On the stability of general Bayesian inference

Jack Jewson[1,*], Jim Q. Smith[1], and Chris Holmes[2]

[1] *University of Warwick, Coventry, CV4 7AL*
[2] *University of Oxford, Oxford, OX1 3LB*
[*] *Correspondence address J.E.Jewson@warwick.ac.uk*

October 2019

Abstract

In modern big-data applications it is increasingly the case that decision makers (DMs) are unable to fully and correctly specify a likelihood function for their observed data in order to conduct a full Bayesian analysis. They are often able to make broad structural judgements, but a level of interpolation is required to produce the likelihood model. As a result, there is likely to be more than one likelihood model that is consistent with the judgements the DM has been able to elicit and no principled reason to choose amongst these. Bayesian machinery must therefore seek to be stable across such equivalence classes. This is a well studied problem with respect to the parametric prior distributions but has not been properly investigated for the likelihood. In this paper we show that traditional Bayesian updating, minimising the Kullback-Leibler Divergence (KLD) between the sample distribution of the observations and the model, is only stable to a very strict class of likelihood models. On the other hand, Bayesian inference aimed at minimising the total-variation divergence (TVD), Hellinger divergence (HD) or the β-divergence (βD) is shown to be stable across interpretable neighbourhoods of likelihood models. We illustrate this for a Bayesian on-line changepoint detection algorithm applied to air pollution data from the City of London

1 Introduction

Given parameter prior, $\pi(\theta)$ and likelihood function $f(z;\theta)$, Bayes' rule provides the provision to update the prior beliefs to posterior beliefs after observing data $\boldsymbol{x}$

$$\pi(\theta|\boldsymbol{x}) = \frac{\pi(\theta)f(\boldsymbol{x};\theta)}{\int \pi(\theta)f(\boldsymbol{x};\theta)d\theta}. \tag{1}$$

However, in order to correctly specify this prior and likelihood a decision maker (DM) must posses the time and infinite introspection to consider all of the information available to them/in the world in order to produce probability specifications in the finest of details. As is pointed out by Goldstein (1990) this requires many more probability specifications to be made at a much higher precision than any DM is ever likely to be able to manage within time constraints of the problem, especially when eliciting continuous densities. This is not to say that a prior or model elicited by a DM is useless. They will often be able to make broad structural judgements - for example O'Hagan (2012) argues that judgements of equal probability such as medians and quartiles can be elicited accurately - but they do not have the introspection or time available to properly and accurately elicit a full probability distribution. As

a result in order to produce the specifications required to implement a full Bayesian analysis some form of interpolating approximation of the belief judgements the DM had been able to make must be used. The question then becomes which interpolating approximation should be used. There will likely be many probability distributions satisfying the judgements the DM has been able to make and no principled reason to choose any such one. As a result the DM actually elicits an equivalence class of models, with only one required for the analysis. Given such an equivalence class and no well founded reason to choose any one model, it is then desirable that Bayesian updating machinery be stable across such a class. We address this problem in this paper. Unlike previous literature we focus on the stability the likelihood function rather than the prior.

It is a well-known fact that arbitrary judgement should not be allowed to overly affect any statistical analysis. This is traditionally analysed using some form of sensitivity analysis. For Bayesian analyses the focus for this has generally been on the prior density $\pi(\theta)$ Jack: [Insert loads of references]. The prior density is the input of the Bayesian machinery that differs from frequentist analyses and as a result is often considered the subjective part of the analysis. We believe this derives from a *M*-CLOSED world assumption (Bernardo & Smith, 2001), that the data was generated from the model i.e. there exists some $\theta_0 \in \Theta$ such that

$$\boldsymbol{x} = x_1, \ldots, x_n \sim f(\cdot; \theta_0). \tag{2}$$

However statisticians are increasingly acknowledging that their inferences are taking pace in an *M*-OPEN world. Here, the model can be seen as one - of an equivalence class of - best guess belief models with all of these being misspecified compared with the process that generated that data. In the *M*-OPEN world it is now reasonable to consider the likelihood function as a subjective input additional to the prior and thus consider how stable Bayesian updating is to its specification.

In this paper we show that Bayesian updating using Bayes' rule, which is known to learn about the parameter of the model minimising the KLD between the data generating process (DGP) and the model, is stable to only a very strict equivalence class of probability models. Instead we consider general Bayesian updating (Bissiri, Holmes, & Walker, 2016) aimed at minimising divergences alternative to the KLD (Jewson, Smith, & Holmes, 2018). This provides a principled and coherent way to update beliefs about the parameter of a model minimising alternative divergences to the KLD. We are able to show that Bayesian updating minimising the total-variation divergence (TVD), Hellinger divergence (HD) and the β-divergence (βD) can be shown to be stable across an interpretable equivalence class of likelihood models. Updating minimising the TVD and HD requires an estimate of the data generating density and is thus not practical for high-dimensional applications with continuous likelihoods. Updating minimising the βD does not suffer from this drawback and thus provides a practically implementable and interpretably stable Bayesian update. We illustrate this on some toy examples and then consider the stability of Bayesian on-line changepoint detection (BOCPD) for air pollution data from the City of London.

The rest of the paper is organised as follows: Section 2 reviews the previous literature addressing the stability of Bayesian analyses. Section 3 introduces general Bayesian updating (Bissiri et al., 2016) and its application to estimating model parameters by minimising divergence (Jewson et al., 2018). Section 4 presents our theoretical contributions surrounding the guarantees on the stability of Bayesian analyses. Lastly Section 5 demonstrates these results, first on some toy examples and then on a BOCPD analysis.

2 Related work

Traditional Bayesian stability analyses have focused on examining the stability of inferential conclusions to the specification of the parameter prior (see Berger et al., 1994) and references within. Focus has been on the prior distribution for the model parameters as this is seen as the subjective part of the analyses differentiating a Bayesian analyses from the a frequentist's analogue.

Here rather than focus on specific cases we look at the inherent or automatic stability that can be guaranteed by the Bayesian learning machine. Gustafson and Wasserman (1995) consider automatic stability of Bayesian inference using some functioning prior f_0 and a 'genuine' prior $q_0^\epsilon = C(f_0, \epsilon, g)$ defined to be a linear, $q_0^\epsilon = (1-\epsilon)f_0 + \epsilon q$, or geometric, $q_0 \propto q^\epsilon f_0^{1-\epsilon}$, contamination of f_0 in the direction of g with contamination of size ϵ. Specifically they investigate the quantity

$$\sup_{g\in\Gamma} \lim_{\epsilon\to 0} \left\{ \frac{\text{TVD}\,(f_n, g_n^\epsilon)}{\text{TVD}\,(f_0, g_0^\epsilon)} \right\} \tag{3}$$

where f_n and g_n^ϵ are the posterior produced by Bayes' rule from n observations with shared likelihood from priors f_0 and g_0^ϵ respectively, and Γ is some class of contaminant priors. The quantity in Eq. (3) provides a worst case difference that can be observed a posteriori across some class of the contaminant priors. Gustafson and Wasserman (1995) prove that for contamination $C(f_0, \epsilon, g)$, being either linear or geometric ϵ-contaminations of the functioning prior f_0 then Eq. (3) divergences at rate $n^{k/2}$ as $n \to \infty$ where k is the dimension of the parameter space Θ. The fact that the rate increases with the dimension of the parameter space is particularly worrying for 'big-data' analyses.

While this result appears particularly alarming, Smith and Rigat (2012) provide conditions for the prior that would ensure posterior stability in terms of TVD. They show that TVD (f_n, g_n) is not actually driven by TVD (f_0, g_0) for large n, it is in fact driven by the roughness of the genuine prior g_0. The neighbourhoods considered by Gustafson and Wasserman (1995) allowed for a 'rough' prior contamination with a spike at the MLE of the observed data encouraging much faster convergence than under the functioning prior. Smith and Rigat (2012) instead consider the following metric defining neighbourhoods of prior densities

$$D_A^R(f, g) := \sup_{\theta,\phi\in A} \left| \frac{f(\theta)g(\phi)}{f(\phi)g(\theta)} - 1 \right|. \tag{4}$$

This condition is a particular way of demanding a level of smoothness in the perturbation from the genuine prior for each small subset $A \subseteq \Theta$. They show that under some mild regularity conditions that provided $D_A^R(f_0, g_0) < \eta$, where A is the small set of parameter values on which the likelihood concentrates, ensures that the posterior under the functioning prior tends in TVD to the posterior under the genuine prior as $\eta \to 0$. This stability results from the fact that the TVD between two posteriors is bounded above by the De Robertis distance between the posteriors. The De Robertis distance then has the special property that the distance between two posteriors using the same likelihood and different priors is equivalent to the distance between the priors. Therefore provided two priors have similar roughness, and are thus close according to De Robertis distance, ensures that the posterior inference produced from the same likelihood is stable in terms of TVD. These results of course assume that the likelihoods used to update priors f_0 and g_0 are the same. This paper, in contrast, is interested when such likelihoods might only be within a neighbourhood of their own.

A natural extension to the work of Gustafson and Wasserman (1995); Smith and Rigat (2012) is to consider for fixed prior on parameters θ, whether Bayesian inference is stable within some neighbourhood of the likelihood model. Smith (2007) briefly covers this topic, in the context of the De Robertis separation, and discovers that the data set can cause divergence in terms of De Robertis distance between the functioning and genuine posterior produced from different likelihoods. Beyond this initial results, work investigating the stability of Bayesian learning across a neighbourhood of likelihood models is limited.

More generally than the prior to posterior stability discussed above, stability of optimal decision making has been considered, largely in economics (Gilboa & Schmeidler, 1989; L. Hansen & Sargent, 2001b; L. P. Hansen & Sargent, 2001a; Whittle & Whittle, 1990) and more recently in statistics (Watson, Holmes, et al., 2016). By considering the stability of optimal decisions, these methods consider only neighbourhoods of posterior beliefs for those elements that enter into the loss function. These will often be posterior predictive distributions, whose neighbourhoods we consider later. These methods consider taking the minimax decision

$$d_C^* := \arg\min_{d\in\mathcal{D}} \sup_{\nu\in\Gamma_C} \mathbb{E}_{\nu(\theta)}\left[\ell(\theta, d)\right] \tag{5}$$

across a KLD neighbourhood of the Bayes' rule posterior beliefs

$$\Gamma_C := \left\{\nu(\theta) : \text{KLD}(\nu(\theta)||\pi(\theta|\boldsymbol{x})) \leq C\right\}. \tag{6}$$

Our criticism of these approaches is that they do not consider the stability of the Bayesian updating machinery. They start with the posterior, the output of the Bayesian updating machine, rather than the subjectively defined inputs, the prior and the likelihood. We argue for considering that the likelihood, in addition to the prior, be considered to have been defined up to some neighbourhood. Therefore a question of interest related to these methods would be what the ball around the likelihood (or prior) looks like in order to guarantee this KLD ball around the posterior (predictive). We try to answer this question in Section 4.4, see Lemma 1.

Alternatively Miller and Dunson (2018) consider producing robustified Bayesian updating by conditioning on data arriving in a neighbourhood of the empirical distribution of the data, $\pi(\theta|d(\boldsymbol{X}_{1:n}, \boldsymbol{x}_{1:n}) < R)$, rather than conditioning on the sample itself. Similarly to above they consider a KLD ball around the empirical distribution of the data, and used this to develop 'coarsened' posteriors. In practise a tractable approximation to these c-posteriors simply results in tempering the likelihood similarly to the work of Holmes and Walker (2017) and P. Grünwald (2016).

While this work is exciting and interesting these methods do not directly answer the questions we ask in this paper. We consider the data to be absolute and a lack of correspondence between the data and the model is the fault with the likelihood function for the data, rather than the data itself. Namely that discrepancies such as outliers are evidence that the model used for inference is misspecified, rather than that the model is correctly specified and that outliers are a problem of the observed data. Hence we try to stick to the Bayesian principle of fixing the observed data exactly, and considering stability across neighbourhoods around the subjectively defined elements of the analysis, the likelihood model and the prior.

Bayes linear methods provide a different approach to this problem (Goldstein, 1999). They take expectations as primitive rather than probabilities, in order to simplify the probability specification required of the DM. When expectation is primitive the DM need only concern themselves with the sub-collection of probabilities and expectations they consider themselves to be able to specify (Goldstein et al., 2006). These judgements can then be updated coherently without the need for the implicit interpolation required to specify a full likelihood model in order to update according to Eq. (1).

3 General Bayesian updating and minimum divergence estimation

A recent technique to conduct Bayesian belief updating without relying on a likelihood model for the data is general Bayesian updating (Bissiri et al., 2016). Consider updating beliefs about the minimiser of a loss function

$$\theta^* := \arg\min_{\theta\in\Theta} \int \ell(\theta, z) dG(z) \tag{7}$$

where $G(\cdot)$ is the data generating distribution of z. Bissiri et al. (2016) argue that in the presence of a prior on the location of θ^*, $\pi(\theta)$, some observed data $\boldsymbol{x} = (x_1, \ldots, x_n)$ and the loss function $\ell(\theta, x_i)$ connecting each observation to the parameter θ. An updating of beliefs must be possible. Such an updating beliefs must aim to minimise the posterior expected loss on the observed data, and be regularised towards the prior using the KLD. The KLD was motivated for this regularisation as it has been shown to be the only divergence ensuring a 'coherent' update of beliefs, where the same posterior was obtained if the updating is done sequentially or in one go. As a result the general Bayesian posterior becomes

$$\begin{aligned}\pi(\theta|\boldsymbol{x}) : &= \arg\min_q \mathbb{E}_{q(\theta)}\left[\sum_{i=1}^n \ell(\theta, x_i)\right] + \mathrm{KLD}(q(\theta)||\pi(\theta)) \\ &\propto \pi(\theta)\exp\left(-\sum_{i=1}^n \ell(\theta, x_i)\right).\end{aligned} \tag{8}$$

Now, if the loss function is the log-score, $\ell(\theta, x_i) = -\log f(x_i;\theta)$, then the general Bayesian update Eq. (8) becomes the traditional Bayes' rule update in Eq. (1). Minimising the log-score can be shown to minimise the KLD between the distribution of the sample and the model. As a result this serves to demonstrate the well known result that Bayesian updating learns about the parameter of the model minimising the KLD (Berk et al., 1966).

However it is also well known that minimising the log-score puts a lot of importance on correctly capturing the tails of the data (Bernardo & Smith, 2001). In fact, Jewson et al. (2018) observed that this can be to the extent that the distribution of most of the data is ignored to accommodate one outlier. Jewson et al. (2018) discuss how this has undesirable consequences from a decision making perspective and instead consider using general Bayesian updating to produce principled updates minimising divergences alternate to the KLD. They take advantage of the following duality between loss functions and divergences Dawid (2007); P. D. Grünwald and Dawid (2004)

$$D(g||f(\cdot;\theta)) = \mathbb{E}_{x\sim g}\left[\ell_D(x, f(\cdot;\theta))\right] - \mathbb{E}_{x\sim g}\left[\ell_D(x, g)\right]. \tag{9}$$

Where $\ell_D(\cdot,\cdot)$ is generally termed a score function rather than a loss function as it takes a probability

distribution as one argument. The score function $\ell_D(\cdot,\cdot)$ can then be substituted into the updating equation, Eq. (8), to produce a Bayesian updating of beliefs about the parameter of the model minimising divergence $D(\cdot||\cdot)$ between the model and the data generating process g. Unsurprisingly when $\ell_D(\cdot,\cdot)$ is the log-score then Eq. (9) is the KLD. The βD is recovered in Eq. (9) for

$$\ell_{(\beta)}(x, f(\cdot;\theta)) = -\frac{1}{\beta-1} f(\cdot;\theta)^{\beta-1} + \frac{1}{\beta}\int f(z;\theta)^{\beta} dz, \tag{10}$$

which was first used by Ghosh and Basu (2016) to produce a robustified Bayesian posterior and has since been used to great effect (see e.g. Higgins et al., 2017; Knoblauch, Jewson, & Damoulas, 2018, 2019). Two other divergences of interest here and considered in Jewson et al. (2018), are the TVD and the HD which are recovered from Eq. (9) using loss functions

$$\ell_{\text{TVD}}(x, f(\cdot;\theta)) = \frac{1}{2}\left|1 - \frac{f(x;\theta)}{g_n(x)}\right| \tag{11}$$

$$\ell_{\text{HD}}(x, f(\cdot;\theta)) = -\frac{\sqrt{f(x;\theta)}}{\sqrt{g_n(x)}} \tag{12}$$

The HD has previously been used by Hooker and Vidyashankar (2014) to produce a robustified updating of beliefs. Note here that in order to obtain the loss function representation for the HD and the TVD we have had to introduce an estimate of the data generating density $g_n(x)$. Hooker and Vidyashankar (2014); Jewson et al. (2018) were able to use kernel density estimation to do this effectively for small dimensional problems. However, we accept that in practice for high-dimensional problems with continuous data, 'local' loss function not requiring an estimate of the data generating density such as the log-score and the βD-loss are much more appealing.

4 Theoretical contribution

Next we consider extending the work of Gustafson and Wasserman (1995); Smith and Rigat (2012) to establish what can be said about the stability of Bayesian updating across neighbourhoods of likelihood models. Proofs of all of the results presented in this section can be found in the Appendix.

4.1 Notions of Stability

At first it seems natural to mimic Gustafson and Wasserman (1995); Smith and Rigat (2012) and investigate whether the posteriors for parameters θ are close for different likelihood functions within some neighbourhood. However, this is not informative. The posterior for two distinct likelihood models, given the same data and as $n \to \infty$, will almost certainly convergence around distinct values of θ. As a result, the posteriors will become very far apart by any divergence measure as the number of data points increases.

Instead when considering stability to the likelihood, it is more natural to consider distributions for the observables. Consider two likelihood models

$$\{f(x;\theta_f) : x \in \mathcal{X} \subset \mathbb{R}^p, \theta_f \in \Theta_f\} \tag{13}$$

$$\{h(x;\theta_h) : x \in \mathcal{X} \subset \mathbb{R}^p, \theta_h \in \Theta_h\}, \tag{14}$$

for the same observables $x \in \mathcal{X} \subset \mathbb{R}^p$, p being the dimension of the observation space. In Definition 1 these are close if for a given set of parameters they produce similar densities for the observables x.

Definition 1 (Neighbourhood of likelihood models)**.** Given divergence $D(\cdot||\cdot)$ and size ϵ, the neighbourhood of likelihood models for observable $x \in \mathcal{X}$ is defined as

$$\mathcal{N}_\epsilon^D := \{(f(\cdot;\theta_f), h(\cdot;\theta_h)) : D(f\left(\cdot;\{\theta_U, \theta_{f\setminus h}\}\right)||h\left(\cdot;\{\theta_U, \theta_{h\setminus f}\}\right)) < \epsilon, \\ \text{for all values of } \theta_U \in \Theta_U, \theta_{f\setminus h} \in \Theta_{f\setminus h}, \theta_{h\setminus f} \in \Theta_{h\setminus f}\} \tag{15}$$

where $\Theta_U := \Theta_f \cap \Theta_h$ is the intersection of the parameter spaces Θ_f and Θ_h for the two likelihood models, $\Theta_f = \{\Theta_U, \Theta_{f\setminus h}\}$ and $\Theta_h = \{\Theta_U, \Theta_{h\setminus f}\}$.

Neighbourhood $\mathcal{N}_\epsilon^D$ demands that when the shared part of the parameter space θ_U is fixed the likelihoods produces similar densities for x measured by divergence D for all values of the unshared parameters $\theta_{f\setminus h}$ and $\theta_{h\setminus f}$.

A Bayesian analysis that is stable across such a neighbourhood must then produce similar posterior predictives for future observables $x' \in \mathbb{R}^p$ given data $\boldsymbol{x}_{1:n} \in \mathbb{R}^{n\times p}$, where n is the number of observations, modelled using two likelihoods in this neighbourhood. This provides a natural analogue to the work of Gustafson and Wasserman (1995); Smith and Rigat (2012), who consider the stability prior to posterior for parameters. We instead focus on observables, the likelihood provides a prior for observables and the predictive is the corresponding posterior.

To this end we consider two particular metrics to compare the posterior predictive inferences from two different likelihood models. The first relates to the divergence between the finite sample predictive distributions arising from the two likelihood models. The second considers the difference between how the predictives, as $n \to \infty$, from the two different likelihood functions approximate the DGP.

One large advantage of considering stability of the distributions for observables rather than parameters is it allows for likelihoods with different dimensions to their parameter space to be considered in the same neighbourhood provided they produce a distribution over the same observables. This is exemplified by the notation used in Definition 1. For example, consider that

$$\begin{aligned} f(\cdot;\theta_f) &= \mathcal{N}\left(x;\mu,\sigma^2\right) \\ h(\cdot;\theta_h) &= 0.95 \times \mathcal{N}\left(x;\mu,\sigma^2\right) + 0.05 \times \mathcal{N}\left(x;\mu_c,\sigma_c^2\right), \end{aligned} \tag{16}$$

then $\theta_U = \{\mu, \sigma^2\}$, $\theta_{f\setminus h} = \emptyset$ and $\theta_{h\setminus f} = \{\mu_c, \sigma_c^2\}$. For fixed value of θ_U and any value of $\theta_{h\setminus f}$ we have that $\text{TVD}(f\left(\cdot;\theta_U\right)||h\left(\cdot;\{\theta_U, \theta_{h\setminus f}\}\right)) < 0.05$.

Jack: [I am slightly unsure about this, at the moment I am fixing the parametrisation and saying the models must be close when i fix the parameters in that parametrisation. Looking at the Maths it may be the case that we just need there to exist a parametrisation where the models are close, which would make the results slightly stronger I think.]

However this neighbourhood condition is unlikely to hold unless Θ_f and Θ_h almost entirely overlap. We additionally note the subtle point that these neighbourhoods are only really meaningful if the parameters that overlap between the two likelihood models maintain the same interpretation across these likelihoods such that it is meaningful that the likelihood models are similar when their parameters

are the same values. This may require reparametrisations from the traditional parametrisations. One example is that they may correspond to a particular moments of the predictive distribution - or in the example above, for (μ, σ^2), the moments of the uncontaminated population.

For readability we present the results of this paper under the assumption that the likelihood models $\{f(\cdot;\theta_f) : \theta_f \in \Theta_f\}$ and $\{h(\cdot;\theta_h) : \theta_h \in \Theta_h\}$ have the same parameter spaces $\Theta_f = \Theta_h = \Theta$. That is to say that $\Theta_{f\setminus h} = \Theta_{h\setminus f} = \emptyset$. This significantly simplifies the notation required. Relaxing this assumption does not change the results in this paper but does change their proofs. Extended proofs of our results to the situations where the parameter spaces are no longer equal can also be found in Appendix (Section 7.2). The next section establishes the notation I will use throughout this the rest of this paper.

4.2 Conditions, notation and assumptions

Before providing any of the results we first introduce the notation and assumptions that will be required.

We define the general Bayesian posterior and corresponding posterior predictive for likelihood model $\{f(x;\theta_f) : \theta_f \in \Theta_f\}$ targeting minimising divergence $D(g||f(\cdot;\theta_f))$ as

$$\pi_f^D(\theta_f|\boldsymbol{x}_{1:n}) = \frac{\pi_f^D(\theta_f)\exp(-\sum_{i=1}^n \ell_D(x_i, f(\cdot;\theta_f)))}{\int \pi_f^D(\theta_f)\exp(-\sum_{i=1}^n \ell_D(x_i, f(\cdot;\theta_f)))d\theta_f} \tag{17}$$

$$m_f^D(y|\boldsymbol{x}_{1:n}) = \int f(y;\theta_f)\pi_f^D(\theta_f|\boldsymbol{x}_{1:n})d\theta_f, \tag{18}$$

where $\ell_D(x,\theta_f)$ is the loss function required to do inference minimising divergence $D(g||f(\cdot;\theta_f))$. We remind the reader here that taking the loss function to be the log-score,

$$\ell_{\text{KLD}}(x_i, f(\cdot;\theta_f))) = -\log f(x_i;\theta_f), \tag{19}$$

recovers standard Bayes' rule updating (Eq. 1) in Eq. (17) and standard one-step-ahead predictive distribution in Eq. (18). Respectively we term the procedure of updating minimising the D = KLD, TVD, HD and the βD as KLD-Bayes, TVD-Bayes, HD-Bayes and βD-Bayes. We assume throughout that the normaliser of the general Bayesian posterior $\int \pi_f^D(\theta_f)\exp(-\sum_{i=1}^n \ell_D(x_i, f(\cdot;\theta_f)))d\theta_f$ is finite. Throughout this section we will use the $\cdot$ notation within divergence functions to indicate the variable that is being integrated over in the divergence, i.e. the divergence does not depend on a value for this variable. Lastly we define

$$\theta_f^D = \arg\min_{\theta_f\in\Theta_f} D(g(\cdot), f(\cdot;\theta_f)), \tag{20}$$

as the parameter of likelihood model $\{f(x;\theta_f) : \theta_f \in \Theta_f\}$ minimising divergence $D(\cdot||\cdot)$ to the data generating density g. This is always assumed to exist and to be unique.

4.3 Metrics

First we demonstrate that divergences that are also metrics have appealing properties for statistical inference. The two metrics considered by Jewson et al. (2018) for inference, the TVD and HD, both require an estimate of the data generating density g_n to be implemented. For the mathematical results of this section we assume we have access to one that is consistent e.g. a KDE for univariate observations, although we acknowledge that in high-dimensional real world problems these are unlikely

to be available. First we list several important properties for the target divergence for inference to satisfy.

Condition 1 (A convex divergence metric)**.** According to its definition divergence $D(g||f)$ must be non-negative everywhere and only 0 when $f = g$. In addition to these we require that the divergence satisfies the following:

M1 Is symmetric $D(g||f) = D(f||g)$.

M2 Satisfies the triangle inequality $D(g, f) \leq D(g, h) + D(h, f)\ \forall h$.

M3 Is convex in both of its arguments. That is to say that for $\lambda \in [0, 1]$

$$D(\lambda g_1 + (1-\lambda)g_2, f) \leq \lambda D(g_1, f) + (1-\lambda)D(g_2, f) \tag{21}$$

$$D(g, \lambda f_1 + (1-\lambda)f_2) \leq \lambda D(g, f_1) + (1-\lambda)D(g, f_2). \tag{22}$$

For such divergences we introduce the following notation

$$D(g||f) = D_M(g, f). \tag{23}$$

Note M1 and M2 of Condition 1 ensure that divergence D_M is a proper distance metric. The triangle inequality in particular will be important in the results to come. The triangle inequality fits naturally with stability. We consider three distributions for the data, the DGP and two candidate likelihoods within some neighbourhood. We seek to analyse the stability of inference trying to minimise a divergence between the DGP and the two likelihood models.

4.3.1 Stability of the limiting predictive approximation of the DGP

Our first results corresponds to the predictive distribution produced from the divergence minimising parameter for each model. We invoke the asymptotic normality of the general Bayesian posteriors (Chernozhukov & Hong, 2003; Lyddon, Holmes, & Walker, 2018) to say that under certain regularity conditions the Bayesian posterior predictive distributions will converge to these as $n \to \infty$. Theorem 1 demonstrates that inference minimising divergence $D_M(\cdot, \cdot)$ is stable across the neighbourhood $\mathcal{N}_\epsilon^{D_M}$ in terms of how the predictive distribution produced using the divergence minimising parameters approximates the DGP.

Theorem 1 (Limiting predictive stability using divergence metrics)**.** Consider the following conditions:

- Divergence $D_M(\cdot, \cdot)$ satisfies M1 and M2 from Condition 1
- We have two likelihood models $\{f(\cdot; \theta_f) : \theta_f \in \Theta_f\}$ and $\{h(\cdot; \theta_h) : \theta_h \in \Theta_h\}$ such that for $\epsilon > 0$ we have that $f, h \in \mathcal{N}_\epsilon^{D_M}$, as defined in Definition 1

Then

$$\left| D_M(g, f(\cdot; \hat{\theta}_f^{D_M})) - D_M(g, h(\cdot; \hat{\theta}_h^{D_M})) \right| \leq \epsilon \tag{24}$$

for all data generating densities g, where $\hat{\theta}_f^{D_M} = \arg\min_\theta D_M(g, f(\cdot; \theta))$ and $\hat{\theta}_h^{D_M} = \arg\min_\theta D_M(g, h(\cdot; \theta))$.

Theorem 1 guarantees that the absolute distance between the divergence from the limiting predictive density of two likelihood models in $\mathcal{N}_{\epsilon}^{D^M}$ to the DGP, is no further than the distance between the two likelihood models a priori. The absolute distance between the divergences to the DGP may seem a strange criteria to look at. However bounding this guarantees stability in the approximation of the model to the DGP across the neighbourhood defined using that divergence. Therefore, the DM can be sure that which ever model they choose within this neighbourhood, they will produce a similar limiting approximation of the DGP, and that this holds completely independently of the DGP and how well specified the two likelihood models are relative to this.

Next we produce a further result investigating the divergence between the finite sample predictive distributions produced from each model.

4.3.2 Stability of the finite sample posterior predictive distributions

In order to prove theorems involving the finite sample posterior predictive distributions, Theorems 2 and 4, we require the following conditions on the observations and the prior specification.

Condition 2 (Concentration of the posterior)**.** For divergence $D(\cdot||\cdot)$ the dataset, $\boldsymbol{x}_{1:n} \sim g(\cdot)$, is of sufficient size and regularity, and the priors $\pi_f^D(\theta)$ and $\pi_h^D(\theta)$ have sufficient prior mass at θ_f^D and θ_h^D such that the posteriors $\pi_{n,f}^D(\theta_f|X_{1:n})$ and $\pi_{n,h}^D(\theta_h|X_{1:n})$ have concentrated to ensure

$$\int_{\Theta_f} D(g||h(y|\theta_f))\pi_f^D(\theta_f|\boldsymbol{x}_{1:n})d\theta_f \geq \int_{\Theta_h} D(g||h(y|\theta_h))\pi_h^D(\theta_h|\boldsymbol{x}_{1:n})d\theta_h \tag{25}$$

$$\int_{\Theta_h} D(g||f(y|\theta_h))\pi_h^D(\theta_h|\boldsymbol{x}_{1:n})d\theta_h \geq \int_{\Theta_f} D(g||f(y|\theta_f))\pi_f^D(\theta_f|\boldsymbol{x}_{1:n})d\theta_f. \tag{26}$$

Condition 2 ensures that n is large enough for the posterior based on the likelihoods f and h to have concentrated sufficiently around their optimal parameter such that the expected divergence under the posterior for θ_k between model $k \in \{f, h\}$ and the DGP is less than the same expected divergence under the posterior for the other model.

The asymptotic normality results of Chernozhukov and Hong (2003); Lyddon et al. (2018) for the general Bayesian posterior concern convergence in distribution and thus one must be slightly careful when evoking these to suggest that there must exist some n such that Condition 2 holds. However, under the assumption that both likelihood models $f(\cdot;\theta_f)$, $h(\cdot;\theta_h)$ and DGP g are all absolutely continuous and provided the weak conditions for asymptotic normality Chernozhukov and Hong (2003); Lyddon et al. (2018) are satisfied, then these results suggest that Condition 2 will be satisfied for large enough n, as by definition $D(g, k(\cdot;, \theta_k^D)) \leq D(g, k(\cdot;, \theta_{k'}^D))$ for $k \in \{f, h\}$ and $k^{'} = \{f, h\} \setminus k$.

Condition 2 is the only part of any of these theorems where the observed data appears. So the following theorems simply require that the Bayesian updating is being done conditional on a dataset satisfying Condition 2. In this sense we consider it formally Bayesian. Extensions could look at whether Condition 2 and the following theorems hold in expectation under the data generating process (DGP).

Theorem 2 (Stability of the posterior predictive using divergence metrics)**.** Consider the following conditions:

- Divergence $D_M(\cdot, \cdot)$ satisfies Condition 1

- We have two likelihood models $\{f(\cdot;\theta_f) : \theta_f \in \Theta_f\}$ and $\{h(\cdot;\theta_h) : \theta_h \in \Theta_h\}$, data generating process g, priors $\pi(\theta_f)$ and $\pi(\theta_h)$ and data $\boldsymbol{x}_{1:n}$ such that Condition 2 holds for divergence $D_M(\cdot,\cdot)$
- The two likelihood models $\{f(\cdot;\theta_f) : \theta_f \in \Theta_f\}$ and $\{h(\cdot;\theta_h) : \theta_h \in \Theta_h\}$ are such that for $\epsilon > 0$ we have that $f, h \in \mathcal{N}_\epsilon^{D_M}$ as defined in Definition 1

Then for $m_f^{D^M}$ and $m_h^{D^M}$ as defined in Eq. (18)

$$D^M(m_f^{D^M}(\cdot|\boldsymbol{x}_{1:n}), m_h^{D^M}(\cdot|\boldsymbol{x}_{1:n}))) \leq \epsilon + R^{D_M}(g, f, h, \boldsymbol{x}_{1:n}), \tag{27}$$

where

$$\begin{aligned} &R^{D_M}(g, f, h, \boldsymbol{x}_{1:n}) := \\ &2\min\left\{\int \left(D_M(g(\cdot), f(\cdot|\theta_f))\right) \pi_f^{D_M}(\theta_f|\boldsymbol{x}_{1:n}) d\theta_f, \int \left(D_M(g(\cdot), h(\cdot|\theta_h))\right) \pi_h^{D_M}(\theta_h|\boldsymbol{x}_{1:n}) d\theta_h\right\}. \end{aligned} \tag{28}$$

Theorem 2 demonstrates that when doing general Bayesian inference minimising divergence $D_M(\cdot,\cdot)$, the divergence between the finite sample posterior predictive distributions produced using each likelihood model is smaller than the prior divergence between these models plus some remainder term $R^{D_M}(g, f, h, \boldsymbol{x}_{1:n})$, provided Condition 2 holds.

Unlike Theorem 1 the upper bound on the stability of the finite sample posterior predictive depends on the data generating density g, and will also not go to 0 as $\epsilon \to 0$. This results from the remainder term $R^{D_M}(g, f, h, \boldsymbol{x}_{1:n})$. This remainder term will be small provided one (or both) of the likelihood models $f(\cdot|\theta_f)$ or $h(\cdot|\theta_h)$ is close to the DGP in terms of divergence $D^M(\cdot,\cdot)$ for some value of their parameter θ_f or θ_h. So here we are able to guarantee stability in the finite sample posterior predictive distributions provided the neighbourhood of models are not too badly specified relative to the DGP.

Unlike the prior misspecification case considered by Gustafson and Wasserman (1995), the divergence between the posterior predictive distributions cannot be expected to converge to 0 as the number of data points grows. However it seems reasonable to demand that if the likelihood models are close then the posterior predictive divergence ought to be bounded, and certainly not divergent in n. Theorem 2 provides this result when inference is designed to minimises a divergence metric.

It is in general hard to say how tight this bound is, for example the remainder term does not depend on ϵ and as a result will not go to 0 as $\epsilon \to 0$. The results in Theorem 1 in the previous section demonstrate that at least as $n \to \infty$ and $\epsilon \to 0$ a different stability metric goes to 0. However the next result, Corollary 1, shows the bound to be as tight as can be expected when the true DGP is contained within the neighbourhood $\mathcal{N}_\epsilon^{D_M}$.

Corollary 1. Consider the following conditions:

- Assume without loss of generality that f is correctly specified for g, that is to say that there exists θ_{f_0} such that $f(\cdot;\theta_{f_0}) = g(\cdot)$.
- Divergence $D_M(\cdot,\cdot)$ satisfies Condition 1 and additionally that the divergence metric satisfies $D^M(h_1, h_2) \leq b < \infty$, $\forall h_1, h_2$

- We have two likelihood models $\{f(\cdot;\theta_f) : \theta_f \in \Theta_f\}$ and $\{h(\cdot;\theta_h) : \theta_h \in \Theta_h\}$, data generating process g, priors $\pi(\theta_f)$ and $\pi(\theta_h)$ and data $\boldsymbol{x}_{1:n}$ such that Condition 2 holds for divergence $D_M(\cdot,\cdot)$
- The two likelihood models $\{f(\cdot;\theta_f) : \theta_f \in \Theta_f\}$ and $\{h(\cdot;\theta_h) : \theta_h \in \Theta_h\}$ are such that for $\epsilon > 0$ we have that $f, h \in \mathcal{N}_\epsilon^{D_M}$, as defined in Definition 1

Then for $m_f^{D^M}$ and $m_f^{D^M}$ as defined in Eq. (18) as $n \to \infty$

$$D^M(m_f^{D_M}(\cdot|\boldsymbol{x}_{1:n}), m_h^{D_M}(\cdot|\boldsymbol{x}_{1:n}))) \leq \epsilon \tag{29}$$

almost surely.

Therefore if we can specify a model that contains the DGP, then we can be sure that if we conduct Bayesian updating aimed at minimising divergence $D_M(\cdot,\cdot)$, then the posterior inferences from the approximate model will be no further from the posterior inferences resulting from the true model than the divergence between the likelihood models a priori. This result is to be expected. However the fact that it follows from Theorem 2 provides some idea of the tightness of the bounds in this theorem.

The results in the previous two sections imply stability comes naturally to inference designed to minimise a divergence that is a proper metric, with satisfying the triangle inequality being particularly important. Two obvious candidates for this as discussed above are the TVD and the HD. Jewson et al. (2018) discussed the desirability of learning using the TVD from a decision theoretic point of view. They identified that minimising the TVD provides guarantees on the approximation of expected utilities by the inferences from the model. The TVD has further desirable properties when thinking about a neighbourhood for likelihood models and practical belief elicitation. 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. It seems a reasonable requirement for a DM to be able to accurately elicit probabilities on a natural scale. Additionally, TVD neighbourhoods contain ϵ-contaminations considered in the context of prior stability by Gustafson and Wasserman (1995).

However there are several difficulties associated with inference targeted at the minimisation of the TVD or the HD. The main one of these being that they both requires an estimate of the data generating density, $g_n(x)$. As a result we seek to use divergences that do not require a density estimate, termed as 'local', to attempt to approximate the stability results concerning divergence metrics. Specifically here we consider the KLD associated with Bayes' rule and the βD.

4.4 The KLD

While inference targeting metrics is inconvenient to implement in practise, inference targeting the KLD (Bayes' rule) is straightforward due to the local property of the log-score. However Lemma 1 shows that stability in terms of the KLD requires unreasonable assumptions on the DGP and the neighbourhood of likelihood models.

Lemma 1 (Limiting stability for the KLD)**.** Defining

$$\hat{\theta}_f^{\mathrm{KLD}} = \arg\min_\theta \mathrm{KLD}(g, f(\cdot;\theta)) \tag{30}$$

$$\hat{\theta}_h^{\mathrm{KLD}} = \arg\min_\theta \mathrm{KLD}(g, h(\cdot;\theta)), \tag{31}$$

we have that for all data generating densities g

$$\begin{aligned}&\left|\text{KLD}(g, f(\cdot;\hat{\theta}_f^{\text{KLD}})) - \text{KLD}(g, h(\cdot;\hat{\theta}_h^{\text{KLD}}))\right| \\ &\leq \max\left\{\int g \log \frac{h(\cdot;\hat{\theta}_h^{\text{KLD}})}{f(\cdot;\hat{\theta}_h^{\text{KLD}})} dx, \int g \log \frac{f(\cdot;\hat{\theta}_f^{\text{KLD}})}{h(\cdot;\hat{\theta}_f^{\text{KLD}})} dx\right\}. \end{aligned} \tag{32}$$

Jack: [Proving you can't easily bound an upper bound doesn't show it's big, however there are actually equalities in here mainly]

Lemma 1 provides an upper bound on the difference in the quality of the KLD approximation to the DGP of two different likelihoods used in Bayes' rule. Standard bounds associated with the natural logarithm are

$$1 - \frac{1}{y} \leq \log y \leq y - 1 \tag{33}$$

which enables us to bound this remainder term

$$\int g \log \frac{h(\cdot;\hat{\theta}_h^{\text{KLD}})}{f(\cdot;\hat{\theta}_h^{\text{KLD}})} dx \leq \int g h(\cdot;\hat{\theta}_h^{\text{KLD}}) + \frac{g}{f(\cdot;\hat{\theta}_h^{\text{KLD}})} dx \tag{34}$$

$$\int g \log \frac{f(\cdot;\hat{\theta}_f^{\text{KLD}})}{h(\cdot;\hat{\theta}_f^{\text{KLD}})} dx \leq \int g f(\cdot;\hat{\theta}_f^{\text{KLD}}) + \frac{g}{h(\cdot;\hat{\theta}_f^{\text{KLD}})} dx. \tag{35}$$

As a result we are able to guarantee the stability of traditional Bayesian inference if we are able to bound $\frac{g}{h(\cdot;\theta_h)}$ and $\frac{g}{f(\cdot;\theta_h)}$ from above. In fact we can see that even if we were to try and apply some reverse Pinsker's inequality to this term the ratios $\frac{g}{h(\cdot;\theta_h)}$ and $\frac{g}{f(\cdot;\theta_h)}$ still remain e.g.

$$\begin{aligned}\int g \log \frac{h(\cdot;\theta_h)}{f(\cdot;\theta_h)} dx &\leq \int \frac{g}{h(\cdot;\theta_h)} h(\cdot;\theta_h) \log \frac{h(\cdot;\theta_h)}{f(\cdot;\theta_h)} dx \qquad (36)\\ &\leq M_h^* \text{KLD}(h(\cdot;\theta_h)||f(\cdot;\theta_h)) \qquad (37)\end{aligned}$$

where $M_h^* = \operatorname{ess\,sup} \frac{g}{h(\cdot;\theta_h)}$. So even if we had conditions on f and h such that a reverse Pinsker's inequality held enabling us to upper bound the $\text{KLD}(h(\cdot;\theta_h)||f(\cdot;\theta_h))$ by the $\text{TVD}(h(\cdot;\theta_h)||f(\cdot;\theta_h))$ and then considered a TVD neighbourhood for our likelihood models, we would still have to bound M_h^*. As a result we conclude that analogously to Smith and Rigat (2012), who demonstrate that a TVD ball around the prior does not impact the posterior stability, a TVD ball around the likelihood model is not sufficient for posterior stability when using Bayes' rule updating.

In fact, posterior stability in the manner we consider here can only be guaranteed if

$$|\log(h(\cdot;\theta_h)) - \log(f(\cdot;\theta_f))| \tag{38}$$

is small in regions where g has density. Without knowledge of g, this requires that Eq. (38) be small everywhere. Therefore in order to be able to produce stable inference as described above, the DM must be able to be confident in the accuracy of their probability statements on the log-scale rather than on the natural scale that we considered for the neighbourhood $\mathcal{N}_\epsilon^{\text{TVD}}$. Logarithms act to inflate the magnitude of small numbers and thus ensuring that $|\log(h(\cdot;\theta_h)) - \log(f(\cdot;\theta_f))|$ 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 their probability specifications as they get further and further into the tails. Something that is known to be very difficult (O'Hagan et al., 2006; Winkler & Murphy, 1968).

We do however note that this notion of stability is with respect to a metric that we know to be intrinsically unstable in a number of ways. For example it is very possible that $h(\cdot;\hat{\theta}_h^{\text{KLD}})$ and $f(\cdot;\hat{\theta}_f^{\text{KLD}})$ could be stable in the sense of Theorem 1 and the TVD metric but still produce very different approximations to g when the quality of the approximation is measured by the KLD. Currently we require that the metric for stability is the same metric we learn using as it easily allows us to say that $D(g||h(\cdot;\hat{\theta}_h^D)) \leq D(g||h(\cdot;\hat{\theta}_f^D))$.

Jack: [I should use the example of the Gaussian and STudents t here, like I do in the presentation]

4.5 The βD

We establish above that it is difficult to specify a neighbourhood of likelihood models such that traditional Bayesian inference minimising the KLD is stable. Here we show that stability can be achieved across the natural $\mathcal{N}_\epsilon^{\text{TVD}}$ neighbourhood of likelihood models when learning using the robust βD.

Firstly we prove an original result connecting the βD and the TVD in a similar manner to a triangle inequality. The result relies on $1 \leq \beta \leq 2$, which places the βD in between the KLD at $\beta = 1$ and the L_2-distance $D_B^{(2)}(g||f) = \frac{1}{2}\int (f-g)^2$. We are yet to come across scenarios where setting β outside this range is appropriate from a practical viewpoint (see e.g. Jewson et al., 2018; Knoblauch et al., 2018). The next results also relies on being able to bound the essential supremum (ess sup) of densities g, f and h. Given base measure μ[1], M is the essential supremum of density $f(x)$, $\operatorname{ess\,sup} f(x) = M$, if the set defined by $f^{-1}(M,\infty)$ has measure 0, i.e. $\mu\left(f^{-1}(M,\infty)\right) = 0$.

Lemma 2 (A triangle inequality relating the βD and the TVD)**.** For densities $f(x)$, $h(x)$ and $g(x)$ with the property that $\max\{\operatorname{ess\,sup} f, \operatorname{ess\,sup} h, \operatorname{ess\,sup} g\} \leq M < \infty$ and $1 < \beta \leq 2$ we have that

$$D_B^{(\beta)}(g||f) \leq D_B^{(\beta)}(g||h) + \frac{M^{\beta-1}}{\beta-1}\text{TVD}(h,f) \tag{39}$$

$$D_B^{(\beta)}(g||h) \leq D_B^{(\beta)}(g||f) + \frac{M^{\beta-1}}{\beta-1}\text{TVD}(h,f) \tag{40}$$

This result is a significant one. It shows an important link between the βD, a convenient divergence to use for inference, and the TVD which we have argued in this paper has properties concerning both accurate decision making and belief specification. We showed above that such an analogous result was not available to connect the KLD with the TVD. Lemma 2 can be used to prove a form of limiting stability for inference using the βD.

Jack: [Can we say something more about the value of M, do I actually only need M to be the maximum at the divergence minimising parameters]

[1] assumed to be the Lebesque measure for continuous random variables and the counting measures for discrete random variables.

4.5.1 Stability of the limiting predictive approximation to the DGP

Using Lemma 2 we firstly seek to bound the absolute distance between the βD between the DGP and each of the limiting predictive distribution produced from two likelihood models within $\mathcal{N}_\epsilon^{\text{TVD}}$.

Theorem 3 (Limiting predictive stability of βD inference)**.** Consider the following conditions:

- $1 < \beta \leq 2$
- We have two likelihood models $\{f(\cdot;\theta_f) : \theta_f \in \Theta_f\}$ and $\{h(\cdot;\theta_h) : \theta_h \in \Theta_h\}$ and data generating process g such that
$$\max\{\operatorname{ess\,sup} f, \operatorname{ess\,sup} h, \operatorname{ess\,sup} g\} \leq M < \infty \tag{41}$$
- The two likelihood models $\{f(\cdot;\theta_f) : \theta_f \in \Theta_f\}$ and $\{h(\cdot;\theta_h) : \theta_h \in \Theta_h\}$ are such that for $\epsilon > 0$ we have that $f, h \in \mathcal{N}_\epsilon^{\text{TVD}}$, as defined in Definition 1.

Then
$$\left| D_B^{(\beta)}(g||f(\cdot;\hat{\theta}_h^{(\beta)})) - D_B^{(\beta)}(g||h(\cdot;\hat{\theta}_f^{(\beta)})) \right| \leq \frac{M^{\beta-1}}{\beta-1}\epsilon \tag{42}$$
where $\hat{\theta}_f^{(\beta)} = \arg\min_\theta D_B^{(\beta)}(g||f(\cdot;\theta))$ and $\hat{\theta}_h^{(\beta)} = \arg\min_\theta D_B^{(\beta)}(g||h(\cdot;\theta))$.

Theorem 3 shows that for two likelihood models in the neighbourhood
$$f(\cdot;\theta_f), h(\cdot;\theta_h) \in \mathcal{N}_\epsilon^{\text{TVD}}, \tag{43}$$
we can be sure that their limiting predictive distribution, $h(\cdot;\hat{\theta}_h^{(\beta)})$ and $f(\cdot;\hat{\theta}_f^{(\beta)})$, aimed at minimising the βD, will be similarly close to the DGP in terms of the βD. So learning using the βD allows us to guarantee two likelihood models that are close in TVD a priori will converge (assuming the regularity conditions of Chernozhukov and Hong (2003); Lyddon et al. (2018)) on predictive inference that is stable with respect to the βD approximation of the DGP. This provides an analogous result to Theorem 1, but now inference is aimed at minimising the βD which is easier to practically implement. We paid particular attention to being able to define the a prior neighbourhood of models in terms of TVD as we believe this is a reasonable neighbourhood with which a DM ought to be able to specify their likelihood up to, see the discussion at the end of Section 4.3

4.5.2 Stability in the finite sample posterior predictive distributions

Next we go one step further and seek to extend this stability in the limiting approximation of the DGP, to being able to bound the βD between the finite sample predictive distributions resulting from two likelihood models in the neighbourhood $\mathcal{N}_\epsilon^{\text{TVD}}$. This result is analogous to Theorem 2

Theorem 4 (Stability of the posterior predictives under the βD learning)**.** Consider the following conditions:

- $1 < \beta \leq 2$
- We have two likelihood models $\{f(\cdot;\theta_f) : \theta_f \in \Theta_f\}$ and $\{h(\cdot;\theta_h) : \theta_h \in \Theta_h\}$, data generating process g satisfying
$$\max\{\operatorname{ess\,sup} f, \operatorname{ess\,sup} h, \operatorname{ess\,sup} g\} \leq M < \infty, \tag{44}$$

and priors $\pi(\theta_f)$ and $\pi(\theta_h)$ and data $\boldsymbol{x}_{1:n}$ such that Condition 2 holds for divergence $D(\cdot,\cdot) = D_B^{(\beta)}(\cdot||\cdot)$

- The two likelihood models $\{f(\cdot;\theta_f) : \theta_f \in \Theta_f\}$ and $\{h(\cdot;\theta_h) : \theta_h \in \Theta_h\}$ are such that for $\epsilon > 0$ we have that $f, h \in \mathcal{N}_\epsilon^{\text{TVD}}$, as defined in Definition 1

Then

$$D_B^{(\beta)}(m_f^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})||m_h^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})) \tag{45}$$
$$\leq \frac{M^{\beta-1}}{\beta-1}\epsilon + \int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h$$
$$D_B^{(\beta)}(m_h^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})||m_f^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})) \tag{46}$$
$$\leq \frac{M^{\beta-1}}{\beta-1}\epsilon + \int\int R(g||h(\cdot;\theta_h)||f(\cdot;\theta_f))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h.$$

where $R(g||f||h)$ and $R(g||h||f)$ were defined in Lemma 4 to be

$$R(g||f||h) = \int (g-f)\left(\frac{1}{\beta-1}h^{\beta-1} - \frac{1}{\beta-1}f^{\beta-1}\right)d\mu \tag{47}$$
$$R(g||h||f) = \int (g-h)\left(\frac{1}{\beta-1}f^{\beta-1} - \frac{1}{\beta-1}h^{\beta-1}\right)d\mu. \tag{48}$$

Theorem 4 shows that the βD-Bayes general Bayesian updating applied to two likelihood models within the neighbourhood $\mathcal{N}_\epsilon^{\text{TVD}}$ produces posterior predictive inferences that are close in terms of the βD between the two posterior predictive densities $m_h^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})$ and $m_f^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})$ provided Condition 2 holds for data $\boldsymbol{x}_{1:n}$ and priors $\pi_f(\theta_f)$ and $\pi_h(\theta_h)$ and the remainder terms

$$\int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h \tag{49}$$
$$\int\int R(g||h(\cdot;\theta_h)||f(\cdot;\theta_f))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h \tag{50}$$

are small. Similarly to the remainder term in Theorem 2 the size of these remainders will depend on the quality of the approximation of the likelihood models $f(\cdot;\theta_f)$ and $h(\cdot;\theta_h)$ to the DGP $g(\cdot)$. Once again we have focussed on proving stability under a TVD neighbourhood a priori dues to relevance and practicality of considering this neighbourhood in actual applications, see the discussion at the end of Section 4.3

5 Experiments

The experiments in this next section serve to demonstrate the impact of the theorems proved in this paper. In general we find that the stability that is observed in practise when using the TVD and the βD is much tighter than the bounds that we have been able to prove.

5.1 Poisson stability

Firstly we consider discrete likelihood models. Conveniently, working with independent and identically distributed discrete data provides a natural estimate of the DGP, $g_n(x)$, the empirical mass function. As a result these examples provide a way to showcase Theorems 2 and 1 when using the TVD, without having to worry about the computability of a estimate of the data generating density. Additionally, it is straightforward to estimate the TVD between two discrete distributions. Even when these do not have finite support, an accurate estimate of the TVD between the two distributions can be gained by truncating the support at a sufficiently large value.

In order to demonstrate these theorems we consider the Poisson likelihood model with parameter λ, $\text{Poi}(\lambda)$. The Poisson likelihood model is not particularly flexible. It has one parameter, λ, and imposes the property that if $X \sim \text{Poi}(\lambda)$ then $\mathbb{E}[X] = \text{Var[X]} = \lambda$. Often unmodelled heterogeneities can lead to the variance of observe data being larger than its mean. This phenomenon is known as over-dispersion. Another issue with real data is that often the number of zeros observed exceeds those that would be predicted under a Poisson model, a phenomenon known as zero-inflation.

A common method to deal with zero-inflation is to consider fitting a mixture model with an extra component modelling counts at 0.

$$\text{Poi}_{ZI}(y; \lambda, \rho) = (1 - \rho)\,\text{Poi}(y; \lambda) + \rho \mathbb{I}_{y=0}, \tag{51}$$

with $\rho \in (0, 1)$. Additionally, unmodelled heterogeneity in the data could be modelled by the addition of a second 'contaminating' Poisson component

$$\text{Poi}_{\text{mix}}(y; \lambda, \lambda_c, \rho) = (1 - \rho)\,\text{Poi}(y; \lambda) + \rho\,\text{Poi}(y; \lambda_c), \tag{52}$$

with $\rho \in (0, 1)$. For both models imposing an upper-bound on the possible value of $\rho < \hat{\rho}$, for example reflecting the subjective belief that at least $(1 - \hat{\rho}) \times 100\%$ of observations come from the Poisson model, places each likelihood model within a TVD neighbourhood of the standard Poisson likelihood of size $\hat{\rho}$.

We apply these three models to two datasets containing over-dispersed Poisson counts

Data 1 BioChemist - the number of articles produced during the last 3 years of a biochemsitry Ph.D by 915 graduate students[2] (Long, 1990)

Data 2 GrouseTicks - the number of ticks on the heads of 403 red grouse chicks[3] (Elston, Moss, Boulinier, Arrowsmith, & Lambin, 2001)

For these two datasets we implement both Bayes' rule updating (KLD-Bayes) and TVD-Bayes for the Poisson, two-component mixture of Poissons and Zero-inflated Poisson models explained above. We use the empirical mass function to estimate the data generating density. For the BioChemist dataset we set $\hat{\rho} = 0.1$ in order to constrain both the Poisson mixture and the zero-inflated Poisson to be within the $\mathcal{N}^{0.1}_{\text{TVD}}$ neighbourhood of the standard Poisson likelihood. For the GrouseTicks dataset we set $\hat{\rho} = 0.2$. For both datasets we consider priors on $\lambda \sim \mathcal{G}(2, 2)$ and $\rho \sim \text{Unif}[0, \hat{\rho}]$.

[2] downloaded from `http://www.stata-press.com/data/lf2/couart2.dta`.

[3] Available in the '*lme4*' package in R

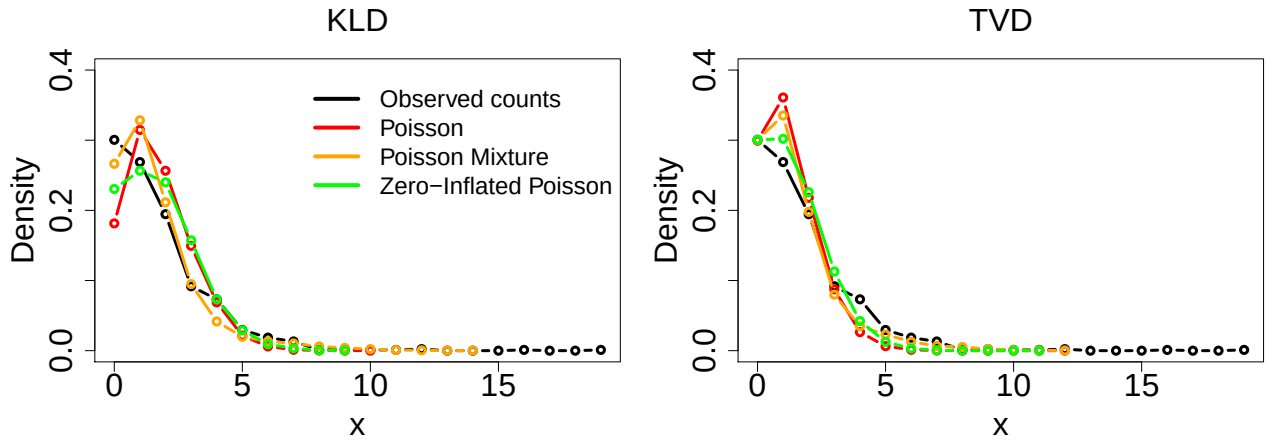

Figure 1: Posterior predictive mass functions for one exchangeable observation when fitting a Poisson likelihood (Poi), a two component mixture of Poissons likelihood (Poi Mix) and a zero-inflated Poisson likelihood (ZI Poi), constrained to fit within the neighbourhood $\mathcal{N}_{\text{TVD}}^{0.1}$, to the **BioChemist dataset**. **Left:** using Bayes' rule (KLD-Bayes) updating. **Right:** using updating aimed at minimising the TVD (TVD-Bayes).

Jack: [Comment in Rossell and Rubio (2018) that TVD of 0.1 is typically irrelevant]

Figure 1 plots the posterior predictive mass functions for one observation obtained by updating using the BioChemist dataset. When using Bayes' rule (KLD-Bayes) the small over-dispersion causes the mean of the Poisson model to be shifted towards the right tail, causing it to fit the modal area of the observed data ($x = 0, 1, 2$) poorly. The zero-inflated Poisson is able to capture this area slightly better while the Poisson mixture appears to provide the best fit for the observed data. As a result these three models provide fairly different approximations to the mode of the observed data, but they do all appear to capture the right tail of the observed data in a similar manner. In contrast, updating using the TVD-Bayes for all three likelihood models provides a more accurate approximation of the data around the mode of the distribution, but fails to capture the heaviness of the right hand tail. The TVD-Bayes fits almost identical posterior predictives around the mode of the observe data for all three likelihood models, only significantly differing at $x = 1$. While the TVD-Bayes appears to be much more stable across the three models near the mode of the data, the KLD-Bayes provides a more stable approximation to the right-hand tail.

These observations are reinforced by the estimates of the TVD between each of these predictive mass function which are presented in Table 1. The observation that the TVD-Bayes achieved greater stability around the mode of the observed data is backed up by uniformly smaller total-variation distances between the predictives when compared with the same distances between the KLD-Bayes predictives. The fact that the TVD values between the Poisson likelihood model and the Poisson mixture, and between the Poisson and the zero-inflated Poisson were both below 0.1, the upper-bound on the distance between the models a priori, demonstrates the result of Theorem 2. We also note this happens despite the posterior density for ρ placing most of its density towards this upper-bound for both the Poisson mixture and the zero-inflated Poisson. Despite the a priori TVD between these likelihood models being upper-bounded by 0.1, the KLD-Bayes predictive distributions from the Poisson and Poisson mixture models have a TVD of greater than 0.1, suggesting here that Bayes' rule is causing these inferences to divergence!

Table 1: Estimates of the TVD between the posterior predictive mass functions for one exchangeable observation when fitting a Poisson likelihood (Poi), a two component mixture of Poissons likelihood (Poi Mix) and a zero-inflated Poisson likelihood (ZI Poi), constrained to fit within the prior neighbourhood $\mathcal{N}_{\text{TVD}}^{0.1}$, to the **BioChemist dataset**. **Left:** using Bayes' rule (KLD-Bayes) updating. **Right:** using updating aimed at minimising the TVD (TVD-Bayes).

KLD-Bayes	Poi	Poi Mix	ZI Poi
Poi	-	0.1283	0.0748
Poi Mix	-	-	0.1317

TVD-Bayes	Poi	Poi Mix	ZI Poi
Poi	-	0.0533	0.0589
Poi Mix	-	-	0.0684

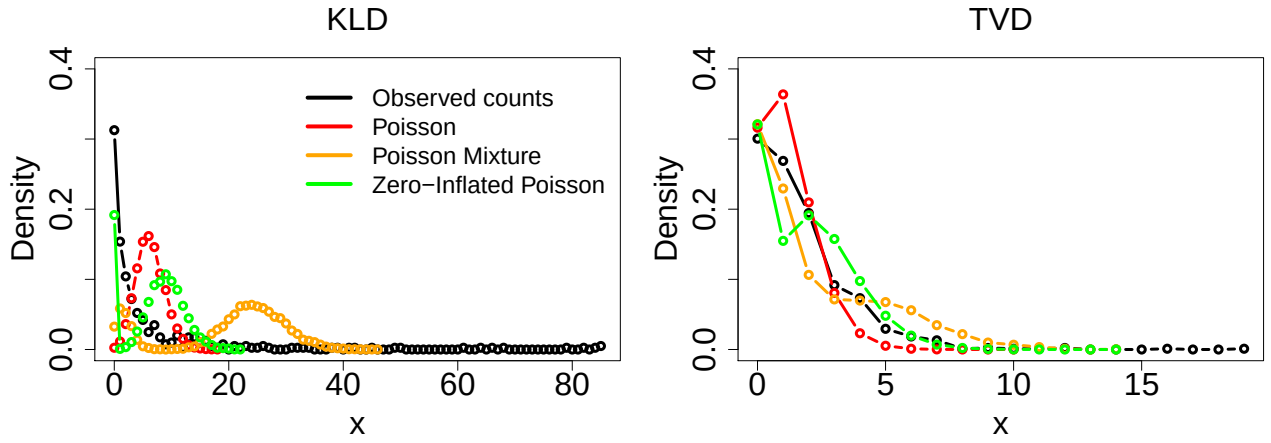

Figure 2: Posterior predictive mass functions for one exchangeable observation when fitting a Poisson likelihood (Poi), a two component mixture of Poissons likelihood (Poi Mix) and a zero-inflated Poisson likelihood (ZI Poi), constrained to fit within the neighbourhood $\mathcal{N}_{\text{TVD}}^{0.2}$, to the **GrouseTicks dataset**. **Left:** using Bayes' rule (KLD-Bayes) updating. **Right:** using updating aimed at minimising the TVD (TVD-Bayes).

Figure 2 shows the corresponding predictive mass functions produced from the GrouseTicks dataset. While for the BioChemist dataset even under the KLD-Bayes, all three models provided a reasonable approximation of to the distribution of the observed data, this is no longer the case for the GrouseTicks data. Under the KLD-Bayes all three models attempt to strike a balance between capturing the large model at 0 and the very long right hand tail. They however achieve this in very different ways, producing very different predictive distributions. In juxtaposition, the TVD-Bayes is able to ignore the long right-hand tail and focus on the mode of the observed data. In doing so all three likelihood models are able to provide much more satisfactory approximations of the observed data and also much greater stability in this approximation across the three likelihoods.

These observations are once again backed up by the TVD values between the predictive distributions presented in Table 2. The TVD values between the KLD-Bayes predictive distributions are huge. In fact they are approaching the upper bound for the TVD at 1. In contrast the values for the TVD-Bayes are much smaller and the fact that the total-variation distances between the Poisson and mixture of Poissons and the Poisson and the zero inflated Poisson are both less than 0.2 again shows the impact of Theorem 2, and the fact that this theorem still holds even when all three models are poorer approximations of the DGP.

Table 2: Estimates of the TVD between the posterior predictive mass functions for one exchangeable observation when fitting a Poisson likelihood (Poi), a two component mixture of Poissons likelihood (Poi Mix) and a zero-inflated Poisson likelihood (ZI Poi), constrained to fit within the prior neighbourhood $\mathcal{N}_{\text{TVD}}^{0.2}$, to the **GrouseTicks dataset**. **Left:** using Bayes' rule (KLD-Bayes) updating, **Right:** using updating aimed at minimising the TVD (TVD-Bayes).

KLD-Bayes	Poi	Poi Mix	ZI Poi
Poi	-	0.8846	0.4636
Poi Mix	-	-	0.8879

TVD-Bayes	Poi	Poi Mix	ZI Poi
Poi	-	0.1373	0.1481
Poi Mix	-	-	0.1659

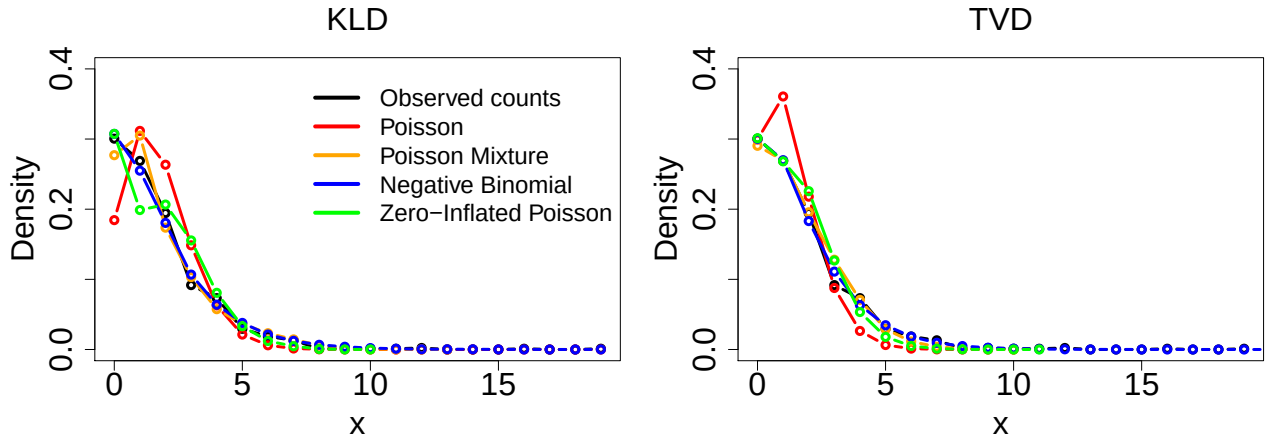

Figure 3: Posterior predictive mass functions for one exchangeable observation when fitting a Poisson likelihood (Poi), a two component mixture of Poissons likelihood (Poi Mix), a zero-inflated Poisson likelihood (ZI Poi) and a negative-binomial likelihood (NB), unconstrained a priori, to the **BioChemist dataset**. **Left:** using Bayes' rule (KLD-Bayes) updating. **Right:** using updating aimed at minimising the TVD (TVD-Bayes).

5.1.1 Unconstrained a priori

In practise a further common way to account for over-dispersion is to use a negative-binomial likelihood model. Negative-binomial models are traditionally interpreted as modelling the the number of failures before a certain number of successes in repeated independent trials. However, they can also be parametrised in terms of a mean number of counts, similar to the Poisson likelihood. It is not straightforward to build a TVD neighbourhood containing the negative-binomial and the Poisson likelihood models and therefore we did not implement this above to illustrate the theorems of this Chapter. Instead we now implement the negative-binomial likelihood alongside the Poisson likelihood, the mixture of Poissons and the zero-inflated Poisson, where we no longer constrain the value of ρ to be less than some threshold. As a result these likelihoods do not fall into any of our a priori neighbourhoods, but Figures 3 and 4 and Tables 3 and 4 show that the TVD-Bayes is still much more stable in terms of TVD than thee KLD-Bayes.

5.2 Fixing the quartiles

Our next example takes inspiration from the approach outlined to belief elicitation in O'Hagan (2012). It is argued there 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 based on the fact that humans are generally able to accurately

Table 3: Estimates of the TVD between the posterior predictive mass functions for one exchangeable observation when fitting a Poisson likelihood (Poi), a two component mixture of Poissons likelihood (Poi Mix), a zero-inflated Poisson likelihood (ZI Poi) and a negative-binomial likelihood (NB), unconstrained a priori, to the **BioChemist dataset**. **Top:** using Bayes' rule (KLD-Bayes) updating. **Bottom:** using updating aimed at minimising the TVD (TVD-Bayes).

KLD-Bayes	Poi	Poi Mix	ZI Poi	NB
Poi	-	0.1470	0.1696	0.1814
Poi Mix	-	-	0.1389	0.0540
ZI Poi	-	-	-	0.0918

TVD-Bayes	Poi	Poi Mix	ZI Poi	NB
Poi	-	0.1225	0.0925	0.1256
Poi Mix	-	-	0.0409	0.0369
ZI Poi	-	-	-	0.0594

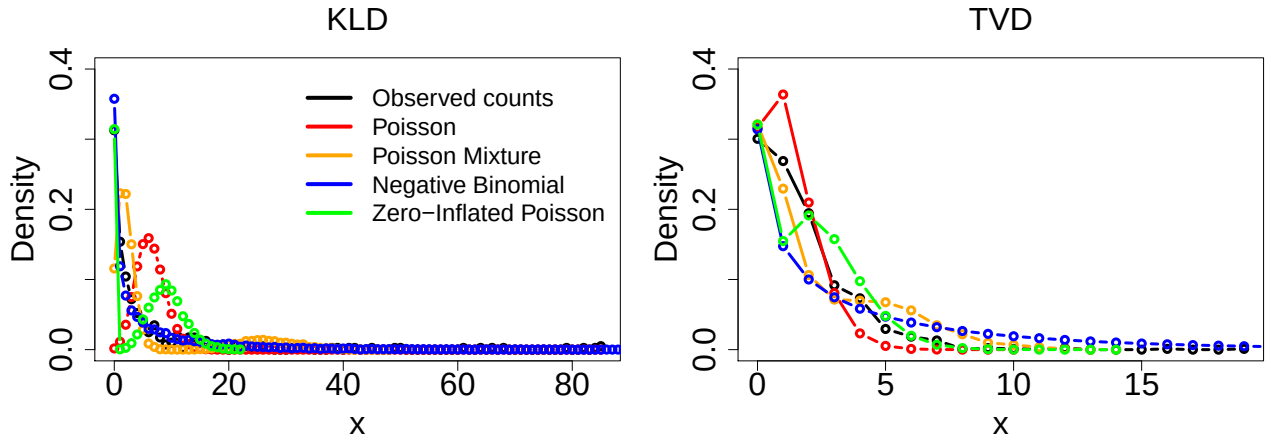

Figure 4: Posterior predictive mass functions for one exchangeable observation when fitting a Poisson likelihood (Poi), a two component mixture of Poissons likelihood (Poi Mix), a zero-inflated Poisson likelihood (ZI Poi) and a negative-binomial likelihood (NB), unconstrained a priori, to the **GrouseTicks dataset**. **Left:** using Bayes' rule (KLD-Bayes) updating. **Right:** using updating aimed at minimising the TVD (TVD-Bayes).

make judgements of equal probability. The rest of this distribution is then filled in arbitrarily by the statistician facilitating the analysis. For example if the upper and lower quartiles are believed to be a similar distance from the median then a Gaussian distribution is typically assumed, where as in principle there are a huge number of distributions sharing these properties.

O'Hagan (2012) justify this 'cavalier' approach of arbitrarily filling in the rest of the density given the medians and quartiles as often "adequate for the purpose for which the elicitation is being performed". The reason for this is that any two distributions with the same mean, modality and quartiles will look very similar, see Figure 5. This may very well be the case if these distributions are going to be directly used to calculate estimates of bounded expected utilities. However Lemma 1 suggests that much more than identical medians and quartiles will be required to ensure the stability of Bayes' rule updating. This example aims to demonstrate this and the stability that can be afforded to such arbitrary assumption when using the βD-Bayes.

Here we consider the stability of Bayesian inference to the choice between a Gaussian and a Student's t-likelihood. The neighbourhood of likelihood models is given by

Table 4: Estimates of the TVD between the posterior predictive mass functions for one exchangeable observation when fitting a Poisson likelihood (Poi), a two component mixture of Poissons likelihood (Poi Mix), a zero-inflated Poisson likelihood (ZI Poi), and a negative-binomial likelihood (NB), unconstrained a priori, to the **GrouseTicks dataset**. **Top:** using Bayes' rule (KLD-Bayes) updating. **Bottom:** using updating aimed at minimising the TVD (TVD-Bayes).

KLD-Bayes	Poi	Poi Mix	ZI Poi	NB
Poi	-	0.7558	0.5112	0.6449
Poi Mix	-	-	0.7976	0.4665
ZI Poi	-	-	-	0.4041

TVD-Bayes	Poi	Poi Mix	ZI Poi	NB
Poi	-	0.2467	0.2272	0.3340
Poi Mix	-	-	0.1988	0.1476
ZI Poi	-	-	-	0.2286

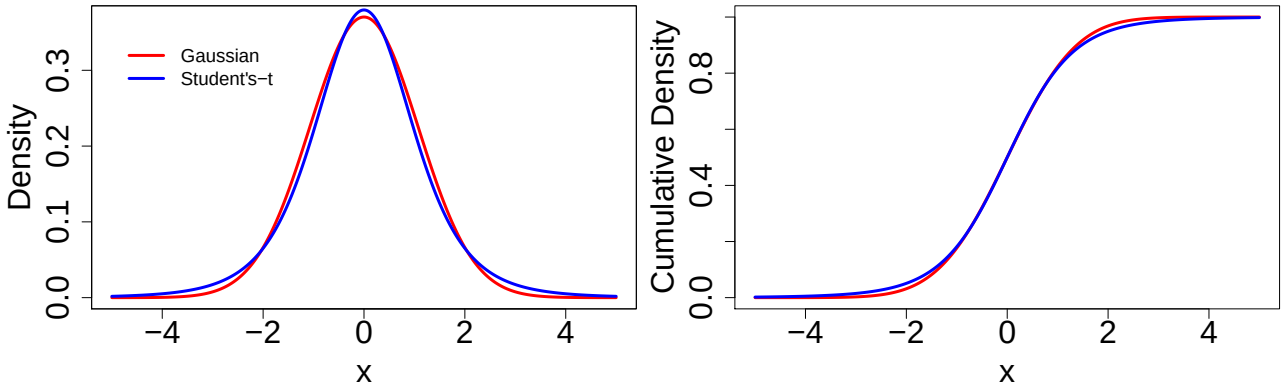

Figure 5: Probability density function (pdf) and cumulative density function (cdf) of a Gaussian $f(x;\theta_f) = \mathcal{N}\left(x;\mu_f,\sigma^2_{adj}\sigma^2_f\right)$ and a Student's-t $h(x;\theta_h) = t_\nu(x;\mu_h,\sigma^2_h)$ random variable, with $\mu = 0$ and $\sigma^2 = 1$. The parameters $\nu = 5$ and $\sigma^2_{adj} = 1.16$ where chosen such that the two densities have the same median and quartiles which conveniently ensured that $(f(x;\theta_f), h(x;\theta_h)) \in \mathcal{N}^{\text{TVD}}_\epsilon$ (defined in (15)) with $\epsilon = 0.043$ for all $\mu = \mu_f = \mu_h$ and $\sigma^2 = \sigma^2_f = \sigma^2_h$. The two likelihoods are accurate within any sensible drawing accuracy. So requiring a DM to distinguish between these two is unreasonable

$$f(x;\theta_f) := \mathcal{N}\left(x;\mu_f,\sigma^2_f \times \sigma^2_{adj}\right) \tag{53}$$
$$h(x;\theta_h) := \text{Student's} - t_\nu\left(x;\mu_h,\sigma^2_h\right) \tag{54}$$

where we set σ^2_{adj} for a given ν to match the quartiles of the two distribution for all $\mu = \mu_f = \mu_h$ and $\sigma^2 = \sigma^2_f = \sigma^2_h$. For $\nu = 5$ we find by optimisation that $\sigma^2_{adj} = 1.16$. In fact we can use the representation

$$\text{TVD}(g,f) = \int_{g>f}(g-f) = \int_{g>f} g - \int_{g>f} f \tag{55}$$

to estimate that this neighbourhood also corresponds to a $\mathcal{N}^{\text{TVD}}_\epsilon$ neighbourhood with $\epsilon = 0.043$ as defined in Eq. (15). Fig 5 plots the probability density function and cumulative distribution function of f and h for $\mu = 0$, $\sigma^2 = 1$, $\nu = 5$ and $\sigma^2_{adj} = 1.16$ defined above. This shows how similar the Gaussian and Student's-t likelihood are. They are clearly within drawing accuracy of each other and it seems unreasonable to require any DM to be able to distinguish between the two.

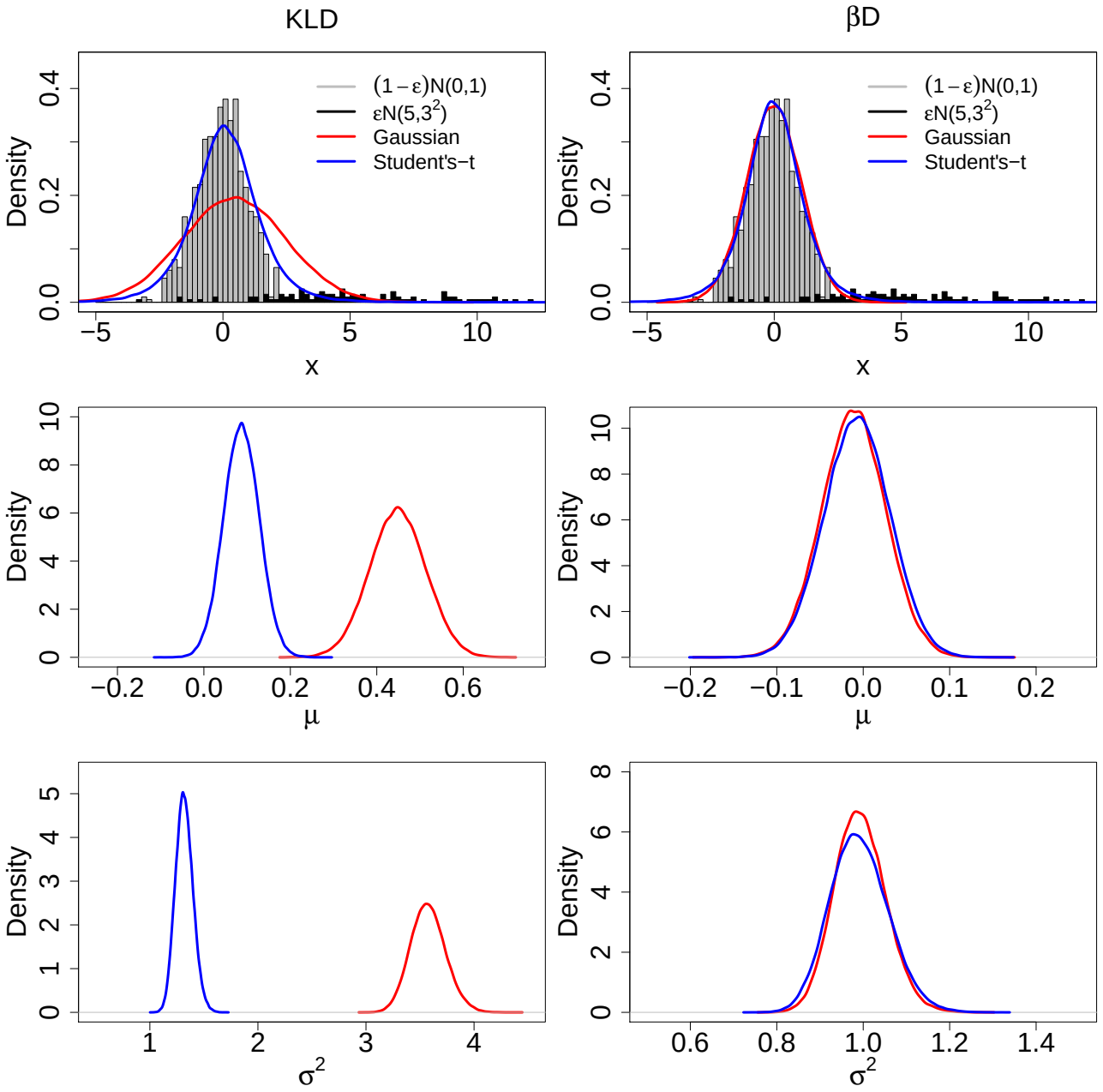

Figure 6: Posterior predictive distributions and parameter posterior distributions for (μ, σ^2) under Bayes' rule updating (KLD-Bayes) (**left**) and βD-Bayes (**right**) under the likelihood functions $f(x;\theta_f) = \mathcal{N}\left(x;\mu,\sigma^2_{adj}\sigma^2\right)$ (red) and $h(x;\theta_h) = t_\nu(x;\mu,\sigma^2)$ blue where $\nu = 5$ and $\sigma^2_{adj} = 1.16$. Both the parameter and predictive inference is stable across $\mathcal{N}_\epsilon^{\text{TVD}}$ under the βD-Bayes and is not under Bayes' rule (KLD-Bayes)

5.2.1 A toy experiment

To investigate the stability of inferences across the neighbourhood $\mathcal{N}_\epsilon^{\text{TVD}}$ of likelihood models, we generated $n = 1000$ observations from the ϵ-contamination model in Example **??** with $(\epsilon = 0.1, \mu_u = 0, \sigma^2_u = 1^2, \mu_c = 5, \sigma^2_c = 3^2)$. We then conducted Bayesian updating under the Gaussian and Student's-t likelihood using both Bayes' rule and the βD-Bayes and 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 $(a_0 = 0.01, b_0 = 0.01, \mu_0 = 0, v_0 = 10)$. Fig. 6 plots the parameter posterior and posterior predictive distributions for both models under both updating mechanisms.

Table 5: Estimates of the energy distance between the Bayesian predictive distributions when using a Gaussian and Student's-t likelihood under Bayes' rule (KLD) and inference minimising the βD

E-distance	KLD	βD
	0.125	2.13×10^{-3}

The left hand side of Fig. 6 demonstrates what most statistical practitioner expect when comparing the performance of a Gaussian and a Student's-t under outlier contamination (O'Hagan, 1979). Under the Student's-t likelihood the inference is much less affected by the outlying contamination than under the Gaussian likelihood. The parameter μ is shifted less towards the contaminant population and the parameter σ^2 is inflated much less by the outlying contamination. In short, very different inferences are produced using a Student's-t and a Gaussian under outlier contamination. Updating using the βD-Bayes presents a striking juxtaposition to this! The βD-Bayes produces almost identical posteriors for both μ and σ^2 resulting in almost identical posterior predictive densities. The βD-Bayes is clearly stable to the selection of either a Gaussian or Student's-t likelihood where Bayes' rule updating is not.

Not only does the βD inference appear to be more stable across $\mathcal{N}_\epsilon^{\text{TVD}}$ here, but we also argue that the βD predictive better captures the majority of the DGP than either of the predictives do under the Bayes' rule (KLD).

Estimating the TVD or the βD between the two predictves distributions is hampered by the fact that they are not available in closed form. However the energy distance Székely and Rizzo (2013) provides a metric that can be easily estimated from samples of the predictive. Table 5 presents the energy distance between the two predictives

Jack: [Should we have both the stability regressions and the changepoint detection or just a properly worked through changepoint detection]

5.2.2 Stability in linear regression

We extend the toy example above to situations where the Gaussian and Student's-t densities are used for error distributions in linear regression. Here some univariate response Y is regressed on a vector of predictors $\boldsymbol{X} = (X_1, \ldots, X_p)$ as follows

$$Y = \boldsymbol{X}\boldsymbol{\theta}^T + \epsilon, \tag{56}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ is a vector of regression coefficients and the errors ϵ are considered independent and identically distributed with mean 0. Similarly to above we consider that the DM is unable to decide between

$$\epsilon \sim \mathcal{N}\left(0, \sigma^2 \times \sigma^2_{adj}\right) \tag{57}$$

$$\epsilon \sim t_\nu\left(0, \sigma^2\right), \tag{58}$$

where we continue to consider $\nu = 5$ and $\sigma^2_{adj} = 1.16$. We apply these two linear models to four datasets from the UCI repository (Lichman et al., 2013), providing a range of sample sizes and number of predictors. The data sets are described below

- Energy: 768 observations seeking to understand the relationship between the cooling load re-

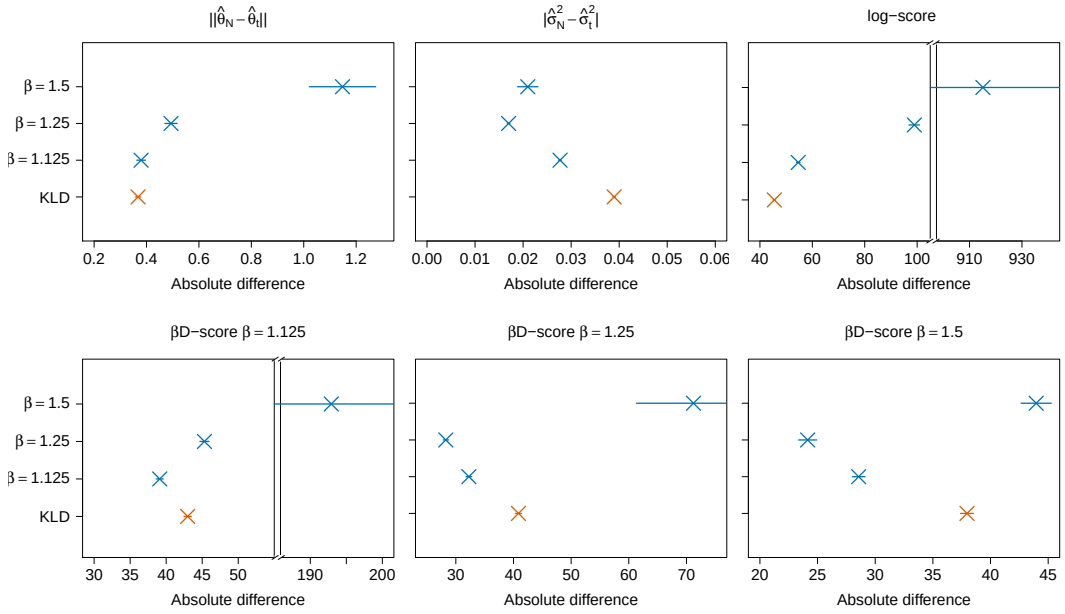

Figure 7: Plots comparing the stability of Bayes' rule and βD-Bayes for $\beta = 1.125, 1.25, 1.5$ inference for linear regression models under either Gaussian or Student's-t error distributions applied to the **Energy dataset**. **From left to right**: L_1 norm of the difference between the posterior means for the regression coefficients $\boldsymbol{\theta}$, absolute difference between the posterior mean estimates for the residual variances σ^2, absolute difference in predictive log-score applied to the training set, absolute difference in predictive βD-score applied to the training set $\beta = 1.125, 1.25, 1.5$. All averaged over 50 subsets of training points.

quirements of buildings as a function of 7 other building parameters.

- Power: 9568 observations seeking to understand the relationship between the electrical output from a combined cycle power plant as a function of 4 other power plant parameters.
- Concrete: 1030 observations seeking to understand the relationship between concrete's compression strength as a function of 8 other features of the concrete.
- BostonHousing: 506 observations seeking to estimate the relationship between the median property value in neighbourhoods of Boston and 13 features of those neighbourhood.

The response and all of the predictors were standardised to each have mean 0 and variance 1.

In order to asses the stability of Bayes' rule updating and updating using the βD-loss we produce $N = 50$ datasets by taking a random 80% of each dataset. The figures below present the absolute difference between several posterior and predictive metrics in order to quantify the stability of the KLD-Bayes and βD-Bayes with $\beta = 1.125$, 1.25 and 1.5 updating under the Gaussian and Student's-t likelihood. These metrics are the L_1 norm of the difference between the posterior mean estimates for the regression parameters θ, the absolute difference between the posterior mean for the residual variances σ^2, the absolute difference between the predictive log-score applied to the training sets and the absolute difference between the predictive βD-loss for $\beta = 1.125$, 1.25, 1.5 also applied to the training sets. We note that the theorems of this chapter say noting about the stability in terms of the parameter estimates and the log-score.

Figure 7 compares these six stability metrics for the four updating methods we consider above when

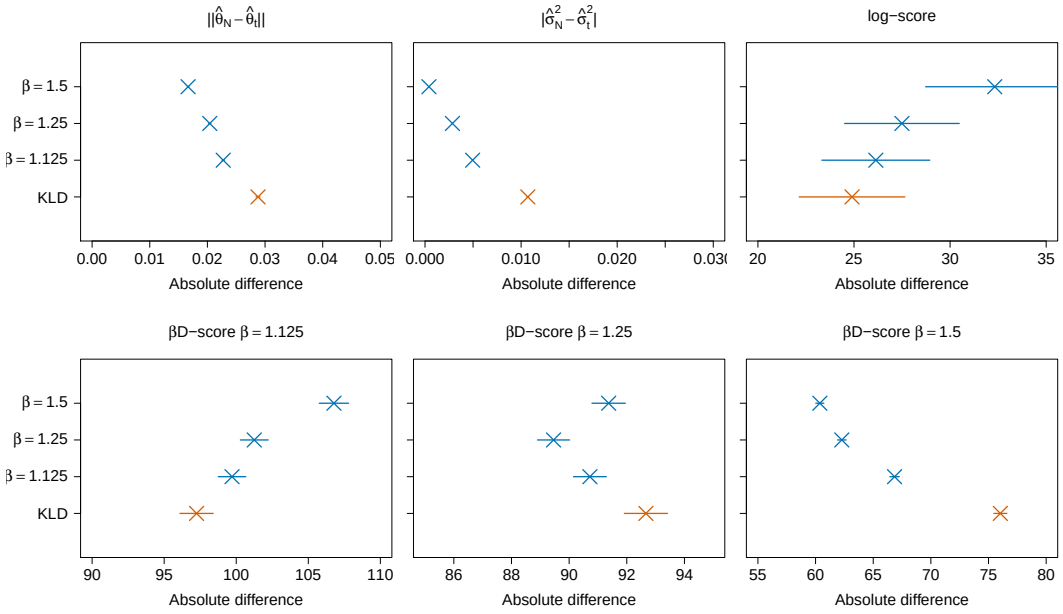

Figure 8: Plots comparing the stability of Bayes' rule and βD-Bayes for $\beta = 1.125, 1.25, 1.5$ inference for linear regression models under either Gaussian or Student's-t error distributions applied to the **Power dataset**. **From left to right**: L_1 norm of the difference between the posterior means for the regression coefficients $\boldsymbol{\theta}$, absolute difference between the posterior mean estimates for the residual variances σ^2, absolute difference in predictive log-score applied to the training set, absolute difference in predictive βD-score applied to the training set $\beta = 1.125, 1.25, 1.5$. All averaged over 50 subsets of training points.

applied to the Energy dataset. The Bayes' rule updating, minimising the KLD, appears to produce the most stable inference according to the log-score and interestingly also in terms of the estimates of the regression parameters θ. It is unsurprising that Bayes' rule updating provides the most stable inference in terms of the log-score . The log-score focuses mainly on how the models approximate the tails of the observed data and therefore this shows that Bayes' rule produces the most stable inference in the tails of the DGP. This is similar to what was observed in the Poisson experiments above. The βD-Bayes for $\beta = 1.125$ and $\beta = 1.25$ produce the most stable inference according to the βD-loss for $\beta = 1.125$ and $\beta = 1.25$ respectively. It appears as though $\beta = 1.5$ is too large for inference in this case as it produces the least stable inference by all metrics. Even the βD-loss using the same β that was used for the updating. As the parameter β of the βD-loss increases away from 1 (the value corresponding to the log-score), the difference in this measure focuses less on the stability of the approximation of the tails of the observed data, and more on the stability of the approximation to the modal part of the data. The fact that the βD-Bayes for $\beta = 1.125$ and $\beta = 1.25$ provide more stable inference than Bayes' rule according to the βD-loss with $\beta = 1.25$ and $\beta = 1.5$ show that these methods are producing more stable inference for the modal areas of the observed data, despite being less stable in terms of the parameter estimate for the mode.

Figure 8 compares these six stability metrics for the four updating methods we consider above when applied to the Power dataset. Here Bayes' rule updating (KLD-Bayes) achieves the most stable inference under the log-score and the also the βD-loss for $\beta = 1.125$. However, the βD-Bayes updating for all three values of β produces more stable inference for the other four metrics, again suggesting that the βD-Bayes produces a more stable approximation around the high density regions of the observed data.

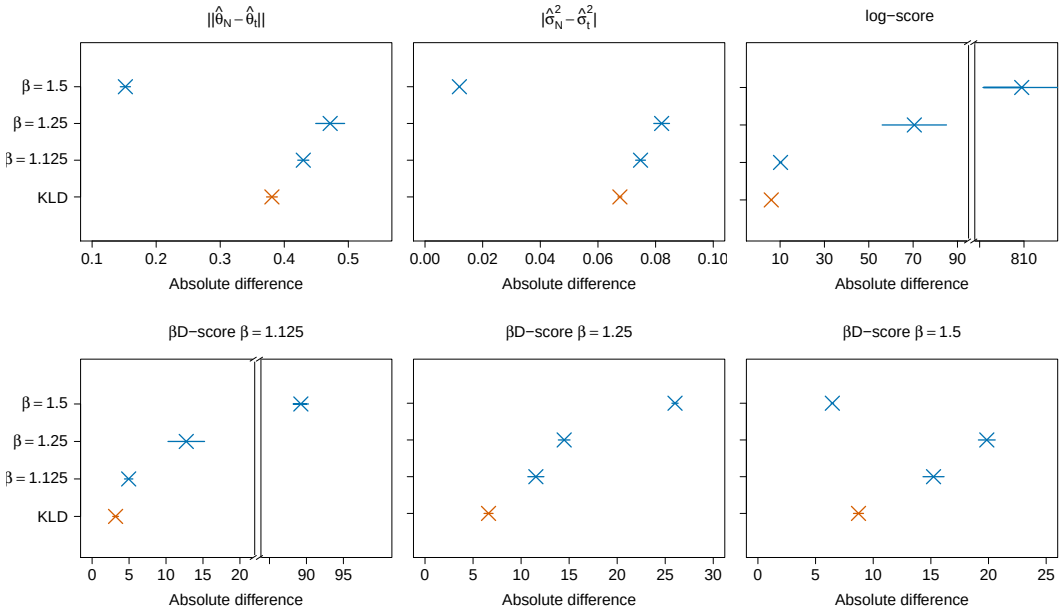

Figure 9: Plots comparing the stability of Bayes' rule and βD-Bayes for $\beta = 1.125, 1.25, 1.5$ inference for linear regression models under either Gaussian or Student's-t error distributions applied to the **Concrete dataset**. **From left to right**: L_1 norm of the difference between the posterior means for the regression coefficients $\boldsymbol{\theta}$, absolute difference between the posterior mean estimates for the residual variances σ^2, absolute difference in predictive log-score applied to the training set, absolute difference in predictive βD-score applied to the training set $\beta = 1.125, 1.25, 1.5$. All averaged over 50 subsets of training points.

We observed something quite different for the Concrete dataset in Figure 9. Here Bayes' rule updating provides the most stable inference according to the βD-loss for $\beta = 1.125$ and $\beta = 1.25$ as well as the log-score, and is only less stable on the other 3 metrics than the βD-Bayes with $\beta = 1.5$. However on all metrics the βD-Bayes for $\beta = 1.125$ and $\beta = 1.25$ are never much less stable than the log-score updating. This suggests that the conditional response of the Concrete dataset is reasonably approximated by either a Gaussian likelihood or a Student's-t likelihood. For example, these likelihoods differ a priori in TVD by $\epsilon = 0.043$. If this difference was observed in terms of the βD-loss function a posteriori for each training data point, then the difference in these training scores would accumulate to just over 35. All four updating methods generally appear to produce inference that is more stable that this threshold!

Lastly, Figure 10 demonstrates the corresponding results for the BostonHousing dataset. Here, the Bayes' rule updating (KLD-Bayes) is only the most stable according to the log-score and is generally much less stable according the the other five metrics. This goes to show that although the KLD-Bayes provides stable inference for the tails of the observed data, it can provide fairly unstable inference when the high density regions of the observed data are of interest. In these cases using the βD-Bayes is shown to be much more stable.

Figures 9 and 10 demonstrate, similarly to the experiments of Chapter **??**, the asymmetric nature of possible gains and losses of using these alternative divergence methods instead of Bayes' rule. When Bayes' rule performs well, these alternative methods can be shown to only be marginally worse, while when Bayes' rule performs poorly, these alternative methods can be shown to improve this performance vastly.

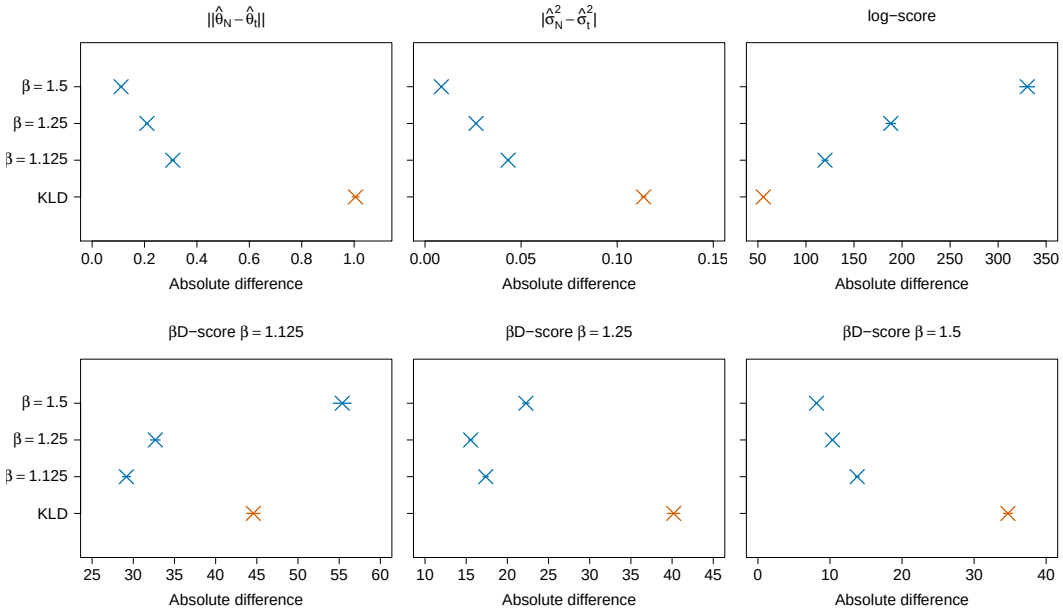

Figure 10: Plots comparing the stability of Bayes' rule and βD-Bayes for $\beta = 1.125, 1.25, 1.5$ inference for linear regression models under either Gaussian or Student's-t error distributions applied to the **BostonHousing dataset**. **From left to right**: L_1 norm of the difference between the posterior means for the regression coefficients $\boldsymbol{\theta}$, absolute difference between the posterior mean estimates for the residual variances σ^2, absolute difference in predictive log-score applied to the training set, absolute difference in predictive βD-score applied to the training set $\beta = 1.125, 1.25, 1.5$. All averaged over 50 subsets of training points.

5.3 The stability of BOCPD

Jack: [I have currently only applied this to the univariate 'real-world' dataset that we used in our NeurIPS paper. The plan would be to apply this to some actual real world data on airpolution that we have which is actually high-dimensional (well 29 dimensions and a few thousad observations). I just need to work this out computationally!]

We next investigate the stability of Bayesian updating in the challenging domain of Bayesian on-line changepoint detection (BOCPD). Consider a multivariate stream of data arriving at discrete time points $\{y_t\}_{t=1}^{\infty} = \{y_1, y_2, \ldots\}$, BOCPD algorithms (Adams & MacKay, 2007; Fearnhead & Liu, 2007; Knoblauch & Damoulas, 2018) produce inference about the location of changes in the sample distribution of the observations in real time, that is as soon as possible after they occur. These methods are able to run on-line by using simple conjugate models for the likelihood and introducing a random-variable called the run-length r_t. At each time point r_t provides the time since the last changepoint. This provides scalability as when a new observation occurs r_t either grows by 1, indicating that the new observation is consistent with the previous r_t observations, or shrinks to 0, indicating that the new observation is from a new regime. Being Bayesian we seek to produce a posterior distribution over the parameters indexing the likelihood density in between changepoints, and on the run-length at each time point. The between changepoint inference happens like standard Bayesian updating, using only those observations within the current segment to produce a posterior for the model parameters. This posterior is then used to produce one-step-ahead posterior predictive densities which in turn quantifies the evidence for a specific run-length available in the next observation. See Adams and MacKay (2007) for a thorough review of BOCPD.

Knoblauch et al. (2018) noticed that BOCPD's formulation using the one-step-ahead predictive densities as a likelihood and their reliance on using simple models for computational convenience lead to it being very non-robust to outliers. They robustified both the model parameter posterior and run-length posteriors using the βD loss function (Eq. 10) (RBOCPD), instead of the log-score associated with traditional Bayesian updating used in BOCPD. This lead to a significant drop in false positive detections of changepoints.

Here we seek to identify whether robustifying the BOCPD with the βD not only made the algorithm robust to outliers, but also produced posterior inference that is more stable to the choice of likelihood model. The simple time series models used by (Adams & MacKay, 2007; Fearnhead & Liu, 2007; Knoblauch & Damoulas, 2018; Knoblauch et al., 2018) are vector auto-regressions and therefore it is natural to consider in this setting whether a Gaussian or a Student's-t distribution are suitable for the errors.

Similarly to above we set the degrees of freedom of the Student's-t likelihood to be $\nu = 5$ and additionally set $\sigma^2_{adj} = 1.16$ in order ensure that the Gaussian and Student's-t likelihoods are in the neighbourhood $\mathcal{N}^{\text{TVD}}_{0.05}$. We then applied the BOCPD and the βD RBOCPD to a section of the the univariate 'well-log' data set, first studied by Ruanaidh, Joseph, and Fitzgerald (1996), and compared the run-length posterior inferences under the two likelihoods. As was mentioned above the run-length posterior uses the one-step-ahead posterior predictive density as its likelihood and therefore, stability of the one-step-ahead posterior predictive distributions should provide stability in the run-length posterior. Additionally the run-length posterior is discrete and has finite support so it is straightforward to estimate the TVD between these.

Figures 11 and 12 plot the extract of the '*well-log*' dataset alongside the run-length posterior from both the Gaussian and Student's-t error models under the KLD-Bayes and the βD-Bayes respectively. Under both the Gaussian and Student's-t the KLD-Bayes posterior initially declares a CP when observing the outlying segment at $t = 390$. However the Student's-t error model is more quickly, compared with the Gaussian error model, able to identify these observations as an outlying segment and therefore the run-length posterior reverts to thinking there was no CP. In fact, under the Gaussian error model even after 50 further non-outlying observations the run-length posterior still has non-negligible mass for a CP at $t = 390$. This is reflected in Figure 13 which plots the TVD between the two run-length posterior from each likelihood. Shortly after the outlying segment the TVD between the two run-length posteriors spikes to nearly 0.5 indicating that the Student's-t no longer thinks there was a CP at $t = 390$ while the Gaussian does. It takes till $t = 450$ for these posterior to stabilise again.

Alternatively, under the βD-Bayes neither of the run-length posterior declares a CP around the outlying segment and as a result the TVD between the two run-length posterior in Figure 13 is very small throughout the segment. The Gaussian error model places slightly more mass at the possibility of a CP at the outlying segment than the Student's-t , which is reflected by the small increase in TVD between the two run-length posteriors, but this is much less than the under the KLD-Bayes and lasts for only several observations. Figure 13 clearly demonstrates the increased stability of the βD-Bayes run-length posterior compared with the KLD-Bayes which is interpreted as deriving from increased stability of the posterior predictive distributions.

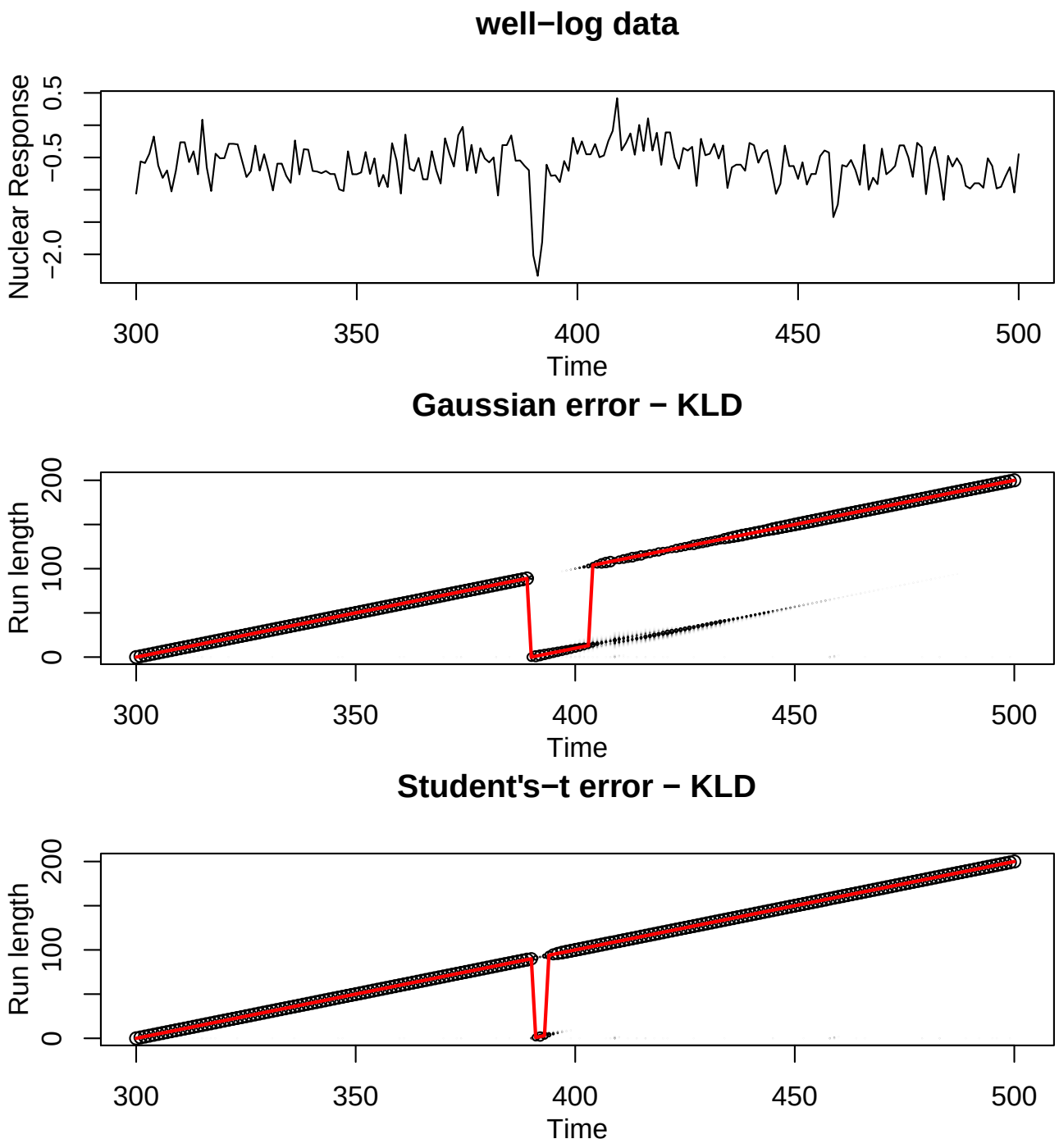

Figure 11: **Top**: Section of the *'well-log'* dataset from $T = 300$ to $T = 500$. **Middle**: run-length posterior distribution under a Gaussian error model and the Bayes' rule updating (KLD-Bayes). **Bottom**: run-length posterior distribution under a Student's-t error model and Bayes' rule updating (KLD-Bayes)

6 Further Work

This paper demonstrates that inference aimed at minimising the TVD and HD (metrics) and also the βD, can be shown to be stable to interpretable neighbourhoods of likelihood models. While, the same cannot be said for inference under Bayes' rule, minimising the KLD. Stability to such modelling selections is a natural and important property for conducting inference in the M-OPEN world. These tell a DM exactly how far they need to go with their belief elicitation to be sure that any interpolation

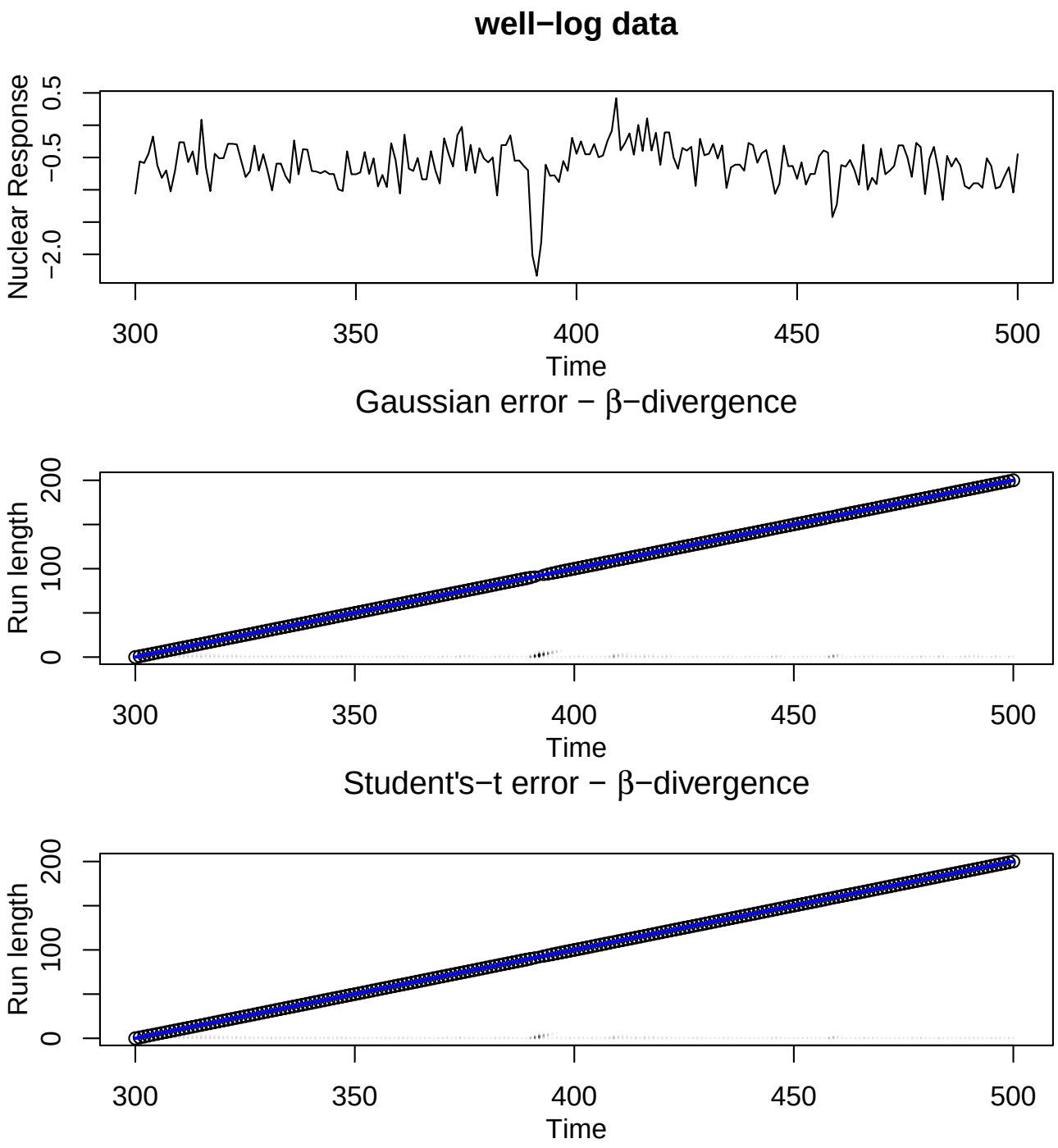

Figure 12: **Top**: Section of the *'well-log'* dataset from $T = 300$ to $T = 500$. **Middle**: run-length posterior distribution under a Gaussian error model and the βD-Bayes updating. **Bottom**: run-length posterior distribution under a Student's-t error model and the βD-Bayes updating

does not significantly effect the posterior conclusions.

References

Adams, R. P., & MacKay, D. J. (2007). Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*.

Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, J. M., Cano, J. A., ... others (1994). An overview of robust bayesian analysis. *Test*, *3*(1), 5–124.

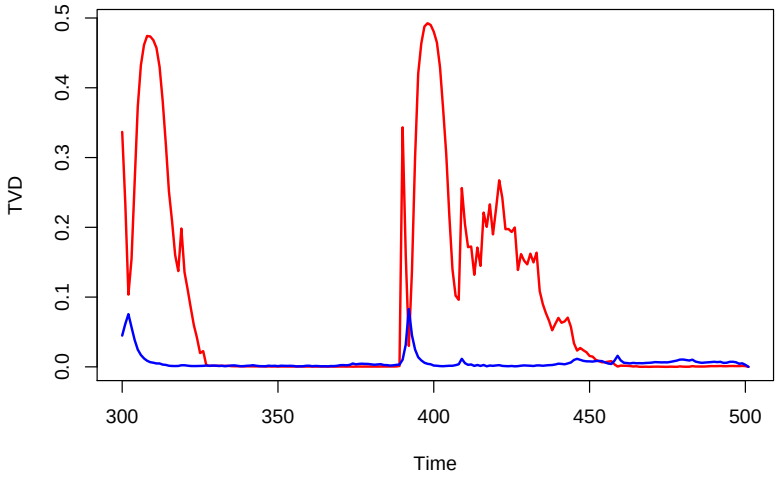

Figure 13: TVD between the run-length posterior when using a Gaussian error model and a Student's-t error model when using the KLD-Bayes and the βD-Bayes

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.

Bernardo, J. M., & Smith, A. F. (2001). *Bayesian theory*. IOP Publishing.

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

Chernozhukov, V., & Hong, H. (2003). An mcmc approach to classical estimation. *Journal of Econometrics*, *115*(2), 293–346.

Dawid, A. P. (2007). The geometry of proper scoring rules. *Annals of the Institute of Statistical Mathematics*, *59*(1), 77–93.

Elston, D., Moss, R., Boulinier, T., Arrowsmith, C., & Lambin, X. (2001). Analysis of aggregation, a worked example: numbers of ticks on red grouse chicks. *Parasitology*, *122*(5), 563–569.

Fearnhead, P., & Liu, Z. (2007). On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(4), 589–605.

Ghosh, A., & Basu, A. (2016). Robust bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, *68*(2), 413–437.

Gilboa, I., & Schmeidler, D. (1989). Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, *18*(2), 141–153.

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

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

Goldstein, M., et al. (2006). Subjective bayesian analysis: principles and practice. *Bayesian Analysis*, *1*(3), 403–420.

Grünwald, P. (2016). Safe probability. *arXiv preprint arXiv:1604.01785*.

Grünwald, P. D., & Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *Annals of Statistics*, 1367–1433.

Gustafson, P., & Wasserman, L. (1995). Local sensitivity diagnostics for bayesian inference. *The Annals of Statistics*, *23*(6), 2153–2167.

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

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

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations* (Vol. 3).

Holmes, C., & Walker, S. (2017). Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, *104*(2), 497–503.

Hooker, G., & Vidyashankar, A. N. (2014). Bayesian model robustness via disparities. *Test*, *23*(3), 556–584.

Jewson, J., Smith, J., & Holmes, C. (2018). Principles of Bayesian inference using general divergence criteria. *Entropy*, *20*(6), 442.

Knoblauch, J., & Damoulas, T. (2018). Spatio-temporal bayesian on-line changepoint detection with model selection. In *International Conference on Machine Learning (ICML).*

Knoblauch, J., Jewson, J., & Damoulas, T. (2018). Doubly robust Bayesian inference for non-stationary streaming data using β-divergences. In *Advances in Neural Information Processing Systems (NeurIPS)* (pp. 64–75).

Knoblauch, J., Jewson, J., & Damoulas, T. (2019). Generalized variational inference. *arXiv preprint arXiv:1904.02063*.

Lichman, M., et al. (2013). *Uci machine learning repository.* Irvine, CA.

Long, J. S. (1990). The origins of sex differences in science. *Social forces*, *68*(4), 1297–1316.

Lyddon, S., Holmes, C., & Walker, S. (2018). General bayesian updating and the loss-likelihood bootstrap. *Biometrika*.

Miller, J. W., & Dunson, D. B. (2018). Robust bayesian inference via coarsening. *Journal of the American Statistical Association*, 1–13.

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

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.

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

Rossell, D., & Rubio, F. J. (2018). Tractable bayesian variable selection: beyond normality. *Journal of the American Statistical Association*, *113*(524), 1742–1758.

Ruanaidh, Ó., Joseph, J., & Fitzgerald, W. J. (1996). Numerical Bayesian methods applied to signal processing.

Smith, J. (2007). Local robustness of bayesian parametric inference and observed likelihoods.

Smith, J., & Rigat, F. (2012). Isoseparation and robustness in finite parameter bayesian inference. *Annals of the Institute of Statistical Mathematics*, *64*, 495–519.

Székely, G. J., & Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, *143*(8), 1249–1272.

Walker, S. G. (2013). Bayesian inference with misspecified models. *Journal of Statistical Planning and Inference*, *143*(10), 1621–1633.

Watson, J., Holmes, C., et al. (2016). Approximate models and robust decisions. *Statistical Science*, *31*(4), 465–489.

Whittle, P., & Whittle, P. R. (1990). *Risk-sensitive optimal control* (Vol. 20). Wiley New York.

Winkler, R. L., & Murphy, A. H. (1968). Evaluation of subjective precipitation probability forecasts. In *Proceedings of the first national conference on statistical meteorology* (pp. 148–157).

7 Appendix

7.1 Proof of theorems for constant parameter space

7.1.1 Proof of Theorem 1

Proof. Define $\Theta = \Theta_f = \Theta_h$. Using the triangle inequality and the definition of $\mathcal{N}_\epsilon^{D_M}$ gives us that for all $\theta \in \Theta$,

$$\begin{aligned} D_M(g, f(\cdot;\theta)) &\leq D_M(h(\cdot;\theta), f(\cdot;\theta)) + D_M(g, h(\cdot;\theta)) && (59)\\ &\leq \epsilon + D_M(g, h(\cdot;\theta)) && (60)\\ D_M(g, h(\cdot;\theta)) &\leq D_M(h(\cdot;\theta), f(\cdot;\theta)) + D_M(g, f(\cdot;\theta)) && (61)\\ &\leq \epsilon + D_M(g, f(\cdot;\theta)). && (62)\end{aligned}$$

Now the definition of the parameters $\hat{\theta}_h^{D_M}$ and $\hat{\theta}_f^{D_M}$ as the parameters of the likelihood models minimising divergence D_M combined with the inequalities above result in

$$\begin{aligned} D_M(g, f(\cdot;\hat{\theta}_f^{D_M})) \leq D_M(g, f(\cdot;\hat{\theta}_h^{D_M})) &\leq \epsilon + D_M(g, h(\cdot;\hat{\theta}_h^{D_M})) && (63)\\ D_M(g, h(\cdot;\hat{\theta}_h^{D_M})) \leq D_M(g, h(\cdot;\hat{\theta}_f^{D_M})) &\leq \epsilon + D_M(g, f(\cdot;\hat{\theta}_f^{D_M})) && (64)\\ \Rightarrow \left| D_M(g, h(\cdot;\hat{\theta}_h^{D_M})) - D_M(g, f(\cdot;\hat{\theta}_f^{D_M})) \right| &\leq \epsilon. && (65)\end{aligned}$$

□

7.1.2 Proof of Theorem 2

Proof. Jensen's inequality can be adapted to show that for convex function ψ, and any function ρ such that $\mathbb{E}_X[|\rho(X)|]$ and $\mathbb{E}_X[|\psi(\rho(X))|]$ are finite, then

$$\psi(\mathbb{E}_X[\rho(X)]) \leq \mathbb{E}_X[\psi(\rho(X))]. \tag{66}$$

Consider applying Jensen's inequality with θ as the random variable of interest with distribution $\pi(\theta|X_{1:n})$, $\rho(\theta) = f(y;\theta)$ for some fixed y and with $\psi(f) = D_M(g, f)$, where g is some fixed probability density, as a convex function. Both $\rho(\cdot)$ and $\psi(\cdot)$ are positive functions so Jensen's inequality is valid providing

$$\mathbb{E}_\theta[f(y;\theta)] = \int f(y;\theta)\pi(\theta|X_{1:n})d\theta < \infty. \tag{67}$$

This simply requires that the Bayesian predictive distribution is defined, and that

$$\mathbb{E}_\theta[D_M(h(\cdot), f(\cdot;\theta))] = \int D_M(h(\cdot), f(\cdot;\theta))\pi(\theta|X_{1:n})d\theta < \infty. \tag{68}$$

By the convexity of $D_M(\cdot,\cdot)$, Jensen's inequality can be applied as described above and therefore

$$D_M(m_f^{D_M}(\cdot), m_h^{D_M}(\cdot))) \leq \int D_M(f(y|X_{1:n}), h(y|\theta_h))\pi_h^{D_M}(\theta_h|X_{1:n})d\theta_h \tag{69}$$

$$\leq \int \left\{ \int D_M(f(y|\theta_f), h(y|\theta_h))\pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f \right\} \pi_h^{D_M}(\theta_h|X_{1:n})d\theta_h. \tag{70}$$

The triangle inequality associated with $D_M(\cdot,\cdot)$ gives that

$$D_M(f,h) \leq D_M(f,g) + D_M(g,h) = D_M(g,f) + D_M(g,h), \tag{71}$$

which is used to show that

$$D_M(m_f^{D_M}(\cdot), m_h^{D_M}(\cdot))) \leq \int \left\{ \int D_M(f(y|\theta_f), h(y|\theta_h))\pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f \right\} \pi_h^{D_M}(\theta_h|X_{1:n})d\theta_h$$

$$\leq \int \left\{ \int D_M(g, f(y|\theta_f)) + D_M(g, h(y|\theta_h))\pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f \right\} \pi_h^{D_M}(\theta_h|X_{1:n})d\theta_h \tag{72}$$

$$= \int D_M(g, f(y|\theta_f))\pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f + \int D_M(g, h(y|\theta_h))\pi_h^{D_M}(\theta_h|X_{1:n})d\theta_h. \tag{73}$$

Now given the first part of Condition 2, equation (25)

$$D_M(m_f^{D_M}(\cdot), m_h^{D_M}(\cdot)))$$

$$\leq \int D_M(g, f(y|\theta_f))\pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f + \int D_M(g, h(y|\theta_h))\pi_h^{D_M}(\theta_h|X_{1:n})d\theta_h$$

$$\leq \int D_M(g, f(y|\theta_f))\pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f + \int D_M(g, h(y|\theta_f))\pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f \tag{74}$$

$$= \int \left(D_M(g, f(y|\theta_f)) + D_M(g, h(y|\theta_f))\right) \pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f. \tag{75}$$

We can add and subtract $D_M(f(y|\theta_f), h(y|\theta_f))$ inside the integral to give

$$D_M(m_f^{D_M}(\cdot), m_h^{D_M}(\cdot))) \leq \int \left(D_M(g, f(y|\theta_f)) + D_M(g, h(y|\theta_f))\right) \pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f$$

$$= \int (D_M(g, f(y|\theta_f)) + D_M(g, h(y|\theta_f)) \tag{76}$$

$$- D_M(f(y|\theta_f), h(y|\theta_f)) + D_M(f(y|\theta_f), h(y|\theta_f)))\pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f.$$

Finally applying the triangle inequality once more gives us that

$$D_M(g,f) + D_M(f,h) \geq D_M(g,h) \Rightarrow D_M(g,f) \geq D_M(g,h) - D_M(f,h) \tag{77}$$

which can be used in combination with the definition of the neighbourhood $\mathcal{N}_\epsilon^{D_M}$ to show that

$$\begin{aligned} D_M(m_f^{D_M}(\cdot), m_h^{D_M}(\cdot))) &\leq \int (D_M(g, f(y|\theta_f)) + D_M(g, h(y|\theta_f)) \\ &\quad - D_M(f(y|\theta_f), h(y|\theta_f)) + D_M(f(y|\theta_f), h(y|\theta_f)))\pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f \\ &\leq \int (2D_M(g, f(y|\theta_f)) + \epsilon)\, \pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f \qquad (78) \\ &= 2\int (D_M(g, f(y|\theta_f)))\, \pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f + \epsilon. \qquad (79) \end{aligned}$$

We note that we could have applied the second part of Condition 2, equation (26), to exchange θ_f for θ_h in line (74) and the triangle inequality also gives us that

$$D_M(g,h) + D_M(f,h) \geq D_M(g,f) \Rightarrow D_M(g,h) \geq D_M(g,f) - D_M(f,h). \qquad (80)$$

Which, in turn can be used to show that

$$D_M(m_f^{D_M}(\cdot), m_h^{D_M}(\cdot))) \leq 2\int (D_M(g, h(y|\theta_h)))\, \pi_h^{D_M}(\theta_h|X_{1:n})d\theta_h + \epsilon \qquad (81)$$

and thus

$$D_M(m_f^{D_M}(\cdot), m_h^{D_M}(\cdot))) \leq R^{D_M}(g, f, h, \boldsymbol{x}_{1:n}) + \epsilon. \qquad (82)$$

where $R^{D_M}(g, f, h, \boldsymbol{x}_{1:n})$ is defined in Eq. 28. □

7.1.3 Proof of Corollary 1

Proof. From Theorem 2 we know that

$$\begin{aligned} &D^M(m_f^{D_M}(\cdot|\boldsymbol{x}_{1:n}), m_h^{D_M}(\cdot|\boldsymbol{x}_{1:n}))) \leq \\ &2\min\left\{\int \left(D^M(g, f(y|\theta_f))\right) \pi_f^{D_M}(\theta_f|\boldsymbol{x}_{1:n})d\theta_f, \int \left(D^M(g, h(y|\theta_h))\right) \pi_h^{D_M}(\theta_h|\boldsymbol{x}_{1:n})d\theta_h\right\} + \epsilon \end{aligned}$$

Additionally as we know that there exists θ_{f_0} such that $f(\cdot;\theta_{f_0}) = g(\cdot)$, which provided the weak conditions for asymptotic normality Chernozhukov and Hong (2003); Lyddon et al. (2018) (Eq. (**??**)) hold, and once again assuming that both g and $f(y;\theta)$ are absolutely continuous, implies there exists n such that

$$D^M(m_f^{D_M}(\cdot|\boldsymbol{x}_{1:n}), m_h^{D_M}(\cdot|\boldsymbol{x}_{1:n}))) \leq 2\int (D(g, f(y|\theta_f)))\, \pi_f^{D}(\theta_1|\boldsymbol{x}_{1:n})d\theta_f + \epsilon, \qquad (83)$$

almost surely. Following the asymptotics of Walker (2013) define

$$A_{\delta,f}^D = \{\theta : D(g(\cdot), f(\cdot;\theta)) \leq \delta\}, \qquad (84)$$

and $A_{\delta,f}^{D_M}{}^c$ its complement. Now for any $\delta > 0$ we have that

$$\begin{aligned} D^M(m_f^{D_M}(y|\boldsymbol{x}_{1:n}), m_h^{D_M}(y|\boldsymbol{x}_{1:n}))) &\leq 2\int \left(D^M(g, f(y|\theta_f))\right) \pi_f^{D_M}(\theta_1|\boldsymbol{x}_{1:n}) d\theta_f + \epsilon \\ &= 2\int_{A_{\delta,f}^{D_M}} \left(D^M(g, f(y|\theta_f))\right) \pi_f^{D_M}(\theta_1|\boldsymbol{x}_{1:n}) d\theta_f \\ &\quad + 2\int_{A_{\delta,f}^{D_M}{}^c} \left(D^M(g, f(y|\theta_f))\right) \pi_f^{D_M}(\theta_f|\boldsymbol{x}_{1:n}) d\theta_f + \epsilon \end{aligned} \tag{85}$$

Now by the definition of $A_{\delta,f}^{D_M}$, for all values of $\theta_f \in A_{\delta,f}^{D_M}$, $D^M(g, f(y|\theta_f)) < \delta$ and therefore the integral over the whole set must be less than δ. And provided the divergence $D^M(\cdot,\cdot)$ is bounded by some finite constant $b < \infty$ for any two distributions (this bound is 1 for the Total Variation and Hellinger divergences) then $\int_{A_{\delta,f}^{D_M}{}^c} \left(D^M(g, f(y|\theta_1))\right) \pi_f^{D_M}(\theta_1|X_{1:n}) d\theta_1 \leq b\Pi_f^{D_M}(A_{\delta,f}^{D_M}{}^c|X_{1:n})$, where $\Pi_f^{D_M}(A_{\delta,f}^{D_M}{}^c|X_{1:n})$ is the density in the set $A_{\delta,f}^{D_M}{}^c$ of the posterior $\pi_f^{D_M}(\theta_f|\boldsymbol{x}_{1:n})$. Therefore,

$$\begin{aligned} D^M(m_f^{D_M}(y), m_h^{D_M}(y))) &\leq 2\int_{A_{\delta,f}^{D_M}} \left(D^M(g, f(y|\theta_1))\right) \pi_f^{D_M}(\theta_1|X_{1:n}) d\theta_1 \\ &\quad + 2\int_{A_{\delta,f}^{D_M}{}^c} \left(D^M(g, f(y|\theta_1))\right) \pi_f^{D_M}(\theta_1|X_{1:n}) d\theta_1 + \epsilon \\ &\leq 2\delta + 2b\Pi_f^{D_M}(A_{\delta,f}^{D_M}{}^c|X_{1:n}) + \epsilon. \end{aligned} \tag{86}$$

Therefore provided that $\Pi_f^{D_M}(A_{\delta,f}^{D_M}{}^c|X_{1:n}) \to 0$ a.s. which is provided by asymptotic normality (Eq. (**??**)) of the posterior (Chernozhukov & Hong, 2003; Lyddon et al., 2018), and since this holds for all δ we have that

$$D^M(m_f^{D_M}(y), m_h^{D_M}(y))) \leq \epsilon. \tag{87}$$

□

7.1.4 Proof of Lemma 1

Proof. The definition of the KLD (Eq. (**??**)) provides that

$$\mathrm{KLD}(g, f(\cdot;\theta_f)) = \mathrm{KLD}(g, h(\cdot;\theta_h)) + \int g \log \frac{h(\cdot;\theta_h)}{f(\cdot;\theta_f)} dx. \tag{88}$$

Now by the definition of $\hat{\theta}_f^{\mathrm{KLD}}$ and $\hat{\theta}_h^{\mathrm{KLD}}$ we can show

$$\mathrm{KLD}(g, f(\cdot;\hat{\theta}_f^{\mathrm{KLD}})) \leq \mathrm{KLD}(g, f(\cdot;\hat{\theta}_h^{\mathrm{KLD}})) \tag{89}$$

$$= \mathrm{KLD}(g, h(\cdot;\hat{\theta}_h^{\mathrm{KLD}})) + \int g \log \frac{h(\cdot;\hat{\theta}_h^{\mathrm{KLD}})}{f(\cdot;\hat{\theta}_h^{\mathrm{KLD}})} dx \tag{90}$$

$$\mathrm{KLD}(g, h(\cdot;\hat{\theta}_h^{\mathrm{KLD}})) \leq \mathrm{KLD}(g, h(\cdot;\hat{\theta}_f^{\mathrm{KLD}})) \tag{91}$$

$$= \mathrm{KLD}(g, f(\cdot;\hat{\theta}_f^{\mathrm{KLD}})) + \int g \log \frac{f(\cdot;\hat{\theta}_f^{\mathrm{KLD}})}{h(\cdot;\hat{\theta}_f^{\mathrm{KLD}})} dx. \tag{92}$$

Combining these two inequalities results in Eq. (32). □

7.1.5 Proof of Lemma 2

In order to prove Lemma 2 we must first prove two smaller lemmas relating the βD triangle inequalities.

Although Bregman divergence, introduced in Eq. (**??**), are not generally metric, a well-known generalisation of the triangle inequality exists for these as follows

Lemma 3 (Bregman Divergence Triangle Inequality)**.** For Bregman divergence $D_\psi(g||f)$ defined in Eq. (**??**) the following generalisation of the triangle inequality holds

$$D_\psi(g||f) + D_\psi(f||h) = D_\psi(g||h) + (g-f)\left(\nabla\psi(h) - \nabla\psi(f)\right) \tag{93}$$

Proof. Following the definition of a Bregman divergence Eq. (**??**)

$$\begin{aligned}
&D_\psi(g||f) + D_\psi(f||h) \\
=&\psi(g) - \psi(f) - (g-f)\psi^{'}(f) + \psi(f) - \psi(h) - (f-h)\psi^{'}(h) && (94)\\
=&\psi(g) - \psi(h) - (-h)\psi^{'}(h) - (g-f)\psi^{'}(f) - (f)\psi^{'}(h) && (95)\\
=&\psi(g) - \psi(h) - (g-h)\psi^{'}(h) - (g-f)\psi^{'}(f) - (f-g)\psi^{'}(h) && (96)\\
=&D_\psi(g||h) + (g-f)\left(\psi^{'}(h) - \psi^{'}(f)\right) && (97)
\end{aligned}$$

□

Applying this specifically for the βD provides the following lemma.

Lemma 4 (Bregman Divergence Triangle Inequality for the βD)**.** The following relationship for the βD holds fo densities g, f and h

$$D_B^{(\beta)}(f||h) = D_B^{(\beta)}(g||h) - D_B^{(\beta)}(g||f) + R(g||f||h) \tag{98}$$
$$D_B^{(\beta)}(h||f) = D_B^{(\beta)}(g||f) - D_B^{(\beta)}(g||h) + R(g||h||f) \tag{99}$$

$$R(g||f||h) = \int (g-f)\left(\frac{1}{\beta-1}h^{\beta-1} - \frac{1}{\beta-1}f^{\beta-1}\right)d\mu \tag{100}$$
$$R(g||h||f) = \int (g-h)\left(\frac{1}{\beta-1}f^{\beta-1} - \frac{1}{\beta-1}h^{\beta-1}\right)d\mu \tag{101}$$

Proof. This follows directly from Lemma 3 and the definition of the βD. □

Now we prove Lemma 2

Proof. Define $A^+ := \{x : h(x) > f(x)\}$ and $A^- := \{x : f(x) > h(x)\}$. Firstly note that

$$\text{TVD}(f,h) = \int_{A^+} (h(x) - f(x))\,dx = \int_{A^-} (f(x) - h(x))\,dx \tag{102}$$

By the definition of the βD we can rearrange

$$\begin{aligned}
&D_B^{(\beta)}(g||f) \\
=&D_B^{(\beta)}(g||h) + \int \frac{1}{\beta}h(x)^\beta - \frac{1}{\beta}f(x)^\beta - \frac{1}{\beta-1}g(x)h(x)^{\beta-1} + \frac{1}{\beta-1}g(x)f(x)^{\beta-1}dx \qquad (103)\\
=&D_B^{(\beta)}(g||h) + \frac{1}{\beta}\int h(x)^\beta - f(x)^\beta dx + \frac{1}{\beta-1}\int g(x)\left(f(x)^{\beta-1} - h(x)^{\beta-1}\right)dx \qquad (104)
\end{aligned}$$

Now by the monotonicity of the function x^β when $1 \leq \beta \leq 2$ we have that

$$\begin{aligned}
\int_{A^-} h(x)^\beta - f(x)^\beta dx < 0 \\
\int_{A^+} g(x)\left(f(x)^{\beta-1} - h(x)^{\beta-1}\right)dx < 0
\end{aligned}$$

therefore removing these two terms provides an upper bound

$$\begin{aligned}
&D_B^{(\beta)}(g||f) \\
=&D_B^{(\beta)}(g||h) + \frac{1}{\beta}\int h(x)^\beta - f(x)^\beta dx + \frac{1}{\beta-1}\int g(x)\left(f(x)^{\beta-1} - h(x)^{\beta-1}\right)dx \\
\leq&D_B^{(\beta)}(g||h) + \frac{1}{\beta}\int_{A^+} h(x)^\beta - f(x)^\beta dx + \frac{1}{\beta-1}\int_{A^-} g(x)\left(f(x)^{\beta-1} - h(x)^{\beta-1}\right)dx. \qquad (105)
\end{aligned}$$

Next $x \in A^+$ ensures $h(x) > f(x)$ and this in turn implies that $h(x)f(x)^{\beta-1} > f(x)^\beta$. As a result we can bound

$$\begin{aligned}
&D_B^{(\beta)}(g||f) \\
\leq&D_B^{(\beta)}(g||h) + \frac{1}{\beta}\int_{A^+} h(x)^\beta - f(x)^\beta dx + \frac{1}{\beta-1}\int_{A^-} g(x)\left(f(x)^{\beta-1} - h(x)^{\beta-1}\right)dx. \\
\leq&D_B^{(\beta)}(g||h) + \frac{1}{\beta}\int_{A^+} h(x)\left(h(x)^{\beta-1} - f(x)^{\beta-1}\right)dx \\
&+ \frac{1}{\beta-1}\int_{A^-} g(x)\left(f(x)^{\beta-1} - h(x)^{\beta-1}\right)dx \qquad (106)\\
=&D_B^{(\beta)}(g||h) + \frac{1}{\beta}\int_{A^+} h(x)^\beta\left(1 - \frac{f(x)^{\beta-1}}{h(x)^{\beta-1}}\right)dx \\
&+ \frac{1}{\beta-1}\int_{A^-} g(x)f(x)^{\beta-1}\left(1 - \frac{h(x)^{\beta-1}}{f(x)^{\beta-1}}\right)dx. \qquad (107)
\end{aligned}$$

Now on A^+ $h(x) > f(x)$ and so $\frac{f(x)}{h(x)}^{\beta-1} > \frac{f(x)}{h(x)}$ for $1 \leq \beta \leq 2$ so

$$\left(1 - \frac{f(x)^{\beta-1}}{h(x)^{\beta-1}}\right) \leq \left(1 - \frac{f(x)}{h(x)}\right) \qquad (108)$$

with the exact same logic holding when $f(x) < h(x)$ for the second integral. We can use this to show

that

$$
\begin{aligned}
&D_B^{(\beta)}(g||f) \\
\leq &D_B^{(\beta)}(g||h) + \frac{1}{\beta}\int_{A^+} h(x)^\beta \left(1 - \frac{f(x)^{\beta-1}}{h(x)^{\beta-1}}\right) d\mu + \frac{1}{\beta-1}\int_{A^-} g(x)f(x)^{\beta-1}\left(1 - \frac{h(x)^{\beta-1}}{f(x)^{\beta-1}}\right) dx \\
\leq &D_B^{(\beta)}(g||h) + \frac{1}{\beta}\int_{A^+} h(x)^\beta \left(1 - \frac{f(x)}{h(x)}\right) dx + \frac{1}{\beta-1}\int_{A^-} g(x)f(x)^{\beta-1}\left(1 - \frac{h(x)}{f(x)}\right) dx \qquad (109) \\
= &D_B^{(\beta)}(g||h) + \frac{1}{\beta}\int_{A^+} h(x)^{\beta-1}\left(h(x) - f(x)\right) dx + \frac{1}{\beta-1}\int_{A^-} g(x)f(x)^{\beta-2}\left(f(x) - h(x)\right) dx \qquad (110)
\end{aligned}
$$

We now use the fact that we defined $\max\{\operatorname{ess\,sup} f, \operatorname{ess\,sup} h, \operatorname{ess\,sup} g\} \leq M < \infty$ to leave

$$
\begin{aligned}
&D_B^{(\beta)}(g||f) \\
= &D_B^{(\beta)}(g||h) + \frac{1}{\beta}\int_{A^+} h(x)^{\beta-1}\left(h(x) - f(x)\right) dx + \frac{1}{\beta-1}\int_{A^-} g(x)f(x)^{\beta-2}\left(f(x) - h(x)\right) dx \qquad (111) \\
\leq &D_B^{(\beta)}(g||h) + \frac{M^{\beta-1}}{\beta}\int_{A^+} \left(h(x) - f(x)\right) dx + \frac{M^{\beta-1}}{\beta-1}\int_{A^-} \left(f(x) - h(x)\right) dx \qquad (112) \\
= &D_B^{(\beta)}(g||h) + \frac{M^{\beta-1}}{\beta}\mathrm{TVD}(h,f) + \frac{M^{\beta-1}}{\beta-1}\mathrm{TVD}(h,f) \qquad (113) \\
= &D_B^{(\beta)}(g||h) + \frac{M^{\beta-1}}{\beta-1}\mathrm{TVD}(h,f). \qquad (114)
\end{aligned}
$$

□

7.1.6 Proof of Theorem 3

Proof. Firstly be the definition of $\hat{\theta}_f^{(\beta)}$ and $\hat{\theta}_h^{(\beta)}$ as the parameters of the likelihood models $f(\cdot;\theta_f)$ and $h(\cdot;\theta_h)$ minimising the βD we have that.

$$
\begin{aligned}
D_B^{(\beta)}(g, f(\cdot;\hat{\theta}_f^{(\beta)})) \leq D_B^{(\beta)}(g, f(\cdot;\hat{\theta}_h^{(\beta)})) \\
D_B^{(\beta)}(g, h(\cdot;\hat{\theta}_h^{(\beta)})) \leq D_B^{(\beta)}(g, h(\cdot;\hat{\theta}_f^{(\beta)})).
\end{aligned} \qquad (115)
$$

Now using the triangle type inequality proven in Lemma 2 and the definition of $\mathcal{N}_\epsilon^{D_M}$ we can show that

$$
\begin{aligned}
&D_B^{(\beta)}(g, f(\cdot;\hat{\theta}_h^{(\beta)})) \\
\leq &\frac{M^{\beta-1}}{\beta-1}\mathrm{TVD}(f(\cdot;\hat{\theta}_h^{(\beta)}), h(\cdot;\hat{\theta}_h^{(\beta)})) + D_B^{(\beta)}(g, h(\cdot;\hat{\theta}_h^{(\beta)})) \\
\leq &\frac{M^{\beta-1}}{\beta-1}\epsilon + D_B^{(\beta)}(g, h(\cdot;\hat{\theta}_h^{(\beta)})) \qquad (116) \\
&D_B^{(\beta)}(g, h(\cdot;\hat{\theta}_f^{(\beta)})) \\
\leq &\frac{M^{\beta-1}}{\beta-1}\mathrm{TVD}(f(\cdot;\hat{\theta}_f^{(\beta)}), h(\cdot;\hat{\theta}_f^{(\beta)})) + D_B^{(\beta)}(g, f(\cdot;\hat{\theta}_f^{(\beta)})) \\
\leq &\frac{M^{\beta-1}}{\beta-1}\epsilon + D_B^{(\beta)}(g, f(\cdot;\hat{\theta}_f^{(\beta)})) \qquad (117) \\
& \qquad (118)
\end{aligned}
$$

Combining these two, results in

$$\Rightarrow \left| D_B^{(\beta)}(g, h(\cdot; \hat{\theta}_h^{(\beta)})) - D_B^{(\beta)}(g, f(\cdot; \hat{\theta}_f^{(\beta)})) \right| \leq \frac{M^{\beta-1}}{\beta - 1}\epsilon \tag{119}$$

□

7.1.7 Proof of Theorem ...

In order to prove the stability of the posterior predictives in the same vein as Theorem 1 we require on last lemma.

Lemma 5 (The convexity of the βD)**.** The βD between two densities $g(x)$ and $f(x)$ is convex in both densities for $1 < \beta \leq 2$, when fixing the other. That is to say that for $\lambda \in [0, 1]$ and fixed f and g

$$D_B^{(\beta)}(\lambda g_1 + (1-\lambda)g_2, f) \leq \lambda D_B^{(\beta)}(g_1, f) + (1-\lambda) D_B^{(\beta)}(g_2, f) \tag{120}$$
$$D_B^{(\beta)}(g, \lambda f_1 + (1-\lambda)f_2) \leq \lambda D_B^{(\beta)}(g, f_1) + (1-\lambda) D_B^{(\beta)}(g, f_2) \tag{121}$$

for $1 < \beta \leq 2$

Proof. First we fix f and look at convexity in the function g. let $\lambda \in [0, 1]$. The function x^p for $x \geq 0$ and $p > 1$ is convex and thus satisfies

$$(\lambda x_1 + (1-\lambda)x_2)^p \leq \lambda x_1^p + (1-\lambda)x_2^p \tag{122}$$

therefore we have that provided $D_B^{(\beta)}(g_1||f) < \infty$ and $D_B^{(\beta)}(g_2||f) < \infty$

$$\begin{aligned}
&D_B^{(\beta)}(\lambda g_1 + (1-\lambda)g_2||f) \\
=&\int \frac{1}{\beta(\beta-1)} \left(\lambda g_1 + (1-\lambda)g_2\right)^\beta + \frac{1}{\beta} f^\beta - \frac{1}{\beta - 1} \left(\lambda g_1 + (1-\lambda)g_2\right) f^{\beta-1} d\mu && (123)\\
\leq&\int \frac{1}{\beta(\beta-1)} \left(\lambda g_1^\beta + (1-\lambda)g_2^\beta\right) + \frac{1}{\beta} f^\beta - \frac{1}{\beta - 1} \left(\lambda g_1 + (1-\lambda)g_2\right) f^{\beta-1} d\mu && (124)\\
=&\lambda D_B^{(\beta)}(g_1||f) + (1-\lambda) D_B^{(\beta)}(g_2||f).
\end{aligned}$$

Next we fix g and look at the convexity in f. Similarly to above we know that when $x \geq 0$ and $1 \leq p \leq 2$ that $\frac{1}{p}x^p$ and $-\frac{1}{p-1}x^{p-1}$ are both convex in x. We therefore have that provided $D_B^{(\beta)}(g||f_1) < \infty$ and $D_B^{(\beta)}(g||f_2) < \infty$

$$\begin{aligned}
&D_B^{(\beta)}(g||\lambda f_1 + (1-\lambda)f_2) \\
=&\int \frac{1}{\beta(\beta-1)} g^\beta + \frac{1}{\beta} \left(\lambda f_1 + (1-\lambda)f_2\right)^\beta - \frac{1}{\beta - 1} g \left(\lambda f_1 + (1-\lambda)f_2\right)^{\beta-1} d\mu && (125)\\
\leq&\int \frac{1}{\beta(\beta-1)} g^\beta + \frac{1}{\beta} \left(\lambda f_1^\beta + (1-\lambda)f_2^\beta\right) - \frac{1}{\beta - 1} g \left(\lambda f_1^{\beta-1} + (1-\lambda)f_2^{\beta-1}\right) d\mu && (126)\\
=&\lambda D_B^{(\beta)}(g||f_1) + (1-\lambda) D_B^{(\beta)}(g||f_2)
\end{aligned}$$

□

We are now able to use the convexity of the βD (Lemma 5), the triangular relationship between the

βD and the TVD (Lemma 2) and the Bregman divergence triangle inequality for the βD (Lemma 4) to extend posterior predictive stability provided by inference targeting metrics (Theorem 2) to inference using the βD in Theorem 4.

Proof. By the convexity of the βD for $1 < \beta \leq 2$ (Lemma 5) we can apply Jensen's inequality as we did in the proof of Theorem 2 to show that

$$D_B^{(\beta)}(m_f^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})||m_h^{(\beta)}(\cdot|\boldsymbol{x}_{1:n}))) \leq \int D_B^{(\beta)}(f(y|X_{1:n})||h(y|\theta_h))\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h \tag{127}$$

$$\leq \int \left\{ \int D_B^{(\beta)}(f(y|\theta_f)||h(y|\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f \right\} \pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h. \tag{128}$$

Now the generalised triangle inequality associated with the βD (Lemma 4) gives us that

$$D_B^{(\beta)}(f||h) = D_B^{(\beta)}(g||h) - D_B^{(\beta)}(g||f) + R(g||f||h) \tag{129}$$

where $R(g||f||h)$ is defined in Eq. (47). Using this here provides

$$D_B^{(\beta)}(m_f^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})||m_h^{(\beta)}(\cdot|\boldsymbol{x}_{1:n}))) \tag{130}$$

$$\leq \int \left\{ \int D_B^{(\beta)}(f(y|\theta_f)||h(y|\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f \right\} \pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h$$

$$= \int \left\{ \int \left[D_B^{(\beta)}(g||h(y|\theta_h)) - D_B^{(\beta)}(g||f(y|\theta_f)) \right.\right.$$

$$\left.\left. + R(g, f(\cdot;\theta_f), h(\cdot;\theta_h)) \right] \pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f \right\} \pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h \tag{131}$$

$$= \int D_B^{(\beta)}(g||h(y|\theta_h))\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h - \int D_B^{(\beta)}(g||f(y|\theta_f))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f$$

$$+ \int \int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f \pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h. \tag{132}$$

Now given the first part of Condition 2, Eq. 25, applied for the $D = D_B^{(\beta)}$ allows us to exchange $\pi_h^{(\beta)}(\theta_h|X_{1:n})$ for $\pi_f^{(\beta)}(\theta_f|X_{1:n})$ in the first integral

$$D_B^{(\beta)}(m_f^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})||m_h^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})))$$

$$\leq \int D_B^{(\beta)}(g||h(y|\theta_h))\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h - \int D_B^{(\beta)}(g||f(y|\theta_f))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f$$

$$+ \int \int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f \pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h$$

$$\leq \int D_B^{(\beta)}(g||h(y|\theta_f))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f - \int D_B^{(\beta)}(g||f(y|\theta_f))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f$$

$$+ \int \int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f \pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h \tag{133}$$

$$= \int \left(D_B^{(\beta)}(g||h(y|\theta_f)) - \int D_B^{(\beta)}(g||f(y|\theta_f)) \right) \pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f$$

$$+ \int \int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f \pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h. \tag{134}$$

where the last line has simply collected the two terms now involving θ_f into one integral. We can now

apply the triangle type inequality from Lemma 2, Eq. 40

$$
\begin{aligned}
&D_B^{(\beta)}(m_f^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})||m_h^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})))\\
\leq &\int \left(D_B^{(\beta)}(g||h(y|\theta_f)) - \int D_B^{(\beta)}(g||f(y|\theta_f)) \right) \pi_f^{(\beta)}(\theta_f|X_{1:n}) d\theta_f\\
&+ \int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h)) \pi_f^{(\beta)}(\theta_f|X_{1:n}) d\theta_f \pi_h^{(\beta)}(\theta_h|X_{1:n}) d\theta_h.\\
\leq &\int \frac{M^{\beta-1}}{\beta-1} \mathrm{TVD}(h(y|\theta_f), f(y|\theta_f)) \pi_f^{(\beta)}(\theta_f|X_{1:n}) d\theta_f\\
&+ \int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h)) \pi_f^{(\beta)}(\theta_f|X_{1:n}) d\theta_f \pi_h^{(\beta)}(\theta_h|X_{1:n}) d\theta_h. \qquad (135)
\end{aligned}
$$

Which given the neighbourhood of likelihood models defined by $\mathcal{N}_\epsilon^{\mathrm{TVD}}$ in Eq. (15)

$$
\begin{aligned}
&D_B^{(\beta)}(m_f^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})||m_h^{(\beta)}(\cdot|\boldsymbol{x}_{1:n}))) \leq \int \frac{M^{\beta-1}}{\beta-1} \mathrm{TVD}(h(y|\theta_f), f(y|\theta_f)) \pi_f^{(\beta)}(\theta_f|X_{1:n}) d\theta_f\\
&+ \int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h)) \pi_f^{(\beta)}(\theta_f|X_{1:n}) d\theta_f \pi_h^{(\beta)}(\theta_h|X_{1:n}) d\theta_h.\\
\leq &\frac{M^{\beta-1}}{\beta-1}\epsilon + \int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h)) \pi_f^{(\beta)}(\theta_f|X_{1:n}) d\theta_f \pi_h^{(\beta)}(\theta_h|X_{1:n}) d\theta_h. \qquad (136)
\end{aligned}
$$

We note that we could have instead considered $D_B^{(\beta)}(m_h^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})||m_f^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})))$. Applied the corresponding version of the Bregman divergence triangle inequality, with remainder $R(g||h||f) = \int (g-h)\left(\frac{1}{\beta-1}f^{\beta-1} - \frac{1}{\beta-1}h^{\beta-1}\right) d\mu$ and used the second part of Condition 2, therefore we also have that

$$
\begin{aligned}
&D_B^{(\beta)}(m_h^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})||m_f^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})))\\
\leq &\frac{M^{\beta-1}}{\beta-1}\epsilon + \int\int R(g||h(\cdot;\theta_h)||f(\cdot;\theta_f)) \pi_f^{(\beta)}(\theta_f|X_{1:n}) d\theta_f \pi_h^{(\beta)}(\theta_h|X_{1:n}) d\theta_h. \qquad (137)
\end{aligned}
$$

□

7.2 Proof of theorems for differing parameter space

Here we extend the results of Chapter **??** to the situation where the parameter space Θ_f and Θ_h of likelihood models $\{f(x;\theta_f) : \theta_f \in \Theta_f\}$ and $\{h(x;\theta_h) : \theta_h \in \Theta_h\}$ are such that $\Theta_f \neq \Theta_h$. Firstly we start with Condition 2, concerning the posterior concentration

Condition 3 (Concentration of the posterior)**.** The data set $\boldsymbol{x}_{1:n} \sim g(\cdot)$ is of sufficient size and regularity, and the priors $\pi_f^D(\theta)$ and $\pi_h^D(\theta)$ have sufficient prior mass at θ_f^D and θ_h^D and that there exists $\theta_{f\setminus h}^*$ and $\theta_{h\setminus f}^*$ such that the posteriors $\pi_{n,f}^D(\theta_f|X_{1:n})$ and $\pi_{n,h}^D(\theta_h|X_{1:n})$ have concentrated to ensure

$$\int_{\Theta_f} D(g, h(y|\left\{\theta_{U,f}, \theta^*_{h\setminus f}\right\}))\pi_f^D(\left\{\theta_{U,f}, \theta_{f\setminus h}\right\}|\boldsymbol{x}_{1:n})d\theta_{U,f}d\theta_{f\setminus h}$$

$$\geq \int_{\Theta_h} D(g, h(y|\left\{\theta_{U,h}, \theta_{h\setminus f}\right\}))\pi_h^D(\left\{\theta_{U,h}, \theta_{h\setminus f}\right\}|\boldsymbol{x}_{1:n})d\theta_{U,h}d\theta_{h\setminus f} \tag{138}$$

$$\int_{\Theta_h} D(g, f(y|\left\{\theta_{U,h}, \theta^*_{f\setminus h}\right\}))\pi_h^D(\left\{\theta_{U,h}, \theta_{h\setminus f}\right\}|\boldsymbol{x}_{1:n})d\theta_{U,h}d\theta_{h\setminus f}$$

$$\geq \int_{\Theta_f} D(g, f(y|\left\{\theta_{U,f}, \theta_{f\setminus h}\right\}))\pi_f^D(\left\{\theta_{U,f}, \theta_{f\setminus h}\right\}|\boldsymbol{x}_{1:n})d\theta_{U,f}d\theta_{f\setminus h}. \tag{139}$$

where $\theta_f = \left\{\theta_{U,f}, \theta_{f\setminus h}\right\}$ and $\theta_h = \left\{\theta_{U,h}, \theta_{h\setminus f}\right\}$

We require the introduction of $\theta^*_{f\setminus h}$ and $\theta^*_{h\setminus f}$ when the size of the parameter spaces for the two likelihood models are not equal and thus we cannot immediate use the posterior for one model in combination with the likelihood of the other. The way in which we define our prior neighbourhoods in these scenarios, makes defining these values straightforward.

Now we prove the extend version of Theorems 2, 1, 3 and 4.

7.2.1 Proof of Theorem 2

Proof. Jensen's inequality can be adapted to show that for convex function ψ, and any function ρ such that $\mathbb{E}_X[|\rho(X)|]$ and $\mathbb{E}_X[|\psi(\rho(X))|]$ are finite, then

$$\psi(\mathbb{E}_X[\rho(X)]) \leq \mathbb{E}_X[\psi(\rho(X))]. \tag{140}$$

Consider applying Jensen's inequality with θ as the random variable of interest with distribution $\pi(\theta|X_{1:n})$, $\rho(\theta) = f(y;\theta)$ for some fixed y and with $\psi(f) = D_M(g, f)$, where g is some fixed probability density, as a convex function. Both $\rho(\cdot)$ and $\psi(\cdot)$ are positive functions so Jensen's inequality is valid providing

$$\mathbb{E}_\theta[f(y;\theta)] = \int f(y;\theta)\pi(\theta|X_{1:n})d\theta < \infty. \tag{141}$$

This simply requires that the Bayesian predictive distribution is defined, and that

$$\mathbb{E}_\theta[D_M(h(\cdot), f(\cdot;\theta))] = \int D_M(h(\cdot), f(\cdot;\theta))\pi(\theta|X_{1:n})d\theta < \infty. \tag{142}$$

By the convexity of $D_M(\cdot,\cdot)$, Jensen's inequality can be applied as described above and therefore

$$D_M(m_f^{D_M}(\cdot), m_h^{D_M}(\cdot))) \leq \int D_M(f(y|X_{1:n}), h(y|\theta_h))\pi_h^{D_M}(\theta_h|X_{1:n})d\theta_h \tag{143}$$

$$\leq \int \left\{\int D_M(f(y|\theta_f), h(y|\theta_h))\pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f\right\}\pi_h^{D_M}(\theta_h|X_{1:n})d\theta_h. \tag{144}$$

The triangle inequality associated with $D_M(\cdot,\cdot)$ gives that

$$D_M(f, h) \leq D_M(f, g) + D_M(g, h) = D_M(g, f) + D_M(g, h), \tag{145}$$

which is used to show that

$$D_M(m_f^{D_M}(\cdot), m_h^{D_M}(\cdot))) \leq \int \left\{ \int D_M(f(y|\theta_f), h(y|\theta_h)) \pi_f^{D_M}(\theta_f|X_{1:n}) d\theta_f \right\} \pi_h^{D_M}(\theta_h|X_{1:n}) d\theta_h$$

$$\leq \int \left\{ \int D_M(g, f(y|\theta_f)) + D_M(g, h(y|\theta_h)) \pi_f^{D_M}(\theta_f|X_{1:n}) d\theta_f \right\} \pi_h^{D_M}(\theta_h|X_{1:n}) d\theta_h \tag{146}$$

$$= \int D_M(g, f(y|\theta_f)) \pi_f^{D_M}(\theta_f|X_{1:n}) d\theta_f + \int D_M(g, h(y|\theta_h)) \pi_h^{D_M}(\theta_h|X_{1:n}) d\theta_h. \tag{147}$$

Now we decompose the parameter for each model as into the part shared by the two models and what is left over we

$$\theta_f = \left\{ \theta_{U,f}, \theta_{f\setminus h} \right\} \tag{148}$$

$$\theta_h = \left\{ \theta_{U,h}, \theta_{h\setminus f} \right\} \tag{149}$$

we can then equivalently write

$$\begin{aligned} & D_M(m_f^{D_M}(\cdot), m_h^{D_M}(\cdot))) \\ \leq & \int D_M(g, f(y|\theta_f)) \pi_f^{D_M}(\theta_f|X_{1:n}) d\theta_f + \int D_M(g, h(y|\theta_h)) \pi_h^{D_M}(\theta_h|X_{1:n}) d\theta_h \\ = & \int D_M(g, f(y|\left\{ \theta_{U,f}, \theta_{f\setminus h} \right\})) \pi_f^{D_M}(\left\{ \theta_{U,f}, \theta_{f\setminus h} \right\}|X_{1:n}) d\theta_{U,f} d\theta_{f\setminus h} \\ & + \int D_M(g, h(y|\left\{ \theta_{U,h}, \theta_{h\setminus f} \right\})) \pi_h^{D_M}(\left\{ \theta_{U,h}, \theta_{h\setminus f} \right\}|X_{1:n}) d\theta_{U,h} d\theta_{h\setminus f} \end{aligned} \tag{150}$$

Now given the first part of Condition 3, equation (138)

$$\begin{aligned} & D_M(m_f^{D_M}(\cdot), m_h^{D_M}(\cdot))) \\ \leq & \int D_M(g, f(y|\left\{ \theta_{U,f}, \theta_{f\setminus h} \right\})) \pi_f^{D_M}(\left\{ \theta_{U,f}, \theta_{f\setminus h} \right\}|X_{1:n}) d\theta_{U,f} d\theta_{f\setminus h} \\ & + \int D_M(g, h(y|\left\{ \theta_{U,h}, \theta_{h\setminus f} \right\})) \pi_h^{D_M}(\left\{ \theta_{U,h}, \theta_{h\setminus f} \right\}|X_{1:n}) d\theta_{U,h} d\theta_{h\setminus f} \\ \leq & \int D_M(g, f(y|\left\{ \theta_{U,f}, \theta_{f\setminus h} \right\})) \pi_f^{D_M}(\left\{ \theta_{U,f}, \theta_{f\setminus h} \right\}|X_{1:n}) d\theta_{U,f} d\theta_{f\setminus h} \\ & + \int D_M(g, h(y|\left\{ \theta_{U,f}, \theta_{h\setminus f}^* \right\})) \pi_f^{D_M}(\left\{ \theta_{U,f}, \theta_{f\setminus h} \right\}|X_{1:n}) d\theta_{U,f} d\theta_{f\setminus h} \end{aligned} \tag{151}$$

$$\begin{aligned} = & \int \left(D_M(g, f(y|\left\{ \theta_{U,f}, \theta_{f\setminus h} \right\})) + D_M(g, h(y|\left\{ \theta_{U,f}, \theta_{h\setminus f}^* \right\})) \right) \\ & \pi_f^{D_M}(\left\{ \theta_{U,f}, \theta_{f\setminus h} \right\}|X_{1:n}) d\theta_{U,f} d\theta_{f\setminus h}. \end{aligned} \tag{152}$$

We can add and subtract $D_M(f(y|\left\{ \theta_{U,f}, \theta_{f\setminus h} \right\}), h(y|\left\{ \theta_{U,f}, \theta_{h\setminus f}^* \right\}))$ inside the integral to give

$$\begin{aligned} & D_M(m_f^{D_M}(\cdot), m_h^{D_M}(\cdot))) \\ \leq & \int \left(D_M(g, f(y|\left\{ \theta_{U,f}, \theta_{f\setminus h} \right\})) + D_M(g, h(y|\left\{ \theta_{U,f}, \theta_{h\setminus f}^* \right\})) \right. \\ & - D_M(f(y|\left\{ \theta_{U,f}, \theta_{f\setminus h} \right\}), h(y|\left\{ \theta_{U,f}, \theta_{h\setminus f}^* \right\})) \\ & \left. + D_M(f(y|\left\{ \theta_{U,f}, \theta_{f\setminus h} \right\}), h(y|\left\{ \theta_{U,f}, \theta_{h\setminus f}^* \right\})) \right) \pi_f^{D_M}(\left\{ \theta_{U,f}, \theta_{f\setminus h} \right\}|X_{1:n}) d\theta_{U,f} d\theta_{f\setminus h}. \end{aligned} \tag{153}$$

Finally applying the triangle inequality once more gives us that

$$D_M(g,f) + D_M(f,h) \geq D_M(g,h) \Rightarrow D_M(g,f) \geq D_M(g,h) - D_M(f,h) \tag{154}$$

which in combination with the definition of the neighbourhood $\mathcal{N}_\epsilon^{D_M}$ can be used to show that

$$\begin{aligned}
&D_M(m_f^{D_M}(\cdot), m_h^{D_M}(\cdot))) \\
\leq &\int \Big(D_M(g, f(y|\left\{\theta_{U,f}, \theta_{f\setminus h}\right\})) + D_M(g, h(y|\left\{\theta_{U,f}, \theta^*_{h\setminus f}\right\})) \\
&- D_M(f(y|\left\{\theta_{U,f}, \theta_{f\setminus h}\right\}), h(y|\left\{\theta_{U,f}, \theta^*_{h\setminus f}\right\})) \\
&+ D_M(f(y|\left\{\theta_{U,f}, \theta_{f\setminus h}\right\}), h(y|\left\{\theta_{U,f}, \theta^*_{h\setminus f}\right\}))\Big) \pi_f^{D_M}(\left\{\theta_{U,f}, \theta_{f\setminus h}\right\}|X_{1:n}) d\theta_{U,f} d\theta_{f\setminus h}. \\
&\leq \int \left(2D_M(g, f(y|\left\{\theta_{U,f}, \theta_{f\setminus h}\right\})) + \epsilon\right) \pi_f^{D_M}(\left\{\theta_{U,f}, \theta_{f\setminus h}\right\}|X_{1:n}) d\theta_f \qquad (155) \\
&= 2\int \left(D_M(g, f(y|\theta_f))\right) \pi_f^{D_M}(\theta_f|X_{1:n}) d\theta_f + \epsilon. \qquad (156)
\end{aligned}$$

We note that we could have applied the second part of Condition 3, equation (139), to exchange $\theta_f = \left\{\theta_{U,f}, \theta_{f\setminus h}\right\}$ for $\left\{\theta_{U,h}, \theta^*_{f\setminus h}\right\}$ in line (151) and the triangle inequality also gives us that

$$D_M(g,h) + D_M(f,h) \geq D_M(g,f) \Rightarrow D_M(g,h) \geq D_M(g,f) - D_M(f,h). \tag{157}$$

Which, in turn can be used to show that

$$D_M(m_f^{D_M}(\cdot), m_h^{D_M}(\cdot))) \leq 2\int \left(D_M(g, h(y|\theta_h))\right) \pi_h^{D_M}(\theta_h|X_{1:n}) d\theta_h + \epsilon \tag{158}$$

and thus

$$D_M(m_f^{D_M}(\cdot), m_h^{D_M}(\cdot))) \leq R^{D_M}(g, f, h, \boldsymbol{x}_{1:n}) + \epsilon. \tag{159}$$

where $R^{D_M}(g, f, h, \boldsymbol{x}_{1:n})$ is defined in Eq. **??**. □

7.2.2 Proof of Theorem 1

Proof. First we decompose

$$\hat{\theta}_f^{D_M} = \left\{\hat{\theta}_{U,f}^{D_M}, \hat{\theta}_{f\setminus h}^{D_M}\right\} \tag{160}$$

$$\hat{\theta}_h^{D_M} = \left\{\hat{\theta}_{U,h}^{D_M}, \hat{\theta}_{h\setminus f}^{D_M}\right\} \tag{161}$$

where in principle it need not be the case that $\hat{\theta}_{U,f}^{D_M} = \hat{\theta}_{U,h}^{D_M}$.

Using the triangle inequality and the definition of $\mathcal{N}_\epsilon^{D_M}$ gives us that $\forall\ \theta_U \in \Theta_U$, $\theta_{f\setminus h} \in \Theta_{f\setminus h}$ and $\theta_{h\setminus f} \in \Theta_{h\setminus f}$

$$\begin{aligned}
D_M(g, f(\cdot; \left\{\theta_U, \theta_{f\setminus h}\right\})) &\leq D_M(h(\cdot; \left\{\theta_U, \theta_{h\setminus f}\right\}), f(\cdot; \left\{\theta_U, \theta_{f\setminus h}\right\})) + D_M(g, h(\cdot; \left\{\theta_U, \theta_{h\setminus f}\right\})) \qquad (162) \\
&\leq \epsilon + D_M(g, h(\cdot; \left\{\theta_U, \theta_{h\setminus f}\right\})) \qquad (163) \\
D_M(g, h(\cdot; \left\{\theta_U, \theta_{h\setminus f}\right\})) &\leq D_M(h(\cdot; \left\{\theta_U, \theta_{h\setminus f}\right\}), f(\cdot; \left\{\theta_U, \theta_{f\setminus h}\right\})) + D_M(g, f(\cdot; \left\{\theta_U, \theta_{f\setminus h}\right\})) \qquad (164) \\
&\leq \epsilon + D_M(g, f(\cdot; \left\{\theta_U, \theta_{f\setminus h}\right\})) \qquad (165)
\end{aligned}$$

which by the definition of the parameter $\hat{\theta}_h^{D_M}$ and $\hat{\theta}_f^{D_M}$ as the parameters of the likelihood models minimising divergence D_M.

$$D_M(g, f(\cdot; \left\{\hat{\theta}_{U,f}^{D_M}, \hat{\theta}_{f\setminus h}^{D_M}\right\})) \leq D_M(g, f(\cdot; \left\{\hat{\theta}_{U,h}^{D_M}, \hat{\theta}_{f\setminus h}^{D_M}\right\})) \tag{166}$$

$$\leq \epsilon + D_M(g, h(\cdot; \left\{\hat{\theta}_{U,h}^{D_M}, \hat{\theta}_{h\setminus f}^{D_M}\right\})) \tag{167}$$

$$D_M(g, h(\cdot; \left\{\hat{\theta}_{U,h}^{D_M}, \hat{\theta}_{h\setminus f}^{D_M}\right\})) \leq D_M(g, h(\cdot; \left\{\hat{\theta}_{U,f}^{D_M}, \hat{\theta}_{h\setminus f}^{D_M}\right\})) \tag{168}$$

$$\leq \epsilon + D_M(g, f(\cdot; \left\{\hat{\theta}_{U,f}^{D_M}, \hat{\theta}_{f\setminus h}^{D_M}\right\})) \tag{169}$$

$$\Rightarrow \left|D_M(g, h(\cdot; \hat{\theta}_h^{D_M})) - D_M(g, f(\cdot; \hat{\theta}_f^{D_M}))\right| \leq \epsilon \tag{170}$$

□

7.2.3 Proof of Theorem 3

Proof. First we decompose

$$\hat{\theta}_f^{(\beta)} = \left\{\hat{\theta}_{U,f}^{(\beta)}, \hat{\theta}_{f\setminus h}^{(\beta)}\right\} \tag{171}$$

$$\hat{\theta}_h^{(\beta)} = \left\{\hat{\theta}_{U,h}^{(\beta)}, \hat{\theta}_{h\setminus f}^{(\beta)}\right\} \tag{172}$$

where in principle it need not be the case that $\hat{\theta}_{U,f}^{(\beta)} = \hat{\theta}_{U,h}^{(\beta)}$.

Firstly be the definition of $\hat{\theta}_f^{(\beta)}$ and $\hat{\theta}_h^{(\beta)}$ as the parameters of the likelihood models $f(\cdot;\theta_f)$ and $h(\cdot;\theta_h)$ minimising the βD we have that for all $\theta_{f\setminus h} \in \Theta_{f\setminus h}$ and $\theta_{h\setminus f} \in \Theta_{h\setminus f}$

$$\begin{aligned} D_B^{(\beta)}(g, f(\cdot; \left\{\hat{\theta}_{U,f}^{(\beta)}, \hat{\theta}_{f\setminus h}^{(\beta)}\right\})) &\leq D_B^{(\beta)}(g, f(\cdot; \left\{\hat{\theta}_{U,h}^{(\beta)}, \theta_{f\setminus h}\right\})) \\ D_B^{(\beta)}(g, h(\cdot; \left\{\hat{\theta}_{U,h}^{(\beta)}, \hat{\theta}_{h\setminus f}^{(\beta)}\right\})) &\leq D_B^{(\beta)}(g, h(\cdot; \left\{\hat{\theta}_{U,f}^{(\beta)}, \theta_{h\setminus f}\right\})). \end{aligned} \tag{173}$$

Using the triangle type inequality proven in Lemma 2 and the definition of $\mathcal{N}_\epsilon^{\text{TVD}}$ gives us

$$\begin{aligned} & D_B^{(\beta)}(g, f(\cdot; \left\{\hat{\theta}_{U,h}^{(\beta)}, \theta_{f\setminus h}\right\})) \\ \leq & \frac{M^{\beta-1}}{\beta-1}\text{TVD}(f(\cdot; \left\{\hat{\theta}_{U,h}^{(\beta)}, \theta_{f\setminus h}\right\}), h(\cdot; \left\{\hat{\theta}_{U,h}^{(\beta)}, \hat{\theta}_{h\setminus f}^{(\beta)}\right\})) + D_B^{(\beta)}(g, h(\cdot; \left\{\hat{\theta}_{U,h}^{(\beta)}, \hat{\theta}_{h\setminus f}^{(\beta)}\right\})) \\ \leq & \frac{M^{\beta-1}}{\beta-1}\epsilon + D_B^{(\beta)}(g, h(\cdot; \left\{\hat{\theta}_{U,h}^{(\beta)}, \hat{\theta}_{h\setminus f}^{(\beta)}\right\})) \end{aligned} \tag{174}$$

$$\begin{aligned} & D_B^{(\beta)}(g, h(\cdot; \left\{\hat{\theta}_{U,f}^{(\beta)}, \theta_{h\setminus f}\right\})) \\ \leq & \frac{M^{\beta-1}}{\beta-1}\text{TVD}(f(\cdot; \left\{\hat{\theta}_{U,f}^{(\beta)}, \hat{\theta}_{f\setminus h}^{(\beta)}\right\}), h(\cdot; \left\{\hat{\theta}_{U,f}^{(\beta)}, \theta_{h\setminus f}\right\})) + D_B^{(\beta)}(g, f(\cdot; \left\{\hat{\theta}_{U,f}^{(\beta)}, \hat{\theta}_{f\setminus h}^{(\beta)}\right\})) \\ \leq & \frac{M^{\beta-1}}{\beta-1}\epsilon + D_B^{(\beta)}(g, f(\cdot; \left\{\hat{\theta}_{U,f}^{(\beta)}, \hat{\theta}_{f\setminus h}^{(\beta)}\right\})) \end{aligned} \tag{175}$$

Combining these two results in

$$\Rightarrow \left|D_B^{(\beta)}(g, h(\cdot; \hat{\theta}_h^{(\beta)})) - D_B^{(\beta)}(g, f(\cdot; \hat{\theta}_f^{(\beta)}))\right| \leq \frac{M^{\beta-1}}{\beta-1}\epsilon \tag{176}$$

□

7.2.4 Proof of Theorem 4

Proof. By the convexity of the βD for $1 < \beta \leq 2$ (Lemma 5) we can apply Jensen's inequality as we did in the proof of Theorem **??** to show that

$$D_B^{(\beta)}(m_f^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})||m_h^{(\beta)}(\cdot|\boldsymbol{x}_{1:n}))) \leq \int D_B^{(\beta)}(f(y|X_{1:n})||h(y|\theta_h))\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h \tag{177}$$

$$\leq \int \left\{ \int D_B^{(\beta)}(f(y|\theta_f)||h(y|\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f \right\} \pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h. \tag{178}$$

Now the generalised triangle inequality associated with the βD (Lemma 4) gives us that

$$D_B^{(\beta)}(f||h) = D_B^{(\beta)}(g||h) - D_B^{(\beta)}(g||f) + R(g||f||h) \tag{179}$$

where $R(g||f||h)$ is defined in Eq. (**??**)and using this here provides

$$D_B^{(\beta)}(m_f^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})||m_h^{(\beta)}(\cdot|\boldsymbol{x}_{1:n}))) \tag{180}$$

$$\leq \int \left\{ \int D_B^{(\beta)}(f(y|\theta_f)||h(y|\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f \right\} \pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h$$

$$= \int \left\{ \int \left[D_B^{(\beta)}(g||h(y|\theta_h)) - D_B^{(\beta)}(g||f(y|\theta_f)) \right.\right.$$

$$\left.\left. + R(g, f(\cdot;\theta_f), h(\cdot;\theta_h)\right] \pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f \right\} \pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h \tag{181}$$

$$= \int D_B^{(\beta)}(g||h(y|\theta_h))\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h - \int D_B^{(\beta)}(g||f(y|\theta_f))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f$$

$$+ \int \int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f \pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h. \tag{182}$$

Now we decompose the parameter for each model as into the part shared by the two models and what is left over we

$$\theta_f = \left\{\theta_{U,f}, \theta_{f\setminus h}\right\} \tag{183}$$

$$\theta_h = \left\{\theta_{U,h}, \theta_{h\setminus f}\right\} \tag{184}$$

we can then equivalently write

$$D_B^{(\beta)}(m_f^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})||m_h^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})))$$

$$\leq \int D_B^{(\beta)}(g||h(y|\theta_h))\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h - \int D_B^{(\beta)}(g||f(y|\theta_f))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f$$

$$+ \int \int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f \pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h \tag{185}$$

$$= \int D_B^{(\beta)}(g||h(y|\left\{\theta_{U,h}, \theta_{h\setminus f}\right\}))\pi_h^{(\beta)}(\left\{\theta_{U,h}, \theta_{h\setminus f}\right\}|X_{1:n})d\theta_{U,h}d\theta_{h\setminus f}$$

$$- \int D_B^{(\beta)}(g||f(y|\left\{\theta_{U,f}, \theta_{f\setminus h}\right\}))\pi_f^{(\beta)}(\left\{\theta_{U,f}, \theta_{f\setminus h}\right\}|X_{1:n})d\theta_{U,f}d\theta_{f\setminus h}$$

$$+ \int \int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f \pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h. \tag{186}$$

Now given the first part of Condition 3, Eq. 138, applied for the $D = D_B^{(\beta)}$ allows us to exchange $\pi_h^{(\beta)}(\{\theta_{U,h}, \theta_{h\setminus f}\}|X_{1:n})$ for $\pi_f^{(\beta)}(\{\theta_{U,f}, \theta_{f\setminus h}\}|X_{1:n})$ in the first integral

$$
\begin{aligned}
& D_B^{(\beta)}(m_f^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})||m_h^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})))\\
\leq & \int D_B^{(\beta)}(g||h(y|\{\theta_{U,h}, \theta_{h\setminus f}\}))\pi_h^{(\beta)}(\{\theta_{U,h}, \theta_{h\setminus f}\}|X_{1:n})d\theta_{U,h}d\theta_{h\setminus f}\\
& - \int D_B^{(\beta)}(g||f(y|\{\theta_{U,f}, \theta_{f\setminus h}\}))\pi_f^{(\beta)}(\{\theta_{U,f}, \theta_{f\setminus h}\}|X_{1:n})d\theta_{U,f}d\theta_{f\setminus h}\\
& + \int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h\\
\leq & \int D_B^{(\beta)}(g||h(y|\left\{\theta_{U,f}, \theta^*_{h\setminus f}\right\}))\pi_f^{(\beta)}(\{\theta_{U,f}, \theta_{f\setminus h}\}|X_{1:n})d\theta_{U,f}d\theta_{f\setminus h}\\
& - \int D_B^{(\beta)}(g||f(y|\{\theta_{U,f}, \theta_{f\setminus h}\}))\pi_f^{(\beta)}(\{\theta_{U,f}, \theta_{f\setminus h}\}|X_{1:n})d\theta_{U,f}d\theta_{f\setminus h}\\
& + \int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h \qquad (187)\\
= & \int \left(D_B^{(\beta)}(g||h(y|\left\{\theta_{U,f}, \theta^*_{h\setminus f}\right\})) - \int D_B^{(\beta)}(g||f(y|\{\theta_{U,f}, \theta_{f\setminus h}\}))\right)\\
& \pi_f^{(\beta)}(\{\theta_{U,f}, \theta_{f\setminus h}\}|X_{1:n})d\theta_{U,f}d\theta_{f\setminus h}\\
& + \int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h. \qquad (188)
\end{aligned}
$$

where the last line has simply collected the two terms now involving $\theta_f = \{\theta_{U,f}, \theta_{f\setminus h}\}$ into one integral. We can now apply the triangle type inequality from Lemma 2, Eq. 40

$$
\begin{aligned}
& D_B^{(\beta)}(m_f^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})||m_h^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})))\\
\leq & \int \left(D_B^{(\beta)}(g||h(y|\left\{\theta_{U,f}, \theta^*_{h\setminus f}\right\})) - \int D_B^{(\beta)}(g||f(y|\{\theta_{U,f}, \theta_{f\setminus h}\}))\right)\\
& \pi_f^{(\beta)}(\{\theta_{U,f}, \theta_{f\setminus h}\}|X_{1:n})d\theta_{U,f}d\theta_{f\setminus h}\\
& + \int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h.\\
\leq & \int \frac{M^{\beta-1}}{\beta-1}\mathrm{TVD}(h(y|\left\{\theta_{U,f}, \theta^*_{h\setminus f}\right\}), f(y|\{\theta_{U,f}, \theta_{f\setminus h}\}))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_{U,f}d\theta_{f\setminus h}\\
& + \int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h. \qquad (189)
\end{aligned}
$$

Which given the neighbourhood of likelihood models defined by $\mathcal{N}_\epsilon^{\mathrm{TVD}}$ in Eq. (15)

$$
\begin{aligned}
& D_B^{(\beta)}(m_f^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})||m_h^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})))\\
\leq & \int \frac{M^{\beta-1}}{\beta-1}\mathrm{TVD}(h(y|\left\{\theta_{U,f}, \theta^*_{h\setminus f}\right\}), f(y|\{\theta_{U,f}, \theta_{f\setminus h}\}))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_{U,f}d\theta_{f\setminus h}\\
& + \int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h.\\
\leq & \frac{M^{\beta-1}}{\beta-1}\epsilon + \int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h. \qquad (190)
\end{aligned}
$$

We note that we could have instead considered $D_B^{(\beta)}(m_h^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})||m_f^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})))$. Applied the corresponding version for the Bregman divergence triangle inequality, with remainder $R(g||h||f) = \int(g -$

$h)\left(\frac{1}{\beta-1}f^{\beta-1}-\frac{1}{\beta-1}h^{\beta-1}\right)d\mu$ and used the second part of Condition 3, therefore we also have that

$$\begin{aligned}&D_B^{(\beta)}(m_h^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})||m_f^{(\beta)}(\cdot|\boldsymbol{x}_{1:n})))\\ \leq&\frac{M^{\beta-1}}{\beta-1}\epsilon+\int\int R(g||h(\cdot;\theta_h)||f(\cdot;\theta_f))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h.\end{aligned}\tag{191}$$

□