Bayesian Analysis (2024) **TBA**, Number TBA, pp. 1–31

On the Stability of General Bayesian Inference

Jack Jewson[*], Jim Q. Smith[†], and Chris Holmes[‡]

Abstract. We study the stability of posterior predictive inferences to the specification of the likelihood model and perturbations of the data generating process. In modern big data analyses, useful broad structural judgements may be elicited from the decision-maker but a level of interpolation is required to arrive at a likelihood model. As a result, an often computationally convenient canonical form is used in place of the decision-maker's true beliefs. Equally, in practice, observational datasets often contain unforeseen heterogeneities and recording errors and therefore do not necessarily correspond to how the data generating process was idealised by the decision-maker. Acknowledging such imprecisions, a faithful Bayesian analysis should ideally be stable across reasonable equivalence classes of such inputs. We are able to guarantee that traditional Bayesian updating provides stability across only a very strict class of likelihood models and data generating processes, requiring the decision-maker to elicit their beliefs and understand how the data was generated with an unreasonable degree of accuracy. On the other hand, a generalised Bayesian alternative using the β-divergence loss function is shown to be stable across practical and interpretable neighbourhoods, providing assurances that posterior inferences are not overly dependent on accidentally introduced spurious specifications or data collection errors. We illustrate this in linear regression, binary classification, and mixture modelling examples, showing that stable updating does not compromise the ability to learn about the data generating process. These stability results provide a compelling justification for using generalised Bayes to facilitate inference under simplified canonical models.

Keywords: stability, generalised Bayes, β-divergence, total variation, generalised linear models.

1 Introduction

Bayesian inferences are driven by the posterior distribution

$$\pi(\theta|y) = \frac{\pi(\theta)f(y;\theta)}{\int \pi(\theta)f(y;\theta)d\theta}, \tag{1}$$

which provides the provision to update parameter prior $\pi(\theta)$ using observed data $y = (y_1, \ldots, y_n) \in \mathcal{Y}^n$ assumed to have been generated according to likelihood $f(\cdot;\theta)$. The quality of such posterior inference depends on the specification of the prior, likelihood, and collection of the data. In controlled experimental environments where time is available to carefully consider such specifications, a posterior calculated in this way might be credible. However, modern applications often involve high-dimensional observational

[*] Department of Econometrics and Business Statistics, Monash University, Clayton VIC 3800, Australia, jack.jewson@monash.edu

[†] Department of Statistics, University of Warwick, Coventry, CV4 7AL, j.q.smith@warwick.ac.uk

[‡] Department of Statistics, University of Oxford, Oxford, OX1 3LB, chris.holmes@stats.ox.ac.uk

https://doi.org/10.1214/24-BA1502

data and are undertaken without the supervision of a trained statistician. In such scenarios, it is natural to question the quality of the specification of $\pi(\theta)$ and $f(\cdot;\theta)$ and the collection of y and we are therefore left to ponder to what extent posterior inference through (1) can be trusted. Much work has previously investigated the stability of (1) to the specification of the prior $\pi(\theta)$, therefore our focus here will be on the likelihood $f(\cdot;\theta)$ and data y.

The likelihood model captures the decision maker's (DM's) beliefs regarding the generation of data y. However, accurately formulating expert judgements as probability densities is difficult. Even for a well-trained expert, so doing requires many more probability specifications to be made at a much higher precision than is possible within the time constraints of a typical problem (Goldstein, 1990). This is not to say that an elicited model is useless. It is certainly possible to reliably elicit important broad structural information from domain experts. However, the resulting "*functional*" model $f(\cdot;\theta)$ generally involves some form of interpolating approximation of the DM's "*true*" beliefs. Doing so is not unreasonable. However, a consequence of such expediency is that the DM does not fully believe all the judgements expressed through their model $f(\cdot;\theta)$. A typical example of the above is when applied practitioners deploy computationally convenient canonical models, for which there are software and illustrative examples available, to their domain specific problems. While the broad structure of such models may be suitable across domains, it is the practitioner's familiarity with its form, its software implementation, or the platform on which it was published that often motivates its use for inference, rather than a careful consideration of how it captures beliefs about the new environment.

Similarly, the data were not necessarily collected exactly how the DM imagined when specifying their model. There may be unforeseen heterogeneities, outliers, or recording errors. Alternatively, the DM may be deploying someone else's carefully elicited model to an analogous but not necessarily exchangeable scenario.

Given the inevitable lack of specificity in f and how the data y were generated, a faithful Bayesian analysis should be able to demonstrate that it is not overly sensitive to their exact specification. Such stability would allow DMs to continue using familiar models in the knowledge that their arbitrary selection is not driving critical posterior inferences. This paper shows that the requirement for such stability necessitates the consideration of an updating rule different from (1).

Consider, for example, a situation where the DM's *true* beliefs for data y corresponds to a Student's-t distribution $t_5(y;\mu,\sigma^2)$ with 5 degrees of freedom. The top left of Figure 1 shows that the ubiquitous Gaussian likelihood, $\mathcal{N}(y;\mu,\sigma^2)$ captures many of the same judgements. The two likelihoods appear almost indistinguishable for all values of their shared μ and σ^2. Therefore, given finite time and introspection the DM may reasonably settle on the Gaussian likelihood as a suitable *functioning* approximation of their beliefs. However, the bottom left of Figure 1 shows that when updating according to (1) using the Gaussian model in place of the Student's-t results in very different posterior inferences. Equally, (1) is not stable to perturbations of the data either, as under the Gaussian model a small proportion of outliers moves the posterior inferences

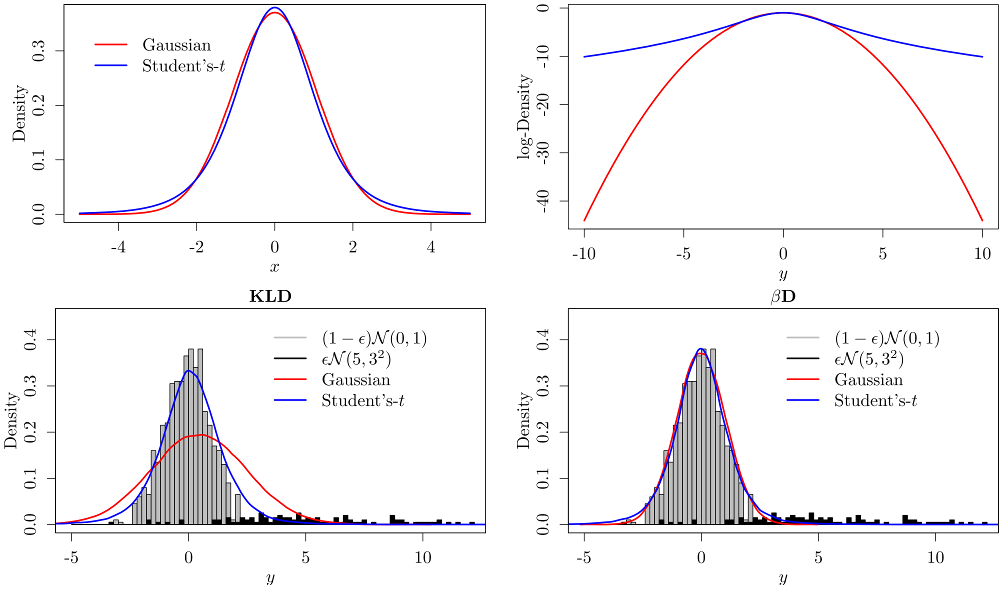

Figure 1: **Top:** Probability density function (PDF) and log-probability density function of a Gaussian $f_{\sigma^2_{adj}}(y;\theta) = \mathcal{N}\left(y;\mu,\sigma^2_{adj}\sigma^2\right)$ and a Student's-t $h_\nu(y;\eta) = t_\nu(y;\mu,\sigma^2)$ random variable, with $\mu = 0$, $\sigma^2 = 1$, $\nu = 5$ and $\sigma^2_{adj} = 1.16$. **Bottom:** The resulting posterior predictive distributions using traditional and βD-Bayes updating with $\beta = 1.22$ on $n = 1000$ observations from an ϵ-contamination model $g(y) = 0.9 \times \mathcal{N}\left(y;0,1\right) + 0.1 \times \mathcal{N}\left(y;5,3^2\right)$.

away from the uncontaminated part of the data generating process (DGP). Section 6.1 contains full details of this example.

We demonstrate that the instability observed in Figure 1 results from the fact that implicitly (1) learns about the parameter of the model minimising the Kullback-Leibler Divergence (KLD) between the data generating process (DGP) and the model, and, as a result, that stability can only be expected when the DM is sure that there is strong agreement between the tails of their functioning and true model specifications and between these and the data. The DM is highly unlikely to be sure of this a priori and therefore, under traditional Bayesian updating, it is left up to the DM to perform some *post hoc* sensitivity analysis to examine the impact their chosen model and particular features of the data had on the inference (see Box, 1980; Berger et al., 1994, and references within). However, such analyses are usually unsystematic and limited to the investigation of a small number of alternative judgements, models, or data points.

An alternative, motivated by the M-open world assumption that the model is misspecified for the DGP (Bernardo and Smith, 2001), is to use general Bayes (Bissiri et al., 2016) to update beliefs about model parameters minimising a divergence different from the KLD (Jewson et al., 2018). A particularly convenient alternative is the β-divergence

(βD) which has previously been motivated as providing inference that is robust to outliers (Basu et al., 1998; Ghosh and Basu, 2016) and desirable from a decision-making point of view (Jewson et al., 2018). In this paper, we extend the motivation for using βD-Bayes further, showing that its posterior predictive inferences are provably stable across an interpretable neighbourhood of likelihood models and DGPs. Such results demonstrate that the βD-Bayes facilitates the safe use of approximate canonical model specification for modern inference problems. As a result, we provide a rigorous justification for the use of βD-Bayes over traditional Bayesian inference that is not restricted to specific/model data scenarios.

While inferences should desirably be stable to small perturbations of f and y, they should still be sensitive the larger changes in order to extract useful inferences about the DGP. Importantly, the stability afforded to βD-Bayes inference does not compromise this. The βD-Bayes has the appealing property that if the model is correctly specified for the DGP, then the data generating parameter will be learned. There exists a growing literature that advocates using the βD for applied analyses (e.g. Knoblauch et al., 2018, 2022; Girardi et al., 2020; Sugasawa, 2020). This is further demonstrated in our experiments. For example, Figure 1 shows that as well as producing similar inference for the Gaussian and Student's-t likelihood models, the βD-Bayes inferences both capture the modal part of the observed data. Further, inferences must also not be overly dependent on the selection of hyperparameter, β, of the βD. We discuss methods to select β and demonstrate reasonable insensitivity to its selection.

Results regarding the stability of (1) have largely focused on the parameter prior. Gustafson and Wasserman (1995) proved that the total variation divergence (TVD) between two posteriors resulting from *functioning* and *true* priors in linear and geometric ϵ-contamination neighbourhoods divergences as $\epsilon \to 0$ at a rate exponential in the dimension of the parameter space. However, Smith and Rigat (2012) showed that the TVD between two posteriors converges to 0 provided the two priors under consideration are close as measured by the local De Robertis distance. Our first results provide analogies to these for the specification of the likelihood model. Gilboa and Schmeidler (1989); Whittle and Whittle (1990); Hansen and Sargent (2001a,b); Watson and Holmes (2016) consider the stability of optimal decision making and consider minimax decision across neighbourhoods of the posterior. However, they do not consider what perturbations of the inputs of (1) would leave a DM in such a neighbourhood *a posteriori*. Most similar to our work is Miller and Dunson (2018), which considers Bayesian updating conditioning on data arriving within a KLD ball of the observed data and results concerning 'global bias-robustness' to contaminating observations, for example of the kernel-Stein discrepancy posteriors of Matsubara et al. (2022). We consider stability to an interpretable neighbourhood of the data which as a special case contains the globally bias-robust contamination.

Bayes linear methods (Goldstein, 1999), which concern only the sub-collection of probabilities and expectations the DM considers themselves to be able to specify (Goldstein et al., 2006), is an alternative to (1) designed to be stable to interpolating approximations. We prefer, however, to adopt the general Bayesian paradigm in this analysis. Firstly, the general Bayesian paradigm includes traditional Bayesian updating as a special case and produces familiar posterior and predictive distributions. Secondly, linear

Bayes requires the elicitation of expectations and variances of unbounded quantities which are themselves unstable to small perturbations (see discussion on Goldstein and Wooff, 1994). Lastly, rather than trying to approximate their beliefs by a single model, the DM could consider several interpolating approximations and let the data guide any decision the they themselves have not able to make using methods such as penalised likelihood approaches (e.g. Akaike, 1973; Schwarz et al., 1978), Bayes' factors (Kass and Raftery, 1995) or Bayesian model averaging (Hoeting et al., 1999). In particular, Williamson and Goldstein (2015) propose methods for combining posterior beliefs across an equivalence class of analyses. However, such methods can be computationally burdensome to compute across even a finite class of models (e.g. Rossell et al., 2021) and can reasonably only consider a handful of models that might fit with the DM's beliefs, all of which contain some level of interpolating approximation.

The rest of the paper is organised as follows: Section 2 presents our inference paradigm, introducing general Bayesian updating (Bissiri et al., 2016), robustified inference with the βD, and defining how we will investigate posterior predictive stability. Section 3 presents our theoretical contributions surrounding the stability of Bayesian analyses to the choice of the likelihood function and Section 4 presents our results on the stability of inference to perturbations of the DGP. Proofs of all of our results are deferred to the supplementary material (Jewson et al., 2024). Section 5 discusses methods to set the β hyperparameter and Section 6 illustrates the stability of the βD-Bayes inference in continuous and binary regression examples from biostatistics and a mixture modelling astrophysics example, where stability is shown not to compromise the model's ability to learn about the DGP. Code to reproduce all of the examples in this paper can be found at `https://github.com/jejewson/stabilityGBI`.

2 A paradigm for inference and stability

2.1 General Bayesian inference

Under the assumption that the model used for inference $f(y;\theta)$ does not exactly capture the DM's beliefs, we find it appealing to adopt the general Bayesian perspective of inference. Bissiri et al. (2016) showed that the posterior update

$$\pi^{\ell}(\theta|y) = \frac{\pi(\theta)\exp\left(-w\sum_{i=1}^{n}\ell(\theta, y_i)\right)}{\int \pi(\theta)\exp\left(-w\sum_{i=1}^{n}\ell(\theta, y_i)\right)d\theta}, \tag{2}$$

provides a coherent means to update prior beliefs after observing data $y \sim g(\cdot)$ about parameter $\theta_g^{\ell} := \arg\min_{\theta\in\Theta}\int \ell(\theta, z)g(z)dz$ without requiring that θ index a model for the data generating density $g(\cdot)$.

The parameter $w > 0$ in (2) calibrates the loss with the prior to accounts for the fact that unlike the likelihood in (1), $\exp(-\ell(\theta, y_i))$ is no longer constrained to integrate to 1. Lyddon et al. (2018) set w to match the asymptotic information in the general Bayesian posterior to that of a sample from the 'loss-likelihood bootstrap', while Giummolè et al. (2019), building on the work of Ribatet et al. (2012), directly calibrate the curvature of the posterior to match that of the frequentist loss minimiser. A Bernstein von-Mises Theorem for generalised posterior (2) was proven in Miller (2021).

We focus on a subset of loss functions, known as proper scoring rules (Gneiting and Raftery, 2007), that depend upon the DM's likelihood model, allowing them to use this to encode their beliefs about the DGP. A scoring rule is proper if it is minimised in expectation at the density that generated the data. It, therefore, provides a means by which the DM can learn about the DGP. Under the log-score, $\ell(\theta, y) = -\log f(y;\theta)$ (2) collapses to (1). The parameter θ_g^ℓ associated with the log-score is the minimiser of the KLD between the distribution of the sample and the model (Berk et al., 1966). We therefore call updating using (1) KLD-Bayes. However, it is well known that minimising the log-score puts large importance on correctly capturing the tails of the data (Bernardo and Smith, 2001) and can have negative consequences for posterior decision making (Jewson et al., 2018). This is demonstrated in the bottom left of Figure 1.

2.2 βD-Bayes

An alternative proper scoring rule is the β-divergence loss (Basu et al., 1998), also known as the Tsallis Score (see e.g. Dawid et al., 2016)

$$\ell^{(\beta)}(y, f(\cdot;\theta)) = -\frac{1}{\beta-1}f(y;\theta)^{\beta-1} + \frac{1}{\beta}\int f(z;\theta)^{\beta}dz, \tag{3}$$

so called as $\arg\min_\theta \mathbb{E}_{y\sim g}\left[\ell^{(\beta)}(y, f(\cdot;\theta))\right] = \arg\min_\theta D_B^{(\beta)}(g||f(\cdot;\theta))$ where $D_B^{(\beta)}(g||f)$ is the β-divergence defined in Section A.1 of the supplementary material. We refer to updating using (2) and loss (3) as βD-Bayes. This was first used by Ghosh and Basu (2016) to produce a robustified Bayesian posterior (βD-Bayes) and has since been deployed for a variety of examples (e.g. Knoblauch et al., 2018, 2022; Girardi et al., 2020; Sugasawa, 2020).

The implicit robustness to outliers exhibited by the βD-Bayes is illustrated in the bottom right of Figure 1, where, unlike the KLD-Bayes, the βD-Bayes continues to captures the distribution of the majority of observations under outlier contamination. Jewson et al. (2018) argued that updating in a manner that is automatically robust to outliers, removes the burden on the DM to specify their beliefs in a way that is robust to the possible existence of occasional outliers. The results of the coming sections provide a formal rationale for adopting this methodology to provide stability to the canonical model choice and departures from the DGP.

While Bayesian inference has been proposed minimising several alternative divergences including the Hellinger divergence, α-divergence, and the TVD (e.g. Hooker and Vidyashankar, 2014; Jewson et al., 2018; Knoblauch and Vomfell, 2020) such methods require a non-parametric density estimate, prohibiting their use for high-dimensional problems with continuous data. We restrict our attention to local methods not requiring such an estimate and in particular to the βD and KLD. The γ-divergence (Fujisawa and Eguchi, 2008) has also been shown to produce robust inference without requiring a non-parametric density estimate (Hung et al., 2018; Knoblauch et al., 2022) and in general behaves very similarly, see Section B.1. There also exists scoring rules that tailor inference towards improved predictive performance (Loaiza-Maya et al., 2021). However, our focus here is on stably learning about the DGP in order to facilitate general decision-making without a specific prediction goal in mind.

2.3 Posterior predictive stability

We investigate the stability of general Bayesian posterior predictive distributions

$$m_f^D(y_{new}|y) = \int f(y_{new};\theta)\pi^D(\theta|y)d\theta, \tag{4}$$

for exchangeable observation $y_{new} \in \mathcal{Y}$ to the specification of the model f, and the DGP g. As a result, we focus on the stability of the posterior distribution for observables $y \in \mathcal{Y}$ to perturbations of the prior for observables, f, and generating distributions for these observables g.

From a decision-making perspective, the posterior predictive is often integrated over to calculate expected utilities, and therefore stable posterior predictive distributions correspond to stable decision-making. Predictive stability is also a more reasonable requirement than say posterior stability. The parameter posteriors for two distinct models/DGPs will generally converge in different places (e.g. Smith, 2007). However, divergent parameter posteriors do not necessarily imply divergent posterior predictives, as we show. Further, focusing on observables allows us to consider interesting cases of neighbouring models with nested parameter spaces (see Section 6.3).

3 Stability to the specification of the likelihood function

In this section, we investigate the stability of inference to the choice likelihood model for a given DGP. We consider that the DM is conducting inference using the functional likelihood model $\{f(\cdot;\theta);\theta \in \Theta \subseteq \mathbb{R}^{q_f}\}$ in place of their true beliefs $\{h(\cdot;\eta);\eta \in \mathcal{A} \subseteq \mathbb{R}^{q_h}\}$ for data $y \in \mathcal{Y}$. We assume that f is an approximation of h in the sense that it captures some of the main aspects of h that the DM has been able to faithfully specify, but interpolates between those in some arbitrary and convenient manner in a way that the DM does not necessarily believe. In this setting, a faithful posterior belief update should not diverge if f or h is used for inference. That is to say that posterior belief updating should be *stable* to the arbitrary specification of minor parts of the likelihood model not driven by the DM's beliefs. In this section we investigate sufficient conditions for how f can approximate h that would ensure such stability. For clarity of argument, we proceed under the assumption that the priors $\pi^D(\theta)$ and $\pi^D(\eta)$ are fixed. All technical conditions are stated in Section A.3 of the supplementary material.

3.1 The stability of the KLD-Bayes

Figure 1 demonstrated that there are examples of models that appear very similar to the naked eye (top left) but can result in substantially different KLD-Bayes posterior predictive inference when conditioning on the same data (bottom left). As a result, we first examine how f must approximate h in order to guarantee stable traditional Bayesian updating (KLD-Bayes). In particular, Lemma 1 investigates how stable the posterior predictive approximation of the DGP g, as measured by the KLD, is to changes in the likelihood model. Condition A.1, stated in Section A.3, requires that there exists mappings $I_f : \Theta \mapsto \mathcal{A}$ and $I_h : \mathcal{A} \mapsto \Theta$ such that the posterior density at parameter values η and θ such that $\text{KLD}(g||h(\cdot; I_f(\theta))) < \text{KLD}(g||h(\cdot;\eta))$ or $\text{KLD}(g||f(\cdot; I_h(\eta))) < \text{KLD}(g||f(\cdot;\theta))$ vanishes exponentially fast.

Lemma 1 (The stability in the posterior predictive approximation of the DGP of KLD-Bayes inference)**.** *For any two likelihood models* $\{f(\cdot;\theta) : \theta \in \Theta \subseteq \mathbb{R}^{q_f}\}$ *and* $\{h(\cdot;\eta) : \eta \in \mathcal{A} \subseteq \mathbb{R}^{q_h}\}$*, and* y*,* $\pi^{KLD}(\theta)$ *and* $\pi^{KLD}(\eta)$ *satisfying Condition A.1 for* $D =$ *KLD, we have that*

$$|KLD(g||m_f^{KLD}(\cdot|y)) - KLD(g||m_h^{KLD}(\cdot|y))| \le C^{KLD}(f,h,y) + \frac{1}{c} + T(f,h,y),$$

where $c := \min\{c_1, c_2\}$*,* $I_f : \Theta \mapsto \mathcal{A}$ *and* $I_h : \mathcal{A} \mapsto \Theta$ *are defined in Condition A.1 and*

$$\begin{aligned} C^{KLD}(f,h,y) :&= \max\left\{ \int KLD(g||f(\cdot;\theta))\pi^{KLD}(\theta|y)d\theta - KLD(g||m_f^{KLD}(\cdot|y)), \right. \\ &\left. \int KLD(g||h(\cdot;\eta))\pi^{KLD}(\eta|y)d\eta - KLD(g||m_h^{KLD}(\cdot|y)) \right\}, \\ T(f,h,y) :&= \max\left\{ \int\int g(y) \log \frac{f(y;\theta)}{h(y; I_f(\theta))} dy \pi^{KLD}(\theta|y)d\theta, \right. \\ &\left. \int\int g(y) \log \frac{h(y;\eta)}{f(y; I_h(\eta))} dy \pi^{KLD}(\eta|y)d\eta \right\}. \end{aligned} \tag{5}$$

As a result, sufficient conditions for KLD-Bayes to provide a stable approximation of the DGP g when using model f in place of model h are that terms $C^{\text{KLD}}(f,h,y)$, $\frac{1}{c}$, and $T(f,h,y)$ are small. The term $C^{\text{KLD}}(f,h,y)$ is the maximal difference between the KLD of the model from g in expectation under the posterior and the KLD of the posterior predictive from g under either model f or h. This is driven by how concentrated the KLD-Bayes posteriors are. Similarly, the term c is the minimal rate associated with Condition A.1. This is driven by how quickly the posteriors concentrate around their KLD minimising parameters. We use Lemma 1 to examine what f must correctly capture about h in order that inference with both achieves similar approximations of the DGP. We therefore investigate some properties of $T(f,h,y)$. Without loss of generality assume that the second term in (5) is the largest. Then, $T(f,h,y)$ being small requires that

$$|\log(h(\cdot;\eta)) - \log(f(\cdot; I_h(\eta)))| \tag{6}$$

is small in regions where $g(\cdot)$ and $\pi^{\text{KLD}}(\eta|y)$ have density. Without knowledge of g, this requires that (6) be small everywhere for all η.

Lemma 1 establishes that if a DM can ensure that (6) is small everywhere then they can use the approximate model f in place of their true beliefs h and be safe in the knowledge that their KLD-Bayes posterior inferences cannot be driven by some arbitrary part of the approximate model. However, this requires the DM to be confident in the accuracy of the probability statements made by f on the log scale. Logarithms act to inflate the magnitude of small numbers thus ensuring that $|\log(h(\cdot;\eta)) - \log(f(\cdot;I_h(\eta)))|$ is small requires that f and h are increasingly similar as their values decrease. This requires the DM to be more and more confident of the accuracy of the probability statements made by f further and further into the tails, something that is known to already be very difficult for low dimensional problems (Winkler and Murphy, 1968; O'Hagan et al., 2006), and becomes increasingly difficult as the dimension of the observation space increases. Tail probabilities definitionally correspond to surprising events and are thus harder to specify accurately. We therefore conclude that this is not a reasonable requirement to ask of any DM.

While Lemma 1 does not indicate the tightness of this bound, the example presented in Figure 1 demonstrates the importance of $T(f, h, y)$ being small for stable inference. Figure 1 (top right) shows that while the Gaussian and Student's-t may appear similar when viewed on the natural scale the difference in their log probabilities is large in their tails. Figure 1, therefore provides an example of two likelihood models where (6) is not small everywhere and a DGP where the two models result in substantially different posterior beliefs (bottom left).

3.2 An interpretable neighbourhood of likelihood models

Motivated by the results of Section 3.1, we consider in what manner a DM might reasonably be able to accurately approximate their beliefs. Firstly, the total variation metric is defined as

$$\mathrm{TVD}(f(\cdot;\theta), h(\cdot;\eta)) := \sup_{Y\in\mathcal{Y}} |f(Y;\theta) - h(Y;\eta)| = \frac{1}{2}\int |f(y;\theta) - h(y;\eta)|\, dy. \qquad (7)$$

Then, a likelihood model f for data $y \in \mathcal{Y}$ is considered 'ϵ-close' to true belief distribution h if Definition 1 is satisfied.

Definition 1 (TVD neighbourhood of likelihood models)**.** *Likelihood models $f(\cdot;\theta)$ and $h(\cdot;\eta)$ for observable $y \in \mathcal{Y}$ are in the neighbourhood $\mathcal{N}_\epsilon^{TVD}$ of size ϵ if*

$$\begin{aligned}&\forall\theta \in \Theta, \exists\eta \in \mathcal{A} \text{ s.t. } TVD(f(\cdot;\theta), h(\cdot;\eta)) \leq \epsilon \quad \text{and}\\ &\forall\eta \in \mathcal{A}, \exists\theta \in \Theta \quad \text{s.t.} \quad TVD(f(\cdot;\theta), h(\cdot;\eta)) \leq \epsilon.\end{aligned}$$

Being in the neighbourhood $\mathcal{N}_\epsilon^{\mathrm{TVD}}$ entails the existence of functions $I_f : \Theta \mapsto \mathcal{A}$ and $I_h : \mathcal{A} \mapsto \Theta$ such that for all θ, $\mathrm{TVD}(f(\cdot;\theta), h(\cdot;I_f(\theta)))$ is small and for all η, $\mathrm{TVD}(h(\cdot;\eta), f(\cdot;I_h(\eta)))$ is also small. This means that there must exist mappings between the two parameter spaces such that for any parameter θ of f, mapping θ to η via I_f leaves $h(\cdot;\eta)$ TVD-close to $f(\cdot,\theta)$. Note that the symmetry of Definition 1 allows Θ and $\mathcal{A}$ to have different dimensions.

The motivation for using the TVD in Definition 1 is three-fold. Firstly, and foremost, the TVD is interpretable. For two likelihoods to be close in terms of TVD requires that the greatest difference in any of the probability statements made by the two likelihoods be small on the natural scale – where elicitation of probabilities and sample distributions usually takes place – and not the log scale. In a practical sense, two densities that appear 'close' to the naked eye will be close according to TVD, while this heuristic will not be sufficient for close log probability. As a result, we believe that specifying a model that is TVD close to their exact beliefs is a feasible and reasonable requirement of a DM.

Further, the TVD is natural in the context of Bayesian decision-making. Two densities that are close in terms of TVD will produce similar estimates of bounded expected utility, and thus lead to similar decisions. This has previously been discussed by Smith (2010) and Jewson et al. (2018). Therefore, a model that is TVD close to the DM's true beliefs will perform similarly from a decision-making perspective a priori. Lastly, the TVD neighbourhood contains ϵ-contamination models which are popular models for investigating prior stability (Gustafson and Wasserman, 1995) and outliers (e.g. Aitkin and Wilson, 1980).

While Pinsker's inequality (Pinkser, 1964) shows that (6) being small everywhere is a sufficient condition for Definition 1, it is not necessary to have close log probabilities to have close absolute probabilities. This is for example evidenced by Figure 1 where the TVD between the two densities is less than 0.05. Therefore, Definition 1 imposes a less strict requirement on the DM than (6) being small everywhere. Section A.6 demonstrates this by presenting an example where a shrinking TVD neighbourhood corresponds to an expanding KLD neighbourhood.

3.3 The stability of the βD-Bayes

In this section, we demonstrate that Definition 1 is a sufficient condition for stable updating under βD-Bayes. In addition to Condition A.1, the results in this section require Condition A.2, stated in Section A.3. This requires the boundedness over the space of data y of the essential supremum of DGP $g(\cdot)$ and models $f(\cdot;\theta)$ and $h(\cdot;\eta)$ for all values of their parameters θ and η. We need this condition to bound the βD and relate it to the TVD. In discrete models, this bound is always 1 and in continuous models such as Gaussian or Student's-t likelihood a bound can be achieved by lower bounding their scale. Theorem 1 provides an analogous result to Lemma 1 but shows that Definition 1 is sufficient for posterior predictive stability.

Theorem 1 (The stability in the posterior predictive approximation of two models to the DGP of βD-Bayes inference)**.** *Assume $1 < \beta \leq 2$ and that the two likelihood models $\{f(\cdot;\theta) : \theta \in \Theta \subseteq \mathbb{R}^{q_f}\}$ and $\{h(\cdot;\eta) : \eta \in \mathcal{A} \subseteq \mathbb{R}^{q_h}\}$ are such that $f, h \in \mathcal{N}_\epsilon^{TVD}$ for $\epsilon > 0$. Then provided Condition A.1 for $D = D_B^{(\beta)}$ is satisfied for y, $\pi^{(\beta)}(\theta)$ and $\pi^{(\beta)}(\eta)$ and there exists $M < \infty$ such that Condition A.2 holds, then*

$$|D_B^{(\beta)}(g||m_f^{(\beta)}(\cdot|y)) - D_B^{(\beta)}(g||m_h^{(\beta)}(\cdot|y))| \leq \frac{M^{\beta-1}(3\beta-2)}{\beta(\beta-1)}\epsilon + \frac{1}{c} + C^{(\beta)}(f,h,y),$$

where $c = \min\{c_1, c_2\}$ are defined in Condition A.1 and

$$C^{(\beta)}(f,h,y) := \max\left\{\int D_B^{(\beta)}(g||f(\cdot;\theta))\pi^{(\beta)}(\theta|y)d\theta - D_B^{(\beta)}(g||m_f^{(\beta)}(\cdot|y)),\right.$$

$$\int D_B^{(\beta)}(g||h(\cdot;\eta))\pi^{(\beta)}(\eta|y)d\eta - D_B^{(\beta)}(g||m_h^{(\beta)}(\cdot|y))\Bigg\}.$$

Additionally, Theorem 2 proves that as well as providing similar approximations to the DGP, the βD between the βD-Bayes posterior predictive distributions themselves can also be bounded.

Theorem 2 (Stability of the posterior predictive distributions of two models under the βD-Bayes inference)**.** *Assume $1 < \beta \leq 2$ and that the two likelihood models $\{f(\cdot;\theta) : \theta \in \Theta \subseteq \mathbb{R}^{q_f}\}$ and $\{h(\cdot;\eta) : \eta \in \mathcal{A} \subseteq \mathbb{R}^{q_h}\}$ are such that $f, h \in \mathcal{N}_\epsilon^{TVD}$ for $\epsilon > 0$. Then provided Condition A.1 for $D = D_B^{(\beta)}$ is satisfied for y, $\pi^{(\beta)}(\theta)$ and $\pi^{(\beta)}(\eta)$ and there exists $M < \infty$ such that Condition A.2 holds, then*

$$\begin{aligned}
D_B^{(\beta)}(m_f^{(\beta)}(\cdot|y)||m_h^{(\beta)}(\cdot|y)) \leq & \frac{M^{\beta-1}(3\beta-2)}{\beta(\beta-1)}\epsilon + \frac{1}{c_1}\\
& + 2\frac{M^{\beta-1}}{\beta-1}\int TVD(g, f(\cdot;\theta))\pi^{(\beta)}(\theta|y)d\theta,\\
D_B^{(\beta)}(m_h^{(\beta)}(\cdot|y)||m_f^{(\beta)}(\cdot|y)) \leq & \frac{M^{\beta-1}(3\beta-2)}{\beta(\beta-1)}\epsilon + \frac{1}{c_2}\\
& + 2\frac{M^{\beta-1}}{\beta-1}\int TVD(g, h(\cdot;\eta))\pi^{(\beta)}(\eta|y)d\eta,
\end{aligned}$$

where c_1 and c_2 are defined in Condition A.1.

Theorem 1 is directly analogous to Lemma 1 with terms $C^{(\beta)}(f, h, y)$ and c having the same interpretation. Corollaries A.1 and A.2 invoke the generalised Bayesian Bernstein von-Mises theorem (Miller, 2021) applied to the βD (Theorem A.1) to show that under very general regularity conditions $c \to \infty$ and $C^{(\beta)}(f, h, y) \to 0$ as $n \to \infty$. Therefore, Theorem 1 establishes that Definition 1 is sufficient for the βD-Bayes posterior predictive distributions under two models to produce similar approximations of DGP g. This allows a DM to proceed using a model that well approximates their beliefs, as measured by the TVD, and know that the imprecision of their beliefs specification cannot lead to substantially different posterior predictive beliefs.

Note that Theorem 1 and Lemma 1 are not directly comparable results. Lemma 1 upper bounds the difference in the KLD approximation of the DGP whereas Theorem 1 bounds the difference in βD approximation of the DGP. The two bounds themselves are therefore not directly comparable only the conditions leading to these bounds. The sufficient conditions for KLD-Bayes to be stable are impractical to satisfy, while the βD-Bayes is provably stable under reasonable conditions that might in practice be plausible to believe. We also do not expect Theorem 1 (and 2) to be tight, however they are not vacuous. Lemma A.8 demonstrates that under Condition A.2, the βD between any two densities is bounded by $\frac{M^{\beta-1}}{\beta-1}$, where M is an upper bound on the model's density or mass function (e.g. 1 if y is discrete). Therefore, provided $\frac{(3\beta-2)}{\beta}\epsilon < 1$, our results provide a tighter upper bound than a trivial bound on the divergence.

Theorem 2 demonstrates that βD-Bayes updating not only provides a stable approximation of the DGP (as in Theorem 1) but also that the βD between posterior predictives under two TVD close models can be bounded above. This result is slightly weaker than Theorem 1 because it requires the TVD between the model and the DGP to be small in expectation under the posterior. A strength of Theorem 1 is that it holds independent of how well either of the models approximates the DGP. Lastly, note that the choice of β away from 1 – the case corresponding to the KLD – is necessary for Theorems 1 and 2 to be practically useful as the bounds in all tend to infinity as $\beta \to 1$.

4 Stability to the data generating process

In this section, we investigate the stability of inference to perturbations of the DGP, the mechanism with which the data was generated. Consider that the DM is conducting inference using likelihood model $\{f(\cdot;\theta); \theta \in \Theta \subseteq \mathbb{R}^{q_f}\}$ that was faithfully elicited to capture beliefs about idealised DGP $g_1(\cdot)$. Whether this corresponds to their true beliefs or an approximation is not relevant to the argument below. Now suppose that, for unforeseen reasons, the data were actually generated according to $g_2(\cdot)$, a perturbation of $g_1(\cdot)$. A useful property to demonstrate would be that if, in some appropriate sense, such perturbations were small, inferences from what was actually observed g_2 would be similar to those had g_1 been observed. Therefore, we investigate sufficient conditions for how $g_2(\cdot)$ can differ from $g_1(\cdot)$ and this stability be achieved. Throughout we consider data sets $y_1 := (y_1, \ldots, y_{n_1}) \sim g_1$. and $y_2 := (y_1, \ldots, y_{n_2}) \sim g_2$. Although not necessary to our argument we assume for simplicity that $n_1 = n_2$. All regularity conditions for these results to hold and their proofs are given in Section A.3 of the supplementary material.

4.1 The stability of the KLD-Bayes

Figure 1 considered a case where the data were generated from $g_2(y) = 0.9 \times \mathcal{N}(y; 0, 1) + 0.1 \times \mathcal{N}\left(y; 5, 3^2\right)$ while the Gaussian model was an accurate representation of $g_1(y) = \mathcal{N}(y; 0, 1)$. Although the DGP was the same for 90% of the observations, KLD-Bayes posterior inference under g_2 differs considerable from what one obtains when fitting f to g_1 – see Figure B.2. Figure 1, therefore, demonstrates that there are examples of models and data where two largely similar DGPs result in substantially different posterior predictive inferences from the same model. As a result, we first investigate how g_2 can differ from g_1 in order to guarantee stable traditional Bayesian updating (KLD-Bayes) for f. Lemma 2 investigates how stable the posterior predictive approximation to the DGP as measured by the KLD is to changes in the DGP. Condition A.3, stated in Section A.3, is analogous to Condition A.1. This requires that the posterior density on regions of $\theta_1|g_1$ and $\theta_2|g_2$ that leaves $f(\cdot;\theta_1)$ KLD closer to g_2 than $f(\cdot;\theta_2)$ (i.e. $\text{KLD}(g_2||f(\cdot;\theta_1)) < \text{KLD}(g_2||f(\cdot;\theta_2)))$ or $f(\cdot;\theta_2)$ KLD closer to g_1 than $f(\cdot;\theta_1)$ (i.e. $\text{KLD}(g_1||f(\cdot;\theta_2)) < \text{KLD}(g_1||f(\cdot;\theta_1)))$ vanishes exponentially fast.

Lemma 2 (The stability in the posterior predictive approximation of two DGPs under the same model for KLD-Bayes inference)**.** *For likelihood model* $\{f(\cdot;\theta) : \theta \in \Theta \subseteq \mathbb{R}^{q_f}\}$

and data sets $y_1 := (y_1, \dots, y_{n_1}) \sim g_1$ *and* $y_2 := (y_1, \dots, y_{n_2}) \sim g_2$ *for* $n_1, n_2 > 0$*, if Condition A.3 holds for* $D = KLD$*,* y_1*,* y_2 *and* $\pi^{KLD}(\theta)$*, then*

$$|KLD(g_1||m_f^{KLD}(\cdot|y_1)) - KLD(g_2||m_f^{KLD}(\cdot|y_2))| \leq C^{KLD}(f, y_1, y_2) + \frac{1}{c} + T_1(g_1, g_2) \\ + T_2(f, y_1, y_2),$$

where $c := \min\{c_{\mathcal{S}^{(1)}}, c_{\mathcal{S}^{(2)}}\}$ *are defined in Condition A.3 and*

$$\begin{aligned} T_1(g_1, g_2) :&= \left| \int g_2(y) \log g_2(y) - g_1(y) \log g_1 dy \right|, \\ T_2(f, y_1, y_2) :&= \max \left\{ \int \int (g_1(y) - g_2(y)) \log f(y; \theta_1) dy \pi^{KLD}(\theta_1|y_1) d\theta_1, \right. \\ &\qquad \left. \int \int (g_2(y) - g_1(y)) \log f(y; \theta_2) dy \pi^{KLD}(\theta_2|y_2) d\theta_2 \right\}, \\ C^{KLD}(f, y_1, y_2) :&= \max \left\{ \int KLD(g_1||f(\cdot; \theta_1)) \pi^{KLD}(\theta_1|y_1) d\theta_1 - KLD(g_1||m_f^{KLD}(\cdot|y_1)), \right. \\ &\qquad \left. \int KLD(g_2||f(\cdot; \theta_2)) \pi^{KLD}(\theta_2|y_2) d\theta_2 - KLD(g_2||m_f^{KLD}(\cdot|y_2)) \right\}. \end{aligned}$$

So KLD-Bayes can certainly be ensured to provide stable approximation to the DGP when using model f to update beliefs on data from g_2 rather than g_1 if terms $C^{\text{KLD}}(f, y_1, y_2)$, $\frac{1}{c}$, $T_1(g_1, g_2)$ and $T_2(f, y_1, y_2)$ are small. The term $C^{\text{KLD}}(f, y_1, y_2)$ is the difference between the KLD of f from g_j in expectation under the posterior and the KLD of the posterior predictive of f from g_j maximised over $j = 1, 2$. This is driven by how concentrated the posteriors are. Similarly, the term c is the minimal rate associated with Condition A.3 and is driven by how quickly the posteriors concentrate around their KLD minimising parameters. We are interested in how g_2 must be close to g_1 for this bound to be small and therefore we focus on terms $T_1(g_1, g_2)$ and $T_2(f, y_1, y_2)$. Small $T_1(g_1, g_2)$ requires g_1 and g_2 to have similar Shannon entropy, a measure of the inherent randomness in the data, which seems a reasonable condition. However, as $f(y; \theta) \rightarrow 0$, $|\log f(y; \theta)| \rightarrow \infty$ therefore small $T_2(f, y_1, y_2)$ requires that $|g_1(y) - g_2(y)|$ gets smaller as $f(y; \theta)$ gets smaller for $\theta \sim \pi^{\text{KLD}}(\theta|y)$. That is to say that, $T_2(f, y_1, y_2)$ being small requires g_1 and g_2 to be increasingly close in their tails.

Such a requirement greatly reduces the generalisability of statistical modelling. The tails of the DGP correspond to rare observations and therefore the KLD-Bayes only generalises across DGPs with similar rare observations. Encountering such situations is not only unlikely in practice, but difficult for any DM to consider following our discussion in Section 3. This, for example, prohibits outlier ϵ-contamination models where the DGP for $(1 - \epsilon)\%$ of the data is the same across g_1 and g_2, but g_2 is contaminated with $\epsilon\%$ of outliers, as seen in Figure 1 and B.2. Such an example also provides an indication that although Lemma 2 is only an upper bound that is not necessarily tight, the absence of small $T_2(f, y_1, y_2)$ results in substantially different KLD-Bayes posterior predictive inferences.

4.2 A plausible neighbourhood of data generating process perturbations

The results of Section 4.1 motivated us to consider what perturbations of the DGP should we reasonably expect our posterior inferences to be stable to. Data generating processes g_1 and g_2 for data $y \in \mathcal{Y}$ are considered 'ϵ-close' if Definition 2 is satisfied.

Definition 2 (TVD Neighbourhood of data generating processes)**.** *Data generating processes g_1 and g_2 for observables $y \in \mathcal{Y}$ are in the neighbourhood $\mathcal{G}_\epsilon^{TVD}$ of size ϵ if $TVD(g_1, g_2) \leq \epsilon$*

Following (7), g_2 is a small perturbation of g_1 according to Definition 2 if the probability statements made by either differ by a maximum of ϵ. This gives equal weight to modal or tail discrepancies rather than overly focusing on having the same tails. As a result, the data generating distribution for two populations will be close if distributions of the majority of the observations are close, rather than the distributions for a few of the observations.

Further, Definition 2 contains ϵ-contamination neighbourhoods as considered by Matsubara et al. (2022) and demands that the data sets were generated under mechanisms that were absolutely close on the natural scale, rather than the log-score considered in the KLD neighbourhoods of Miller and Dunson (2018).

4.3 The stability of the βD

We now demonstrate that Definition 2 is a sufficient condition on g_1 and g_2 to bound the consequences of generalising βD-Bayes inference for f from g_1 to g_2. In addition to Condition A.3, the results in this section require Condition A.4 which is analogous to Condition A.2 and requires the bounding of the essential supremum over the space y of DGPs $g_1(\cdot)$ and $g_2(\cdot)$ and model $f(y;\theta)$ for all θ. Theorem 3 is an analogous result to Lemma 2 showing that Definition 2 is sufficient for posterior predictive stability

Theorem 3 (The stability in the posterior predictive approximation of two DGPs under the same model of βD-Bayes inference)**.** *Assume $1 < \beta \leq 2$, likelihood model $\{f(\cdot;\theta) : \theta \in \Theta \subseteq \mathbb{R}^{q_f}\}$ and that the two data sets $y_1 := (y_1, \ldots, y_{n_1}) \sim g_1$ and $y_2 := (y_1, \ldots, y_{n_2}) \sim g_2$ for $n_1, n_2 > 0$ are such that $\{g_1, g_2\} \in \mathcal{G}_\epsilon^{TVD}$. Then provided that Condition A.3 holds for $D = D_B^{(\beta)} y_1$, y_2 and $\pi^{(\beta)}(\theta)$ and there exists $M < \infty$ such Condition A.4 holds, then*

$$|D_B^{(\beta)}(g_1||m_f^{(\beta)}(\cdot|y_1)) - D_B^{(\beta)}(g_2||m_f^{(\beta)}(\cdot|y_2))| \leq \frac{M^{\beta-1}(\beta+2)}{\beta(\beta-1)}\epsilon + \frac{1}{c} + C^{(\beta)}(f, y_1, y_2),$$

where $c := \min\{c_{\mathcal{S}^{(1)}}, c_{\mathcal{S}^{(2)}}\}$ are defined in Condition A.4 and

$$\begin{aligned} C^{(\beta)}(f, y_1, y_2) : = \max\Bigg\{ &\int D_B^{(\beta)}(g_1||f(\cdot;\theta_1))\pi^{(\beta)}(\theta_1|y_1)d\theta_1 - D_B^{(\beta)}(g_1||m_f^{(\beta)}(\cdot|y_1)), \\ &\int D_B^{(\beta)}(g_2||f(\cdot;\theta_2))\pi^{(\beta)}(\theta_2|y_2)d\theta_2 - D_B^{(\beta)}(g_2||m_f^{(\beta)}(\cdot|y_2))\Bigg\}. \end{aligned}$$

Additionally, Theorem 4 proves that as well as providing similar approximations to the DGPs, the βD between the βD-Bayes posterior predictive distributions themselves can also be bounded.

Theorem 4 (The stability of the posterior predictive distribution under two DGPs of the βD-Bayes inference)**.** *Assume $1 < \beta \leq 2$, likelihood model $\{f(\cdot;\theta) : \theta \in \Theta \subseteq \mathbb{R}^{q_f}\}$ and that the two data sets $y_1 := (y_1, \ldots, y_{n_1}) \sim g_1$ and $y_2 := (y_1, \ldots, y_{n_2}) \sim g_2$ for $n_1, n_2 > 0$ are such that $\{g_1, g_2\} \in \mathcal{G}_\epsilon^{TVD}$. Then provided there exists $M < \infty$ such that Condition A.3 hold, Condition A.4 holds for $D = D_B^{(\beta)} y_1$, y_2 and $\pi^{(\beta)}(\theta)$, then*

$$\begin{aligned} D_B^{(\beta)}(m_f^{(\beta)}(\cdot|y_1)||m_f^{(\beta)}(\cdot|y_2)) &\leq 2\frac{M^{\beta-1}}{\beta-1}\epsilon + \frac{1}{c_{\mathcal{S}^{(1)}}} \\ &\quad + 2\frac{M^{\beta-1}}{\beta-1}\int TVD(g_1, f(\cdot;\theta_1))\pi^{(\beta)}(\theta_1|y_1)d\theta_1, \\ D_B^{(\beta)}(m_f^{(\beta)}(\cdot|y_2)||m_f^{(\beta)}(\cdot|y_1)) &\leq 2\frac{M^{\beta-1}}{\beta-1}\epsilon + \frac{1}{c_{\mathcal{S}^{(2)}}} \\ &\quad + 2\frac{M^{\beta-1}}{\beta-1}\int TVD(g_2, f(\cdot;\theta_2))\pi^{(\beta)}(\theta_2|y_2)d\theta_2, \end{aligned}$$

where $c_{\mathcal{S}^{(1)}}$ and $c_{\mathcal{S}^{(2)}}$ are defined in Condition A.4.

Theorems 3 and 4 are the analogous result to Theorems 1 and 2 respectively with terms $C^{(\beta)}(f, y_1, y_2)$ and c having the same interpretation. The value M is still easy to bound here and Corollaries A.1 and A.2 demonstrate that $C^{(\beta)}(f, y_1, y_2) \rightarrow 0$ and $\frac{1}{c} \rightarrow 0$, as $n \rightarrow \infty$. Therefore, Theorem's 3 and 4 establish that βD-Bayes inferences will be similar for any two DGPs satisfying Definition 2. This allows a DM to use their model and know that small unforeseen perturbations of the DGP will not drive substantially different posterior inference or alternatively use a default model or software from the literature and know that as long as their application area is similar, the model's generalisation will not overly affect posterior inferences.

Once again, we do not invoke a comparison of the bounds from Lemma 2 and Theorem 3, as they are bounding different quantities. Instead, we consider the strength of the sufficient conditions required for boundedness. KLD-Bayes requires strict conditions for the DGPs that are difficult for the DM to know would be satisfied, while the βD-Bayes is stable across a reasonable generalisation of the DGP.

5 Setting β

To implement βD-Bayes inference it is obviously necessary to choose an appropriate value of β. We briefly review a variety of methods that have been proposed to do this and comment on how these relate to the results of this paper. We then demonstrate that inference is not too sensitive to this choice provided that β is not chosen close to one.

5.1 Data driven methods

One approach is to try to learn a value for β that is 'optimal' in some sense for the DM's functioning likelihood model and the particular observed data at hand. Once the DM has decided upon model $f(\cdot;\theta)$, the value β regulates the trade-off between robustness and efficiency (e.g. Basu et al., 1998). Minimising the KLD ($\beta = 1$) provides the most efficient inference but is very sensitive to outliers. Increasing β away from 1 gains robustness to outliers at a cost to efficiency. Warwick and Jones (2005); Ghosh and Basu (2015); Basak et al. (2021) seek to optimise the robustness-efficiency trade-off by estimating β to minimise the mean squared error (MSE) of estimated model parameters, Toma and Broniatowski (2011); Kang and Lee (2014) minimise the maximum perturbation of the parameter estimates resulting from replacing one observation by the population estimated mean, and Yonekura and Sugasawa (2023) build on the work of Jewson and Rossell (2022) to estimate β minimising the Fisher's divergence to the DGP. The intuition behind these methods is that values estimated close to $\beta = 1$ indicate the model f is pretty well specified for the data at hand, while larger values indicate increasing large levels of possible model misspecification. We use the method of Yonekura and Sugasawa (2023) to learn the value of β in the example in Section 6.1.

While each of the above methods use different criteria with different interpretations to select β, the results of this paper provide a DM using one of these a unifying interpretation via an upper bound for how sensitive their posterior inferences could be to the specification of their model and the data. For example, a DM learning a larger value for β knows that the term $\frac{M^{\beta-1}}{(\beta-1)}\epsilon$ will be small for any value of ϵ and that even large departures from their model or data would result in similar inferences. This suggests they will only be able to learn slowly about the DGP. On the other hand, a DM estimating a very small β knows that their posterior inference may be very dependent on the precise model class they have chosen.

5.2 User specified

Other works have advocated for the subjective specification for the value of β (e.g. Jewson et al., 2018) and the results in this paper help to facilitate this by interpreting β in terms of the level of stability it brings. The results of this paper demonstrate that β controls the amount that imprecision in the specification of the model or data can be magnified into the posterior, allowing for the interpretation of β as a meta prior for the DM's confidence in their elicited model or data collection. The less confident they are, the greater β will need to be to prevent non-negligible *a posteriori* divergence. Eliciting β as such requires the DM to reflect on the value of ϵ associated with their beliefs or the quality of the data. For the neighbourhoods of Definition 1, this can be obtained by considering for a given set of parameters what the largest possible error in any of the probability statements could be, or for Definition 2 by considering the minimal proportion of a population that they believe is consistent with the DGP.

A default implementation, however, would be to set β such that $\frac{M^{\beta-1}(3\beta-2)}{\beta(\beta-1)} = U$ ensuring that the posterior predictive imprecision as measured by Theorem 1 is only

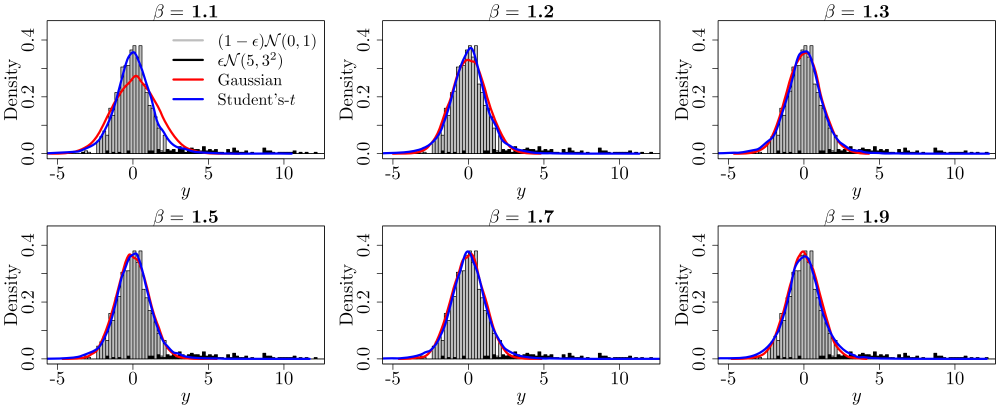

Figure 2: Posterior predictive distributions using βD-Bayes updating on $n = 1000$ observations from an ϵ-contamination model $g(y) = 0.9 \times \mathcal{N}(y; 0, 1) + 0.1 \times \mathcal{N}\left(y; 5, 3^2\right)$ for different values of β.

$U > 1$ times the level of prior imprecision ϵ. We demonstrate such an approach for the example in Section 6.2 for $U = 2$. Importantly, a DM could not hope to set β to provide maximal stability. Maximum stability, i.e. minimising the right-hand side of the bounds in Theorems 1 and 3 would set $\beta \to \infty$ and result in the posterior under any model and data collapsing to the prior, providing absolutely stable inference but not learning anything from the data. For minimally efficient learning to take place, posterior beliefs should not be closer, in the worst case, than the models were a priori.

5.3 Sensitivity

Finally, βD-Bayes inference appears not to be overly sensitive to the exact value of β. Figure 2 demonstrates that for the example introduced in Section 1, inference for the Gaussian and Student's-t models is almost identical for values of $\beta \geq 1.2$. Section B.1 provides further demonstration of this.

6 Experiments

6.1 Gaussian and Student's-t likelihood

We revisit the Gaussian and Student's-t example briefly introduced in Section 1. The likelihood models considered here are

$$f_{\sigma^2_{adj}}(y; \theta) := \mathcal{N}\left(y; \mu, \sigma^2 \times \sigma^2_{adj}\right) \text{ and } h_\nu(y; \eta) := \text{Student's} - t_\nu\left(y; \mu, \sigma^2\right). \tag{8}$$

Hyperparameters, $\nu = 5$ and $\sigma^2_{adj} = 1.16$ are fixed to match the quartiles of the two distributions for all μ and σ^2. These were inspired by O'Hagan (2012), who argued that for absolutely continuous probability distributions, it is only reasonable to ask an expert to make a judgement about the median and the quartiles of a distribution along with maybe a few specially selected features. This is justified as adequate as any two distributions with similar percentiles will look very similar, see for example Figure 1. However, Section 3.1 suggests that greater precision is required to ensure the stability of Bayes' rule updating. On the other hand, the likelihoods in (8) are contained in $\mathcal{N}^{\text{TVD}}_{0.043}$. We generated $n = 1000$ observations from the ϵ-contamination model $g(x) = 0.9 \times \mathcal{N}(y; 0, 1) + 0.1 \times \mathcal{N}\left(y; 5, 3^2\right)$ contained within the $\mathcal{G}^{\text{TVD}}_{0.1}$ neighbourhood of $\mathcal{N}(y; 0, 1)$. We then conducted Bayesian updating under the Gaussian and Student's-t likelihood using both Bayes' rule and the βD-Bayes under shared priors $\pi(\mu, \sigma^2) = \mathcal{N}\left(\mu; \mu_0, v_0\sigma^2\right) \mathcal{IG}(\sigma^2; a_0, b_0)$, with hyperparameters $(a_0 = 0.01, b_0 = 0.01, \mu_0 = 0, v_0 = 10)$. We used the method of Yonekura and Sugasawa (2023) to set $\beta = 1.22$ when using the Gaussian distribution and use the same value for the Student's-t. Figure 1 and Figure B.1, which plots the parameter posterior distributions for both models under both updating mechanisms, clearly demonstrate the stability of the βD-Bayes across these two models and the lack of stability of traditional Bayesian updating. Not only is the βD inference more stable across $\mathcal{N}^{\text{TVD}}_{\epsilon}$, the βD predictive better captures the majority of the DGP than either of the KLD-Bayes predictives. The capturing of the $\mathcal{N}(y; 0, 1)$ mode further illustrates the βD-Bayes' stability across neighbourhoods of the DGP.

While the predictive distributions and divergence measures are not available in closed form, we can use Markov chain Monte Carlo (MCMC) samples and adaptive quadrature (Piessens et al., 2012), respectively, to estimate the necessary quantities and verify that Lemma 1 and Theorems 1 and 2 hold for this example. For the KLD-Bayes $|\text{KLD}(g||m^{\text{KLD}}_f(\cdot|y)) - \text{KLD}(g||m^{\text{KLD}}_h(\cdot|y))|$ is estimated to be 0.220 which is smaller than our estimate of $T(f, h, y)$ which was 0.617. For the βD-Bayes $|D^{(\beta)}_B(g||m^{(\beta)}_f(\cdot|y)) - D^{(\beta)}_B(g||m^{(\beta)}_h(\cdot|y))|$ is estimated as 0.041, $D^{(\beta)}_B(m^{(\beta)}_f(\cdot|y)||m^{(\beta)}_h(\cdot|y))$ as 0.006, and $D^{(\beta)}_B(m^{(\beta)}_h(\cdot|y)||m^{(\beta)}_f(\cdot|y))$ as 0.010 which are all smaller than $\frac{M^{\beta-1}(3\beta-2)}{\beta(\beta-1)}\epsilon = 0.219$ for $M = 1/\sqrt{2\pi}$. Verification of Lemma 2 and Theorems 3 and 4 for this example is presented in Section B.1.

Figure 3 plots influence functions (West, 1984) for the KLD-Bayes and βD-Bayes under the Gaussian and Student's-t model. Influence functions are the gradient of the loss function evaluated at parameter estimates as a function of the observations and show the impact that observation had on the analysis. Under the βD-Bayes, the influence functions of the Gaussian and Student's-t likelihoods are closer for almost every y, illustrating the stability to the model, and additionally, the influence functions for both models under the βD-Bayes vary less with y, illustrating stability to the DGP.

DLD data

We consider a ribonucleic acid (RNA) sequencing data set from Yuan et al. (2016) measuring gene expression for $n = 192$ patients with different types of cancer. Rossell and Rubio (2018) studied the impact of 57 predictors on the expression of DLD, a gene that can perform several functions such as metabolism regulation. To illustrate our

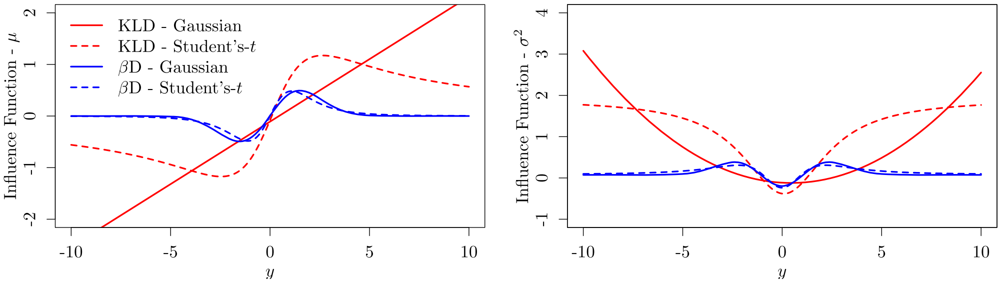

Figure 3: Influence functions for parameter μ and σ^2 of the Gaussian and Student's-t likelihood models under the KLD-Bayes and βD-Bayes with $\beta = 1.22$.

results, we selected the 15 variables with the 5 highest loadings in the first 3 principal components, and fitted regression models using the neighbouring models in (8) for the residuals. Section B.1 lists the selected variables. Once again, we used the method of Yonekura and Sugasawa (2023) to set $\beta = 1.34$ when using the Gaussian distribution, and use the same value for the Student's-t.

Figure 4 demonstrates that βD-Bayes produces more stable estimates of the fitted residuals (top-left), the estimated density of the residuals (top-right), parameter estimates (bottom-left), and posterior predictive density for the observed data (bottom-right) than the traditional Bayesian inference. Rossell and Rubio (2018) found evidence that this data is heavy-tailed, further demonstrated in Figure B.5, which caused the KLD-Bayes to estimate very different densities under the Gaussian and Student's-t model, while the βD-Bayes is stable to this feature of the data. Figure B.4 shows the fit of the models to the posterior mean estimates of the standardised residuals, showing that as well as being stable, the βD-Bayes produces good estimation around the mode of the DLD data under both models. Section B.1 considers a further regression example showing that even when one of the models under consideration is 'well-specified' for the data, the βD-Bayes inference continues to perform adequately.

6.2 Binary classification

Binary classification models predict $y \in \{0, 1\}$ from p-dimensional regressors X. The canonical model in such a setting is logistic regression where

$$P_{LR}(y = 1|X, \theta) = \frac{1}{1 + \exp\left(-X\theta\right)}, \quad P_{LR}(y = 0|X, \theta) = 1 - P_{LR}(Y = 1|X, \theta),$$

where $\theta \in \mathbb{R}^p$ are the regression parameters. Alternative, less ubiquitous models include probit regression, which uses an alternative generalised linear model (GLM) link function depending on the standard Gaussian cumulative distribution function (CDF) $\Phi(\cdot)$, 'heavier tailed' t-logistic regression (Ding and Vishwanathan, 2010; Ding et al., 2013) and a mixture type model that explicitly models the chance of mislabelling of the

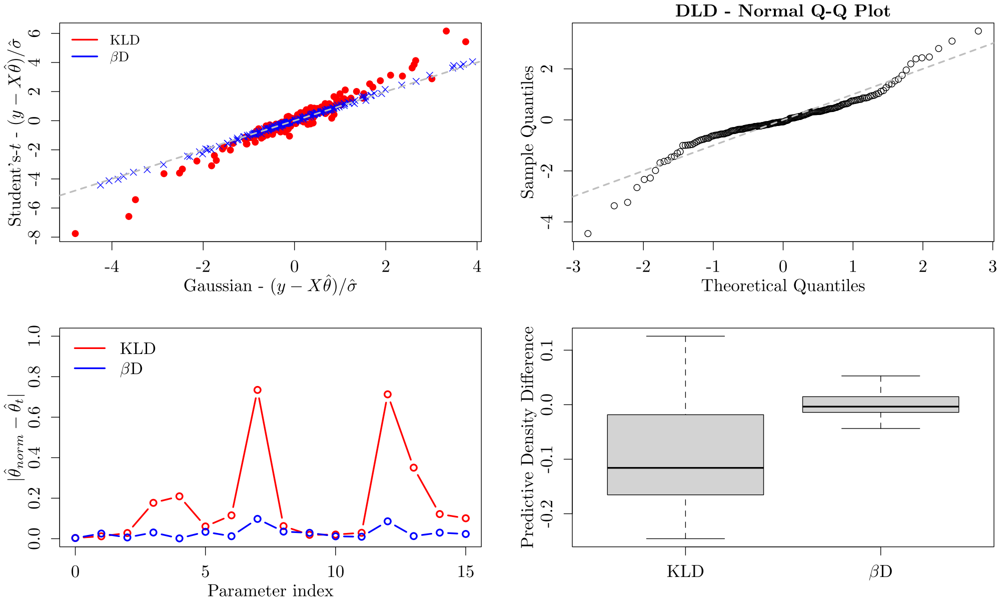

Figure 4: Posterior mean estimates of standardised residuals (**top left**), posterior predictive residual distributions (**top-right**), absolute difference in posterior mean parameter estimates (**bottom left**) and difference in posterior predictive densities of the observations (**bottom right**) under the Gaussian and Student's-t model of KLD-Bayes and βD-Bayes ($\beta = 1.34$) for the DLD data.

observed classes.

$$\begin{aligned} P_{PR}(y=1|X,\eta) &= \Phi(w_{PR}X\theta),\\ P_{tLR}(y=1|X,\eta) &= \exp_t((0.5w_{tLR}X\theta - G_t(w_{tLR}X\theta))),\\ P_{ML}(y=1|X,\eta) &= (1-\nu_1)P_{LR}(y=1|X,\theta) + \nu_0(1-P_{LR}(y=1|X,\theta)), \end{aligned}$$

where $0 < t < 2$, $0 < \nu_0, \nu_1 < 1$, '$\exp_t$' is the so-called t-exponential and G_t ensures that $P_{tLR}(y=1|X,\eta)$ is normalised, both are defined in Section B.3. Setting $t > 1$ results in heavier-tailed probabilities than the logistic model. For the probit and t-logistic models parameters θ are scalar multiples $w_{PR}, w_{tLR} \in \mathbb{R}$ of the logistic regression parameters $\theta \mapsto w\theta$. These are calculated in order to minimise the *a priori* TVD between the models and the logistic regression baseline according to $\mathcal{N}_\epsilon^{\text{TVD}}$ (see Section B.3). We upper bound ν_0 and ν_1 by 0.05 making $\epsilon = 0.05$ for these models. Figure 5 plots $P(y=1|X,\theta)$ as a function of $X\theta$ for all four models (left) and the TVD between each alternative model and the logistic regression (right), demonstrating that all four produce very similar binary probabilities.

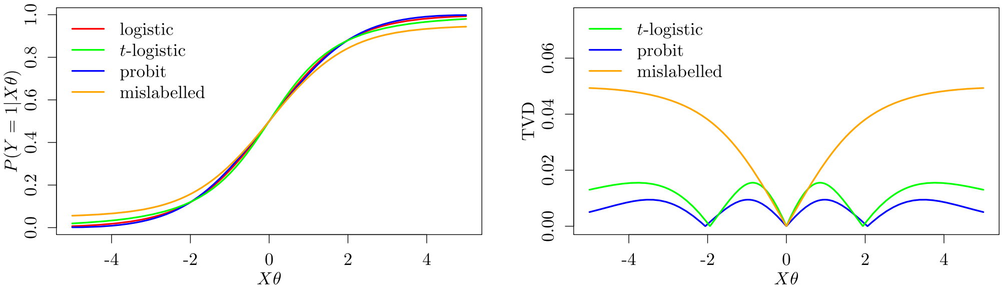

Figure 5: **Left**: $P(y=1|X,\theta)$ for logistic, probit, t-logistic and mislabelled models. **Right**: TVD between the logistic regression canonical model and the probit, t-logistic and mislabelled models. The θ parameters of the probit and t-logistic models are scalar multiplied in a fashion that minimise the TVD to the logistic regression.

Colon cancer dataset

To investigate the stability of posterior predictive inferences across the logistic, probit, t-logistic, and mislabelled binary regression models we consider the colon cancer dataset of Alon et al. (1999). The dataset contains the expression levels of 2000 genes from 40 tumours and 22 normal tissues and there is purportedly evidence that certain tissue samples may have been cross-contaminated (Tibshirani and Manning, 2013). Rather than consider the full 2000 genes we first run a frequentist least absolute shrinkage and selection operator (LASSO) procedure, estimating the hyperparameter via cross-validation, and focus our modelling only on the nine genes selected by this procedure. We understand that such post-model selection biases parameter estimates, but the stability of the predictive inference is our focus here. We set $\beta=2$ so that $U:=\frac{M^{\beta-1}(3\beta-2)}{\beta(\beta-1)}=2$ with $M=1$ as was proposed in Section 5.2.

Figure 6 compares the *a posteriori* TVD distance between the posterior predictive distributions for each observation with the *a priori* TVD distance between each of the models (top) and the difference between the posterior mean regression parameter estimates of the two models (bottom) under the KLD-Bayes and βD-Bayes. The stability of the βD-Bayes is once again demonstrated here. For almost every observation and every pair of models, the posterior predictive inference is as stable as it was *a priori*, while the KLD-Bayes inference is more often divergent. For the t-logistic and mislabelled models the predictive stability of the βD-Bayes also provides greater stability in the posterior mean parameter estimates.

6.3 Mixture modeling

An advantage of considering the stability of the distributions for observables rather than parameters is that it allows 'neighbouring' models to have different dimensions to their parameter space. For example, consider initial model $f(\cdot;\theta)$ and then 'neighbouring'

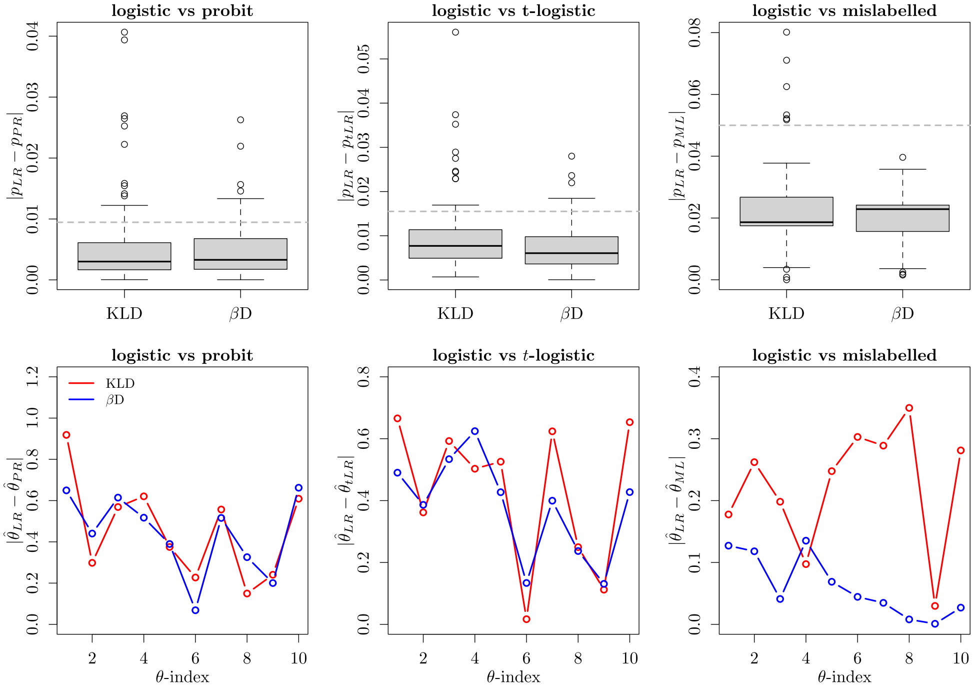

Figure 6: Colon Cancer Data. **Top**: TVD between the posterior predictive estimated probabilities for each observation of the probit (**left**), t-logistic (**centre**) and mislabelled (**right**) models and the canonical logistic regression under the KLD-Bayes and βD-Bayes ($\beta = 2$). The dotted line represented the *a priori* TVD distance between the models. **Bottom**: Absolute differences between posterior mean parameter estimates and those of the logistic regression.

model

$$h(\cdot;\eta) = (1-\omega) \times f(\cdot;\theta) + \omega \times h'(\cdot;\kappa),$$

for $\eta = \{\theta, \kappa, \omega\}$. Here, $h(\cdot;\eta)$ is a mixture model combining the likelihood model $f(\cdot;\theta)$, which could itself already be a mixture model, and some other density $h'(\cdot;\kappa)$ with additional parameters κ. For all $\theta \in \Theta$ and any $\kappa \in K$ we have that $\mathrm{TVD}(f(\cdot;\theta), h(\cdot;\{\theta,\kappa,\omega\})) < \omega$ and therefore a TVD neighbourhood can be defined by upper bounding ω.

Shapley galaxy dataset

We examine the Shapley galaxy dataset of Drinkwater et al. (2004), recording the velocities of 4215 galaxies in the Shapley supercluster, a large concentration of gravitationally-interacting galaxies; see Figure 7. The clustering tendency of galaxies continues to be a subject of interest in astronomy. Miller and Dunson (2018) investigate this data using Gaussian mixture models and use their coarsened posterior to select the number of mixture components, finding considerable instability in the number of estimated components K under different specifications of the coarsening parameter. See Cai et al. (2021) for further issues with estimating the number of components in mixture models.

We estimate Gaussian mixture models of the form

$$f(y;\theta) = \sum_{k=1}^{K} \omega_j \mathcal{N}(y;\mu_j,\sigma_j),$$

under the KLD-Bayes and βD-Bayes, considering number of components $K \in \{2,3,4,5,6\}$ and using the normal-inverse Wishart priors of Fúquene et al. (2019) (full details available in Section B.2). βD-Bayes inference for such one-dimensional mixture models is easy to implement using adaptive quadrature to approximate the necessary integral term $\frac{1}{\beta}\int h(z;\eta)^{\beta}dz$. We do not formally place any constraint on the estimation of ω_k, however, any model that estimates a component with small ω_k can be seen as a neighbour of a model with one fewer component.

Figure 7 shows the posterior predictive approximation to the histogram of the data of the Gaussian mixture models under the KLD-Bayes and βD-Bayes and Table 1 records the TVD between the posterior predictive distribution of recursively adding components to the model. The βD-Bayes inference for $\beta = 1.25$ and 1.5 is more stable to the addition of an extra component. In particular, for $K \geq 3$ the βD-Bayes inference stably estimates the biggest components of the data centered approximately at $5,000$ and $15,000$ km/s, while the KLD-Bayes produces very different inference for these modes depending on the number of clusters selected.

Method	$K = 2$ vs 3	$K = 3$ vs 4	$K = 4$ vs 5	$K = 5$ vs 6
KLD	0.27	0.12	0.13	0.08
βD ($\beta = 1.25$)	0.26	0.06	0.06	0.05
βD ($\beta = 1.5$)	0.22	0.04	0.07	0.02

Table 1: Total variation distances between posterior predictive distributions for different number of mixture components K under the KLD-Bayes and βD for $\beta = 1.25$ and 1.5.

7 Discussion

This paper investigated the posterior predictive stability of traditional Bayesian updating and a generalised Bayesian alternative minimising the βD. In practice, the model used for inference is usually a convenient and canonical interpolation of the broad belief statements made by the DM and the observed data was not necessarily collected in the manner the DM imagined. We proved that βD-Bayes inference is provably stable across a class of likelihood models and data generating processes whose probability statements are absolutely close, a TVD neighbourhood, by establishing bounds on how far their predictive inferences can diverge. On the other hand, our results require the DM to be

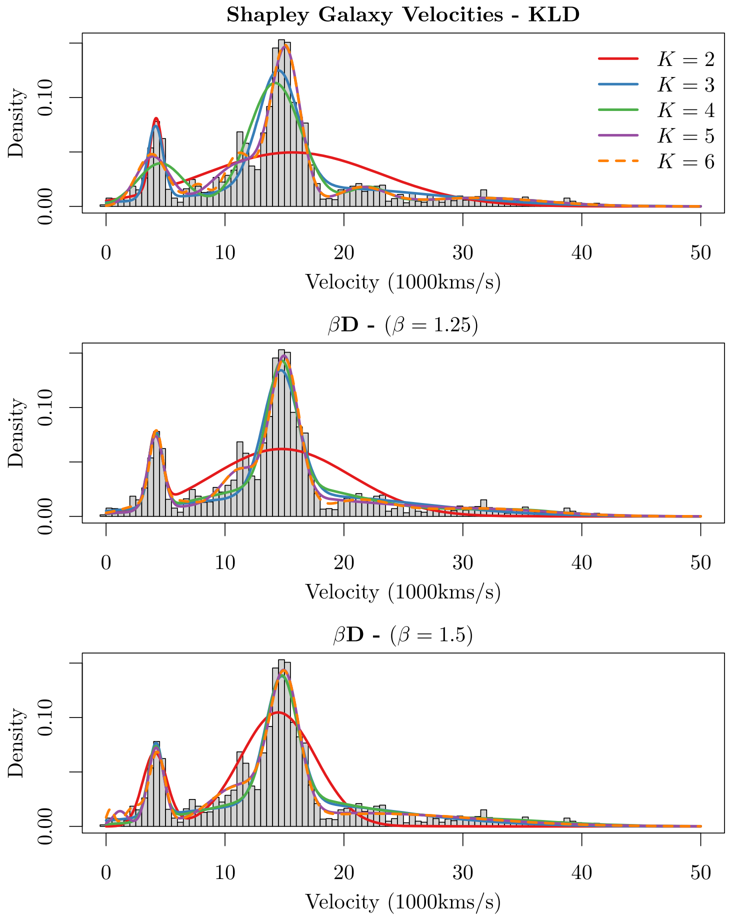

Figure 7: Shapley Galaxy Data: Histograms of the data, in units of 1,000 km/s, excluding a small amount of data extending in a tail up to 80,000 km/s, and posterior predictive distributions of the fitted Gaussian mixture models with $K = 2 - 6$ components under the KLD-Bayes (**top**), βD-Bayes with $\beta = 1.25$ (**middle**) and βD-Bayes with $\beta = 1.5$ (**bottom**).

sure about the tail properties of their beliefs and the DGP to guarantee stability for standard Bayesian inference.

The results of this paper simplify the process of belief elicitation for the βD-Bayes, bounding the *a posteriori* consequences for a given level of *a priori* inaccuracy, leaving the DM free to use the best guess approximation of their beliefs that they are most comfortable with, rather than switch to a less familiar model with better outlier rejection properties (O'Hagan, 1979). Such stability is achieved through a minimal amount of extra work compared with traditional Bayes' rule inference, and it provides a similarly recognisable output. We hope such results help to justify the increased use of the βD to make robust inferences in statistics and machine learning applications.

A key issue motivating the departure from standard Bayesian methods here is a lack of concordance between the likelihood model and the data. Such an issue can be attributed to either a failure of the modeller to think carefully enough about the DGP, or errors in data collection. However, we treat these results separately to exemplify two different manifestations of the instability of Bayes' rule.

The main limitation of our work is that we do not consider a universal measure of posterior predictive stability. Lemmas 1 and 2 use the KLD divergence to the DGP and Theorems 1 and 3 use the βD. It could of course be reasonably argued that the TVD should be used directly. However, the TVD is notoriously difficult to compute directly for large problems and is complicated by the intractability of the KLD and βD-Bayes posterior distributions. So instead, we focused on comparing the strength of the sufficient conditions required by each method for some measure of stability and used examples to indicate that these translate into meaningful differences.

The feasibility of βD-Bayes is dependent on the model likelihood being available in closed form – although robust general Bayesian method exists to deal with cases when it is not (Matsubara et al., 2022) – and the integral term in (3) being either available in closed form or fast to approximate accurately. These conditions are met by many standard models including exponential family and Student's-t models. When they are not then there are various methods available to make such calculations. For example, quadrature can be used for low-dimensional data. This integral is over the data not parameters and is therefore invariant to the parametrisation of the model. Further, one contribution of this paper is to show that the βD-Bayes allows a DM to use a canonical model, where this integral would be available, in place of their true beliefs and know that any approximate probabilistic specifications this might make will not have had an undue influence on their inference.

Future work could explore the applicability of such results in multivariate settings where belief specification and data collection are harder, and further investigate our KLD-Bayes results. While we argued when you could guarantee the stability of such methods, identifying for which statements KLD-Bayes is not stable would provide important and useful results to facilitate more focused belief elicitation.

To continue to facilitate the deployment of βD-Bayes methods in practice, more work is required to study and build upon existing methods to select β, particularly in high dimensions. While it is clear that considerable gains can be made over standard methods in certain scenarios, an adversarial analysis of the βD performance compared with its KLD-Bayes analogue would further motivate its wider applications. Other, interesting theoretical developments could seek to extend the posterior predictive stability to the stability of marginal posterior distributions in the case where there is an interpretable parameter of interest that is shared across models.

Acknowledgments

The authors would like to thank Danny Williamson, Christian Robert, and Sebastian Vollmer for their insightful discussions on the topics in this paper. We would also like to thank two anonymous reviewers, the Associated Editor and the Editor for their help in improving this paper.

Funding

JJ was partially funded by the Ayudas Fundación BBVA a Equipos de Investigación Científica 2017, the Government of Spain's Plan Nacional PGC2018-101643-B-I00, and a Juan de la Cierva Formación fellowship FJC2020-046348-I. CH was supported by the EPSRC Bayes4Health programme grant and both CH and JQS were supported by the Alan Turing Institute, UK.

Supplementary Material

Supplementary Material: On the Stability of General Bayesian Inference
(DOI: 10.1214/24-BA1502SUPP; .pdf). The supplementary material contains the proofs of the results of the paper as well as additional experimental results.

References

Aitkin, M. and Wilson, G. T. (1980). "Mixture models, outliers, and the EM algorithm." *Technometrics*, 22(3): 325–331. 10

Akaike, H. (1973). "Information theory and an extension of the maximum likelihood principle." In *Second International Symposium on Information Theory*, 267–281. MR0483125. 5

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays." *Proceedings of the National Academy of Sciences*, 96(12): 6745–6750. 21

Basak, S., Basu, A., and Jones, M. (2021). "On the 'optimal' density power divergence tuning parameter." *Journal of Applied Statistics*, 48(3): 536–556. MR4205987. doi: https://doi.org/10.1080/02664763.2020.1736524. 16

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. (1998). "Robust and efficient estimation by minimising a density power divergence." *Biometrika*, 85(3): 549–559. MR1665873. doi: https://doi.org/10.1093/biomet/85.3.549. 4, 6, 16

Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., De la Horra, J., Martín, J., Ríos-Insúa, D., Betrò, B., et al. (1994). "An overview of robust Bayesian analysis." *Test*, 3(1): 5–124. 3

Berk, R. H. et al. (1966). "Limiting behavior of posterior distributions when the model is incorrect." *The Annals of Mathematical Statistics*, 37(1): 51–58. MR0189176. doi: https://doi.org/10.1214/aoms/1177699477. 6

Bernardo, J. M. and Smith, A. F. (2001). *Bayesian theory*. John Wiley & Sons. MR1274699. doi: https://doi.org/10.1002/9780470316870. 3, 6

Bissiri, P., Holmes, C., and Walker, S. G. (2016). "A general framework for updating belief distributions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. MR3557191. doi: https://doi.org/10.1111/rssb.12158. 3, 5

Box, G. E. (1980). "Sampling and Bayes' inference in scientific modelling and robustness." *Journal of the Royal Statistical Society. Series A (General)*, 383–430. MR0603745. doi: https://doi.org/10.2307/2982063. 3

Cai, D., Campbell, T., and Broderick, T. (2021). "Finite mixture models do not reliably learn the number of components." In *International Conference on Machine Learning*, 1158–1169. PMLR. 23

Dawid, A. P., Musio, M., and Ventura, L. (2016). "Minimum scoring rule inference." *Scandinavian Journal of Statistics*, 43(1): 123–138. MR3466997. doi: https://doi.org/10.1111/sjos.12168. 6

Ding, N. and Vishwanathan, S. (2010). "t-Logistic regression." *Advances in Neural Information Processing Systems*, 23. 19

Ding, N., Vishwanathan, S., Warmuth, M., and Denchev, V. S. (2013). "T-logistic regression for binary and multiclass classification." *Technical Report*, 1–55. URL https://sites.google.com/site/ssnding/ 19

Drinkwater, M. J., Parker, Q. A., Proust, D., Slezak, E., and Quintana, H. (2004). "The large scale distribution of galaxies in the shapley supercluster." *Publications of the Astronomical Society of Australia*, 21(1): 89–96. 22

Fujisawa, H. and Eguchi, S. (2008). "Robust parameter estimation with a small bias against heavy contamination." *Journal of Multivariate Analysis*, 99(9): 2053–2081. MR2466551. doi: https://doi.org/10.1016/j.jmva.2008.02.004. 6

Fúquene, J., Steel, M., and Rossell, D. (2019). "On choosing mixture components via non-local priors." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(5): 809–837. MR4025398. doi: https://doi.org/10.1111/rssb.12333. 23

Ghosh, A. and Basu, A. (2015). "Robust estimation for non-homogeneous data and the selection of the optimal tuning parameter: the density power divergence approach." *Journal of Applied Statistics*, 42(9): 2056–2072. MR3371040. doi: https://doi.org/10.1080/02664763.2015.1016901. 16

Ghosh, A. and Basu, A. (2016). "Robust Bayes estimation using the density power divergence." *Annals of the Institute of Statistical Mathematics*, 68(2): 413–437. MR3464228. doi: https://doi.org/10.1007/s10463-014-0499-0. 4, 6

Gilboa, I. and Schmeidler, D. (1989). "Maxmin expected utility with non-unique prior." *Journal of Mathematical Economics*, 18(2): 141–153. MR1000102. doi: https://doi.org/10.1016/0304-4068(89)90018-9. 4

Girardi, P., Greco, L., Mameli, V., Musio, M., Racugno, W., Ruli, E., and Ventura, L. (2020). "Robust inference for non-linear regression models from the Tsallis score: Application to coronavirus disease 2019 contagion in Italy." *Stat*, 9(1): e309. MR4193414. doi: https://doi.org/10.1002/sta4.309. 4, 6

Giummolè, F., Mameli, V., Ruli, E., and Ventura, L. (2019). "Objective Bayesian inference with proper scoring rules." *Test*, 28(3): 728–755. MR3992136. doi: https://doi.org/10.1007/s11749-018-0597-z. 5

Gneiting, T. and Raftery, A. E. (2007). "Strictly proper scoring rules, prediction, and estimation." *Journal of the American Statistical Association*, 102(477): 359–378. MR2345548. doi: https://doi.org/10.1198/016214506000001437. 6

Goldstein, M. (1990). "Influence and belief adjustment." *Influence Diagrams, Belief Nets and Decision Analysis*, 143–174. 2

Goldstein, M. (1999). "Bayes linear analysis." *Wiley StatsRef: Statistics Reference Online*. 4

Goldstein, M. and Wooff, D. A. (1994). "Robustness measures for Bayes linear analyses." *Journal of statistical planning and inference*, 40(2-3): 261–277. MR1294983. doi: https://doi.org/10.1016/0378-3758(94)90125-2. 5

Goldstein, M. et al. (2006). "Subjective Bayesian analysis: principles and practice." *Bayesian Analysis*, 1(3): 403–420. MR2221272. doi: https://doi.org/10.1214/06-BA116. 4

Gustafson, P. and Wasserman, L. (1995). "Local sensitivity diagnostics for Bayesian inference." *The Annals of Statistics*, 23(6): 2153–2167. MR1389870. doi: https://doi.org/10.1214/aos/1034713652. 4, 10

Hansen, L. and Sargent, T. J. (2001b). "Robust control and model uncertainty." *American Economic Review*, 91(2): 60–66. 4

Hansen, L. P. and Sargent, T. J. (2001a). "Acknowledging misspecification in macroeconomic theory." *Review of Economic Dynamics*, 4(3): 519–535. MR0937259. 4

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). "Bayesian model averaging: a tutorial." *Statistical science*, 382–401. MR1765176. doi: https://doi.org/10.1214/ss/1009212519. 5

Hooker, G. and Vidyashankar, A. N. (2014). "Bayesian model robustness via disparities." *Test*, 23(3): 556–584. MR3252095. doi: https://doi.org/10.1007/s11749-014-0360-z. 6

Hung, H., Jou, Z.-Y., and Huang, S.-Y. (2018). "Robust mislabel logistic regression without modeling mislabel probabilities." *Biometrics*, 74(1): 145–154. MR3777935. doi: https://doi.org/10.1111/biom.12726. 6

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 (Statistical Methodology)*. MR4515553. 16

Jewson, J., Smith, J., and Holmes, C. (2018). "Principles of Bayesian inference using general divergence criteria." *Entropy*, 20(6): 442. MR3879894. doi: https://doi.org/10.3390/e20060442. 3, 4, 6, 10, 16

Jewson, J., Smith, J., and Holmes, C. (2024). "Supplementary Material: On the Stability of General Bayesian Inference." *Bayesian Analysis*. doi: https://doi.org/10.1214/24-BA1502SUPP. 5

Kang, J. and Lee, S. (2014). "Minimum density power divergence estimator for Poisson autoregressive models." *Computational Statistics & Data Analysis*, 80: 44–56. MR3240474. doi: https://doi.org/10.1016/j.csda.2014.06.009. 16

Kass, R. E. and Raftery, A. E. (1995). "Bayes factors." *Journal of the american statistical association*, 90(430): 773–795. MR3363402. doi: https://doi.org/10.1080/01621459.1995.10476572. 5

Knoblauch, J., Jewson, J., and Damoulas, T. (2018). "Doubly Robust Bayesian Inference for Non-Stationary Streaming Data using β-Divergences." In *Advances in Neural Information Processing Systems (NeurIPS)*, 64–75. 4, 6

Knoblauch, J., Jewson, J., and Damoulas, T. (2022). "An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference." *Journal of Machine Learning Research*, 23(132): 1–109. URL http://jmlr.org/papers/v23/19-1047.html MR4577084. 4, 6

Knoblauch, J. and Vomfell, L. (2020). "Robust Bayesian Inference for Discrete Outcomes with the Total Variation Distance." *arXiv preprint* arXiv:2010.13456. 6

Loaiza-Maya, R., Martin, G. M., and Frazier, D. T. (2021). "Focused Bayesian prediction." *Journal of Applied Econometrics* 517–543. MR4309597. doi: https://doi.org/10.1002/jae.2810. 7

Lyddon, S., Holmes, C., and Walker, S. (2018). "General Bayesian updating and the loss-likelihood bootstrap." *Biometrika*. MR3949315. doi: https://doi.org/10.1093/biomet/asz006. 5

Matsubara, T., and Knoblauch, J., and Briol, F-X and Oates, C. J. (2022). "Robust generalised Bayesian inference for intractable likelihoods" *Journal of the Royal Statistical Society. Series B (Methodological)*, 997–1022. MR4460583. doi: https://doi.org/10.1111/rssb.12500. 4, 14, 25

Miller, J. W. and Dunson, D. B. (2018). "Robust Bayesian inference via coarsening." *Journal of the American Statistical Association*, 1–13. MR4011766. doi: https://doi.org/10.1080/01621459.2018.1469995. 4, 14, 23

Miller, J. W. (2021). "Asymptotic normality, concentration, and coverage of generalized posteriors" *The Journal of Machine Learning Research*, 22(168):1–53. URL https://www.jmlr.org/papers/v22/20-469.html. MR4318524. 6, 11

O'Hagan, A. (1979). "On outlier rejection phenomena in Bayes inference." *Journal of the Royal Statistical Society. Series B (Methodological)*, 358–367. MR0557598. 24

O'Hagan, A. (2012). "Probabilistic uncertainty specification: Overview, elaboration techniques and their application to a mechanistic model of carbon flux." *Environmental Modelling & Software*, 36: 35–48. 17

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons. 9

Piessens, R., de Doncker-Kapenga, E., Überhuber, C. W., and Kahaner, D. K. (2012). "Quadpack: a subroutine package for automatic integration." MR0712135. doi: https://doi.org/10.1007/978-3-642-61786-7. 18

Pinkser, M. (1964). "Information and Information Stability of Random Variables and Processes." MR0213190. 10

Ribatet, M., Cooley, D., and Davison, A. C. (2012). "Bayesian inference from composite likelihoods, with an application to spatial extremes." *Statistica Sinica*, 813–845. MR2954363. 5

Rossell, D., Abril, O., and Bhattacharya, A. (2021). "Approximate Laplace approximations for scalable model selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4): 853–879. MR4320004. 5

Rossell, D. and Rubio, F. J. (2018). "Tractable bayesian variable selection: beyond normality." *Journal of the American Statistical Association*, 113(524): 1742–1758. MR3902243. doi: https://doi.org/10.1080/01621459.2017.1371025. 18, 19

Schwarz, G. et al. (1978). "Estimating the dimension of a model." *The Annals of Statistics*, 6(2): 461–464. MR0468014. 5

Smith, J. (2007). "Local robustness of Bayesian parametric inference and observed likelihoods." 7

Smith, J. and Rigat, F. (2012). "Isoseparation and robustness in finite parameter Bayesian inference." *Annals of the Institute of Statistical Mathematics*, 64: 495–519. MR2880867. doi: https://doi.org/10.1007/s10463-011-0334-9. 4

Smith, J. Q. (2010). *Bayesian decision analysis: principles and practice*. Cambridge University Press. MR2828346. doi: https://doi.org/10.1017/CBO9780511779237. 10

Sugasawa, S. (2020). "Robust empirical Bayes small area estimation with density power divergence." *Biometrika*, 107(2): 467–480. MR4126288. doi: https://doi.org/10.1093/biomet/asz075. 4, 6

Tibshirani, J. and Manning, C. D. (2013). "Robust logistic regression using shift parameters (long version)." *arXiv preprint* arXiv:1305.4987. MR2634010. 21

Toma, A. and Broniatowski, M. (2011). "Dual divergence estimators and tests: robustness results." *Journal of Multivariate Analysis*, 102(1): 20–36. MR2729417. doi: https://doi.org/10.1016/j.jmva.2010.07.010. 16

Warwick, J. and Jones, M. (2005). "Choosing a robustness tuning parameter." *Journal of*

Statistical Computation and Simulation, 75(7): 581–588. MR2162547. doi: https://doi.org/10.1080/00949650412331299120. 16

Watson, J. and Holmes, C. (2016). "Approximate models and robust decisions." *Statistical Science*, 31(4): 465–489. MR3598725. doi: https://doi.org/10.1214/16-STS592. 4

West, M. (1984). "Outlier models and prior distributions in Bayesian linear regression." *Journal of the Royal Statistical Society. Series B (Methodological)*, 431–439. MR0790630. 18

Whittle, P. and Whittle, P. R. (1990). *Risk-sensitive optimal control*, volume 20. Wiley New York. MR1093001. 4

Williamson, D. and Goldstein, M. (2015). "Posterior belief assessment: Extracting meaningful subjective judgements from Bayesian analyses with complex statistical models." *Bayesian Analysis*, 10(4): 877–908. MR3432243. doi: https://doi.org/10.1214/15-BA966SI. 5

Winkler, R. L. and Murphy, A. H. (1968). "Evaluation of subjective precipitation probability forecasts." In *Proceedings of the first national conference on statistical meteorology*, 148–157. American Meteorological Society Boston. 9

Yonekura, S. and Sugasawa, S. (2023). "Adaptation of the tuning parameter in general Bayesian inference with robust divergence." *Statistics and Computing*, 33, 39 MR4544336. doi: https://doi.org/10.1007/s11222-023-10205-7. 16, 18, 19

Yuan, T., Huang, X., Woodcock, M., Du, M., Dittmar, R., Wang, Y., Tsai, S., Kohli, M., Boardman, L., Patel, T., et al. (2016). "Plasma extracellular RNA profiles in healthy and cancer patients." *Scientific reports*, 6(1): 1–11. 18