Computational Statistics
https://doi.org/10.1007/s00180-023-01390-0

ORIGINAL PAPER

Multi-pass Bayesian estimation: a robust Bayesian method

Yeming Lei[1,2] · Shijie Zhou[2] · Jerzy Filar[1] · Nan Ye[1]

Received: 30 January 2023 / Accepted: 8 July 2023

Abstract

The prior plays a central role in Bayesian inference but specifying a prior is often difficult and a prior considered appropriate by a modeler may be significantly biased. We propose multi-pass Bayesian estimation (MBE), a robust Bayesian method capable of adjusting the prior's influence on the inference result based on the prior's quality. MBE adjusts the relative importance of the prior and the data by iteratively performing approximate Bayesian updates on the given data, with the number of updates determined using a cross-validation method. The repeated use of the data resembles the data cloning method, but data cloning performs maximum likelihood estimation (MLE), while MBE interpolates between standard Bayesian inference and MLE; there are also algorithmic differences in how MBE and data cloning make repeated use of the data. Alternatively, MBE can be considered a method for constructing a new prior from the given initial prior and the data. We additionally provide a new non-asymptotic bound on the convergence of data cloning, and provide an MBE-like iterative heuristic approach which achieves faster convergence speed by boosting posterior variance. In numerical simulations on several simulated and real-world datasets, MBE provides robust inference results as compared to standard Bayesian inference and MLE.

Keywords Bayesian method · Multi-pass Bayesian estimation · Maximum likelihood estimation · Prior

✉ Nan Ye
nan.ye@uq.edu.au

Yeming Lei
yeming.lei@uq.edu.au

Shijie Zhou
shijie.zhou@csiro.au

Jerzy Filar
j.filar@uq.edu.au

[1] School of Mathematics and Physics, The University of Queensland, St Lucia, QLD 4067, Australia

[2] Oceans and Atmosphere, CSIRO, 306 Carmody Road, St Lucia, QLD 4067, Australia

Published online: 01 August 2023

1 Introduction

Bayesian analysis has diverse applications (Punt and Hilborn 1997; Corander et al 2008; Monnahan et al 2017; Brown and Hund 2018), partly due to its ability to incorporate useful information through the prior. However, specifying a prior is often a difficult decision process involving many subtleties. In general, the prior is often constructed by collating relevant pieces of information, assessing their veracity, and translating them into a mathematical form. There is a considerable ongoing debate on how this should be done, with no consensus on what information should be incorporated, and how the information should be incorporated. For example, when little information is available, we may use a flat or diffuse prior, but there is no universally accepted good flat prior. The uniform distribution is a natural choice but is not invariant to reparametrization. Objective priors (Yang and Berger 1996; Kass and Wasserman 1996; Ghosh 2011), including Jeffrey's priors (Jeffreys 1946) and reference priors (Bernardo 1979; Berger and Bernardo 1992), are often used as alternatives to uniform priors. These priors are considered to make the analysis less subjective by the advocates, but this is a point of dispute among Bayesians (Lindley 1983). As another example, many Bayesians advocate eliciting subjective knowledge of the domain experts and converting the elicited knowledge into an informative prior, but there is no universally accepted method on how this should be done (Mikkola et al 2021).

Various forms of bias may be present even in a prior obtained from a very careful analysis, leading to undesirable decisions. First, the bias may stem from mathematical subtleties. For example, while various diffuse priors are commonly used when there is little external information, they are not necessarily non-informative (Lemoine 2019). Second, a researcher may have a strong prior belief that could be incorrect, which can produce an unreasonable posterior (Bolstad and Curran 2016, Section 16.1) and (Yang and Berger 1996). For example, when developing priors for data-poor fish stocks using knowledge of data-rich fish stocks, biased informative priors may be used due to systematic differences in stock status and characteristics. This is because well studied stocks are typically large, high-yielding and well-managed stocks, while data-poor stocks usually have little management (Punt and Hilborn 1997).

In this paper, we introduce a Bayesian technique that is more robust with respect to the bias in the chosen prior. Specifically, if the prior is well-specified, we want to exploit the prior. On the other hand, if the prior is misspecified or biased, we want to reduce the prior's influence. Our proposed method, referred to as multi-pass Bayesian estimation (MBE), adjusts the prior's influence by iteratively performing approximate Bayesian updates on the given data. This gradually reduces the importance of the prior. A cross-validation method is used to determine the number of Bayesian updates so that the relative importance of the prior is commensurate with its quality.

Our work is closely related to the data cloning method and the empirical Bayes method. Data cloning creates multiple copies of the dataset and uses Markov chain Monte Carlo (MCMC) methods to estimate the model parameters. When the number of copies is large, the mean of the posterior distribution converges to the maximum likelihood estimate. This provides an elegant method for computing the maximum likelihood estimates (MLEs) for complex models through Bayesian framework (Lele et al 2007, 2010). Similarly to data cloning, MBE makes repeated use of the data. However, data cloning computes the MLE, while MBE interpolates between standard Bayesian inference and MLE. In addition, data cloning performs Bayesian inference on a single dataset consisting of multiple clones of the original dataset using MCMC, and the time complexity grows quickly in the number of clones. However, MBE performs approximate incremental Bayesian updates and the time complexity grows linearly in the number of repetitions. Empirical Bayes optimizes the hyperparameters of the prior by maximizing the likelihood (Carlin and Louis 2000) which can be viewed as learning a data-dependent prior. Similarly, MBE can also be viewed as an alternative method for learning a data-dependent prior. However, MBE avoids the difficult problem of computing the maximum likelihood estimates.

We previously performed a preliminary investigation on the benefits of adjusting the relative importance of the prior and the data for estimating fishery models, by performing a fixed number of Bayesian updates (Lei et al 2021). This paper provides a general framework with a cross-validation procedure for automatically adjusting the relative importance of the prior and the data, and a MBE-like heuristic approach for computing MLEs. We further provided some theoretical analysis and performed extensive experiments to demonstrate the effectiveness of both procedures across several inference problems.

The remainder of this paper is organized as follows. Section 2 describes MBE, and introduces an MBE-like heuristic iterative approach that has a faster convergence rate than data cloning for computing MLE by boosting posterior variance. Section 3 provides a new non-asymptotic theoretical analysis of the convergence of data cloning. Section 4 demonstrates that MBE provides robust inference results as compared to standard Bayesian inference and maximum likelihood estimation on several simulated and real-world datasets. In addition, our MBE-like approach with variance boosting achieves a faster convergence speed for computing the MLE compared to data cloning and similar MBE-like approaches. Section 5 concludes the paper.

2 Multi-pass Bayesian estimation

MBE is a robust Bayesian method that alleviates the bias in the prior: it aims to maximize the use of both the information from the prior and the information from the data, while minimizing the influence of potential bias in the prior. To this end, it combines the prior and the data in a way that automatically adjusts the relative importance of the prior and the data to produce a distribution that interpolates between standard Bayesian inference and MLE. MBE produces the standard Bayesian posterior on one extreme, and the MLE on the other extreme.

It is instructive to first briefly review how the prior and the data are used in standard Bayesian inference, MLE, and data cloning. Formally, we are given a prior $p_0(\theta)$ and a likelihood model $p(D \mid \theta)$, where θ denotes an element from the parameter space Θ, and D denotes a dataset. Standard Bayesian inference combines the prior and the data using the Bayes' rule by computing a posterior that is proportional to the product of the prior $p_0(\theta)$ and the likelihood $\ell(\theta) = p(D \mid \theta)$:

$$p_1(\theta) = P(\theta \mid D) \propto p_0(\theta)\ell(\theta). \tag{1}$$

MLE seeks to maximize the likelihood $\ell(\theta)$ only and is thus not using the prior at all. Data cloning computes the MLE using MCMC, which provides a nice link between standard Bayesian inference and MLE. Specifically, given a large number of j clones of D, say $D_1 = D_2 = \ldots = D_j = D$, data cloning first computes the posterior using the dataset consisting of these clones:

$$p_j(\theta) = P(\theta \mid D_1, \ldots, D_j) \propto p_0(\theta)\ell(\theta)^j, \tag{2}$$

then it uses the posterior mean as an MLE. Intuitively, this works because, as j becomes larger, the maximum likelihood estimator $\theta_{ML} \in \arg\max_\theta \ell(\theta)$ acquires larger and larger density in the posterior.

The distributions $p_0, p_1, p_2, \ldots$ form a spectrum of distributions as j increases, the prior becomes less important while the data becomes more important in p_j. In particular, p_0 is the initial prior, p_1 is the standard Bayesian posterior, and p_∞ is the MLE. Figure 1 provides a schematic illustration of this spectrum. Naively, we can try to select the best distribution in the spectrum for performing inference, but using MCMC to compute p_j exactly becomes too expensive as j increases. Our MBE algorithm provides a general framework for efficiently generating such a spectrum of distributions, and choosing the best one among them for inference.

Specifically, MBE uses the data and the prior to incrementally construct a sequence of distributions where the relative importance of the prior and the data varies and chooses the one that is most suitable for inference. This is done based on two key components: an update operator $\mathcal{U}(p, D)$ and a validation score $\mathcal{S}(p, D)$. The update operator $\mathcal{U}(p, D)$ takes in a distribution p on Θ and a dataset D, and then outputs a distribution on Θ where the data is relatively more important. This allows incremental construction of a spectrum of distributions on Θ where the relative importance of the prior and the data varies. The validation score $\mathcal{S}(p, D)$ evaluates whether a posterior distribution p has good predictive power on D. We provide details on the update operator $\mathcal{U}$ and the validation score $\mathcal{S}$ in Sects. 2.1 and 2.2. Examples of how the update operators and the validation score are implemented on some problems are provided in Sect. 4.5.

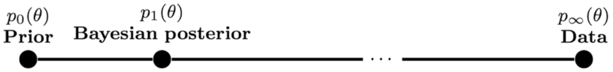


Fig. 1 A spectrum of distributions where the relative importance of the prior and the data varies

Algorithm 1 Multi-pass Bayesian Estimation (MBE)

```
Require: prior $p_0(\theta)$, data $D$
Ensure: a distribution on $\Theta$ for inference
 1: $j^* \leftarrow$ SelectImportance($p_0, D$).
 2: for j=1 to $j^*$ do
 3:     $p_j \leftarrow \mathcal{U}(p_{j-1}, D)$
 4: end for
 5: return $p_{j^*}$.

 6: procedure SelectImportance($p_0, D$)
 7:     Partition $D$ into two subsets $D_1$ and $D_2$.
 8:     $j \leftarrow 0$.
 9:     $p_{0,1} = p_{0,2} = p_0$.
10:     while termination criterion is not met do
11:         $j \leftarrow j + 1$.
12:         $p_{j,1} \leftarrow \mathcal{U}(p_{j-1,1}, D_1)$.
13:         $p_{j,2} \leftarrow \mathcal{U}(p_{j-1,2}, D_2)$.
14:         $s_j = \log(\mathcal{S}(p_{j,1}, D_2)) + \log(\mathcal{S}(p_{j,2}, D_1))$.
15:     end while
16:     $j^* \leftarrow \arg\max_j s_j$, where max is over all computed scores.
17: return $j^*$
18: end procedure
```

Algorithm 1 displays the pseudocode for 2-fold cross-validation MBE, which consists of two phases. The first phase (line 1) determines the optimal relative importance using cross-validation by the SelectImportance function (line 5 to line 15). To determine the optimal relative importance, we use a 2-fold cross-validation method. We first partition the dataset D into two subsets D_1 and D_2 (line 6). MBE then iteratively constructs a sequence of distributions $p_{1,1}, p_{2,1}, \ldots$ using the update operator $\mathcal{U}$, a prior $p_{0,1} = p_0$ and the dataset D_1 (lines 9 to line 11). For each $p_{j,1}$, we compute its validation score $\mathcal{S}(p_{j,1}, D_2)$ on D_2 (first part of line 13). Similarly, we reverse the role of D_1 and D_2, and compute a sequence of distributions $p_{1,2}, p_{2,2}, \ldots$. The validation score for j is then computed as $\log(\mathcal{S}(p_{j,1}, D_2)) + \log(\mathcal{S}(p_{j,2}, D_1))$. The construction stops when a certain termination criterion is met. Finally, we choose the j value leading to the largest validation score as the optimal relative importance.

The second phase (lines 2–3) computes the posterior for inference using the entire dataset D and the optimal relative importance is determined in the first phase.

We make a few comments on the cross-validation procedure used in the first phase. First, we used 2 fold cross-validation for computational efficiency in the first phase, but more than 2 folds could be used. We provide the pseudocode for N-fold cross-validation MBE in Appendix A and show that 2-fold cross-validation perform similarly as 5-fold cross-validation on our benchmark problems. Second, we followed the practice of using equal-sized subsets as in the standard cross-validation procedure. In fact, we tried the alternative splits of 30/70 and 40/60 on the simple Gaussian estimation problem described in Sect. 4.1.1, and did not observe significant difference in MBE's performance, thus we only consider equal-sized splits in this paper. Third, for time series data, the importance score is calculated in a slightly different way, as discussed in Sect. 2.2.

Our MBE-like approach for computing the MLE is the same as MBE, but replaces the first phase with setting $j^* \to \infty$. We provide a theoretical discussion on this approach in Sect. 3.

Details of the update operator, the validation score, and the termination criterion are described below.

2.1 The update operator

There are various choices for the update operator. We describe three update operators below.

The first one is the *exact Bayesian update* operator $\mathcal{B}$ defined by

$$\mathcal{B}(p, D)(\theta) = p(\theta)p(D \mid \theta) \Big/ \int p(\theta')p(D \mid \theta')d\theta'.$$

The exact Bayesian update operator is conceptually simple and, in some cases, a closed-form formula can be given for it. When using the exact Bayesian update operator, the distributions $p_1, p_2, \ldots$ are the same as those defined by Eq. 2, and are equivalent to the posteriors obtained by running the Bayesian filter with identical observations at different time steps. However, in general, there is no closed-form formula for the exact Bayesian update operator and it is not possible to compute the exact Bayesian update.

The second update operator is what we call the *moment matching Bayesian update* operator $\mathcal{M}(p, D)$, which is the default update operator that we use in this paper. The operator first runs an MCMC algorithm to calculate a set of particles representing the posterior $q(\theta) \propto p(\theta)p(D \mid \theta)$ and then fits a distribution $\tilde{p}$ on the set of particles using moment matching. The particular form of $\tilde{p}$ can be user-specified and must be a distribution that can be used as input to the chosen MCMC algorithm. When there are multiple parameters to be estimated, we choose $\tilde{p}$ as a product of univariate distributions and learn the parameters of the univariate distributions using moment matching. In general, we do not have to use moment matching for estimating the parameters of $\tilde{p}$, and we can use alternative parameter estimation methods such as maximum likelihood estimation. We choose moment matching because this often leads to simple estimates that can be efficiently computed. For the MCMC algorithm, we used Gibbs sampling implemented in R2jags (Su et al 2015), an R package to the JAGS (Plummer et al 2003) library. R2jags supports the use of the Gelman-Rubin convergence diagnostic (Gelman and Rubin 1992; Brooks and Gelman 1998) to detect convergence, which can allow terminating the simulation of Markov chains once convergence has occurred, instead of simulating the Markov chains for the maximum number of allowed steps.

The third update operator is what we call the *variance-boosted Bayesian update* operator $\mathcal{V}$. The operator first runs MCMC as in the moment-matching Bayesian update, and then performs moment matching to fit a distribution $\tilde{p}$ with the variance of each univariate distribution equal to $\max\{\sigma^2, v\}$, where σ^2 is the variance of the considered univariate variable in the particles and v is a user-specified constant. The variance-boosted Bayesian update is motivated by our observation that the moment-matching Bayesian update operator $\mathcal{M}$ sometimes leads to only slightly different updated distributions in a single step. Thus we "boost" the variance of the updated distribution whenever it falls below a threshold, so that the posterior can move faster towards the MLE. In this paper, we set the v value for each parameter separately so that the coefficient of variation of each parameter is at most 0.1, i.e., $v = 0.1^2\mu^2$, where μ is the mean value of the parameter. The threshold value 0.1 was chosen on the basis that it is neither too small nor too large. This offered some speed-up for computing MLE in our experiments (see Sect. 4.4). 1 includes a sensitivity study on the threshold value. The results indicate that a larger threshold may lead to faster convergence to the MLE in terms of the number of iterations for the variance-boosted update. However, for the Bayesian update at each iteration, MCMC sometimes converges very slowly for a large threshold, and even fail to converge for a very large threshold.

2.2 The validation score

To determine the posterior p_j that is optimal for inference, we develop a validation score to measure the quality of the relative importance index j.

When the dataset D consists of i.i.d. examples, we randomly split the dataset into two disjoint sets D_1 and D_2, and then compute the posterior $p_{j,1} = \mathcal{U}(p_{j-1,1}, D_1)$ and $p_{j,2} = \mathcal{U}(p_{j-1,2}, D_2)$ on D_1 and D_2 respectively using the same initial prior. The cross-validation score of the relative importance index j is calculated as

$$s_j = \log(\mathcal{S}(p_{j,1}, D_2)) + \log(\mathcal{S}(p_{j,2}, D_1)), \tag{3}$$

where $\mathcal{S}(p, D) = \int p(D \mid \theta)p(\theta)d\theta$ measures on average how well a random model θ distributed according to $p(\theta)$ can predict D. The exact computation of the validation score is hard. A natural idea is to use the Monte Carlo approximation to compute the validation scores. For example,

$$\mathcal{S}(p_{j,1}, D_2) \approx \frac{1}{N}\sum_{i=1}^{N} p(D_2 \mid \theta_i^{(1)}), \tag{4}$$

where the $\theta_i^{(1)}$'s are independently sampled from posterior $p_{j,1}$. When using MCMC to compute $p_{j,1}$, these $\theta_i^{(1)}$'s values can be taken as a random subset of the particles representing $p_{j,1}$. However, the Monte Carlo approximation is too expensive on some datasets. A more efficient approximation is

$$\mathcal{S}(p_{j,1}, D_2) \approx p(D_2 \mid \theta_{j,\mathrm{med}}^{(1)}), \tag{5}$$

where the $\theta^{(1)}_{j,\mathrm{med}}$ is the parameter vector where each component takes its median value according to its marginal distribution given by $p_{j,1}$. We used the median approximation to calculate $\mathcal{S}(p_{j,1}, D_2)$ and $\mathcal{S}(p_{j,2}, D_1)$ in this paper, even though the Monte Carlo approximation is unbiased but the median approximation is generally not. This is because the Monte Carlo approximation yields highly variable estimates — for example, for the simple problem of estimating a Gaussian in Sect. 4, even using $N = 10,000$ models from the posterior gives estimates of the validation score that are too variable to be useful for selecting a good relative importance index. In fact, our empirical results suggest that the median approximation is effective (see Sect. 4). This is likely because the median approximation ranks the relative importance indices similarly as the exact validation score.

It is possible to use an alternative validation score, such as $\log(\mathcal{S}(p_{j,1}, D_2) + \mathcal{S}(p_{j,2}, D_1))$. In 1, we compared this alternative validation score to the one used in this paper and found that they produced similar results. We used the sum of log-likelihood values as our validation score in this paper for the simplicity of its implementation. This is because in $\mathcal{S}(p_{j,1}, D_2) + \mathcal{S}(p_{j,2}, D_1)$, the two scores can be very small and a straightforward implementation may often yield 0 as the output due to numerical underflow. On the other hand, the log-likelihood values do not suffer from numerical underflow.

When the dataset is a time series, we use a slightly different procedure to calculate the validation score of each importance index j. We first take D_1 as the first half of the time series, and D_2 as the second half. The validation score of j is then computed as

$$s_j = \mathcal{S}(p_{j,1}, D_2) \tag{6}$$

However, there is a subtlety in this case: evaluating $p(D_2 \mid \theta)$ may be computationally quite expensive due to the presence of latent variables. For some models, it helps to consider the posterior distribution $p(\theta, \theta_{\mathrm{aux}})$ of θ and some of the latent variables θ_{aux} instead of the posterior distribution $p(\theta)$. In MCMC, such distribution is already computed, and thus there is no additional cost of obtaining it. However, it can be more efficient to compute $p(D_2 \mid \theta, \theta_{\mathrm{aux}})$ as compared to computing $p(D_2 \mid \theta)$, and thus we can use the following approximation:

$$\mathcal{S}(p_{j,1}, D_2) \approx p(D_2 \mid \theta^{(1)}_{j,\mathrm{med}}, \theta^{(1)}_{j,\mathrm{aux},\mathrm{med}}), \tag{7}$$

where $\theta^{(1)}_{j,\mathrm{aux},\mathrm{med}}$ is the parameter vector where each component takes its median value according to $p_{j,1}(\theta_{\mathrm{aux}})$. We illustrate the above idea using a fishery dynamics model in Sect. 4.1.2.

2.3 Termination criterion

The simplest termination criterion is to run the updates for a fixed number of J iterations. In our experiments, we used $J = 100$ updates, but we noticed that the convergence of the posterior median parameter values often occurs before that. Thus the algorithm may be more efficient if it stops further updates once such convergence happens. An empirical investigation of the effectiveness of early stopping will be an interesting question for further study.

3 Theoretical analysis

We provide some theoretical analysis on the posterior spectrum $p_0, p_1, p_2, \ldots$ defined by $p_j = \mathcal{U}(p_{j-1}, D)$ for $j \geq 1$.

We first consider the case when the update operator is the exact Bayesian update operator $\mathcal{B}$. The data cloning paper (Lele et al 2007) showed that the mean value of the data-cloned posterior converges to the maximum likelihood estimates as the number of data clones increases. We provide a more refined convergence result: under mild conditions, the posterior distribution in Eq. 2 essentially converges to a distribution over the maximum likelihood estimates. Note that in the data cloning paper, the authors (Lele et al 2007) assumed that there was a unique MLE, while we make no such assumption in our proof.

Theorem 1 *Let* $\Theta_{ML} = \arg\max_\theta \ell(\theta)$ *be the set of maximum likelihood estimates, and* ℓ^* *be the maximum likelihood value. Let* $p_j = \mathcal{B}(p_{j-1}, D)$. *Assume that there exists* $\epsilon > 0, \delta > 0, \nu > \zeta > 0$ *such that*

(a) *for any* θ *at a distance of at least* ϵ *away from* Θ_{ML}, $\ell(\theta) \leq \ell^* - \delta$, *and*
(b) *for any* θ *within a distance of* ϵ *away from* Θ_{ML}, *the largest eigenvalue of the Hessian* $\nabla^2 \ell(\theta)$ *is at most* $-\zeta$ *and at least* $-\nu$.

Then for any $r > 0$, $p_j(d(\theta, \Theta_{ML}) \leq r) \to 1$ as $j \to \infty$, where $d(\theta, \Theta_{ML}) = \inf_{\theta' \in \Theta_{ML}} \|\theta - \theta'\|$. That is, as j becomes larger, the probability of θ lying within an arbitrarily small distance from maximum likelihood estimates tends to 1.

Proof Note that if for some r, $p_j(d(\theta, \Theta_{ML}) \leq r) \to 1$ as $j \to \infty$, then the result holds for any larger r too. Thus it suffices to show that the result holds for r values smaller than some fixed value. In particular, we consider $r \in (0, r_0]$, where r_0 is the largest number such that $r_0 \leq \epsilon$, and $\zeta r_0^2 \leq 2\delta$.

We first bound the likelihood of parameters that are at a distance of at most ϵ away from Θ_{ML}. Consider any θ such that $d(\theta, \Theta_{ML}) < \epsilon$, then there exists $\theta_0 \in \Theta_{ML}$ such that $d = \theta - \theta_0$ has a norm less than ϵ. From Taylor's theorem, we have

$$\ell(\theta) = \ell(\theta_0 + d) = \ell(\theta_0) + d^\top \nabla \ell(\theta_0) + \frac{1}{2} d^\top \nabla^2 \ell(\theta_0 + cd) d, \tag{8}$$

for some $c \in [0, 1]$. Since θ_0 maximizes the likelihood, we have $\nabla \ell(\theta_0) = 0$. In addition, by assumption (b), we have $-\nu \|d\|^2 \leq d^\top \nabla^2 \ell(\theta_0 + cd) d \leq -\zeta \|d\|^2$. Thus we have

$$\ell^* - \frac{\nu}{2} \|d\|^2 \leq \ell(\theta_0 + d) \leq \ell^* - \frac{\zeta}{2} \|d\|^2. \tag{9}$$

Now choose any $r' < r$ such that $vr'^2 \leq \zeta r^2/2$. We partition the parameter space into the following three sets:

$$\begin{aligned}\Theta_1 &= \{\theta : d(\theta, \Theta_{ML}) < r'\},\\ \Theta_2 &= \{\theta : d(\theta, \Theta_{ML}) \in [r', r]\},\\ \Theta_3 &= \{\theta : d(\theta, \Theta_{ML}) > r\},\end{aligned}$$

and we use $p_j(\Theta_i)$ to denote the probability that $\theta \in \Theta_i$ according to p_j.

Using the likelihood bound above, we have

$$\begin{aligned}\theta \in \Theta_1 &\Rightarrow \ell(\theta) > \ell^* - \frac{\zeta r^2}{4},\\ \theta \in \Theta_3 &\Rightarrow \ell(\theta) < \ell^* - \frac{\zeta r^2}{2}.\end{aligned}$$

It follows that $\frac{p_j(\Theta_1)}{p_j(\Theta_3)} > \frac{\ell^* - \zeta r^2/4}{\ell^* - \zeta r^2/2} \frac{p_{j-1}(\Theta_1)}{p_{j-1}(\Theta_3)} > \left(\frac{\ell^* - \zeta r^2/4}{\ell^* - \zeta r^2/2}\right)^j \frac{p_0(\Theta_1)}{p_0(\Theta_3)}$ which tends to ∞ as $j \to \infty$, as the ratio $\frac{\ell^* - \zeta r^2/4}{\ell^* - \zeta r^2/2}$ is larger than 1.

This implies that $\frac{p_j(\Theta_1) + p_j(\Theta_2)}{p_j(\Theta_3)} \to \infty$, thus the probability $p_j(d(\theta, \Theta_{ML}) \leq r) = p_j(\Theta_1 \cup \Theta_2) \to 1$. □

As a corollary of the above proof, we can show that the convergence occurs at an exponential rate.

Corollary 1 *Under the assumptions of Theorem* 1, *there exists constants* $a > 0$, *such that for any* $r > 0$, *there exists* $s \in (0, 1)$ *satisfying*

$$p_j(d(\theta, \Theta_{ML}) \leq r) \geq 1 - as^j.$$

Proof From the proof of Theorem 1, we have $\frac{p_j(\Theta_1)}{p_j(\Theta_3)} > \left(\frac{\ell^* - \zeta r^2/4}{\ell^* - \zeta r^2/2}\right)^j \frac{p_0(\Theta_1)}{p_0(\Theta_3)}$. Thus we have

$$p_j(\Theta_3) \leq \frac{p_j(\Theta_3)}{p_j(\Theta_1)} < \left(\frac{\ell^* - \zeta r^2/2}{\ell^* - \zeta r^2/4}\right)^j \frac{p_0(\Theta_3)}{p_0(\Theta_1)} = as^j,$$

where $a = \frac{p_0(\Theta_3)}{p_0(\Theta_1)}$, and $s = \frac{\ell^* - \zeta r^2/2}{\ell^* - \zeta r^2/4} < 1$. Therefore, $p_j(d(\theta, \Theta_{ML}) \leq r) = p_j(\Theta_1 \cup \Theta_2)$ $= 1 - p_j(\Theta_3) \geq 1 - as^j$. □

Note that any distribution p^* on the MLEs is a fixed point of $\mathcal{B}$ in the sense of

$$p^* = \mathcal{B}(p^*, D). \tag{10}$$

Thus from Theorem 1, we can view the repeated application of the exact Bayesian update operator as a fixed-point iteration for computing a fixed point of the operator, that is the sequence of distributions computed converges to a fixed-point p^* of $\mathcal{B}$.

For the moment-matching Bayesian update operator $\mathcal{M}$ and the variance-boosted Bayesian update operator $\mathcal{V}$, we conjecture that the limiting distributions computed by repeated applications of these operators also converge to fixed points of these operators in general, and such fixed points depend mostly on data. While we do not have a formal result, we believe that further investigation of this conjecture is important, considering how the variance-boosted Bayesian update operator can be used to efficiently compute MLEs, as demonstrated in Sect. 4.4.

4 Numerical experiments

We perform three sets of experiments in this section. First, we compare MBE with the standard Bayesian method and MLE on how they perform under different priors on several inference problems, so as to investigate their robustness against biased priors and their ability to exploit information in the prior. Second, we investigate the performance of the different update operators used in MBE, so as to identify the best variant of MBE. Third, we demonstrate the use of variance boosting for efficiently computing MLEs.

We describe the details of the inference problems and the experimental settings used in Sect. 4.1, then we present results for the three sets of experiments in the following subsections.

4.1 Inference problems and experimental settings

We consider estimating the parameters of three different models: the Gaussian distribution, a fishery stock assessment model known as the Schaefer model, and logistic regression. For standard Bayesian and MBE, we use the posterior median as the estimated parameter. We generate synthetic data for estimating Gaussians and the Schaefer model, thus we know the ground truth model for these two cases, and we can evaluate the accuracy of the estimated parameters. For logistic regression, we use several real-world binary classification datasets. We do not know the ground truth model in this case, thus we split a dataset into a training set and a test set, and we are interested in the test accuracy of the learned logistic regression model.

For each of the three models, we consider three different prior settings: misspecified prior, non-informative or weak prior, and well-specified prior. A weak prior is chosen to be flat or uniform so that it does not favour any particular values. For the Gaussian and the Schaefer model, we know the true parameters generating the synthetic data, thus a well-specified prior is chosen to be a probability distribution centered around the true value, and a misspecified prior's probability distribution is chosen to center around a value sufficiently different from the ground truth. For logistic regression, we take the MLE computed on the test set as the ground truth when specifying the well-specified prior and the misspecified prior.

4.1.1 Gaussian model

We consider inferring the mean μ and standard deviation σ of a Gaussian distribution $N(\mu, \sigma^2)$ using an i.i.d. sample $x_1, \dots, x_n$ drawn from the distribution. We are interested in two different settings: only μ is unknown, and both μ and σ are unknown. The prior of the mean μ is chosen to be a Gaussian $N(\mu_0, \sigma_0^2)$. In the case that σ is unknown, its prior is chosen to be a lognormal distribution $LN({\mu'}_0, {\sigma'}_0^2)$. The ground truth values of μ and σ and their priors used in our experiments are given in Table 1.

We compared the inference methods on 100 datasets, each consisting of $n = 50$ observations. The quality of inference is measured in terms of the absolute errors of the estimates, that is the absolute values of the differences between the estimates and the true values.

4.1.2 Schaefer model

The Schaefer model is a popular fishery stock assessment model which can produce estimates of maximum sustainable yield and other associated fisheries reference points (e.g., see (Winker et al 2018)). It belongs to the class of Surplus Production Models, which describe how the fishery biomass evolves according to the generic difference equation

$$B_{t+1} = B_t + SP_t - C_t, \tag{11}$$

where B_t is the biomass in year t, C_t is the catch in year t, and the surplus production term SP_t aggregates the effects of recruitment, growth, and natural mortality.

We consider the following stochastic version of the Schaefer model

$$B_{t+1} = \left[B_t + rB_t\left(1 - \frac{B_t}{K}\right) - C_t\right]e^{\varepsilon_B}. \tag{12}$$

where r is the intrinsic growth rate of population, and K is the carrying capacity which can be interpreted as the maximum biomass that can be sustained by the environment, and $\varepsilon_B \sim N(0, \sigma_B^2)$ is a random process error. In addition, we assume that the catch C_t is related to the fishing effort E_t and the biomass B_t via

$$C_t = qE_tB_te^{\varepsilon_C}, \tag{13}$$

Table 1 Ground truth values and priors for the parameters of the Gaussian distribution

Parameter	Ground truth	Misspecified prior	Weak prior	Well-specified prior
μ	0	$N(-1, 0.1^2)$	$N(-1, 2^2)$	$N(0, 0.1^2)$
σ	1	$LN(0.1, 0.1^2)$	$LN(0.1, 2^2)$	$LN(1, 0.1^2)$

where q is known as the catchability constant, and $\varepsilon_C \sim N(0, \sigma_C^2)$ models the random variation in the catch process. Finally, we assume that the observed catch per unit effort (CPUE) I_t at time t is

$$I_t = \frac{C_t}{E_t} e^{\varepsilon_I} = qB_t e^{\varepsilon_C + \varepsilon_I} = qB_t e^{\varepsilon_{IC}}, \tag{14}$$

where E_t is the effort value, $\varepsilon_I \sim N(0, \sigma_I^2)$ is used to model measurement errors, and $\varepsilon_{IC} \sim N(0, \sigma_{IC}^2)$ with $\sigma_{IC}^2 = \sigma_I^2 + \sigma_C^2$.

Our inference problem is to estimate the parameters $(K, q, r, \psi, \sigma_B^2, \sigma_{IC}^2)$ from the catch time series $\{C_t\}$ and the CPUE time series $\{I_t\}$ generated according to Eqs. 12,13 and 14, where $\psi = B_0/K$ is known as the initial biomass depletion. In our experiments, we used $K = 2000$, $r = 0.4$, $\psi = 0.8$, $q = 0.004$, $\sigma_B = 0.1$, $\sigma_C = 0.1$, $\sigma_I = 0.1$, and a dataset includes 30 years of catch and CPUE values.

Log-normal (LN) priors are used for K, r, ψ and q, because these parameters should be non-negative. We used the same coefficient of variation (CV) for all the parameters in each prior setting. The priors used are listed in Table 2. Note that we used weak priors for those Gaussian noises for all prior settings, by assuming that the variances σ_B^2 and σ_{IC}^2 follow an inverse gamma distribution, invgamma($\alpha = 4, \beta = 0.01$), which is commonly used in fisheries research (Millar and Meyer 2000; Winker et al 2018).

As in the case of the Gaussian distribution, we also compared the inference methods on 100 randomly generated datasets. The quality of the inference is measured using the absolute relative error $|\theta_{med} - \theta_{true}|/\theta_{true}$, where θ_{med} is the posterior median and θ_{true} is ground truth. Relative errors are used because the parameters are of very different scales.

We make three remarks on the implementation of MBE on the Schaefer model. First, since we are dealing with time series data, we compute the validation score as in Eq. 6. That is, we only consider how well the posterior $p_{j,1}(\theta)$ computed using the first part of data can predict the second part of data D_2. Second, computing the likelihood for the Schaefer model is hard due to the presence of latent variables, thus approximation is needed. Given a dataset D consisting of the observed catches $C_{1:n}$ and abundance indices $I_{1:n}$, the likelihood of the parameters $\theta = (K, B_1, r, q, m, \sigma_B^2, \sigma_C^2, \sigma_I^2)$ is given by

Table 2 Ground truth values and priors for the parameters of the Schaefer model

	True value	Misspecified prior	Weak prior	Well-specified prior
K	2000	$LN(8000, CV = 0.1)$	$LN(8000, CV = 2)$	$LN(2000, CV = 0.1)$
r	0.4	$LN1, CV = 0.1)$	$LN(1, CV = 2)$	$LN(0.4, CV = 0.1)$
ψ	0.8	$LN(1, CV = 0.1)$	$LN(1, CV = 2)$	$LN(0.8, CV = 0.1)$
q	0.004	$LN(0.01, CV = 0.1)$	$LN(0.01, CV = 2)$	$LN(0.004, CV = 0.1)$

$$\begin{aligned}
L(\theta \mid D) =& p\,(I_{1:n} \mid \theta, C_{1:n}) \\
=& \int p(I_{1:n}, B_{2:n} \mid \theta, C_{1:n}) dB_{2:n} \\
=& \int p(B_{2:n} \mid \theta, C_{1:n}) p(I_{1:n} \mid \theta, C_{1:n}, B_{2:n}) dB_{2:n} \\
=& \int \left[\prod_{t=1}^{n-1} p(B_{t+1} \mid B_t, C_t, K, r, m) \prod_{t=1}^{n} p(I_t \mid B_t, q) \right] dB_{2:n} \\
=& \int \left[\prod_{t=1}^{n-1} f_{LN}(B_{t+1} \mid \mu = \ln[B_t + SP(B_t, K, r, m) - C_t], \sigma^2 = \sigma_B^2) \times \right. \\
& \left. \prod_{t=1}^{n} f_{LN}(I_t \mid \mu = \ln(qB_t), \sigma^2 = \sigma_C^2 + \sigma_I^2) \right] dB_{2:n},
\end{aligned} \tag{15}$$

where $f_{LN}(x \mid \mu, \sigma^2)$ is the probability density function of the log-normal distribution $LN(\mu, \sigma^2)$. We approximate $L(\theta \mid D)$ using a simple Monte Carlo estimate:

$$L(\theta \mid D) \approx \frac{1}{M} \sum_{i=1}^{M} p(I_{1:n} \mid \theta, C_{1:n}, B_{2:n}^{(i)}), \tag{16}$$

where each $B_{2:n}^{(i)}$ is independently sampled using the parameters θ according to Eqs (11) and (13). Third, as mentioned in Sect. 2.2, the validation score in Eq. 6 is computed via Eq. 7. The auxiliary variable $\theta_{\text{aux,med}}$ is simply chosen to be the median of the last biomass in the first part of the time series. This then allows us to calculate each summand in Eq. 7 using the Monte Carlo estimate in Eq. 16. In this paper, we used $M = 100000$ to approximate the $L(\theta \mid D)$ in Eq. 16.

4.1.3 Logistic regression

The logistic regression model is a commonly used classification model. We consider estimating the parameters of a binary logistic regression model

$$p(Y \mid \mathbf{x}, \boldsymbol{\beta}) = \frac{1}{1 + e^{-\mathbf{x}^\top \boldsymbol{\beta}}}, \tag{17}$$

where $\mathbf{x} \in \mathbb{R}^{d+1}$ is the input vector (including a dummy variable with value 1 for modeling the bias), Y is the class label which takes either value 0 or 1, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)$ consists of the model parameters.

We used six real binary classification datasets from the LIBSVM datasets (Chang and Lin 2011), which cover diverse domains: breast-cancer, diabetes, german.numer, heart, svmguide1, and svmguide3. Note that we only used the training sets for svmguide1 and svmguide3. Tabel 3 shows the datasets and their sizes.

Table 3 The binary classification datasets and their sizes

Dataset	Numbers of examples	Numbers of features
Breast-cancer	683	10
Diabetes	768	8
German.numer	1000	24
Heart	270	13
svmguide1	3089	4
svmguide3	1243	21

For each dataset, we consider inferring the model parameters using a random 50% of the full dataset as the training set, and evaluating the inferred model's predictive accuracy using the other 50% of the data as the test set.

We used a Gaussian prior for each β_i with the same coefficient of variation (CV) applied to account for the varying scales of the parameters since these parameters are on very different scales, the same σ^2 in $N(\mu, \sigma^2)$ would represent very different degrees of uncertainties for the parameters. The misspecified prior and well-specified prior are chosen to be $N(-\beta_i^*, CV = 0.1)$ and $N(\beta_i^*, CV = 0.1)$ respectively, where $(\beta_0^*, \beta_1^*, \dots, \beta_d^*)$ is the MLE on the test set. For the weak prior, we used the Gaussian distribution with mean 0 and large variance of $\sigma^2 = 100$. The priors are given in Table 4.

We compared the inference methods on 100 random train-test splits on each dataset. The quality of inference is measured using the test set accuracy of the estimated method.

4.2 Robustness of MBE

We compared MBE using moment matching Bayesian update operator against standard Bayesian and MLE in this section. We consider this operator because as we shall see in Sect. 4.3, it has better or comparable performance than the other updater operators. For the computation of MLE, we used the best possible method available for each model. For the Gaussian model, the MLEs can be computed exactly. For the Schaefer model, computing the MLEs is hard, therefore we used our MBE-like approach to approximate the MLEs. This involved applying 100 Bayesian updates using the variance-boosted update operator on three different priors. The estimates with the highest likelihood were chosen as the approximate MLEs. We used variance-boosted Bayesian update operator to approximate MLEs because as we shall see in Sect. 4.4 it computes the MLEs more efficiently than alternative methods. For logistic regression, the MLEs can be easily

Table 4 Priors used in the logistic regression model, where β^* is the MLE on the test set

Parameter	Misspecified prior	Weak prior	Well-specified prior
β_i	$N(-\beta_i^*, CV = 0.1)$	$N(0, \sigma^2 = 100)$	$N(\beta_i^*, CV = 0.1)$

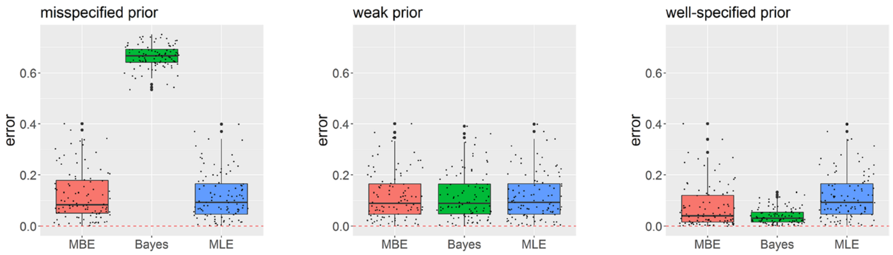


Fig. 2 Absolute errors of MBE, standard Bayesian and MLE for estimating the mean μ only for a Gaussian distribution. A boxplot shows the absolute errors on 100 datasets with 50 observations each

Table 5 Average absolute errors of MBE, standard Bayesian and MLE for estimating the mean μ for a Gaussian distribution. The average is computed over 100 datasets with 50 observations each

	MBE	Bayesian	MLE
Misspecified prior	0.121	0.671	0.117
Weak prior	0.117	0.117	0.117
Well-specified prior	0.075	0.039	0.117
Average	0.104	0.276	0.117

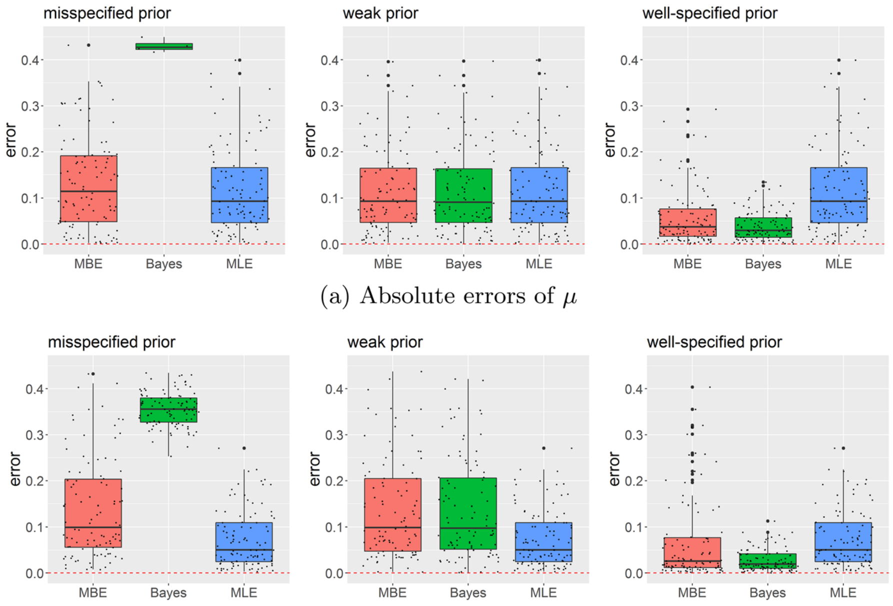
computed using optimization algorithms implemented in various standard libraries — we used the *glm* function in R to compute the MLEs. In the remainder of this section, we shall first present the results on the three inference problems separately, and then discuss the overall performance of the inference methods.

4.2.1 Gaussian model

We conducted a comparison of standard Bayesian, MBE, and MLE on 100 randomly generated datasets, each consisting of 50 observations, by measuring the absolute errors of the estimates, which are the absolute differences between the estimates and the true values.

When only the mean needs to be estimated, we present the boxplots and averages of the absolute errors are shown in Fig. 2 and Tabel 5 respectively. When both the mean and the standard deviation need to be estimated, we show the the boxplots and averages of the absolute errors in Fig. 3 and Table 6.

In cases where the prior is misspecified, the MBE method performs similarly to MLE and outperforms the standard Bayesian. When the prior is weak, MBE and standard Bayesian have similar performance to MLE. With a well-specified prior, MBE and standard Bayesian perform similarly and both outperform MLE. Overall, the performance of MBE is comparable to the best-performing method (either standard Bayesian or MLE) in all three scenarios, while standard Bayesian's performance degrades significantly with a misspecified prior, and MLE is unable to take advantage of a well-specified prior. In addition, MBE outperforms standard Bayesian and MLE in terms of the average performance across the three prior settings.



(a) Absolute errors of μ

(b) Absolute errors of σ

Fig. 3 Absolute errors of MBE, standard Bayesian and MLE for estimating the mean μ and the standard deviation σ for a Gaussian distribution. A boxplot shows the absolute errors on 100 datasets with 50 observations each

Table 6 Average absolute errors of MBE, standard Bayesian and MLE for estimating both the mean μ and the standard deviation σ for a Gaussian distribution. The average is computed over 100 datasets with 50 observations each

	μ			σ		
	MBE	Bayesian	MLE	MBE	Bayesian	MLE
Misspecified prior	0.129	0.345	0.117	0.075	0.494	0.074
Weak prior	0.117	0.117	0.117	0.074	0.073	0.074
Well-specified prior	0.059	0.040	0.117	0.051	0.036	0.074
Average	0.102	0.167	0.117	0.067	0.201	0.074

The results suggest that MBE is capable of appropriately balancing the influence of the prior and data, regardless of whether the prior is useful or misleading.

To gain a deeper understanding of the MBE method, we illustrate in Fig. 4 how the posterior evolves as the number of iterations increases, for the problem of estimating the mean only. As the number of Bayesian updates increases, the posterior moves away from the misspecified prior (shown in light green) towards MLE (which is centered around 0 in this case).

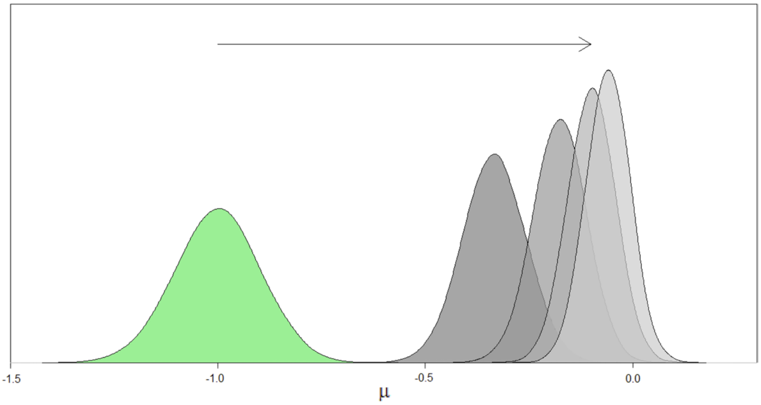
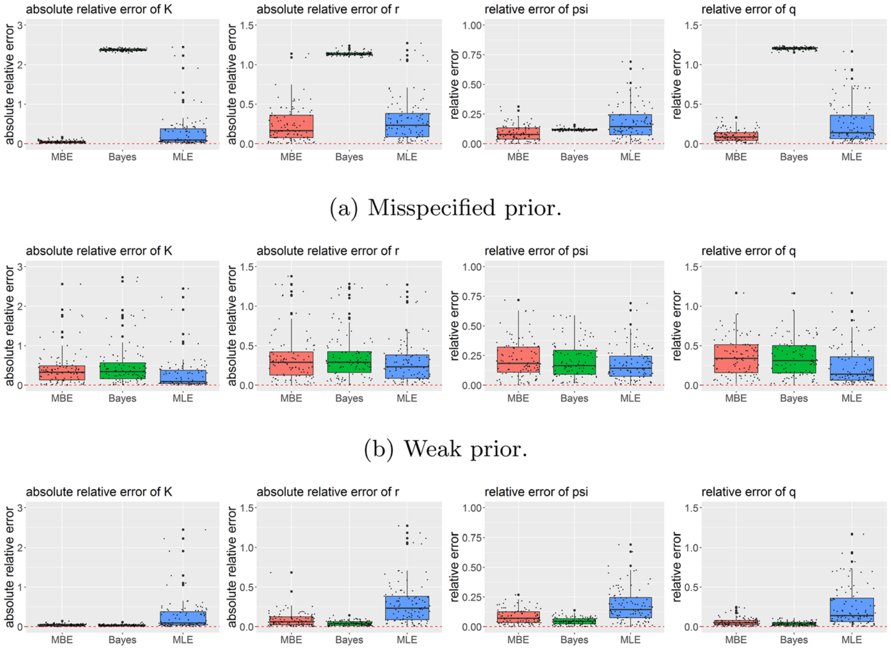


Fig. 4 An illustration of how the posterior evolves when applying MBE to a Gaussian distribution. The posterior for the mean μ (shown in grey with lighter color corresponding to more updates) moves away from the misspecified prior (shown in light green) along the direction of the arrow towards the MLE as the number of iterations increases



(a) Misspecified prior.

(b) Weak prior.

(c) Well-specified prior.

Fig. 5 Absolute relative errors of MBE, standard Bayesian and MLE for estimating the Schaefer model. Each boxplot shows the absolute relative errors for an estimated parameter on 100 datasets with 30 observations each

Table 7 Average absolute relative errors of MBE, standard Bayesian and MLE for estimating the Schaefer model. The average is computed over 100 datasets with 30 observations each

	MBE	Bayesian	MLE
Misspecified prior	0.118	1.209	0.258
Weak prior	0.354	0.360	0.258
Well-specified prior	0.068	0.040	0.258
Average	0.180	0.537	0.258

4.2.2 Schaefer model

We compare the inference methods in terms of their absolute relative errors on 100 randomly generated datasets. Figure 5 and Tabel 7 show the boxplots and averages of the absolute relative errors of the four key parameters (K, q, r, ψ).

Overall, our results for the fisheries data are similar to those for the Gaussian examples, in that MBE is more robust to the priors when compared with standard Bayesian and MLE. It is noteworthy that when using a misspecified prior, MBE has better performance than both the standard Bayesian and MLE methods.

Table 8 Accuracies of standard Bayesian, MBE, and MLE on six classification datasets. The largest average accuracies for the three prior settings are highlighted in bold

	Breast-cancer			Diabetes		
Prior	MBE	Bayesian	MLE	MBE	Bayesian	MLE
Misspecified prior	0.968	0.817	0.957	0.770	0.322	0.766
Weak prior	0.960	0.960	0.957	0.769	0.768	0.766
Well-specified prior	0.973	0.973	0.957	0.782	0.782	0.766
Average	**0.967**	0.917	0.957	**0.774**	0.624	0.766
	German.numer			Heart		
Prior	**MBE**	Bayesian	MLE	**MBE**	Bayesian	MLE
Misspecified prior	0.752	0.349	0.749	0.834	0.161	0.824
Weak prior	0.755	0.753	0.749	0.821	0.819	0.824
Well-specified prior	0.789	0.789	0.749	0.868	0.870	0.824
Average	**0.765**	0.630	0.749	**0.841**	0.617	0.824
	Svmguide1			Svmguide3		
Prior	**MBE**	Bayesian	MLE	**MBE**	Bayesian	MLE
Misspecified prior	0.951	0.924	0.948	0.797	0.546	0.812
Weak prior	0.952	0.952	0.948	0.818	0.816	0.812
Well-specified prior	0.953	0.953	0.948	0.829	0.830	0.812
Average	**0.952**	0.943	0.948	**0.815**	0.731	0.812

4.2.3 Logistic regression

We compared the test set accuracy of the logistic models estimated using standard Bayesian, MBE, and MLE. Table 8 shows the average test set accuracy of standard Bayesian, MBE, and MLE on different datasets, where for each dataset, the average is computed over 100 random train-test splits of the dataset. If a misspecified prior is used, the standard Bayesian method will have the lowest accuracy, while the MBE and MLE methods will have similar accuracy. It is worth noting that for some datasets, e.g. breast-cancer, diabetes, german.numer, heart, and svmguide1, MBE outperforms the standard Bayesian and MLE on the misspecified prior at the same time. This is probably because MBE uses cross-validation to select the optimal number of updates, which allows it to choose a model with better generalization performance. In cases where the prior is weak, the three methods perform similarly. When a well-specified prior is used, the standard Bayesian method and MBE tend to have higher accuracy than MLE on all datasets. Overall, the accuracy of the MBE method demonstrates a higher level of accuracy, tending to be closer to the best-performing method or highest when different priors are used. Note that on svmguide1 and svmguide3, MBE and MLE perform very similarly. This is likely because the two datasets are the largest ones, thus allowing MLE to learn a very good model.

4.2.4 Overall performance

We briefly discuss the average performance of the inference methods across the three prior settings. On the Gaussian datasets, MBE and MLE perform similarly and they are substantially better than standard Bayesian. On the Schaefer datasets, MBE and standard Bayesian perform similarly and they are substantially better than MLE. On the logistic regression datasets, MBE is generally slightly better than MLE, which is in turn substantially better than standard Bayesian. Taking these results together, we can see that MBE demonstrates robustness against bias in the prior and the capability to exploit the information in a prior. In contrast, standard Bayesian can fail poorly due to the bias in the prior, and MLE is not able to exploit any useful information in the prior.

4.3 Comparison of MBE variants

In this section, we compared the performance of the variants of MBE on estimating the Gaussian and Shaefer parameters. We also included a variant that performs a fixed number of updates to demonstrate the need to select a suitable number of updates. To make it easier to follow, we use the following acronyms to refer to different variants of the MBE method: E-MBE for MBE using an exact Bayesian update, MM-MBE for MBE using a moment-matching update, VB-MBE for MBE using a variance-boosted Bayesian update, and FIX-MBE for MM-MBE with a fixed

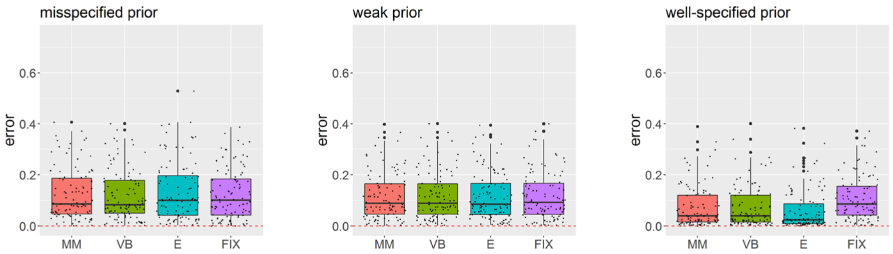
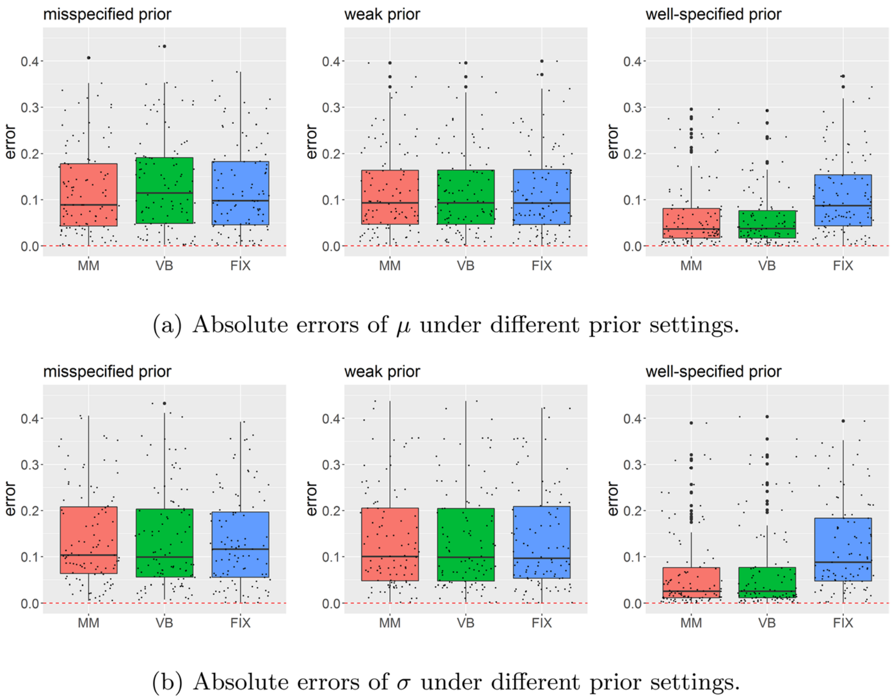


Fig. 6 Absolute errors of MM-MBE, VB-MBE, E-MBE and FIX-MBE for estimating the mean μ only for a Gaussian distribution. Each boxplot shows the absolute errors on 100 datasets with 50 observations each

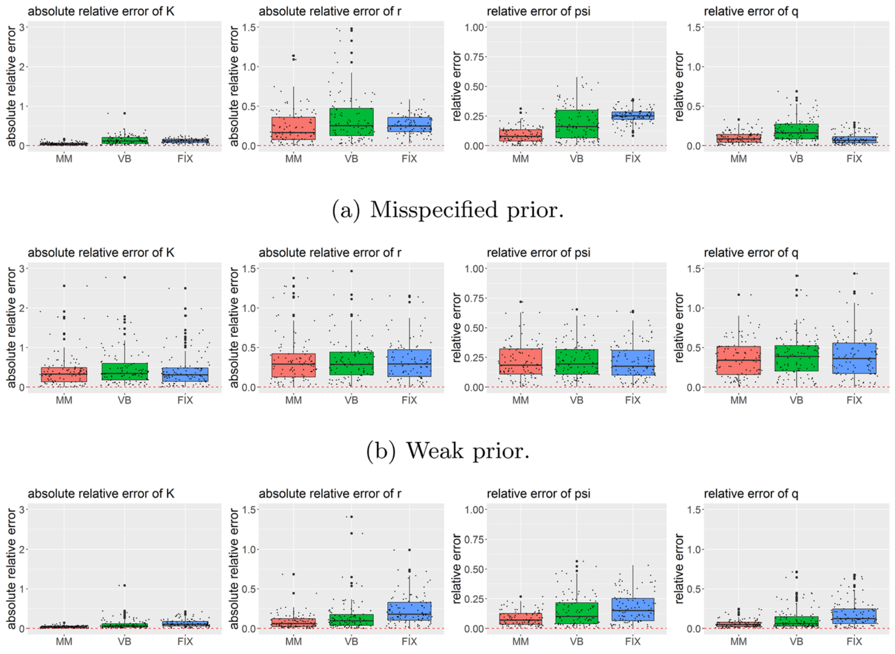
number of 25 updates. Note that 25 was used in our previous preliminary study (Lei et al 2021), but in general using other constants did not seem to perform well across different prior settings in our ad hoc experiments. This is expected because for well-specified priors, we should only perform a small number of updates to fully exploit the prior information, while for misspecified priors, we should perform a larger number of updates to reduce the influence of the prior.



(a) Absolute errors of μ under different prior settings.

(b) Absolute errors of σ under different prior settings.

Fig. 7 Absolute errors of MM-MBE, VB-MBE, and FIX-MBE for estimating both the mean μ and the standard deviation σ for a Gaussian distribution. A boxplot shows the absolute errors on 100 datasets with 50 observations each



(a) Misspecified prior.

(b) Weak prior.

(c) Well-specified prior.

Fig. 8 Absolute errors of MM-MBE, VB-MBE, and FIX-MBE for estimating a Schaefer model. Each boxplot shows the absolute errors on 100 datasets with 30 observations each

In Fig. 6, we compare the absolute errors of three MBE variants (E-MBE, MM-MBE, and VB-MBE) with MM-MBE with fixed numbers of iterations when only the mean of the Gaussian needs to be estimated. When estimating both the mean and standard deviation of the Gaussian and the parameters of the Schaefer model, it is not feasible to perform exact Bayesian update, thus we only compare the absolute errors of MM-MBE, VB-MBE, and FIX-MBE in Figs. 7 and 8.

In the Gaussian experiments, all four MBE variants show similar results. E-MBE performs slightly better in well-specified prior settings compared to VB-MBE and MM-MBE, and FIX-MBE has the largest absolute errors. For the Schaefer model, MM-MBE outperforms VB-MBE and FIX-MBE for the misspecified and the well-specified prior, and the three methods perform similarly for the weak prior with FIX-MBE to be slightly worse. Overall, MM-MBE has better or comparable performance than the other variants of MBE.

4.4 Approximating MLE

In this section, we evaluate the effectiveness of our MBE-like approach for computing the MLE when using different update operators. The exact Bayesian update is not always possible because we often do not have a closed form for the posterior.

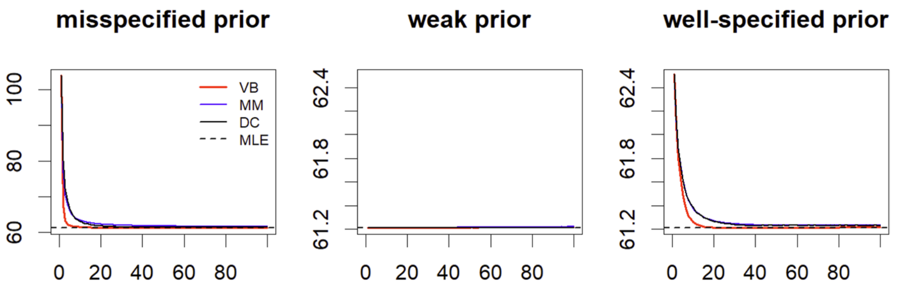


Fig. 9 Plots of NLL against iteration for VB-MBE, MM-MBE, DC, and true MLE on Gaussian data. Note that the y axis ranges for the weak prior and well-specified prior are the same, but these are different from the y axis range for the misspecified prior

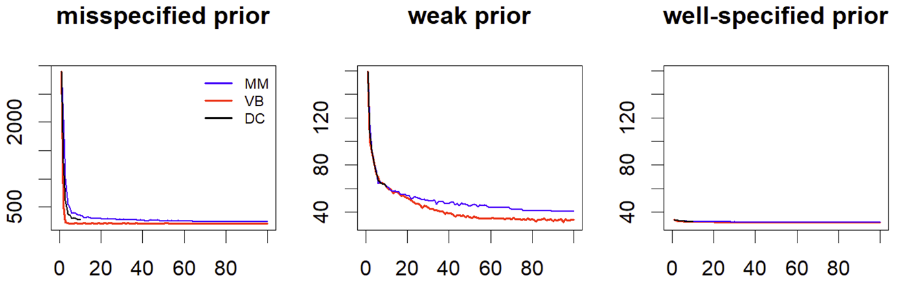


Fig. 10 Plots of NLL against iteration for VB-MBE, MM-MBE, and DC on the Schaefer model. Note that the y-axis ranges for the weak prior and well-specified prior are the same, but these are different from the y-axis range for the misspecified prior

We thus use data cloning (DC) to approximate multiple exact Bayesian updates, that is, we approximate p_j obtained using j exact Bayesian updates by the posterior computed using MCMC on a dataset consisting of j copies of the data. We used the posterior medians as parameter estimates for our MBE-like methods. We shall simply use the MBE names to refer to the corresponding MBE-like methods for computing the MLE below. We compared the performance of VB-MBE, MM-MBE and DC on the three models in Sect. 4.1 by plotting the negative log-likelihood (NLL) value against the iteration.

We make two remarks on our experimental setup. First, the MLE for the Gaussian distribution can be exactly computed using a closed-form formula, and the MLE for logistic regression can be accurately and efficiently computed using the *glm* function in R. The experiments are mainly conducted to better understand how the estimation methods perform in general. Experiments on these two models additionally allow us to study whether the estimation methods converge to the true MLE. Second, DC's computation time increases rapidly with the number of copies of data used, thus we only run it for 10 iterations on the Schaefer model and 20 iterations on the logistic regression models.

For the Gaussian distribution, we consider estimating both the mean μ and the standard deviation σ. Figure 9 plots the average NLL values on 100 random datasets against the number of iterations for different priors. The results indicate that, for the Gaussian experiments, VB-MBE has a faster convergence rate than DC and MM-MBE when misspecified priors and well-specified priors are used. However, when weak priors are used, all three methods quickly converged to the MLE.

For the Schaefer model, Fig. 10 plots the average NLL values on 100 random datasets against the number of iterations. VB-MBE shows faster convergence rate and lower final NLL than MM-MBE and DC in all three prior settings. All three methods seem to converge to different values for different priors. For example, VB-MBE converges to 235.5, 37.8, and 37.2 respectively for misspecified, weak, and well-specified prior. This is possibly because the convergence rates are too slow for the misspecified priors, or the convergence theory does not apply for the Schaefer model.

For the logistic regression model, we are interested in computing the MLE of the test set instead of the entire dataset, because the quality of the priors in Table 4 are set with respect to the test set MLE. Instead of computing the average NLL values over 100 random datasets as for the Gaussian distribution and the Schaefer model, we compute the averages over 10 random datasets, because the experiments are very time consuming. However, the performances of the inference methods are quite similar on the individual datasets, thus we expect the observations below to hold in general.

Figure 11 plots the average NLL values against the number of iterations. For misspecified priors, VB-MBE converges to the MLE faster than MM-MBE and DC for all datasets. For weak priors, DC generally performs better than VB-MBE and MM-MBE. However, note that the differences between these methods are actually quite small in general (much less than 1). They appear to be large because the y-axis ranges are very small — much smaller than the y-axis range for the misspecified case. For well-specified priors, all three methods perform similarly. Besides, when using the weak prior or well-specified prior, the inferences of the three methods at the first iteration are already very close to the MLE.

Overall, VB-MBE has a faster or competitive convergence rate to a lower or similar NLL value than MM-MBE and DC in most cases. This is particularly so for the most challenging case of computing the MLE for the Schaefer model. It should be noted that for all methods, the estimates may be sensitive to the prior for a complex model. We recommend running VB-MBE with both weak and informative priors and selecting the inference result with the highest likelihood.

4.5 Implementation details

We provide some details on how the Bayesian update operators and the validation scores are implemented on our benchmark problems.

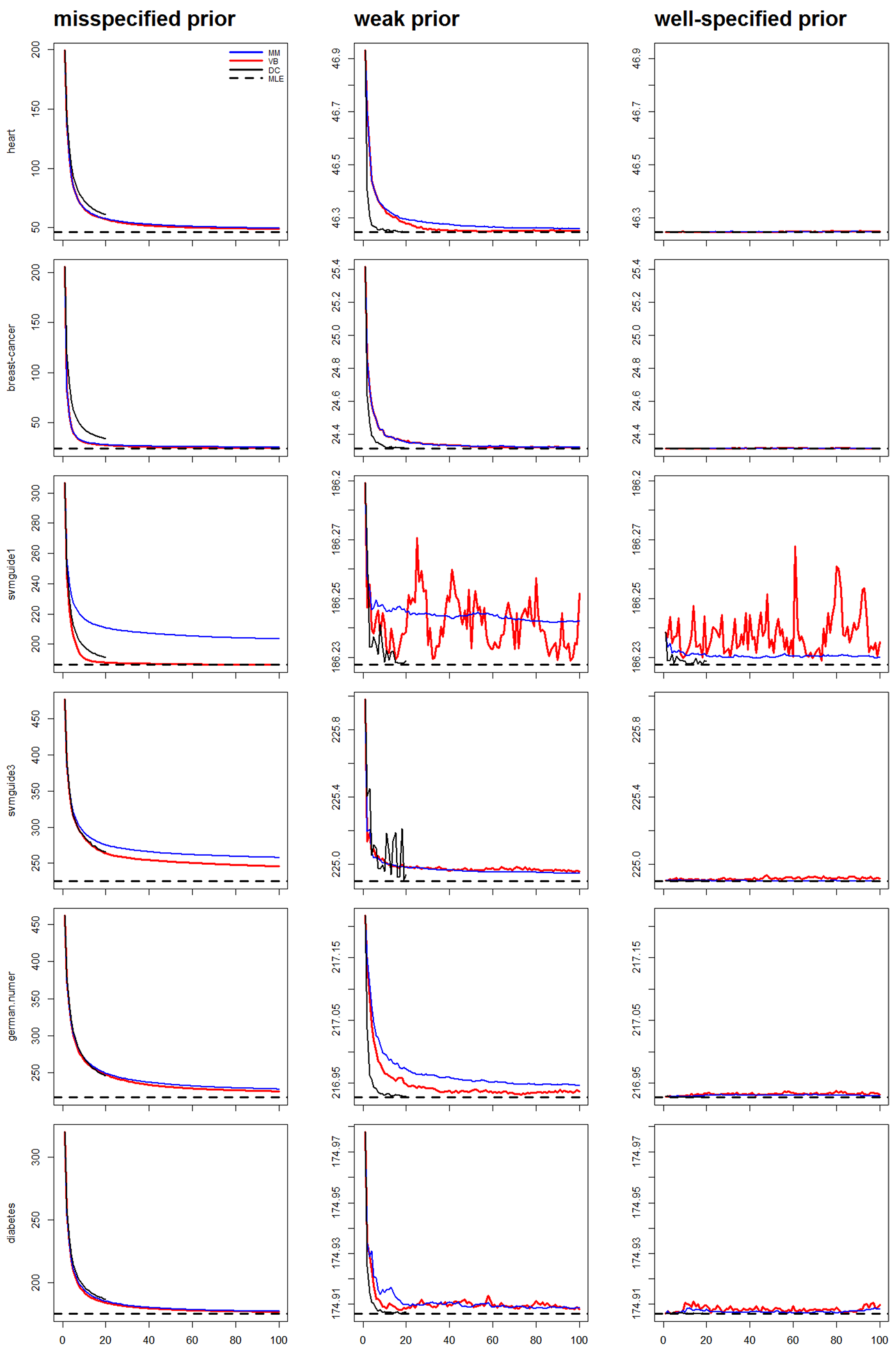


Fig. 11 Plots of NLL against iteration for MM-MBE, VB-MBE, and DC on the logistic regression model. Note that the y-axis ranges for the weak prior and well-specified prior are the same, but these are different from the y-axis range for the misspecified prior

4.5.1 Exact Bayesian update

The exact Bayesian update has a closed-form formula when estimating the mean of a Gaussian distribution with a known σ. Specifically, when estimating the mean μ of a Gaussian with a known σ from a dataset $x_1, \dots, x_n$ and a prior $N(\mu_0, \sigma_0^2)$ for μ, the posterior at the j-th iteration for E-MBE is given by

$$p_j(\mu) = N\left(\mu; \frac{jn\tau\bar{x} + \tau_0\mu_0}{jn\tau + \tau_0}, \frac{1}{jn\tau + \tau_0}\right), \tag{18}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean, and $\tau = \sigma^{-2}$ and $\tau_0 = \sigma_0^{-2}$ are the precision parameters.

For our other benchmark problems, there is no closed-form formula for the exact Bayesian update.

4.5.2 Moment-matching update

Moment-matching updates in MBE admit simple and efficient implementations, because a posterior is computed using MCMC and thus represented using a set of particles. To illustrate, consider estimating a single parameter θ. Assume that we have a posterior $q(\theta)$ approximately represented by a sample $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(n)}$. If we want to choose a log-normal distribution $\tilde{p} = \text{lognormal}(\mu, \sigma^2)$ to approximate the posterior, one possible way to do this is to match the mean and the variance with the sample for the posterior, that is,

$$e^{\mu+\sigma^2/2} = \bar{\theta}, \quad (e^{\sigma^2} - 1)e^{2\mu+\sigma^2} = S^2, \tag{19}$$

where $e^{\mu+\sigma^2/2}$ and $(e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$ are the expectation and variance of $\tilde{p}$, and $\bar{\theta} = \frac{\sum_i \theta^{(i)}}{n}$ and $S^2 = \frac{\sum_i (\theta^{(i)} - \bar{\theta})^2}{n}$ are the sample mean and the variance of the sample for $q(\theta)$. Solving the above two equations, we have

$$.\mu = \log\left(\frac{\bar{\theta}^2}{\sqrt{S^2 + \bar{\theta}^2}}\right), \quad \sigma^2 = \log\left(\frac{S^2}{\bar{\theta}^2} + 1\right) \tag{20}$$

Similar calculations can often be easily done for other distributions used in our benchmark problems.

When θ has multiple parameters, we can perform moment matching for multivariate distributions. However, in this paper, as pointed out in Sect. 2.1, we choose $\tilde{p}$ as a product of univariate distributions and learn the parameters of the univariate distributions using moment matching.

4.5.3 Variance-boosted Bayesian update

Variance-boosted Bayesian update adds an additional *variance boosting* step to moment-matching Bayesian update: it increases the variance of the distribution from moment matching if it falls below a certain user-specified threshold ν. This is straightforward to implement. For example, the variance-boosted Bayesian update for the log-normal distribution example above is given by

$$\mu = \log\left(\frac{\bar{\theta}^2}{\sqrt{S^2 + \bar{\theta}^2}}\right), \quad \sigma^2 = \max\left\{\log\left(\frac{S^2}{\bar{\theta}^2} + 1\right), \nu\right\}. \tag{21}$$

4.5.4 Validation score

As discussed in Sect. 2.2, exact computation of the validation score $\mathcal{S}(p, D) = \int p(D \mid \theta)p(\theta)d\theta$ is generally not possible, and we use the median approximation $\mathcal{S}(p, D) \approx p(D \mid \theta_{\text{med}})$, where θ_{med} is the parameter vector where each component takes its median value according to its marginal distribution given by p.

For Gaussians and logistic regression, computing the median approximation can be done exactly using closed-form likelihood formulas. For the Schaefer model, a Monte Carlo estimate of the likelihood is used as discussed at the end of Sect. 4.1.2.

5 Conclusion

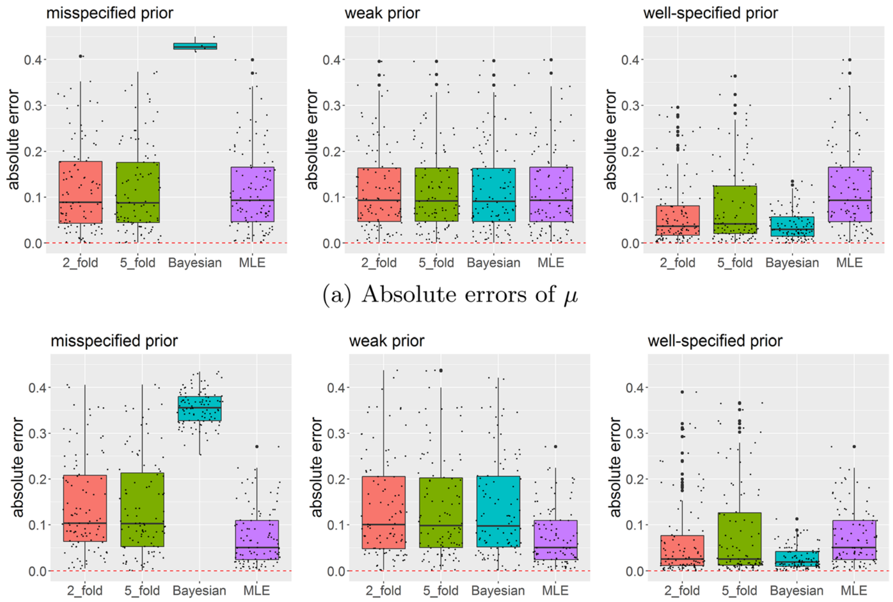
In this paper, we proposed a new Bayesian technique called MBE that is robust against bias in the prior. MBE performs iterative approximate Bayesian updates and aims to optimally adjust the relative importance between the prior and data. When the prior is biased, MBE essentially ignores the prior; and when the prior is informative, MBE is able to exploit the information from the prior. This allows MBE to achieve better overall performance across different prior settings than both standard Bayesian and MLE.

We also introduce an MBE-like approach to estimate the MLE. We present a theoretical analysis for our approach, which contains a new convergence result that is applicable to data cloning. Our MBE-like approach using variance boosting is efficient and demonstrates fast convergence rate as compared to data cloning, particularly on a complex model and on misspecified priors.

Appendix A. *N*-fold Cross-Validation MBE

Algorithm 2 shows an N-fold cross-validation version of the MBE algorithm. This is a direct generalization of the 2-fold cross-validation MBE algorithm shown in Algorithm 1 where the 2-fold cross-validation procedure is replaced by an N-fold one.

5-fold cross-validation and 2-fold cross-validation perform similarly on the Gaussian dataset (Fig. 12) and the logistic regression datasets (Table 9) — note that we did not do the experiments on the Schaefer model due to the high computational cost of computing the validation score. However, for some datasets, 2-fold



(a) Absolute errors of μ

(b) Absolute errors of σ

Fig. 12 Absolute errors of MBE with 2-fold and 5-fold cross-validation, standard Bayesian and MLE for estimating the mean μ and the standard deviation σ for a Gaussian distribution

Table 9 Accuracy of MBE with 2-fold and 5-fold cross-validation, standard Bayesian, and MLE on six classification datasets. The largest average accuracies across the three prior settings are highlighted in bold

	Breast-cancer				Diabetes			
Prior	2-fold	5-fold	Bayes	MLE	2-fold	5-fold	Bayes	MLE
Misspecified	0.968	0.968	0.817	0.957	0.770	0.772	0.322	0.766
Weak	0.960	0.964	0.960	0.957	0.769	0.771	0.768	0.766
Well-specified	0.973	0.969	0.973	0.957	0.782	0.779	0.782	0.766
Average	**0.967**	**0.967**	0.917	0.957	**0.774**	**0.774**	0.624	0.766
	German.numer				Heart			
Prior	2-fold	5-fold	Bayes	MLE	2-fold	5-fold	Bayes	MLE
Misspecified	0.752	0.748	0.349	0.749	0.834	0.834	0.161	0.824
Weak	0.755	0.751	0.753	0.749	0.821	0.815	0.819	0.824
Well-specified	0.789	0.779	0.789	0.749	0.868	0.861	0.870	0.824
Average	**0.765**	0.759	0.630	0.749	**0.841**	0.837	0.617	0.824
	Svmguide1				Svmguide3			
Prior	2-fold	5-fold	Bayes	MLE	2-fold	5-fold	Bayes	MLE
Misspecified	0.951	0.951	0.924	0.948	0.797	0.793	0.546	0.812
Weak	0.952	0.953	0.952	0.948	0.818	0.813	0.816	0.812
Well-specified	0.953	0.951	0.953	0.948	0.829	0.822	0.830	0.812
Average	**0.952**	**0.952**	0.943	0.948	**0.815**	0.809	0.731	0.812

cross-validation is slightly better, possibly because the validation sets in 5-fold cross-validation are small and do not provide reliable estimates on the quality of a relative importance index. Additionally, 2-fold cross-validation is significantly faster than 5-fold cross-validation.

Algorithm 2 MBE with N-fold cross-validation

```
Require: prior $p_0(\theta)$, data $D$
Ensure: a distribution on $\Theta$ for inference
$j^* \leftarrow$ SelectImportance$(p_0, D)$.
for j=1 to $j^*$ do
    $p_j \leftarrow \mathcal{U}(p_{j-1}, D)$
end for
return $p_{j^*}$.

procedure SelectImportance$(p_0, D)$
    Split $D$ into $N$ non-overlapping folds: $(D_1, D_2, \ldots, D_N) \leftarrow Split(D, N)$.
    $j \leftarrow 0$.
    $p_{0,1} = p_{0,2} = \cdots = p_{0,n} = p_0$.
    while termination criterion is not met do
        $j \leftarrow j + 1$.
        for $n = 1$ to $N$ do
            $p_{j,n} \leftarrow \mathcal{U}(p_{j-1,n}, D_1, D_2, \ldots, D_{n-1}, D_{n+1}, \ldots, D_N)$
            $S_{j,n} \leftarrow \log(\mathcal{S}(p_{j,n}, D_n))$
        end for
        $s_j = \sum_{n=1}^{N} S_{j,n}$.
    end while
    $j^* \leftarrow \arg\max_j s_j$, where max is over all computed scores.
return $j^*$
end procedure
```

Appendix B. Sensitivity of Variance Boosting to the Variance Threshold

We used a threshold of 0.1 for the coefficient of variation in variance boosting for the experiments in Sect. 4. We provide additional results on how VB-MBE's NLL changes in the number of iterations for different threshold values in this appendix. Specifically, we considered the thresholds 0.05, 0.1, 0.2, 0.4 and 1. The results for the Gaussian distribution, the Schaefer model, and logistic regression are shown in Figs. 13, 14 and 15 respectively. In general, as the threshold increases, VB-MBE converges faster in terms of the number of iterations, but may result in slower convergence and thus longer computational time for MCMC in each iteration. In

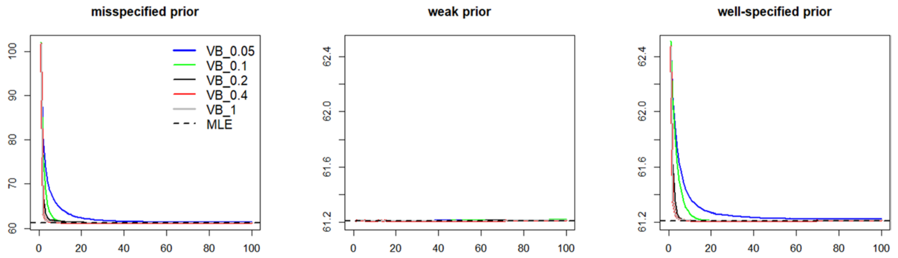


Fig. 13 Plots of NLL against iteration for VB-MBE with different threshold values on Gaussian data. Note that the y-axis ranges for the weak prior and well-specified prior are the same, but these are different from the y-axis range for the misspecified prior

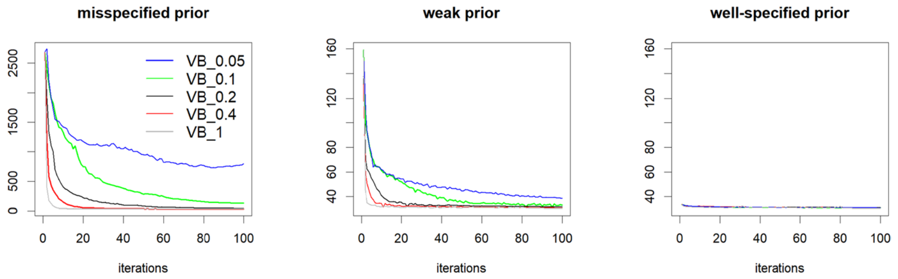


Fig. 14 Plots of NLL against iteration for VB-MBE with different threshold values on the Schaefer model. Note that the y-axis ranges for the weak prior and well-specified prior are the same, but these are different from the y-axis range for the misspecified prior

addition, we observed that MCMC failed to converge in some cases for a large threshold value such as 100, even with a larger number of MCMC iterations. Overall, these results suggest that an alternative threshold on the coefficient of variation may make VB-MBE more efficient for approximate maximum likelihood estimation. Note that a fast convergence of VB-MBE implies that the spectrum of posterior distribution is likely to lack diversity.

Appendix C. Alternative validation score of relative importance indices

We used $s_j = \log(\mathcal{S}(p_{j,1}, D_2)) + \log(\mathcal{S}(p_{j,2}, D_1))$ as the validation score for the relative importance index j in Sect. 4. In this appendix, we compared this default validation score with an alternative validation score $s_j = \log(\mathcal{S}(p_{j,1}, D_2) + \mathcal{S}(p_{j,2}, D_1))$. These two scores will be called MBE_1 and MBE_2 respectively below. The results on the Gaussian distribution (Fig. 16) and logistic regression (Table 10) show that the two validation scores yield similar results, thus even though their values are generally different, they may provide similar rankings for the relative importance

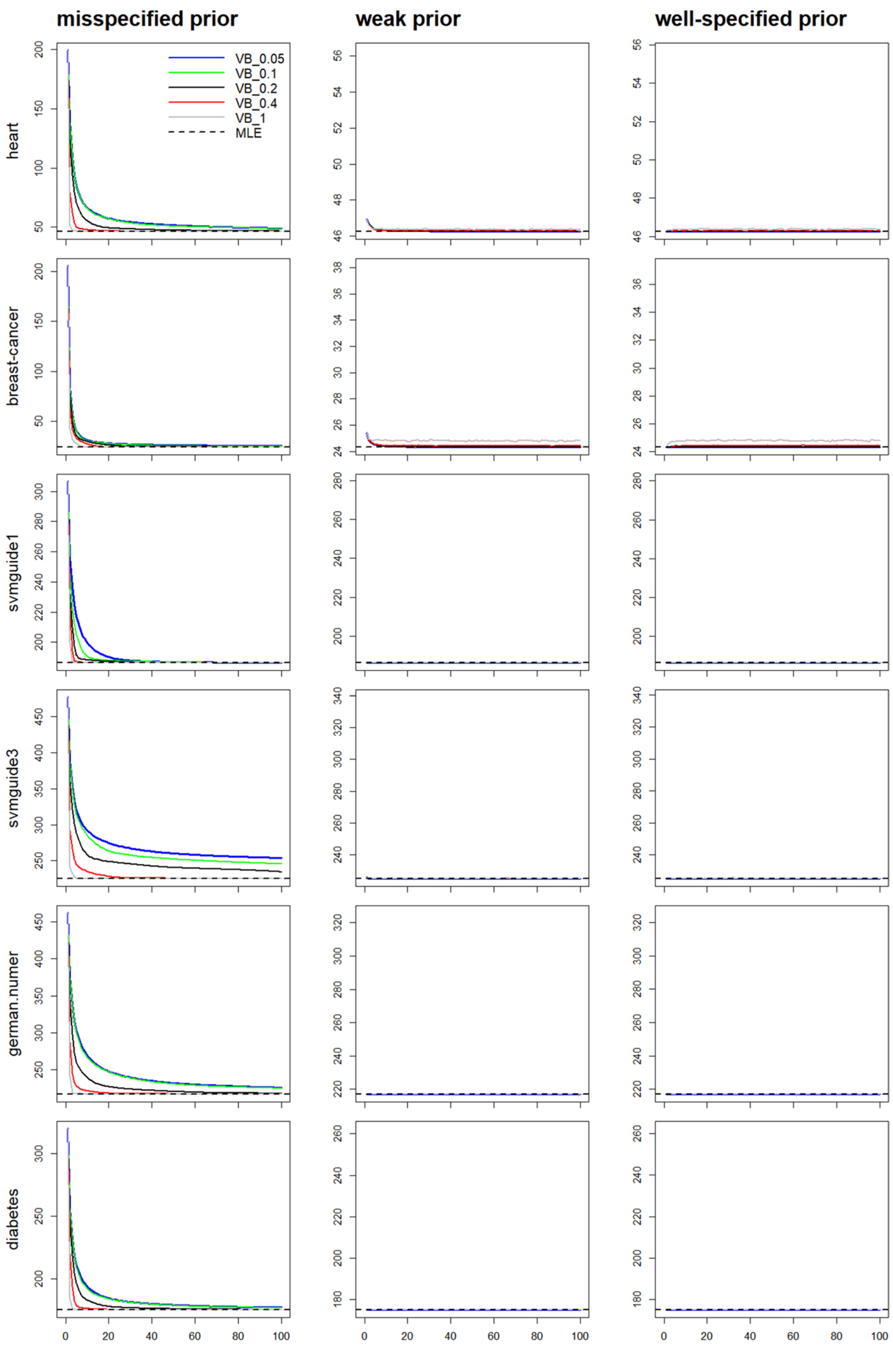
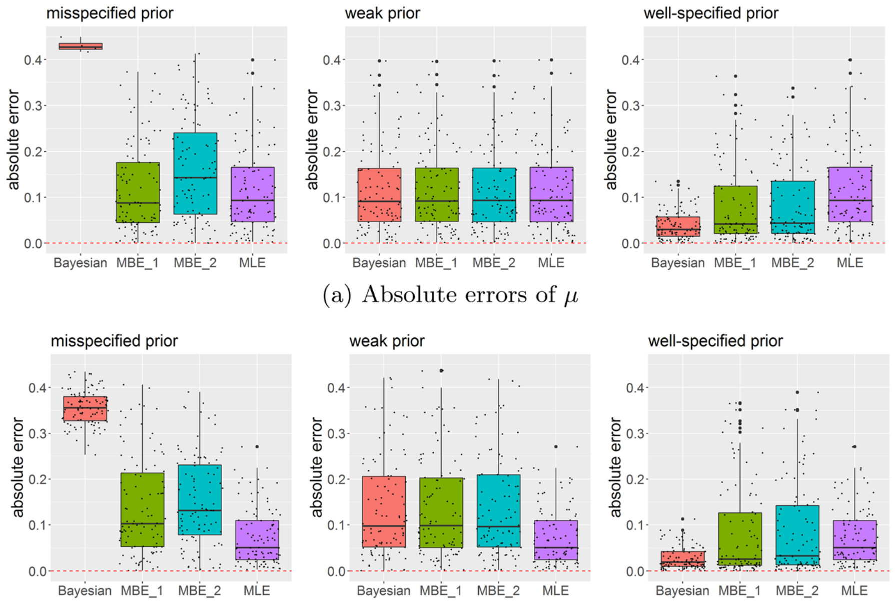


Fig. 15 Plots of NLL against iteration for VB-MBE with different threshold values on the logistic regression model. Note that the y-axis ranges for the weak prior and well-specified prior are the same, but these are different from the y-axis range for the misspecified prior



(a) Absolute errors of μ

(b) Absolute errors of σ

Fig. 16 Absolute errors of MBE with different validation scores, standard Bayesian, and MLE for estimating the mean μ and the standard deviation σ for a Gaussian distribution. A boxplot shows the absolute errors on 100 datasets with 50 observations each

Table 10 Accuracies of MBE with different validation scores, standard Bayesian, and MLE on six classification datasets. The largest average accuracies across the three prior settings are highlighted in bold

	Breast-cancer				Diabetes			
Prior	MBE_1	MBE_2	Bayes	MLE	MBE_1	MBE_2	Bayes	MLE
Misspecified	0.968	0.967	0.817	0.957	0.772	0.772	0.322	0.766
Weak	0.964	0.964	0.960	0.957	0.771	0.770	0.768	0.766
Well-specified	0.969	0.971	0.973	0.957	0.779	0.779	0.782	0.766
Average	**0.967**	**0.967**	0.917	0.957	**0.774**	**0.774**	0.624	0.766
	German.numer				Heart			
Prior	MBE_1	MBE_2	Bayes	MLE	MBE_1	MBE_2	Bayes	MLE
Misspecified	0.748	0.748	0.349	0.749	0.834	0.834	0.161	0.824
Weak	0.751	0.752	0.753	0.749	0.815	0.814	0.819	0.824
Well-specified	0.779	0.782	0.789	0.749	0.861	0.867	0.870	0.824
Average	0.759	**0.761**	0.630	0.749	0.837	**0.838**	0.617	0.824
	Svmguide1				Svmguide3			
Prior	MBE_1	MBE_2	Bayes	MLE	MBE_1	MBE_2	Bayes	MLE
Misspecified	0.951	0.951	0.924	0.948	0.793	0.793	0.546	0.812
Weak	0.953	0.953	0.952	0.948	0.813	0.814	0.816	0.812
Well-specified	0.951	0.952	0.953	0.948	0.822	0.820	0.830	0.812
Average	**0.952**	**0.952**	0.943	0.948	**0.809**	**0.809**	0.731	0.812

indices. Note that we did not do the experiments on the Schaefer model due to the high computational cost of calculating the validation score.

Acknowledgements This work is partially supported by the 2022 Coastal & Ocean Collaboration Funding, developed jointly by The University of Queensland (UQ) and the Commonwealth Scientific and Industrial Research Organisation (CSIRO) Oceans and Atmosphere. We thank the anonymous reviewers for their helpful feedback.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

References

Berger JO, Bernardo JM (1992) On the development of the reference prior method. Bayesian stat 4(4):35–60

Bernardo JM (1979) Reference posterior distributions for Bayesian inference. J Roy Stat Soc: Ser B (Methodol) 41(2):113–128

Bolstad WM, Curran JM (2016) Introduction to Bayesian statistics. Wiley, Hoboken, NJ, USA

Brooks SP, Gelman A (1998) General methods for monitoring convergence of iterative simulations. J Comput Graph Stat 7(4):434–455

Brown JL, Hund LB (2018) Estimating material properties under extreme conditions by using Bayesian model calibration with functional outputs. J Roy Stat Soc: Ser C (Appl Stat) 67(4):1023–1045

Carlin BP, Louis TA (2000) Empirical Bayes: past, present and future. J Am Stat Assoc 95(452):1286–1289

Chang C, Lin C (2011) LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST) 2(3):1–27

Corander J, Sirén J, Arjas E (2008) Bayesian spatial modeling of genetic population structure. Comput Stat 23:111–129

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7(4):457–472

Ghosh M (2011) Objective priors: an introduction for frequentists. Stat Sci 26(2):187–202

Jeffreys H (1946) An invariant form for the prior probability in estimation problems. Proc R Soc Lond A 186(1007):453–461

Kass RE, Wasserman L (1996) The selection of prior distributions by formal rules. J Am Stat Assoc 91(435):1343–1370

Lei Y, Zhou S, Ye N (2021) Prior versus data: A new Bayesian method for fishery stock assessment. In: Proceedings of the 24th International Congress on Modelling and Simulation, pp 43–49

Lele SR, Dennis B, Lutscher F (2007) Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. Ecol Lett 10(7):551–563

Lele SR, Nadeem K, Schmuland B (2010) Estimability and likelihood inference for generalized linear mixed models using data cloning. J Am Stat Assoc 105(492):1617–1625

Lemoine NP (2019) Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. Oikos 128(7):912–928

Lindley DV (1983) Theory and practice of Bayesian statistics. J Royal Stat Soc Ser D (The Statistician) 32(1/2):1–11

Mikkola P, Martin OA, Chandramouli S, et al (2021) Prior knowledge elicitation: The past, present, and future. arXiv preprint arXiv:2112.01380

Millar RB, Meyer R (2000) Non-linear state space modelling of fisheries biomass dynamics by using Metropolis-Hastings within-Gibbs sampling. J Roy Stat Soc: Ser C (Appl Stat) 49(3):327–342

Monnahan CC, Thorson JT, Branch TA (2017) Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. Methods Ecol Evol 8(3):339–348

Plummer M, et al (2003) JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd international workshop on distributed statistical computing, Vienna, Austria., 125.10, pp 1–10

Punt AE, Hilborn R (1997) Fisheries stock assessment and decision analysis: the Bayesian approach. Rev Fish Biol Fish 7(1):35–63

Su Y, Yajima M, Su MY (2015) Package ‘r2jags’. R package version 003-08. https://cran.r-project.org/web/packages/R2jags/

Winker H, Carvalho F, Kapur M (2018) JABBA: just another Bayesian biomass assessment. Fish Res 204:275–288

Yang R, Berger JO (1996) A catalog of noninformative priors. Institute of Statistics and Decision Sciences, Duke University Durham, Durham, NC, USA