Bayesian Analysis (2018) **13**, Number 1, pp. 59–83

Approximation of Bayesian Predictive p-Values with Regression ABC

David J. Nott[*¶], Christopher C. Drovandi[†], Kerrie Mengersen[‡], and Michael Evans[§]

Abstract. In the Bayesian framework a standard approach to model criticism is to compare some function of the observed data to a reference predictive distribution. The result of the comparison can be summarized in the form of a p-value, and computation of some kinds of Bayesian predictive p-values can be challenging. The use of regression adjustment approximate Bayesian computation (ABC) methods is explored for this task. Two problems are considered. The first is approximation of distributions of prior predictive p-values for the purpose of choosing weakly informative priors in the case where the model checking statistic is expensive to compute. Here the computation is difficult because of the need to repeatedly sample from a prior predictive distribution for different values of a prior hyperparameter. The second problem considered is the calibration of posterior predictive p-values so that they are uniformly distributed under some reference distribution for the data. Computation is difficult because the calibration process requires repeated approximation of the posterior for different data sets under the reference distribution. In both these problems we argue that high accuracy in the computations is not required, which makes fast approximations such as regression adjustment ABC very useful. We illustrate our methods with several examples.

Keywords: ABC, Bayesian inference, Bayesian p-values, posterior predictive check, prior predictive check, weakly informative prior.

1 Introduction

We consider Bayesian inference for a parameter θ with prior $p(\theta)$, and a parametric model $p(y|\theta)$ for data y with observed value y_{obs}. An established approach to model criticism in the Bayesian setting involves comparing some function of the observed data to a reference distribution, such as the prior predictive (Box, 1980) or posterior predictive distribution (Guttman, 1967; Rubin, 1984; Gelman et al., 1996). The result of the comparison is usually summarized by a p-value, describing how far out in the tails of the reference predictive distribution the observed data lies. A small p-value indicates surprise and a possible need to reformulate the model. Computation of Bayesian predictive p-values can be challenging, and in this work we consider some approximate

[*]Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, standj@nus.edu.sg

[†]School of Mathematical Sciences, Queensland University of Technology, Brisbane, Queensland, 4072 Australia

[‡]School of Mathematical Sciences, Queensland University of Technology, Brisbane, Queensland, 4072 Australia

[§]Department of Statistics, University of Toronto, Toronto, Ontario, M5S 3G3, Canada

[¶]Corresponding author.

 https://doi.org/10.1214/16-BA1033

methods for that task in some settings where high accuracy is not needed and approximate methods are very attractive. Our methods are based on regression adjustment approximate Bayesian computation (ABC) approaches (Beaumont et al., 2002; Blum, 2010; Blum and François, 2010). In the applications we consider here, unlike the usual ones for ABC methods, it is useful to consider both situations in which the likelihood is tractable as well as when it is not. While the computations are approximate, it is to be noted that all models are in fact wrong and it is primarily gross violations that worry us. The ABC approach seems well-suited to this.

Two problems are considered. The first problem concerns weakly informative prior choice. The notion of a prior weakly informative with respect to a given proper "base" prior was formalized by Evans and Jang (2011), inspired by Gelman (2006), using distributions of p-values for measuring prior-data conflict. The randomness in the distribution of p-values comes from repeated sampling of data from the prior predictive distribution under the base prior. A prior is weakly informative compared to the base prior if prior-data conflicts happen less often when an analysis is done under the alternative prior rather than the base prior, for data simulated under the base prior. We explain the idea in more detail in Section 4. Suppose S is some summary statistic for the data and that a class of priors $p(\theta|\lambda)$ indexed by a hyperparameter λ is considered. We want to make a weakly informative choice of λ compared to some baseline value λ_0. Approximation of distributions of conflict p-values for appropriate test statistics for characterizing weak informativity involves repeated sampling from the prior predictive distributions $p(S|\lambda)$ for a large number of different values λ and this is computationally expensive when simulation of S is expensive. We suggest the use of regression adjustment ABC methods to approximate the simulation step to ease the computational burden. The choice of a good value for λ in order to define a weakly informative prior is simply a screening computation. After an appropriate λ value is chosen, a more accurate calculation for the finally chosen prior can be performed to see if our approximate computations characterized weak informativity well enough.

The second main contribution of the paper concerns calibration of posterior predictive p-values in model checking so that they are uniformly distributed under some reference distribution for the data, such as the prior predictive distribution. For some choices of the statistic used for model checking the corresponding posterior predictive p-values can have a distribution that is far from uniform, clustering around a value of 0.5. This makes it difficult to decide when a certain posterior predictive check has produced a surprising result. Because of this many authors have discussed the need for calibration of posterior predictive p-values to set an interpretable scale for them (Robins et al., 2000; Hjort et al., 2006; Steinbakk and Storvik, 2009). The difficulty is that the calibration process usually involves repeated approximation of the posterior distribution for different data sets under the reference distribution, and this is computationally expensive. A further contribution of this paper is to suggest performing this repeated posterior approximation using regression adjustment ABC methods, which is computationally thrifty since it involves only fitting regression models. We show that the corresponding approximate model check has an interesting interpretation and role regardless of whether the regression approach approximates the calibrated posterior predictive p-value well or not. The interpretation is based on using a regression model

to capture relationships between the parameter, data and a data replicate under the prior, and then seeing whether a certain pseudo-observation for this regression model based on the observed data is an outlier in the regression.

The paper is organized as follows. In the next section we review basic ideas of prior and posterior predictive checks. Section 3 introduces basic ideas of regression ABC and Section 4 considers the problem of weakly informative prior choice. Section 5 applies regression adjustment ABC to calibration of posterior predictive p-values and Section 6 concludes.

2 Prior and posterior predictive checks

A common approach to Bayesian model criticism uses Bayesian predictive model checking. Denoting the observed value of the data y by y_{obs}, we consider some model checking statistic or discrepancy measure $D(y)$, and then for some reference predictive distribution for the data $r(y)$ we consider the distribution of $D(y)$ for $y \sim r(y)$ and determine how far out in the tails of this distribution $D(y_{obs})$ lies. We can summarize the comparison by a p-value,

$$p = P(D(y) \geq D(y_{obs})),$$

where $D(y)$ is defined in such a way that a large value indicates a possibly interesting departure from the model. One choice for the reference predictive distribution $r(y)$ is the prior predictive distribution

$$p(y) = \int p(\theta)p(y|\theta)d\theta.$$

The use of prior predictive p-values in model checking was advocated by Box (1980). Box (1980) suggested use of the statistic $D(y) = 1/p(y)$, and some refinements of Box's approach are suggested by Evans and Moshonov (2006). Prior predictive p-values cannot be used when the prior is improper. However, when the prior is proper, prior predictive p-values based on a minimal sufficient statistic provide one natural way to characterize the informativeness of a prior, a point that has been made in Evans and Moshonov (2006) and Evans and Jang (2011). We use the methods developed in Section 4 in order to approximate computation of the prior predictive p-value distributions they suggest for characterizing weak informativity of one prior with respect to another. Although many authors have developed approaches to detecting prior-data conflict (O'Hagan, 2003; Marshall and Spiegelhalter, 2007; Dahl et al., 2007; Gåsemyr and Natvig, 2009; Scheel et al., 2011; Presanis et al., 2013) they have not been concerned with using conflict to characterize weak informativity of priors.

An alternative choice for the reference distribution $r(y)$ in model checking is the posterior predictive distribution,

$$p(y^*|y_{obs}) = \int p(y^*|\theta)p(\theta|y_{obs})d\theta,$$

where y^* is a predictive replicate of the observed data sharing the same value of the parameter θ. This approach can be useful when the prior is improper and it is also quite easy to implement once a sample from the posterior distribution has already been

generated. If the posterior predictive distribution for the replicate is conflicting with the observed data y_{obs}, then the fitted model is inconsistent with the observed data in some way, and this suggests changing the model. The distribution of the posterior predictive p-value is not necessarily uniform under sampling from the marginal distribution of y, and it can sometimes be helpful to calibrate such p-values to set an interpretable scale for them. This is the problem we take up in Section 4 where the computational difficulties involved in this calibration process are described. Extending the above discussion somewhat, in posterior predictive model checking it makes sense to consider a discrepancy measure which is a function of both the data and parameters, $D(y, \theta)$ say (Gelman et al., 1996). We compare the values of $D(y_{obs}, \theta)$ to the values of $D(y^*, \theta)$ under the joint posterior distribution $p(\theta, y^*|y_{obs})$ for (θ, y^*) where

$$p(\theta, y^*|y_{obs}) \propto p(\theta)p(y_{obs}|\theta)p(y^*|\theta).$$

The comparison of $D(y^*, \theta)$ with $D(y_{obs}, \theta)$ is formalized through the posterior predictive p-value

$$Q(y_{obs}) = P(D(y^*, \theta) \geq D(y_{obs}, \theta)|y_{obs}).$$

Various kinds of replication can be considered within this posterior predictive checking framework, particularly in relation to hierarchical models, and this can be appropriate in different contexts. See Gelman et al. (1996) for further discussion of this.

Crucial to any approach to predictive model checking is the choice of an appropriate discrepancy. For the applications of prior predictive checks to detection of prior-data conflict, it is argued in Evans and Moshonov (2006) that making the discrepancy a function of a minimal sufficient statistic is the right thing to do. This is because dependence of the discrepancy on aspects of the data that don't affect the likelihood can have nothing to do with whether a prior-data conflict occurs. For checking the data model using posterior predictive p-values the choice of discrepancy measure will depend on what we wish to use the model for and it is difficult to give general guidelines. In a later example considered by Hjort et al. (2006) we use the discrepancy considered in their analysis of the same data, namely a kind of generalized Pearson statistic. Global goodness of fit measures such as this are one interesting choice for discrepancy functions, but other choices that might probe more local departures from the model for specific observations or groups of observations as well as discrepancies that reflect the intended use of the model will also be appropriate.

3 Regression adjustment ABC

Computation of Bayesian predictive p-values can sometimes involve expensive and repeated approximations of various conditional distributions; in the problems we consider in this paper, these are either posterior distributions for different data sets or prior predictive distributions for different values of a prior hyperparameter. Since regression analysis is a standard tool for estimation of conditional distributions, it is sensible to ask whether regression can be useful here for the needed computations. Suitable methods already exist in the approximate Bayesian computation (ABC) literature (Marin et al., 2011) and we now explain these.

ABC methods are used for approximate Bayesian inference in situations where simulation from the model is easy but where the likelihood is difficult or impossible to calculate. The most basic ABC methods are based on rejection sampling ideas, but there are more sophisticated variants of the basic approach. We describe only the regression adjustment method of Blum and François (2010), which is an extension of the local linear method of Beaumont et al. (2002); see Marin et al. (2011) for a broader coverage of ABC methods. Suppose as before we have a likelihood $p(y_{obs}|\theta)$ and prior $p(\theta)$. We want to approximate the posterior distribution $p(\theta|y_{obs})$. In the method of Blum and François (2010), as in most ABC methods, we first assume that we can reduce the data y_{obs} to a low-dimensional summary statistic $s_{obs} = S(y_{obs})$ which is informative for θ. A sufficient statistic would be an ideal choice, but practically unattainable in most contexts where ABC is used. Next, suppose we simulate parameters and data sets as $(\theta_i, y_i) \sim p(\theta)p(y|\theta)$, and we write $s_i = S(y_i)$ for the summary statistics corresponding to the y_i, $i = 1, \ldots, n$. Given these simulations we may consider using regression to estimate $p(\theta|s_{obs})$ from the data (θ_i, s_i), $i = 1, \ldots, n$ with θ as response and the summaries s as predictors. For simplicity suppose θ is univariate. Blum and François (2010) consider the model

$$\theta_i = \mu(s_i) + \sigma(s_i)\epsilon_i,$$

where ϵ_i, $i = 1, \ldots, n$ are zero mean, variance one, independent and identically distributed errors, and $\mu(s)$ and $\sigma(s)$ are flexible mean and standard deviation functions. Blum and François (2010) parameterize $\mu(s)$ and $\sigma(s)$ using neural networks and then after fitting to the data obtain estimates $\hat{\mu}(s)$ and $\hat{\sigma}(s)$. Let $\hat{\epsilon}_i$ denote the empirical residual $\hat{\epsilon}_i = \hat{\sigma}(s_i)^{-1}(\theta_i - \hat{\mu}(s_i))$. Approximating the posterior distribution $p(\theta|s_{obs})$ using the fitted regression model at s_{obs} and the empirical residuals gives that

$$\begin{aligned}\theta_i^a &= \hat{\mu}(s_{obs}) + \hat{\sigma}(s_{obs})\hat{\epsilon}_i \\ &= \hat{\mu}(s_{obs}) + \hat{\sigma}(s_{obs})\hat{\sigma}(s_i)^{-1}(\theta_i - \hat{\mu}(s_i)),\end{aligned}$$

$i = 1, \ldots, n$ comprise an approximate sample from $p(\theta|s_{obs})$ if the regression model is correct. A multivariate extension is possible, as well as localization of the fit with a kernel, usually with support chosen to include a certain number of nearest neighbours of s_{obs}. The number of neighbours receiving positive weight is often chosen as a fraction of n. The regression adjusted sample is constructed only using the points given positive weight by the kernel. In the regression adjustment approach approximating the posterior distribution for any value of s_{obs} is easy once the regression model has been fitted, as it involves only moving particles around by mean and scale adjustments. This gives us a fast approximate method for approximating posterior distributions for different data sets based on the same samples from the prior. We will use this approach in Section 4 for calibration of posterior predictive p-values, where we need to approximate a posterior distribution for many different data sets.

Conventional ABC computational algorithms are used to perform Bayesian inference without evaluating the likelihood, using only simulations from the data model. In applications where simulation from the data model is expensive, several authors have noted that conventional ABC computations may be very difficult or impractical and the use of regression methods to replace the data simulation step has been considered (see, for

example, Moores et al. (2015)). The method of Moores et al. (2015) is also related to the synthetic likelihood ABC method of Wood (2010), which uses a working normal model for summary statistics where the mean and covariance are estimated at each parameter value by simulation. In our description of regression ABC above we regressed θ on the summary statistics S in order to approximate the posterior distribution for θ for many different values for S. Now we regress S on θ instead to approximate the distribution of S for many different values of θ. Again suppose we have samples (θ_i, S_i) from $p(\theta)p(S|\theta)$ for θ and some summary statistics S. Fit a regression of S on θ similar to before, using the method of Blum and François (2010) but with the roles of parameters and summary statistics inverted:

$$S_i = \mu(\theta_i) + \sigma(\theta_i)\epsilon_i.$$

Next, approximate the distribution of S given θ using the fitted regression model and the empirical residuals: an approximate sample from $S|\theta$ is

$$S_i^a(\theta) = \hat{\mu}(\theta) + \hat{\sigma}(\theta)\hat{\sigma}(\theta_i)^{-1}(S_i - \hat{\mu}(\theta_i)),$$

$i = 1, \ldots, n$. After the regression model is fitted, generation from the data model for a given θ can be approximated by choosing at random an i uniformly in $\{1, \ldots, n\}$ and returning $S_i^a(\theta)$. This regression approximated data generation step can be used within a conventional ABC algorithm. Some localization of the regression model could be performed within this procedure – i.e. if fitting the regression model is cheap compared to a data generation step, then it could still be attractive to fit a different regression model locally around each θ to generate pseudo-data for every θ value where this is required. In the next section we will focus on using the above idea to undertake repeated generation from a marginalized model $p(S|\lambda)$ where $p(S|\lambda)$ is a prior predictive distribution corresponding to a prior $p(\theta|\lambda)$ and λ is a hyperparameter. That is, $p(S|\lambda) = \int p(S|\theta)p(\theta|\lambda)d\theta$ and we need to generate samples from $p(S|\lambda)$ for many different values of λ. The application we consider is the use of conflict p-values in the spirit of Evans and Moshonov (2006) and Evans and Jang (2011) for characterizing weak informativity of one prior with respect to another. Evans and Jang (2011) consider the situation in which there is a certain base prior which represents our current best information, but where that choice is tentative and we would like to assess sensitivity by finding an alternative prior that is less informative relative to the base prior or to replace the base prior when it is in conflict with the data. Their notion of weak informativity is an attempt to make precise a similar idea suggested in Gelman (2006).

4 Weakly informative prior selection

Evans and Moshonov (2006) and Evans and Jang (2011) consider a decomposition of the data distribution into different components that have different roles in model checking, and suggest that one should check separately for lack of model fit (which means that there is no parameter value for which the observed data is not surprising) and prior-data conflict (which means that there are parameter values providing a good fit to the data but the prior does not give any weight to them). These considerations lead them to suggest using the distribution of a certain conflict prior predictive p-value based on

a minimal sufficient statistic for quantifying weak informativity of a prior distribution. In particular, if a minimal sufficient statistic is denoted by $T = T(Y)$, they suggest that an appropriate p-value for measuring prior-data conflict is the prior predictive p-value with the discrepancy measure $D(t) = 1/p(t)$, where $p(t)$ is the prior predictive distribution of T. They also recommend conditioning on a maximal ancillary statistic when available to remove variation unrelated to the prior, with different choices for the maximal ancillary corresponding to different ways of checking for conflict. Although the use of sufficient statistics might make it seem like the approach is of limited applicability, we note that later we use the maximum likelihood estimator, or an approximation to it, as a general asymptotically sufficient statistic T.

With this way of measuring prior-data conflict, Evans and Jang (2011) consider the distribution of the conflict p-values when the data are distributed according to the prior predictive distribution for the base prior as a tool for evaluating the informativeness of different priors compared to the base prior. Suppose that the class of prior distributions under consideration is $p(\theta|\lambda)$ where λ is a parameter to be chosen, and that the base prior is $p_B(\theta)$. We want to choose λ such that $p(\theta|\lambda)$ is weakly informative with respect to $p_B(\theta)$. The conflict prior predictive p-value is a function of the data. If the data are random, then so is the prior predictive p-value. Evans and Jang (2011) suggest that considering the data as distributed according to the prior predictive distribution under the base prior is natural if the base prior is the best, though perhaps tentative, current representation of prior knowledge. Their prior predictive checking statistic depends on the prior distribution used for the analysis in their approach, and the prior predictive p-value distribution for a given λ is computed using $p(\theta|\lambda)$ in the model checking statistic (i.e. $D(t) = 1/p(t|\lambda)$). Prior-data conflict is characterized by a certain cutoff, γ say, for the conflict p-value. Weak informativity is defined by less prior-data conflict for $p(\theta|\lambda)$ than for $p_B(\theta)$. More precisely, weak informativity at level γ means that the γ quantile of the p-value distribution for $p(\theta|\lambda)$ is greater than the corresponding γ quantile for the p-value distribution for $p_B(\theta)$. As well as characterizing weak informativity for a certain cutoff level γ, one can define uniform weak informativity in various senses as described by Evans and Jang (2011).

4.1 Regression adjustment for exploring weak informativity

Regression adjustment ABC methods are useful in this problem of characterizing weak informativity because we need to repeatedly simulate from the prior predictive distribution of a minimal sufficient statistic T for a grid of values for the hyperparameter λ. That is, we need to repeatedly simulate from the marginalized model $p(T|\lambda)$ for many different values of λ and this can be computationally expensive. Regression adjustment ABC methods can represent a computationally thrifty approximation to the data generation step based on fitting a single regression model. The choice of a good parameter λ to use for a weakly informative prior is simply a screening computation; after an appropriate λ value is chosen we can do a more accurate calculation for the finally chosen prior to see if our approximate regression screening computations characterized weak informativity well enough. High accuracy is not needed in the initial computation and this makes fast approximate approaches very attractive.

Suppose we have some pseudo-prior for λ, $p(\lambda)$, to generate design points for λ in fitting this regression approximation. This pseudo-prior is not used for inference about λ in any way and a deterministic design could be used instead. Let $(\lambda_i, \theta_i, T_i)$, $i = 1, \ldots, n$ be a sample from $p(\lambda)p(\theta|\lambda)p(T|\theta)$. Suppose that a regression model has been fitted,

$$T_i = \mu(\lambda_i) + \sigma(\lambda_i)\epsilon_i.$$

Then for any value of λ an approximate sample from $p(T|\lambda)$ can be obtained as

$$T_i^a(\lambda) = \hat{\mu}(\lambda) + \hat{\sigma}(\lambda)\hat{\sigma}(\lambda_i)^{-1}(T_i - \hat{\mu}(\lambda_i)),$$

$i = 1, \ldots, n$. A kernel estimate based on the samples $T_i^a(\lambda)$ is needed to approximate $p(T|\lambda)$. Write this kernel estimate as $\hat{p}(T|\lambda)$. Then for a sample $T_1^0, \ldots, T_n^0$ generated from the prior predictive distribution under the base measure (this is done exactly, not using the regression approximation), we approximate the distribution of conflict p-values for $p(\theta|\lambda)$ by the empirical distribution of $\hat{P}(T_1^0, \lambda), \ldots, \hat{P}(T_n^0, \lambda)$ where

$$\hat{P}(T_j^0, \lambda) = n^{-1} \sum_{i=1}^{n} I(\hat{p}(T_i^a(\lambda)|\lambda) \leq \hat{p}(T_j^0|\lambda)).$$

This is easily computed for any λ once a training sample $(\lambda_i, \theta_i, T_i)$ has been generated and the regression fitted. It is also possible to fit the regression locally around each value λ in obtaining the values $T_i^a(\lambda)$ but whether this is worthwhile depends on the cost of regression model fitting relative to data generation.

Although the methods we suggest based on regression provide an order of magnitude improvement in terms of computation time compared to the corresponding methods which do not use regression adjustment, the methods are still very computationally intensive. However, the ABC computations are embarrassingly parallelizable so that these methods may become more attractive with improvements in the ease of implementation of parallel computation methods. Implementing these methods on just a few examples can also often provide insights into weakly informative prior choices that may be useful for whole classes of models, so that the value of the methods may extend beyond their application to just a particular example.

4.2 Normal location model

We illustrate the approach first for a simple location normal model where the distribution of the prior predictive p-values can be computed analytically. This is useful as a way of obtaining understanding of the definition of weak informativity, and to show that regression approximations are able to give the correct answer when it is known.

Suppose that $y \sim N(\mu, 1)$ and the base prior for the unknown mean μ is $N(0, 1)$. We consider weak informativity of the priors $N(0, \lambda^2)$ with respect to this base prior for $\lambda \in [0.5, 3]$. In this example a larger variance in the prior is a reasonable characterization of weak informativity; however this isn't always the case as illustrated in the next subsection. A minimal sufficient statistic here is y, and $p(y|\lambda)$ is $N(0, 1 + \lambda^2)$. For

an observed value y_{obs}, the conflict p-value is $P(\log p(y|\lambda) \leq \log p(y_{obs}|\lambda))$ for $y \sim N(0, 1+\lambda^2)$. This probability is $P(y^2 \geq y_{obs}^2) = 2\Phi(\frac{-|y_{obs}|}{\sqrt{1+\lambda^2}})$ where $\Phi(\cdot)$ denotes the standard normal distribution function. The distribution function of this p-value, for y_{obs} distributed according to the $N(0,2)$ prior predictive distribution under the base measure, is

$$\begin{aligned} P\left(2\Phi\left(\frac{-|y_{obs}|}{\sqrt{1+\lambda^2}}\right) \leq p\right) &= P\left(\frac{-|y_{obs}|}{\sqrt{2}} \leq \sqrt{\frac{1+\lambda^2}{2}}\Phi^{-1}\left(\frac{p}{2}\right)\right) \\ &= 2\Phi\left(\sqrt{\frac{1+\lambda^2}{2}}\Phi^{-1}\left(\frac{p}{2}\right)\right). \end{aligned}$$

We can compare this with our approximation of this distribution following the numerical procedure of the last section. Our pseudo-prior for λ is uniform on $[0.5, 3]$. We simulated 100,000 samples of (λ, μ, y) from the prior and then for $\lambda = 0.5, 1, 2$ and 3 we used the 1,000 nearest neighbours for each of these λ to fit a local linear regression model to approximate a sample of size 1,000 from the prior predictive at these λ values. Although we estimate the distribution of p-values only at 4 values of λ from the prior samples we have generated, it is easy, using these same samples, to estimate this distribution for any λ value in $[0.5, 3]$. We used the default implementation of the procedure of Beaumont et al. (2002) in the `abc` package in `R` for the regression adjustment (Csilléry et al., 2012). In using the software for the purpose we describe, the role of the summary statistics and the parameters needs to be reversed compared to the usual ABC applications. We used a kernel estimate (the `density` function in `R` with the default bandwidth selection) to approximate the conflict p-values for a sample of 1000 points sampled from the prior predictive for the base prior.

Figure 1 shows the empirical distribution function (quantile-quantile plots of the p-values versus approximate expected order statistics for the uniform distribution) for these 1000 approximate conflict p-values versus the analytically derived p-value distribution for $\lambda = 0.5, 1, 2$ and 3. The agreement is very good and shows that, for the larger variances $\lambda = 2$ and 3, these priors are weakly informative compared to the $\lambda = 1$ base prior (because the distribution function lying below the diagonal line means that, for every γ, the γ quantile of the p-value distribution is larger than the corresponding quantile for the p-value distribution under the base prior, which is uniform). Here $\lambda = 1$ gives the base prior and the p-value distribution is uniform, whereas for $\lambda = 0.5$ we have a more informative prior with prior-data conflicts being produced more often (quantiles of the p-value distribution under this prior being smaller than quantiles for the uniform distribution).

4.3 Logistic regression example

The following example is considered in Evans and Jang (2011) and is concerned with analysis of a bioassay experiment using a logistic regression model. The issue of sensible default proper priors for logistic regression has received a lot of recent attention in the literature (see Gelman et al. (2008) and the references therein) and so examining weak informative priors for this model is of interest. The bioassay experiment considered here

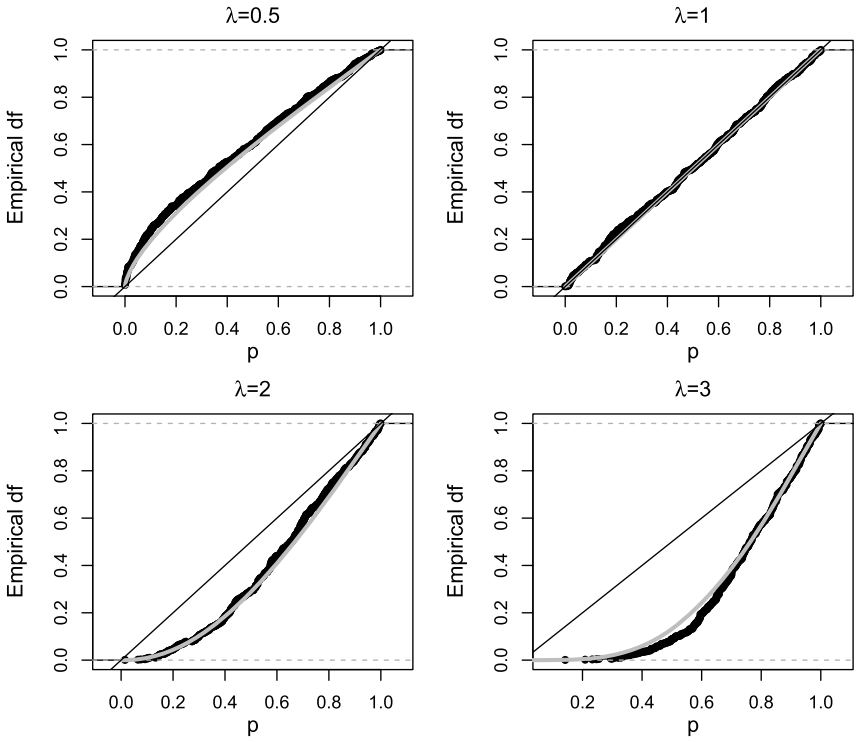


Figure 1: Estimated distribution of conflict p-value for $\lambda = 0.5, 1, 2$ and 3 for the normal location model. In each plot black is the regression estimated distribution and grey is the exact answer.

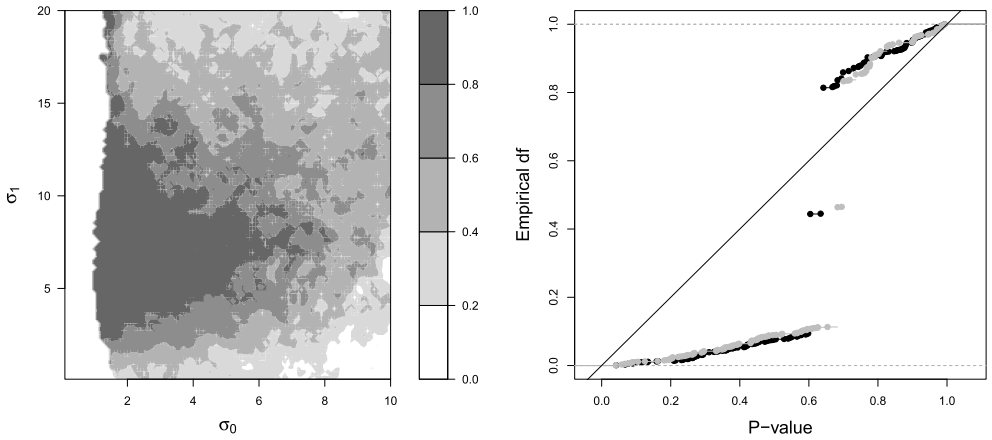
is described more fully in Racine et al. (1986) and Gelman et al. (2008). Five animals at each of four dose levels were exposed to a toxin and the number of deaths were recorded. Let x_i be the dose level (suitably transformed to log scale and then centred and scaled as in Gelman et al. (2008)) and let y_i be the number of deaths out of 5 at dose x_i. A logistic regression model for the data is $y_i \sim \text{Bin}(5, p_i)$ with $\text{logit}(p_i) = \beta_0 + \beta_1 x_i$ where we order the x_i so that $x_1 < x_2 < x_3 < x_4$. We consider a prior where β_0 and β_1 are independent, $\beta_0 \sim N(0, \sigma_0^2)$, $\beta_1 \sim N(0, \sigma_1^2)$. Our base prior puts $\sigma_0 = 10$, $\sigma_1 = 2.5$ which is similar to Gelman et al. (2008), except that they use Cauchy priors instead of normal with these scale parameters. We will investigate weak informativity with respect to the base prior as σ_0 and σ_1 vary for the alternative prior. In Figure 4 of Evans and Jang (2011), four cases are considered for exploring weak informativity in this example. There is a normal or Cauchy choice for the base prior, and normal or Cauchy choices for the alternative priors. We have chosen to consider the first of these (a normal base prior and normal alternative priors) to illustrate our regression approximation methodology.

We use an approximation to the maximum likelihood estimator (MLE) as the basis for an approximate sufficient statistic since the MLE is asymptotically sufficient. Since

the MLE suffers from non-existence for some potential datasets, and these occur in simulations from the prior predictive distribution, we consider the posterior mode for the prior with $\sigma_0 = \sigma_1 = 10$. This posterior mode will be similar to the MLE in non-degenerate cases but the regularization provided by the prior ensures existence and stabilizes the optimization even in degenerate settings. The choice of $\sigma_0 = \sigma_1 = 10$ gives a fairly flat prior over the part of the parameter space corresponding to reasonable sized effects with standardized covariates and is a reasonable prior for the purpose of getting a posterior mode estimate that always exists but is similar to the MLE in non-degenerate settings. Evans and Jang (2011) consider the exact sufficient statistic (y_1, y_2, y_3, y_4) but we use the posterior mode $(\hat{\beta}_0, \hat{\beta}_1)$ for the dimension reduction that this brings and to explore point estimators similar to the MLE as generic choices for statistics for defining conflict in models where there might be no non-trivial minimal sufficient statistic. We also treat the distribution of the mode as continuous even though strictly speaking it is discrete since the data are discrete. In application of our procedure we don't use the values $(\hat{\beta}_0, \hat{\beta}_1)$ directly, but rather transform to the fitted probabilities at the covariate values x_2, x_3. That is, we use for our approximate sufficient statistics $(\hat{p}_2, \hat{p}_3)$ where $\hat{p}_i = 1/(1 + \exp(-\hat{\beta}_0 - \hat{\beta}_1 x_i))$, $i = 2, 3$. Making the approximate sufficient statistic a nearly linear function of the data has the advantage of making the conflict p-value considered here similar to a certain invariantized version of it considered in Evans and Jang (2010). In this example, focusing on $\hat{p}_2$, $\hat{p}_3$ makes sense since these values are plausibly linearly related to the actual deaths at dose levels x_2 and x_3.

The pseudo-prior we use for (σ_0, σ_1) in our procedure is uniform on $[0.1, 10] \times [0.1, 20]$. The choice of the ranges for σ_0 and σ_1 was guided by the analysis in Evans and Jang (2011) which showed that this region was interesting from the point of view of covering the range of hyperparameter values indicating weak informativity with respect to the base prior. We generated 400,000 values for (σ_0, σ_1) and the corresponding $\hat{p}_2, \hat{p}_3$ values from the corresponding prior predictive distribution. For each λ on a 100×100 regular grid covering the support of the hyperprior we used the local linear regression adjustment method of Beaumont et al. (2002) based on applying the default implementation in the `abc` package in `R` (Csilléry et al., 2012) and using the 1000 nearest neighbours at each grid point to get a pseudo-sample from the prior predictive of size 1000. We then considered a kernel density estimate based on these samples and 1,000 samples of $(\hat{p}_2, \hat{p}_3)$ simulated under the prior predictive for the base measure to get an approximation to the distribution of the conflict p-value at each grid point. The generation of summary statistics took 67 hours of CPU time on a quad processor Windows PC 3.10 GHz workstation. Note that if we were to generate 1,000 samples at each of the 10,000 grid points directly we would require $10{,}000{,}000$ samples which would increase the required computational effort by an order of magnitude. The two-dimensional kernel estimation was implemented using the `sm.density` function in the `sm` package in `R` (Bowman and Azzalini, 2014) with the default bandwidth choice. We do not make any adjustment for boundary bias due to the compact support for $(\hat{p}_2, \hat{p}_3)$.

Evans and Jang (2011) suggest that one way to measure the degree of informativity of a prior with respect to the base prior is the following. Choose γ to be a cutoff value for the conflict p-value that defines the degree of conflict of interest (we use $\gamma = 0.05$ here). Let p_γ be the γ quantile of the conflict p-value distribution for the



(a) (b)

Figure 2: (a) Estimated degree of weak informativity at level 0.05 for logistic regression example. (b) Estimated distribution of conflict p-value without using regression adjustment (black) and using regression adjustment (grey).

base prior. Let q_γ be the probability of a conflict p-value less than or equal to p_γ under the alternative prior. Then measure the degree of weak informativity of the alternative prior by $1 - q_\gamma/p_\gamma$ when $q_\gamma \leq p_\gamma$. We define the degree of weak informativity to be 0 if $q_\gamma > p_\gamma$. Figure 2(a) plots the degree of weak informativity with respect to the base prior over the grid of points for (σ_0, σ_1) for $\gamma = 0.05$. This plot is similar to Figure 5 of Evans and Jang (2011) and the result is qualitatively similar, but note that they should not be expected to be exactly the same since we are basing our definition of weak informativity on a posterior mode estimator here which summarizes the likelihood information in a similar way to the MLE while being nondegenerate. Based on the plot, $(\sigma_0, \sigma_1) = (2, 5)$ would seem to be a good choice for a weakly informative prior. Simulating 10,000 values from the prior predictive directly for this alternative prior (that is, not using regression adjustment) and approximating the distribution of conflict p-values gives the black points in Figure 2(b); the grey points in the same figure show the distribution obtained using regression adjustment. The comparison shows that the approximation error induced by the use of the regression approach here is small, in the sense that the two distributions match closely in the lower tail corresponding to small p-values, which are those indicative of conflict. The large lumps of discrete mass in the p-value distributions occur since the base prior here is very diffuse which has the effect of putting a large amount of prior predictive mass on simulated data of $(0, 0, 0, 0)$ and $(5, 5, 5, 5)$. This corresponds to summary statistics $(\hat{p}_2, \hat{p}_3)$ that are very close to either $(0, 0)$ or $(1, 1)$ and so we have large discrete masses on these two summary statistic values. Since the same conflict p-value is returned for the same value of the summary statistic, we get corresponding discrete masses in the distribution of the p-values. These values really have no effect on the assessment of weak informativity of alternative priors with respect to the base prior. This is because these aren't values that produce conflicts

under the base prior and the continuity approximation still works well for the purpose considered here as can be seen by the similarity of Figure 2(a) to Figure 5 of Evans and Jang (2011). Of course there is some disagreement in the Monte Carlo approximations about exactly how much discrete mass is assigned to these points and this is the reason for the disagreement between the curves corresponding to p-values close to 0.5.

5 Calibration of posterior predictive p-values

As discussed in Section 2, the distribution of posterior predictive p-values tends not to be uniform under repeated sampling from the prior predictive distribution. The issue of calibration of such p-values has been discussed in the literature to set an interpretable scale for them. Posterior predictive p-values which are not calibrated tend to have a distribution which is concentrated towards 0.5, which might be thought of as a kind of conservatism if we were to incorrectly interpret the p-value as being drawn from a uniform distribution under ideal model conditions. This section discusses how regression ABC methods can be used for calibration, a task which is computationally difficult due to the need to repeatedly sample from posterior distributions for different data.

5.1 The need for calibration

As in Section 2 we write $Q(y)$ for a posterior predictive p-value based on data y. We suppress dependence on the discrepancy measure chosen in the notation. The idea of calibration of posterior predictive p-values is to compare $Q(y_{obs})$ to $Q(y')$ for $y' \sim m(y)$ where $m(y)$ is some distribution for y such as the prior predictive distribution. An adjusted or calibrated posterior predictive p-value is then given by

$$Q'(y_{obs}) = P(Q(y') \leq Q(y_{obs})).$$

It may be easier to interpret a calibrated posterior predictive p-value, since we know what is expected for it under repeated sampling from the reference predictive distribution.

Computation of this calibrated posterior predictive p-value is difficult. The usual approach (see, for example, Hjort et al. (2006)) is to generate a large number M of data sets from $m(y)$, $y_1, \ldots, y_m$ say, to calculate for each of these corresponding unadjusted posterior predictive p-values $Q(y_1), \ldots, Q(y_m)$, and then to approximate $Q'(y_{obs})$ by the fraction of $Q(y_j)$ less than $Q(y_{obs})$. The difficulty is that computation of each $Q(y_j)$ involves a calculation for a different posterior distribution, so that we must somehow approximate the posterior distribution for M different datasets. Regression adjustment ABC methods can be used to quickly approximate all the required posterior distributions at once with regression calculations. An alternative approach is to use importance sampling, but this does not work well when the data sets are very different. McVinish et al. (2013) consider a modified version of importance sampling that performs better in this respect. Clearly using ABC methods for model criticism is related to the use of ABC methods for model choice, and there is a very active recent literature on this very interesting problem (see the recent review by Marin et al. (2015)).

The issue of calibration of posterior predictive checks is somewhat controversial; some authors (for example, Bayarri and Berger (2000) or Bayarri and Castellanos (2007)) have suggested that the conservatism of posterior predictive p-values is due to their using the data twice, since the posterior distribution based on y_{obs} is being used to predict some function of y_{obs}. Others have argued that posterior predictive p-values have a valid interpretation as a posterior probability with the data conditioned on only once (Gelman, 2013). The conservatism of a particular check can vary according to the choice of the discrepancy. For a general choice of model checking statistic, various methods have been suggested to adjust the reference predictive distribution so that the resulting p-value is approximately uniform, usually by conditioning on some functions of the data in the likelihood to remove some of the information about θ (Bayarri and Berger, 2000; Robins et al., 2000; Bayarri and Castellanos, 2007). Evans and Moshonov (2006) give a decomposition of the joint distribution of (θ, y) into components which corresponds to different sources of information available for model checking, checking the prior and inference for θ, and suggest that this decomposition might be used to understand a bit more precisely when a posterior predictive check could be uninformative. In any case, we believe that the idea of calibration of posterior predictive p-values certainly can perform a useful role in some problems.

5.2 The basic idea

We now describe our ABC approach to approximate calibration of posterior predictive p-values. Suppose we simulate data $(\theta_i, y_i, y_i^*) \sim p(\theta)p(y|\theta)p(y^*|\theta)$ from the prior distribution for θ, y and a predictive replicate y^* and that we have a near sufficient statistic $S(y)$ and a discrepancy measure $D(y, \theta)$. Then from the simulations (θ_i, y_i, y_i^*) we can construct the values $(D(y_i^*, \theta_i) - D(y_{obs}, \theta_i), S(y_i)) = (D_i, s_i)$, $i = 1, \ldots, n$. We will approximate the posterior distribution for the difference of discrepancies $D(y^*, \theta) - D(y_{obs}, \theta)$ directly in computing the p-value since this makes the problem into a univariate one. We consider a regression model

$$D_i = \mu(s_i) + \sigma(s_i)\epsilon_i,$$

and to approximate the distribution of $D(y^*, \theta) - D(y_{obs}, \theta)$ given y_{obs} we use the set of samples

$$D_i^a(s_{obs}) = \hat{\mu}(s_{obs}) + \hat{\sigma}(s_{obs})\hat{\sigma}(s_i)^{-1}(D_i - \hat{\mu}(s_i)),$$

$i = 1, \ldots, n$ where as before $s_{obs} = S(y_{obs})$. Then the unadjusted posterior predictive p-value $Q(y_{obs}) = P(D(y^*, \theta) \geq D(y_{obs}, \theta)|y_{obs}) = P(D(y^*, \theta) - D(y_{obs}, \theta) \geq 0|y_{obs})$ can be approximated by, in cases where $Q(y_{obs})$ cannot be computed analytically,

$$\hat{Q}(y_{obs}) = n^{-1} \sum_{i=1}^{n} I(D_i^a(s_{obs}) \geq 0),$$

where $I(\cdot)$ denotes the indicator function. This is easily calculated for any value of s_{obs} and, if we have datasets $y^{(1)}, \ldots y^{(M)}$ simulated from a reference distribution $m(y)$, we can easily compute $\hat{Q}(y^{(1)}), \ldots \hat{Q}(y^{(M)})$ using the same single fitted regression model.

This idea can also be implemented with the regression models fitted locally, and if the regression calculations are inexpensive this will still be a simple computation. Hence we can approximate the adjusted posterior predictive p-value in a computationally thrifty way. Our estimated adjusted posterior predictive p-value is

$$\hat{Q}'(y_{obs}) = M^{-1} \sum_{i=1}^{M} I(\hat{Q}(y_{obs}) \leq \hat{Q}(y^{(j)})).$$

An anonymous reviewer has asked whether it is also possible to perform checks of a data model based on the conditional distribution of the data given the value of an informative (ideally sufficient) statistic, with ABC methods used to perform the conditioning. Evans and Moshonov (2006) suggest that this is an appropriate way to check this part of the model. The use of ABC methods in this context is an intriguing suggestion, but goes beyond our purpose here of calibrating posterior predictive p-values.

The regression calibration method just suggested seems to rely on the accuracy of the regression model for approximating the posterior distribution of $D(y^*, \theta) - D(y_{obs}, \theta)$ and we might be reluctant to place much faith in this in complicated high-dimensional settings. We now give an alternative motivation for the calculation of $\hat{Q}'(y_{obs})$ as a useful quantity for model criticism and argue that these p-values are interesting regardless of whether we are able to approximate the posterior distribution of $D(y^*, \theta) - D(y, \theta)$ accurately by regression.

5.3 An alternative motivation and some limitations

$\hat{Q}(y_{obs})$ counts the proportion of observations i in the regression training sample for which

$$\hat{\mu}(s_{obs}) + \hat{\sigma}(s_{obs}) \frac{D_i - \hat{\mu}(s_i)}{\hat{\sigma}(s_i)} \geq 0.$$

This inequality can be written

$$\frac{D_i - \hat{\mu}(s_i)}{\hat{\sigma}(s_i)} \geq \frac{-\hat{\mu}(s_{obs})}{\hat{\sigma}(s_{obs})}.$$

The expression on the left is the standardized residual for the ith observation in the regression training sample. To interpret the expression on the right, note that if the data y_{obs} were observed again as the replicate, then this would make the discrepancy measure $D(y^*, \theta) - D(y_{obs}, \theta)$ equal to zero. Hence

$$\frac{-\hat{\mu}(s_{obs})}{\hat{\sigma}(s_{obs})} = \frac{0 - \hat{\mu}(s_{obs})}{\hat{\sigma}(s_{obs})}$$

is the standardized residual in our fitted model for the situation where $(y, y^*, \theta) = (y_{obs}, y_{obs}, \theta)$. Since for the observed data an actual replicate is not observed, assuming the replicate is the same as the observed data minimizes the degree of conflict between the observed data and the replicate. Hence if a value of $(y, y^*, \theta) = (y_{obs}, y_{obs}, \theta)$ is considered surprising, that suggests that y_{obs} is a surprising value under the assumed model.

Our calibration procedure can be seen as adjusting for the conservative assumption that y_{obs} is observed again for the replicate.

In effect, through an outlier analysis of residuals in a regression model fitted to simulations from the prior, an uncalibrated p-value is calculated by comparing residuals for $(y_{obs}, y_{obs}, \theta)$ with (y_i, y_i^*, θ), $i = 1, \ldots, n$ to decide whether y_{obs} is unusual. Furthermore, the calibration procedure does a similar calculation, but to develop a reference distribution for the unusualness of $(y_{obs}, y_{obs}, \theta)$ it considers instead the residuals from observations (y_i, y_i, θ) in order to appropriately account for the fact that we assumed y_{obs} was observed again for the replicate. The main idea, then, is that we can perform model criticism through regression diagnostics for a regression model fitted to simulations from the prior. This is a sensible thing to do regardless of whether the calibrated p-value approximates the calibrated posterior predictive p-value well.

An anonymous reviewer expressed some very natural concerns about the calibration procedure suggested here using regression ABC. One concern relates to whether the regression approximation involves extrapolation, and whether that may occur especially in the situations where there is a model failure and where we most need the method to work well. We argue that this is not the case for appropriate choices of the summary statistics in the ABC procedure and when the posterior predictive check is focused on exploring the appropriateness of the data model. In ABC we try to choose summary statistics that are informative about the parameter and in the present setting a very natural choice is the maximum likelihood estimator (MLE) or some other point estimate of the parameters. The MLE is, after all, asymptotically sufficient, and we use it as the summary statistic in the example of the next subsection. The MLE would not usually be available as a summary statistic in ABC analyses because it requires being able to evaluate the likelihood, but in the model checking applications considered here we are not assuming that the likelihood is unavailable. If the data are highly informative about the parameter, and the MLE is a consistent estimator of the parameter, the prior predictive distribution of this statistic will be similar to the prior distribution on the parameter itself. The value s_{obs} is very different from $s_1, \ldots, s_n$ in the regression adjustment if the MLE for the observed data is very different to values of the MLE simulated from the prior predictive. If the data are very informative about the MLE this occurs when the MLE for the observed data lies out in the tails of the prior on the parameter. This is a question of prior-data conflict and has nothing to do with the correctness of the data model. A check for prior-data conflict can be done separately. Hence if there is no prior-data conflict, and if we are using a posterior predictive p-value to examine failures of the data model, such failures should not result in the need for extrapolation in the regression model and hence our methodology should not fail for that reason, at least for suitable choices of the ABC summary statistics. In the above we discussed the use of the MLE as the summary statistic, but the same reasoning applies with any summary statistic that gives an informative estimate of the model parameters.

Another natural question to ask is: when does the regression approximated calibrated p-value accurately approximate the true calibrated p-value? Short of actually computing both we do not believe we can give any general guarantees about accuracy for the regression approximation. However, one can say that if the uncalibrated p-value (easily

computed from a Markov chain Monte Carlo (MCMC) analysis say for the observed data) agrees with the uncalibrated p-value from the ABC approach, then that would give some degree of confidence in the ABC calibrated p-value. Of course it's also possible that the calibrated p-values may be close even when the uncalibrated p-values are not. We have argued that for model checking computations high accuracy is not needed; if the uncalibrated p-value does not indicate surprise but the regression calibration indicates a possible need to reformulate the model, then we can simply fit an alternative model and consider more formal methods of model choice for comparing the new model with the old. The computational expense involved in doing this is almost certainly much less than performing the calibration procedure for the original model using MCMC. So calibration with ABC followed by fitting an alternative model is something that might be feasible and useful when the full calibration with MCMC is not possible. Furthermore, if the discrepancy for the check has been constructed to examine the adequacy of a certain aspect of the model specification the failure of the check may suggest a direction in which the current model should be expanded. Clearly the exact calibrated posterior predictive p-value would be preferable if it could be obtained, but if not the regression approximated version would be useful under some circumstances and might lead to the fitting of an improved alternative model.

5.4 Capture–recapture example

We consider a capture–recapture dataset on the European Dipper (*Cinclus cinclus*) collected by Marzolin (1988). The data is collected over six years and is shown in Table 1. Lebreton et al. (1992) apply various Cormack–Jolly–Seber (CJS) survivor models to this data. The most general model can be described as follows. Let i and j be two indices related to a particular year relative to the year the experiment was initiated. For example, since the experiment began in 1981, here $i = 1$ denotes 1981 and $j = 2$ denotes 1982 etc. Let ϕ_i be the probability that an animal survives from year i to $i+1$ for $i = 1, \ldots, 6$, p_j be the probability that an animal is captured in the jth year for $j = 2, \ldots, 7$ and $\tilde{p}_j = 1 - p_j$. A data point y_{ij} consists of the number of animals caught in year j out of the number of animals released in year i, R_i. The number of animals that are never caught during the experimental study that are released in year i is thus given by $r_i = R_i - \sum_{j=i+1}^{7} y_{ij}$.

Each row in Table 1 can be assumed to be an independent draw from a multinomial distribution with the number of trials given by R_i and probabilities as shown in Table 1 together with the probability χ_i of never being captured if released in year i. Thus χ_i is simply one minus the sum of the probabilities in the row (and is therefore a function of the model parameters). Hence the likelihood of the data is given by

$$p(y|\theta) \propto \prod_{i=1}^{6} \chi_i^{r_i} \prod_{j=i+1}^{7} \left(\phi_i p_j \prod_{k=i+1}^{j-1} \phi_k \tilde{p}_k \right)^{y_{ij}},$$

where θ is the vector of model parameters and $y = \{y_{i,j} | i = 1, \ldots, 6, j = 2, \ldots, 7\}$. We consider two models. The first is the previously described model with 12 parameters (referred to as the so-called T/T model, $\theta = (\phi_1, \ldots, \phi_6, p_2, \ldots, p_7)$) while the second

i	R_i	1982	1983	1984	1985	1986	1987
1	22	11 $\phi_1 p_2$	2 $\phi_1 p_3 \prod_{j=2}^{2} \phi_j \tilde{p}_j$	0 $\phi_1 p_4 \prod_{j=2}^{3} \phi_j \tilde{p}_j$	0 $\phi_1 p_5 \prod_{j=2}^{4} \phi_j \tilde{p}_j$	0 $\phi_1 p_6 \prod_{j=2}^{5} \phi_j \tilde{p}_j$	0 $\phi_1 p_7 \prod_{j=2}^{6} \phi_j \tilde{p}_j$
2	60		24 $\phi_2 p_3$	1 $\phi_2 p_4 \prod_{j=3}^{3} \phi_j \tilde{p}_j$	0 $\phi_2 p_5 \prod_{j=3}^{4} \phi_j \tilde{p}_j$	0 $\phi_2 p_6 \prod_{j=3}^{5} \phi_j \tilde{p}_j$	0 $\phi_2 p_7 \prod_{j=3}^{6} \phi_j \tilde{p}_j$
3	78			34 $\phi_3 p_4$	2 $\phi_3 p_5 \prod_{j=4}^{4} \phi_j \tilde{p}_j$	0 $\phi_3 p_6 \prod_{j=4}^{5} \phi_j \tilde{p}_j$	0 $\phi_3 p_7 \prod_{j=4}^{6} \phi_j \tilde{p}_j$
4	80				45 $\phi_4 p_5$	1 $\phi_4 p_6 \prod_{j=5}^{5} \phi_j \tilde{p}_j$	2 $\phi_4 p_7 \prod_{j=5}^{6} \phi_j \tilde{p}_j$
5	88					51 $\phi_5 p_6$	0 $\phi_5 p_7 \prod_{j=6}^{6} \phi_j \tilde{p}_j$
6	98						52 $\phi_6 p_7$

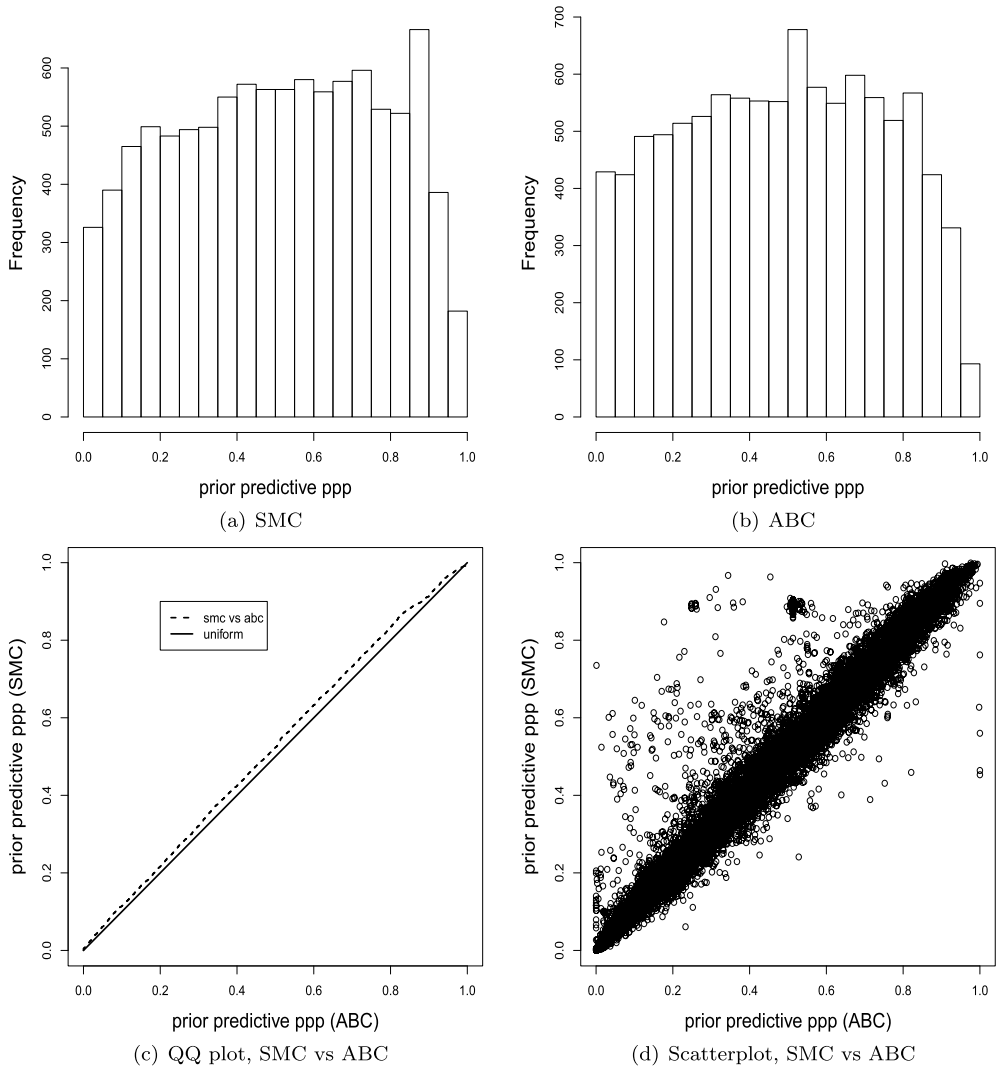
Table 1: Capture–recapture data. Also shown are the probabilities (under the general CJS model) that an animal contributes to each cell in the table if released in a certain year and caught in a subsequent year.

model is constrained, with $\phi_i = \phi$ and $p_j = p$ (the so-called C/C model, $\theta = (\phi, p)$). Both of these models are considered by Hjort et al. (2006), who estimate both the posterior predictive p-value (ppp) and calibrated posterior predictive p-value (cppp) based on the discrepancy

$$D(y, \theta) = \sum_{i,j} (y_{ij}^{1/2} - e_{ij}^{1/2})^2.$$

Here e_{ij} is the expected number of captured animals for the i, jth cell, which involves the model parameters θ and the release numbers R_i. Here we repeat the analysis of Hjort et al. (2006) but consider ABC to speed up the calculations. All the full posterior computations we use here for comparison with the ABC results are based on a sequential Monte Carlo (SMC) algorithm, which is built upon the base algorithm of Chopin (2002) (see also Del Moral et al. (2006) for a reference on SMC for static models). Details of the SMC algorithm are given in the supplementary materials (Nott et al., 2016). We chose SMC here as the algorithm that we use does not require any tuning, and is thus suitable for analysing datasets simulated from a prior predictive distribution, which might lead to very different posteriors.

For the C/C model, independent and uniform priors over the unit interval are placed over ϕ and p. SMC is run using $N = 10{,}000$ particles, which produces a ppp of roughly 0.061 (consistent with Hjort et al. (2006)). The cppp is estimated using the double simulation approach with 10,000 simulated datasets from the prior predictive distribution.



(a) SMC (b) ABC

(c) QQ plot, SMC vs ABC (d) Scatterplot, SMC vs ABC

Figure 3: Prior predictive ppp distributions for the C/C model based on (a) SMC, (b) ABC and (c), (d) QQ-plot and scatterplot of SMC versus ABC.

Using the SMC approach with $N = 1{,}000$ particles on each of these datasets results in a prior predictive distribution of ppps as shown in Figure 3(a). Following this process we obtain a cppp of 0.043. Hjort et al. (2006) obtain a value of 0.022 using 500 prior predictive datasets in their calculations. Overall about 32 hours of computation was required.

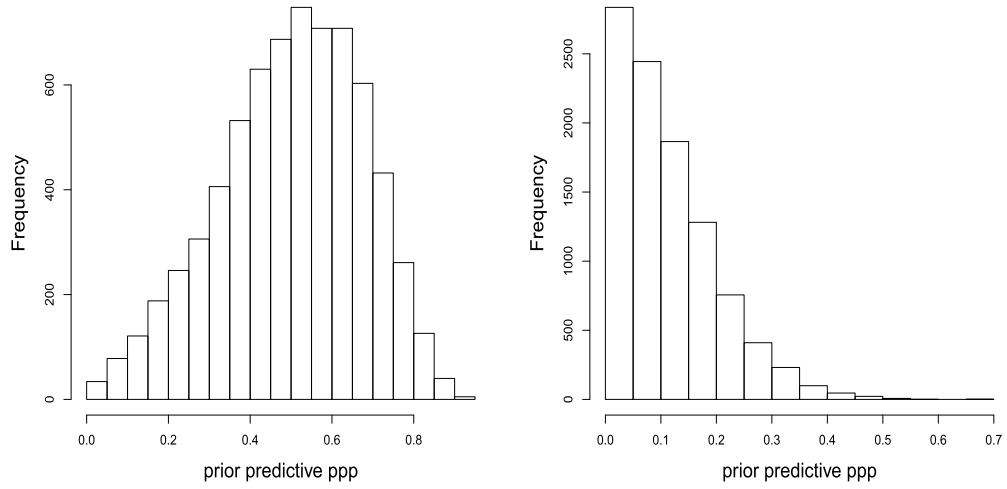
The ABC approach is based on 100,000 draws from the prior. For each dataset, 1,000 nearest neighbours are kept in implementing the ABC regression algorithm. Note that

this is the same number of particles used in the SMC sampler. Neural network regression is then applied in order to refine the distribution of the discrepancy values. The default implementation of the `abc R` package (Csilléry et al., 2012) is used. The summary statistics chosen in the ABC approach are the maximum likelihood estimates, which are asymptotically sufficient and low dimensional and seem to yield a good ABC approximation, in the sense that the ABC posterior distribution is similar to that obtained using non-ABC methods without summary statistics. Using this process we obtain an unadjusted ppp of 0.057 (0.061 for the likelihood-based SMC approach). For the double simulation, we use exactly the same 10,000 prior predictive datasets as used in the SMC approach. With the same ABC settings used to obtain the ppp, the estimated cppp was 0.047 (compared to 0.043 obtained using SMC). Figure 3(b) shows the distribution of prior predictive ppps obtained and Figure 3(d) shows a good correspondence between the ppp values obtained under SMC and ABC for the 10,000 datasets. We do notice, however, that there is some piling up of p-values near zero and one for the ABC approach. The ABC approach is roughly 4–5 times faster than the SMC method, requiring 7 hours of computation. The ABC approach is much faster still if a simpler regression adjustment approach is used (say linear rather than the neural network method) but the neural network adjustment is most effective in this case: a similar plot to Figure 3(b) using local linear regression adjustment showed a much poorer agreement between the ABC and SMC answers for this problem.

For the T/T model, using the SMC approach with $N = 10{,}000$, we obtain an unadjusted ppp of 0.070 (compared to 0.075 obtained in Hjort et al. (2006)). We simulate 10,000 datasets from the prior predictive distribution for calibration purposes. Due to the high dimensionality of the parameter, the SMC approach is rather expensive, and we stop the process after 6859 prior predictive datasets (the distribution of prior predictive ppps is shown in Figure 4(a)). The SMC approach is roughly 10 times slower than the ABC approach. Using $N = 1{,}000$, we obtain a cppp of 0.008 (compared to the value 0.002 obtained by Hjort et al. (2006)).

For this model the sufficient statistics are given by the row and column sums of Table 1. Thus we use these as the summary statistics in our ABC approach. This time we use 1,000,000 parameters from the prior for ABC rejection and keep 1,000 samples in the localization step of the ABC algorithm. Again we refine using a neural network with the default implementation in the `abc R` package. The estimate of the unadjusted ppp from the ABC approach is 0.021, which is quite different to the estimated 0.070 obtained from SMC. This is not surprising given that ABC approximations typically deteriorate in higher dimensional problems, due to the difficulty in generating simulated data very close to the observed. The distribution of prior predictive ppps is also markedly different to that generated from SMC (see Figure 4(b)). Under this distribution of ppps, the cppp was estimated at 0.115. The lack of agreement between the cppp and ABC based approximation to it is not necessarily a problem for the reasons discussed in Section 5.3: the ABC approach represents a different kind of check with a sensible interpretation.

Finally we discuss the Monte Carlo (MC) variability of the results. For both the C/C and T/T examples, we replicated the entire ABC calibration process 10 times,



(a) SMC (b) ABC

Figure 4: Prior predictive ppp distributions for the T/T model based on (a) SMC and (b) ABC.

which includes generating a new set of prior predictive simulations each time. We find that there is a reasonable amount of MC variability in estimating individual ppps. Our experience suggests that this variability is due to the neural network regression, which is a stochastic procedure because of the random initialization used in the optimization of the model parameters and the existence of multiple modes in the relevant objective function. Interestingly, though, despite the variability in individual ppps, we find that the prior distribution of ppps, the distribution of values that we use for calibration, is very similar across the 10 runs. That is, variability inherent in the neural network, while it does affect the variability of individual ppp values, does not really affect the overall distribution of prior predictive ppp values. For the C/C model, if we naively repeat the whole calibration process 10 times we obtain a mean estimated cppp of 0.047 with a standard deviation of 0.016. However, if we fix the unadjusted ppp for the observed data and compare it with the 10 distributions of prior predictive ppps we obtain a mean estimated cppp with a standard error of roughly 0.002 so that the MC variability is dominated by the between-run variability of the unadjusted ppp estimates for the observed data. Therefore, our suggestion is to replicate the neural network process several times and take the average in order to obtain an unadjusted ppp for the observed data with a low Monte Carlo error, but not to do this in the estimation of the prior distribution of the ppps used for calibration. The unadjusted ppp for the C/C and the T/T models presented above is obtained by averaging the results of 10 and 20 neural network regression fits, respectively; since such averaging is not done for the distribution of prior predictive ppps, this only results in a small amount of extra computation.

6 Discussion

We have explored the potential for using regression ABC methods in calculation of Bayesian predictive p-values in some cases where high accuracy of the computations is not required. This is a new application of ABC methods as far as we are aware. The methods of weakly informative prior choice that we have suggested are also easily applied to models which themselves require an ABC treatment for inference – for these models it is difficult to derive the usual weakly informative prior choices as those may require a knowledge of the likelihood, such as the ability to compute the Fisher information. In this work we have used the non-invariant conflict p-value of Evans and Moshonov (2006) rather than its invariant counterpart proposed in Evans and Jang (2010). It would be interesting to see if computation of the invariant p-value could be routinely attempted using similar methods to the ones we have developed.

Supplementary Material

Supplementary material for "Approximation of Bayesian predictive p-values with regression ABC" (DOI: 10.1214/16-BA1033SUPP; .pdf).

References

Bayarri, M. J. and Berger, J. O. (2000). "P values for composite null models (with discussion)." *Journal of the American Statistical Association*, 95: 1127–1142. MR1804239. doi: https://doi.org/10.2307/2669749. 72

Bayarri, M. J. and Castellanos, M. E. (2007). "Bayesian checking of the second levels of hierarchical models." *Statistical Science*, 22: 322–343. MR2416808. doi: https://doi.org/10.1214/07-STS235. 72

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). "Approximate Bayesian computation in population genetics." *Genetics*, 162: 2025–2035. 60, 63, 67, 69

Blum, M. G. B. (2010). "Approximate Bayesian computation: a nonparametric perspective." *Journal of the American Statistical Association*, 105(491): 1178–1187. MR2752613. doi: https://doi.org/10.1198/jasa.2010.tm09448. 60

Blum, M. G. B. and François, O. (2010). "Non-linear regression models for approximate Bayesian computation." *Statistics and Computing*, 20: 63–75. MR2578077. doi: https://doi.org/10.1007/s11222-009-9116-0. 60, 63, 64

Bowman, A. W. and Azzalini, A. (2014). "R package `sm`: nonparametric smoothing methods (version 2.2–5.4)." http://azzalini.stat.unipd.it/Book_sm 69

Box, G. E. P. (1980). "Sampling and Bayes' inference in scientific modelling and robustness (with discussion)." *Journal of the Royal Statistical Society, Series A*, 143: 383–430. MR0603745. doi: https://doi.org/10.2307/2982063. 59, 61

Chopin, N. (2002). "A sequential particle filter method for static models." *Biometrika*,

89(3): 539–551. MR1929161. doi: https://doi.org/10.1093/biomet/89.3.539. 76

Csilléry, K., François, O., and Blum, M. G. B. (2012). "ABC: an R package for approximate Bayesian computation (ABC)." *Methods in Ecology and Evolution*, 3: 475–479. 67, 69, 78

Dahl, F. A., Gåsemyr, J., and Natvig, B. (2007). "A robust conflict measure of inconsistencies in Bayesian hierarchical models." *Scandinavian Journal of Statistics*, 34: 816–828. MR2396940. doi: https://doi.org/10.1111/j.1467-9469.2007.00560.x. 61

Del Moral, P., Doucet, A., and Jasra, A. (2006). "Sequential Monte Carlo samplers." *Journal of the Royal Statistical Society, Series B*, 68(3): 411–436. MR2278333. doi: https://doi.org/10.1111/j.1467-9868.2006.00553.x. 76

Evans, M. and Jang, G. H. (2010). "Invariant P-values for model checking." *The Annals of Statistics*, 38: 512–525. MR2589329. doi: https://doi.org/10.1214/09-AOS727. 69, 80

Evans, M. and Jang, G. H. (2011). "Weak informativity and the information in one prior relative to another." *Statistical Science*, 26: 423–439. MR2917964. doi: https://doi.org/10.1214/11-STS357. 60, 61, 64, 65, 67, 68, 69, 70, 71

Evans, M. and Moshonov, H. (2006). "Checking for prior-data conflict." *Bayesian Analysis*, 1: 893–914. MR2282210. doi: https://doi.org/10.1016/j.spl.2011.02.025. 61, 62, 64, 72, 73, 80

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1: 1–19. MR2221284. 60, 64

Gelman, A. (2013). "Two simple examples for understanding posterior p-values whose distributions are far from unform." *Electronic Journal of Statistics*, 7: 2595–2602. MR3121624. doi: https://doi.org/10.1214/13-EJS854. 72

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). "A weakly informative default prior distribution for logistic and other regression models." *The Annals of Applied Statistics*, 2: 1360–1383. MR2655663. doi: https://doi.org/10.1214/08-AOAS191. 67, 68

Gelman, A., Meng, X.-L., and Stern, H. (1996). "Posterior predictive assessment of model fitness via realized discrepancies." *Statistica Sinica*, 6: 733–807. MR1422404. 59, 62

Gåsemyr, J. and Natvig, B. (2009). "Extensions of a conflict measure of inconsistencies in Bayesian hierarchical models." *Scandinavian Journal of Statistics*, 36: 822–838. MR2573310. doi: https://doi.org/10.1111/j.1467-9469.2009.00659.x. 61

Guttman, I. (1967). "The use of the concept of a future observation in goodness-of-fit problems." *Journal of the Royal Statistical Society, Series B*, 29: 83–100. MR0216699. 59

Hjort, N. L., Dahl, F. A., and Steinbakk, G. H. (2006). "Post-processing posterior predictive p-values." *Journal of the American Statistical Association*, 101: 1157–1174. MR2324154. doi: https://doi.org/10.1198/016214505000001393. 60, 62, 71, 76, 77, 78

Lebreton, J.-D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). "Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies." *Ecological Monographs*, 62(1): 67–118. 75

Marin, J.-M., Pudlo, P., and Robert, C. P. (2015). "Likelihood-free Model Choice." arXiv:1503.07689. 71

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. (2011). "Approximate Bayesian computational methods." *Statistics and Computing*, 21: 289–291. MR2992292. doi: https://doi.org/10.1007/s11222-011-9288-2. 62, 63

Marshall, E. C. and Spiegelhalter, D. J. (2007). "Identifying outliers in Bayesian hierarchical models: a simulation-based approach." *Bayesian Analysis*, 2: 409–444. MR2312289. doi: https://doi.org/10.1214/07-BA218. 61

Marzolin, G. (1988). "Polygynie du Cincle plongeur (Cinclus cinclus) dans les côtes de Lorraine." *Oiseau et la Revue Francaise d'Ornithologie*, 58(4): 277–286. 75

McVinish, R., Mengersen, K., Nur, D., Rousseau, J., and Guihenneuc-Jouyaux, C. (2013). "Recentered importance sampling with applications to Bayesian model validation." *Journal of Computational and Graphical Statistics*, 22: 215–228. MR3044331. doi: https://doi.org/10.1080/10618600.2012.681239. 71

Moores, M. T., Drovandi, C. C., Mengersen, K., and Robert, C. P. (2015). "Preprocessing for approximate Bayesian computation in image analysis." *Statistics and Computing*, 25: 23–33. MR3304900. doi: https://doi.org/10.1007/s11222-014-9525-6. 64

Nott, D. J., Drovandi, C. C., Mengersen, K., and Evans, M. (2016). "Supplementary material for "Approximation of Bayesian predictive p-values with regression ABC"." *Bayesian Analysis*. doi: https://doi.org/10.1214/16-BA1033SUPP. 76

O'Hagan, A. (2003). "HSS model criticism (with discussion)." In Green, P. J., Hjort, N. L., and Richardson, S. T. (eds.), *Highly Structured Stochastic Systems*, 423–453. Oxford University Press. MR2082403. 61

Presanis, A. M., Ohlssen, D., Spiegelhalter, D. J., and Angelis, D. D. (2013). "Conflict diagnostics in directed acyclic graphs, with applications in Bayesian evidence synthesis." *Statistical Science*, 28: 376–397. MR3135538. doi: https://doi.org/10.1214/13-STS426. 61

Racine, A., Grieve, A. P., Flühler, H., and Smith, A. F. M. (1986). "Bayesian methods in practice: experiences in the pharmaceutical industry." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 35: 93–150. MR0868007. doi: https://doi.org/10.2307/2347264. 68

Robins, J. M., van der Vaart, A., and Ventura, V. (2000). "Asymptotic distribution of p-values in composite null models." *Journal of the American Statistical Association*, 95: 1143–1156. MR1804240. doi: https://doi.org/10.2307/2669750. 60, 72

Rubin, D. B. (1984). "Bayesianly justifiable and relevant frequency calculations for the applied statistician." *Annals of Statistics*, 12: 1151–1172. MR0760681. doi: https://doi.org/10.1214/aos/1176346785. 59

Scheel, I., Green, P. J., and Rougier, J. C. (2011). "A graphical diagnostic for identifying influential model choices in Bayesian hierarchical models." *Scandinavian Journal of Statistics*, 38(3): 529–550. MR2833845. doi: https://doi.org/10.1111/j.1467-9469.2010.00717.x. 61

Steinbakk, G. H. and Storvik, G. O. (2009). "Posterior predictive p-values in Bayesian hierarchical models." *Scandinavian Journal of Statistics*, 36: 320–336. MR2528987. doi: https://doi.org/10.1111/j.1467-9469.2008.00630.x. 60

Wood, S. N. (2010). "Statistical inference for noisy nonlinear ecological dynamic systems." *Nature*, 466(7310): 1102–1104. 64

Acknowledgments

David Nott was supported by a Singapore Ministry of Education Academic Research Fund Tier 2 grant (R-155-000-143-112). Christopher Drovandi was supported by an Australian Research Council's Discovery Early Career Researcher Award funding scheme DE160100741 and by a Queensland University of Technology Early Career Travel Fellowship. Kerrie Mengersen was supported by the Australian Research Council. Kerrie Mengersen and Christopher Drovandi acknowledge the hospitality of National University of Singapore in 2013 and 2014, in order to collaboratively progress this research. The authors thank the Editor, Associate Editor and two referees for comments that greatly improved the presentation of the paper.