Bayesian Analysis (2024) **TBA**, Number TBA, pp. 1–29

Cutting Feedback and Modularized Analyses in Generalized Bayesian Inference

David T. Frazier* and David J Nott†

Abstract. This work considers Bayesian inference under misspecification for complex statistical models comprised of simpler submodels, referred to as modules, that are coupled together. Such "multi-modular" models often arise when combining information from different data sources where there is a module for each data source. When some of the modules are misspecified, the challenges of Bayesian inference under misspecification can sometimes be addressed by using "cutting feedback" methods, which modify conventional Bayesian inference by limiting the influence of unreliable modules. Here we investigate cutting feedback methods in the context of generalized posterior distributions built from loss functions. We make three main contributions. First, we describe how cutting feedback methods can be defined in the generalized Bayes setting, and discuss the appropriate scaling of the loss functions in this context. Second, we derive a novel type of conditional Laplace approximation that accurately describes the behavior of the posterior for a given module's parameters when we condition on parameters in other modules. Third, we leverage this novel result to provide several convenient diagnostics for Bayesian modular inference, which we then apply to examples in the literature on cut model inference.

Keywords: cutting feedback, model misspecification, modularization, semi-modular inference, generalized Bayesian inference.

1 Introduction

Complex statistical models are sometimes composed of smaller sub-models, which we call modules, that are interconnected. This modular structure is common when integrating information from multiple data sources, where each data source is associated with a separate sub-model. When a model with a modular structure is correctly specified, Bayesian inference has desirable properties, regardless of the number or complexity of the modules. However, when there is misspecification, conventional Bayesian inference may need to be modified. This paper explores new forms of a method called "cutting feedback" for modified Bayesian inference under misspecification.

It is well-known that misspecification of an assumed model compromises the use and interpretation of Bayesian inference; see, e.g., Grünwald (2012) for examples. Nonetheless, when dealing with a multi-modular model, a researcher may suspect that only some modules are grossly misspecified. In such cases, modified Bayesian analysis methods

*Department of Econometrics and Business Statistics, Monash University, Australia, david.frazier@monash.edu

†Department of Statistics and Applied Probability, National University of Singapore, Singapore, standj@nus.edu.sg

 https://doi.org/10.1214/24-BA1448

like cutting feedback, which is the focus of this paper, can limit the influence of unreliable modules, thereby preserving the validity of inferences on parameters in correctly specified modules. This can make model criticism easier and ensure that estimates of parameters in the misspecified modules retain a useful interpretation (Liu et al., 2009). For discussion on the wide-ranging applications of cutting feedback and modularized Bayesian inference methods, see Jacob et al. (2017) and Liu et al. (2009), with the latter paper focusing on applications in the analysis of computer models.

The current literature on cutting feedback mainly focuses on fully specified parametric models. However, if a parametric model is misspecified, researchers can still produce useful Bayesian inferences by using a posterior based on a loss function that captures the features of the data that are most important. Such generalized Bayesian inference methods (see, for example, Bissiri et al., 2016), have become increasingly popular in statistical inference. They recover conventional Bayesian inference as a special case when the loss function used in their construction is the negative log likelihood. This paper combines the use of cutting feedback methods with generalized Bayesian inference, resulting in an attractive approach to modular Bayesian inference. Our framework allows a targeted loss function to be used for modules which are misspecified, instead of relying on the negative log likelihood function. Meanwhile, we can continue to use the negative log likelihood function as the loss for modules that are well specified. The generalized Bayes perspective on modular inference is useful in model improvement. Starting with a flawed parametric model specification, we can replace the negative log likelihood for suspect modules with other loss functions to see whether this resolves any incompatibility between the "cut posterior" produced by cutting feedback methods and full posterior inferences.

Our work makes three main contributions to the literature on generalized Bayesian inference and cutting feedback. Firstly, we describe how to define cutting feedback in the generalized Bayesian setting, and discuss how to appropriately scale loss functions for different modules to each other and to the prior. Secondly, we derive a novel large sample result that allows us to express the posterior for the parameters of a given module conditional on the parameters of the remaining modules. In contrast, the only existing result on the large sample behaviour of cut posteriors of which we are aware (Pompe and Jacob, 2021) presents a joint analysis of the cut posterior. As we argue in Section 2.3, the conventional normal approximation to the joint cut posterior provides only limited insight into propagation of uncertainty in cutting feedback, because conditioning on a subset of variables in a multivariate normal distribution results in a conditional covariance matrix that doesn't depend on the conditioning variables. In contrast, our results provide convenient normal approximations for conditional posterior distributions where covariance matrices change with the values for the conditioning variables, giving useful insights into uncertainty propagation in cut posteriors.

Finally, we use this large sample approximation to develop easily computable diagnostics for understanding the coupling of the modules, and to conduct different forms of semi-modular inference (SMI) as in Carmona and Nicholls (2020). While estimation of the SMI (and cut) posterior generally requires a computationally burdensome nested Markov chain Monte Carlo (MCMC) method, we show how our large sample

results can be used as fast approximations to the SMI and cut posteriors that ease this computational burden: these approximations can be used directly in computationally onerous settings, or as proposals in MCMC or importance sampling. We illustrate our methodology in two benchmark examples found in the literature on cutting feedback.

Notation. Here we define notation used in the remainder of the paper. The term $\|\cdot\|$ denotes the Euclidean norm, while $|\cdot|$ denotes the absolute value function. C denotes an arbitrary positive constant that can change from line-to-line. For $x = (x_1^\top, x_2^\top)^\top \in \mathbb{R}^d$ and a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, we let $\nabla_x f(x)$ denote the gradient of $f(x)$ wrt x, and $\nabla^2_{xx} f(x)$ the Hessian. Let $N\{\mu, \Sigma\}$ denote the normal distribution with mean μ and covariance matrix Σ, with $N\{x; \mu, \Sigma\}$ the corresponding normal density at the point x. For $\mathcal{D}$ some known distribution, and $x = (x_1^\top, x_2^\top)^\top \in \mathbb{R}^d$ a d-dimensional random variable, the notation $x \sim \mathcal{D}$ signifies that the law of x is $\mathcal{D}$, while $x_1|x_2 \sim \mathcal{D}$ signifies that the conditional law of x_1 given x_2 is $\mathcal{D}$. The measure $P_0^{(n)}$ denotes the true unknown probability measure generating the data, and $\Rightarrow$ denotes weak convergence (under $P_0^{(n)}$).

2 Motivation and Framework

Figure 1 is a graphical representation of the model we will use throughout this paper. The data come from two data sources, denoted $\boldsymbol{z}$ and $\boldsymbol{w}$, and we denote the complete data by $\boldsymbol{y} = (\boldsymbol{z}^\top, \boldsymbol{w}^\top)^\top$. The data source for each observation is known; we do not consider problems such as mixture modelling for which there is some unobserved allocation of observations to mixture components representing the different data sources. We write $\boldsymbol{z} = (z_1, \ldots, z_{n_1})^\top$, $z_i \in \mathcal{Z}$, $\boldsymbol{w} = (w_1, \ldots, w_{n_2})^\top$, $w_i \in \mathcal{W}$, and $n = n_1 + n_2$ for the number of observations in $\boldsymbol{y}$.

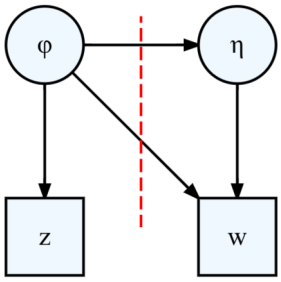

Figure 1: Graphical structure of the two-module system. The red dashed line indicates the cut.

There are parameters $\theta = (\varphi^\top, \eta^\top)^\top$, where $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$, $\varphi \in \Phi \subseteq \mathbb{R}^{d_\varphi}$, $\eta \in \mathcal{E} \subseteq \mathbb{R}^{d_\eta}$, and we write $d_\theta = d_\varphi + d_\eta$ for the dimension of θ. We consider Bayesian inference, with a prior density $\pi(\theta) = \pi(\varphi)\pi(\eta|\varphi)$. In Figure 1, the parameter φ is shared in the models for $\boldsymbol{z}$ and $\boldsymbol{w}$, but η is specific to the model for $\boldsymbol{w}$. There is a likelihood for $\boldsymbol{z}$, $p(\boldsymbol{z}|\varphi)$, that we believe is well specified and which we trust. There is a likelihood for $\boldsymbol{w}$, $p(\boldsymbol{w}|\varphi, \eta)$ which we suspect is misspecified and which we do not trust. This means that

Bayesian inference for φ using $\pi(\varphi|\boldsymbol{z})$ is trusted, but Bayesian inference using $\pi(\varphi|\boldsymbol{y})$ is not.

Because of our lack of trust in $\pi(\varphi|\boldsymbol{y})$, we may wish to use the "cut" joint posterior density

$$\pi_{\text{cut}}(\varphi, \eta|\boldsymbol{z}, \boldsymbol{w}) := \pi(\varphi|\boldsymbol{z})\pi(\eta|\varphi, \boldsymbol{y}), \tag{2.1}$$

instead of the conventional posterior density. In (2.1) marginal inferences for φ are performed using $\pi(\varphi|\boldsymbol{z})$, but conditional inference for η given φ is the same as for the conventional joint posterior density, based on $\pi(\eta|\varphi, \boldsymbol{y})$. Using (2.1) rather than the conventional posterior is referred to as "cutting feedback", and this approach was initially developed in specific fields of application as a way of dealing with potential model misspecification of the suspect likelihood $p(\boldsymbol{w}|\varphi, \eta)$.

Another motivation for cut posteriors is given in Lemma 1 of Yu et al. (2023), which proved that the cut posterior density is the closest approximation of the conventional Bayesian posterior density, in Kullback-Leibler divergence, subject to the constraint that the φ-marginal is the "trusted" $p(\varphi|\boldsymbol{z})$. Further, there is also a foundational justification for cutting feedback in terms of a generalization of Bayesian conditioning referred to as Jeffrey conditionalization (Jeffrey, 1965) originating in the philosophy of science literature. A recent discussion of the connections with cutting feedback accessible to a statistical audience is given in Hahn and Herren (2023, Section 1.3).

Cutting feedback is a widely used Bayesian modular inference method. Modular inference splits a complex model into smaller submodels called modules that are coupled together. If we are concerned about model misspecification affecting only some modules, it may be sensible to limit the interaction between the modules when making inference. Modules are defined here to be subsets of the terms used in specifying a joint Bayesian model. For the model of Figure 1, we consider two modules. The first is $\{\pi(\varphi), p(\boldsymbol{z}|\varphi)\}$ (hereafter module 1) represented by nodes to the left of the red line in Figure 1. The second is $\{\pi(\eta|\varphi), p(\boldsymbol{w}|\varphi, \eta)\}$ (hereafter module 2) and these terms are represented by the nodes to the right of the red line. Inference from the cut posterior is modular, in the sense that marginal inference for φ is based on module 1 only, $\pi(\theta|\boldsymbol{z}) \propto \pi(\varphi)p(\boldsymbol{z}|\varphi)$, and conditional inferences about η given φ are only based on module 2, $\pi(\eta|\varphi, y) \propto \pi(\eta|\varphi)p(\boldsymbol{w}|\varphi, \eta)$. For an explicit definition of modules and cutting feedback for a general model taking the form of a directed acyclic graph (DAG), we refer the reader to Liu and Goudie (2022a). We describe in the supplementary material (Frazier and Nott, 2024) how the model of Figure 1 and our definitions of the modules relate to the definitions in Liu and Goudie (2022a).

As further motivation for the use of cutting feedback, we introduce a simple example discussed in more detail later in Section 4.1. This example was given in Plummer (2015), and is based on a real epidemiological study (Maucort-Boulch et al., 2008). The model here consists of two modules. Module 1 incorporates survey data from 13 countries on high-risk human papillomavirus (HPV) prevalence for women in a certain age group. Denote by z_i the number of women with high-risk HPV in country i in a survey of N_i individuals, $i = 1, \ldots, 13$, and assume that $z_i \sim \text{Binomial}(N_i, \varphi_i)$, where $\varphi_i \in [0, 1]$ is a

country-specific prevalence probability. The parameters φ_i are assumed independent in their prior, with $\varphi_i \sim U[0,1]$. Write $\varphi = (\varphi_1, \ldots, \varphi_{13})^\top$.

Module 2 incorporates cervical cancer incidence data $\boldsymbol{w}$, with w_i the number of cervical cancer cases in T_i woman years of follow-up in country i, $i = 1, \ldots, 13$. The relationship between cervical cancer incidence and HPV prevalence is described by a Poisson regression model, $w_i \sim \text{Poisson}(T_i\rho_i)$, where $\log \rho_i = \eta_1 + \eta_2\varphi_i$. For these data the Poisson regression model can be criticized on the grounds of an incorrect specification of the mean model or link function, and a failure to account for overdispersion. Because φ_i is appearing as a covariate in the Poisson regression, inference about φ_i is influenced by the misspecification in the second module. Estimation of φ adapts to the misspecification, distorting inference about these parameters, which also results in uninterpretable inference about the regression parameters η used to summarize the relationship between HPV prevalence and the rate of cancer incidence. The main interest of the analysis lies in understanding this relationship. The top left graph in Figure 2 in Section 4.1 shows posterior samples for η from the conventional posterior (blue) and the cut posterior (green). As we can see, the interpretation of the main parameter of interest changes markedly when cut inferences are used to account for the misspecification in the second module.

2.1 Related Literature and Motivation

The "two module" system of Figure 1 was introduced by Plummer (2015), with the motivation of clarifying previous implicit definitions of cutting feedback methods based on modifying MCMC algorithms. One implementation of cutting feedback implicitly is through the `cut` function of the WinBUGS and OpenBUGS software packages (Lunn et al., 2009). If a Bayesian model is defined through a DAG, and a Gibbs sampler is considered for sampling the posterior distribution using the DAG parameter nodes as blocks, then "cuts" can be defined for some links of the graph. Each cut corresponds to leaving out a certain term in the joint model when forming the full conditional posterior density for one of the parameter nodes. Once modified full conditional distributions have been constructed, a modified Gibbs sampler iteratively samples from these, and the cut posterior distribution is defined as the stationary distribution of the resulting Markov chain. See Lunn et al. (2009) or Plummer (2015) for a more detailed description. The implicit definition corresponds with the explicit definition (2.1) for the model of Figure 1, if the cut involves leaving out $p(\boldsymbol{w}|\varphi, \eta)$ when forming the full conditional density for φ.

Lunn et al. (2009) note that the modified full conditional distributions are not the full conditional distributions of any well-defined joint distribution but argue that the use of such inconsistent conditional distributions can be sensible. If modified Gibbs steps are replaced by Metropolis-within-Gibbs updates in the sampling process, Plummer (2015) observed that the stationary distribution of the Markov chain can depend on the proposal used, and went on to define the "two-module" system of Figure 1 where an explicit definition of the cut posterior distribution can be given, clarifying some aspects of the implicit cut approach. This two module system is general enough for many applications of Bayesian modular inference in which there might be one suspect model component of particular concern. This two module system is also fundamental

to the recent work of Liu and Goudie (2022a) where multi-modular systems and cut posteriors are defined generally. Liu and Goudie (2022a) define two module systems first, based on a partitioning of the observables into two parts, and then consider recursively splitting existing modules into two in order to define more complex multi-modular representations. In what follows, we will focus our discussion on cutting feedback in two-module systems, given their usefulness in applications and their role in defining multi-modular models with more than two modules. We define modules and cutting feedback precisely in the context of this two module system, and refer the interested reader to Liu and Goudie (2022a) and the supplementary material for a more general discussion.

The rest of the paper is concerned with the two-module system described in Figure 1, and generalized Bayesian extensions where the likelihood terms are replaced by loss-likelihoods. While the two-module system may seem quite specific, it has found many uses in empirical settings and in Section 2.1 of the supplementary appendix, we briefly discuss several applications of the two-modular system in the recent literature.

2.2 Generalized Posteriors

In the system of Figure 1, we will discuss two methods that can guard against compromised inferences on φ due to potential misspecification of the second module. The first method involves using a loss function rather than a parametric model to capture the important features of the data for the second module; a generalized posterior is constructed based on the loss function for the parameters of interest. The second approach is to employ cutting feedback methods. Generalized Bayesian methods and cutting feedback are not used here as approximations to conventional Bayesian inference; they are alternative inferential approaches intended to address the issue of misspecification and having a sound statistical justification in their own right. Approximate methods for computation may be of interest, but this is discussed later in Section 3, based on the asymptotic results we develop there. In this article, we combine cutting feedback and generalized Bayesian updating to produce robust Bayesian inferences on θ, and we explain generalized Bayesian inferences first.

When a Bayesian model is misspecified, standard Bayesian approaches can deliver inferences that are poor or unreliable (see, e.g., Grünwald and Van Ommen, 2017 for specific examples, as well as Kleijn and van der Vaart, 2012 for general results in parametric models). Specifying full probabilistic models for complex data can be difficult, and it would be attractive if Bayesian inference could be done only for parameters of interest appearing in a loss function. Under some mild conditions on the loss, Bissiri et al. (2016) justify a Bayesian analysis in this setting in which the likelihood in the usual Bayesian update is replaced with a "loss likelihood" with a highly constrained form. The target parameter of interest is then the population minimizer of the loss.

In a standard generalized posterior analysis without modular structure, there is a parameter θ and data $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$. Since we are not in the modular setting, $\boldsymbol{y}$ does not decompose into two data sources $\boldsymbol{y} = (\boldsymbol{z}^\top, \boldsymbol{w}^\top)^\top$. The prior $\pi(\theta)$ is to be updated into a generalized posterior $\pi(\theta|\boldsymbol{y})$, where the belief update depends on $\boldsymbol{y}$ only

through a loss function $q_n(\theta) = \sum_{i=1}^{n} q(y_i;\theta)$, where $q(y_i;\theta)$ is the loss for the ith observation. A remarkable argument in Bissiri et al. (2016) specifies the form that the belief update must take, under some mild conditions. They consider the requirement of order coherence, where if the data $\boldsymbol{y}$ are split into two parts and an update is done sequentially, then the result should be the same as if a single update were done using all the data. Order coherence is enough to determine the form of the generalized posterior density, which is

$$\pi(\theta|\boldsymbol{y}) \propto \pi(\theta)\exp\left\{-\nu q_n(\theta)\right\},$$

where $\nu \geq 0$ is called the learning rate, and scales the information in the loss function appropriately relative to the information in the prior. While the generalized Bayesian update of Bissiri et al. (2016) is motivated by Bayesian notions of coherence, the choice of learning rate gives the opportunity to bring in other considerations such as information matching in the update (Holmes and Walker, 2017; Lyddon et al., 2019) or achieving good frequentist performance for estimating functionals of interest (Syring and Martin, 2018). A generalization of the arguments in Bissiri et al. (2016) relevant to the justification of parametric cutting feedback methods is discussed in Nicholls et al. (2022). Generalized Bayesian updating is also related to PAC-Bayes methods; see Alquier (2021) for an introduction.

For modular Bayesian inference, the decomposition of the statistical model into two distinct modules, containing data $\boldsymbol{z}$ and $\boldsymbol{w}$ respectively, implies that we are free to choose separate loss functions for each module. Let $\ell : \mathcal{Z} \times \Phi \to \mathbb{R}$ denote the loss function for module one, involving a parameter φ, and $m : \mathcal{W} \times \mathcal{E} \times \Phi \to \mathbb{R}$ denote the loss function for module two, involving parameters η and φ. In the following we write

$$Q_n(\theta) = L_{n_1}(\varphi) + M_{n_2}(\eta,\varphi),\ L_{n_1}(\varphi) = -\sum_{i=1}^{n_1}\ell(z_i,\varphi),\ M_{n_2}(\eta,\varphi) = -\sum_{i=1}^{n_2} m(w_i,\eta,\varphi), \tag{2.2}$$

so that $-L_{n_1}(\varphi)$ and $-M_{n_2}(\eta,\varphi)$ are the empirical loss functions for the first and second modules respectively. When the two sample sizes are equal, i.e., $n_1 = n_2$, we abuse notation and simply denote the criteria as $L_n(\varphi)$ and $M_n(\eta,\varphi)$.

Consider first a belief update of the prior density $\pi(\varphi)$ using $\boldsymbol{z}$ and the first module loss function $\ell(\cdot)$. The order coherence argument of Bissiri et al. (2016) implies that the generalized posterior density $\pi(\varphi|\boldsymbol{z})$ takes the form $\pi(\varphi|\boldsymbol{z}) \propto \pi(\varphi)\exp\{\nu L_{n_1}(\varphi)\}$, where $\nu \geq 0$ is a learning rate for the first module that needs to be chosen. If the loss function is the negative log-likelihood, and we take $\nu = 1$, this is the conventional Bayesian update.

Once $\pi(\varphi|\boldsymbol{z})$ is obtained, suppose we now take $\pi(\theta|\boldsymbol{z}) = \pi(\varphi|\boldsymbol{z})\pi(\eta|\varphi)$ as the "prior" for a Bayesian update using the information in the second module. Again following the order coherence argument of Bissiri et al. (2016), and its extensions in Nicholls et al. (2022), the generalized posterior density $\pi(\theta|\boldsymbol{z},\boldsymbol{w})$ given $\boldsymbol{z}$ and $\boldsymbol{w}$ takes the form

$$\begin{aligned}\pi(\theta|\boldsymbol{z},\boldsymbol{w}) &\propto \pi(\theta|\boldsymbol{z})\exp\{\nu' M_{n_2}(\eta,\varphi)\}\\ &\propto \pi(\varphi)\pi(\eta|\varphi)\exp\{\nu L_{n_1}(\varphi) + \nu' M_{n_2}(\eta,\varphi)\},\end{aligned} \tag{2.3}$$

where $\nu' \geq 0$ is an additional learning rate. If $\nu' = 1$ and the loss function $m(\cdot)$ is the log-likelihood for $\boldsymbol{w}$, this is a conventional Bayesian update using the data for the second module. The full belief update (2.3), takes the form of Bayesian updating where the likelihood has been replaced by the loss likelihood $\exp\{\nu L_{n_1}(\varphi) + \nu' M_{n_2}(\eta, \varphi)\}$.

Loss likelihoods are often used to target a parameter of interest directly when the attempt to specify a full probabilistic model might result in severe misspecification. Our use of generalized Bayesian methods within the modular inference framework is slightly different, since here any model misspecification is structured. By this we mean that misspecification occurs in only some modules. From this perspective, it is attractive to use a loss likelihood for some of the modules, while retaining a probabilistic model for the data in other modules. If the parameter of interest occurs in a correctly specified module for which we use a probabilistic model for the data, then this parameter is meaningful, regardless of whether misspecification occurs in other modules. In this situation, the motivation to use a loss likelihood in a misspecified module is to ensure that the parameter of interest retains its intended meaning in the analysis. So the loss likelihood might be used not to target the primary parameter of interest, but to achieve sensible inference about a nuisance parameter, without which the primary inference of interest could be compromised. The comparison of the full and cut versions of the generalized Bayesian posterior may tell us whether the interpretation of the inference changes when we cut, and whether the use of a certain loss likelihood was successful in achieving a consistent inferential interpretation in the cases of only one or both data sources being used. In conventional generalized Bayesian inference, it may sometimes be difficult to define the parameter of interest as the minimizer of a loss, or to formulate priors for such a parameter, although similar difficulties occur with misspecified generative models too. The modular setting makes the specification of a loss easier in structured problems, since the loss is only required in certain modules.

In generalized Bayesian inference the choice of the learning rate is very important, and this is true in the case of modular inference considered here also. See Wu and Martin (2020) for a review and comparison of different methods. If the learning rate is not carefully chosen, uncertainty quantification by the generalized posterior distribution can be very poor. Generalizing similar ideas to Holmes and Walker (2017) and Lyddon et al. (2019), later we suggest choosing ν and ν' based on an information matching argument. Although we have introduced two learning rates, there are several special cases of interest in our later discussion. In our examples we use generative models for the first module that are well-specified, and use the negative log-likelihood as the loss. Here it makes sense to choose $\nu = 1$. Conventional generalized Bayesian analyses without modular structure would correspond to $\nu = \nu'$. In our later theory, for consistency with the rest of the generalized Bayesian literature, we assume this. There is no loss of generality in doing so, or even in omitting learning rates altogether, since any chosen learning rates can always be absorbed into the definition of the loss function. Our discussion of the choice of learning rates in Section 3.2 is general, however, describing the setting where separate learning rates for the two modules must be chosen.

2.3 Cutting Feedback with Generalized Posteriors

Our confidence in the accuracy of the first module means that the criterion $L_{n_1}(\varphi)$ can be chosen as the negative log-likelihood. However, since we are working with generalized posteriors, we only maintain that $L_{n_1}(\varphi)$ produces "reliable inferences" for φ. Our lack of confidence in the specification of the second module means we are concerned that incorporating this module may contaminate our inferences for φ. In such situations, extending cutting feedback methods to a generalized Bayesian framework can be helpful, and can yield more reliable inferences than conducting standard Bayesian inference using the joint likelihood.

In the two module system of Figure 1 for a probabilistic model, the first module is $\{\pi(\varphi), p(\boldsymbol{z}|\varphi)\}$, and the second module is $\{\pi(\eta|\varphi), p(\boldsymbol{z}|\eta,\varphi)\}$. In our generalized Bayesian analysis the modules are $\{\pi(\varphi), \exp\{\nu L_{n_1}(\varphi)\}$ and $\{\pi(\eta|\varphi), \exp\{\nu M_{n_2}(\eta,\varphi)\}$, if there is a single learning rate for both modules.

Generalized Bayesian analyses have been used in the context of two module system previously, but only as a justification for parametric cutting feedback methods when a probabilistic model is specified. Carmona and Nicholls (2020) considered order coherence for cut and semi-modular inference methods, and Nicholls et al. (2022) observed that the implicit loss function used in these approaches is not additive as required in the theory of Bissiri et al. (2016). Nicholls et al. (2022) generalize the existing theory to "prequentially additive" loss functions, which is enough to justify standard parametric cut inference as valid and order coherent generalized Bayesian updating. In contrast to this work, our aim is not to justify cutting feedback methods for probabilistic multi-modular models as coherent in some sense, but to consider situations where there may be no probabilistic model for the data, but only loss functions to connect module data to parameters.

To present cutting feedback for generalized posteriors, decompose $\pi(\theta|\boldsymbol{y})$ in (2.3) as the product of a marginal posterior for $\varphi|\boldsymbol{z}$, a conditional posterior for $\eta|\boldsymbol{w},\varphi$, and a "feedback term":

$$\pi(\theta|\boldsymbol{y}) = \pi_{\text{cut}}(\varphi|\boldsymbol{z})\pi(\eta|\boldsymbol{w},\varphi)\tilde{p}(\boldsymbol{w}|\varphi), \tag{2.4}$$

where $\pi_{\text{cut}}(\varphi|\boldsymbol{z}) \propto \pi(\varphi)\exp\{\nu L_{n_1}(\varphi)\}$, $\pi(\eta|\boldsymbol{w},\varphi) := \pi(\eta|\varphi)\exp\{\nu M_{n_2}(\eta,\varphi)\}/m_\eta(\boldsymbol{w}|\varphi)$, and

$$\tilde{p}(\boldsymbol{w}|\varphi) \propto m_\eta(\boldsymbol{w}|\varphi), \quad m_\eta(\boldsymbol{w}|\varphi) = \int_{\mathcal{E}} \pi(\eta|\varphi)\exp\{\nu M_{n_2}(\eta,\varphi)\}\mathrm{d}\eta. \tag{2.5}$$

The feedback term $\tilde{p}(\boldsymbol{w}|\varphi)$ derives its name from representing the influence of module two on the marginal posterior for φ. To understand this better, consider integrating out η in (2.4), to obtain $\pi(\varphi|\boldsymbol{y}) = \pi_{\text{cut}}(\varphi|\boldsymbol{z})\widetilde{p}(\boldsymbol{w}|\varphi)$. Since $\pi_{\text{cut}}(\varphi|\boldsymbol{z})$ represents the posterior density for φ based only on the first module data $\boldsymbol{z}$, we see that $\widetilde{p}(\boldsymbol{w}|\varphi)$ modifies this posterior based on the second module data to give the φ marginal of $\pi(\theta|\boldsymbol{y})$. Dropping the feedback term $\tilde{p}(\boldsymbol{w}|\varphi)$ in (2.4) produces a "generalized cut posterior", extending (2.1):

$$\pi_{\text{cut}}(\theta|\boldsymbol{z},\boldsymbol{w}) := \pi_{\text{cut}}(\varphi|\boldsymbol{z})\pi(\eta|\boldsymbol{w},\varphi).$$

In this joint cut posterior, marginal posterior inferences for φ are obtained based on module one only, and the conditional posterior density of η given φ is the same as

for $\pi(\theta|\boldsymbol{y})$ and based on module two only. Our discussion of cut inference is in the generalized Bayesian framework, but if we use the negative log likelihood as the loss for an assumed probabilistic model, our definition of the cut posterior reduces to (2.1).

Obtaining samples from the cut posterior $\pi_{\text{cut}}(\theta|\boldsymbol{z},\boldsymbol{w})$ is challenging. Since

$$\pi_{\text{cut}}(\theta|\boldsymbol{z},\boldsymbol{w}) \propto \pi(\varphi)\exp\{\nu L_{n_1}(\varphi)\}\frac{\pi(\eta|\varphi)\exp\{\nu M_{n_2}(\eta,\varphi)\}}{m_\eta(\boldsymbol{w}|\varphi)},$$

if MCMC is used to sample from $\pi_{\text{cut}}(\theta|\boldsymbol{z},\boldsymbol{w})$, we must evaluate the term $m_\eta(\boldsymbol{w}|\varphi)$. This term is similar to a"marginal likelihood" for η conditional on a fixed φ, and is generally not available in closed form outside of toy examples. In principle, even though we are in the case of generalized posteriors, the computationally intensive methods proposed by Plummer (2015), and Jacob et al. (2017) to deal with the intractable term $m_\eta(\boldsymbol{w}|\varphi)$ could be used to sample from the cut posterior.

While sampling from $\pi_{\text{cut}}(\theta|\boldsymbol{z},\boldsymbol{w})$ is difficult, draws from $\pi(\eta|\boldsymbol{w},\varphi)$ for any fixed φ can be made without the need to compute $m_\eta(\boldsymbol{w}|\varphi)$. This suggests the following sequential algorithm to obtain draws from $\pi_{\text{cut}}(\theta|\boldsymbol{w},\boldsymbol{z})$: first, sample $\varphi' \sim \pi_{\text{cut}}(\varphi|\boldsymbol{z})$; then, sample $\eta' \sim \pi(\eta|\boldsymbol{w},\varphi')$. At the first stage, draws from $\pi_{\text{cut}}(\varphi|\boldsymbol{z})$ could be obtained by running an MCMC chain targeting the posterior density $\pi_{\text{cut}}(\varphi|\boldsymbol{z})$. The conditional draws of η given φ are then performed by running a separate MCMC chain for each sample, which is computationally burdensome. The approach is reminiscent of multiple imputation algorithms, and was originally suggested by Plummer (2015), who also discussed a related tempering method of similar computational complexity. The sequential sampling approach above can also be thought of as implementing a modified Gibbs sampling algorithm with blocks φ and η, but where the likelihood term from the second module is dropped when forming the full conditional distribution for φ. As mentioned earlier, the resulting modified conditional distributions are not the full conditional distributions of any joint distribution in general, and if we attempt to replace the usually intractable direct sampling of the modified conditional distributions with Metropolis-within-Gibbs steps, then the stationary distribution of the MCMC sampler depends on the proposal used. A number of other authors have investigated computation for cutting feedback (Jacob et al., 2020; Liu and Goudie, 2022b; Yu et al., 2023; Carmona and Nicholls, 2022) and this remains an active area of research.

The sequential definition of the cut posterior distribution in the two-module system suggests that analysis of cut procedures should study $\pi_{\text{cut}}(\varphi|\boldsymbol{z})$ to understand the behavior of cut inference for φ, and the conditional posterior of $\pi(\eta|\boldsymbol{w},\varphi)$ to understand how uncertainty about φ propagates to marginal cut inferences about η. This is the strategy we follow in the next section. Such an analysis is complicated by the fact that $\pi_{\text{cut}}(\theta|\boldsymbol{z},\boldsymbol{w})$ does not arise as a posterior for a generative model, and therefore we must use techniques employed in the study of generalized posteriors to analyze $\pi_{\text{cut}}(\theta|\boldsymbol{z},\boldsymbol{w})$.

3 The Behavior of $\pi_{\text{cut}}(\theta|\boldsymbol{z},\boldsymbol{w})$

In this section, we explore the behavior of $\pi_{\text{cut}}(\theta|\boldsymbol{z},\boldsymbol{w})$ by separately analysing $\pi_{\text{cut}}(\varphi|\boldsymbol{z})$, and then analysing $\pi(\eta|\boldsymbol{w},\varphi)$ when we condition on an observed value of φ within the

high probability region of $\pi_{\text{cut}}(\varphi|\boldsymbol{z})$. This yields useful insights into the behavior of cut posteriors and allows us to develop new diagnostic tools for examining these posteriors. The approximations implied by our asymptotic results are also valuable for cut posterior computation, which can be time-consuming and difficult to implement in practice. As discussed in Section 2.3, a common way to sample the cut joint posterior distribution involves a nested MCMC scheme where a separate MCMC chain is run to draw a sample of η from its posterior conditional density for each marginal cut posterior sample φ. If this MCMC step can be replaced by a draw from a normal approximation, or the normal approximation is used to obtain a good proposal density for MCMC or importance sampling, then this can reduce the computational burden of commonly used methods for cut posterior computation.

3.1 Maintained Assumptions and Main Results

To deduce the behavior of $\pi_{\text{cut}}(\theta|\boldsymbol{z},\boldsymbol{w})$, we must extend the assumptions often used to analyze generalized posteriors (see, e.g., Miller, 2021) so as to account for the two-step posterior updating, the different roles the loss functions play in the posterior, and the different sample sizes for the modules. The first assumption we maintain requires that the sample sizes in the two modules grow in rough proportion.

Assumption 1. $\zeta := \lim_{n_1,n_2\to\infty} n_1/n_2$ exists and is such that $0 < \zeta < \infty$.

The following conditions are sufficient to demonstrate that the cut posterior $\pi_{\text{cut}}(\varphi|\boldsymbol{z})$ is asymptotically well-behaved.

Assumption 2. (i) There exist $\mathbb{L}(\varphi)$ such that $\sup_{\varphi\in\Phi}|n_1^{-1}L_{n_1}(\varphi)-\mathbb{L}(\varphi)| = o_p(1)$. (ii) There is a unique $\varphi^\star \in \text{Int}(\Phi)$ such that for every $\delta > 0$ there exists $\epsilon(\delta) > 0$ so that $\sup_{\|\varphi-\varphi^\star\|\geq\delta}\{\mathbb{L}(\varphi)-\mathbb{L}(\varphi^\star)\} \leq -\epsilon(\delta)$. (iii) $\pi(\varphi)$ is continuous on Φ, with $\pi(\varphi^\star) > 0$, and $\int_\Phi \|\varphi\|\pi(\varphi)\text{d}\varphi < \infty$. (iv) For an arbitrary $\delta > 0$, and $\|\varphi-\varphi^\star\| \leq \delta$, $\mathbb{L}(\varphi)$ and $L_{n_1}(\varphi)$ are twice continuously differentiable, with $\sup_{\|\varphi-\varphi^\star\|\leq\delta}\|\nabla^2_{\varphi\varphi}L_{n_1}(\varphi)/n_1 - \nabla^2_{\varphi\varphi}\mathbb{L}(\varphi)\| = o_p(1)$, and $-\nabla^2_{\varphi\varphi}\mathbb{L}(\varphi^\star)$ positive-definite. (v) $\nabla_\varphi L_{n_1}(\varphi^\star)/\sqrt{n_1} = O_p(1)$.

Assumption 2 resembles standard conditions employed to deduce asymptotic posterior normality, see, e.g., Lehmann and Casella (2006), but where $L_{n_1}(\varphi)$ is an arbitrary loss function. These assumptions enforce smoothness on $L_{n_1}(\varphi)$ and identification to ensure that $\pi_{\text{cut}}(\varphi|\boldsymbol{z})$ concentrates onto $\varphi^\star$ – the value that minimizes the limit loss function $\mathbb{L}(\varphi)$, and the parameter value of interest in generalized Bayesian inference. Due to space restrictions, we forgo a detailed discussion of these assumptions until Section 2 in the supplementary material.

Define $\Sigma_{11} := -\nabla_{\varphi\varphi}\mathbb{L}(\varphi^\star)$, $Z_{n_1}(\varphi^\star) := -\Sigma_{11}^{-1}\nabla_\varphi L_{n_1}(\varphi^\star)/\sqrt{n_1}$, the local parameter $\phi := \sqrt{n_1}(\varphi-\varphi_\star)$ and its posterior $\pi(\phi|\boldsymbol{z}) = \pi(\varphi^\star+\phi/\sqrt{n_1}|\boldsymbol{z})/\sqrt{n_1}^{d_\varphi}$, which has support $\Phi_{n_1} := \{\phi : \sqrt{n_1}(\varphi-\varphi^\star) \in \Phi\}$. Lemma 1 states that the cut posterior $\pi_{\text{cut}}(\phi|\boldsymbol{z})$ behaves like a Gaussian density with mean $Z_{n_1}(\varphi^\star)$, and covariance $[\nu\Sigma_{11}]^{-1}$.

Lemma 1. Under Assumptions 1-2, $\int_{\Phi_{n_1}}\|\phi\|\left|\pi_{\text{cut}}(\phi|\boldsymbol{z}) - N\{\phi; Z_{n_1}(\varphi^\star), [\nu\Sigma_{11}]^{-1}\}\right|\text{d}\phi = o_p(1)$.

A possible interpretation of the conditional cut posterior $\pi(\eta|\varphi, \boldsymbol{w})$ suggested by an Associate Editor is that by first computing $\pi_{\text{cut}}(\varphi|\boldsymbol{z})$, the cut posterior $\pi(\eta|\varphi, \boldsymbol{w})$ can be viewed as a generalized posterior based on a "penalised loss function": values of η that do not agree with the value of φ – obtained under the cut posterior $\pi_{\text{cut}}(\varphi|\boldsymbol{z})$, and as measured by $M_{n_2}(\eta, \varphi)$ – produce a larger loss and thus are "penalised" in the posterior update. This interpretation further clarifies that the behavior of $\pi(\eta|\varphi, \boldsymbol{w})$ critically depends on the behavior of $M_{n_2}(\eta, \varphi)$ only when $\varphi \in \Phi_\delta$, with Φ_δ a region of high-posterior probability under $\pi_{\text{cut}}(\varphi|\boldsymbol{z})$. For example, uncertain quantification of η via $\pi(\eta|\boldsymbol{w}, \varphi)$ critically depends on the value of φ and thus the behavior of $\pi_{\text{cut}}(\varphi|\boldsymbol{z})$.

To formally demonstrate this behavior, we view $M_{n_2}(\eta, \varphi)$ as being indexed by a fixed $\varphi \in \Phi_\delta$, and to reinforce this perspective we use the notation $M_{n_2}(\eta|\varphi) := M_{n_2}(\eta, \varphi)$. Let $\Phi_\delta := \{\varphi \in \Phi : \|\varphi - \varphi^\star\| \leq \delta\}$ denote an arbitrary δ-neighborhood of $\varphi^\star$, and consider the following regularity conditions on $M_{n_2}(\eta|\varphi)$.

Assumption 3. (i) There exists $\mathbb{M}(\eta|\varphi)$ with $\sup_{\varphi\in\Phi_\delta, \eta\in\mathcal{E}} |{n_2}^{-1} M_{n_2}(\eta|\varphi) - \mathbb{M}(\eta|\varphi)| = o_p(1)$ for some $\delta > 0$. (ii) Given $\delta_1 > 0$, for each $\varphi \in \Phi_{\delta_1}$ there is an $\eta^\star_\varphi \in \text{Int}(\mathcal{E})$ such that for any $\delta_2 > 0$ there exist $\epsilon(\delta_1, \delta_2) > 0$, so that $\sup_{\varphi\in\Phi_{\delta_1}} \sup_{\|\eta-\eta^\star_\varphi\|\geq\delta_2} \{\mathbb{M}(\eta|\varphi) - \mathbb{M}(\eta^\star_\varphi|\varphi)\} \leq -\epsilon(\delta_1, \delta_2)$.

Assumption 3 imposes conditions on $M_{n_2}(\eta|\varphi)$ that ensure the conditional cut posterior $\pi(\eta|\boldsymbol{w}, \varphi)$ concentrates onto $\eta^\star_\varphi$ – the limit minimizer of $\mathbb{M}(\eta|\varphi)$ for a fixed $\varphi \in \Phi_\delta$. These conditions imply that if we study $\pi(\eta|\varphi, \boldsymbol{w})$ when φ is restricted to a neighbourhood of $\varphi^\star$, then the conditional posterior should concentrate mass near $\eta^\star_\varphi$. By "conditioning on" $\varphi \in \Phi_\delta$ we can view $M_{n_2}(\eta|\varphi)$ as being indexed by a fixed parameter value, φ, and verify the regularity conditions in Assumption 3 using tools from empirical processes theory at the point $\varphi \in \Phi_\delta$; we refer to Portier (2016) for a discussion and several examples. In Section 2 of the supplementary material, we give a more detailed discussion of Assumption 3 and how it can be verified in certain examples.

Assumption 4. (i) For some $\delta > 0$, and each $\varphi \in \Phi_\delta$, $\pi(\eta|\varphi)$ is continuous in η. (ii) $\sup_{\varphi\in\Phi_{\delta_1}} \int_{\mathcal{E}} \|\eta\|\pi(\eta|\varphi)\text{d}\eta < \infty$.

Assumption 4(i) is a standard regularity condition, while Assumption 4(ii) implies that the conditional prior has sufficient moments. A sufficient condition for the latter condition is prior independence $\pi(\eta|\varphi) = \pi(\eta)$ and $\int \|\eta\|\pi(\eta)\text{d}\eta < \infty$.

Assumption 5. For some $\delta_1, \delta_2 > 0$, the following are satisfied. There exist a vector function $\Delta_{n_2}(\eta|\varphi)$, and matrix function $J(\eta|\varphi)$ such that

$$M_{n_2}(\eta|\varphi) - M_{n_2}(\eta^\star_\varphi|\varphi) = (\eta - \eta^\star_\varphi)^\top \Delta_{n_2}(\eta^\star_\varphi|\varphi) - \frac{n_2}{2}(\eta - \eta^\star_\varphi)^\top J(\eta|\varphi)(\eta - \eta^\star_\varphi) + R_{n_2}(\eta, \varphi).$$

(i) for all $\varphi \in \Phi_{\delta_1}$, $\Delta_{n_2}(\eta^\star_\varphi|\varphi)/\sqrt{n_2} = O_p(1)$;

(ii) the map $\eta \mapsto J(\eta|\varphi)$ is continuous for all $\|\eta - \eta^\star_\varphi\| \leq \delta_2$, for each $\varphi \in \Phi_{\delta_1}$, and $J(\eta^\star_\varphi|\varphi)$ is positive-definite for each $\varphi \in \Phi_{\delta_1}$;

(iii) for any $\delta_2 > 0$, $\sup_{\varphi\in\Phi_{\delta_1}} \sup_{\|\eta-\eta^\star_\varphi\|\leq\delta_2} R_{n_2}(\eta, \varphi)/[1 + n_2\|\eta - \eta^\star_\varphi\|^2] = o_p(1)$.

Assumption 5 is required to ensure that $\pi(\eta \mid \boldsymbol{w}, \varphi)$ concentrates in a Gaussian manner as the sample size diverges, and ensures that $M_{n_2}(\eta|\varphi)$ admits a valid quadratic expansion around $\eta^\star_\varphi$ for each $\varphi \in \Phi_\delta$, with a remainder term that can be suitably controlled in a probabilistic sense. The expansion underlying Assumption 5 is akin to the local asymptotic normality conditions often assumed when proving asymptotic posterior normality in exact Bayesian inference (see, e.g., Chapter 7 of van der Vaart, 2000 for a textbook example and discussion). Due to space restrictions, we must forgo a more detailed discussion of this assumption until Section 2 in the supplementary material.

The above assumptions allow us to study the large sample behavior of the cut posterior $\pi(\eta|\boldsymbol{w}, \varphi)$. To present this behavior as succinctly as possible, define $Z_{n_2}(\eta^\star_\varphi|\varphi) := J(\eta^\star_\varphi|\varphi)^{-1}\Delta_{n_2}(\eta^\star_\varphi|\varphi)/\sqrt{n_2}$, as well as the local parameter $t := \sqrt{n_2}(\eta - \eta^\star_\varphi)$ and its posterior $\pi(t|\boldsymbol{w}, \varphi) = \pi(\eta^\star_\varphi + t/\sqrt{n_2}|\boldsymbol{w}, \varphi)/\sqrt{n_2}^{d_\eta}$, which has support where $\mathcal{E}_{n_2} := \{t = \sqrt{n_2}(\eta - \eta^\star_\varphi) : \eta \in \mathcal{E}, \varphi \in \Phi_\delta\}$.

Theorem 1. If for some $\delta > 0$, Assumptions 1-5 are satisfied for $\varphi \in \Phi_\delta$, then $\int_{\mathcal{E}_{n_2}} \|t\| \left|\pi(t|\boldsymbol{w}, \varphi) - N\{t; Z_{n_2}(\eta^\star_\varphi|\varphi), [\nu J(\eta^\star_\varphi|\varphi)]^{-1}\}\right| \mathrm{d}t = o_p(1)$.

Theorem 1 demonstrates that in large samples $\pi(\eta|\boldsymbol{w}, \varphi)$ behaves like a Gaussian density with a mean and variance *that both depend on* φ. This result is useful for at least two reasons. First, the only other result on the behavior of cut posteriors of which we are aware, Pompe and Jacob (2021), demonstrates that in large samples the cut posterior for $\theta = (\varphi^\top, \eta^\top)^\top$ is jointly Gaussian with a variance *that depends on a fixed* $\varphi^\star$ *and* $\eta^\star = \eta^\star_{\varphi^\star}$, and is similar to other results for generalized posteriors, see, e.g., Chernozhukov and Hong (2003), Zhang (2006), and Miller (2021), where a conventional multivariate normal (Laplace) approximation is produced. However, this joint approximation has the immediate drawback that the induced conditional posterior (approximation) has a covariance matrix *that does not depend on the conditioning value* φ. Since in small-to-medium sample sizes the conditional posterior $\pi(\eta|\boldsymbol{w}, \varphi)$ will have a mean and variance that changes with the value of φ, a global approximation of this kind is unlikely to be accurate.

Second, the conditional approximation in Theorem 1 can be directly used in cases where accessing $\pi(\eta|\boldsymbol{w}, \varphi)$ may be difficult but where $Z_{n_2}(\eta^\star_\varphi|\varphi)$ and $J(\eta^\star_\varphi|\varphi)$ can be easily estimated. The latter occurs, for example, in cases where the MCMC sampler has a difficult time sampling $\pi(\eta|\boldsymbol{w}, \varphi)$ at the particular value of φ on which we are conditioning. In such cases, the normal approximation in Theorem 1 can be useful as a proposal distribution for MCMC or for importance sampling.

3.2 Calibration of Learning Rates

The uncertainty quantification of the generalized cut posterior density $\pi_{\text{cut}}(\theta|\boldsymbol{z}, \boldsymbol{w})$ depends crucially on the choice of learning rates, which we now discuss. Consider the loss likelihood term in (2.3), where ν and ν' need to be chosen. Lyddon et al. (2019), inspired by an earlier method of Holmes and Walker (2017), suggest to choose learning rates by matching the Fisher information number for the generalized Bayes update to

the Fisher information number from an update based on a loss likelihood bootstrap approach, asymptotically. We do not describe here in detail the reasoning behind the method of Lyddon et al. (2019), but the key to its application here for estimation of multiple learning rates is to exploit the modular structure of the model. We set the first learning rate ν based on the prior to posterior update for φ in the first module, and set the second learning rate ν' based on the conditional prior to conditional posterior update for η in the second module, fixing φ to an estimate based on module one.

To state the idealized learning rates we require some additional notation. Let

$$\Sigma_{11} = -\nabla^2_{\varphi\varphi}\mathbb{L}(\varphi^*), \quad \Sigma_{22} = -\nabla^2_{\eta\eta}\mathbb{M}(\eta^*|\varphi^*), \quad \Sigma_{12} = \nabla^2_{\varphi\eta}\mathbb{M}(\eta^*|\varphi^*),$$

$$\Psi_{11} = \lim_{n\to\infty} \mathrm{Cov}(L_{n_1}(\varphi^*)/\sqrt{n_1}), \quad \Psi_{22} = \lim_{n\to\infty} \mathrm{Cov}(\Delta_{n_2}(\eta^*|\varphi^*)/\sqrt{n_2}).$$

With this notation, if we apply the method of Lyddon et al. (2019) for choosing ν based on the update for the parameter φ using the first module only, we obtain the ideal choice

$$\nu = \mathrm{tr}(\Sigma_{11}{\Psi_{11}}^{-1}\Sigma_{11})/\mathrm{tr}(\Sigma_{11}).$$

We can estimate Σ_{11} by $-{n_1}^{-1}\nabla^2_{\varphi\varphi}L_{n_1}(\widehat{\varphi})$, where $\widehat{\varphi} = \arg\max_\varphi L_{n_1}(\varphi)$. To estimate Ψ_{11}, we could use ${n_1}^{-1}\sum_{i=1}^{n_1}\nabla_\varphi\ell(z_i;\widehat{\varphi})\nabla_\varphi\ell(z_i;\widehat{\varphi})^\top$, although Ψ_{11} can also be estimated in other ways.

After calibrating ν based on the first module, we can calibrate ν' by considering a conditional update of our beliefs for η in the second module, conditional on an estimate of φ from the first module, $\varphi = \widehat{\varphi}$ say. Matching the Fisher information number suggests choosing ν' as $\nu' = \mathrm{tr}(\Sigma_{22}{\Psi_{22}}^{-1}\Sigma_{22})/\mathrm{tr}(\Sigma_{22})$. To estimate Σ_{22} we can use $-{n_2}^{-1}\nabla^2_{\eta\eta}M_{n_2}(\widehat{\eta}_{\widehat{\varphi}}|\widehat{\varphi})$, where $\widehat{\eta}_{\widehat{\varphi}} = \arg\max_\eta M_{n_2}(\eta|\widehat{\varphi})$, and Ψ_{22} can be estimated by ${n_2}^{-1}\sum_{i=1}^{n_2}\nabla_\eta m(w_i;\widehat{\eta}_{\widehat{\varphi}},\widehat{\varphi})\nabla_\eta m(w_i;\widehat{\eta}_{\widehat{\varphi}},\widehat{\varphi})^\top$, or using some other method.

In a conventional generalized Bayesian analysis, there is only one learning rate to choose, but here there are two. This makes choosing learning rates more difficult, but also makes the modular generalized Bayesian approach more flexible. The way that marginal inferences about φ and conditional inferences for η given φ can be done separately in a modular approach for two different loss functions makes the choice of two learning rates feasible. We thank two anonymous referees for their insight in encouraging us to explore further the choice of separate learning rates for different modules.

3.3 Diagnostics for $\eta|w,\varphi$: Understanding Uncertainty Propagation

Theorem 1 demonstrates that even in large samples the behavior of $\pi(\eta|\boldsymbol{w},\varphi)$ depends on the value of φ on which we are conditioning. Moreover, for different values of φ, the resulting mean and variance can vary substantially. Both the cut and full marginal posterior density for η are obtained by integrating out φ in the conditional posterior density $\pi(\eta|\varphi,\boldsymbol{y})$, but using different distributions for φ: $\pi(\varphi|\boldsymbol{z})$ for the cut posterior, and $\pi(\varphi|\boldsymbol{y})$ for the full posterior. Hence it is the different uncertainty quantification for φ, and the way this propagates when we integrate out φ in $\eta|\varphi,\boldsymbol{y}$, that determines the different inferences for η in the cut and full posterior densities. We now discuss

a number of diagnostics to understand uncertainty propagation in this sense, using the result of Theorem 1 about the behaviour of the conditional posterior density for convenient computation.

Most simply, if η is low-dimensional, we can visualise the impact of φ on the posterior for $\eta|\boldsymbol{w}, \varphi$ by viewing the kernel

$$|J(\eta^\star_\varphi|\varphi)|^{1/2} \exp\left\{-n_2 \cdot \nu \cdot (\eta - \eta^\star_\varphi)^\top J(\eta^\star_\varphi|\varphi)(\eta - \eta^\star_\varphi)/2\right\},$$

across a given range of values for φ. The resulting plot will demonstrate how the cut posterior for η changes as the conditioning value of φ changes.

The above approximation cannot be directly accessed, since $\eta^\star_\varphi$ and $J(\eta^\star_\varphi|\varphi)$ are unknown in practice. However, in cases where $M_{n_2}(\eta|\varphi)$ is twice continuously differentiable in η, it is simple to estimate $\eta^\star_\varphi$ and $J(\eta^\star_\varphi|\varphi)$ by their empirical counterparts $\widehat{\eta}_\varphi := \text{argmax}_\eta\, M_{n_2}(\eta|\varphi)$, and $J_{n_2}(\widehat{\eta}_\varphi|\varphi) := {n_2}^{-1}\nabla^2_{\eta\eta} M_{n_2}(\widehat{\eta}_\varphi|\varphi)$ respectively.

Understanding how the value of φ impacts uncertainty quantification of $\pi(\eta|\boldsymbol{w}, \varphi)$ is of particular interest. For $\alpha \in (0, 1)$, let $C^\eta_\alpha(\varphi) \subset \mathcal{E}$ be such that $\int_{C^\eta_\alpha(\varphi)} \pi(\eta|\boldsymbol{w}, \varphi)\mathrm{d}\eta = 1-\alpha$. Since generating samples from the cut posterior $\pi(\theta|\boldsymbol{w}, \boldsymbol{z})$ directly can be difficult, construction of $C^\eta_\alpha(\varphi)$ at many difference values of φ is also difficult. However, the conditional large sample approximation in Theorem 1 can be used to easily construct an estimate of $C^\eta_\alpha(\varphi)$ across several values of φ, and allows us to understand how the cut posterior $\pi(\eta|\boldsymbol{w}, \varphi)$ quantifies uncertainty about η at various values of φ; see Appendix C for an algorithmic implementation of this procedure. We illustrate the construction of credible sets for η conditional on φ in the example of Section 4.1.

The large sample approximation in Theorem 1 can also be used to visualize the behavior of specific functionals of interest, e.g., moments of $\eta|\boldsymbol{w}, \varphi$, without running an MCMC sampling algorithm to obtain draws of $\eta|\boldsymbol{w}, \varphi$. As an example, suppose that η is a scalar for simplicity and that we are interested in understanding how the variance of its cut posterior depends on φ. Using the law of total variance, we can write

$$\text{Var}(\eta) = E(\text{Var}(\eta|\varphi)) + \text{Var}(E(\eta|\varphi)),$$

(where expectations in this expression are with respect to the cut posterior) and for draws $\varphi^{(s)} \sim \pi_{\text{cut}}(\varphi|\boldsymbol{z})$, $s = 1, \ldots, S$, we can plot histograms of $\text{Var}(\eta|\varphi^{(s)})$ and $E(\eta|\varphi^{(s)})$ to understand how variability in η relates to φ. The conditional means and variances can be approximated by the normal approximations obtained from Theorem 1. In the example in Section 4.1, we discuss diagnostics of this type, as well as methods for understanding posterior skewness in the parameter in the second module.

Finally, another way of understanding how differing uncertainty quantification for φ propagates in inference about η is to consider semi-modular inference (SMI) (Carmona and Nicholls, 2020). Carmona and Nicholls proposed SMI as an extension of cut model inference. Nicholls et al. (2022) extend this construction to prequentially additive loss functions and Carmona and Nicholls (2022) investigate the use of normalizing flows for their computation.

Consider once again the two module system, and point estimation for the shared parameter φ based on full and cut posterior distributions. The intuition behind SMI is that if the degree of misspecification is not severe, then the bias of the full posterior estimator may only be moderate, while its variance might be greatly reduced compared to the cut posterior estimator. In this case, full posterior estimates may have better frequentist performance in managing a bias-variance trade-off. If the misspecification is serious, however, full posterior estimation may have a large bias, and estimation based on the cut model may be preferred. Instead of making a binary choice between the full and cut posterior density, it might be better to modulate the influence of the misspecified module in a more continuous way, using an "influence parameter" denoted here as $\gamma \in [0, 1]$. In the proposal of Carmona and Nicholls (2020), the choice $\gamma = 0$ results in the cut posterior, whereas $\gamma = 1$ corresponds to the full posterior, so that the SMI posterior interpolates between cut and full posterior based on the influence parameter. Nicholls et al. (2022) also explore some more Bayesian properties of validity and order-coherence of SMI posteriors for their original approach and some alternatives. In the supplementary material we describe how the large sample normal approximations of Theorem 1 can be used to perform efficient computation for the original method of Carmona and Nicholls (2020), but also what Nicholls et al. (2022) call the γ-SMI posterior, although we prefer the name marginal SMI posterior in the rest of the paper. Of the latter, the authors write: "However it is very awkward computationally and in fact we have no idea how to implement it in practice." The conditional Laplace approximations we have developed provide one practical implementation. Examining either SMI posterior for a grid of values for γ that interpolates between the cut and full posterior gives information about how much information from the suspect second module can be used before inferences of interest are affected. If a single choice of γ is needed, this can be done on predictive grounds, as described in Carmona and Nicholls (2020). The γ-SMI method is demonstrated for the example of Section 4.1 in the supplementary material.

While the diagnostics that we suggest for understanding propagation of uncertainty in this section are practical to compute, we admit that they are most useful in the setting where the conditional posterior distribution of η given φ is close to Gaussian, although the normal approximations can provided by Theorem 1 can also be used as proposal distributions in MCMC and importance sampling for situations where the normal approximation does not suffice on its own.

4 Examples

In this section we consider two examples. The first example illustrates our large sample approximations for cut posterior computation, and for implementing diagnostics for understanding uncertainty propagation between modules. We consider both probabilistic model specifications as well as a generalized Bayesian analysis using a quasi-likelihood. Our second example also considers a generalized Bayesian analysis, for which the learning rate for the second module needs to be carefully chosen. We illustrate a situation where an appropriate choice of the loss function can resolve conflict between cut and full posterior inferences, giving insight into how an initially flawed parametric model may need to be improved.

4.1 HPV Prevalence

We now return to the HPV prevalence example introduced in Section 2.1, and consider cut posterior computation, diagnostics and a generalized posterior analysis.

Cut Posterior Computation With Large Sample Approximation

In the HPV model specified in Section 2.1, it is straightforward to obtain posterior samples from $\pi_{\text{cut}}(\varphi|\boldsymbol{z})$. This is because the likelihood for each z_i is binomial, and the priors for the parameters φ_i are conjugate. In $\pi_{\text{cut}}(\varphi|\boldsymbol{z})$ the φ_i are independent, with $\pi_{\text{cut}}(\varphi_i|\boldsymbol{z})$ a beta density, $\text{Beta}(z_i+1, n_i - z_i + 1)$. We generate samples $\varphi^{(s)}$, $s = 1, \ldots, S = 1000$, from $\pi_{\text{cut}}(\varphi|\boldsymbol{z})$ by direct Monte Carlo sampling. To generate samples $\eta^{(s)}$ so that $(\varphi^{(s)}, \eta^{(s)})$ is a draw from the joint cut posterior density, we do the following. By Theorem 1, we can approximate the conditional posterior density of η given $\varphi^{(s)}$, w and z by a normal density with mean $\mu(\varphi^{(s)}) = \widehat{\eta}_{\varphi^{(s)}}$ and covariance matrix $\Sigma(\varphi^{(s)}) = n^{-1}\nabla_{\eta\eta} M_n(\widehat{\eta}_{\varphi^{(s)}}|\varphi^{(s)})$. For each $\varphi^{(s)}$, we generate $1,000$ proposal samples for η from a multivariate t-distribution with mean $\mu(\varphi^{(s)})$, scale matrix $\Sigma(\varphi^{(s)})$, and 5 degrees of freedom, and draw a single sample $\eta^{(s)}$ from these proposals using sampling importance resampling (SIR).

For comparison, we can also draw an approximate sample $\widetilde{\eta}^{(s)}$ say from the conditional normal approximation directly. For practical purposes the SIR samples can be considered near-exact, and Figure 2 (top row) shows the marginal posterior samples for $\eta = (\eta_1, \eta_2)$ for the two approaches. A sample based estimate of the 1-Wasserstein distance between the posterior marginal cut distributions estimated by the exact SIR and approximate conditional normal methods is 0.004 and 0.062 for η_1 and η_2 respectively, showing that our large sample conditional normal approximations result in accurate cut posterior computation. We can see that the marginal cut posterior density for η is non-Gaussian, but this is captured very well in the approximate sampling approach where the conditional posterior density for η is close to normal. It is the uncertainty about φ that is propagated in making marginal inferences about η that results in the non-Gaussian structure in the marginal posterior distribution for η. Also shown in Figure 2 are samples from the usual Bayesian posterior distribution, obtained via MCMC using the `rstan` package (Carpenter et al., 2017). The full and cut posterior inferences differ substantially, demonstrating how much the misspecification of the second module changes the inference about η here. The bottom row of the figure compares the univariate marginals for the cut and full posterior densities for η_1 and η_2.

Generalized Posterior Analysis

The middle row of Figure 2 shows samples from the generalized cut posterior distribution obtained when the Poisson likelihood is replaced by a quasi-likelihood (Wedderburn, 1974), which allows for overdispersion with respect to the Poisson model. When using the negative log quasi-likelihood as the loss for the second module, it is sensible to choose a learning rate $\nu' = 1$. For the first module we use the same parametric model as before. The overdispersion parameter in the quasi-likelihood is denoted by λ, and

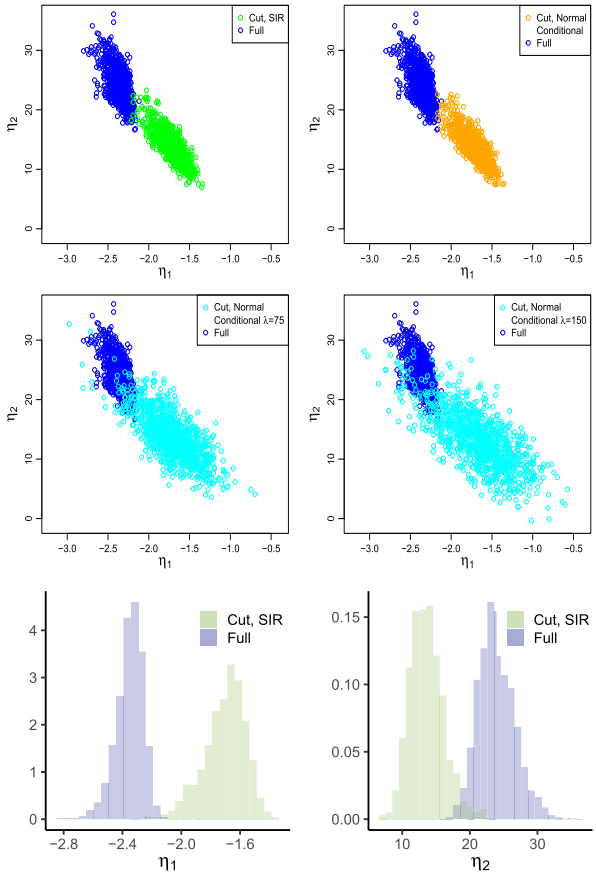

Figure 2: Top: marginal posterior samples for η for full posterior density (blue) obtained by MCMC, and cut posterior samples by SIR (green) and approximation by conditional normal sampling (orange). Middle: marginal posterior samples for η for full posterior density (blue) obtained by MCMC, and cut posterior samples from conditional normal sampling (cyan) for quasi-likelihood loss for the second module with $\lambda = 75$ (left) and $\lambda = 150$ (right). Bottom: histogram density estimates for η_1 (left) and η_2 (right) for full posterior (blue) and cut posterior samples by SIR (green).

instead of making the Poisson assumption that the mean and variance are equal, it is assumed that the variance is λ times the mean for each w_i. The left plot in the middle row is for $\lambda = 75$, and the right plot is for $\lambda = 150$. We can see that even if we assume a standard deviation for the w_i that is more than 10 times that implied by a Poisson mean-variance relationship, the full posterior samples do not become plausible under the cut distribution. Yu et al. (2023) have elaborated on the comparison of the cut and full posterior distributions as a kind of conflict check, and the lack of consistency of the cut and full posterior inferences here suggests that altering the parametric Poisson regression to another parametric model incorporating multiplicative overdispersion will not result in an adequate generative model for the data unless the degree of overdispersion is very

large. The samples in the quasi-likelihood analysis were generated using the conditional normal approximation for the density of η given φ.

Uncertainty Propagation

Figure 3 shows, for 5 samples from the marginal cut posterior distribution of φ, a 95% probability ellipsoid of minimal volume for the conditional normal approximations of $p(\eta|\varphi, y)$. The 5 φ samples are selected from $1,000$ cut posterior samples according to the 0.1, 0.3, 0.5, 0.7 and 0.9 quantiles of the determinant of the estimated conditional covariance matrix of η given φ. The variation in the shape of these ellipsoids is substantial as φ changes. The variation of the volumes of the ellipsoids with location shows a dependence between the conditional mean and variance of $\eta|\varphi$, which helps explain marginal skewness in η through propagation of uncertainty between the two modules.

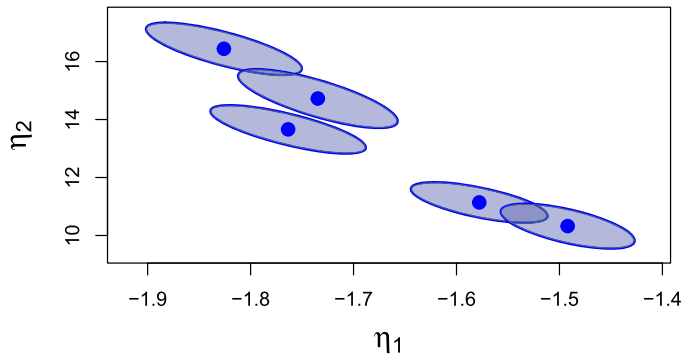

Figure 3: 95% probability ellipsoids of minimal volume for the normal approximation to the conditional posterior density of η given φ for 5 draws from the marginal cut posterior distribution of φ. The 5 φ samples are selected from $1,000$ cut posterior samples according to the 0.1, 0.3, 0.5, 0.7 and 0.9 quantiles of the determinant of the estimated conditional covariance matrix of η given φ.

We can also use the normal approximation to the conditional posterior density as a diagnostic to understand the way that the uncertainty in φ propagates into the second module, for both the cut and full posterior density. Noting that

$$\text{Var}(\eta_j) = E(\text{Var}(\eta_j|\varphi)) + \text{Var}(E(\eta_j|\varphi)), \tag{4.1}$$

we could plot histograms of the values $\mu(\varphi^{(s)})_j$, $s = 1, \ldots, S$ and $\Sigma(\varphi^{(s)})_{jj}$, $s = 1, \ldots, S$ for $j = 1, 2$ to understand how uncertainty in φ propagates into η. In (4.1) the expectations can be defined as with respect to either the full posterior distribution or with respect to the cut posterior distribution. The mean of the samples in a histogram of $\Sigma(\varphi^{(s)})_{jj}$ relates to the first term on the right-hand side of (4.1). The variability of the samples in a histogram of $\mu(\varphi^{(s)})_j$ assesses variability propagated to η_j from the second term on the right-hand side of (4.1).

Generalizing (4.1) to third central moments using the law of total cumulance (Brillinger, 1969), we can also write

$$E((\eta_j - E(\eta_j))^3) = E(E((\eta_j - E(\eta_j|\varphi))^3|\varphi)) + E((E(\eta_j|\varphi) - E(\eta_j))^3) + 3\text{Cov}(E(\eta_j|\varphi), \text{Var}(\eta_j|\varphi)). \quad (4.2)$$

Once again, the expectations in the above expression can be defined as with respect to either the full posterior distribution or with respect to the cut posterior distribution. If the conditional posterior for η_j given φ is approximately symmetric, then the first term on the right-hand side of (4.2) can be neglected. Then the posterior skewness of η_j depends on the second and third terms. These terms relate to the skewness of the conditional expectation $E(\eta_j|\varphi)$ (considered as a function of φ) and the covariance between the conditional mean and conditional variance. The skewness of the conditional expectation can be assessed from looking at the skewness in a histogram of $\mu(\varphi^{(s)})_j$, while plotting the samples $(\mu(\varphi^{(s)})_j, \Sigma(\varphi^{(s)})_{jj})$, $s = 1, \ldots, S$, is helpful for assessing the $\text{Cov}(E(\eta|\varphi), \text{Var}(\eta|\varphi))$ term in (4.2).

Figure 4 shows a scatterplot of $(\mu(\varphi^{(s)})_1, \Sigma(\varphi^{(s)})_{11})$, $s = 1, \ldots, S$, with histograms of each variable on the axes, for η_1. The plot on the left is for the cut posterior density, and the plot on the right is for the full posterior density. There is a strong negative relationship between the conditional posterior mean of φ and its conditional variance, as well as negative skewness in the histogram of $\mu(\varphi^{(s)})_1$, which by (4.2) explains the negative skew in the marginal distribution for η_1 evident in Figure 2. This is so for both the cut and full posterior densities.

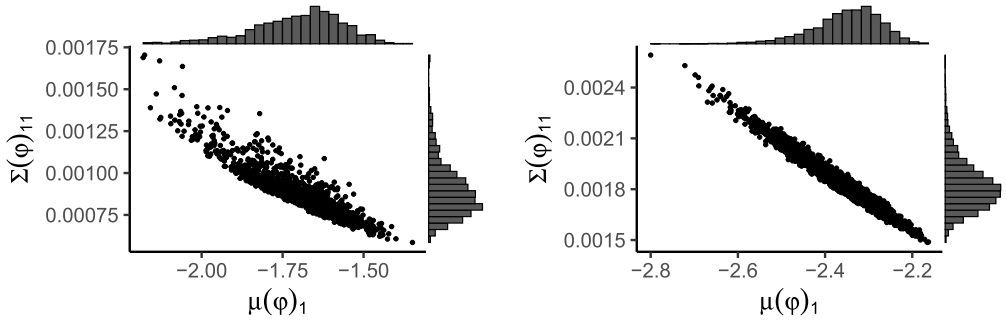

Figure 4: Scatterplot of $(\mu(\varphi^{(s)})_1, \Sigma(\varphi^{(s)})_{11})$, $s = 1, \ldots, S$, for cut posterior (left) and full posterior (right) samples. Histograms of each variable are shown on the axes.

Figure 5 shows a similar plot to Figure 4 for the parameter η_2. In this case, there is a strong positive relationship between the conditional posterior mean of φ and its conditional variance, and positive skewness in the histogram of $\mu(\varphi^{(s)})_2$, which explains the positive skew in the marginal distribution of η_2, in both the cut and full posterior densities, as shown in Figure 2. The dependence between $\mu(\varphi)_j$ and $\Sigma(\varphi)_{jj}$ in Figures 4 and 5 relates directly to the way the conditional variance of η depends on φ, which is exactly what is being captured in the conditional perspective taken in the theory

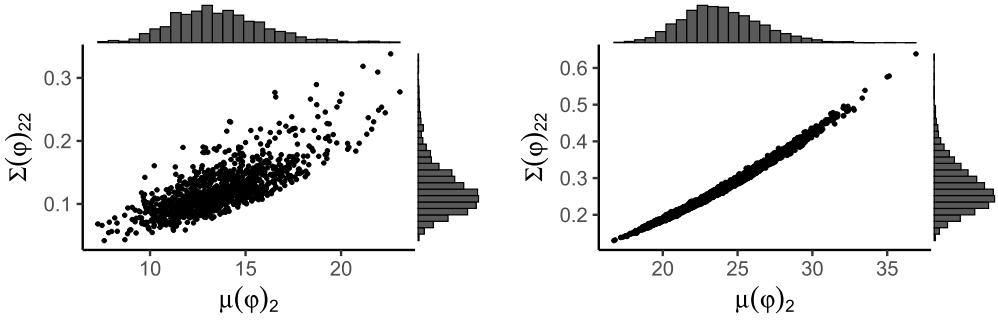

Figure 5: Scatterplot of $(\mu(\varphi^{(s)})_2, \Sigma(\varphi^{(s)})_{22})$, $s = 1, \ldots, S$, for cut posterior (left) and full posterior (right) samples. Histograms of each variable are shown on the axes.

of Section 3.1. Understanding this dependence is particularly useful for explaining the marginal posterior shape for η in the full and cut posterior distributions. In the supplementary material we describe the application of SMI for this example, comparing a new marginal SMI approach we introduce with the SMI of Carmona and Nicholls (2020). The two methods give similar results in this example for inference about η, and both demonstrate that even using a small amount of the information from the misspecified module changes the information substantially.

4.2 A Random Effects Model

Our second example, discussed in Liu et al. (2009), considers a random effects model. The data are denoted by Y_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, J$, where i indexes groups, and j indexes observations within groups. The data for group i is modelled as

$$Y_{ij}|\beta_i, \varphi_i \overset{iid}{\sim} N(\beta_i, \varphi_i^2), \quad j = 1, \ldots, J,$$

where β_i is a random effect, and φ_i is a group standard deviation. The prior density for β is

$$\beta_i|\psi \overset{iid}{\sim} N(0, \psi^2),$$

$i = 1, \ldots, N$, where ψ is the random effects standard deviation. Liu et al. (2009) consider this example to demonstrate a problem that can occur for some hierarchical models, in which there is a model for the random effects with thin tails, such as Gaussian. In the model above, if there is an outlying value for one of the random effects, this can lead to poor inference for the corresponding group standard deviation, and overshrinkage in estimating the random effect. The difficulty is most pronounced when the number of replicates J is small compared to N. Liu et al. (2009) give an insightful discussion that exploits the simple form of the model to do analytic calculations. We do not repeat their analysis here, but demonstrate the problem numerically and illustrate the utility of our generalized Bayes approaches to modular inference.

First, we will set up the model so that it takes the form of a two module system. Write $\beta = (\beta_1, \ldots, \beta_N)^\top$ and $\varphi = (\varphi_1, \ldots, \varphi_N)^\top$. Let $\eta = (\beta^\top, \psi)^\top$. We use similar

priors to Liu et al. (2009), although we parametrize our model in terms of standard deviations rather than variances and transform priors appropriately. Components of φ are independent in the prior, with marginal densities $\pi(\varphi_i) \propto \varphi_i^{-1}$. For the prior on ψ, we use $\pi(\psi|\varphi_i) \propto (\bar{\varphi}^2/J + \psi^2)^{-1}\psi$, where $\bar{\varphi}^2 = N^{-1}\sum_{i=1}^{N}\varphi_i^2$.

We will reduce the full data down to sufficient statistics. Let $w_i = J^{-1}\sum_{j=1}^{J} Y_{ij}$, $z_i = \sum_{j=1}^{J}(Y_{ij} - z_i)^2$, $i = 1, \ldots, N$, and write $\boldsymbol{z} = (z_1, \ldots, z_n)^\top$, $\boldsymbol{w} = (w_1, \ldots, w_n)^\top$. It is easily seen that $\boldsymbol{z}$ and $\boldsymbol{w}$ are sufficient for $\theta = (\varphi^\top, \eta^\top)^\top$, with $\boldsymbol{z}$ and $\boldsymbol{w}$ being independent of each other. The density of $\boldsymbol{z}|\varphi$, written $p(\boldsymbol{z}|\varphi)$, depends only on φ, with

$$z_i|\varphi_i \sim \text{Gamma}\left(\frac{J-1}{2}, \frac{1}{2\varphi_i^2}\right),$$

independently for $i = 1, \ldots, N$. Similarly, write $p(\boldsymbol{w}|\varphi, \eta)$ for the density of $\boldsymbol{w}$, and

$$w_i|\beta_i, \varphi_i \sim N\left(\beta_i, \frac{\varphi_i^2}{J}\right),$$

independently, for $i = 1, \ldots, N$. The model for the sufficient statistics is a two-module system. The first module consists of $p(\boldsymbol{z}|\varphi)$ and $p(\varphi)$, and the second module comprises $p(\boldsymbol{w}|\varphi, \eta)$ and $p(\eta|\varphi)$.

We simulate a dataset from the model, with $N = 100$, $J = 10$, $\psi = 1$ and $\varphi_i = 0.5$, $i = 1, \ldots, N$. The random effects vector β is simulated from its prior, except for β_1, which is fixed at 10. Since β_1 is inconsistent with the hierarchical prior, this leads to poor estimation of φ_1 when J is small compared to N, and poor estimation of β_1. Figure 6 (left) compares the posterior distributions of φ_1 from the conventional parametric and the cut posterior distributions. The boxplots are for 1,000 posterior samples in each case. The horizontal line shows the true value. The accuracy of the conventional posterior is poor, and inconsistent with the cut posterior inferences which are more accurate.

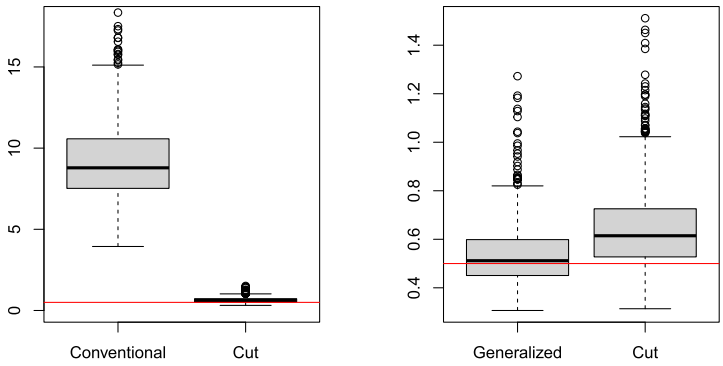

Figure 6: Left: comparison of posterior samples for φ_1 for conventional posterior and cut posterior for parametric model specification. Right: comparison of posterior samples for φ_1 for generalized Bayes analysis with Tukey's loss for the second module, $\kappa = 5$ and $\nu' = 3.3$ with cut posterior. All boxplots summarize 1,000 posterior samples, and the horizontal line is the true value $\varphi_1 = 0.5$.

Generalized Posterior Analysis

This example can be thought of as relating to a prior-data conflict (Evans and Moshonov, 2006) due to the simulated data being generated with a random effect for one group that is unusual with respect to the hierarchical prior. However, if only summary data were given, it would not be possible to tell if the problem with the model lies with the prior or the likelihood for a group having an unusual value for the sample mean. If an unusual sample mean summary was due to a single outlier in the data for one group, this would suggest the likelihood is at fault, whereas an unusual value for a random effect would result in all observations in the group being affected. So replicate data is essential to distinguish between a problem with the likelihood and a problem with the prior in model checking. Although knowing sufficient statistics is enough for inference if the model is correct, for model checking we may need non-sufficient information lying in the replicates directly. Since the situation of summary data only being available is common in some applications (in meta- analysis for example, where only summary data might be published) we feel it is interesting to analyze this standard example from the cut literature from the point of view of a possibly misspecified likelihood when only summary data is available.

With this motivation, it is interesting to replace the normal model for w_i in module 2 with a loss likelihood, to see whether this resolves the inconsistency between the cut and full generalized posterior inferences. Here we consider a slightly extended version of Tukey's loss (Beaton and Tukey, 1974), which was recently used for a generalized Bayesian analysis by Jewson and Rossell (2022),

$$m(u) = \begin{cases} u^2/2 - u^4/(4\kappa) - u^6/(6\kappa), & \text{if } |u| \leq \kappa \\ \kappa^2/6 \end{cases},$$

for $\kappa \geq 0$; Tukey's original loss did not contain the additional term $u^6/(6\kappa)$, which further penalizes large values. As pointed out by Jewson and Rossell (2022), Tukey's loss can be useful when an analyst knows the distribution of the data has heavy tails, but a precise knowledge of the tail behaviour is difficult to formalize. Writing $w_i' = w_i'(\varphi_i, \beta_i) = (w_i - \beta_i)/(\phi_i/\sqrt{J})$, in our generalized Bayesian analysis we replace the Gaussian negative log-likelihood terms

$$-\log p(w_i|\varphi_i, \beta_i) = \frac{1}{2}\log\frac{2\pi\varphi_i^2}{J} - \frac{1}{2}w_i'^2,$$

with the modified version of Tukey's loss

$$m(w_i; \eta, \varphi) = \frac{1}{2}\log\frac{2\pi\varphi_i^2}{J} + \begin{cases} \frac{w_i'^2}{2} - \frac{w_i'^4}{4\kappa^2} - \frac{w_i'^6}{6\kappa^4} & \text{if } |w_i'| \leq \kappa \\ \frac{\kappa^2}{6}, \end{cases}$$

for $i = 1, \ldots, N$, where κ is a tuning parameter controlling the degree of robustness to departures from normality. As $\kappa \to \infty$, Tukey's loss approaches the Gaussian negative log-likelihood, whereas small values for κ result in greater robustness to outliers. There are a variety of ways to choose κ, but here we fix $\kappa = 5$. Jewson and Rossell (2022)

describe a way of choosing κ and other loss parameters using a so-called $\mathcal{H}$-posterior based on the Hyvärinen score, and also consider model choice for loss functions, but these directions are not pursued here. For Tukey's loss, the corresponding loss likelihood is not integrable in w, so it does not correspond to any probabilistic model.

Our generalized Bayesian analysis requires a choice of the learning rates ν and ν' as discussed in Section 3.2. Recall that ν calibrates the module 1 loss to the prior, and ν' can be thought of as calibrating the module 2 loss to the conditional prior for $\eta|\varphi$. Since we use the original probabilistic specification for module 1, we choose the learning rate ν to be 1, and the generalized Bayes and conventional cut posterior densities for φ are the same. To choose ν', we use the method discussed in Section 3.2. However, noting that only the parameters β appear in the loss function and not the prior hyperparameter ψ, we calibrate ν' by considering matching the Fisher information number for updates for β asymptotically with ψ fixed, for loss likelihood bootstrap and generalized Bayes. Since the matching is done asymptotically, the choice of ψ makes no difference to the value of ν' obtained. To estimate the matrix Ψ_{22} in estimating ν' in Section 3.2, we used a Bayesian bootstrap applied to the original data groups, since it is not possible otherwise to estimate Ψ_{22} from the data sufficient statistics. This is because there is no replication that can be used, with β_i appearing only in the model for w_i. The learning rate obtained for the second module for the analysis was $\nu' = 3.3$.

Figure 6 (right) compares the posterior distributions of φ_1 for the generalized Bayes posterior and the cut posterior distributions. Once again, the boxplots are for 1,000 posterior samples, and the horizontal line shows the true value. The cut posterior is the same as for the conventional posterior for the parametric model, as we are still using the negative log-likelihood as the loss for module 1. We see that now the cut and full posterior inferences are consistent with each other, so that the Tukey's loss, which accommodates heavy-tailed data, resolves the conflict between different parts of the model.

For computations in this example, we used the `rstan` package (Carpenter et al., 2017) for both the conventional and generalized posterior densities. We ran four chains with 1000 iterations burn-in and 4000 sampling iterations, thinning the output so that 1000 samples are retained. The cut posterior density for φ_1^2 is inverse gamma, and was sampled directly to get $1,000$ cut posterior samples for φ_1.

5 Discussion

This paper combines generalized posterior inference with cutting feedback methods for flexible Bayesian modular inference. Starting from a parametric model, we suggest to replace the negative log likelihood of unreliable modules with different choices of a loss function to resolve any incompatibility between cut and full posterior inferences. We have also studied the large sample behaviour of the generalized cut posterior distribution from a novel conditional perspective. Our main theoretical result describes the asymptotic behaviour of the conditional (cut) posterior distribution of a module's parameters given parameters in other modules, formally justifying a type of conditional

Laplace approximation. This conditional perspective allows the approximation to depend on the module's parameters that are being conditioned on, which is in contrast to the conventional joint (Gaussian) Laplace approximation whose conditional covariance matrix is fixed. We describe how this large sample approximation is useful for computing diagnostics, describing uncertainty propagation between modules, as well as for the efficient implementation of a new approach to semi-modular inference.

In the framework for modular inference that we have developed, the loss function is a sum of loss functions associated with different modules. We considered calibrating the different component loss functions in one example, but more research is needed on the best way to do this for different purposes. With a single loss function, there are different methods of calibrating the loss to the prior, and the best method to use may depend on the goals of the analysis. A similar remark applies in generalized Bayesian modular inference. An anonymous referee has also asked about the connections with the "restricted likelihood" approach to dealing with misspecification, discussed recently in Lewis et al. (2021). Restricted likelihood reduces the data to an insufficient summary statistic, to discard information that cannot be matched under the assumed model. The method can be implemented computationally using likelihood-free inference algorithms, and modular inference has been considered in this context by Chakraborty et al. (2023).

Another anonymous referee raises interesting questions about what is lost when a conventional Bayesian analysis is replaced with a generalized Bayesian one based on a loss likelihood. Although we are convinced that generalized Bayesian methods are helpful, we are uncertain about how they will be used in the future as these techniques become more widely known. Below we discuss three uses for generalized Bayesian inference in order of increasing controversy from a conventional Bayesian point of view.

For cutting feedback methods based on a fully specified probabilistic model, generalized Bayesian justifications for them exist (Nicholls et al., 2022), and are one form of support for their use. However, there are other justifications too from a conventional Bayesian perspective. For example, a comparison of a cut and conventional posterior might be considered as a kind of model check. Yu et al. (2023) explain why certain calibrated comparisons between cut and full posterior densities based on the Kullback-Leibler divergence have the logical features one would require for some model checking tasks.

A deeper use of generalized Bayesian reasoning involves replacing some log-likelihood components with general loss functions which are not derived from any probabilistic model. An analyst might still think of these methods as having diagnostic value in a conventional Bayesian framework. If a probabilistic specification of some module is found to be flawed, and if determining an alternative probabilistic model is difficult, then the use of a loss likelihood could help in deciding how to expand the current model. If a generalized Bayesian analysis with a certain loss likelihood reduces the tension between cut and full posterior inferences compared to the original probabilistic model, this may suggest ways in which the initial model can be extended to a new probabilistic model that might require significant effort and thought to specify. The simpler loss-likelihood analysis might tell us whether it is worth the effort. Sometimes the loss function can be obtained by modifying the negative log likelihood loss for the initial parametric model;

our random effects example illustrates this, where setting $\kappa = \infty$ in the Tukey loss recovers the negative log likelihood for the initially assumed normal model.

Taking an even more permissive view of Bayesian reasoning as something useful in predictive problems where no scientific inference is involved, the main appeal of generalized Bayes methods might be their attractiveness for dealing with nuisance parameters by integration, rather than through optimization, and there may also be a strong preference for using a particular loss function in some applications (see, e.g., Loaiza-Maya et al. (2021) for a specific example). Here elements of Bayesian thinking are being used to achieve improved prediction or to target a meaningful parameter based on the chosen loss function. Although there is a precise and limited sense in which the generalized Bayesian methods are coherent, in other respects their behaviour might deviate from that of conventional Bayesian inference.

Acknowledgments

We thank the editorial team for their help in greatly improving the manuscript.

Funding

David Nott's research was supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 2 (MOE-T2EP20123-0009), and he is affiliated with the Institute of Operations Research and Analytics at the National University of Singapore. David Frazier was supported by the Australian Research Council's Discovery Early Career Researcher Award funding scheme (DE200101070).

Supplementary Material

Supplementary Material: Cutting feedback and modularized analyses in generalized Bayesian inference.
(DOI: 10.1214/24-BA1448SUPP; .pdf). The supplementary material contains: additional discussion of cutting feedback and modular inference, discussion on the assumptions used in the main text to obtain the stated results, proofs of all results stated in the main text, detailed algorithms for conditional credible interval construction and MCMC sampling for marginal SMI, and an empirical comparison of marginal SMI with the SMI method of Carmona and Nicholls (2020) in an example.

References

Alquier, P. (2021). "User-friendly introduction to PAC-Bayes bounds." arXiv:2110.11216. 7

Beaton, A. E. and Tukey, J. W. (1974). "The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data." *Technometrics*, 16(2): 147–185. 23

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). "A general framework for updating belief distributions." *Journal of the Royal Statistical Society. Series B, Statistical*

methodology, 78(5): 1103. MR3557191. doi: https://doi.org/10.1111/rssb.12158. 2, 6, 7, 9

Brillinger, D. R. (1969). "The calculation of cumulants via conditioning." *Annals of the Institute of Statistical Mathematics*, 21(1): 215–218. 20

Carmona, C. and Nicholls, G. (2020). "Semi-Modular Inference: enhanced learning in multi-modular models by tempering the influence of components." In *International Conference on Artificial Intelligence and Statistics*, 4226–4235. PMLR. 2, 9, 15, 16, 21, 26

Carmona, C. and Nicholls, G. (2022). "Scalable semi-modular inference with variational meta-posteriors." *arXiv preprint arXiv:2204.00296*. 10, 15

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). "Stan: A Probabilistic Programming Language." *Journal of Statistical Software*, 76(1): 1–32. Number: 1. URL https://www.jstatsoft.org/index.php/jss/article/view/v076i01 17, 24

Chakraborty, A., Nott, D. J., Drovandi, C. C., Frazier, D. T., and Sisson, S. A. (2023). "Modularized Bayesian analyses and cutting feedback in likelihood-free inference." *Statistics and Computing*, 33(1): 33. MR4537429. doi: https://doi.org/10.1007/s11222-023-10207-5. 25

Chernozhukov, V. and Hong, H. (2003). "An MCMC approach to classical estimation." *Journal of Econometrics*, 115(2): 293–346. MR1984779. doi: https://doi.org/10.1016/S0304-4076(03)00100-3. 13

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. 23

Frazier, D. T. and Nott, D. J. (2024). "Supplementary Material for "Cutting feedback and modularized analyses in generalized Bayesian inference"." *Bayesian Analysis*. doi: https://doi.org/10.1214/24-BA1448SUPP. 4

Grünwald, P. (2012). "The safe Bayesian: learning the learning rate via the mixability gap." In *Algorithmic Learning Theory: 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings 23*, 169–183. Springer. MR3042889. doi: https://doi.org/10.1007/978-3-642-34106-9_16. 1

Grünwald, P. and Van Ommen, T. (2017). "Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it." *Bayesian Analysis*, 12(4): 1069–1103. MR3724979. doi: https://doi.org/10.1214/17-BA1085. 6

Hahn, P. R. and Herren, A. (2023). "Comment on "Causal Inference Under Mis-Specification: Adjustment Based on the Propensity Score (with Discussion)," by David A. Stephens. Widemberg S. Nobre. Erica E. M. Moodie. Alexandra M. Schmidt." *Bayesian Analysis*, 18(2): 639 – 694. MR4609024. doi: https://doi.org/10.1214/22-ba1322. 4

Holmes, C. C. and Walker, S. G. (2017). "Assigning a value to a power likelihood in

a general Bayesian model." *Biometrika*, 104(2): 497–503. MR3698270. doi: https://doi.org/10.1093/biomet/asx010. 7, 8, 13

Jacob, P. E., Murray, L. M., Holmes, C. C., and Robert, C. P. (2017). "Better together? Statistical learning in models made of modules." *arXiv preprint arXiv:1708.08719*. 2, 10

Jacob, P. E., O'Leary, J., and Atchadé, Y. F. (2020). "Unbiased Markov chain Monte Carlo methods with couplings (with discussion)." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3): 543–600. MR4112777. doi: https://doi.org/10.1111/rssb.12336. 10

Jeffrey, R. C. (1965). *The logic of decision*. McGraw-Hill Book Co., New York-Toronto-London. MR0233448. 4

Jewson, J. and Rossell, D. (2022). "General Bayesian Loss Function Selection and the use of Improper Models." *Journal of the Royal Statistical Society Series B*, 84(5): 1640–1665. MR4515553. 23

Kleijn, B. J. and van der Vaart, A. W. (2012). "The Bernstein-von-Mises theorem under misspecification." *Electronic Journal of Statistics*, 6: 354–381. MR2988412. doi: https://doi.org/10.1214/12-EJS675. 6

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media. MR1639875. 11

Lewis, J. R., MacEachern, S. N., and Lee, Y. (2021). "Bayesian Restricted Likelihood Methods: Conditioning on Insufficient Statistics in Bayesian Regression." *Bayesian Analysis*, 1(1): 1–38. MR4381137. doi: https://doi.org/10.1214/21-BA1257. 25

Liu, F., Bayarri, M. J., and Berger, J. O. (2009). "Modularization in Bayesian analysis, with emphasis on analysis of computer models." *Bayesian Analysis*, 4(1): 119–150. MR2486241. doi: https://doi.org/10.1214/09-BA404. 2, 21, 22

Liu, Y. and Goudie, R. J. B. (2022a). "A General Framework for Cutting Feedback within Modularized Bayesian Inference." *arXiv preprint arXiv:2211.03274*. 4, 6

Liu, Y. and Goudie, R. J. B. (2022b). "Stochastic Approximation Cut Algorithm for Inference in Modularized Bayesian Models." *Statistics and Computing*, 32(7): 1–15. MR4350200. doi: https://doi.org/10.1007/s11222-021-10070-2. 10

Loaiza-Maya, R., Martin, G. M., and Frazier, D. T. (2021). "Focused Bayesian prediction." *Journal of Applied Econometrics*, 36(5): 517–543. MR4309597. doi: https://doi.org/10.1002/jae.2810. 26

Lunn, D., Best, N., Spiegelhalter, D., Graham, G., and Neuenschwander, B. (2009). "Combining MCMC with 'sequential' PKPD modelling." *Journal of Pharmacokinetics and Pharmacodynamics*, 36: 19–38. 5

Lyddon, S. P., Holmes, C. C., and Walker, S. G. (2019). "General Bayesian updating and the loss-likelihood bootstrap." *Biometrika*, 106(2): 465–478. MR3949315. doi: https://doi.org/10.1093/biomet/asz006. 7, 8, 13, 14

Maucort-Boulch, D., Franceschi, S., and Plummer, M. (2008). "International correlation between human papillomavirus prevalence and cervical cancer incidence." *Cancer Epidemiology and Prevention Biomarkers*, 17(3): 717–720. 4

Miller, J. W. (2021). "Asymptotic normality, concentration, and coverage of generalized posteriors." *Journal of Machine Learning Research*, 22(168): 1–53. MR4318524. 11, 13

Nicholls, G. K., Lee, J. E., Wu, C.-H., and Carmona, C. U. (2022). "Valid belief updates for prequentially additive loss functions arising in Semi-Modular Inference." *arXiv preprint arXiv:2201.09706*. 7, 9, 15, 16, 25

Plummer, M. (2015). "Cuts in Bayesian graphical models." *Statistics and Computing*, 25(1): 37–43. MR3304902. doi: https://doi.org/10.1007/s11222-014-9503-z. 4, 5, 10

Pompe, E. and Jacob, P. E. (2021). "Asymptotics of cut distributions and robust modular inference using Posterior Bootstrap." *arXiv preprint arXiv:2110.11149*. 2, 13

Portier, F. (2016). "On the asymptotics of Z-estimators indexed by the objective functions." *Electronic Journal of Statistics*, 10(1): 464 – 494. URL https://doi.org/10.1214/15-EJS1097 MR3466190. doi: https://doi.org/10.1214/15-EJS1097. 12

Syring, N. and Martin, R. (2018). "Calibrating general posterior credible regions." *Biometrika*, 106(2): 479–486. MR3949316. doi: https://doi.org/10.1093/biomet/asy054. 7

van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press. MR1652247. doi: https://doi.org/10.1017/CBO9780511802256. 13

Wedderburn, R. W. M. (1974). "Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method." *Biometrika*, 61(3): 439–447. MR0375592. doi: https://doi.org/10.1093/biomet/61.3.439. 17

Wu, P.-S. and Martin, R. (2020). "A comparison of learning rate selection methods in generalized Bayesian inference." *arXiv preprint arXiv:2012.11349*. MR4515727. doi: https://doi.org/10.1214/21-ba1302. 8

Yu, X., Nott, D. J., and Smith, M. S. (2023). "Variational Inference for Cutting Feedback in Misspecified Models." *Statistical Science*, 38(3): 490 – 509. URL https://doi.org/10.1214/23-STS886 MR4630957. doi: https://doi.org/10.1214/23-sts886. 4, 10, 18, 25

Zhang, T. (2006). "Information-theoretic upper and lower bounds for statistical estimation." *IEEE Transactions on Information Theory*, 52(4): 1307–1321. MR2241190. doi: https://doi.org/10.1109/TIT.2005.864439. 13