $i = 1, \ldots$, to be the ones that define a stick breaking process in the usual way. If we take the random variable Y_i to indicate that the i-th atom of F_X is drawn from one of δ_{x_k}, $k \in \{1, \ldots, n\}$, then the parameters $a = n$ and $b = \phi + n$, and π represents the total weights assigned to δ_{x_k} for a random F_X. If, instead, the variable Y_i is used to indicate that the i-th atom of F_X is drawn from G_0', then the parameters $a = \phi$ and $b = \phi + n$.

By the symmetry in the resulting Beta distribution for π, we deduce that, if we define a sequence of random vectors $(Y_{i,1}, \ldots, Y_{i,n+1}), i = 1, 2, \ldots$ where

$$\begin{aligned} &Y_{i,1} = \mathbb{I}[i\text{-th atom of } F_X \text{ is drawn from } G_0'], \\ &Y_{i,k+1} = \mathbb{I}[i\text{-th atom of } F_X \text{ is drawn from } \delta_{x_k}], \quad k = 1, \ldots, n, \end{aligned} \tag{A.6}$$

then, the SBP in Eqn. (A.5), can thought of as tagged according to $(Y_{i,1}, \ldots, Y_{i,n+1})$. We have that

$$\begin{aligned} V_i &\sim \text{Beta}(1, \phi + n), \quad i = 1, \ldots \\ (Y_{i,1}, \ldots, Y_{i,n+1}) &\sim \text{Categorical}\left(\frac{\phi}{\phi+n}, \frac{1}{\phi+n}, \ldots, \frac{1}{\phi+n}\right), \quad i = 1, \ldots \\ \pi_l &= \sum_{i=1}^{\infty} \left\{ V_i \prod_{j=1}^{i-1} (1 - V_j) \mathbb{I}[Y_{i,l} = 1] \right\}, \quad l = 1, \ldots, n+1, \end{aligned}$$

where $(\pi_1, \ldots, \pi_{n+1})$ has a Dirichlet$(\phi, 1, \ldots, 1)$ distribution. Note that π_1 represents the weight assigned to all atoms of F_X sampled from G_0'.

Let $(B_1, \ldots, B_m)$ be a measurable partition of the sample space of the Dirichlet process described in Eqn. (A.5). Suppose we now expand the tagging variable to be

$$(Y^\star_{i,1}, \ldots, Y^\star_{i,m+n}), \quad i = 1, 2, \ldots$$

where

$$\begin{aligned}
Y^\star_{i,j} &= \mathbb{I}[i\text{-th atom of } F_X \text{ is drawn from } G'_0 \text{ and its value falls in } B_j], \quad j = 1, \ldots, m \\
Y^\star_{i,m+k} &= \mathbb{I}[i\text{-th atom of } F_X \text{ is drawn from } \delta_{x_k}], \quad k = 1, \ldots, n.
\end{aligned}$$

We tag the SBP corresponding to the DP of Eqn. (A.5) with $(Y^\star_{i,1}, \ldots, Y^\star_{i,m+n})$. We have that

$$\begin{aligned}
&V_i \sim \text{Beta}(1, \phi + n), \quad i = 1, \ldots \\
&(Y^\star_{i,1}, \ldots, Y^\star_{i,m+n}) \\
&\qquad \sim \text{Categorical}\left(\frac{\phi}{\phi + n} G'_0(B_1), \ldots, \frac{\phi}{\phi + n} G'_0(B_m), \frac{1}{\phi + n}, \ldots, \frac{1}{\phi + n}\right), \quad i = 1, \ldots \\
&p_l = \sum_{i=1}^{\infty} \left\{ V_i \prod_{j=1}^{i-1} (1 - V_j) \mathbb{I}[Y^\star_{i,l} = 1] \right\}, \quad l = 1, \ldots, m + n,
\end{aligned}$$

which leads to $(p_1, \ldots, p_{m+n})$ having a Dirichlet$(\phi G'_0(B_1), \ldots, \phi G'_0(B_m), 1, \ldots, 1)$ distribution. By elementary properties of the Dirichlet distribution, this means

$$\begin{aligned}
\left(\frac{p_1}{1 - \sum_{j=m+1}^{m+n} p_j}, \ldots, \frac{p_m}{1 - \sum_{j=m+1}^{m+n} p_j}\right) &|(p_{m+1}, \ldots, p_{m+n}) \\
&\sim \text{Dirichlet}(\phi G'_0(B_1), \ldots, \phi G'_0(B_m)).
\end{aligned}$$

Since $(B_1, \ldots, B_m)$ is any partition of the sample space of G'_0, and $p_j, j \in \{1, \ldots, m\}$ represents the weights of all atoms of a random F_X that is attributable to G'_0 and falls within partition B_j, by definition of the Dirichlet process, we have that the conditional random measure generated by the normalized sub-sequence of SBP with atoms drawn from G'_0 follows a DP(ϕ, G'_0) model; we denote this conditional measure as $F_X | \sum_{k=1}^n (Y_{i,k+1}) = 0$, noting equivalence of $Y^\star_{i,m+k}$ and $Y_{i,k+1}$ defined in Eqn. (A.6). A similar line of reasoning leads us to conclude that the conditional measure $F_X | \sum_{k=1}^n (Y_{i,k+1}) = 1$ is supported on the observed data values with weight vector distributed according to the Dirichlet$(1, \ldots, 1)$ distribution.

We can draw an approximate sample of the conditional measure $F_X | \sum_{k=1}^{n}(Y_{i,k+1}) = 0$ by a truncated SBP. An algorithm for efficient approximate sampling from the posterior DP given by Eqn. (A.5) therefore alternates between drawing of conditional random measure $F_X | \sum_{k=1}^{n}(Y_{i,k+1}) = 0$ from G_0', and $F_X | \sum_{k=1}^{n}(Y_{i,k+1}) = 1$ from $\hat{F}_n$, and then combining the two parts via weighting by a sample of

$$\pi_{obs} \sim \text{Beta}(n, \phi + n),$$

that is,

$$F_X = \left(F_X | \sum_{k=1}^{n}(Y_{i,k+1}) = 0 \right) (1 - \pi_{obs}) + \left(F_X | \sum_{k=1}^{n}(Y_{i,k+1}) = 1 \right) \pi_{obs},$$

to construct a draw of random measure F_X from approximately the posterior DP.

A.4 Approximating the DPM when there is practically no clustering

As an approximation to the DPM, we propose the following. Let $\tilde{\eta}_n := (\eta_1, \ldots, \eta_n)$ denote the collection of latent kernel parameters for the data points, where η_i is the latent kernel parameter for the observation x_i. With slight change in notation to match that of Model 5.1, we note that, based on p. 144 of Ghosh and Ramamoorthi (2003), given the latent kernels $\tilde{\eta}_n$, the posterior distribution of the kernel parameter is

$$\text{DP}\left(\phi + n, \frac{\phi}{\phi + n} G_0 + \frac{1}{\phi + n} \sum_{i=1}^{n} \delta_{\eta_i} \right).$$

This suggests that we may first sample $\tilde{\eta}_n | \tilde{x}_n$ according to the Gibbs sampling algorithm for the Polya urn, then, given a sample of $\tilde{\eta}_n | \tilde{x}_n$, we sample additionally

$$F_X | \tilde{\eta}_n, \tilde{x}_n \sim \text{DP}\left(\phi + n, \frac{\phi}{\phi + n} \phi G_0 + \frac{1}{\phi + n} \sum_{i=1}^{n} \delta_{\eta_i} \right),$$

for a joint sample of $(F_X, \tilde{\eta}_n) | \tilde{x}_n$. The limitation here is that F_X can only be sampled approximately. However, using the algorithm in Appendix A.3, we can efficiently obtain an approximation to F_X with a small number of stick breaking moves even if the number of data points is large.

A.5 DPM with uniform kernel

Let the kernel of a DPM model be

$$\mathcal{K}_U(x;\mu,h) = \frac{1}{h}\mathbb{I}\left[x \in \left[\mu - \frac{1}{2}h, \mu + \frac{1}{2}h\right]\right].$$

For the kernel parameter vector (μ, h) we assume it is sampled from the Dirichlet process, with the base distribution,

$$\begin{aligned}
G_0 &= G_{0\mu} \times G_{0h} \\
G_{0\mu}(\mu) &= \sum_{j=1}^{J} p_j \frac{1}{(b_j - a_j)} \mathbb{I}[\mu \in (a_j, b_j)] \\
G_{0h} &= \text{Gamma}(\alpha_h, \beta_h).
\end{aligned}$$

To implement the blocked Gibbs sampling algorithm with truncated DP, we require the following expressions. Given latent label S_i for each observation x_i. For each cluster k,

$$\begin{aligned}
f(\mu_k, h_k|\tilde{S}_n, \tilde{x}_n) &\propto G_{0\mu}(\mu_k) \times G_{0h}(h_k) \prod_{S_i=k} \mathcal{K}(x_i;\mu_k,h_k) \\
&= \left(\sum_{j=1}^{J} p_j \frac{\mathbb{I}[\mu_k \in (a_j,b_j)]}{(b_j - a_j)}\right) f_h(h_k) \prod_{S_i=k} \frac{1}{h_k}\mathbb{I}\left[x_i \in \left[\mu_k - \frac{1}{2}h_k, \mu_k + \frac{1}{2}h_k\right]\right] \\
&= \sum_{j=1}^{J} p_j \frac{\mathbb{I}[\mu_k \in (a_j,b_j)]}{(b_j - a_j)} f_h(h_k) \frac{1}{h^{n_k}} \left(\prod_{S_i=k} \mathbb{I}\left[x_i \in \left[\mu_k - \frac{1}{2}h_k, \mu_k + \frac{1}{2}h_k\right]\right]\right).
\end{aligned}$$

Therefore conditionally,

$$\begin{aligned}
f(h_k|\mu_k, \tilde{S}_n, \tilde{x}_n) &\propto \left(\sum_{j=1}^{J} p_j \frac{\mathbb{I}[\mu_k \in (a_j, b_j)]}{(b_j - a_j)}\right) f_h(h_k) \frac{1}{h_k^{n_k}} \left(\prod_{S_i=k} \mathbb{I}\left[x_i \in \left[\mu_k - \frac{1}{2}h_k, \mu_k + \frac{1}{2}h_k\right]\right]\right) \\
&\propto f_h(h_k)\frac{1}{h_k^{n_k}} \mathbb{I}\left[h_k > 2(\mu_k - X_{(1)}) \text{ and } h_k > 2(X_{(n_k)} - \mu_k)\right] \\
&= \mathbb{I}\left[h_k > 2(\mu_k - X_{(1)}) \text{ and } h_k > 2(X_{(n_k)} - \mu_k)\right] \times \frac{f_h^{\star}(h_k)}{Z(\mu_k; X_{(1)}, X_{(n_k)})},
\end{aligned}$$

where $f_h^{\star}$ is the Gamma distribution with the updated parameter $\alpha^{\star} = \alpha_h - n_k$, $\beta^{\star} = \beta_h$, and

$$Z(\mu_k; X_{(1)}, X_{(n_k)}) = \int (\mathbb{I}\left[h_k > 2(\mu_k - X_{(1)}) \text{ and } h_k > 2(X_{(n_k)} - \mu_k)\right]) f_h^{\star}(h) \mathrm{d}h.$$

The conditional distribution of h_k given $\mu_k, \tilde{S}_n, \tilde{x}_n$ is therefore a truncated Gamma distribution which can be sampled by the inverse CDF method. Note that α_h has to be greater than n_k in order to be a proper distribution.

As for the marginal distribution of $f(\mu_k|\tilde{S}_n, \tilde{x}_n)$, it is

$$\begin{aligned}
f(\mu_k|\tilde{S}_n, \tilde{x}_n) \propto &\left(\sum_{j=1}^{J} p_j \frac{\mathbb{I}[\mu_k \in (a_j, b_j)]}{(b_j - a_j)}\right) \times \\
&\int f_h(h) \frac{1}{h^{n_k}} \mathbb{I}\left[h_k > 2(\mu_k - X_{(1)}) \text{ and } h_k > 2(X_{(n_k)} - \mu_k)\right] \mathrm{d}h \\
\propto &\left(\sum_{j=1}^{J} p_j \frac{\mathbb{I}[\mu_k \in (a_j, b_j)]}{(b_j - a_j)}\right) Z(\mu_k; X_{(1)}, X_{(n_k)}).
\end{aligned}$$

This distribution is piecewise continuous, and supported on a finite range. It can also be sampled via inverse CDF method. Numerical integration can be highly time consuming, we may wish to do this via MCMC, e.g. Metropolis-Hastings.

As for the multivariate kernel for modeling observable (X, C), $X \in \mathbb{R}, C \in \{0, 1\}$, we may choose $\mathcal{K}(x, c) = \mathcal{K}_U(x; \mu, h) \times \left[\theta^c (1-\theta)^{(1-c)}\right]$

To update θ_k for a cluster k, it is

$$\begin{aligned}
Pr(\theta_k|\tilde{S}_n, \tilde{c}_n) &\propto G_{0\theta}(\theta_k) \times \prod_{S_i=k} \left[\theta_k^{c_i}(1-\theta_k)^{(1-c_i)}\right] \\
&= G_{0\theta}(\theta_k) \theta_k^{\sum_{S_i=k} c_i} (1-\theta_k)^{n_k - \sum_{Si=k} c_i}.
\end{aligned}$$

Let us further assume that $G_{0\theta}$ is a Beta$(\alpha_\theta, \beta_\theta)$ distribution,

$$\theta_k | \tilde{S}_n, \tilde{c}_n \sim \text{Beta}(a_\theta + \sum_{S_i = k} c_i, b_\theta + n_k - \sum_{Si=c} c_i).$$

A.6 Expressions involved in the logistic regression estimating equation for DPM with uniform kernel

Assume the following kernel: $\mathcal{K}(x, c) = \frac{1}{h}\mathbb{I}\left[x \in \left[\mu - \frac{1}{2}h, \mu + \frac{1}{2}h\right]\right] \times [\theta^c(1-\theta)1-c]$

To find the transformation $\psi(F_{XC})$ where F_{XC} are mixtures of the kernel above, the following integrals are required.

For each kernel k, the integrals of interest appearing the estimating equation for logistic regression are:

$$\begin{aligned}\mathbb{E}_k\left[\frac{1}{1+\exp\left(-(\psi_0+\psi_1 X)\right)}\right] &= \int_{\mu_k - 0.5h_k}^{\mu_k + 0.5h_k} \frac{1}{h_k}\frac{1}{1+\exp\left(-(\psi_0+\psi_1 x)\right)}\mathrm{d}x \\ &= \frac{1}{h_k}\frac{1}{\psi_1}\log\left(1+e^{(\psi_0+\psi_1 x)}\right)\Big|_{a_k}^{b_k},\end{aligned}$$

where $b_k = \mu_k + 0.5h_k$, $a_k = \mu_k - 0.5h_k$. This expression was obtained via Wolfram Alpha and Mathematica.

Similarly, we found that

$$\begin{aligned}\mathbb{E}_k\left[\frac{X}{1+\exp\left(-(\psi_0+\psi_1 X)\right)}\right] &= \int_{\mu_k - 0.5h_k}^{\mu_k + 0.5h_k} \frac{1}{h_k}\frac{x}{1+\exp\left(-(\psi_0+\psi_1 x)\right)}\mathrm{d}x \\ &= \frac{1}{h_k}\frac{1}{\psi_1^2}\left\{\text{Li}_2\left(-e^{(\psi_0+\psi_1 x)}\right) + \psi_1 x \log\left(1+e^{(\psi_0+\psi_1 x)}\right)\right\}\Big|_{a_k}^{b_k},\end{aligned}$$

where Li_2 denotes the dilogarithm function.

A.7 Comparison of Bayesian semiparametric methods for linear regression

As yet another comparison of the performance of various Bayesian semiparametric methods for inferring functional parameters, we conducted a simulation study with the target

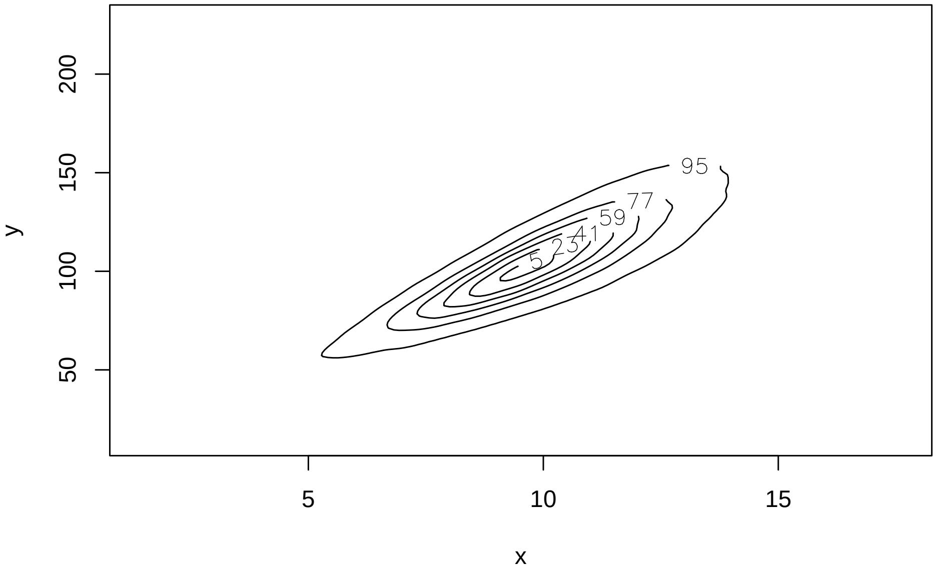


Figure A.1: Comparison study of Section A.7 estimating linear regression parameters: contour plot of the data-generating distribution as given by Model A.1.

of inference being the parameters of linear regression. The data was generated according Model A.1, with contour plot for (X, Y) under this model shown in Figure A.1.

Model A.1. (Data-generating mechanism for Appendix A.7)

$$\begin{aligned} X &\sim \mathcal{N}(\mu = 10, \sigma^2 = 9), \\ Y &\sim 5 + 10x + \epsilon \\ \epsilon &\sim \text{Normal}(\mu = 10, \sigma^2 = X^2). \end{aligned}$$

The functional

$$\beta(F_{XY}) = \underset{(b_0,b_1)\in\mathbb{R}^2}{\arg\min} \int (y - (b_0 + b_1 x))^2 \, \mathrm{d}F_{XY} \in \mathbb{R}^2$$

defines the parameter of interest. As the structure of $F_{Y|X}$ from the data generating model is indeed linear, we expect good Bayesian procedures to be consistent for the true regression parameter values, that is, $(5, 10)$.

We let the subjective prior distribution $p(\beta_0, \beta_1)$ be

Model A.2. (Subjective prior in Appendix A.7)

$$
\begin{aligned}
p_{\beta_0,\beta_1}(\beta_0,\beta_1) &= p_0(\beta_0) \times p_1(\beta_1) \\
p_0(\beta_0) &= \text{Normal}(\mu = 5, \sigma^2 = 100) \\
p_1(\beta_1) &= \text{Normal}(\mu = 10, \sigma^2 = 25).
\end{aligned}
$$

Whereas, TAB proposal model was chosen to be the following DP model:

Model A.3. (Prior proposal model for θ-augmented Bayes method in Appendix A.7)

$$
\begin{aligned}
F_{XY} &\sim DP(\phi, G_0') \\
\phi &= 0.5 \\
G_0' &= \text{ Discretized version of the distribution } G_0\text{,} \\
&\quad \text{with mass assigned to points } \{(ih, jh) | i, j \in \mathbb{Z}\} \text{ for fixed bin width } h \text{ s.t.} \\
&\quad G_0'(X = ih, Y = jh) = \int \mathbb{1}\left[x \in \left(ih - \frac{h}{2}, ih + \frac{h}{2}\right)\right] \times \\
&\qquad\qquad \mathbb{1}\left[y \in \left(jh - \frac{h}{2}, jh + \frac{h}{2}\right)\right] \mathrm{d}G_0(x, y) \\
G_0 &= \text{BivariateNormal}\left(\mu = (0,0)^\top, \Sigma = \begin{bmatrix} 40^2 & 0 \\ 0 & 0.5^2 \end{bmatrix}\right) \\
h &= 1 \times 10^{-4}.
\end{aligned}
$$

The Bayesian prior for modelling the observables was:

Model A.4. (Bayesian prior for observables in Section A.7)

$$
\begin{aligned}
(X, Y) &\sim F_{XY} \\
F_{XY} &\sim \text{TA}(m = p_{\beta_0,\beta_1} / q^{\Pi}_{\beta_0,\beta_1}, \mathcal{P}_{\Pi}).
\end{aligned}
$$

We sampled from the TAB target posterior via the MCMC algorithm documented in Section 4.2. A contour plot of the Statistician's prior p_{β_0,β_1} compared to that of an estimate of the distribution $q^{\Pi}_{\beta_0,\beta_1}$ can be found in Figure A.2. The plot shows that $q^{\Pi}_{\beta_0,\beta_1}$ was much more spread out than p_{β_0,β_1}, which suggests that the sampling mechanism described in

Section 4.2 functioned well. Each TAB posterior was obtained via an MCMC chain with $4e5$ runs. The GB posterior and BB posterior were obtained through direct sampling with $1e6$ samples. BEL posteriors were obtained via MCMC with 10^4 runs due to the time-consuming nature of the algorithm. The results are shown in Tables A.1 to A.3.

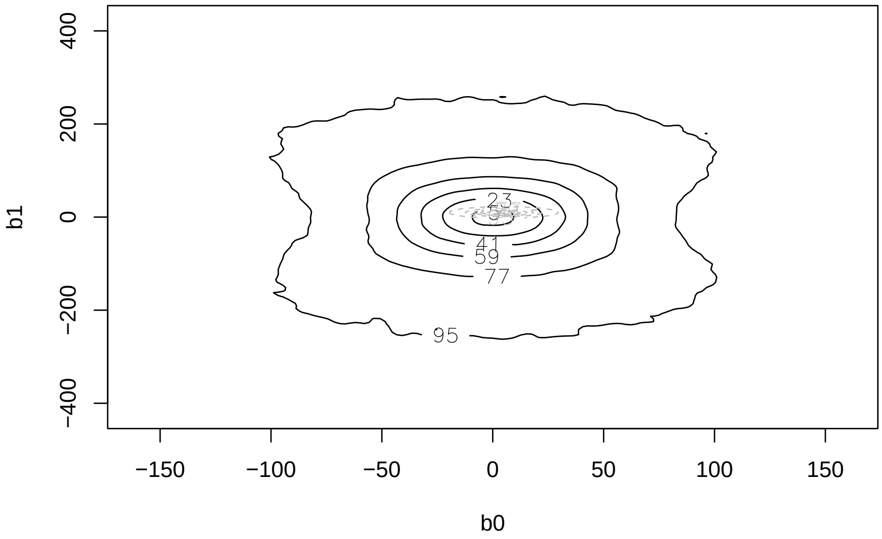


Figure A.2: Comparison study of Section A.7 estimating linear regression parameters: contour plot of the prior over regression coefficients induced by Π (in solid lines) compared that of the required subjective prior distribution (in grey dotted lines).

Examining the results for joint inference, TAB and GB performed very well, with coverage probabilities close to nominal level and small credible regions. Whereas, the coverage probabilities of the BB were much lower than nominal level.

However, the playing field was much more level when it came to marginal inference of the regression coefficients. TAB, GB and BEL methods all showed over-coverage of the credibility intervals. This seems to suggest that the subjective prior that was used was extremely informative relative to the data. The use of BB was adequate for marginal inference in terms of coverage, but the average length of confidence intervals were much longer than the competitors, due to not incorporating the subjective prior. Based on these

Table A.1: Comparison study of Section A.7 estimating linear regression parameters: results for joint inference. Performance metrics were computed based on 300 resampling events at each sample size n.

	Method	Estimated coverage probability of 95% CR	Average size of 95% CR
$n = 20$	TAB	0.957	34.416
	BB	0.833	43.131
	BEL	0.917	28.669
	GB	0.946	31.641
$n = 50$	TAB	0.953	17.651
	BB	0.887	19.949
	BEL	0.945	17.369
	GB	0.940	16.687

Table A.2: Comparison study of Section A.7 estimating linear regression parameters: results for marginal inference of β_0. Performance metrics were computed based on 300 resampling events at each sample size n.

	Method	Estimated coverage probability of 95% CI	Average size of 95% CI	Estimated bias	Average quadratic risk
$n = 20$	TAB	1.000	34.371	-0.200	95.725
	BB	0.910	46.457	-0.860	349.842
	BEL	0.973	28.752	-0.045	85.707
	GB	0.997	30.566	-0.156	82.957
$n = 50$	TAB	1.000	26.424	0.355	61.926
	BB	0.940	30.896	0.381	137.768
	BEL	1.000	25.499	0.0746	63.709
	GB	0.983	24.948	0.256	64.588

Table A.3: Comparison study of Section A.7 estimating linear regression parameters: results for marginal inference of β_1. Performance metrics were computed based on 300 resampling events at each sample size n.

	Method	Estimated coverage probability of 95% CI	Average size of 95% CI	Estimated bias	Average quadratic risk
$n = 20$	TAB	1.000	3.496	0.016	1.077
	BB	0.897	4.710	0.078	3.649
	BEL	0.960	2.991	-0.004	0.979
	GB	0.987	3.075	0.012	0.945
$n = 50$	TAB	1.000	2.696	-0.036	0.685
	BB	0.917	3.181	-0.035	1.477
	BEL	0.99	2.632	0.003	0.687
	GB	0.967	2.478	-0.023	0.698

results it appears that the problem and data-generating mechanism were well-behaved as to allow all semiparametric methods we examined to perform relatively well.