Bayesian Inference in the M-open world

by

Jack Edward Jewson

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

September 2019

THE UNIVERSITY OF WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

Firstly and foremostly I must thank my supervisor Prof. Jim Q. Smith for mentoring me throughout my PhD. I doubt whether many PhD. students are as lucky as I have been to have a supervisor so caring and generous with their time. Our conversations have always been entertaining and have inspired me to think deeply about the philosophy of Bayesian statistics. I am also indebted to Prof. Chris Holmes at the University of Oxford for his co-supervision. His paper on General Bayesian Updating provided the spark that started this PhD thesis and I have, and will, always look forward to our brainstorming session.

This PhD thesis would certainly not have progressed as far as it has done without my collaborations with my brilliant colleague and friend Jeremias Knoblauch and his supervisor Dr. Theo Damoulas. Their knowledge of applied problems and desire for real world impact provided me with a platform to implement my ideas and helped to convince me they might actually work.

Additionally I would like to thank Prof. Stephen Walker for hosting me for an enjoyable and educational 3 months at the University of Texas at Austin, and for his continued mentoring. Other thanks must be given to Prof. Simon French, who supervised my masters thesis, encouraged me to pursue a PhD, and co-authored my first publication, and to Prof. Christian Robert for providing interesting feedback and discussion on early parts of this work, for encouraging another of my early submissions and for acting as my internal examiner. I am also extremely grateful to my external examiner, Dr. Danny Williamson, whose interest and thorough reading of this thesis not only helped me to fix many typos and improve the presentation of some key arguments, but offered interesting discussion in the viva and a different

perspective on some of the philosophical arguments that have ultimately enriched the thesis greatly.

I must also thank my friends and peers Giuseppe, Nathan, Joe, Paul and last but not least Beniamino, for providing the company that is required to keep one sane when undertaking a PhD. Lastly I must thank my parents, Tony and Sheila, without whose constant prompting, I may never have even started writing this thesis.

Declarations

I declare that I have written and developed this PhD thesis entitled **"Bayesian Inference in the M-open world"** completely by myself, under the supervision of Prof. Jim Q. Smith and Prof. Chris Holmes, for the degree of Doctor of Philosophy in Statistics. I have not used sources or means without declaration in the text. I also confirm that this thesis has not been submitted for a degree at any other university.

During my PhD I have written the following articles, listed in order of writing:

1. **"Subjective Bayesian updating"** a self authored discussion piece on the Read Journal of the Royal Statistical Society Series A (JRSSA) article "Beyond subjective and objective in statistics" [Gelman and Hennig, 2015]. Additionally this appeared as part of some complied discussion pieces in "Some discussions on the Read Paper *Beyond subjective and objective in statistics by A. Gelman and C. Hennig*" [Celeux, Jewson, Josse, Marin, and Robert, 2017]. This article can be found at `https://arxiv.org/abs/1705.03727`.

2. **"Principles of Bayesian Inference Using General Divergence Criteria"** [Jewson, Smith, and Holmes, 2018], in collaboration with my supervisors Prof. Jim Q. Smith and Prof. Chris Holmes. This article has been peer reviewed and published in a special edition of the journal 'Entropy' titled 'Foundations of Statistics'. This article can be found at `https://www.mdpi.com/1099-4300/20/6/442`.

3. **"Doubly Robust Bayesian Inference for Non-Stationary Streaming Data with β-Divergences"** [Knoblauch, Jewson, and Damoulas, 2018], in collaboration with fellow PhD student Jeremias Knoblauch and his supervisor

Dr. Theo Damoulas. This article has been peer reviewed and published in the 32nd proceedings of the Advances in Neural Information Processing Systems (NeurIPS). This article can be found at `https://papers.nips.cc/paper/7292-doubly-robust-bayesian-inference-for-non-stationary-streaming-data-with-beta-divergences.pdf`. The division of labour in this project between myself and Jeremias was fairly clear. Together we decided to investigate whether the methods discussed in Jewson et al. [2018] could help rectify some of the problems with Knoblauch and Damoulas [2018]. I conceptualised the algorithm (Sections 5.3 and 5.4), derived the variational inference mechanism (Section 5.4.4) and proved the theorems in the paper (Theorems 11 and 12). Jeremias implemented the algorithm and adapted it to run in real time with real data (Section 5.4.6). He also produced all of the results beyond toy examples (Section 5.5). I proposed the initialisations for β and we worked together to derive the on-line optimisation for β (Section 5.4.6).

4. **"Generalised Variational Inference"** [Knoblauch, Jewson, and Damoulas, 2019], in collaboration with fellow PhD student Jeremias Knoblauch and his supervisor Dr. Theo Damoulas. This article has been submitted to a peer reviewed journal and at the time of writing is still under review. A preprint of this article can be found at `https://arxiv.org/pdf/1904.02063.pdf`. The division of labour between Jeremias and I was slightly less clear for this piece of work. Our work with variational inference in Knoblauch et al. [2018] lead us both independently to notice that the ELBO for variational inference was a constrained optimisation of the general Bayesian objective function for Bayes' rule (Section 4.2.2). Jeremias then proposed we extend this to attempt to fix the issues associated with VI. We conceptualised the axioms together, but Jeremias proved the theorems associated with these axioms (Section 4.3.2). I conducted and analysed the toy experiments looking at the impact of changing the prior regularising divergence (Section 4.4.2) and the multi-modal label switching example demonstrating the importance of our axioms (Section

4.5.1). However Jeremias identified and analysed the problems F-VI has with hyperparameetr optimisation (Section 4.8.1). I proved and analysed the theorems providing an interpretation for GVI as a posterior approximation (Section 4.6). However, Jeremias derived the black box implementation of GVI (Section 4.7) and applied this to the Bayesian neural networks and the deep Gaussian processes (Section 4.8).

Chapter 1 provides the necessary background for this thesis, often presented with different notation form the literature (references are provided). Chapter 2 is an extended version of the content of Jewson et al. [2018]. Chapters 4 and 5 are extended versions of the content of Knoblauch et al. [2019] and Knoblauch et al. [2018] respectively, where I particularly aim to focus on my contributions and how I interpret my joint work with Jeremias. Chapter 3 has not been published at the time of writing but constitutes work I have done on my own with my supervisors. We intend to submit this for publication in the near future. Many of the figures and text in this thesis have been taken from the aforementioned articles.

Celeux, Gilles and Jewson, Jack and Josse, Julie and Marin, Jean-Michel and Robert, Christian P (2017). Some discussions on the Read Paper" Beyond subjective and objective in statistics" by A. Gelman and C. Hennig. *arXiv preprint arXiv:1705.03727*, 1–5.

Jewson, Jack and Smith, Jim and Holmes, Chris (2018). Principles of Bayesian inference using general divergence criteria. Entropy, 20, 6, 442–466.

Knoblauch, Jeremias and Jewson, Jack and Damoulas, Theodoros (2018). Doubly Robust Bayesian Inference for Non-Stationary Streaming Data using β-Divergences. Advances in Neural Information Processing Systems (NeurIPS), 64–75

Knoblauch, Jeremias and Jewson, Jack and Damoulas, Theodoros. (2019). Generalized Variational Inference. *arXiv preprint arXiv:1904.02063*, 1–61.

Abstract

This thesis examines Bayesian inference and its suitability for modern statistical applications. Motivated by the vast quantities of data currently available for analysis, we forgo the $\mathcal{M}$-CLOSED assumption that the model used for inference is correctly specified and place ourselves in the more realistic $\mathcal{M}$-OPEN world. Here, we assume that the model used for statistical inference is at best an approximation.

In the $\mathcal{M}$-OPEN world Bayes' rule updating has been shown [Berk et al., 1966; Bissiri et al., 2016] to learn about the model parameters minimising the log-score, or equivalently the Kullback-Leibler divergence (KLD) to the data generating process (DGP). It is also known that minimising the log-score puts great emphasis on correctly capturing the tails of the sample distribution of the data. We observe, that this emphasis is so great, that the majority of the data can be ignored to sufficiently account an outlier. This is purportedly desirable when inference is the goal of the analysis. However, in Chapter 2 we show that when informed decision making via the minimisation of expected losses is the goal of the statistical analysis, as it so often is, Bayes' rule inferences are less desirable. This motivates us to consider minimising alternative divergences to the KLD.

Bayesian updating minimising alternative divergences to the KLD has briefly been considered in the literature. However, those methods are neither sufficiently well motivated or properly justified as a principled updating of beliefs. We are able to use the foundations of general Bayesian inference (GBI) to produce belief updates minimising any statistical divergence. This allows us to consider the divergence as a subjective judgement and motivate several divergences from a decision making

perspective.

Chapter 3 extends the motivation for minimising divergences alternative to the KLD. Here, we consider the model to be one among a equivalence class of belief models all respecting the belief judgements the decision maker (DM) has been able to make. It is therefore desirable for inference to be stable across this equivalence class. This is a well studied problem with respect to the prior component of the Bayesian analysis, but we believe we are one of the first to consider extending these result to the likelihood model. We prove that, unlike Bayes' rule updating, inference designed at minimising the total-variation divergence (TVD), the Hellinger divergence (HD), and the β-divergence (βD), are able to provide provably stable inferences.

Chapter 4 is inspired by the computation required to infer posteriors in modern Bayesian inference. We derive a generalised optimisation problem defining Bayesian inference. This is axiomatically motivated and contains Bayes' rule inference, GBI and variational inference (VI) as special cases. This generalised Bayesian inference problem is composed of three interpretable components: a loss function defining the limiting parameter of interest for the analysis; a prior regularising divergence describing how the posterior should quantify uncertainty; and a set of admissible posterior densities to optimise over. Chapters 2 and 3 examined changing the target parameter of inference to deal with model misspecification. Chapter 4 then shows that changing the prior regularising divergence can resolve VI's tendency to allow posteriors to over-concentrate, we call these methods generalised variational inference (GVI). We also show situations where methods failing to satisfy our axioms produces undesirable and non-transparent inference. We show that GVI is able improve upon state of the art performances for deep Gaussian processes and Bayesian neural networks.

The final chapter considers the challenging and widely applicable problem of detecting regime changes in multi-dimensional on-line streaming data, Bayesian online changepoint detection (BOCPD). BOCPD must use simple computable models in order to run in real time. The current methodology allows model misspecifications and outliers associated with these simple models to cause the detection of spurious

changepoints (CP). We robustify this analysis using the βD. We are able to prove results demonstrating that greater evidence is required in order to force the declaration of a CP when using the βD instead of the KLD. Additionally, we deploy a type of GVI algorithm to produces fast and accurate posterior inference that are suitable for on-line application. Applying this robustified algorithm to data recording air pollution in London finds a changepoint around the introduction of the congestion charge but, unlike previous methods does not detect any further regime changes.

Notation and Abbreviations

General Notation

$\pi(\cdot)$, $\pi(\cdot\|\mathbf{x})$	are distributions for θ
$\pi(\cdot\|\cdot)$	denotes a posterior distribution conditional on some observables
g	denotes the data generating process
$f(\cdot;\cdot)$	denotes the conditioning of a likelihood on its parameters or hyperparameters. Unless otherwise stated assume that all likelihood functions and data generating processes are absolutely continuous on their support
$\mathcal{N}(\mu, \sigma^2)$	denotes a univariate Gaussian distribution with mean μ and variance σ^2
$\mathcal{N}_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$	denotes a multivariate Gaussian distribution of dimension $p \geq 2$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathcal{IG}(a, b)$	denotes a inverse gamma distribution with shape a and scale b
$\mathcal{G}(a, b)$	denotes a gamma distribution with shape a and scale b

Divergence Abbreviations

$D(p\|\|q)$	A general divergence between p and q, assumed to be asymmetric
$D(p, q)$	A symmetric divergence between p and q
KLD	The Kullback-Leibler Divergence
TVD	The Total Variation Divergence
HD	The Hellinger Divergence
$\alpha\beta\gamma$D	The $\alpha\beta\gamma$-Divergence (referred to in text)
$D_G^{(\alpha,\beta,r)}$	The $\alpha\beta\gamma$-Divergence (referred to in equations with parameters α, β and γ)
αD	The α-Divergence (referred to in text)
$D_A^{(\alpha)}$	The α-Divergence (referred to in equations with parameter α)
Rényi-αD	The Rényi α-Divergence (referred to in text)
$D_{AR}^{(\alpha)}$	The Rényi α-Divergence (referred to in equations with parameter α)
βD	The β-Divergence (referred to in text)
$D_B^{(\beta)}$	The β-Divergence (referred to in equations with parameter β)
γD	The γ-Divergence (referred to in text)
$D_G^{(\gamma)}$	The γ-Divergence (referred to in equations with parameter γ)

General Abbreviations

IID	Independently and Identically Distributed
DM	The Decision Maker
DGP	The Data Generating Process
M-CLOSED	The M-closed world
M-OPEN	The M-open world
GBI	General Bayesian Inference
LLB	Loss Likelihood Bootstrap
MLE	Maximum Likelihood Estimate
MCMC	Markov Chain Monte Carlo
MDE	Minimum Divergence Estimation
MAP	Maximum A Posteriori
BMA	Bayesian Model Averaging
VI	Variational Inference
ELBO	Evidence Lower Bound
EP	Expectation Propagation
F-VI	Variational approximation to a posterior minimising divergence F between the approximating family and the actual posterior
VAE	Variational Auto-Encoder
GVI	Generalised Variational Inference
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MEDSE	Median Squared Error
NIG	Normal Inverse Gamma
BLR	Bayesian Linear Regression
AR	Auto-Regression
VAR	Vector Auto-Regression
BVAR	Bayesian Vector Auto-Regression
FDR	False Discovery Rate
SGD	Stochastic Gradient Descent
BNN	Bayesian Neural Network
GP	Gaussian Process
KDE	Kernel Density Estimate
CP	Changepoint
PPM	Product Partition Model
BOCPD	Bayesian On-line Changepoint Detection
RBOCPD	Robust Bayesian On-line Changepoint Detection
NOX	Nitrous Oxide

Chapter 1

Introduction

This thesis adopts the Bayesian paradigm of statistical analysis. There have been several fundamental developments of Bayesian statistics, the most notable of these belong to De Finetti [1931], Savage [1972] and most recently to Bernardo and Smith [2001]. Here, for simplicity we follow the methods as expressed in Bernardo and Smith [2001] and use this as a springboard to redevelop some of their approach.

1.1 Bayesian Decision Making

According to Bernardo and Smith [2001], coherent and rational decision making should take place via a Bayesian analysis. A decision problem consists of a set Θ of possible decisions that could be made in the face of a set of possible outcomes/states of the world $\mathcal{Z}$. The decision maker's (DM's) uncertainty over the space of outcomes should be characterised by a probability distribution, π, on $\mathcal{Z}$. The consequences (reward) of each possible outcome, $z \in \mathcal{Z}$, given decision $\theta \in \Theta$ should be characterised using a loss (utility) function, $\ell : \mathcal{Z} \times \Theta \to \mathbb{R}$. The loss function and belief distribution should then be combined to provide a preference ordering over decisions based on each decision's expected loss. The Bayes' optimal decision (often called the Bayes estimate) is the decision minimising this expected loss

$$\hat{\theta} := \arg\min_{\theta \in \Theta} \mathbb{E}_z \left[\ell(z, \theta) \right]. \tag{1.1}$$

One popular procedure to help the DM frame their beliefs about the unknown outcome, especially if this belief is going to be informed by some data $\boldsymbol{x}$ [O'Hagan, 2012], is to use Bayes' rule. Bayes' rule says that given initial/prior beliefs about z characterised as probability distribution $\pi(z)$, and a likelihood specifying the probability of observing data $\boldsymbol{x}$ under a particular state of the world z, $p(\boldsymbol{x}|z)$, then the updating of beliefs from before to after observing data $\boldsymbol{x}$ should follow

$$\pi(z|\boldsymbol{x}) = \frac{\pi(z)p(\boldsymbol{x}|z)}{p(\boldsymbol{x})}. \tag{1.2}$$

In Eq. (1.2), $\pi(z|\boldsymbol{x})$ is called the posterior belief or posterior density and the normalising constant, often called the marginal likelihood or model evidence, is $p(\boldsymbol{x}) = \int p(\boldsymbol{x}|z)\pi(z)dz$. In this way, Bayes' rule can be used by a DM to combine what they learn from the data with their beliefs prior to seeing the data in order to calculate the expectation in Eq. (1.1).

1.1.1 Inferential Decision Problems

Next we present how Bernardo and Smith [2001] express statistical inference within their framework, placing specific focus on the intrinsic relationship this has with the Kullback-Leibler divergence and the log-score. We aim to augment this methodology in later chapters of this thesis.

Bernardo and Smith [2001] frame statistical inference as a specific form of the decision making problem. Here the decision θ to be made is the parameter of a probability density $f(\cdot;\theta)$ for future observable z. This setting is convenient as the de Finetti representation theorem [De Finetti, 1931] states that under the assumption of exchangeable observations $(\boldsymbol{x}, z)$ that there exists a parameter θ such that $f(\boldsymbol{x}, z;\theta) = f(z;\theta)\prod_{i=1}^{n} f(x_i;\theta)$. Therefore this parameter θ provides conditional independence between previously observed data $\boldsymbol{x}$ and future exchangeable observation z. Unless otherwise stated we assume that $f(\boldsymbol{x};\theta)$ is absolutely continuous on the support $\mathcal{X}$.

In this situation Bernardo and Smith [2001] argue that the loss function associated with scoring probabilistic predictions (often called a 'scoring rule') should be proper and local. A proper scoring rule results in the DM's expected loss being minimised when the DM truly quotes their beliefs, and a local scoring rule is one where the score only depends on the quoted probability of the actual observed outcome and nothing else. It was proved long ago that the only proper, local scoring rule is the logarithmic scoring rule (log-score) [Good, 1952]:

Definition 1 (The logarithmic scoring rule (log-score) [Good, 1952])**.** For quoted probability density $f(\cdot;\theta)$ the log-score for a set of observations $\boldsymbol{z}$ in the support of f, is

$$\ell(\boldsymbol{z}, \theta) = \ell(\boldsymbol{z}, f(\cdot;\theta)) = \sum_{i=1}^{n} -\log\left(f(z_i;\theta)\right). \tag{1.3}$$

A quantity of interest for inferential decisions and the log-score is the expected extra loss when believing the data was distributed according to $f(\cdot;\theta)$ when it was actually distributed according to distribution g:

$$\mathbb{E}_{Z\sim g(\cdot)}[-\log(f(z;\theta))] - \mathbb{E}_{Z\sim g(\cdot)}[-\log(g(z))] =: \text{KLD}(g||f_\theta), \tag{1.4}$$

where KLD $(g||f_\theta)$ is the Kullback-Leibler divergence from the model f_θ to the data generating density g. As a result, minimising the log-score in expectation over observed data is equivalent to minimising the KLD to the data generating density g.

Definition 2 (The Kullback-Leibler Divergence (KLD) [Kullback and Leibler, 1951])**.** The KLD between probability densities $g(x)$ and $f(x)$ is given by

$$\text{KLD}(g||f) = \int g(x)\log\frac{g(x)}{f(x)}dx. \tag{1.5}$$

As is well known but often forgotten, someone who seeks to produce predictions minimising the log-score, or the KLD, will have to beware of approximating the probability of an event by 0 [Bernardo and Smith, 2001], since if they are wrong, they will incur an infinite loss.

This thesis focuses on the M-OPEN world, where the density $f(z;\theta)$ will be considered as misspecified. Precisely what this means is captured by the following definition provided by Bernardo and Smith [2001].

Definition 3 (M-CLOSED and M-OPEN world [Bernardo and Smith, 2001])**.** The M-CLOSED world assumption assumes there exists θ_0 such that the observed data was generated from the model with parameter θ_0, i.e.

$$\boldsymbol{x}_{1:n} = x_1, \ldots, x_n \sim f_{\theta_0}. \tag{1.6}$$

On the other hand the M-OPEN world assumes

$$\boldsymbol{x}_{1:n} = x_1, \ldots, x_n \sim g \tag{1.7}$$

and that there may well not exist a θ_0 such that $g = f_{\theta_0}$

While Definition 3 presents the M-CLOSED and M-OPEN world from the point of view of independently and identically distributed (IID) observations. Both regression and time series models can also be formulated using such a framework. Here, the observations are considered IID conditional on the value of some information, in regression this is the value of the corresponding predictors of that observation while in time series this corresponds to the filtration of information prior the observation. We elaborate on what exactly g might refer to in Section 1.1.2.

1.1.2 The $\mathcal{M}$-open world

While Bernardo and Smith [2001] provide the recipe for coherent and rational decision making in the face of uncertainty, they are quite explicit that their recipe is prescriptive rather than descriptive. Bernardo and Smith [2001] describe how people should make decision in an ideal world, not how people do make decisions in the real world. One particular part of this recipe that we focus on in this thesis is the DM's ability to specify their beliefs in terms of a probability distribution. Bernardo and Smith [2001] prescribe that this can be done by comparing a bet on the uncertain outcome, with a bet against some objectively defined probability, for example a roulette wheel or a coin flip. The prescription of Bernardo and Smith [2001] requires that this comparison can be done exactly and for as many probability statements that need to be elicited. When Bayes' rule is being used, this elicitation must be perfected for both the prior and the likelihood. In this thesis we will focus mainly on the likelihood specification, obtained after such an elicitation.

In Goldstein [1990] the author identifies that in order to correctly condition upon observed data in a Bayesian updating of beliefs, a model for the whole world in which the observations occur must be constructed. This requires many more probability specifications than any DM is ever going to be able to make, especially when eliciting continuous densities. So from a practical point of view, in order to implement a Bayesian analysis we must use some interpolating approximation of what the DM believes were they to have time to express it. So the formal analysis uses probabilistic densities that are not those the DM would use were they to have the infinite time necessary to reflect on them. This defines the subjectivist opinion of the $\mathcal{M}$-OPEN world - the model used for the belief updating is only ever feasibly an approximation of the DM's true beliefs.

Throughout this thesis I make the $\mathcal{M}$-OPEN world assumption. Within this framework we acknowledge that any class of models is unlikely to capture either the actual sampling distribution or exactly what the DM believes this to be. The model is then at best an approximate description the DM's beliefs, or of the underlying real world process. The celebrated quote of George Box that

"All models are wrong but some are useful"

epitomises the aims of this thesis. We acknowledge that the model used for inference is almost certainly misspecified relative to what we believe the data generating distribution really is, but proceed in the belief that we can still make use of a misspecified model.

The data generating process

Often in this thesis we use the phrase the "data generating process" (DGP). The data generating process is a widespread term in the literature and appears to suggest that 'Nature' or 'God' is using a simulator to generate observations. While this may fit nicely with many theoretical contributions, it becomes difficult to argue for in reality.

In this thesis we consider the DGP to represent the DM's true beliefs about the sample distribution of the observations. As is discussed above, in order correctly specify these the 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. Further to this while the DM was taking infinite time to consider their beliefs they would surely be obtaining further information that may very well change judgements they have already elicited. As a result these genuine beliefs must be approximated. In the special case when the data is the result of a draw from a known probability model - a common initial step in validating a methodology - then this thoughtful analysis and"a data generating process" obviously coincide.

Henceforth we use "the data generating process" in this sense to align our terminology as closely as possible with that in common usage. It is clearly debatable whether such a distribution exists or could ever be obtained. However, we feel considering the existence of such a distribution makes it more straightforward to present ideas relating to the degree of concordance between the DM's model and their actual beliefs.

We note that this definition does not prohibit the existence of an objectively defined data generating density encompassing all available information, but merely outlines that in reality we accept that different parties have different information and given finite time constraints and introspection prioritises different areas of the specification, providing a range of 'subjective' beliefs.

1.1.3 $\mathcal{M}$-open inference

Walker [2013] provides a principled justification for Bayes' rule in the $\mathcal{M}$-OPEN world.

This requires the DM to think about the KLD minimising parameter. When the model is correctly specified, the Bayesian learns about the parameter θ_0 that generated the data. By definition any statistical divergence $D(g||f)$ is minimised at 0 only when $g = f$. Therefore, in the $\mathcal{M}$-CLOSED world learning the parameter, θ_0, is equivalent to minimising $D(g(\cdot)||f(\cdot;\theta))$, for any statistical divergence $D(\cdot||\cdot)$. When the model is considered to be incorrect, there is no longer any formal relationship between the parameter θ and the data. The likelihood no longer represents the probability of the observed data conditioned on the parameter and Bayes' rule no longer provides the correct way to update beliefs based on conditional probability.

Therefore, in order for $\mathcal{M}$-OPEN inference to be meaningful a divergence measure must be chosen and the parameter of interest can then defined as

$$\theta^D = \arg\min_{\theta\in\Theta} D(g(\cdot)||f(\cdot;\theta)). \tag{1.8}$$

Walker [2013] then states that once the parameter of interest has been defined as the minimiser of some divergence, it is then possible for a DM to define their prior beliefs about where this may lie. The DM's final task is to ensure that their Bayesian learning machine is learning about the same parameter with which they defined their prior belief.

Recall that Bayesian updating, via Eq. (1.2), learns the parameters of the model which minimises the KLD of the model from the data generating density [e.g. Berk et al., 1966; Bissiri et al., 2016]. This enables the DM to continue to conduct belief updating in a principled fashion. Provided they are interested in, and specify prior beliefs about, the parameter θ^{KL} minimising the KLD of the model from the data generating process, then they should continue to use the standard Bayesian prior to posterior updating formula.

In summary, it is no longer possible or meaningful to learn about the parameter which generated the data in the $\mathcal{M}$-OPEN world. A statistical divergence measure must instead be specified between the fitted model and the genuine one in order to define the parameter targeted by the inference [Walker, 2013]. If a DM is using Bayes' rule then this divergence is the KLD

1.1.4 Current solutions to the $\mathcal{M}$-open world

Once the DM acknowledges that they are in the $\mathcal{M}$-OPEN world, they have several options currently available to them:

1. Proceed as though the model class does contain the true sample distribution and conduct some suite of *a posteriori* sensitivity analyses.

2. Modify the model class in order to improve its robustness properties.

3. Abandon the given parametric class and appeal to more data driven techniques.

4. Augment the model class with alternative plausible models and perform model averaging.

Method 1 is how Box [1980] recommends approaching parametric model estimation and is the most popular approach amongst statisticians. Although it is acknowledged that the model is only approximate, Bayesian inference is applied as though the statistician believes the model to be correct. The results are then checked to examine how sensitive these are to the approximations made. Authors Berger et al. [1994] provide a thorough review while Watson et al. [2016] consider this in a decision focused manner.

Method 2 corresponds to the classical robustness approach. Within this approach, one model within the parametric class may be substituted for a model providing heavier tails [O'Hagan, 1979; Berger et al., 1994]. Alternatively different estimators, for example M-estimators [Huber and Ronchetti, 1981; Hampel et al., 2011], see Greco et al. [2008] for a Bayesian analogue, are used instead of those justified by the model class.

The third possibility is to abandon any parametric description of the probability space. Examples of this solution include, empirical likelihood methods [Owen, 1991]; a decision focused general Bayesian update [Bissiri et al., 2016]; Bayesian non-parametric methods; or to appeal to statistical learning methods such as neural networks or support vector machines. Such methods simply substitute the assumptions and structure associated with the model class, for a much broader class of models. Thus, such non-parametric methods are not as free of misspecification related worries as they may seem.

The fourth option embeds the elicited model $f(\cdot;\theta)$ within a larger class of models, $\mathcal{M}$. Inference can then be conducted for all of these models and then averaged. Particularly relevant to this thesis is Bayesian model averaging (BMA) [Hoeting et al., 1999]. BMA averages the inferences from each models using weights given by the Bayesian model posterior

$$\pi(M_i|\boldsymbol{x}_{1:n}) \propto \pi(M_i)p_{M_i}(\boldsymbol{x}) = \pi(M_1)\int \pi(\theta)\prod_{i=1}^{n} f_{M_1}(x_i;\theta)d\theta, \tag{1.9}$$

produced by combining the marginal likelihood of each model with a prior on the model space. This is done in the hope that a larger class of models provides a more accurate representation of the DM's uncertainty and thus has a greater chance of containing the DGP.

In this sense BMA pretends that a $\mathcal{M}$-OPEN problem is $\mathcal{M}$-CLOSED. This can be seen from the fact that BMA is known to converge on the model whose inference are closest to the DGP in terms of KLD [see e.g. Rossell, 2018; Yao et al., 2018b] and thus only one model is used asymptotically. This extended model class may very well be able to get closer to the DGP but, for the reasons outlined above, in practice any finite model class is never going to contain the DGP. As a result this does not solve the problems of misspecification. Additionally BMA requires both careful elicitation and expensive computations to be done for many further models and thus may not be feasible in practice.

1.1.5 Bayes Linear methods

Here we provide a small exposition into Bayes line methods as they provides one of a few Bayesian methods specifically designed to produce principled posterior beliefs under the $\mathcal{M}$-OPEN world assumption and do not easily fit into the methods described above. General Bayesian updating provides a further, more recent example of such methodology and we introduce this at much greater length in Section 1.2.

Bayes linear methods [Goldstein, 1999] provide tools to move away from making a full probability specification in a principled manner. 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. (1.2).

In addition to this argument, Goldstein et al. [2006] carefully points out that there exists no result suggesting that conditioning as specified by Eq. (1.2) is the correct way to update beliefs. Using Bayes' rule describes now, how a DM's beliefs would change if they observed some data, but by the time this data has been observed there is inevitably more information the DM would wish to incorporate into their posterior belief, and Bayes' rule provides no provision for this.

A particularly interesting use for Bayes linear methods inline with the themes of this thesis was recently proposed by Williamson et al. [2015]. They observe that while it is prudent to conduct Bayesian sensitivity analysis in the $\mathcal{M}$-OPEN world to consider how model misspecification affects the inference, it is not clear how a subjective Bayesian should use the results of a sensitivity analysis to update their beliefs. Williamson et al. [2015] demonstrate that Bayes linear methods provides a principled tool to combine the results of several alternative analysis and produce provably superior beliefs for the posterior expectations of interest. This approach appears similar to that of BMA described above, but the Bayes linear approach is specifically designed and motivated considering the $\mathcal{M}$-open world. Additional work generalising BMA to the $\mathcal{M}$-open world can be found for example in Yao et al. [2018b] and the reference within.

1.2 General Bayesian Updating

Next we examine the general Bayesian updating of Bissiri et al. [2016] as this provides an important methodological tool that we use to extend the foundations of Bayesian inference beyond those discussed in Section 1.1.1.

The difficulties associated with specifying beliefs about the sample distribution of observed data $\boldsymbol{x}$ (see Section 1.1.3) motivate the general Bayesian updating of Bissiri, Holmes, and Walker [2016]. Bissiri et al. [2016] consider only conducting inference about the factors of the world/problem, θ, that matter to the DM and enter into their loss function $\ell(\theta, \boldsymbol{x})$, the DM's small world [Savage, 1972]. That is to say unlike Bayes' rule, θ need no longer be a parameter indexing a probability density. They produce a general Bayesian update - a coherent method to produce a posterior distribution over some quantity of interest without relying on a full model for the observations.

Firstly consider the Bayes act associated with beliefs corresponding to the data generating distribution $G(\cdot)$ and quantity of interest θ as

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{z \sim G}[\ell(\theta, z)] = \arg\min_{\theta} \int \ell(\theta, z) dG(z). \tag{1.10}$$

We note that this no longer needs to be an inferential decision problem, θ can be any quantity of interest. Rather than elicit a model representing the DM's beliefs over z, Bissiri et al. [2016] argue that given a prior, some data $\boldsymbol{x}$ and a loss function connecting the parameter to the observations, an updating of beliefs about θ must be possible in the absence of a model for the sampling distribution. They suggest that the optimal posterior distribution resulting from such an updating of beliefs should depend only on the prior and the information in the data through the loss function. As a result the updating must have the form

$$\pi(\theta|\boldsymbol{x}) = \psi\left\{\ell(\theta, \boldsymbol{x}), \pi(\theta)\right\}. \tag{1.11}$$

Additionally, they impose the following assumptions:

Assumption 1: Coherence (Bayesian Additivity)

$$\psi\left[\ell(\theta,\boldsymbol{x}_2),\psi\left\{\ell(\theta,\boldsymbol{x}_1),\pi(\theta)\right\}\right]=\psi\left\{\ell(\theta,\boldsymbol{x}_2)+\ell(\theta,\boldsymbol{x}_1),\pi(\theta)\right\}, \tag{1.12}$$

the posterior from part of the data becomes the prior for the rest.

Assumption 2: Invariance to a priori and a posteriori conditioning. For any set $A\subset\Theta$

$$\frac{\psi\left\{\ell(\theta,\boldsymbol{x}),\pi(\theta)\right\}}{\int_A\psi\left\{\ell(\theta,\boldsymbol{x}),\pi(\theta)\right\}d\theta}=\psi\left\{\ell(\theta,\boldsymbol{x}),\pi_A(\theta)\right\} \tag{1.13}$$

where $\pi_A(\theta)=\frac{\pi(\theta)\mathbf{1}(\theta\in A)}{\int_A\pi(\theta)d\theta}$. It does not matter if θ is restricted to be on A a priori or a posteriori.

Assumption 3: Larger loss (lower evidence) should result in lower posterior probability for a fixed prior. For $A\subset\Theta$, $\ell(\theta,\boldsymbol{x})>\ell(\theta,\mathbf{y})$ for $\theta\in A$ and $\ell(\theta,\boldsymbol{x})=\ell(\theta,\mathbf{y})$ for $\theta\in\Theta\setminus A$, then

$$\int_A\psi\left\{\ell(\theta,\boldsymbol{x}),\pi(\theta)\right\}d\theta<\int_A\psi\left\{\ell(\theta,\mathbf{y}),\pi(\theta)\right\}d\theta. \tag{1.14}$$

Assumption 4: No information provides no updating. If $\ell(\theta,\boldsymbol{x})$ is constant then $\psi\left\{\ell(\theta,\boldsymbol{x}),\pi(\theta)\right\}=\pi(\theta)$. If the sample provides no additional information about θ then the posterior should be equal to the prior.

Assumption 5: Invariance of the loss to an additive constant. If $\tilde{\ell}(\theta,\boldsymbol{x})=\ell(\theta,\boldsymbol{x})+c$ for some constant c then

$$\psi\left\{\tilde{\ell}(\theta,\boldsymbol{x}),\pi(\theta)\right\}=\psi\left\{\ell(\theta,\boldsymbol{x}),\pi(\theta)\right\}. \tag{1.15}$$

The existence of an updating rule and the above axioms, allow Bissiri et al. [2016] to construct a loss function $L(q|\pi,\ell,\boldsymbol{x})$ that elicits an optimal updating of beliefs. In the absence of a likelihood, Assumption 3 enforces that the optimal posterior should be 'close' the data in the sense that it minimises the expected loss of the observed data under the proposed posterior. Assumption 4 enforces that the optimal posterior should also be close to the prior, where the closeness between two distributions for θ can naturally be measured by a divergence $D(\cdot||\cdot)$. Bissiri et al. [2016] prove that KLD between posterior and prior is the only divergence which can guarantee that Assumption 1 and 2 are always satisfied. Lastly Assumption 5 enforces that the measures of closeness to the observed data and closeness to the prior are combined additively (We provide a proof for this result in Chapter 4, Theorem 9). Definition 4 defines the general Bayesian updating of beliefs and Theorem 1 demonstrates this is the solution to an objective function satisfying the formulation of Bissiri et al. [2016] and the above assumptions.

Definition 4 (The general Bayesian posterior)**.** The general Bayesian posterior given prior $\pi(\theta)$, data $\boldsymbol{x}$ and loss function $\ell(\cdot,\cdot)$ is

$$\pi^{GB}(\theta|\boldsymbol{x}) \propto \pi(\theta)\exp(-w\ell(\theta,\boldsymbol{x})) \tag{1.16}$$

where $w > 0$ is a calibration weight.

Theorem 1 (Derivation of the general Bayesian posterior)**.** The general Bayesian posterior defined in Eq. (1.16) is the solution to the following optimisation problem

$$\arg\min_{q\in\mathcal{P}} L(q|\pi,\ell,\boldsymbol{x}) = \arg\min_{q\in\mathcal{P}} \left\{\mathbb{E}_{\theta\sim q(\theta)}\left[w\ell(\theta;\boldsymbol{x})\right] + \textsc{kld}(q(\theta)||\pi(\theta))\right\} \tag{1.17}$$

where $\mathcal{P} = \left\{q(\theta) : \int q(\theta)d\theta = 1\right\}$

Proof. Rearranging the objective in Eq. (1.17) provides

$$\begin{aligned}
&\arg\min_{q\in\mathcal{P}} \left\{\mathbb{E}_{\theta\sim q(\theta)}\left[w\ell(\theta;\boldsymbol{x})\right] + \textsc{kld}(q(\theta)||\pi(\theta)\right\}\\
&= \arg\min_{q\in\mathcal{P}} \int \left\{w\ell(\theta;\boldsymbol{x}) + \log\frac{q(\theta)}{\pi(\theta)}\right\} q(\theta)d\theta\\
&= \arg\min_{q\in\mathcal{P}} \int \left\{\log\frac{q(\theta)}{\pi(\theta)\exp\left(-w\ell(\theta;\boldsymbol{x})\right)}\right\} q(\theta)d\theta\\
&= \arg\min_{q\in\mathcal{P}} \int \left\{\log\frac{q(\theta)}{\frac{\pi(\theta)\exp(-w\ell(\theta;\boldsymbol{x}))}{Z}}\right\} q(\theta)d\theta\\
&= \arg\min_{q\in\mathcal{P}} \textsc{kld}\left(q(\theta)||\frac{\pi(\theta)\exp\left(-w\ell(\theta;\boldsymbol{x})\right)}{Z}\right) \qquad (1.18)
\end{aligned}$$

Where $Z = \int \pi(\theta)\exp\left(-w\ell(\theta;\boldsymbol{x})\right)d\theta$ ensures that Eq. (1.18) is the KLD between two normalised densities. Lastly, by the definition of a divergence the $\textsc{kld}(p||q)$ is uniquely minimised when $p = q$ and as a result

$$\pi(\theta|\boldsymbol{x}) = \frac{\pi(\theta)\exp\left(-w\ell(\theta,\boldsymbol{x})\right)}{\int \pi(\theta)\exp\left(-w\ell(\theta,\boldsymbol{x}\right)d\theta} \tag{1.19}$$

□

The posterior in equation (1.19) is not an approximation or a pseudo-posterior, but rather a valid, coherent representation of subjective uncertainty in the minimiser of the loss function in Eq. (1.10). Often $\ell(\theta, \boldsymbol{x}) = \sum_{i=1}^{n} l(\theta, x_i)$ is the cumulative loss of the current data set. Updating using the cumulative loss amounts to replacing integrating over the data generating distribution in equation (1.10), with an empirical integration over the data whose distribution is $G(z)$.

In the above formulation the loss function is calibrated against the prior, by multiplying the loss by some positive constant w. While any probability density is defined so that it integrates to 1, loss functions are characterised by their minimisers and thus need only be defined up to a monotonic transformation. As a result loss functions can be scaled to be arbitrarily large or small. For example $\ell_1(x, \theta) = |x-\theta|$ and $\ell_2(x, \theta) = 10|x-\theta|$ can both be used to produce posterior distributions about the location of the median but will produce very different quantifications of uncertainty about the location of this median. Therefore, it is important that the weight the loss function has in the updating process is calibrated against the prior such that the posterior distributions resulting from the general Bayesian update provides a meaningful quantification of uncertainty in the absence of a model. One approach to do so is to match the width of the posteriors with those of the frequentists confidence intervals Lyddon et al. [2018]; Syring and Martin [2019]. We discuss Lyddon et al. [2018] in detail in Section 1.2.2 which also provides a Bayesian interpretation.

1.2.1 Recovering Bayes' Rule

It is straightforward to see that if the log-score is used in the general Bayesian update, the traditional Bayes' rule update is recovered:

$$\pi(\theta|\boldsymbol{x}) \propto \pi(\theta) \exp(-\sum_{i=1}^{n} -\log(f_\theta(x_i))) = \pi(\theta) \prod_{i=1}^{n} f_\theta(x_i). \tag{1.20}$$

This echoes the well-known result that the Bayesian predictive distribution finds the distribution that is closest to the DGP in terms of KLD [Berk et al., 1966]. While Bissiri et al. [2016] provide the framework to update a probability belief distribution using a loss function, they only consider the log-score in an inferential scenario. This however serves to demonstrate that in the $\mathcal{M}$-OPEN world Bayes rule is no longer updating conditional probabilities, instead it is providing the optimal posterior according to the above assumptions and objective function.

1.2.2 Calibrating the loss

Even before the introduction of GBI, setting the calibration weight $w \neq 1$ has been considered under the log-score. In this context the weight acts to temper the likelihood in Bayes' rule

$$\pi(\theta|\boldsymbol{x}) \propto \pi(\theta) \exp\left(-w \sum_{i=1}^{n} - \log\left(f(x_i;\theta)\right)\right) = \pi(\theta) \prod_{i=1}^{n} f(x_i;\theta)^w. \tag{1.21}$$

Likelihood tempering has been mainly used in associated with Monte-Carlo Methods [see e.g. Del Moral et al., 2006]. However, it has more recently started to appear in the model misspecification literature. The argument for this is that Bayes' rule only provides a valid quantification of posterior uncertainty if the model is correctly specified for the sample distribution of the data [Holmes and Walker, 2017]. When the model is misspecified often Bayes' rule learns too quickly [Grünwald, 2016; Holmes and Walker, 2017; Miller and Dunson, 2018]. That is to say it concentrates too quickly around the KLD minimising parameter. Posteriors over-concentrating can cause a decrease in predictive performance and is often considered a robustness issue associated with Bayes' rule. This can be combated by setting $w < 1$. For a thorough list of references see Holmes and Walker [2017].

When the loss function is not the log-score, the most compelling way to set w comes from Lyddon et al. [2018]. They set w by matching the asymptotic information in the GBI posterior with additive loss function $w\ell(\theta,\boldsymbol{x}) = w\sum_{i=1}^{n} \ell(\theta, x_i)$, with that of a sample from the 'loss-likelihood bootstrap' (LLB), a generalisation of Bayesian-bootstrap [Rubin, 1981] to general loss functions. The loss-likelihood bootstrap produces a sample of B parameter estimates using the following algorithm:

Algorithm 1 (The loss-likelihood bootstrap (LLB) [Lyddon et al., 2018])**.** Initalise number of samples B:

- For $j = 1 : B$
  - Sample $v_1, \dots, v_n \sim$ Dirichlet $(1, \dots, 1)$
  - Set $\theta^{(j)} = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} v_i \ell(\theta, x_i)$
- Output $\left(\theta^{(1)}, \dots, \theta^{(B)}\right)$

Lyddon et al. [2018] interpret the Dirichlet weights as being drawn from the posterior predictive of the Bayesian non-parametric Dirichlet process model for the data generating density with hyperparameter $\alpha = 0$. As a result in the absence of a model Lyddon et al. [2018] argue the LLB provides principled uncertainty quantification about the minimiser of $\int \ell(\theta, x)dG(x)$. This is further justified by the fact that under $\ell(\theta, x) = -\log f(x;\theta)$ and the Jeffrey's prior the asymptotic distribution of the LLB agree with the posterior produced by Bayes' rule if the likelihood is correctly specified. Moreover the optimisation done for each LLB iteration is independent of any weight w multiplying the loss function $\ell(\theta, x)$. As a result the LLB provides weight free uncertainty quantification. However the LLB does not leave room to input prior information. Lyddon et al. [2018] therefore suggest using the GBI posterior but setting w to match the asymptotic covariances of the GBI and LLB distributions.

1.3 Divergence and Loss Functions

Although the log-score and the KLD are currently intrinsic to Bayesian inference, there exist a multitude of other divergences, several of which can be conveniently related to a loss function. This section reviews the definition of a divergence measure and introduces several divergences of particular interest in this thesis.

1.3.1 Statistical Divergences

Throughout this section and the rest of the thesis we assume we have probability spaces whose measures are absolutely continuous with respect to some base measure, the Lebesgue measure for continuous state spaces and the counting measure for discrete spaces. As a result we are able to simplify the definitions of the divergences below in terms of normalised probability densities. Cichocki and Amari [2010] provide more general definitions. By default we assume we have absolutely continuous densities here.

In the $\mathcal{M}$-OPEN world it is no longer possible for an estimated statistical model to correctly capture the sample distribution of the data. As a result an important concept when considering inference in the $\mathcal{M}$-OPEN world is measuring the discrepancy between two distributions [see e.g. Walker, 2013]. Discrepancy measures of this kind are called statistical divergences [see e.g. Amari, 1985].

Definition 5 (Statistical Divergences [Eguchi et al., 1985])**.** A statistical divergence $D(g||f)$ is a measure of discrepancy between two probability densities f and g with the following two properties

1. $D(g||f) \geq 0, \forall f, g$

2. $D(g||f) = 0$ if and only if $g = f$

We note this definition is not sufficient for $D(g||f)$ to be a metric. Divergences are often asymmetrical and do not necessarily satisfy the triangle-inequality. Arguably the most famous divergence is the Kullback-Leibler Divergence (KLD) introduced in the seminal paper of Kullback and Leibler [1951] and appearing throughout the statistics literature. This was defined above in Eq. (1.5). However, there are many other families of divergences. One well-known family of divergences containing the KLD are the Csiszár–Morimoto ϕ-divergence [Csiszár, 1964; Morimoto, 1963]

Definition 6 (ϕ-divergence [Csiszár, 1964; Morimoto, 1963])**.** A ϕ-divergence, generated by convex function $\phi(\cdot)$ with $\phi(1) = 0$ is

$$D_{\phi}(g||f) = \int f(x)\phi\left(\frac{g(x)}{f(x)}\right) dx. \tag{1.22}$$

The restriction on ϕ to be convex and equal to 0 at 1 ensure these are proper divergences. The KLD is recovered as a member of the ϕ-divergence family with $\phi(t) = t\log t$. Other members of the ϕ-divergence family of particular interest here are the Total-Variation (TVD), Hellinger (HD) and α-Divergence (αD).

Definition 7 (The Total-Variation Divergence (TVD))**.** A TVD is given by

$$\text{TVD}(g||f) = \sup_{A\in\mathcal{F}} |g(A) - f(A)| = \frac{1}{2}\int |g(x) - f(x)|\, dx. \tag{1.23}$$

The TVD is a member of the ϕ-divergence family with $\phi(t) = \frac{1}{2}|t-1|$.

Definition 8 (The Hellinger Divergence (HD) [Hellinger, 1909])**.** A HD is given by

$$\begin{aligned}\text{HD}(g||f) &= \frac{1}{2}\int \left(\sqrt{g(x)} - \sqrt{f(x)}\right)^2 dx\\ &= 1 - \int \sqrt{g(x)}\sqrt{f(x)}dx. \end{aligned}\tag{1.24}$$

The HD is a member of the ϕ-divergence family with $\phi(t) = 1 - \sqrt{t}$.

Both the TVD and the HD are symmetric and satisfy the triangle inequality and are therefore proper metrics.

Definition 9 (The α-divergence (αD) [Chernoff et al., 1952; Liese and Vajda, 1987; Amari, 1985])**.** The αD is defined as

$$D_A^{(\alpha)}(g(x)||f(x)) = \frac{1}{\alpha(1-\alpha)}\left\{1 - \int g(x)^{\alpha} f(x)^{1-\alpha} dx\right\}, \tag{1.25}$$

where $\alpha \in \mathbb{R} \setminus \{0,1\}$. $D_A^{(\alpha)}$ is a member of the ϕ-divergence family with $\phi(t) = \frac{t-t^{\alpha}}{\alpha(1-\alpha)}$. When $\alpha = 1$, $D_A^{(1)}(g(x)||f(x)) = \text{KLD}(g(x)||f(x))$ and when $\alpha = 0$, $D_a^{(0)}(g(x)||f(x)) = \text{KLD}(f(x)||g(x))$.

There are several parametrisations of the αD, for example the Amari notation parametrised by α_A with $\alpha = \frac{1-\alpha_A}{2}$, or the Cressi-Read notation [Cressie and Read, 1984] with $\alpha = \lambda + 1$.

Another well-known family of divergences also containing the KLD are the Bregman-divergences [Bregman, 1967].

Definition 10 (Bregman-divergences [Bregman, 1967])**.** A Bregman-divergence, generated by strictly convex continuously differentiable function $\psi(\cdot)$ with first derivative $\nabla\psi(\cdot)$ is

$$D_{\psi}(g||f) = \int \left[\psi(g(x)) - \psi(f(x)) - (g(x) - f(x))\,\nabla\psi\,(f(x))\right] dx. \tag{1.26}$$

An important member of the Bregman divergence family for this thesis is the β-Divergence (βD) [Basu et al., 1998; Mihoko and Eguchi, 2002].

Definition 11 (The β-divergence ($D_B^{(\beta)}$) [Basu et al., 1998; Mihoko and Eguchi, 2002])**.** The βD is defined as

$$\begin{aligned} &D_B^{(\beta)}(g(x)||f(x)) \\ &= \frac{1}{\beta(\beta-1)}\int g(x)^{\beta}dx + \frac{1}{\beta}\int f(x)^{\beta}dx - \frac{1}{\beta-1}\int g(x)f(x)^{\beta-1}dx, \end{aligned} \tag{1.27}$$

where $\beta \in \mathbb{R} \setminus \{0,1\}$. $D_B^{(\beta)}$ is a member of the Bregman-divergence family with $\psi(t) = \frac{1}{\beta(\beta-1)}t^{\beta}$. When $\beta = 1$, $D_B^{(1)}(g(x)||f(x)) = \text{KLD}(g(x)||f(x))$.

The βD has often been referred to as the Density-Power Divergence (DPD) in the statistics literature [Basu et al., 1998] where it is often parametrised as $\beta = \beta_{DPD} + 1$.

A third, less-known divergence family, containing the KLD and cases of ϕ- and Bregman divergences of interest here, is the $\alpha\beta\gamma$-divergence ($\alpha\beta\gamma$D) family introduced in Cichocki and Amari [2010]. The $\alpha\beta\gamma$D family contains the KLD, the

αD and the βD as well as the γ-divergence (γD) [Fujisawa and Eguchi, 2008] and the Rényi α-divergence (RÉNYI-αD) [Rényi et al., 1961]

Definition 12 (The $\alpha\beta\gamma$-divergence $D_G^{(\alpha,\beta,r)}$ [Cichocki and Amari, 2010])**.** The $\alpha\beta\gamma$-divergence $D_G^{(\alpha,\beta,r)}$ Cichocki and Amari [2010] takes the form

$$\begin{aligned} &D_G^{(\alpha,\beta,r)}(g(x)||f(x)) \\ &= \frac{1}{\alpha(\beta-1)(\alpha+\beta-1)r}\left[\left(\tilde{D}_G^{(\alpha,\beta)}(g(x))||f(x))+1\right)^r-1\right] \end{aligned} \tag{1.28}$$

where $r > 0$ and

$$\begin{aligned} &\tilde{D}_G^{(\alpha,\beta)}(g(x)||f(x)) \\ &= \int\left(\alpha g(x)^{\alpha+\beta-1}+(\beta-1)f(x)^{\alpha+\beta-1}-(\alpha+\beta-1)g(x)^{\alpha}f(x)^{\beta-1}\right)dx \end{aligned}$$

with $\alpha \neq 0$, $\beta \neq 1$. When $r = \alpha = \beta = 1$, $D_G^{(1,1,1)}(g(x)||f(x)) = \text{KLD}(g(x)||f(x))$.

This exposition is a summary of the excellent review conducted in Cichocki and Amari [2010]. We note that the parametrizations of these divergences may vary throughout the literature.

Remark 1 (The αD and βD as a members of the $\alpha\beta\gamma$D)**.** The αD and the βD are contained within the $D_G^{(\alpha,\beta,r)}$:

- The αD is recovered from $\alpha\beta\gamma$D when $r = 1$ and $\beta = 2 - \alpha$.
- The βD is recovered from $\alpha\beta\gamma$D when $r = \alpha = 1$.

Definition 13 (The Rényi α-divergence ($D_{AR}^{(\alpha)}$) [Rényi et al., 1961])**.** The Rényi [Rényi et al., 1961] α-divergence is defined as

$$D_{AR}^{(\alpha)}(g(x)||f(x)) = \frac{1}{\alpha(\alpha-1)}\log\left(\int g(x)^{\alpha}f(x)^{1-\alpha}dx\right), \tag{1.29}$$

where $\alpha \in \mathbb{R} \setminus \{0, 1\}$. $D_{AR}^{(\alpha)}$ is recovered from $D_G^{(\alpha,\beta,r)}$ in the limit as $r \to 0$ and $\beta = 2 - \alpha$. When $\alpha = 1$, $D_{AR}^{(1)}(g(x)||f(x)) = \text{KLD}(g(x)||f(x))$ while for $\alpha = 0$, $D_{AR}^{(0)}(g(x)||f(x)) = \text{KLD}(f(x)||g(x))$. Note that we use the scaled version of the RÉNYI-αD proposed by Liese and Vajda [1987] and frequently used in the literature [e.g. Cichocki and Amari, 2010]

In fact the RÉNYI-αD is an invertible function of the αD since

$$D_{AR}^{(\alpha)}(g(x)||f(x)) = \frac{1}{\alpha(\alpha-1)}\log\left(1-\alpha(1-\alpha)D_A^{(\alpha)}(g(x)||f(x))\right).$$

Definition 14 (The γ-divergence ($D_G^{(\gamma)}$) Fujisawa and Eguchi [2008])**.** The γ-divergence Fujisawa and Eguchi [2008] is defined as

$$D_G^{(\gamma)}(g(x)||f(x)) = \frac{1}{\gamma(\gamma-1)} \log \frac{\left(\int g(x)^\gamma dx\right)\left(\int f(x)^\gamma dx\right)^{\gamma-1}}{\left(\int g(x) f(x)^\gamma dx\right)^{\frac{1}{\gamma-1}}}, \tag{1.30}$$

where $\gamma \in \mathbb{R} \setminus \{0, 1\}$. $D_G^{(\gamma)}$ is recovered from $D_G^{(\alpha,\beta,r)}$ in the limit as $r \to 0$, $\alpha = 1$ and $\beta = \gamma$. When $\gamma = 1$, $D_G^{(1)}(g(x)||f(x)) = \text{KLD}(g(x)||f(x))$.

The γD can be shown to be generated from the βD by applying the following transformation

$$c_0 \int g(x)^{c_1} f(x)^{c_2} dx \to c_0 \log \int g(x)^{c_1} f(x)^{c_2} dx,$$

to all three terms of βD. The Rényi-αD can be shown to be generated by αD under the same transformation of its two terms.

Remark 2 (Recovering the KLD)**.** As is pointed out above, the parametrised divergences αD, Rényi-αD, βD and γD all recover the KLD in the limit as $\alpha = \beta = \gamma \to 1$. This can be shown using the 'replica trick':

$$\lim_{x \to 0} \frac{Z^x - 1}{x} = \log(Z).$$

1.3.2 Loss functions

When attempting to estimate a model by minimising one of the above divergences [Walker, 2013], $D(g(x), f(x;\theta))$, one never has direct access to $g(x)$. One does however usually have access to samples $x_1, \ldots, x_n$ assumed to have been 'generated' by g. As a result, one can hope to measure closeness between $f(\cdot;\theta)$ and the empirical measure $\hat{g}_n$ of the sample. Fortunately some divergences have a natural interpretation in terms of loss functions [e.g. Bernardo and Smith, 2001; Grünwald and Dawid, 2004; Dawid, 2007]. For these divergences, there exists a function $\ell : \mathcal{X} \times \mathcal{F}(\mathbf{\Theta}) \to \mathbb{R}$ mapping from the samples space $\mathcal{X}$ and the space of probability measures $\mathcal{F}(\mathbf{\Theta})$ on $\mathbf{\Theta}$ into the real numbers such that one can write

$$D(g(x)||f(x;\theta)) = \mathbb{E}_{g(x)}\left[\ell(x, f(\cdot;\theta))\right] - \mathbb{E}_{g(x)}\left[\ell(x, g)\right]. \tag{1.31}$$

The loss function interpretation is then as follows: A divergence taking the form of Eq. (1.31) is the excess expected penalty incurred for believing x was distributed according to $f(\cdot;\theta)$ when it was actually distributed according to g. Section 1.1.1 showed that the KLD can be written in this form using the log-score. The 'entropy' of the DGP g, $\mathbb{E}_{g(x)}\left[\ell(x,g)\right]$, is unaffected by f. Hence, finding f minimising $\text{KLD}(g||f(\cdot;\theta))$ is equivalent to finding f minimising $\mathbb{E}_{g(x)}\left[-\log f(x;\theta)\right]$. Lastly the expectation under the data generating distribution g can be approximated by the empirical distribution of the sample $\mathbb{E}_{g(x)}\left[-\log f(x;\theta)\right] \approx \mathbb{E}_{\hat{g}_n}(x)\left[-\log f(x;\theta)\right] = \frac{1}{n}\sum_{i=1}^{n} -\log f(x_i;\theta)$.

Not all divergence naturally take the form of Eq. (1.31). Another divergence that does, and is of interest for this thesis is the βD.

$$\begin{aligned} D_B^{(\beta)}(g||f(\cdot;\theta)) &= \frac{1}{\beta}\int f(x;\theta)^{\beta}dx - \frac{1}{(\beta-1)}\int f(x;\theta)^{\beta-1}g(x)dx + \frac{1}{\beta(\beta-1)}\int g(x)^{\beta}dx \\ &= \mathbb{E}_{g(x)}\left[\frac{1}{\beta}\int f(z;\theta)^{\beta}dz - \frac{1}{(\beta-1)}f(x;\theta)^{\beta-1}\right] \qquad (1.32) \\ &\quad -\mathbb{E}_{g(x)}\left[\frac{1}{\beta}\int g(z)^{\beta}dz - \frac{1}{(\beta-1)}g(x)^{\beta-1}\right]. \end{aligned}$$

So

$$\ell_{\beta}(x, f(\cdot;\theta)) = \frac{1}{\beta}\int f(z;\theta)^{\beta}dz - \frac{1}{(\beta-1)}f(x;\theta)^{\beta-1}, \qquad (1.33)$$

which is available in closed form for many exponential families and only depends on the form of $f(x;\theta)$.

1.3.3 Relationships between Divergences

Important throughout this thesis is going to be how the divergences introduced in Section 1.3.1 relate to one an other for fixed g and f, and specifically how they relate to the TVD. This section provides a summary of the well-known bounds currently available in the literature. Pinsker's inequality [Pinkser, 1964] is a well-known bound relating the KLD and the TVD.

$$\text{TVD}(g,f) \leq \sqrt{\frac{1}{2}\text{KLD}(g||f)} \qquad (1.34)$$

Much work has been put into finding so called reverse Pinsker's inequalities. Defining $a = \operatorname{ess\,inf}\frac{dF}{dG}$ and $b = \operatorname{ess\,sup}\frac{dF}{dG}$ Sason and Verdu [2016]; Binette [2019] show that

$$\text{KLD}(g||f) \leq C_{g,f}\text{TVD}(g,f), \text{ where } C_{g,f} = \left(\frac{\log(a)}{a-1} + \frac{\log(b)}{1-b}\right) \qquad (1.35)$$

provided a and b are well defined. However for two continuous densities with support $(-\infty,\infty)$ a and b are often not well defined, see Section 2.3.

The Hellinger divergence can be shown to bound the TV-divergence both above and below [Devroye and Gyorfi, 1985; Liese and Vajda, 1987]:

$$\textsc{hd}(g,f) \leq \textsc{tvd}(g,f) \leq \sqrt{\textsc{hd}(g,f)}\sqrt{2-\textsc{hd}(g,f)} \leq \sqrt{2\textsc{hd}(g,f)}. \tag{1.36}$$

Lastly the αD is bounded above by the TVD ([Prop. 2.35 Liese and Vajda, 1987],[Cor. 1 Sason and Verdu, 2016])

$$\alpha(1-\alpha)d_\alpha(g,f) \leq d_{TV}(g,f). \tag{1.37}$$

1.4 Summary

This introductory chapter has introduced the foundations of Bayesian inference through the lens of decision theory and has identified the intrinsic link between Bayes' rule and the KLD. We have introduced the $\mathcal{M}$-OPEN world, where the model used for inference is considered to be misspecified and outlined current approaches to conduct statistical inference in this paradigm. We have specifically emphasised general Bayesian updating, a principled, state of the art method to produce posterior beliefs about the minimiser of a general loss function. Lastly we have introduced several additional divergences to the KLD and presented some relationships between these.

The following chapters will build upon the foundations laid here. Chapter 2 will use general Bayesian updating to conduct Bayesian parametric model inference aimed at minimising divergences alternative to the KLD. This is motivated by a desire to produce inference that is useful for a general decision problem, instead of solely for the sake of inference. Chapter 3 will extend the motivation behind using these alternative divergences to consider how stable their inferences are to the subjective choice of a likelihood model across an equivalence class. Chapter 4 tackles some of the computational issues of producing posterior inferences when moving away from Bayes' rule and embeds the methods discussed in Chapters 2 and 3 within a generalised inference framework. Lastly Chapter 5 uses the developments of Chapters 2, 3 and 4 to produce an algorithm that conducts robust on-line changepoint detection. This algorithm is applied to a dataset recording air pollution levels in London and is able to detect the introduction of the London congestion change but ignore spurious changepoints resulting from model misspecification.

Chapter 2

Principles of Bayesian Inference Using General Divergence Criteria

In this chapter we will discuss the merits of conducting traditional Bayesian inference in the $\mathcal{M}$-OPEN world. We will take a decision theoretic approach throughout, assuming that the DM is performing inference in order to produce estimates of expected utility. Under this paradigm we establish that Bayes' rule may not provide a suitable method to adjust beliefs when observing data. We instead propose learning by minimising alternative divergence to the KLD. Several of the developments proposed in this chapter have been published in Jewson, Smith, and Holmes [2018].

This chapter is organised as follows: Section 2.1 demonstrates the lack of robustness provided by Bayes' rule when the observed data contains outliers. Section 2.2 outlines our opinion that this is the fault of the belief updating mechanism rather than the specific model used for inference. Section 2.3 identifies that the lack of robustness from Bayes' rule is a result of the motivation for Bayes' rule as solving an inferential decision problem. Instead, we motivate $\mathcal{M}$-OPEN inference as desirable if it is useful for a general decision problem. This in turn motivates minimising alternative divergences to the KLD. Section 2.4 reviews existing literature minimising divergence alternative to the KLD and illustrates the flaws in their justification in the $\mathcal{M}$-OPEN world. Section 2.5 uses general Bayesian updating to provide a principled justification for a belief update minimising any divergence. Section 2.6 considers several divergences that could be used for inference, motivating them by their ability to produce accurate estimates of expected utilities. Lastly, Section 2.7 illustrates the impact changing the divergence can have in the $\mathcal{M}$-OPEN world, whilst also demonstrating their performance is not much worse than Bayes' rule in the $\mathcal{M}$-CLOSED world.

2.1 Bayes' rule in the $\mathcal{M}$-open world

In the $\mathcal{M}$-CLOSED world, where the fitted model class is implicitly assumed to contain the sample distribution from which the data came, Bayesian updating is highly compelling. However, modern statisticians are increasingly acknowledging that their inference is taking place in the $\mathcal{M}$-OPEN world [Bernardo and Smith, 2001]. Remarkably, as we pointed out in Chapter 1, in the $\mathcal{M}$-OPEN world standard Bayesian updating can be seen as a method which learns a model by minimising the KLD from the model from which the data were sampled [Berk et al., 1966; Bissiri et al., 2016]. Therefore traditional Bayesian updating turns out to still be a well principled method for belief updating provided the decision maker (DM) concerns themselves with the parameters of the model that are closest to the data as measured by KLD [Walker, 2013].

Viewed conversely, Walker [2013] identified that in order to do principled inference, a DM must aim to produce predictions that are closest to the DGP in terms of KLD. However as is well known - but also often forgotten - correctly capturing the tail characteristics of the sample distribution will be of up-most importance to the DM if they aim to minimise the KLD [Bernardo and Smith, 2001]. Minimising the KLD $(g||f_\theta)$ is equivalent to minimising

$$\mathbb{E}_{g(x)}\left[-\log f(x;\theta)\right] \approx \frac{1}{n}\sum_{i=1}^{n} -\log f(x_i;\theta), \tag{2.1}$$

see Section 1.3.2. Now $\lim_{f(x;\theta)\rightarrow 0} -\log f(x;\theta) = \infty$ so assigning low probability to any observation will incur a large penalty according to the KLD.

This phenomenon is particularly evident when the DGP has heavier tails than the model, a phenomenon commonly caused by, but not limited to, the presence of outliers in the dataset. Outliers are observations that appear inconsistent with the rest of the data, which is equivalent to saying they occur in the tails of a distribution fitted to the rest of the data, i.e. unlikely given the rest of the data. In order to illustrate this consider a simple motivating example.

Example (ϵ-contamination)**.** For $\epsilon \in (0, 0.5)$

$$g(x) = (1-\epsilon)\mathcal{N}\left(x;\mu_u,\sigma_u^2\right) + \epsilon\mathcal{N}\left(x;\mu_c,\sigma_c^2\right)$$
$$f(x;\mu,\sigma^2) = \mathcal{N}\left(x;\mu,\sigma^2\right),$$

where $\theta = (\mu, \sigma^2)$ are the model parameters and $(\epsilon, \mu_u, \sigma_u^2, \mu_c, \sigma_c^2)$ are fixed features of the DGP.

Here the model has the flexibility to match at least $(1-\epsilon) \times 100\%$ of the data but the remaining $\epsilon \times 100\%$ corresponds to an outlying sub-population. Fig. 2.1 shows a histogram of $n = 1000$ observations from $g(x)$ with $(\epsilon = 0.1, \mu_u = 0, \sigma_u^2 = 1^2, \mu_c = 5, \sigma_c^2 = 3^2)$ along with the traditional Bayesian predictive distribution under model $f(x;\mu,\sigma^2)$ using Normal-Inverse-Gamma (NIG) prior, $\pi(\mu,\sigma^2) = \mathcal{N}\left(\mu;\mu_0,v_0\sigma^2\right)\mathcal{IG}(\sigma^2;a_0,b_0)$, with hyperparameters $(a_0 = 3, b_0 = 2, \mu_0 = 0, v_0 = 10)$.

Fig. 2.1 demonstrates how the tail observations resulting from the ϵ-contamination force the Bayesian predictive's variance to increase and its mean to shift towards the right-hand tail, away from 90% of the data. In fact the log-density plots shows that the Bayesian predictive and data generating models agree at the mean of the contaminant population suggesting that Bayes' rule is solely concerned with capturing the outlying contamination at the expense of matching to the majority of observations.

This example may seem contrived but we simply aim for this to act as a caricature of situations that often arise in real world applications. All that is required for an ϵ-contamination is that there is some unmodelled sub-population of the data who behave differently. One might argue this sub-population is easy to spot in our caricature. However, the ability to discern the contaminating population becomes increasingly difficult as the dimension of the sample space increases to those encountered in real, high-dimensional problems. See Section 2.7.5 and 5.5.2 for a further discussion on this.

In the seminal book of Huber [2011] robustness is defined as "an insensitivity to small deviations from the assumption". This simple example goes to show that the usual Bayesian belief adjustment using Bayes' rule is non-robust in this sense.

2.2 The Model

In Section 1.1.4 we discussed current approaches to dealing with the lack of robustness in $\mathcal{M}$-OPEN Bayesian inference. Although alternatives 2 and 3 can be shown to

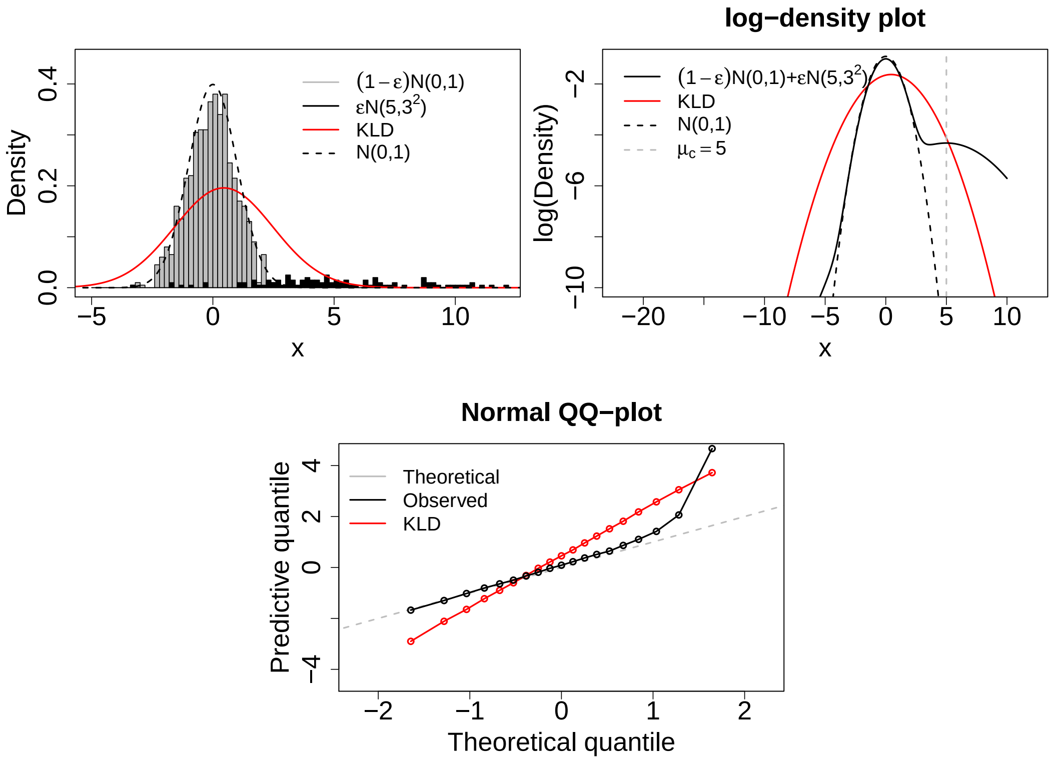

Figure 2.1: Traditional Bayesian updating applied under ϵ-contamination. $n = 1000$ observations generated from $g(x) = 0.9\mathcal{N}\left(x; 0, 1^2\right) + 0.1\mathcal{N}\left(x; 5, 3^2\right)$ with Bayesian model $f(x; \mu, \sigma^2) = \mathcal{N}\left(x; \mu, \sigma^2\right)$ under NIG prior $\pi(\mu, \sigma^2) = \mathcal{N}\left(\mu; 0, 10\sigma^2\right)\mathcal{IG}(\sigma^2; 2, 2)$. **Left:** Histogram of the data, separating the 90% **uncontaminated** points from the 10% **contaminated** points, with the **Bayesian predictive** overlayed. **Right:** Log-Density plot comparing the log-density of $g(x)$ with the traditional Bayes predictive and a standard Gaussian. Note how the densities of the **Bayesian predictive** and $g(x)$ agree at the mean of the contamination. **Bottom:** Gaussian QQ plot comparing the observed quantiles of the data with that of the fitted Bayesian model and standard Gaussian

be very powerful in certain scenarios, it is our opinion that these will not in general be entirely satisfactory. In an applied statistical problem the model(s) represents the DM's best guess for the sample distribution of the data. The model provides the only opportunity to input not only structural, but quantitative expert judgements about the domain, something that is often critical to a successful analysis - see Lazer et al. [2014] for example. The model provides an interpretable and transparent explanation about how different factors might be related to each other. This type of evidence is often essential when advising on decisions to policy makers. Often, simple assumptions play important roles in providing interpretability to the model and in particular preventing it from over fitting to any non-generic features contained in any one data set.

For the above reasons an unambiguous statement of a model, however simple, is in our opinion an essential element for much of applied statistics. In light of this, statistical methodology should also be sufficiently flexible to cope with the fact that the DM's model is inevitably an approximation both of their beliefs and the real world process. Currently, standard Bayesian statistics sometimes struggles to do this well. We argue below that this is because it implicitly minimises the KLD to the underlying process. Namely we believe a lack of robustness is the fault of the current Bayesian updating machinery rather than any one particular model.

2.3 Inferential Procedure or Decision Problem

Throughout this chapter we assume that inference is being done in order to facilitate principled decision making. That is to say that the goal of the Bayesian analysis is to learn posterior (predictive) distributions for some future uncertain quantity in order to take the optimal decision whose consequences (utility) depend on this future uncertainty.

This is the situation considered by the foundations of Bayesian statistics, but this is all too often forgotten. In this situation desirable inferences are going to be those that provide accurate estimates of the expected utilities associated with each possible decision. We call this our principle of M-OPEN inference for decision making. We argue that inferential procedures, at least for the purposes discussed in this thesis, should be designed in the knowledge that the posterior judgements will eventually be used to minimise an expected loss. In the M-CLOSED world we can be assured that Bayes' rule will be targeting the distribution that generated that data and as a result the expected loss estimates we produce will be accurate (at least asymptotically). We show here that in the M-OPEN world it is no longer clear that this is happening.

The log-score, which Bayes' rule is minimising, is supposedly motivated as reflecting the fact that the "tails of the distribution are, generally speaking, extremely important in pure inference problems..." [Bernardo and Smith, 2001, Ch. 2, p. 76]. Exactly what is meant by a pure inference problem is hard to interpret. It suggests that inference is only being done to best estimate the DGP and with no use for this estimate of the DGP in mind. In such situations, the properties of propriety and locality (Section 1.1.1) associated with the log-score may well be all a DM cares about. However, such situations are surely only hypothetical. If there is no use in mind for inference then why is it being done in the first place? We interpret the fact that the log-score is supposedly optimal for inference problems as a consequence of the intimate link between the log-score and Bayes' rule inference, outlined in Section 1.2.1. If one truly believe the probabilities specified by their likelihood and prior at the instant the updating takes place then Bayes' rule is probabilistically the correct way to update.

Clearly there do exist decision problems where the tails of the fitted probability distribution are important. Particularly, in situations where the losses are unbounded, for example in gambling and odds setting scenarios, correctly capturing tail behaviour will be essential. However, many loss functions connected to real decisions may be less concerned with how the fitted model approximated the tails of the observed data. Bernardo and Smith [2001] proceed to point out that "...this is in contrast to many practical decision problems where the form of the utility (loss) function often makes the solution robust with respect to changes in the tails of the distribution assumed". For entirely reasonable practical reasons these rare tail events are precisely ones which the DM will find hard to accurately elicit [O'Hagan et al., 2006], see Winkler and Murphy [1968] for a demonstration of this in the context of forecasting the probability of precipitation. As will be demonstrated next, inferential procedures assuming an unbounded loss function such as the log-score associated with the KLD, provide no guarantees about performance in more general decision making scenarios.

One guarantee that a (predictive) distribution provides accurate estimates of a bounded expected loss is provided by the TVD. If $\text{TVD}(g, f) \leq \epsilon$ then for any loss function bounded between 0 and 1, the expected loss $\mathbb{E}_{z\sim g}\left[\ell\left(z, \theta_f^*\right)\right]$ under g of making decision θ_f^*, the optimal decision believing the data was distributed according to f, is at most 2ϵ greater than the expected loss incurred from making the optimal decision under g, $\mathbb{E}_{z\sim g}\left[\ell\left(z, \theta_g^*\right)\right]$ (see Smith [2010] for example). Therefore if the predictive distribution is close to the DGP in terms of TVD then the consequences of the model misspecification in terms of the expected loss of the decision made is small. This result suggests that when informed decision making is the goal of the statistical analysis, closeness in terms of TVD ought to be the canonical criteria the DM demands. We note that the assumption of a bounded loss function is not that strong. As Watson et al. [2016] correctly point out, it is almost always possible to bound any 'real-world' loss function, for example using some arbitrary maximum or minimum, and actually that any method using MCMC to sample from the posterior/predictive must be assuming a bounded loss function

The KLD forms an upper bound on the TVD through Pinsker's inequality Pinkser [1964] (see Section 1.3.3). As a result if the KLD between the DGP and the fitted model is small, then we can be sure the TVD will be. The reverse Pinsker's inequality (Section 1.3.3) provides an upper bound for the KLD in terms of the TVD but this bound depends on a further multiplier $C_{g,f}$. Without greater knowledge about the unknown density of the DGP, g, and how it is approximated by f_θ, we cannot be sure of the size of the $C_{g,f}$, with it very possibly being infinite.

As a result there are situations where the TVD is very small but the KLD is very large. Specifically this occurs when $\frac{dF}{dG}$ is very small, which is a consequence of the tails of F being lighter than the tails of G, as in ϵ-contamination example. In this scenario, a predictive distribution whose associated optimal decisions achieve a close to optimal expected loss estimate (as the distribution is close in TVD) will receive very little posterior mass when using Bayes' rule. This is clearly undesirable in a decision making context.

It is well known that when the KLD is large, which is almost inevitable in high dimensional problems, the KLD minimising models can give a very poor approximation when our main interest is in getting the central part of the posterior model uncertainty well estimated. This is a direct consequence of the KLD giving overarching priority to correctly specifying the tails of the generating sample distribution. As a result in the M-OPEN world when decision making with respect to a bounded utility function is the goal of the analysis, Bayes' rule is no longer necessarily the best way to update beliefs. The log-score provides a poor surrogate for the actual loss functions at hand!

2.3.1 Moving away from KL-divergence in the M-open world

Once the consequences of using Bayes' rule to solve decision problems in the M-OPEN world is understood, we believe that DMs may reasonably desire alternative options for parameter updating that are as well principled as Bayes rule, but place less importance on tail misspecification and are more focused towards their decision problem. In this new era of "big data" it becomes increasingly likely that the model used for inference is misspecified, especially in the tails of the process - see Section 2.7.5. We believe many DMs would consider it desirable for models that approximate the distribution of the majority of the data not to be disregarded because they poorly fit a few outlying observations.

By fitting model parameter in a way that can be non-robust, the DM is having to combine their best guess belief model with something that will be robust to the parameter fitting. The DM must seek the best representation of their beliefs about a process in order to make future predictions. However under traditional Bayesian updating they must also give consideration to how robust these beliefs are. This seems an unfair task to ask of the DM. We therefore propose to decouple what the DM believes about the DGP, from how the DM wishes the model to be fitted. To enact this we add the following option to the list of solutions to the $\mathcal{M}$-OPEN world (Section 1.1.4):

5. Acknowledge that the model class is only approximate, but is the best available, and seek to infer the model parameters in a way that are most useful to the decision maker.

Option 5 suggests that the DM may actually want to explicitly target a more robust divergence than the KLD when conducting inference, a framework commonly known as Minimum Divergence Estimation (MDE), see Basu et al. [2011]. Minimum divergence estimation is of course a well-developed field by frequentists, with Bayesian contributions coming more recently. However when the realistic assumption of being in the $\mathcal{M}$-OPEN world is considered the currently proposed Bayesian minimum divergence posteriors fail to fully comply with the principled justification and motivation required to produce a coherent updating of beliefs. A Bayesian cannot therefore make principled inference using currently proposed methods in the $\mathcal{M}$-OPEN setting, except in a way that Miller and Dunson [2018] describe as "tend(ing) to be either limited in scope, computationally prohibitive, or lacking a clear justification". In order to make principled inference it appears as though the DM must currently concern themselves with minimising the KLD.

However in this chapter we remove this reliance upon the KLD by providing a justification for Bayesian updating minimising alternative divergences, both theoretically and ideologically. Our updating of beliefs does not produce an approximate or pseudo posterior, but uses general Bayesian updating [Bissiri et al., 2016] to produce the coherent posterior beliefs of a decision maker who wishes to produce predictions from a model that provide an explanation of the data that is as good as possible in terms of some pre-specified divergence measure. By doing this the principled statistical practice of fitting model parameters to produce predictions is adhered to, but the parameter fitting is done so acknowledging the $\mathcal{M}$-OPEN nature of the problem.

Another principled alternative to traditional Bayesian updating when it is difficult to fully specify a model for the DGP is Bayes linear methods [Goldstein, 1999]. These only require the subjective specifications of expectations and covariances for various quantities the DM is well informed about and interested in in order to do inference. Our alternative approach - which provides outputs more familiar to the typical Bayesian - instead chooses a more robust divergence. Using this divergence instead of the KLD provides a different approach which updates beliefs whilst being robust to routine assumptions and so makes a full probability specification a much less strenuous task. Chapter 3 will aim to be more precise about this claim.

2.4 Existing Bayesian Minimum Divergence estimation

Here we review some of the literature associated with minimum divergence estimation. Minimum divergence estimation (MDE) considers making inference about the parameters θ of a parametric model $\{f(x;\theta) : \theta \in \Theta\}$ by minimising the divergence between $f(x;\theta)$ and the data generating distribution of the observed data. By its very nature MDE has in the past often been addressed from a frequentist standpoint. MDE can be split into two categories, those that conduct inference by minimising local proper scores and those that conduct inference by minimising a member of the class of disparities. Local proper scoring rules depend only on the likelihood of the observed data under the model and are designed to be minimised at the DM's true beliefs (the DGP). These include the Tsallis (or βD-loss) [Basu et al., 1998], the Hyvarinen loss [Hyvärinen, 2005] and more recently the γD-loss [Hung et al., 2018]. Dawid et al. [2016] provide general theory for proper scoring rule inference.

Alternatively, disparity based methods first build a non-parametric density estimate $g_n(x)$ of the DGP from the data (often a Kernel Density Estimate (KDE)), and then conduct inference by minimising the divergence between the model and the estimate $g_n(x)$. Beran [1977] proposes using the HD in order to discover robust parameter estimates. They observe that when the data is distributed according to one member of the model class, $g(x) = f(x;\theta_0)$, the HD between the model and a data generating process is approximately equal to the KLD between the model a data generating process, when n is large. Therefore the minimum HD estimate (MHDE) is asymptotically equivalent to the MLE, the estimate achieved by minimising the KLD. By the same argument Beran [1977] propose that the MHDE will also be asymptotically efficient. Additionally, Kuchibhotla and Basu [2015] considers the αD and the generalised-KLD. Results regarding the robustness and efficiencies of general discrepancy measures can be found in Lindsay [1994]; Basu and Lindsay [1994]; Kuchibhotla and Basu [2015, 2016]. Basu et al. [2011] provides a comprehensive review of minimum divergence estimation.

The frequentist literature in this area is vast. In this thesis we choose to focus on the Bayesian contributions. These have been limited but come first from Hooker and Vidyashankar [2014] and more recently from Ghosh and Basu [2016] and Ghosh and Basu [2017]. Authors Hooker and Vidyashankar [2014] use the approximate equality between the KLD and the HD to produce a more robust posterior distribution over the model parameters, a direct Bayesian analogue to the MHDE. Ghosh and Basu [2016] considers the Bayesian analogue to Basu et al. [1998] minimising the proper Tsallis score associated with the βD in order to produce a 'pseudo-posterior' that does not require a density estimate. Ghosh and Basu [2017] demonstrate exponential converge results illustrating that this posterior is asymptotically optimal in exactly the same exponential rate as the traditional Bayesian posterior when the model is correct [Barron, 1988].

2.4.1 Why the current justification is not enough

Bayesian methods are traditionally motivated by Bayes' rule, principled belief updating and conditional probability. Following this, both Hooker and Vidyashankar [2014] and Ghosh and Basu [2016] seek to justify their 'pseudo' or 'approximate' posteriors by asserting that they approximate what might be obtained using Bayes' rule, but are more robust to the existence of outliers in finite samples.

The asymptotic approximation of the HD and the KLD when the data comes from the model is used by Hooker and Vidyashankar [2014] to justify replacing the KLD by the HD to produce posteriors of the form

$$\pi(\theta|\mathbf{x}) \propto \pi(\theta)\exp(-n\mathrm{HD}(g_n, f(\cdot;\theta))), \tag{2.2}$$

where g_n is an estimate of the data generating density g. Firstly define the KLD minimising parameter as

$$\hat{\theta}^{\mathrm{KLD}} = \arg\min_{\theta\in\Theta} \mathrm{KLD}(g_n(x), f(\cdot;\theta)). \tag{2.3}$$

This asymptotic equivalence between the KLD and HD can be seen by taking Taylor expansions of the two divergences about the KLD minimising parameter. For the KLD we then have that

$$\mathrm{KLD}(g_n(x)||f(x;\theta^{'})) = A_{\mathrm{KLD}} - B_{\mathrm{KLD}} - C_{\mathrm{KLD}} + \cdots, \tag{2.4}$$

where

$$A_{\text{KLD}} := \int g_n(x) \log \left(\frac{g_n(x)}{f(x; \hat{\theta}^{\text{KLD}})} \right) dx, \quad B_{\text{KLD}} := (\theta^{'} - \hat{\theta}^{\text{KLD}}) \int \frac{\nabla_\theta f(x; \hat{\theta}^{\text{KLD}})}{f(x; \hat{\theta}^{\text{KLD}})} g_n(x) dx$$
$$C_{\text{KLD}} := \frac{(\theta^{'} - \hat{\theta}^{\text{KLD}})^2}{2} \int \left(\frac{\nabla_\theta^2 f(x; \hat{\theta}^{\text{KLD}})}{f(x; \hat{\theta}^{\text{KLD}})} - \frac{(\nabla_\theta f(x; \hat{\theta}^{\text{KLD}}))^2}{f(x; \hat{\theta}^{\text{KLD}})^2} \right) g_n(x) dx.$$

While, for the HD we have that

$$4\text{HD}(g_n(x), f(x; \theta^{'})) = A_{\text{HD}} - B_{\text{HD}} - C_{\text{HD}} + \cdots, \tag{2.5}$$

where

$$A_{\text{HD}} := 4 \int \left(1 - \frac{\sqrt{f(x; \hat{\theta}^{\text{KLD}})}}{\sqrt{g_n(x)}} \right) g_n(x) dx$$
$$B_{\text{HD}} := (\theta^{'} - \hat{\theta}^{\text{KLD}}) \int 2 \frac{\nabla_\theta f(x; \hat{\theta}^{\text{KLD}})}{\sqrt{f(x; \hat{\theta}^{\text{KLD}})} \sqrt{g_n(x)}} g_n(x) dx$$
$$C_{\text{HD}} := \frac{(\theta^{'} - \hat{\theta}^{\text{KLD}})^2}{2} \int \left(2 \frac{\nabla_\theta^2 f(x; \hat{\theta}^{\text{KLD}})}{\sqrt{f(x; \hat{\theta}^{\text{KLD}})} \sqrt{g_n(x)}} - \frac{(\nabla_\theta f(x; \hat{\theta}^{\text{KLD}}))^2}{\sqrt{g_n(x)} (f(x; \hat{\theta}^{\text{KLD}}))^{3/2}} \right) g_n(x) dx.$$

Now following the same arguments as used in Hooker and Vidyashankar [2014]: The terms A_{KLD} and A_{HD} do not depend on $\theta^{'}$ and the terms B_{KLD} and B_{HD} must be 0 as by definition $\nabla_\theta f(x; \hat{\theta}^{\text{KLD}}) = \nabla_\theta \log f(x; \hat{\theta}^{\text{KLD}}) = 0$. Now if $g_n(x)$ is consistent for $g = f(\cdot; \hat{\theta}^{\text{KLD}}) = f(\cdot; \theta_0)$ then the left hand term in the integrals of C_{KLD} and C_{HD} Eq. (2.4) and (2.5) will be 0, and the right hand terms are equivalent in the limit as $n \to \infty$. This second order equivalence was noted by Beran [1977] as demonstrating the frequentist efficiency of a minimum HD estimate.

However when $\hat{\theta}^{\text{KLD}} \neq \theta_0$ is not the data generating parameter because the model class is misspecified, and if $g_n(x)$ is still converging to $g(x) \neq f(x; \theta_0)$ as $n \to \infty$ then terms C_{KLD} and C_{HD} in Eq. (2.4) and (2.5) will be different. Specifically, $\sqrt{f(\theta; \hat{\theta}^{\text{KLD}})} \sqrt{g_n(x)}$ will no longer converge to $g_n(x)$ as $n \to \infty$. In this setting the current literature gives no foundational reasoning why updating using the Hellinger divergence constitutes a principled updating of beliefs.

The 'pseudo-posterior' of Ghosh and Basu [2016], obtained instead by substituting the βD into Eq. (2.2), is justified either using the exponential converge results, which similar to above relies on the $\mathcal{M}$-CLOSED assumption that the model is correctly specified, or as the traditional posterior originating from an alternative belief model $\tilde{f}(x;\theta) \propto \exp\left(nD_B^{(\beta)}(g||f(\cdot;\theta))\right)$. However, this alternative model will not even correspond to the DM's approximate beliefs about the data generating process and in fact may well not even be normalisable. There is therefore, a lack of formal justification for a DM to update beliefs using Bayes rule on this object.

In the next sections we seek to unify these approaches under a principled and coherent updating regime, that does not merely seek to approximate some Bayes' rule inference but is principled and desirable when acknowledging the $\mathcal{M}$-OPEN world.

2.5 Principled Bayesian minimum divergence estimation

We next take a foundational approach to theoretically justify an updating of beliefs targeting the parameters minimising any statistical divergence between the model and the data generating density. Consider the general inference problem of wanting to estimate the parameters θ of the parametric model $\{f(\cdot;\theta) : \theta \in \Theta\}$. The model here can be considered as the DM's best guess at the DGP. We consider it important to continue to use a model even after the acknowledging that it is inevitably misspecified for the reasons outlined in Section 2.2. Following Bernardo and Smith [2001] and Walker [2013] we consider fitting these parameters in a decision theoretic manner, by minimising some divergence function. We specifically restrict ourselves to divergences that can be written in terms of expected losses as in Eq. (1.31). These divergence functions naturally admit a loss function that can be used to quantify how well a model fits data (see Section 1.3.2). The corresponding inferential goal is to solve the decision problem

$$\theta^* = \arg\min_{\theta} D(g(\cdot)||f(\cdot;\theta)) = \arg\min_{\theta} \int \ell_D(x, f(\cdot;\theta))dG(x). \tag{2.6}$$

Here the entropy term in the definition of the divergence is removed from the minimisation because it does not depend on θ. In the formulation of Walker [2013] this divergence was the KLD in order to justify the continued use of Bayes' rule under model misspecification. However, such a justification and the continued use of Bayes' rule forces the Bayesian into concerns associated with non-robustness.

Alternatively, demanding a solution to equation (2.6) is analogous to the approach taken by Bissiri et al. [2016], outlined in Chapter 1, when producing their general Bayesian update. The only difference now is that the loss function depends on the parameter through a model so is a scoring rule. By minimising a combination of the expected loss of the data under the posterior and the KLD between the prior and the posterior, Bissiri et al. [2016] come up with an optimal updating of beliefs for any decision problem. As a result general Bayesian updating provides a coherent updating of beliefs that target the parameters minimising the general divergence $D(\cdot||\cdot)$ as

$$\pi^{(D)}(\theta|\mathbf{x}) = \underset{q\in\mathcal{P}}{\arg\min}\left\{\mathbb{E}_{\theta\sim q(\theta)}\left[w\sum_{i=1}^{n}\ell_D(x_i, f(\cdot;\theta))\right] + \textsc{kld}(q(\theta)||\pi^{(D)}(\theta))\right\} \tag{2.7}$$

$$\propto \pi^{(D)}(\theta)\exp(-w\sum_{i=1}^{n}\ell_D(x_i, f(\cdot;\theta))). \tag{2.8}$$

The calibration weight w [Bissiri et al., 2016] is discussed further in Section 2.5.1. In order to stay consistent with Walker [2013], we use the notation $\pi^{(D)}$ to indicate that the prior and posterior belief distributions correspond to beliefs about the parameter minimising divergence $D(\cdot||\cdot)$. An example of how this distinction might manifest itself is as follows: minimising the KLD compared with one of the more robust divergences we consider later makes observing a large variance more likely as a small proportion of observations can drastically inflate the predictive variance, e.g. see Fig. 2.1. Thus the corresponding prior should have a longer right tail.

Directly eliciting beliefs according to the geometries of different divergences will be very difficult, even for those with a high level of mathematical training. However, this distinction is necessary in the M-OPEN world as the DM can not possibly express beliefs about the data-generating parameter as they know this does not exist. The chosen divergences relates the parameters of the model to the observations and thus allows the DM to think about their parameter in terms of observables for which they can conceivably have beliefs about [Gelman et al., 2017; Goldstein and Wooff, 2007; Williamson et al., 2015]. In practise however, the DM is likely to do so only using only some vague, high-level notion of closeness between the model and the DGP, rather than the explicit form of some divergence. As a result, we argue it may actually be easier to consider beliefs about the parameters minimising some of the more robust divergences introduced below than the KLD consider by Bayes' rule updating under missepcification which defines close by accuracy of tail specification. This thesis generally focuses on likelihoods rather than priors and as a result we do not consider this point further. Though whether there are divergences that simplify prior elicitation is an interesting areas for future research.

Additionally, Eq. (2.7) illustrates that although changing the loss functions allows these generalised minimum divergence posteriors to learn about the parameters of the model minimising a divergence different from the KLD to the DGP, they are still derived by also seeking to minimise the KLD of the prior from the posterior[Bissiri et al., 2016]. This ensures the posterior in Eq. (2.8) is still coherent. In Chapter 4 we go further and explore changing this divergence also.

Applying the general Bayesian update in an inferential scenario like this, provides a compromise between purely loss based general Bayesian inference and traditional Bayesian updating using Bayes' rule. General Bayesian updating as presented by Bissiri et al. [2016] produces posterior beliefs directly about the minimiser of the loss function connecting the decision to the data. In this case θ is not the parameter indexing a likelihood but some value directly relevant to the decision problem at hand. On the other hand, traditional Bayesian updating is undertaking an inference problem completely independently to any actual decision problem. Bayes' rule simply aims to produce the best characterisation of the DGP, where best is canonically defined in terms of the KLD.

Our approach introduced here is still concerned with estimating model parameters based on how well the model's predictions reflect the DGP. However, the criteria for closeness between predictions and reality can be chosen with the knowledge that the resulting inferences will be used to inform a decision problem. As a result, the minimum divergence posteriors can be seen as producing inferences that will be suitable for a broad range of loss functions. Unlike in the original general Bayesian update, the DM is no longer required to precisely define the loss function associated with their decision problem at the inference stage. They need only consider broadly how robust to tail misspecifications they want their inference to be in order to define the target divergence (see Section 2.6.7). The general Bayesian updating proposed above allows the goals of the statistical analysis to be coupled together with the parameter updating, something not previously possible under traditional Bayesian statistics.

Considering the realistic $\mathcal{M}$-OPEN nature of the model class provides a further justification for a middle ground between purely loss-based general Bayes and Bayes' rule. The original general Bayesian update, assumes absolutely no information about the DGP when simply using a loss function to produce posterior beliefs. In contrast using Bayes rule traditionally[1] assumes the DGP is known precisely. As we point out, it is actually more likely that the decision maker is able to express informative but not exact beliefs about the DGP and therefore a half-way-house between these two is appropriate in reality.

The loss function associated with minimising the βD is given in Eq. (1.33)

[1] Prior to the interpretation provided by Walker [2013] explained in section 1.1.3

while the loss function associated with the HD is found by observing that

$$\begin{aligned}\text{HD}(g, f(\cdot;\theta)) &= 1 - \int \sqrt{f(x;\theta)}\sqrt{g(x)}dx &(2.9)\\ &= \mathbb{E}_g\left[\ell_H(x, f(\cdot))\right] - \mathbb{E}_g\left[\ell_H(x, g(\cdot;\theta))\right] &(2.10)\end{aligned}$$

where

$$\ell_H(x, f(\cdot;\theta)) = -\frac{\sqrt{f(x;\theta)}}{\sqrt{g(x)}} \qquad (2.11)$$

Plugging Eq. (2.11) and (1.33) into Eq. (2.8) produce the general Bayesian posteriors minimising the HD and βD as

$$\begin{aligned}\pi^H(\theta|\mathbf{x}) &\propto \pi^H(\theta)\exp\left(\sum_{i=1}^n \frac{\sqrt{f(x_i;\theta)}}{\sqrt{g_n(x_i)}}\right) &(2.12)\\ \pi_B^{(\beta)}(\theta|\mathbf{x}) &\propto \pi_B^{(\beta)}(\theta)\exp\left(\sum_{i=1}^n\left\{\frac{1}{\beta-1}f(x_i;\theta)^{\beta-1} - \frac{1}{\beta}\int f(y;\theta)^\beta dy\right\}\right). &(2.13)\end{aligned}$$

Equation (2.12) introduces $g_n(\cdot)$ to estimate the data generating density g (see Section 2.6.8 for more on this). As a result, the general Bayesian updating is being conducted using an empirical loss function,

$$\hat{\ell}_H(x, f(\cdot;\theta)) = -\frac{\sqrt{f(x;\theta)}}{\sqrt{g_n(x)}}, \qquad (2.14)$$

which approximates the true loss function required to minimise the Hellinger divergence between the model and the data generating process ℓ_H in Eq. (2.11).

Equation (2.13) is exactly the distribution resulting from the robust parameter update of Ghosh and Basu [2016] (with their $\alpha = \beta - 1$), while equation (2.12) is similar to the posterior produced by Hooker and Vidyashankar [2014] except the divergence function has been decomposed into its score and entropy term here. This demonstrates that the posteriors above are not pseudo posteriors - as Ghosh and Basu [2016] suggests - or approximations of posteriors - as Hooker and Vidyashankar [2014] suggests - they provide a principled method for the DM to update their prior beliefs about the parameter minimising an alternative divergence to the KLD.

As well as being philosophically appealing, we can be reassured that these generalised model posteriors posses similarly convenient mathematical properties to the traditional Bayes' rule posterior. Chernozhukov and Hong [2003]; Lyddon et al. [2018] provide regularity conditions under which the general Bayesian posterior (Eq.

(1.16)) for general loss function $\ell(x,\theta)$ is asymptotically normal. They show that

$$\sqrt{n}\left(\theta-\hat{\theta}_n\right)\rightarrow z^{'}, \tag{2.15}$$

where convergence is in distribution, $z^{'}\sim\mathcal{N}\left(0,\frac{1}{w}J\left(\theta^*\right)^{-1}\right)$ and

$$\theta^*=\underset{\theta\in\Theta}{\arg\min}\int\ell(\theta,x)dG(x),\qquad \hat{\theta}_n=\underset{\theta\in\Theta}{\arg\min}\frac{1}{n}\sum_{i=1}^{n}\ell(\theta,x_i) \tag{2.16}$$

$$J(\theta)=\int\nabla_\theta^2\ell(\theta,x)dG(x),\qquad \hat{J}(\theta)=\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta^2\ell(\theta,x_i), \tag{2.17}$$

where $G(\cdot)$ is the data generating distribution and ∇_θ^2 is the second derivative with respect to θ. This agrees with the posterior asymptotic normality results presented in Hooker and Vidyashankar [2014] (who restrict $g=f_{\theta_0}$) and Ghosh and Basu [2016] for the HD and βD loss functions respectively, who also note that the result still holds when $J\left(\theta^*\right)$ is replaced by the empirical $\hat{J}\left(\hat{\theta}_n\right)$. We note that the result for the HD requires some conditions on the density estimate used. Additionally Basu et al. [2011] prove consistency of the empirical divergence minimiser $\hat{\theta}_n$ to the exact divergence minimiser to θ^*.

One convenient consequence of these results is that if by chance we are in the $\mathcal{M}$-CLOSED world, i.e. there exists θ_0 such that $g(x)=f(x;\theta_0)$, then property ii) in Definition 5 of a statistical divergence ensures that $D(q||f(\cdot;\theta))$ must be uniquely minimised at $g=f(\cdot;\theta_0)$. As a result, in this case the general Bayesian posterior (Eq. (2.8)) will still be learning about the data generating parameter θ_0. We examine the frequentist efficiency in this scenario in Section 2.7.4.

2.5.1 Calibration

The two posteriors in equations (2.12) and (2.13) can be seen as using the canonical form of the HD and βD loss function and setting the general Bayesian calibration weight $w=1$. Unlike probability distributions, loss functions are invariant to a linear change in scale. Increasing the weight on the loss function, $w\ell_D(\theta,x)$, does not change the loss minimising parameter but will cause the posterior in Eq. (2.8) to concentrate further and receive less influence from the prior. The converse happens when decreasing w. As a result w needs to be set to calibrate the loss function.

One solution to this is to argue that $w=1$ is correct here. The posteriors above both use well-defined models and the canonical form of well-defined divergence functions and as a result there is no arbitrariness in the size of the loss.

This is further demonstrated by the fact that Bayes' rule corresponds to using the canonical form of the KLD and a probability model with weight $w = 1$. When the model is correct, this causes the posterior variance of these alternative methods to be increased for finite data samples relative to the traditional Bayesian posterior. This is to be expected. Zellner [1988] showed that Bayes' rule processes information optimally when the model is correctly specified and therefore produces the most precise posterior distributions. The posterior is simply a subjective reflection of the DM's uncertainty after seeing the data, and if they believe their model is incorrect and therefore target a more robust divergence, they are likely to have greater posterior uncertainty than if they naively believe their model to be correct and proceed accordingly.

This argument however relies on the existence of canonical forms of divergences. In practice these do not exist. For example the HD has a multiplying constant of 1 when viewed in isolation but has $\frac{1}{0.5(1-0.5)} = 4$ when viewed as a member of the αD family with $\alpha = 0.5$.

The loss-likelihood bootstrap (LLB) method of Lyddon et al. [2018] described in Section 1.2.2 provides a more principled procedure to set w. This matches the asymptotic information in the general Bayesian posterior for a given loss function with the asymptotic information in a sample from the LLB. The LLB is interpreted by Lyddon et al. [2018] as producing a non-parametric, Dirichlet process estimate of the DGP. This can then be repeatedly sampled from, and these samples used to minimise the loss function. This produces a sample of parameter values quantifying uncertainty about the minimiser of the loss. As a result, the LLB is motivated by Lyddon et al. [2018] as a principled quantification of uncertainty in the absence of a model for the DGP. As the only operation involving the loss is minimisation, the LLB is invariant to a scalar weighting of the loss unlike the general Bayesian posterior. However, the general Bayesian posterior depends on a prior distribution where the LLB does not. As a result, w is set to match the asymptotic information in both methods, where the prior has little influence, but the general Bayesian posterior is preferred for finite n. For the next two chapters we use this method of Lyddon et al. [2018] to set w for the minimum divergence posterior.

However, we note that the approach of Lyddon et al. [2018] is designed for general loss functions rather than model based divergence loss functions. In the presence of a model, one desirable property for minimum divergence Bayesian updating might be that if the model is correct then the posterior agrees with the posterior obtained using Bayes' rule. In fact this is similar to the justification provided for the update of Hooker and Vidyashankar [2014]. This can be shown not to be the case when using the LLB and therefore provides an area of further research for tailoring the selection of w to loss functions depending upon a likelihood model.

2.5.2 The likelihood principle and Bayesian additivity

Authors Hooker and Vidyashankar [2014] identify that their posterior distribution minimising the Hellinger divergence, Eq. (2.12), no longer satisfies Bayesian additivity. That is to say that the posteriors would look different if the sample was used as a whole to update the posterior, or if the updating happens incrementally on subsets of the observations. This is because the posterior now depends on the estimate of the data generating density, which depends in a non-additive way on the data. This constitutes a departure from the likelihood principle underpinning traditional Bayesian statistics. The likelihood principle says that the likelihood is sufficient for the data. When the model is perfectly specified this is a sensible principle, the likelihood of the observed data under the correct model represents all the information in the data. However, in the M-open world the likelihood principle is no longer a reasonable requirement. When the model is only considered as an approximation of the DGP, it is reasonable to suspect that the data contains more (or less) information than is represented in the likelihood of an incorrect model. When the model is correct we can reasonably expect a sample from the fitted model to resemble the observed data. When the model is misspecified this is no longer the case.

Bayesian additivity (defined as coherence by Bissiri et al. [2016]) is a central principle of the general Bayesian update. It was the demand for coherence that motivated the KLD as the divergence used to penalise the posterior to prior divergence. Which in turn, elicited the form of the general Bayesian update Eq. (2.8). We argue such a demand does not prohibit the use of the HD loss function. The general Bayesian posterior minimising the Hellinger divergence, Eq. (2.12), combines the loss function for each observation in an additive way, which is consistent with the additivity demanded by Bissiri et al. [2016], but a different density estimate is produced when the data is considered as a whole or in parts. This causes the empirical loss function, $\hat{\ell}_H(x, f(\cdot;\theta))$ in Eq. (2.14), used for the updating to be different when the data arrives incrementally, as opposed to one go. If the data generating density, $g(x)$, were available then the exact loss function associated with minimising the Hellinger divergence, $\hat{\ell}_H(x, f(\cdot;\theta))$ in Eq. (2.11), could be calculated and the Bayesian update would be additive. However because an approximation of the loss function associated with the Hellinger divergence is used, additivity must be sacrificed.

In contrast the posterior of Ghosh and Basu [2016], Eq. (2.13), does not require a density estimate and is thus able to produce an additive belief update. However, the likelihood is also no longer considered to be sufficient for the data. The posterior also depends on the integral, $\frac{1}{\beta}\int f(z;\theta)^{\beta}dz$.

When using an alternative divergence to the KLD in order to conduct a Bayesian updating, it is necessary for additional information to the 'local' information provided by the likelihood of the observed data to be incorporated into the loss function. For the HD this information comes from the data by way of a density estimate, while for the βD this comes from the model through $\frac{1}{\beta}\int f(z;\theta)^{\beta}dz$.

2.5.3 Subjectivity

Demonstrating that principled inference can be made using alternative divergence measures than the KLD, enables us to consider the selection of the divergence used for updating to be a subjective judgement made by the DM, alongside the prior and model, to help tailor the inference to the specific problem. Celeux et al. [2017] observed that while Gelman and Hennig [2015] advocate greater freedom for subjective judgements to impact statistical methodology, they fail to consider the possibility of subjective Bayesian parameter updating. In the utopic $\mathcal{M}$-CLOSED world Bayes' rule and the log-score provide objectively the correct way to update probability beliefs, given that the DM believes all of the judgements made by their prior at the time the updating takes place. This thesis in general focuses on the likelihood aspect of the Bayesian updating and thus we make the very liberal assumption of a correctly specified prior for simplicity. In the more realistic $\mathcal{M}$-OPEN world we have argued above that this is no longer so clear. Very few problems seek answers that are connected with a specific dataset or model, they seek answers about the real world process underpinning these. Authors Goldstein et al. [2006], focusing on belief statements, demonstrate that subjective judgements help to generalise conclusions from the model and the data to the real world process. In a similar spirit we argue that carefully selecting an appropriate divergence measure and documenting the reasons for doing this provides a further way of introducing subjective judgement into a statistical analysis.

2.5.4 A decision theoretic view of probabilities

The minimum divergence general Bayesian updating introduced in this section enables us to produce a completely different interpretation for the likelihood in a Bayesian analysis. Under Bayes' rule the likelihood is the probability of observing the data under certain parametric conditions. This requires the DM who elicits this likelihood to consider every aspect of the world in which the data was generated, making many more judgements than they can possibly be expected to have the introspection to make [Goldstein, 1990]. We discussed how this inevitably leads to misspecifications in Section 1.1.2.

In contrast, in this general Bayesian minimum divergence paradigm the likelihood no longer need be considered in such a strict generative sense. The DM almost certainly does not believe the data was generated from their model. Instead, the likelihood is combined with a loss function allowing us to interpret the likelihood predictively. Rather than considering the probability of generating the observed data under given parametric conditions, we can asses the loss/reward the DM would have received had they used the same parametric conditions to produce a predictive distribution for the observed the data.

Interpreting the likelihood predictively using a loss function then allows for much greater flexibility and subjectivity when conducting inference. Not only can the loss function be changed to allow the DM greater control over how their inferences are designed to fit the data, e.g. favouring correctly modelling most observations and ignoring the fact that an outlier was poorly modelled, but the DM can also now only consider scoring predictions on margins of the observation space they are interested in. Rather than producing a joint likelihood for every dimension of the observed data, the DM can now view their likelihood predictively and just score it on the margins that are important for them to predict. This constitutes an important area of further research, possibly providing a powerful dimension reduction tool to help robustify high-dimensional analyses.

2.5.5 Posterior prediction and the model

Often when performing parametric model inference the values of the parameters are not necessarily of interest to the DM. These are often just artificial constructs to separate past data from future data. In these situations the DM is usually interested in the distribution of future observations given the past data. Given the acknowledgement that the model is misspecified it is not exactly clear what form this distribution should take. One candidate that we consider in this thesis for the general Bayesian minimum divergence paradigm is

$$f(y|\boldsymbol{x}) = \int f(y;\theta)\pi^{(D)}(\theta|\boldsymbol{x})d\theta. \tag{2.18}$$

When $D =$ KLD this is the familiar posterior predictive. However, even when using a different divergence, we still use the model likelihood to produce predictions. This is justified following the "all models are wrong but some are useful" assumption discussed in Section 2.2. Although the model is not a correct representation of our beliefs, we continue to use the model in the belief that it captures some important structural components of the DM's beliefs that they surely want to impact both updating and future prediction. However, unlike in the M-CLOSED world the DM is not compelled by the rules of conditional probability to use this. Seeing the data for example could well provide information that would alter the form for which the DM might want to predict and thus the proposal above is considered a default plug-in with more careful thought about the distribution possibly required.

However, we explicitly do not attempt to interpret the minimum divergence general Bayesian posteriors, Eq. (2.8), as conducting Bayes' rule updating on a robustified belief model $\tilde{f}(x;\theta) \propto \exp\left(-\ell_D(x, f(\cdot;\theta))\right)$ [Ghosh and Basu, 2017]. If this were interpreted as an alternative belief model then predictions should be produced using the alternative belief model $\tilde{f}(y;\theta)$, which in many cases will not correspond to a normalised probability density, rather than the original model $f(y;\theta)$.

The separation of loss function used for inference and the model used for prediction in Eq. (2.18) allows us to produce inference that is insensitive to outliers without having a model which generates outliers [O'Hagan, 1979].

2.6 Possible Divergences to consider

Section 1.3 introduced several well-known families of divergences and their corresponding loss function interpretations, where available. Here we motivate some of these for use in inference. We do not claim this list is exhaustive but merely contains ones we have considered and experimented with for inference. Throughout, we seek to motivate these in a decision theoretic sense based on their ability to produce accurate estimates of expected utility.

2.6.1 Total-Variation Divergence (TVD)

General Bayesian updating targeting the minimisation of the TVD can be produced using loss function

$$\ell_{TV}\left(x, f(\cdot;\theta)\right) = \frac{1}{2}\left|1 - \frac{f(x;\theta)}{g_n(x)}\right|. \tag{2.19}$$

Notice, similarly to Eq. (2.14), Eq. (2.19) also requires an estimate of the data generating density g_n and is thus an empirical approximation to the loss function minimising the TVD. Our principle of $\mathcal{M}$-OPEN inference suggests that under a bounded utility function TVD provides the canonical criteria for inference, see Section 2.3.

Unlike the log-score, Eq. (2.19) is bounded as $f(x;\theta) \to 0$. Authors Hooker and Vidyashankar [2014] identify some drawbacks of having a bounded score function. The score function being upper bounded means that there is some limit to the score that can be incurred in the tails of the posterior distribution. Under the log-score, as $f(x;\theta) \to 0$, $-\log(f(x;\theta)) \to \infty$ and as a result in this region

$$\pi(\theta|x) \propto \pi(\theta)\exp\left(-\log(f(x;\theta))\right) \to 0. \tag{2.20}$$

Now under a bounded loss function we have that for fixed $g_n(x)$, $f(x;\theta) \to 0$ $\ell(x,\theta) \to B < \infty$ and as a result in this region

$$\pi(\theta|x) \propto \pi(\theta)\exp\left(-w\ell(x,\theta)\right) \to \pi(\theta). \tag{2.21}$$

Therefore the tails of the posterior will be equivalent to the tails of the prior. As a result the DM is required to think more carefully about their prior distribution. In fact, when changing the divergence targeted by inference the DM actually has a completely different prior to construct, see the discussion in Section 2.5 Not only are improper priors prohibited, but more data is required to move away from a poorly specified prior. This can be somewhat mitigated by an appropriately chosen calibration weight, see Tables 2.5 and 2.6. However the loss function is not monotonic or strictly convex which we believe harms the finite sample efficiency when the DGP is within the chosen model class, see Table 2.5 and discussion.

2.6.2 Hellinger Divergence (HD)

As was mention in Section 1.3, Devroye and Gyorfi [1985]; Liese and Vajda [1987] observed, that the HD can be used to bound the TVD both above and below. As a result, the HD and the TVD are geometrically equivalent. Thus, if one of them is small, the other is small and similarly if one of them is large the other is also. So if one distribution is close to another in terms of TVD, then the two distributions will be close in terms of HD as well. Authors Beran [1977] first noted that minimising the HD gave a robust alternative to minimising the KLD, while Hooker and Vidyashankar [2014] proposed a Bayesian alternative (Eq. (2.12)) that has been discussed at length above. While Hooker and Vidyashankar [2014] motivated their posterior through asymptotic approximations, identifying the geometric equivalence between the HD and TVD proposes further justification for a robust Bayesian updating of beliefs similar to that of Hooker and Vidyashankar [2014]. Specifically, if being close in terms of TVD is the ultimate robust goal, then being close in HD will guarantee closeness in TVD. The HD can therefore serve as a proxy for TVD that retains some desirable properties of the KLD: it is possible to compute the HD between many known families [Smith, 1995] and the score associated with the HD has a similar strictly convex shape, this is discussed further in Section 2.7.4. A posterior targeting the HD does suffer from the same drawbacks associated with having a bounded scoring function that are mentioned at the end of the previous section and requires an estimate of the data generating density. Lastly along with TVD, the HD is also a metric.

2.6.3 α-Divergence (αD)

The HD and KLD can be smoothed between by αD. Setting $\alpha = 1$ corresponds to a KLD limiting case (see Definition 9, Remark 1) while subbing $\alpha = 0.5$ into Eq. (2.22) recovers 4 times the HD in Eq. (2.11). We think of these as two extremes of efficiency and robustness within the αD family with which a DM would want to conduct inference between. The parameter α thus controls this trade-off. The loss function used to target minimising the αD is

$$\ell_A^{(\alpha)}(x, f(\cdot;\theta)) = \frac{1}{\alpha(1-\alpha)} g_n(x)^{\alpha-1} f(x;\theta)^{1-\alpha}. \tag{2.22}$$

It was demonstrated in [Prop. 2.35 Liese and Vajda, 1987],[Cor. 1 Sason and Verdu, 2016] that for $\alpha \in (0,1)$ the αD can be bounded above by TVD, see Eq. (1.37). Therefore, similarly to the HD if the TVD is small then the αD will be small (provided $\alpha \neq \{0,1\}$). So a predictive distribution that is close to the data generating density in terms of TVD will receive high posterior mass under an update targeting the αD. Cichocki et al. [2011] provide intuition about how the parameter α impact the influence each observation x has on the inference about $f(\cdot;\theta)$ when the data comes from $g(x)$. They consider influence to be how the observations impact the estimating equation of θ, this is a frequentist setting but the intuition remains useful here.

$$\begin{cases} \alpha > 1, \text{ down-weights } x \text{ with small } g(x)/f(x;\theta). \\ \alpha < 1, \text{ down-weights } x \text{ with large } g(x)/f(x;\theta). \end{cases} \tag{2.23}$$

The size of $g(x)/f(x;\theta)$ for an observation x defines how outlying (large values) or inlying (small values) the observation is. Choosing $\alpha < 1$ ensures outliers relative to the model have little influence on the inference. We can project these notions of influence into the Bayesian paradigm using the influence plots described in Section 2.6.7 (see Figure 2.2). Once again the loss function is bounded in θ for $\alpha < 1$ and an estimate of the data generating density is required to operationalise this loss function.

2.6.4 β-Divergence (βD)

Another one parameter divergence family containing the KLD is the βD family. Eq. 2.13 above presents the posterior targeting minimising the βD. Both Basu et al. [1998] and Dawid et al. [2016] noticed that inference can be made using the βD without requiring a density estimate. This was used in Ghosh and Basu [2016] to produce a robust posterior distribution that did not require an estimate of the data generating density, which has been extensively discussed in previous sections. Cichocki et al. [2011] again uses estimating equations to asses the impact observations can have on the βD.

$$\begin{cases} \beta > 1, \text{ down weights } x \text{ where } f(x;\theta) \text{ is small.} \\ \beta < 1, \text{ down weights } x \text{ where } f(x;\theta) \text{ is large.} \end{cases} \tag{2.24}$$

Taking $\beta > 1$ results in the influence of observations that have low predicted probability $f(x;\theta)$ under the model being down-weighted. When minimising the KLD at $\beta = 1$, as is done by Bayes' rule, the influence of an observation x is inversely related to its probability under the model. Raising β above 1 will decrease the influence of the smaller values of $f(x;\theta)$, robustifying the inference to tail specification. However this results in a decrease in efficiency relative to methods minimising the KLD [Ghosh and Basu, 2017]. We provide an illustration of this phenomenon suitable for the Bayesian paradigm via the influence curves in Figure 2.2.

Next we provide some motivation for learning using the βD in a decision making setting by showing that under certain conditions the βD can be bounded above by the TVD.

Bounding the βD using the TVD

We are not aware of any previous results relating Bregman divergences to the TVD. Below we prove that the TVD can be used to form an upper bound on the βD under certain conditions. We make no assertions that this bound is at all tight but argue it allows us to identify when closeness in terms of TVD will guarantee closeness in the βD. The following theorem requires that we can bound the essential supremum (ess sup) of densities g and f. Given base measure[2] μ, M is the essential supremum of density $f(x)$, $\operatorname{ess\,sup} f(x) = M$, if the set defined by $f^{-1}((M, \infty))$, with (M, ∞) the open interval between M and infinity, has measure 0, i.e. $\mu\left(f^{-1}((M, \infty))\right) = 0$.

Theorem 2. Suppose the densities g and f satisfy the following, $\operatorname{ess\,sup} g = M_g$, $\operatorname{ess\,sup} f = M_f$, and that there exists $M \geq \max(M_g, M_f)$ then for $1 < \beta < 2$ we have that

$$D_B^{(\beta)}(g||f) \leq \left(\frac{M^{\beta-1}}{\beta-1}\right) \text{TVD}(g, f). \tag{2.25}$$

Proof. Firstly this proof makes use of the following identity for the TVD. Defining $A^- = \{x : g(x) < f(x)\}$ and $A^+ = \{x : g(x) \geq f(x)\}$ the TVD can be rewritten as

$$\text{TVD}(g, f) = \int_{A_-} (f(x) - g(x))\, dx = \int_{A_+} (g(x) - f(x))\, dx. \tag{2.26}$$

Now the βD can be rearranged to give

$$\begin{aligned}
& (\beta - 1)\, \beta D_B^{(\beta)}(g||f) && (2.27)\\
= \;& (\beta - 1) \int f(x)^{\beta-1} (f(x) - g(x))\, dx + \int \left(g(x)^{\beta-1} - f(x)^{\beta-1}\right) g(x) dx \\
= \;& (\beta - 1) \int_{A_-} f(x)^{\beta-1} (f(x) - g(x))\, dx + \int_{A_-} \left(g(x)^{\beta-1} - f(x)^{\beta-1}\right) g(x) dx \\
& + (\beta - 1) \int_{A_+} f(x)^{\beta-1} (f(x) - g(x))\, dx + \int_{A_+} \left(g(x)^{\beta-1} - f(x)^{\beta-1}\right) g(x) dx \\
\leq \;& (\beta - 1) \int_{A_-} f(x)^{\beta-1} (f(x) - g(x))\, dx + \int_{A_+} \left(g(x)^{\beta-1} - f(x)^{\beta-1}\right) g(x) dx,
\end{aligned}$$

as on A_+ we know that $(f(x) - g(x)) < 0$ and on A_- $g(x) < f(x) \Rightarrow g^{\beta-1}(x) < f^{\beta-1}(x)$ for $1 \leq \beta \leq 2$. Since $f \leq M$,

$$\int_{A_-} f(x)^{\beta-1} (f(x) - g(x))\, dx \leq M^{\beta-1} \int_{A_-} (f(x) - g(x))\, dx \leq M^{\beta-1} \text{TVD}(f, g).$$

On A_+ we have that $g(x) > f(x)$ which implies that $\frac{f(x)}{g(x)} < 1$ and that when

[2] we discussed in Section 1.3.1 that this is assumed to be the Lebesgue measure for continuous random variables and the counting measures for discrete random variables.

$1 < \beta < 2$, $\frac{f(x)}{g(x)}^{\beta-1} > \frac{f(x)}{g(x)}$ so

$$\begin{aligned}
\int_{A_+}\left(g(x)^{\beta-1} - f(x)^{\beta-1}\right)g(x)dx &= \int_{A_+} g(x)^{\beta-1}\left(1 - \left(\frac{f(x)}{g(x)}\right)^{\beta-1}\right)g(x)dx \\
&\leq \int_{A_+} g(x)^{\beta-1}\left(1 - \left(\frac{f(x)}{g(x)}\right)\right)g(x)dx \qquad (2.28)\\
&= \int_{A_+} g(x)^{\beta-1}\left(g(x) - f(x)\right)dx \qquad (2.29)\\
&\leq M^{\beta-1}\text{TVD}(f,g), \qquad (2.30)
\end{aligned}$$

since $g \leq M$ we have that. Folding the two bounds together leaves

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

which rearranged proves the theorem. □

The implications of Theorem 2 are as follows. Provided $\left(\frac{M^{\beta-1}}{\beta-1}\right)$ does not get too small, we can be confident that the predictive distributions that are close to the data generating density in terms of TVD, and thus produce accurate estimates of bounded expected losses, will receive high posterior mass under an update targeting the βD. We discuss when $\left(\frac{M^{\beta-1}}{\beta-1}\right)$ might becomes too small in Section 2.7.5

2.6.5 γ-Divergence (γD)

Another divergence eliciting a loss function that does not require an estimate of the data generating density is the γD. Now, the γD as it is introduced in Eq. (1.30) does not naturally allow for the interpretation provided by Eq. (1.31)

$$\begin{aligned}
D_G^{(\gamma)}(g||f(\cdot;\theta)) &= \frac{1}{\gamma}\log\int f(x;\theta)^{\gamma}dx - \frac{1}{(\gamma-1)}\log\int f(x;\theta)^{\gamma-1}g(x)dx \\
&+\frac{1}{\gamma(\gamma-1)}\log\int g(x)^{\gamma}dx \\
&= \frac{1}{\gamma}\log\int f(x;\theta)^{\gamma}dx - \frac{1}{(\gamma-1)}\log\mathbb{E}_{g(x)}\left[f(x;\theta)^{\gamma-1}\right] \qquad (2.31)\\
&+\frac{1}{\gamma(\gamma-1)}\log\int g(x)^{\gamma}dx
\end{aligned}$$

However minimising the $D_G^{(\gamma)}$ for θ allows us to ignore the entropy term and is therefore equivalent to minimising

$$\begin{aligned}
&\frac{1}{\gamma}\log\int f(x;\theta)^\gamma dx - \frac{1}{(\gamma-1)}\log\mathbb{E}_{g(x)}\left[f(x;\theta)^{\gamma-1}\right]\\
&=\log\frac{\left(\int f(x;\theta)^\gamma dx\right)^{\frac{1}{\gamma}}}{\left(\mathbb{E}_{g(x)}\left[f(x;\theta)^{\gamma-1}\right]\right)^{\frac{1}{(\gamma-1)}}}
\end{aligned}\tag{2.32}$$

Viewing inference on θ purely as optimisation, one can apply any monotonic transform to a Eq. (2.32) without changing the location of its minimiser. The function $\gamma\exp(x)$ is monotonically increasing on $\mathbb{R}$ for $\gamma > 0$. Similarly, the function $h(x) = x^{1-\gamma}$ is monotonic on $\mathbb{R}^+$ and decreasing (increasing) for $\gamma > 1$ ($\gamma < 1$). The multiplier $\frac{1}{\gamma-1}$ is negative when $\gamma < 1$ and ensures that $h(x) = x^{1-\gamma}$ is increasing in x. As a result, (for $\gamma > 0$) minimising Eq. (2.32) is equivalent to minimising,

$$-\frac{\gamma}{\gamma-1}\frac{\mathbb{E}_{g(x)}\left[f(x;\theta)^{\gamma-1}\right]}{\left(\int f(x;\theta)^\gamma dx\right)^{\frac{\gamma-1}{\gamma}}} \approx -\frac{\gamma}{\gamma-1}\frac{1}{n}\sum_{i=1}^{n}\frac{f(x_i;\theta)^{\gamma-1}}{\left(\int f(z;\theta)^\gamma dz\right)^{\frac{\gamma-1}{\gamma}}}\tag{2.33}$$

with $x_1,\ldots,x_n \sim g(\cdot)$. Eq. (2.33) provides γ times the loss function of Futami et al. [2017]; Hung et al. [2018].

Substituting Eq. (2.33) into Eq. (1.31) results in an scalar multiple of the alternative definition of the γ-divergence used in Hung et al. [2018].

$$\frac{\gamma}{(\gamma-1)}\left\{I_{g,\gamma}(\theta)^{\frac{1}{\gamma}} - I_{f,\gamma}(\theta)^{\frac{1-\gamma}{\gamma}}\int g(x)f(x;\theta)^{\gamma-1}dx\right\},\tag{2.34}$$

where $I_{f,\gamma}(\boldsymbol{\theta}) = \int f(x;\boldsymbol{\theta})^\gamma dx$. This divergence appears to be different from $D_G^{(\gamma)}$ as defined by Cichocki and Amari [2010], but the above derivation shows that both versions will be minimised for the same value of $\boldsymbol{\theta}$.

The γD loss function in Eq. (2.33) is very similar to the βD loss used to produce the posterior in Eq. (2.13). Both raise the likelihood to the power greater than 0 and then correct for this using the integral of the likelihood to a similar power, the βD does this additively while the γD does this multiplicatively. The multiplicative nature of the γD score and the fact that it is always positive is appealing for computational reasons, see Section 4.8.2. As yet we have not proven any relationship between the γD and the TVD. However, Hung et al. [2018] show that under a linear ϵ-contamination minimising the γD can estimate the parameters of the uncontaminated model with negligible bias.

2.6.6 The Divergence Hyperparameters

The hyperparameters α, β or γ control the trade-off between robustness and efficiency and selecting these hyperparameters is part of the subjective judgement associated with selecting that divergence. These values can be considered as a sort of meta-prior on the confidence the DM has in their model. This being the case we feel that to ask for multiple hyperparameter to be set by the DM is perhaps over ambitious. For this reason we have neglected to mention any 2 parameter divergence families for example the $\alpha\beta$-divergence [Cichocki et al., 2011]. We discuss the setting of these divergence hyperparameters further in Section 2.7.5. The influence curves introduced in the next section provide a way to understand the impact of selecting a particular divergence and its hyperparameter.

Once we consider the divergence and its associated hyperparameters as part of the subjective specification of the analysis, one can then consider investigating the sensitivity of the analysis to this exact specification. When moving into the $\mathcal{M}$-OPEN world there is no notion of the correct divergence, or the correct value of any hyperparameters associated with the target divergence. However, the difference between two analyses with different divergences or hyperparameters can be very informative about the relationship between the likelihood model under consideration and the sample distribution of the data. For example, if traditional Bayesian inference using the KLD produces very different inferences to those using the robust HD or TVD for example, then this is informative about a lack of tail correspondence between the model and the sample distribution of the data. On the other hand agreement between the inference minimising different divergences can reassure the DM that their model sufficiently captures the DGP. Additionally, considering the stability of inference minimising the αD or βD to the value of the α or β can be very informative about the location of outliers. If α (β) is set sufficiently lower (higher) than 1 to ignore all outliers then decreasing (increasing) it further should only affect the inference by sacrificing a small amount of efficiency with no further robustness gained. The framework of Williamson et al. [2015] provides a possible approach to formalise these ideas above and use such alternative analyses to improve a DM's belief specification.

2.6.7 Influence curves

Here we introduce an important tool for understanding the role the target divergence has on the inference problem, the influence curve. To gain intuition about why these alternative divergences are able to produce robust inference we need to assess the relative impact any observation could have on a posterior. To do this we deploy a technique from the Bayesian outlier detection literature. Bayesian outlier detection methods aim to identify observations that have too great an influence on the posterior. This can be done by measuring the divergence between the posterior with and without each observation [Peng and Dey, 1995; Kurtek and Bharath, 2015]. That is to say define the influence of observation x_i on the posterior as

$$I(x_i) := D\left(\pi(\theta|x_{1:n})||\pi(\theta|x_{1:n\setminus i}\right) \tag{2.35}$$

for some divergence D. Peng and Dey [1995] consider ϕ-divergence (Definition 6), while Kurtek and Bharath [2015] use the non-parametric Fisher-Rao metric

$$D_{\text{FR}}\left(g,f\right) = \arccos\left\{\int\sqrt{g}\sqrt{f}d\mu\right\} \tag{2.36}$$

The square-root ensures the densities live on the "positive orthant of a Hilbert sphere" which in turn allows for the analytic computation of geodesic distances between the densities. Here we consider the Fisher-Rao metric for its appealing geometric interpretation and the fact that it is symmetric. We also note that the Fisher-Rao metric is an invertible transformation of the HD. A further advantage of these methods is their computational efficiency; sampling only from the full posterior is enough to produce a Monte-Carlo estimate of the influence for each observation.

We extend the use of these influence functions, to produce an influence curve. Rather than looking at the influence of observed data points, we evaluate the influence of any potential future observation to produce an influence curve

$$I(z) := D\left(\pi(\theta|x_{1:n}, z)||\pi(\theta|x_{1:n})\right) \tag{2.37}$$

We produce influence curves for all of the divergences we considered for inference in the previous section and explain what these tell us about the subjective selection of that divergence.

The influence curves in Figure 2.2 demonstrate the influence one observation m posterior standard deviations away from the posterior mean have on a posterior produced fitting the model $\mathcal{N}(\mu,\sigma^2)$ to $n = 1000$ observations from $\mathcal{N}(0,1)$. The KLD has a strictly increasing influence function in the number of posterior standard deviations from the mean. This demonstrates the fact that tail observations have large influence over the posterior. As a result the KLD is suitable if tails are important to the decision problem at hand, but increasing influence characterises a lack of robustness when tails are not important. Alternatively all of the robust divergence

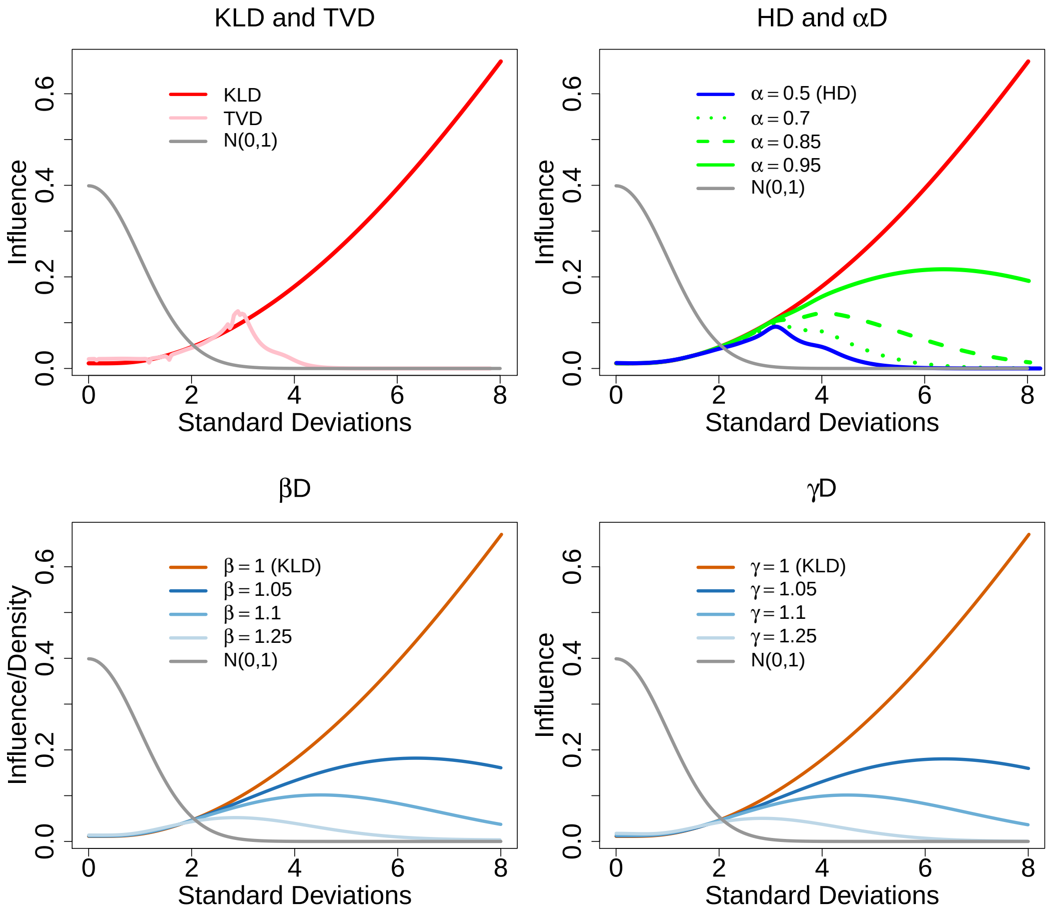

Figure 2.2: Influence Plots: The Fisher-Rao divergence between a posterior with and without a new observation at varying posterior standard deviations away from the previous posterior mean. **Top:** Bayesian inference minimising the KLD, TVD, HD and αD. **Bottom:** Bayesian inference minimising the KLD, βD and γD

measures listed above have concave influence functions. The influence of an observation increases as it moves away from the mean, mimicking the behaviour under the KLD initially, but then decreases as the observation is increasingly declared an outlier. The curves for the αD, βD and γD show that changing the divergence parameter allows a practitioner to change at what point observations are declared as outliers and thus control the level of robustness to tail observations. For smaller α (bigger β and γ) the influence functions start to decrease closer to the posterior mean characterising greater robustness to the tail specification of the model.

The influence curves for the αD and βD provide a Bayesian analogy to the frequentist measure of influence based on the estimating equation analysis discussed in Sections 2.6.3 and 2.6.4. What is not obvious from these influence curves is the fact the βD is down-weighting the influence of each observation based solely on the magnitude of its predicted likelihood $f(x;\theta)$, while the αD (and HD) is able to do so based on the ratio of $f(x;\theta)/g_n(x)$. As a result, these influence curves for the βD and a given β are dependent on the variance and dimension of the observation space. While, the αD is able to down-weight influence at a constant rate independent of the dimension and variance of the observation space. Increasing the dimension and/or variance, makes $f(x;\theta)$ small causing the βD to down-weight the influence of observations more. In contrast, as the variance and dimension increase $g_n(x)$ will also decrease allowing the ratio of $f(x;\theta)/g_n(x)$ to be maintained. This is examined further at the end of Section 2.7.4.

2.6.8 Density estimation

As has been mentioned before, for the TVD, HD and αD, it is not possible to exactly calculate the loss function associated with any value of θ and x because the data generating density $g(x)$ will not be available. In this case, a density estimate of $g(x)$ is required to produce an empirical loss function. The Bayesian can consider the density estimate as providing additional information to the likelihood from the data (see Section 2.5.2's discussion on the likelihood principle), and can thus consider their general Bayesian posterior inferences to be made conditional upon the density estimate as well as the data. The general Bayesian update is a valid update for any loss function, and therefore conditioning on the density estimate as well as the data still provides a valid posterior. However, how well this empirical loss function approximates the exact loss function associated with each divergence is of interest. The exact loss function is of course the loss function the DM would prefer to use having made the subjective judgement to minimise that divergence. If the density estimate is consistent to the data generating process, then provided the sample size is large the density estimate will converge to the data generating density, and the empirical loss function will then correctly approximate the loss function associated with that divergence. It is this fact that ensures the consistency of the posterior estimates of the minimum Hellinger posterior Hooker and Vidyashankar [2014].

Authors Hooker and Vidyashankar [2014] use a fixed width kernel density estimate (FKDE) to estimate the underlying data generating density and in our examples in Section 2.7 we adopt this practice using a Radial Basis Function (RBF) kernel for simplicity and convenience. These off-the-shelf methods are shown to work remarkably well for the low-dimensional problems considered here. However we note that applying such techniques to medium and high-dimensional problems is not so straightforward. For example, Silverman [1986] identifies practical drawbacks of FKDEs, including their inability to correctly capture the tails of the data generating process whilst not over smoothing the centre, as well as the number of data points required to fit these accurately in medium to high dimensions. In addition to this Tamura and Boos [1986] observe that the variance of the FKDE when using a density kernel in high dimensions lead to asymptotic bias in the estimate that is larger than $\mathcal{O}\left(n^{-1/2}\right)$. Alternatives include using a kernel with better mean-squared error properties (Epanechnikov [1969], Rosenblatt et al. [1976]), variable width adaptive KDEs [Abramson, 1982], which Hwang et al. [1994] show to be promising in high dimensions, piecewise-constant (alternatively tree based) density estimation [Ram and Gray, 2011; Lu et al., 2013] which are also promising in high dimensions, or a fully Bayesian Gaussian process as is recommended in Li and Dunson [2016]. In general we note that medium and high-dimensional density estimation is an active and important area of research. Developments in this field will further help facilitate the implementation of some of these robust methods to medium and high dimensional applications

2.7 Illustrations

In this section we aim to illustrate some of the qualitative features associated with conducting inference targeting the minimisation of the different divergences identified in Section 2.6. Throughout these experiments *stan* [Carpenter et al., 2016], implementing the No-U-Turn sampler [Hoffman and Gelman, 2014], is used to produce fast and efficient samples from the general Bayesian posteriors of interest. We demonstrate the impact model misspecifications can have on a traditional Bayesian analysis for simple inference, regression and time series analysis, and that superior robustness can be obtained by minimising alternative divergences to the KLD. In Section 2.7.4 we also show that when the observed data is in fact generated from the model, these methods can be shown not to lose too much precision. At this stage we have deliberately restricted ourselves to simple demonstrations designed to provide a transparent illustration of the impact that changing the divergence measure can have on inferential conclusions. However we discuss how robust methodology becomes more important as problems and models become more complex and high dimensional and thus encourage practitioners to experiment with this methodology in practice. In Chapter 4 and 5 we investigate the performance of some of these divergence measures for more complicated real world examples. For all of the experiments in this section we set w according to Lyddon et al. [2018][3]

2.7.1 $\mathcal{M}$-open robustness

Simple inference

The experiments below demonstrate the robustness of the general Bayesian update targeting KLD (red), HD (blue), TVD (pink), αD (green), βD (orange) and the γD (light blue). In the future these may be referred to as KLD-Bayes, HD-Bayes, TVD-Bayes, αD-Bayes, βD-Bayes and γD-Bayes respectively. For illustrative purposes we have fixed $\alpha = 0.75$ for the αD-Bayes and $\beta = \gamma = 1.5$ for the βD-Bayes and the γD-Bayes.

Firstly we consider again the ϵ-contamination example (p. 23). Figure 2.1 demonstrated that Bayes' rule was very non-robust to ϵ-contamination. Figure 2.3 plots the posterior predictive originating from fitting the same normal model $f(\cdot;\theta) = \mathcal{N}(\mu, \sigma^2)$ but using the robust divergences mention in the previous sections. Additionally, Figure 2.3 investigates the performance of the same model on a dataset consisting of $n = 1000$ simulations from a Student's t-distribution with degrees of freedom 3 and the real data set, tracks1, taken as the first variable from the 'Geographical Original of Music Data Set'[4][5] [Zhou et al., 2014]. This dataset contains information about $n = 1059$ music tracks with the aim to determine if 68 audio features can be used to predict the country of origin of the artist. Here we consider simply learning the data-generating density of the first variable, where a KDE of the data appeared to be approximately normally distributed. Prior distributions $\sigma^2 \sim \mathcal{IG}(0.001, 0.001)$ and $\mu|\sigma^2 \sim \mathcal{N}(0, 10^2\sigma^2)$ were used for all examples.

In contrast to Figure 2.1, the top left of Figure 2.3 shows that for the ϵ-contamination example the Bayesian inference targeting minimising the HD, βD and γD appears to correctly capture the distribution for 90% of the data. Their ability to down-weight the influence of outlying observations enables them to almost entirely ignore the outlying contamination. Minimising the TVD and αD fit marginally larger variances demonstrating that these methods give more influence to outlying observations but they still capture the distribution for the majority of the observations much more closely the the Bayes' rule predictive does.

The Student's t-distribution has consistently heavier tails than the normal

[3] For the TVD loss we use the absolute approximation $|x - a| \approx \frac{2}{k}\log(1 + \exp(k(x - a))) - x - \frac{2}{k}\log(2)$ [Schmidt et al., 2007] with $k = 1000$ to ensure stability of the gradients and Hessians when estimating w.

[4] downloaded from https://archive.ics.uci.edu/ml/datasets/Geographical+Original+of+Music

[5] the data set was transformed by adding $\min(tracks1) + 0.001$ to every value in order to make the data strictly positive so the gamma and log-Normal distributions could be applied

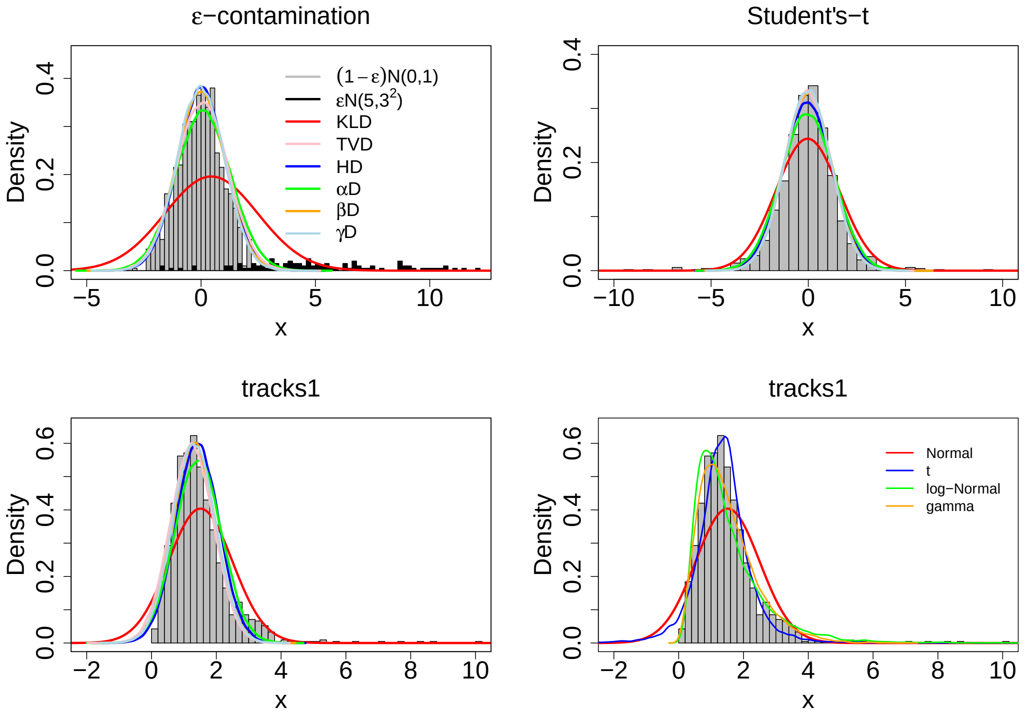

Figure 2.3: Posterior predictive distributions (smoothed from a sample) arising from Bayesian minimum divergence estimation fitting a normal distribution $\mathcal{N}(\mu, \sigma^2)$ to an ϵ-contaminated normal $0.9\mathcal{N}(0,1)+0.1\mathcal{N}(5,3^2)$ (top left), a t-distribution t_3 (top right) and the tracks1 dataset (bottom left) using inference targeting minimising the KLD (red), HD (blue), TVD (pink), αD (green), βD (orange) and the γD (light blue). The bottom right plots the posterior predictive distributions (smoothed from a sample) from alternative models using Bayes' rule, Normal (red), Student's-t (blue), logNormal (green) and gamma (orange).

distribution. Thus the top right hand plot of Figure 2.3 more clearly demonstrates the importance placed on tail misspecification by each method. The Bayes' rule predictive fits the largest variance to correctly capture these tails, the αD is able to fit a slightly smaller variance because of its bounded nature. However the variance of the αD predictive is still larger than that of the other methods because it gives greater influence to observations further from mean as depicted in Figure 2.2. Lastly targeting the minimisation of the HD, TVD, βD and γD places the least weight on tail misspecification and therefore these methods are able to fit a smaller variance and produce a predictives more closely resembling the data generating process for the majority of the data. A similar phenomenon is observed for the 'tracks' data. Here there is an ordering from βD, γD, TVD, HD, αD and KLD on both bias towards the right tail and on the size of the predictive variance, in response to the slight positive skew of the histogram of the data. Here it is clear to see that the TVD, HD, βD and γD produce much better fits of the majority of the data than the other methods do. Lastly, the bottom right plot demonstrates how several possible alternative models to the Gaussian perform on the 'tracks1' data set when updating using Bayes' rule. This shows that a Gaussian distribution was actually the best fit for the bulk of the data, and the poor fit achieved is down to the importance placed on tail misspecification by the estimating procedure rather than the model selected.

In order to demonstrate the improved fit of the minimum divergence methods in the context of a decision problem, consider classifying individuals as regular, if they lie between the 10% and 90% quantiles and irregular if they are outside this region. We use the 0-1 classification loss, where 0 loss is obtained for correctly classifying an observation and a loss of 1 is obtained for an incorrect classification, to compare the performance of the predictive distributions produced using the KLD, HD, TVD, αD, βD, γD from the optimal classification under the known DGP for the ϵ-contamination and Student's-t data. We use sampling from the model to estimate the 10% and 90% quantiles from the ϵ-contamination while these are available analytically for the Student's-t distribution. Table 2.1 presents the sum of the absolute difference between the optimal classification loss knowing g and the classification loss incurred under the KLD-Bayes, HD-Bayes, TVD-Bayes, αD-Bayes, βD-Bayes, γD-Bayes. For both the ϵ-contamination and the Student's-t datasets the KLD performs the worst with the alternative divergence methods all performing much better.

Table 2.1: Sum of the absolute difference between optimal 0-1 classification as 'regular' (between 10% and 90% quantile) or 'irregular' (outside 10% and 90% quantile) under known DGP and 0-1 classification produced under the KLD-Bayes, HD-Bayes, TVD-Bayes, αD-Bayes, βD-Bayes, γD-Bayes for the ϵ-contamination and Student's-t datasets fo size $n = 1000$.

	KLD	**HD**	**TVD**	$\boldsymbol{\alpha}$**D**	$\boldsymbol{\beta}$**D**	$\boldsymbol{\gamma}$**D**
ϵ-cont	111	58	54	47	73	74
Stu-t	90	8	18	27	16	20

2.7.2 Regression under heteroscedasticity

In addition to the simple inference examples above we consider how changing the divergence can affect inference in a regression example. From the previous examples we can see that when the tails of the model are misspecified the KLD minimising predictive distribution inflates the variance of the fitted model to ensure no observations are predicted with low probability. Further, we suspect that placing large weight on tail observations, which occur with low probability, will cause large variance across repeat sampling. This may explain why frequentist parameter estimates in linear regression under heteroscedasticity errors have larger variance. We seek to demonstrate these claims here and investigate whether the alternative methods mentioned above can improve this performance.

The alternative Bayesian minimum divergence methods place less weight on tail observations. These have been shown to be able to produce inferences with a smaller predictive variance and should also result in smaller estimation variance across repeat sampling. While repeat sampling and estimation variance are not problems in the Bayesian paradigm, these results do show that traditional Bayesian inference can be somewhat imprecise when the tails are misspecified. This is clearly undesirable when conditioning on observed data. We note that similar results have been observed in the context of Bayesian variable selection by Rossell and Rubio [2018]

In order to demonstrate this we simulated $n = 200$ data points with $N = 50$ repeats from the following heteroscedastic linear model

$$y \sim \mathcal{N}\left(X\boldsymbol{\beta}, \sigma(X_1)^2\right), \text{ where } \sigma(X_1) = \exp\left(\frac{2X_1}{3}\right). \tag{2.38}$$

For the experiments we simulated $X \sim \mathcal{N}_p\left(\mathbf{0}, I\right)$ and $\beta_i \sim$ Unif$[-2, 2]$ for $i = 1, \ldots, p$. We held the β's constant across repeat experiments. Figure 2.4 plots the observed data Y vs X_1 for one of these repeats. This shows how larger values of X_1 are associated with greater variance in Y. For each repeat we also simulated a testing set, $\tilde{X} \sim \mathcal{N}_p\left(\mathbf{0}, I\right)$, of size $\tilde{n} = 100$. We then calculated the data-generating means for the test set observations as $\tilde{Y} = \tilde{X}\boldsymbol{\beta}$. These characterise the part of the model that was is correctly specified for the DGP.

We then conducted Bayes' rule updating and general Bayesian updating using the five divergences mentioned above with priors $\sigma^2 \sim \mathcal{IG}(2, 0.5)$ and $\beta_i|\sigma^2 \sim (0, 5\sigma^2)$. The medians across the 50 repeats of the posterior mean estimate for the residual variance for the different methods are presented in Table 2.2, while Table 2.3 presents the median squared errors (MedSE)[6] across the 50 repeats for the posterior means of the parameters $\hat{\beta}_i$, $\sum_{i=1}^{p}\left(\hat{\beta}_i - \beta_i\right)^2$, as well as the MedSE for the

[6] Medians were used rather than means to reduce the sensitivity of these values to outlying experiments which we believe result from problems with the computation rather than the divergences properties for inference

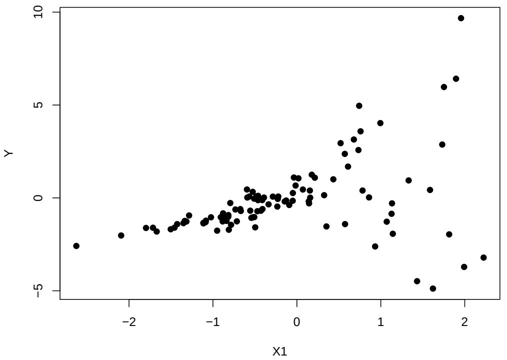

Figure 2.4: A data sets simulated from the heteroscedastic linear model $y \sim \mathcal{N}\left(X\boldsymbol{\beta}, \sigma(X_1)^2\right)$, with $\sigma(X_1) = \exp(2X_1/3)$ and $p = 1$.

predictive means on a test set $\tilde{Y}$, $\sum_{i=1}^{100}\left(\hat{Y}_i - \tilde{Y}_i\right)^2$. In order to apply the HD-Bayes, TVD-Bayes and αD-Bayes to a regression problem an estimate of the conditional density of the response given the covariates is required. We follow authors Hooker and Vidyashankar [2014] and implement conditional KDEs to approximate the true data generating density. For simplicity, the familiar two-stage bandwidth estimation process of Hansen [2004] was used to find the optimal bandwidth parameters.

Table 2.2: Table of posterior mean values for the residual variance of a standard linear model fitted to data from the heteroscedastic linear model $y \sim \mathcal{N}\left(X\boldsymbol{\beta}, \sigma(X_1)^2\right)$, with $\sigma(X_1) = \exp(2X_1/3)$. Estimates are medians across $N = 50$ repeats of the posterior means produced from datasets of size $n = 200$ with increasing dimension of the predictor space $p = 1, 5, 10, 15, 20$, under the Bayesian minimum divergence technology. Bayes' rule (KLD) fits a large predictive variance in order to accommodate the large variance in the DGP when X_1 is large. The alternative methods produce more accurate estimate of the residual variance across the majority of the space of X_1.

$\hat{\sigma}^2$	**KLD**	**HD**	**TVD**	α**D**	β**D**	γ**D**
$p = 1$	2.22	0.75	0.56	1.19	0.94	0.86
$p = 5$	2.19	0.43	0.37	0.95	0.89	0.83
$p = 10$	2.18	0.36	0.37	0.89	0.97	0.87
$p = 15$	2.25	0.32	0.59	0.82	0.89	0.84
$p = 20$	2.16	0.28	0.72	0.81	0.87	0.80

Table 2.2 demonstrates that the alternative divergences learn a smaller estimate of the residual variance, $\hat{\sigma}^2$, than Bayes' rule does under heteroscedastic errors. Minimising the KLD requires capturing the extremes of the variance of the response given the predictions which occurs when X_1 is large. However this value is a terrible

Table 2.3: Table of posterior mean Median Squared Errors (MedSE) values for parameters, $\boldsymbol{\beta}$, and test set means, $\tilde{Y}$, of a standard linear model fitted to the heteroscedastic linear model $y \sim \mathcal{N}\left(X\boldsymbol{\beta}, \sigma(X_1)^2\right)$, with $\sigma(X_1) = \exp(2X_1/3)$. Estimates are medians across $N = 50$ repeats of the posterior means produced from datasets of size $n = 200$ with increasing dimension of the predictor space $p = 1, 5, 10, 15, 20$, under the Bayesian minimum divergence technology. The alternative divergences estimate the parameters of the underlying mean function more accurately which allows them to perform better predictively.

MedSE	**KLD**		**HD**		**TVD**	
	β	$\tilde{Y}$	β	$\tilde{Y}$	β	$\tilde{Y}$
$p = 1$	0.02	1.66	0.02	1.83	0.04	4.24
$p = 5$	0.07	6.65	0.04	4.15	0.05	5.50
$p = 10$	0.14	13.84	0.09	9.44	0.13	12.44
$p = 15$	0.23	21.26	0.16	15.27	0.17	15.16
$p = 20$	0.28	23.95	0.21	20.05	0.20	18.32

MedSE	α**D**		β**D**		γ**D**	
	β	$\tilde{Y}$	β	$\tilde{Y}$	β	$\tilde{Y}$
$p = 1$	0.01	1.10	0.01	1.08	0.01	1.13
$p = 5$	0.04	3.98	0.03	3.23	0.03	3.34
$p = 10$	0.08	7.60	0.08	7.68	0.08	7.74
$p = 15$	0.12	11.30	0.12	11.07	0.12	11.09
$p = 20$	0.15	13.09	0.16	15.09	0.16	15.12

estimate of the residual variance of $Y|X$ for most of the support of X. On the other hand, the alternative divergences are able to ignore the areas of extreme variance and estimate a residual variance that captures that data generating variance more closely across a greater range of the predictor space.

We note that for these reasons the estimate of the $\hat{\sigma}^2$ will no longer correspond to an estimate of the residual variance of $Y|X$ as this will be dominated by the large variance terms when X_1 is large. This should rather be interpreted predictively as the variance of the predictive distribution which is closest to the data generating distribution in terms of that alternative divergence. The KLD estimate will correctly capture the variance for large X_1 but drastically over-estimate this for the majority of X_1's support. On the other hand the alternative methods will underestimate the variance for large X_1's put produce a much more accurate estimate fo the variance in the DGP across the majority of the support of X_1. We discussed in Section 2.5.5 that in the $\mathcal{M}$-open world the DM does not necessarily have to use the model to make future predictions. However, we believe it provides a sensible default and provides an interpretation of the inferred parameters in terms of observables.

Table 2.3 illustrates the impact that the fitting of a large variance has on the parameter estimates of the mean function. Placing less influence on outliers allows all of the alternative divergences to produce more precise estimates of the parameters of the underlying linear relationship that was correctly specified by the model. This then leads to better performance when predicting the uncontaminated test set. Clearly the errors for all of the methods will increase as p increases because the same amount of data is used to estimate more parameters. However it is clear that the errors under the KLD are rising more rapidly. By being less sensitive to the error distribution the alternative divergences are better able to capture the true underlying process for the mean. This may very well explain why under misspecification the traditional Bayesian marginal likelihood loses power to detect truly active coefficients under misspecification [Rossell and Rubio, 2018].

2.7.3 Time series analysis

In order to further demonstrate how inflating the variance by targeting the KLD under misspecification can damage inference, we consider a time series example. We simulate $x_1, \dots, x_T$ from an auto-regressive process of order L (AR(L)), and then consider additive independent generalised auto-regressive conditionally heteroscedastic errors of order (1,1) (GARCH(1,1)), $e_1, \dots, e_t$. The data generating model can be summarised as follows

$$x_t = \sum_{i=1}^{L} \mu_i x_{t-i} + \epsilon \text{ with } \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{2.39}$$

$$e_t = \psi_t \epsilon_k \text{ with } \epsilon \sim \mathcal{N}(0, 1) \tag{2.40}$$

$$\psi_t^2 = \omega + \alpha_1 e_{t-1}^2 + \beta_1 \psi_{t-1}^2 \tag{2.41}$$

$$y_t = x_t + e_t \tag{2.42}$$

where $\omega > 0$, $\alpha_i > 0$, $\beta_j \geq 0$. GARCH processes are used to model non-stationary, chaotic time series where the variance of the process depends on the magnitude and sign of the previous observations of the process. Eliciting a GARCH process from a DM is a difficult task. It is far from obvious how this GARCH process behaves as a function of its parameters and selecting a lag length for the AR process as well as two lag lengths for the GARCH process increases the complexity of the model selection problem. Therefore it seems conceivable that a DM could want to fit a simple AR process to noisy time series data in order to investigate the underlying process. One situation where this may be desirable is in financial time series applications where large amounts of data can arrive at a very high frequency.

In order to investigate how the minimum divergence methods perform in this scenario we simulated 3 data sets with $T = 1000$, and fitted an AR(L) process to these, where L was chosen to match the underlying AR process. The 3 data sets were given by

1. an AR(3) with $\mu = (0.25, 0.4, 0.2, 0.3)$
2. an AR(1) with $\mu = (0, 0.9)$ with GARCH(1,1) errors $\omega = 2, \alpha_1 = 0.99, \beta_1 = 0.01$.
3. an AR(1) with $\mu = (0, 0.9)$ with GARCH(1,1) errors $\omega = 1, \alpha_1 = 0.75, \beta_1 = 0.01$

The plots in Figure 2.5 demonstrate the one-step ahead posterior predictive performance of the minimum divergence posteriors on a test set $T = 100$, simulated from the underlying AR process. Under misspecification we show only the inference under the HD-Bayes to avoid cluttering the plots. Results not presented here showed that the other minimum divergence posteriors perform similarly. We use the same priors as the regression example and once again conditional density estimates were used for the loss functions of the HD, TVD and αD divergences.

The top plots demonstrate that when the model is correctly specified the minimum divergence posteriors produce similar time series inference to Bayes' rule. This is most easily seen in the right hand plot which shows the difference in the squared prediction errors between the Bayes' rule predictive and the predictive minimising the HD is mostly around 0 and Table 2.4 which demonstrates the root mean squared error across the test set is the same under both methods. The middle plots demonstrate how the Bayes' rule predictive and the predictive minimising the HD perform under 'extreme' volatility in the error distribution. The tails of the model being very poorly specified causes Bayes' rule to fit a huge variance with the average posterior predictive variance across the test dataset being slightly above 26. Fitting such a high variance makes the inference on the auto-regressive parameters μ insensitive to the data. As a result the underlying trend in the data is completely missed. In contrast the posterior predictive minimising the HD have a much smaller variance of around 7. This allows the inference on the lag parameters to be much more sensitive to the underlying AR process. Clearly the predictive minimising the HD is unable to exactly fit the truth because the model is misspecified. However it does a much better job of capturing the broad features of the underlying dependence

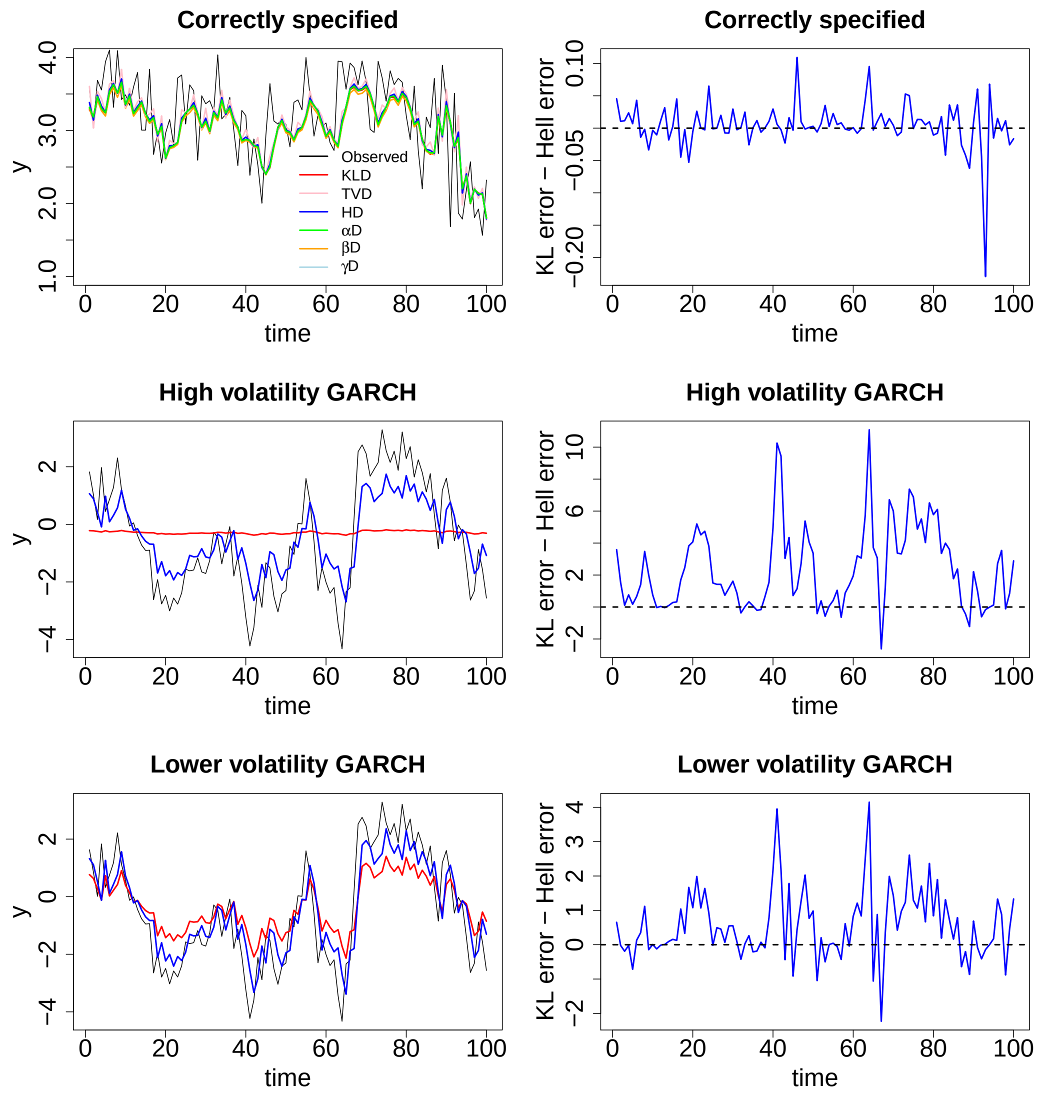

Figure 2.5: **Left**: One step ahead posterior predictions arising from Bayesian minimum divergence estimation fitting AR models with the correctly chosen lags to an AR(3) with no additional error (**top**), an AR(1) with GARCH(1,1) errors, $\alpha = 0.99$, $\omega = 2$ (**middle**) and a AR(1) with GARCH(1,1) errors, $\alpha = 0.75$, $\omega = 1$ (**bottom**) using inference targeting minimising the KLD (red), HD (blue), TVD (pink), αD (green), βD (orange) and the γD (light blue). **Right**: the difference in one step ahead posterior squared prediction errors when using Bayes' rule and minimising the HD. When the model is correctly specified all of the methods appear to perform similarly. Under misspecification minimising the HD does a much better job of correctly capturing the underlying dependence in the data.

between time points. This is clearly demonstrated by the considerably lower root mean squared predictive error showed in Table 2.4 and by the error differences plot on the right of Figure 2.5 being mostly large and positive.

The bottom plots demonstrates how the Bayes' rule predictive and predictive minimising the HD perform when the error distribution is less volatile. When the volatility is smaller the Bayes' rule predictive variance is also smaller. Therefore the inference is more sensitive to the underlying trend in the data than in the previous example. However, the true dependence in the data is still under estimated relative to the inference minimising the HD. This is again demonstrated in Table 2.4 and the error difference plot on the right of Figure 2.5. Once again this shows that the way in which Bayes' rule deals with misspecification, increasing the predictive variance, can mask some of the underlying trends in the data which can be discovered by other methods. Table 2.4 plots the root mean squared errors (RMSE) for Bayes' rule and the HD-Bayes on the test set to quantify their correspondence to the data.

Table 2.4: Root mean squared errors (RMSE) for Bayes' rule and HD minimising posterior mean predictions when fitting an AR model to an AR(3) with no additional error, an AR(1) with 'high volatility' GARCH(1,1) errors, and a AR(1) with 'low volatility' GARCH(1,1) errors, for 100 test data points from the underlying AR model.

RMSE	Correctly Specified	High Volatility	Low Volatility
KL	0.49	1.87	1.21
Hell	0.48	1.07	0.94

2.7.4 $\mathcal{M}$-closed efficiency

The examples in Section 2.7, demonstrate how inference designed to minimise an alternative divergence to the KLD can lead to superior robustness to tail misspecifications. It was also observed in Section 2.5 that when inference is done in the $\mathcal{M}$-CLOSED world, where the model is correctly specified, these alternative divergence methods are still able to learn about the data generating parameter. However, by placing less importance on tail misspecifications in order to gain improved robustness, the DM must trade-off a decrease in efficiency in the case when the model class does in fact contain the DGP. Minimising the KLD uses Bayes' rule and thus conditions on the data coming from the model. As a result minimising the KLD is guaranteed to perform the best when the model class contains the DGP. However, we demonstrate here that this trade-off between robustness and efficiency is asymmetric in the favour of these robust methods, a lot of robustness can be gained without losing too much efficiency.

In order to examine the frequentist efficiency, the observed mean squared error (MSE) $\frac{1}{N}\sum_{j=1}^{N}(\hat{\theta}_n^j - \theta)^2$, over N repeats of a simulated experiment are examined. The MSEs are examined on data generated from $X \sim \mathcal{N}(0, 10)$ under fitting the model $X \sim \mathcal{N}(\mu, \sigma^2)$ and two prior regimes $\sigma_i \sim \mathcal{IG}(2.1, 4)$, for $i = 1, 2$ and $\mu_1 \sim \mathcal{N}(0, 100\sigma_1^2)$ and $\mu_2 \sim \mathcal{N}(20, 100\sigma_2^2)$. The prior for μ_1 is centred on the data while the prior for μ_2 is not. Table 2.5 plots the observed MSEs over $N = 100$ replications of this experiment.

Table 2.5: Table of posterior mean MSE values when estimating $\mathcal{N}(\mu_i, \sigma_i^2)$ from data sets simulated from the model of size $n = 50, 100, 200, 500$ under the Bayesian minimum divergence technology.

	MSE	**KLD**		**HD**		**TVD**	
		$\mu_{01} = 0$	$\mu_{02} = 20$	$\mu_{01} = 0$	$\mu_{02} = 20$	$\mu_{01} = 0$	$\mu_{02} = 20$
μ	$n = 50$	0.19	0.19	0.23	0.22	0.61	0.61
	$n = 100$	0.08	0.08	0.08	0.09	0.14	0.14
	$n = 200$	0.03	0.03	0.03	0.03	0.05	0.05
	$n = 500$	0.02	0.02	0.02	0.02	0.03	0.03
σ^2	$n = 50$	4.30	4.22	13.10	11.41	17.72	16.79
	$n = 100$	2.26	2.24	6.69	6.14	6.61	6.62
	$n = 200$	1.09	1.08	3.45	3.27	2.82	2.82
	$n = 500$	0.45	0.45	1.21	1.17	1.46	1.46
	MSE	α**D**		β**D**		γ**D**	
		$\mu_{01} = 0$	$\mu_{02} = 20$	$\mu_{01} = 0$	$\mu_{02} = 20$	$\mu_{01} = 0$	$\mu_{02} = 20$
μ	$n = 50$	0.20	0.20	0.25	0.25	0.25	0.25
	$n = 100$	0.08	0.08	0.10	0.10	0.10	0.10
	$n = 200$	0.03	0.03	0.04	0.04	0.04	0.04
	$n = 500$	0.02	0.02	0.02	0.03	0.03	0.03
σ^2	$n = 50$	5.62	5.26	5.91	5.78	6.30	5.90
	$n = 100$	2.89	2.77	3.05	3.02	3.24	3.13
	$n = 200$	1.44	1.40	1.44	1.42	1.49	1.45
	$n = 500$	0.51	0.50	0.60	0.60	0.62	0.62

Table 2.5 demonstrates several interesting points about these different methods. Unsurprisingly Bayes' rule, minimising the KLD, provides the lowest MSE for both parameters and all values of n. For large n the MSEs for μ are generally the same and very small for all of the methods. αD-Bayes with $\alpha = 0.75$ performs only slightly worse than Bayes' rule and provides more efficient finite sample inference than the HD-Bayes ($\alpha = 0.5$). This is unsurprising as α regulates the robustness efficiency trade-off and the previous experiments show the HD-Bayes to be more robust than the αD-Bayes. The TVD-Bayes provides the worst finite sample efficiency of all of the methods. We suspect this is probably a result of the irregularities of the

TVD score function which were discussed in Section 2.6.1 and are now also shown in Figure 2.6. There is also evidence from Table 2.5 that the prior has much less influence on the TVD-Bayes than it does on the other methods. Table 2.6 suggest this may be to do with the exceedingly large calibration weight estimated for the TVD loss, which we believe again to result from the irregularities of the TVD loss function. Lastly the βD-Bayes and γD-Bayes appear to perform marginally worse than the αD-Bayes and better than the HD-Bayes, which is particularly impressive given they do not require a density estimate. For fixed $\beta = \gamma = 1.5$ it seems as though the βD-Bayes is more efficient than the γD-Bayes, which agrees with the findings of Jones et al. [2001]. Setting the calibration weight appears to alleviate the fears of Section 2.6 that these methods with 'smaller' loss functions might be much more sensitive to the prior. Bizarrely better performance appears to be achieved when the prior is specified away from the data here.

Table 2.6: Table of values for the calibration weight w estimated by the method of Lyddon et al. [2018] when fitting $\mathcal{N}(\mu_i, \sigma_i^2)$ to a dataset of size $n = 500$ under the Bayesian minimum divergence technology.

$n = 500$	KLD	HD	TVD	αD	βD	γD
w	1.00	1.07	15.74	0.58	4.40	1.70

Examining whether the estimates of σ^2 were generally above or below the data generating parameter demonstrates how these different methods learn, see Table 2.7. Firstly it appears as though the prior, with mean $\frac{4}{2.1-1} \approx 3.6$ and variance $\frac{4^2}{(2.1-1)^2(2.1-2)} \approx 132$, encourages an estimate of σ^2 smaller than the data generating value, $\sigma^2 = 10$, as all methods generally underestimate this for small n. Under the Bayes' rule, βD-Bayes and γD-Bayes, the effect of this appears to decrease as n increases, as is expected with prior influence. However the HD-Bayes and αD-Bayes still consistently underestimate σ^2 for larger n while the TVD-Bayes consistently overestimates σ^2 in-spite of the prior specifications. We note there does not appear to be any negligible biases in the estimation of μ for any of the methods under either prior.

Figure 2.6 compares the score functions used in the general Bayesian update targeting the minimisation of the KLD, TVD, HD and αD. As α increases, the upper bound on the score associated with the αD increases and the score functions become more convex, tending towards the log-score when $\alpha = 1$. In contrast, the score associated with the HD is closer to being linear. We believe that this is responsible for causing the smaller variance estimates when minimising the HD and αD when $\alpha = 0.75$. The closer the score is to being linear the more similar the penalty is for

Table 2.7: Table of sums of posterior mean over (positive) or under (negative) estimation $\mathcal{N}(\mu_i, \sigma_i^2)$ from data sets simulated from the model of size $n = 50, 100, 200, 500$ under the Bayesian minimum divergence technology

	MSE	**KLD**		**HD**		**TVD**	
		$\mu_{01} = 0$	$\mu_{02} = 20$	$\mu_{01} = 0$	$\mu_{02} = 20$	$\mu_{01} = 0$	$\mu_{02} = 20$
μ	$n = 50$	0	0	2	0	4	6
	$n = 100$	-18	-18	-1	-10	-8	-8
	$n = 200$	-6	-6	-16	-8	-8	-8
	$n = 500$	-4	-4	-2	0	-4	-4
σ^2	$n = 50$	-28	-26	-92	-90	8	10
	$n = 100$	-18	-14	-90	-90	36	36
	$n = 200$	-16	-6	-94	-94	46	46
	$n = 500$	8	8	-80	-80	64	64

	MSE	α**D**		β**D**		γ**D**	
		$\mu_{01} = 0$	$\mu_{02} = 20$	$\mu_{01} = 0$	$\mu_{02} = 20$	$\mu_{01} = 0$	$\mu_{02} = 20$
μ	$n = 50$	0	0	-2	-4	-6	-6
	$n = 100$	-18	-18	-14	-14	-16	-14
	$n = 200$	-6	-4	-10	-10	-10	-10
	$n = 500$	-4	0	-6	-8	-2	-4
σ^2	$n = 50$	-64	-64	-44	-40	-48	-46
	$n = 100$	-60	-58	-18	-18	-24	-22
	$n = 200$	-54	-50	-18	-16	-20	-20
	$n = 500$	-42	-42	-6	-6	-8	-6

over and under predicting the probability of an observation when compared with reality. The bounded nature of the scores prevents the penalty for under predicting being too high, and therefore more posterior predictive mass is placed near the MAP of the posterior predictive distribution. In contrast, a score function with greater convexity will penalise under prediction compared with the data generating density to a greater extent spreading the posterior mass out further. However, one must keep in mind that it is the severe nature of the penalty for under prediction incurred by the KLD that renders it non-robust. On the other hand the TVD-Bayes generally over predict the variance. The score function for the TVD-Bayes is the only score function considered here that is not monotonic in the predicted probability of each observation. While under the other scores, the score for each individual observation can be increased by predicting that observation with greater probability, this is not the case for the TVD-Bayes. We believe this causes the TVD-Bayes to produce variance estimates that are too large.

We emphasised above that the superior small sample efficiency of the βD-

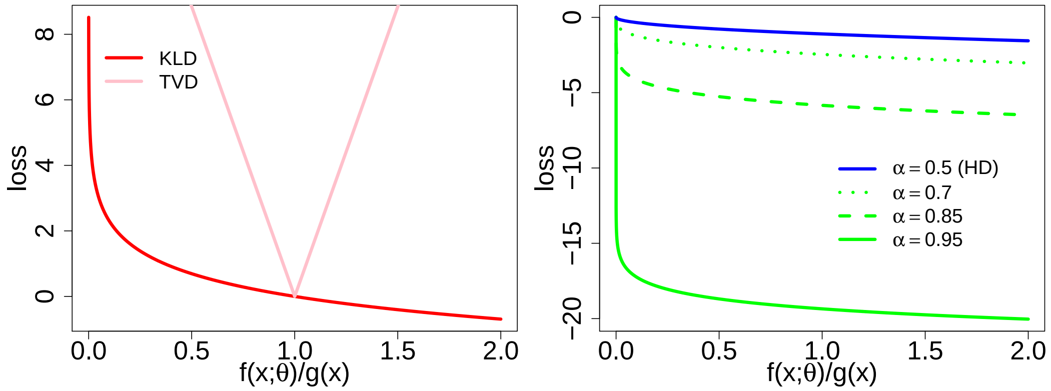

Figure 2.6: The loss functions of the KLD and TVD (**left**) and the HD and αD (**right**) for different values of $f(x;\theta)/g(x)$.

Bayes and γD-Bayes compared with the HD-Bayes was dependent on being able to select a value of β or γ correctly managing the robustness efficiency trade-off. This is evidenced in Table 2.8 where the values of β and γ are increased above 1.5, and the small sample efficiency can be seen to deteriorate.

Table 2.8: Table of posterior mean MSE values when estimating $\mathcal{N}(\mu_i, \sigma_i^2)$ from data sets simulated from the model of size $n = 50, 100, 200, 500$ under the βD-Bayes and γD-Bayes for increasing values of β and γ.

	MSE		β**D**			γ**D**	
		$\beta = 1.5$	$\beta = 2.5$	$\beta = 4$	$\gamma = 1.5$	$\gamma = 2.5$	$\gamma = 3$
μ	$n = 50$	0.25	0.40	1.05	0.25	0.51	0.70
	$n = 100$	0.10	0.18	0.29	0.10	0.21	0.32
	$n = 200$	0.04	0.09	0.16	0.04	0.10	0.14
	$n = 500$	0.02	0.05	0.09	0.03	0.05	0.07
σ^2	$n = 50$	5.91	8.09	8.38	6.30	20.19	32.06
	$n = 100$	3.05	4.58	5.28	3.24	9.37	15.45
	$n = 200$	1.44	2.06	2.38	1.49	4.21	7.12
	$n = 500$	0.60	0.89	1.11	0.62	1.63	2.69

As was explained in Section 2.6, the use of a density estimate allows the HD and αD to down-weight the influence of observations based on the ratio f/g. The βD-Bayes and γD-Bayes do not need a density estimate and therefore down-weights the influence of observations solely based on their predicted probability under the model f. Therefore, taking β or γ too high can result in a serious loss of small sample efficiency.

We note that all of the scores remain proper and therefore for large samples they should all recover the DGP in this M-CLOSED scenario. The observations of this section simply concern how the different methods perform for small sample sizes.

2.7.5 Application to higher dimensions

The examples in this chapter only demonstrate the performance of these minimum divergence techniques for relatively small dimensional problems. These have been specifically designed so as to clearly demonstrate the effect that tail misspecifications can have. However, it is as the dimension and complexity of the problem increases that we believe these methods become more and more important. This can be put down to two aspects. The first of these is that outliers or highly influential contaminant data-points become hard to identify in high dimensions. In our examples it is clear from looking at the KDE or histogram of the data that there are going to be outlying observations, but in many dimensions visualising the data in this way is not possible. In addition to this, automatic methods for outlier detection struggle in high dimensions [Filzmoser et al., 2008].

The second factor in requiring robust inference in high dimensions, is that not only are outliers harder to identify, they are more likely to occur. The occurrence of outliers indicates that the DM's belief model is misspecified in the tails. These misspecifications should be unsurprising. We have already discussed that specifying beliefs about tail behaviour requires thinking about very small probabilities which is known to be difficult [Winkler and Murphy, 1968; O'Hagan et al., 2006], and often routine assumptions (for example Gaussianity) may be applied. As the dimension of the space increase the tails of the distribution account for a greater proportion of the overall density, increasing the chance of seeing observations that differ from the practitioners beliefs. The impact these Bayesian minimum divergence methods can have in high-dimensional real world problems is investigated further in Chapters 4 and 5.

Unfortunately there is no free lunch when it comes to applying any of these methods to complex, high dimensions problems. We briefly discuss some of the issues here.

An estimate of the data generating density: Minimising the HD, TVD and αD requires an estimate of the data generating density. Section 2.6.8 includes references to demonstrate that it is possible to do this in high dimensions, but these are not necessarily straightforward to implement. We eagerly await further research in this area.

The divergence hyperparameter: However, given an estimate of the data generating density for the αD, α can then be selected based on how important tail misspecifications are, with a guarantee on some reasonable efficiency. We consider $\alpha = 0.5$ and the HD as a realistic lower bound for how robust one might desire to be and this has been shown in Table 2.5 to have reasonable small sample efficiency. Although minimising the βD or γD has the computational advantage of not requiring an estimate of the data generating density, there is no longer any guarantee that a reasonable level of efficiency will be obtained in high dimensions. As the dimension increases the predicted probability of each (multivariate) observation shrinks towards 0. The βD and γD down-weight the influence of observations with small predicted probabilities, and as a result β or γ needs to be selected very carefully in order to prevent the analysis from disregarding the majority of the data. This is precisely what damages the small sample efficiencies in Table 2.8 when β and γ are too high in one dimension. As the dimension increases the small sample efficiency will deteriorate for smaller values of β or γ and happen at a faster rate. In fact, we identified that the bound on the $D_B^{(\beta)}(g||f)$ using the TVD (g, f) (Theorem 2) was only useful if $\frac{M}{\beta-1}$, where $M = \max\{\operatorname{ess\,sup} g, \operatorname{ess\,sup} f\}$, was not too small. Increasing the dimension of the observation space will generally lead to a decrease in M. To see this consider a multivariate Gaussian density with diagonal covariance matrix and fixed variance acorss dimension

$$M = \frac{1}{(2\pi)^{d/2}\,(\sigma^2)^{d/2}} \tag{2.43}$$

which is clearly decreasing in d.

The technology available in the literature for setting β effectively is limited. This is a price that is paid in order to not require a non-parametric density estimate. Restricted mainly to the βD, Section 5.4.6 provides an empirical way to consider setting β using the influence curves above and on-line optimisation. A promising area of research here may be to consider the choice of β as a type of 'meta-prior' based on how confident a DM is that their likelihood model for the data is correctly specified. However, ways to transparently covey the impact of a certain choice of β must be developed in order for this 'meta-prior' to be practically useful.

Computation Lastly minimising an alternative divergence to the KLD results in conjugate posterior distributions no longer being available. The solution to this problem adopted in this chapter was to use Monte-Carlo methods to sample from the minimum divergence posteriors. In particular we employed the *stan* probabalistic programming language [Carpenter et al., 2016] implementing the No-U-Turn sampler [Hoffman and Gelman, 2014]. However, we note here that as the dimension increases these become computationally expensive also [Beskos et al., 2013]. Chapters 4 and 5 attempts to address this problem by introducing a convenient 'approximation' technique to supplement this vanilla inferential technique.

2.8 Further work

This chapter uses general Bayesian updating [Bissiri et al., 2016] in order to theoretically justify a Bayesian update that targets the parameters of a model that minimise a statistical divergence to the data generating process that is not the KLD. When the $\mathcal{M}$-OPEN world is considered, moving away from targeting the minimisation of the KLD can provide an important tool in order to robustify a statistical analysis. The desire for robustness ought to only increase as increasingly bigger models are built to approximate more complex real world processes. This chapter outlines to the statistical practitioner a principled justification through which they can select the divergence they use for their analysis in a subjective manner allowing them the potential to make more useful predictions from their best approximate belief model.

The next chapters seek to address some of the further work inspired by this chapter. Chapter 3 seeks to produce further theoretical results motivating switching from minimising the KLD via Bayes' rule to an alternative, more robust divergence. Specifically we investigate how stable each of these updating rules is to the particular choice of misspecified model which may help to motivate the choice of divergence. We discussed in Section 2.7.5 that even for simple models not minimising the KLD breaks the conjugacy property of Bayes' rule updating. Chapter 4 takes a foundational look at methods to improve the computability of these methods and produces a particularly convenient inference algorithm tailored towards minimising the βD that we implement in Chapter 5. Lastly, the investigation into how these minimum divergence methods perform empirically at the end of this chapter has been limited to analysing their performance in simulated examples that, by modern standard are relatively simple and small. Chapters 4 and 5 investigate how the robustness-efficiency trade-off associated with some of these methods manifests itself in real world high-dimensional examples. In doing so initial progress towards methods to guide the selection of divergence hyperparameters is made in Chapter 5. In particular these make use of the influence curves introduced in in Section 2.6.7.

Chapter 3

The Stability of Bayesian Inference

Chapter 2 considered producing the most useful inference, in a decision making capacity, from the DM's single best guess belief model when acknowledging that inferences are taking place in the M-OPEN world. In fact, when we consider the M-OPEN world it is more likely that there exists a whole set of equally preferable belief models all representing the beliefs the DM has about the DGP. Chapter 3 considers this scenario. Given such an equivalence class of beliefs models it is then natural to investigate what can be said about the stability of inference across this class of models. This chapter shows that very little can be said about the stability of Bayes' rule inferences, but that inferences using divergences that are proper metrics and inference minimising the βD can be shown to provide stability guarantees.

An outline of this chapter is as follows: firstly in Sections 3.1-3.3 we motivate demanding stability to the subjective specification of the likelihood function for a Bayesian analysis, analogously to the subjective specification of the prior. Sections 3.3.1 and 3.3.3 will establish what it means for inference to be stable with respect to the specification of the likelihood. We will proceed to provide results guaranteeing the stability of both the finite sample predictive, Section 3.3.4, and their ability to approximate the DGP as $n \to \infty$, Section 3.3.5, for inference using divergences that are symmetric and satisfy the triangle inequality (divergences that are metrics e.g. the TVD and HD). However inference targeting the minimisation of metrics generally require non-parametric density estimates and are therefore not necessarily straightforward to implement. We thus develop methods to approximate the stability results available for metrics using divergences whose associated score function is local meaning that it does not require a non-parametric density estimates. Section

3.4.1 demonstrates the difficulty in ensuring stability when using Bayes' rule. However, we are able to present very promising results for the stability of inference when using the βD both in terms of the predictive distributions, Section 3.4.2, and their limiting ability to approximate the DGP as $n \to \infty$, Section 3.4.2. Lastly Section 3.5 illustrates the impact these result can have in some simple model specification examples.

3.1 A set of approximate Models

Chapter 2 considers inference from the point of view of making informed decisions under the model misspecification paradigm. In this setting we assumed that the model used for inference was the DM's best guess approximation (either of their own beliefs or of the DGP), capturing some of the important structure of the underlying process. In reality the DM is unlikely to have one best guess belief model. There will be many likelihood models consistent with the judgements the DM has been able to elicit. This observation is by no means a recent one: de Finetti for example was quoted by Dempster [1975] as saying "Subjectivists should feel obligated to recognise that any opinion (so much more the initial one) is only vaguely acceptable... So it is important not only to know the exact answer for an exactly specified initial problem, but what happens changing in a reasonable neighbourhood the assumed initial opinion.", while Savage [1972] remarked "... in practice the theory of personal probability is supposed to be an idealization of one's own standard of behaviour; that the idealization is often imperfect in such a way that an aura of vagueness is attached to many judgements of personal probability...".

It is possible to formalise the notions above by suggesting that any belief specification (and particularly specification of absolutely continuous densities) requires some level of interpolation. There may be several judgements the DM is accurately able to make, but in order to produce the densities or distributions required to fulfil a full Bayesian analysis will require an interpolation between these. For example, O'Hagan [2012] argues that only judgements of medians and quantiles can be reliably made while the rest of the distribution must be filled in arbitrarily.

As a result inference is being done with one possible candidate model but with no principled reason for choosing that particular model, besides maybe mathematical and computational convenience. When this is acknowledged it is then natural to seek to analyse the sensitivity of inference to such arbitrary decisions. This chapter seeks to investigate how stable both traditional Bayesian inference through Bayes' rule and the robustified inference introduced in Chapter 2 are across such a set of admissible likelihood models.

One solution to this is to let the data help guide any decision the DM themselves is not able to make. That is to say, formulate any judgements the DM is uncertain about as alternative belief distributions for the data and use the data to decide which of these is best reflected. This is naturally encompassed in a Bayesian analysis through Bayesian model posteriors and the resulting model averaging (BMA) and selection. Section 1.1.4 discussed BMA as a solution to the $\mathcal{M}$-OPEN world, concluding that its traditional form [Hoeting et al., 1999] was not suitable for the $\mathcal{M}$-OPEN world but that alternatives for example the posterior belief assessment of Williamson et al. [2015] or other methods such as stacking [see e.g. Yao et al., 2018b, and references within] could be promising.

However, these can be computationally prohibitive. Firstly it is unlikely that the DM has a small discrete set of models they equivalently feel describe their beliefs. In practice the set of models consistent with the beliefs they have been able to elicit will be enormous. Additionally each likelihood model requires the careful elicitation of a parameter prior and even once this has been done the within-model computations are non-trivial.

In this chapter we take a different approach. We seek to investigate the automatic stability guarantees possessed by the Bayesian updating machinery introduced in Chapter 2 across a class of models. As a result this chapter complements works on model averaging and selection. Proving that Bayesian updating is stable across a class of models removes the need to consider these alternative analyses and thus relives some of the computational burden associated with averaging and selection procedures.

3.2 Prior Stability

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

$$\sup_{g\in\Gamma}\lim_{\epsilon\to 0}\left\{\frac{\text{TVD}\,(f_n, g_n^\epsilon)}{\text{TVD}\,(f_0, g_0^\epsilon)}\right\} \tag{3.1}$$

where f_n and g_n^ϵ are the posteriors produced by Bayes' rule from n observations with shared likelihood and respective priors f_0 and g_0^ϵ, and Γ is some class of contaminant priors. The quantity in Eq. (3.1) 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. (3.1) diverges 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 neighbourhoods of prior densities defined using the local De Robertis distance

$$D_A^R(f, g) := \sup_{\theta,\phi\in A}\left|\frac{f(\theta)g(\phi)}{f(\phi)g(\theta)} - 1\right|. \tag{3.2}$$

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 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 chapter, in contrast, is interested when such likelihoods might only be within a neighbourhood of their own.

More generally than the prior to posterior stability discussed above, stability of optimal decision making has been considered, largely in economics [Hansen and Sargent, 2001a,b; Whittle and Whittle, 1990; Gilboa and Schmeidler, 1989] and more recently in statistics [Watson 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

$$d_C^* := \arg\min_{d\in\mathcal{D}} \sup_{\nu\in\Gamma_C} \mathbb{E}_{\nu(\theta)}\left[\ell(\theta, d)\right] \tag{3.3}$$

across a KLD neighbourhood of the Bayes' rule posterior beliefs

$$\Gamma_C := \left\{\nu(\theta) : \text{KLD}(\nu(\theta)||\pi(\theta|\boldsymbol{x})) \leq C\right\}. \tag{3.4}$$

Watson et al. [2016] are actually able to derive the form of

$$\pi_C^{\sup} := \arg\sup_{\nu\in\Gamma_C} \mathbb{E}_{\nu(\theta)}\left[\ell(\theta, d)\right] \tag{3.5}$$

showing similarities to the form of the general Bayesian update Eq. (1.16). 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 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 this chapter, 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, $\pi(\theta|d(x_{1:n}, x_{1:n}) < R)$, 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, Eq. (1.21), similarly to the work of Holmes and Walker [2017] and Grünwald [2016]

While this work is exciting and interesting these methods do not directly answer the questions we ask in this thesis. The way we talk about outliers in Section 2.1 frames them as a fault with the likelihood function for the data, rather than the data itself. Namely that 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 with the 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.

3.3 Likelihood Stability

A natural extension to the work of Smith and Rigat [2012]; Gustafson and Wasserman [1995] 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 and discovers that the data can cause divergence in terms of De Robertis distance between the functioning and genuine posterior produced from different likelihoods. Beyond these initial results, work investigating the stability of Bayesian learning across a neighbourhood of likelihood models is limited. I believe this to be a consequence of the $\mathcal{M}$-CLOSED world assumption. In the $\mathcal{M}$-CLOSED world of controlled experimental conditions it is reasonable to consider the prior as the only subjective element of the analysis whose sensitivity must be checked. However in the $\mathcal{M}$-OPEN world it is now also reasonable to consider the likelihood in this way. Henceforth we acknowledge that the model forms part of the subjective prior specification of the analyses. In this chapter we take inspiration from the results of Smith and Rigat [2012]; Gustafson and Wasserman [1995] and seek to analyse the inherent stability of the Bayesian updating machine to the choice of likelihoods.

3.3.1 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 converge 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. Two likelihoods are defined to be close if for a given set of parameters they produce similar densities for the observables

$x \in \mathbb{R}^p$, where p is the dimension of the observation space, we elaborate on this further in Definition 15. A stable Bayesian analysis must then produce similar posterior predictives for future observables $x^{'} \in \mathbb{R}^p$ given data $x_{1:n} \in \mathbb{R}^{n \times p}$, 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 of the parameter posteriors across neighbourhoods of parameter priors. 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 advantages 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. For example mixture distributions with extra mixture components can be considered within these neighbourhoods.

The next section establishes the notation I will use throughout the rest of this chapter.

3.3.2 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

$$\pi_f^D(\theta_f|x_{1:n}) = \frac{\pi_f^D(\theta_f)\exp(-\sum_{i=1}^n \ell_D(x_i, f(\cdot;\theta_f)))}{\int \pi_f^D(\theta_f)\exp(-\sum_{i=1}^n \ell_D(x_i, f(\cdot;\theta_f)))d\theta_f} \tag{3.6}$$

$$m_f^D(y|x_{1:n}) = \int f(y;\theta_f)\pi_f^D(\theta_f|x_{1:n})d\theta_f, \tag{3.7}$$

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,

$$\ell_{\text{KLD}}(x_i, f(\cdot;\theta_f))) = -\log f(x_i;\theta_f), \tag{3.8}$$

recovers standard Bayes' rule updating in Eq. (3.6) and standard one-step-ahead predictive distribution in Eq. 3.7. 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. If divergence $D(g||f(\cdot;\theta_f))$ requires a density estimate of g for its loss function we assume we have access to one that is consistent e.g. a KDE for univariate observations. Throughout this section we will use the $\cdot$ 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

$$\theta_f^D = \arg\min_{\theta_f \in \Theta_f} D(g(\cdot), f(\cdot;\theta_f)), \tag{3.9}$$

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.

3.3.3 A neighbourhood of likelihood models

First we define exactly what we mean by a neighbourhood of likelihood models. We consider two likelihood models

$$\{f(x;\theta_f) : x \in \mathcal{X} \subset \mathbb{R}^p, \theta_f \in \Theta_f\} \tag{3.10}$$
$$\{h(x;\theta_h) : x \in \mathcal{X} \subset \mathbb{R}^p, \theta_h \in \Theta_h\}, \tag{3.11}$$

for the same observables $x \in \mathcal{X}$. Defining $\Theta_U := \Theta_f \cap \Theta_h$ the intersection of the parameter spaces Θ_f and Θ_h for the two likelihood models, we then write $\Theta_f = \left\{\Theta_U, \Theta_{f\setminus h}\right\}$ and $\Theta_h = \left\{\Theta_U, \Theta_{h\setminus f}\right\}$. We therefore define a neighbourhood of likelihood models as follows.

Definition 15 (Neighbourhood of likelihood models)**.** The neighbourhood of likelihood models for observable $x \in \mathcal{X}$ is defined as

$$\begin{aligned}\mathcal{N}_\epsilon^D := &\left\{(f(\cdot;\theta_f), h(\cdot;\theta_h)) : D(f\left(\cdot;\left\{\theta_U, \theta_{f\setminus h}\right\}\right) || h\left(\cdot;\left\{\theta_U, \theta_{h\setminus f}\right\}\right)) < \epsilon,\right.\\ &\left.\text{for all values of } \theta_U \in \Theta_U, \theta_{f\setminus h} \in \Theta_{f\setminus h}, \theta_{h\setminus f} \in \Theta_{h\setminus f}\right\}\end{aligned} \tag{3.12}$$

Neighbourhood $\mathcal{N}_\epsilon^D$ demands that when we fix the shared part of their parameter spaces θ_U the likelihoods produce similar densities for x measured by divergence D for all values of the unshared parameters $\theta_{f\setminus h}$ and $\theta_{h\setminus f}$. This neighbourhood condition is unlikely to hold unless θ_f and θ_h almost entirely overlap. However, we add this notation to allow for some special cases where the likelihood models have different parameter dimensions. For example, consider that

$$f(\cdot;\theta_f) = \mathcal{N}\left(x;\mu,\sigma^2\right) \tag{3.13}$$

$$h(\cdot;\theta_h) = 0.95\times\mathcal{N}\left(x;\mu,\sigma^2\right) + 0.05\times\mathcal{N}\left(x;\mu_c,\sigma_c^2\right) \tag{3.14}$$

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

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 chapter 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 $\Theta_{f\backslash h} = \Theta_{h\backslash f} = \emptyset$. This significantly simplifies the notation required. An extension of these results to the situations where the parameter spaces are no longer equal can be found in Appendix Chapter 7.

3.3.4 The stability of the predictive distribution

First we list several important properties for the divergence targeted by inference to satisfy in order for our initial results to hold. We relax some of these in future sections.

Condition 1 (A convex divergence metric)**.** According to Definition 5 a divergence 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 $D(g,f)$ is convex in both of its arguments. That is to say that for $\lambda \in [0,1]$

$$D(\lambda g_1 + (1-\lambda)g_2, f) \leq \lambda D(g_1, f) + (1-\lambda)D(g_2, f) \tag{3.15}$$

$$D(g, \lambda f_1 + (1-\lambda)f_2) \leq \lambda D(g, f_1) + (1-\lambda)D(g, f_2). \tag{3.16}$$

For such divergences we introduce the following notation

$$D(g||f) = D_M(g,f). \tag{3.17}$$

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.

The first result relates to the stability of the posterior predictive distribution. In order to prove theorems involving the finite sample posterior predictive, Theorems 3 and 6, we require the following to hold which places conditions on the observations and the prior specification.

Condition 2 (Concentration of the posterior)**.** For divergence $D(\cdot||\cdot)$ the dataset, $x_{1:n} \sim g(\cdot)$, 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 $\pi_f^D(\theta_f|x_{1:n})$ and $\pi_h^D(\theta_h|x_{1:n})$ have concentrated to ensure

$$\int_{\Theta_f} D(g||h(\cdot;\theta_f))\pi_f^D(\theta_f|x_{1:n})d\theta_f \geq \int_{\Theta_h} D(g||h(\cdot;\theta_h))\pi_h^D(\theta_h|x_{1:n})d\theta_h \tag{3.18}$$

$$\int_{\Theta_h} D(g||f(\cdot;\theta_h))\pi_h^D(\theta_h|x_{1:n})d\theta_h \geq \int_{\Theta_f} D(g||f(\cdot;\theta_f))\pi_f^D(\theta_f|x_{1:n})d\theta_f. \tag{3.19}$$

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] (Eq. (2.15)) 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] (Eq. (2.15)) 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 3 (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 $x_{1:n}$ such that Condition 2 holds for divergence $D_M(\cdot,\cdot)$
- For the two likelihood models $\{f(\cdot;\theta_f) : \theta_f \in \Theta_f\}$ and $\{h(\cdot;\theta_h) : \theta_h \in \Theta_h\}$ there exists $\epsilon > 0$ such that $f, h \in \mathcal{N}_\epsilon^{D_M}$ as defined in Definition 15.

Then for $m_f^{D_M}$ and $m_h^{D_M}$ as defined in Eq. (3.7)

$$D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n}))) \leq R^{D_M}(g, f, h, x_{1:n}) + \epsilon, \tag{3.20}$$

where

$$\begin{aligned} R^{D_M}(g, f, h, x_{1:n}) := 2\min\Bigg\{ &\int \left(D_M(g, f(\cdot;\theta_f))\right) \pi_f^{D_M}(\theta_f|x_{1:n}) d\theta_f, \qquad (3.21)\\ &\int \left(D_M(g, h(\cdot;\theta_h))\right) \pi_h^{D_M}(\theta_h|x_{1:n}) d\theta_h \Bigg\}. \end{aligned}$$

Theorem 3 demonstrates that general Bayesian inference using a proper divergence **metric** produces stable posterior predictive inferences, where stability is measured with respect to that divergence, providing the remainder term,

$R^{D_M}(g, f, h, x_{1:n})$, defined in Eq. (3.21) is small, and the priors and data are sufficient for Condition 2 to hold for that divergence metric. 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.

Unlike the prior misspecification case considered by Gustafson and Wasserman [1995], when the model is wrong 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. Next we prove Theorem 3

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

$$\psi(\mathbb{E}_X[\rho(X)]) \leq \mathbb{E}_X[\psi(\rho(X))]. \tag{3.22}$$

Consider applying this with θ_f as the random variable of interest with distribution $\pi_f^{D_M}(\theta_f|x_{1:n})$, $\rho(\theta) = f(y; \theta)$ 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 the Bayesian predictive distribution is defined,

$$m_f^{D_M}(z|x_{1:n}) = \mathbb{E}_{\pi_f^{D_M}(\theta_f|x_{1:n})}[f(z; \theta_f)] = \int f(z; \theta_f)\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f < \infty, \quad \forall z \tag{3.23}$$

and that

$$\mathbb{E}_{\pi_f^{D_M}(\theta_f|x_{1:n})}[D_M(h(\cdot), f(\cdot; \theta_f))] = \int D_M(h(\cdot), f(\cdot; \theta_f))\pi_f^{D_M}(\theta_f|x_{1:n})d\theta < \infty. \tag{3.24}$$

We note that by symmetry we could exchange f for h above. Therefore, by the convexity of $D_M(\cdot, \cdot)$, Jensen's inequality can be applied as described above, first to $m_h^{D_M}(\cdot|x_{1:n})$ and then to $m_f^{D_M}(\cdot|x_{1:n})$. Therefore,

$$D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n}))) \leq \int D_M(m_f^{D_M}(\cdot|x_{1:n}), h(\cdot; \theta_h))\pi_h^{D_M}(\theta_h|x_{1:n})d\theta_h \tag{3.25}$$

$$\leq \int \left\{ \int D_M(f(\cdot; \theta_f), h(\cdot; \theta_h))\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f \right\} \pi_h^{D_M}(\theta_h|x_{1:n})d\theta_h. \tag{3.26}$$

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

$$D_M(f,h) \leq D_M(f,g) + D_M(g,h) = D_M(g,f) + D_M(g,h), \tag{3.27}$$

which can be used to show that

$$\begin{aligned}
&D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n}))) \\
&\quad \leq \int \left\{ \int D_M(f(\cdot;\theta_f), h(\cdot;\theta_h))\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f \right\} \pi_h^{D_M}(\theta_h|_{1:n})d\theta_h \\
&\quad \leq \int \left\{ \int D_M(g, f(\cdot;\theta_f)) + D_M(g, h(\cdot;\theta_h))\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f \right\} \pi_h^{D_M}(\theta_h|x_{1:n})d\theta_h
\end{aligned} \tag{3.28}$$

$$\quad = \int D_M(g, f(\cdot;\theta_f))\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f + \int D_M(g, h(\cdot;\theta_h))\pi_h^{D_M}(\theta_h|x_{1:n})d\theta_h. \tag{3.29}$$

Now given the first part of Condition 2, equation (3.18)

$$\begin{aligned}
&D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n}))) \\
&\leq \int D_M(g, f(\cdot;\theta_f))\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f + \int D_M(g, h(\cdot;\theta_h))\pi_h^{D_M}(\theta_h|x_{1:n})d\theta_h \\
&\leq \int D_M(g, f(\cdot;\theta_f))\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f + \int D_M(g, h(\cdot;\theta_f))\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f \qquad (3.30) \\
&= \int \left(D_M(g, f(\cdot;\theta_f)) + D_M(g, h(\cdot;\theta_f))\right)\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f. \qquad (3.31)
\end{aligned}$$

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

$$\begin{aligned}
&D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n}))) \\
&\quad \leq \int \left(D_M(g, f(\cdot;\theta_f)) + D_M(g, h(\cdot;\theta_f))\right)\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f \\
&\quad = \int (D_M(g, f(\cdot;\theta_f)) + D_M(g, h(\cdot;\theta_f)) \qquad (3.32) \\
&\qquad\quad - D_M(f(\cdot;\theta_f), h(\cdot;\theta_f)) + D_M(f(\cdot;\theta_f), h(\cdot;\theta_f)))\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f.
\end{aligned}$$

Finally applying the triangle inequality once more gives us that

$$D_M(g,f) + D_M(f,h) \geq D_M(g,h) \Rightarrow D_M(g,f) \geq D_M(g,h) - D_M(f,h) \tag{3.33}$$

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

to show that

$$
\begin{aligned}
D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n}))) &\leq \int (D_M(g, f(\cdot;\theta_f)) + D_M(g, h(\cdot;\theta_f)) \\
&\quad - D_M(f(\cdot;\theta_f), h(\cdot;\theta_f)) + D_M(f(\cdot;\theta_f), h(\cdot;\theta_f)))\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f \\
&\leq \int (2D_M(g, f(\cdot;\theta_f)) + \epsilon)\, \pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f \qquad (3.34) \\
&= 2\int (D_M(g, f(\cdot;\theta_f)))\, \pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f + \epsilon. \qquad (3.35)
\end{aligned}
$$

We note that we could have applied the second part of Condition 2, equation (3.19), to exchange θ_f for θ_h in line (3.30) and the triangle inequality also gives us that

$$D_M(g, h) + D_M(f, h) \geq D_M(g, f) \Rightarrow D_M(g, h) \geq D_M(g, f) - D_M(f, h). \qquad (3.36)$$

Which, in turn can be used to show that

$$D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n}))) \leq 2\int (D_M(g, h(\cdot;\theta_h)))\, \pi_h^{D_M}(\theta_h|x_{1:n})d\theta_h + \epsilon \qquad (3.37)$$

and thus

$$D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n}))) \leq R^{D_M}(g, f, h, x_{1:n}) + \epsilon. \qquad (3.38)$$

where $R^{D_M}(g, f, h, x_{1:n})$ is defined in Eq. 3.21. □

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 4 in the next 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 $x_{1:n}$ such that Condition 2 holds for divergence $D_M(\cdot,\cdot)$

- For the two likelihood models $\{f(\cdot;\theta_f):\theta_f\in\Theta_f\}$ and $\{h(\cdot;\theta_h):\theta_h\in\Theta_h\}$ there exists $\epsilon>0$ such that $f,h\in\mathcal{N}_\epsilon^{D_M}$ as defined in Definition 15.

Then for $m_f^{D_M}$ and $m_f^{D_M}$ as defined in Eq. (3.7) as $n\to\infty$

$$D_M(m_f^{D_M}(\cdot|x_{1:n}),m_h^{D_M}(\cdot|x_{1:n})))\leq\epsilon \tag{3.39}$$

almost surely.

Therefore, if we can specify a model that is close to the DGP in either TVD or HD (or any other metric), then Bayesian updating aimed at minimising the same divergence will produce posterior inferences from the approximate model that are no further from the posterior inferences that would have resulted from using 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 3 provides some idea of the tightness of the bounds in this theorem. Next we prove Corollary 1.

Proof. From Theorem 3 we know that

$$D_M(m_f^{D_M}(\cdot|x_{1:n}),m_h^{D_M}(\cdot|x_{1:n})))\leq\epsilon \quad + \\ 2\min\left\{\int\left(D_M(g,f(\cdot;\theta_f))\right)\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f,\int\left(D_M(g,h(\cdot;\theta_h))\right)\pi_h^{D_M}(\theta_h|x_{1:n})d\theta_h\right\}$$

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. (2.15)) hold, and once again assuming that both g and $f(y;\theta)$ are absolutely continuous, implies there exists n such that

$$D_M(m_f^{D_M}(\cdot|x_{1:n}),m_h^{D_M}(\cdot|x_{1:n})))\leq 2\int\left(D(g,f(\cdot;\theta_f))\right)\pi_f^D(\theta_f|x_{1:n})d\theta_f+\epsilon, \tag{3.40}$$

almost surely. Following the asymptotics of Walker [2013] define

$$A_{\delta,f}^D=\{\theta:D(g(\cdot),f(\cdot;\theta))\leq\delta\}, \tag{3.41}$$

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

$$
\begin{aligned}
D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n}))) &\leq 2\int \left(D_M(g, f(\cdot;\theta_f))\right)\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f + \epsilon \\
&= 2\int_{A_{\delta,f}^{D_M}} \left(D_M(g, f(\cdot;\theta_f))\right)\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f \qquad (3.42)\\
&\quad + 2\int_{{A_{\delta,f}^{D_M}}^c} \left(D_M(g, f(\cdot;\theta_f))\right)\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f + \epsilon
\end{aligned}
$$

Now by the definition of $A_{\delta,f}^{D_M}$, for all values of $\theta_f \in A_{\delta,f}^{D_M}$, $D_M(g, f(\cdot;\theta_f)) < \delta$ 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 $\int_{{A_{\delta,f}^{D_M}}^c} \left(D_M(g, f(\cdot;\theta_f))\right)\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f \leq b\Pi_f^{D_M}({A_{\delta,f}^{D_M}}^c|x_{1:n})$, where $\Pi_f^{D_M}({A_{\delta,f}^{D_M}}^c|x_{1:n})$ is the probability of being in the set ${A_{\delta,f}^{D_M}}^c$ under the posterior $\pi_f^{D_M}(\theta_f|x_{1:n})$. Therefore,

$$
\begin{aligned}
D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n}))) &\leq 2\int_{A_{\delta,f}^{D_M}} \left(D_M(g, f(\cdot;\theta_f))\right)\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f \\
&\quad + 2\int_{{A_{\delta,f}^{D_M}}^c} \left(D_M(g, f(\cdot;\theta_f))\right)\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f + \epsilon \\
&\leq 2\delta + 2b\Pi_f^{D_M}({A_{\delta,f}^{D_M}}^c|x_{1:n}) + \epsilon. \qquad (3.43)
\end{aligned}
$$

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. (2.15)) of the posterior [Chernozhukov and Hong, 2003; Lyddon et al., 2018], and since this holds for all δ we have that

$$
D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n}))) \leq \epsilon. \qquad (3.44)
$$

□

3.3.5 Stability in the limit

While the above results provide a bound on how far apart the predictives from two likelihood models in the same neighbourhood are when learning using a proper divergence metric $D_M(\cdot,\cdot)$, it is not clear what happens as $\epsilon \to 0$ unless the neighbourhood contains the DGP (Corollary 1). Theorem 4 provides a limiting stability result that only depends on ϵ.

Theorem 4 (Limiting predictive stability using divergence metrics)**.** Consider the following conditions:

- Divergence $D_M(\cdot,\cdot)$ satisfies M1 and M2 from Condition 1
- For the two likelihood models $\{f(\cdot;\theta_f) : \theta_f \in \Theta_f\}$ and $\{h(\cdot;\theta_h) : \theta_h \in \Theta_h\}$ there exists $\epsilon > 0$ such that $f, h \in \mathcal{N}_\epsilon^{D_M}$ as defined in Definition 15.

Then

$$\left|D_M(g, f(\cdot;\hat{\theta}_f^{D_M})) - D_M(g, h(\cdot;\hat{\theta}_h^{D_M}))\right| \leq \epsilon \tag{3.45}$$

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

Theorem 4 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.

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 $\theta \in \Theta$,

$$\begin{aligned} D_M(g, f(\cdot;\theta)) &\leq D_M(h(\cdot;\theta), f(\cdot;\theta)) + D_M(g, h(\cdot;\theta)) && (3.46)\\ &\leq \epsilon + D_M(g, h(\cdot;\theta)) && (3.47)\\ D_M(g, h(\cdot;\theta)) &\leq D_M(h(\cdot;\theta), f(\cdot;\theta)) + D_M(g, f(\cdot;\theta)) && (3.48)\\ &\leq \epsilon + D_M(g, f(\cdot;\theta)). && (3.49) \end{aligned}$$

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

$$\begin{aligned} D_M(g, f(\cdot;\hat{\theta}_f^{D_M})) \leq D_M(g, f(\cdot;\hat{\theta}_h^{D_M})) \leq \epsilon + D_M(g, h(\cdot;\hat{\theta}_h^{D_M})) && (3.50)\\ D_M(g, h(\cdot;\hat{\theta}_h^{D_M})) \leq D_M(g, h(\cdot;\hat{\theta}_f^{D_M})) \leq \epsilon + D_M(g, f(\cdot;\hat{\theta}_f^{D_M})) && (3.51)\\ \Rightarrow \left|D_M(g, h(\cdot;\hat{\theta}_h^{D_M})) - D_M(g, f(\cdot;\hat{\theta}_f^{D_M}))\right| \leq \epsilon. && (3.52) \end{aligned}$$

□

3.4 Approximating Metrics

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 are the TVD and the HD. We discussed in Section 2.3 the desirability of learning using the TVD from a decision theoretic point of view. 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 is small on the natural scale. We believe that although the DM is never going to be able to exactly elicit a full likelihood model, they may well be able to elicit judgements that are accurate on the natural scale. We juxtapose this with the unreasonable requirement of increasingly accurate estimation of tail probabilities in order to guarantee stability of traditional Bayesian updating in the next section. 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, $g_n(x)$. Although we identify in Section 2.6.8 that this is an ongoing area of research, current methods struggle to scale to high dimensions. 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 robust βD.

3.4.1 The KLD

While inference targeting metrics is inconvenient to implement in practise, inference targeting the KLD, using 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

$$\hat{\theta}_f^{\text{KLD}} = \arg\min_{\theta} \text{KLD}(g, f(\cdot;\theta)) \tag{3.53}$$

$$\hat{\theta}_h^{\text{KLD}} = \arg\min_{\theta} \text{KLD}(g, h(\cdot;\theta)), \tag{3.54}$$

we have that for all data generating densities g

$$\begin{aligned}
&\left|\textsc{kld}(g, f(\cdot;\hat{\theta}_f^{\textsc{kld}})) - \textsc{kld}(g, h(\cdot;\hat{\theta}_h^{\textsc{kld}}))\right|\\
&\leq \max\left\{\int g\log\frac{h(\cdot;\hat{\theta}_h^{\textsc{kld}})}{f(\cdot;\hat{\theta}_h^{\textsc{kld}})}dx, \int g\log\frac{f(\cdot;\hat{\theta}_f^{\textsc{kld}})}{h(\cdot;\hat{\theta}_f^{\textsc{kld}})}dx\right\}. \qquad (3.55)
\end{aligned}$$

Proof. The definition of the KLD (Eq. (1.5)) provides that

$$\textsc{kld}(g, f(\cdot;\theta_f)) = \textsc{kld}(g, h(\cdot;\theta_h)) + \int g\log\frac{h(\cdot;\theta_h)}{f(\cdot;\theta_f)}dx. \qquad (3.56)$$

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

$$\begin{aligned}
\textsc{kld}(g, f(\cdot;\hat{\theta}_f^{\textsc{kld}})) &\leq \textsc{kld}(g, f(\cdot;\hat{\theta}_h^{\textsc{kld}})) && (3.57)\\
&= \textsc{kld}(g, h(\cdot;\hat{\theta}_h^{\textsc{kld}})) + \int g\log\frac{h(\cdot;\hat{\theta}_h^{\textsc{kld}})}{f(\cdot;\hat{\theta}_h^{\textsc{kld}})}dx && (3.58)\\
\textsc{kld}(g, h(\cdot;\hat{\theta}_h^{\textsc{kld}})) &\leq \textsc{kld}(g, h(\cdot;\hat{\theta}_f^{\textsc{kld}})) && (3.59)\\
&= \textsc{kld}(g, f(\cdot;\hat{\theta}_f^{\textsc{kld}})) + \int g\log\frac{f(\cdot;\hat{\theta}_f^{\textsc{kld}})}{h(\cdot;\hat{\theta}_f^{\textsc{kld}})}dx. && (3.60)
\end{aligned}$$

Combining these two inequalities results in Eq. (3.55). □

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, proven using Bernoulli's inequality [Bernoulli, 1689], associated with the natural logarithm are

$$1 - \frac{1}{y} \leq \log y \leq y - 1 \qquad (3.61)$$

which enables us to bound this remainder term

$$\begin{aligned}
\int g\log\frac{h(\cdot;\hat{\theta}_h^{\textsc{kld}})}{f(\cdot;\hat{\theta}_h^{\textsc{kld}})}dx &\leq \int gh(\cdot;\hat{\theta}_h^{\textsc{kld}}) + \frac{g}{f(\cdot;\hat{\theta}_h^{\textsc{kld}})}dx && (3.62)\\
\int g\log\frac{f(\cdot;\hat{\theta}_f^{\textsc{kld}})}{h(\cdot;\hat{\theta}_f^{\textsc{kld}})}dx &\leq \int gf(\cdot;\hat{\theta}_f^{\textsc{kld}}) + \frac{g}{h(\cdot;\hat{\theta}_f^{\textsc{kld}})}dx. && (3.63)
\end{aligned}$$

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.

$$\int g \log \frac{h(\cdot;\theta_h)}{f(\cdot;\theta_h)} dx \leq \int \frac{g}{h(\cdot;\theta_h)} h(\cdot;\theta_h) \log \frac{h(\cdot;\theta_h)}{f(\cdot;\theta_h)} dx \tag{3.64}$$

$$\leq M_h^* \text{KLD}(h(\cdot;\theta_h)||f(\cdot;\theta_h)) \tag{3.65}$$

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 $|\log(h(\cdot;\theta_h)) - \log(f(\cdot;\theta_f))|$ is small in regions where g has mass. Without knowledge of g, this requires that $|\log(h(\cdot;\theta_h)) - \log(f(\cdot;\theta_f))|$ is 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 [Winkler and Murphy, 1968; O'Hagan et al., 2006].

We do however note that this notion of stability is with respect to a metric that we have already shown in this thesis 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 4 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 $D(g||h(\cdot;\hat{\theta}_h^D)) \leq D(g||h(\cdot;\hat{\theta}_f^D))$.

3.4.2 The βD

We established 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 a series of lemmas showing how the βD relates to the TVD in a similar fashion to the triangle inequality associated with metric divergences. Although Bregman divergences, introduced in Eq. (1.26), are not generally metrics, they do posses the following "three-point property" [Cichocki and Amari, 2010]

Lemma 2 (Three-point property of Bregman Divergences)**.** For Bregman divergence $D_\psi(g||f)$ defined in Eq. (1.26) the following generalisation of the triangle inequality holds

$$D_\psi(g||f) + D_\psi(f||h) = D_\psi(g||h) + (g - f)\left(\nabla\psi(h) - \nabla\psi(f)\right), \tag{3.66}$$

where $\nabla\psi(\cdot)$ is the first derivative of the function $\psi(\cdot)$.

Proof. Following the definition of a Bregman divergence Eq. (1.26)

$$\begin{aligned}
&D_\psi(g||f) + D_\psi(f||h)\\
=&\psi(g) - \psi(f) - (g - f)\nabla\psi(f) + \psi(f) - \psi(h) - (f - h)\nabla\psi(h) &(3.67)\\
=&\psi(g) - \psi(h) - (-h)\nabla\psi(h) - (g - f)\nabla\psi(f) - (f)\nabla\psi(h) &(3.68)\\
=&\psi(g) - \psi(h) - (g - h)\nabla\psi(h) - (g - f)\nabla\psi(f) - (f - g)\nabla\psi(h) &(3.69)\\
=&D_\psi(g||h) + (g - f)\left(\nabla\psi(h) - \nabla\psi(f)\right) &(3.70)
\end{aligned}$$

□

Applying Lemma 2 specifically for the βD provides the following lemma.

Lemma 3 (Three-point property of the βD)**.** The following relationship for the βD holds for densities g, f and h

$$D_B^{(\beta)}(f||h) = D_B^{(\beta)}(g||h) - D_B^{(\beta)}(g||f) + R(g||f||h) \tag{3.71}$$

$$R(g||f||h) = \int (g - f)\left(\frac{1}{\beta - 1}h^{\beta-1} - \frac{1}{\beta - 1}f^{\beta-1}\right)d\mu \tag{3.72}$$

Proof. This follows directly from Lemma 2 and the definition of the βD. □

Next 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 Chapters 4 and 5. The next result also relies on being able to bound the essential supremum (ess sup) of densities g, f and h. We explained exactly what we mean by the ess sup in Section 2.6.4.

Lemma 4 (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

$$D_B^{(\beta)}(g||h) \leq D_B^{(\beta)}(g||f) + \frac{M^{\beta-1}}{\beta-1}\text{TVD}(h, f) \tag{3.73}$$

The symmetry of the TVD ensures Lemma 4 will also hold if we swap h and f. 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 thesis has desirable 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. Next we prove Lemma 4.

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

$$\begin{aligned}\text{TVD}(f, h) &= \frac{1}{2}\int_{A^+} (h(x) - f(x))\, dx + \frac{1}{2}\int_{A^-} (f(x) - h(x))\, dx && (3.74)\\ &= \int_{A^+} (h(x) - f(x))\, dx = \int_{A^-} (f(x) - h(x))\, dx. && (3.75)\end{aligned}$$

To see this consider $L_{f,h} : \mathcal{X} \to \mathbb{R}$ with $L_{f,h}(x) := \min(f(x), h(x))$ as the lower of the two probability densities for every x. Given that both f and h are probability densities an thus integrate to 1 we have that

$$\begin{aligned}\int_{A^+} (h(x) - f(x))dx &= 1 - \int L_{f,h}(x)dx && (3.76)\\ \int_{A^-} (f(x) - h(x))dx &= 1 - \int L_{f,h}(x)dx. && (3.77)\end{aligned}$$

The two right hand sides are identical and therefore the two left hand sides must be equal. By the definition of the βD we can rearrange

$$\begin{aligned}
&D_B^{(\beta)}(g||h)\\
=&D_B^{(\beta)}(g||f) + \left(\int \left[\frac{1}{\beta}h(x)^\beta - \frac{1}{\beta}f(x)^\beta - \frac{1}{\beta-1}g(x)h(x)^{\beta-1} + \frac{1}{\beta-1}g(x)f(x)^{\beta-1}\right] dx\right) \qquad (3.78)\\
=&D_B^{(\beta)}(g||f) + \left(\frac{1}{\beta}\int \left[h(x)^\beta - f(x)^\beta\right] dx + \frac{1}{\beta-1}\int g(x)\left(f(x)^{\beta-1} - h(x)^{\beta-1}\right) dx\right) \qquad (3.79)
\end{aligned}$$

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

$$\begin{aligned}
&D_B^{(\beta)}(g||h)\\
=&D_B^{(\beta)}(g||f) + \frac{1}{\beta}\int \left[h(x)^\beta - f(x)^\beta\right] dx + \frac{1}{\beta-1}\int g(x)\left(f(x)^{\beta-1} - h(x)^{\beta-1}\right) dx\\
\leq&D_B^{(\beta)}(g||f) + \frac{1}{\beta}\int_{A^+} \left[h(x)^\beta - f(x)^\beta\right] dx + \frac{1}{\beta-1}\int_{A^-} g(x)\left(f(x)^{\beta-1} - h(x)^{\beta-1}\right) dx. \qquad (3.80)
\end{aligned}$$

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

$$\begin{aligned}
&D_B^{(\beta)}(g||h)\\
\leq&D_B^{(\beta)}(g||f) + \frac{1}{\beta}\int_{A^+} \left[h(x)^\beta - f(x)^\beta\right] dx + \frac{1}{\beta-1}\int_{A^-} g(x)\left(f(x)^{\beta-1} - h(x)^{\beta-1}\right) dx.\\
\leq&D_B^{(\beta)}(g||f) + \frac{1}{\beta}\int_{A^+} h(x)\left(h(x)^{\beta-1} - f(x)^{\beta-1}\right) dx\\
&+\frac{1}{\beta-1}\int_{A^-} g(x)\left(f(x)^{\beta-1} - h(x)^{\beta-1}\right) dx \qquad (3.81)\\
=&D_B^{(\beta)}(g||f) + \frac{1}{\beta}\int_{A^+} h(x)^\beta\left(1 - \frac{f(x)^{\beta-1}}{h(x)^{\beta-1}}\right) dx\\
&+\frac{1}{\beta-1}\int_{A^-} g(x)f(x)^{\beta-1}\left(1 - \frac{h(x)^{\beta-1}}{f(x)^{\beta-1}}\right) dx. \qquad (3.82)
\end{aligned}$$

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

$$\left(1 - \frac{f(x)^{\beta-1}}{h(x)^{\beta-1}}\right) \leq \left(1 - \frac{f(x)}{h(x)}\right) \tag{3.83}$$

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

$$\begin{aligned}
&D_B^{(\beta)}(g||h)\\
\leq &D_B^{(\beta)}(g||f) + \frac{1}{\beta}\int_{A^+} h(x)^{\beta}\left(1 - \frac{f(x)^{\beta-1}}{h(x)^{\beta-1}}\right)dx\\
&+ \frac{1}{\beta-1}\int_{A^-} g(x)f(x)^{\beta-1}\left(1 - \frac{h(x)^{\beta-1}}{f(x)^{\beta-1}}\right)dx\\
\leq &D_B^{(\beta)}(g||f) + \frac{1}{\beta}\int_{A^+} h(x)^{\beta}\left(1 - \frac{f(x)}{h(x)}\right)dx + \frac{1}{\beta-1}\int_{A^-} g(x)f(x)^{\beta-1}\left(1 - \frac{h(x)}{f(x)}\right)dx \quad (3.84)\\
= &D_B^{(\beta)}(g||f) + \frac{1}{\beta}\int_{A^+} h(x)^{\beta-1}\left(h(x) - f(x)\right)dx\\
&+ \frac{1}{\beta-1}\int_{A^-} g(x)f(x)^{\beta-2}\left(f(x) - h(x)\right)dx \quad (3.85)
\end{aligned}$$

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

$$\begin{aligned}
&D_B^{(\beta)}(g||h)\\
= &D_B^{(\beta)}(g||f) + \frac{1}{\beta}\int_{A^+} h(x)^{\beta-1}\left(h(x) - f(x)\right)dx\\
&+ \frac{1}{\beta-1}\int_{A^-} g(x)f(x)^{\beta-2}\left(f(x) - h(x)\right)dx \quad (3.86)\\
\leq &D_B^{(\beta)}(g||f) + \frac{M^{\beta-1}}{\beta}\int_{A^+}\left(h(x) - f(x)\right)dx + \frac{M^{\beta-1}}{\beta-1}\int_{A^-}\left(f(x) - h(x)\right)dx \quad (3.87)\\
= &D_B^{(\beta)}(g||f) + \frac{M^{\beta-1}}{\beta}\text{TVD}(h,f) + \frac{M^{\beta-1}}{\beta-1}\text{TVD}(h,f) \quad (3.88)\\
= &D_B^{(\beta)}(g||f) + \frac{M^{\beta-1}}{\beta-1}\text{TVD}(h,f). \quad (3.89)
\end{aligned}$$

□

Lemma 4 can now be used to prove a form of limiting stability for inference using the βD.

Stability in the limit

Using Lemma 4 we firstly seek to bound the absolute distance between the βD of each of the limiting predictive distribution produced from two likelihood models within $\mathcal{N}_\epsilon^{\text{TVD}}$ from the DGP.

Theorem 5 (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

$$\max\{\operatorname{ess\,sup} f, \operatorname{ess\,sup} h, \operatorname{ess\,sup} g\} \leq M < \infty \tag{3.90}$$

- For the two likelihood models $\{f(\cdot;\theta_f) : \theta_f \in \Theta_f\}$ and $\{h(\cdot;\theta_h) : \theta_h \in \Theta_h\}$ there exists $\epsilon > 0$ such that $f, h \in \mathcal{N}_\epsilon^{D_M}$ as defined in Definition 15.

Then

$$\left| D_B^{(\beta)}(g||f(\cdot;\hat{\theta}_h^{(\beta)})) - D_B^{(\beta)}(g||h(\cdot;\hat{\theta}_f^{(\beta)}))\right| \leq \frac{M^{\beta-1}}{\beta-1}\epsilon \tag{3.91}$$

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 5 shows that for two likelihood models in the neighbourhood

$$\{f(\cdot;\theta_f), h(\cdot;\theta_h)\} \in \mathcal{N}_\epsilon^{\text{TVD}}, \tag{3.92}$$

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. Additionally, similarly to Theorem 4, Theorem 5 hold without any conditions on the DGP besides bounding its essential supremum. We paid particular attention to being able to define the a priori 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 in Section 3.4. Next we prove Theorem 5.

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.

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

Now using the triangle type inequality proven in Lemma 4 and the definition of $\mathcal{N}_\epsilon^{D_M}$ we can show that

$$\begin{aligned}
&D_B^{(\beta)}(g, f(\cdot;\hat{\theta}_h^{(\beta)}))\\
\leq&\frac{M^{\beta-1}}{\beta-1}\text{TVD}(f(\cdot;\hat{\theta}_h^{(\beta)}), h(\cdot;\hat{\theta}_h^{(\beta)})) + D_B^{(\beta)}(g, h(\cdot;\hat{\theta}_h^{(\beta)}))\\
\leq&\frac{M^{\beta-1}}{\beta-1}\epsilon + D_B^{(\beta)}(g, h(\cdot;\hat{\theta}_h^{(\beta)})) \qquad (3.93)\\
&D_B^{(\beta)}(g, h(\cdot;\hat{\theta}_f^{(\beta)}))\\
\leq&\frac{M^{\beta-1}}{\beta-1}\text{TVD}(f(\cdot;\hat{\theta}_f^{(\beta)}), h(\cdot;\hat{\theta}_f^{(\beta)})) + D_B^{(\beta)}(g, f(\cdot;\hat{\theta}_f^{(\beta)}))\\
\leq&\frac{M^{\beta-1}}{\beta-1}\epsilon + D_B^{(\beta)}(g, f(\cdot;\hat{\theta}_f^{(\beta)})) \qquad (3.94)
\end{aligned}$$

Combining these two, results in

$$\Rightarrow \left|D_B^{(\beta)}(g, h(\cdot;\hat{\theta}_h^{(\beta)})) - D_B^{(\beta)}(g, f(\cdot;\hat{\theta}_f^{(\beta)}))\right| \leq \frac{M^{\beta-1}}{\beta-1}\epsilon \qquad (3.95)$$

□

Stability of the posterior predictives

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 $\mathcal{N}_\epsilon^{\text{TVD}}$. In order to prove the stability of the posterior predictives in the same vein as Theorem 4 we require one 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

$$D_B^{(\beta)}(\lambda g_1 + (1-\lambda)g_2, f) \leq \lambda D_B^{(\beta)}(g_1, f) + (1-\lambda)D_B^{(\beta)}(g_2, f) \qquad (3.96)$$
$$D_B^{(\beta)}(g, \lambda f_1 + (1-\lambda)f_2) \leq \lambda D_B^{(\beta)}(g, f_1) + (1-\lambda)D_B^{(\beta)}(g, f_2) \qquad (3.97)$$

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

$$(\lambda x_1 + (1-\lambda)x_2)^p \leq \lambda x_1^p + (1-\lambda)x_2^p \tag{3.98}$$

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

$$\begin{aligned}
&D_B^{(\beta)}(\lambda g_1 + (1-\lambda)g_2||f)\\
=&\int \frac{1}{\beta(\beta-1)}\left(\lambda g_1 + (1-\lambda)g_2\right)^\beta + \frac{1}{\beta}f^\beta - \frac{1}{\beta-1}\left(\lambda g_1 + (1-\lambda)g_2\right)f^{\beta-1}d\mu \qquad (3.99)\\
\leq&\int \frac{1}{\beta(\beta-1)}\left(\lambda g_1^\beta + (1-\lambda)g_2^\beta\right) + \frac{1}{\beta}f^\beta - \frac{1}{\beta-1}\left(\lambda g_1 + (1-\lambda)g_2\right)f^{\beta-1}d\mu \qquad (3.100)\\
=&\lambda D_B^{(\beta)}(g_1||f) + (1-\lambda)D_B^{(\beta)}(g_2||f).
\end{aligned}$$

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 $D_B^{(\beta)}(g||f_2) < \infty$

$$\begin{aligned}
&D_B^{(\beta)}(g||\lambda f_1 + (1-\lambda)f_2)\\
=&\int \frac{1}{\beta(\beta-1)}g^\beta + \frac{1}{\beta}\left(\lambda f_1 + (1-\lambda)f_2\right)^\beta - \frac{1}{\beta-1}g\left(\lambda f_1 + (1-\lambda)f_2\right)^{\beta-1}d\mu \qquad (3.101)\\
\leq&\int \frac{1}{\beta(\beta-1)}g^\beta + \frac{1}{\beta}\left(\lambda f_1^\beta + (1-\lambda)f_2^\beta\right) - \frac{1}{\beta-1}g\left(\lambda f_1^{\beta-1} + (1-\lambda)f_2^{\beta-1}\right)d\mu \qquad (3.102)\\
=&\lambda D_B^{(\beta)}(g||f_1) + (1-\lambda)D_B^{(\beta)}(g||f_2)
\end{aligned}$$

□

We are now able to use the convexity of the βD (Lemma 5), the triangular relationship between the βD and the TVD (Lemma 4) and the three-point property the βD (Lemma 3) to extend posterior predictive stability provided by inference targeting metrics (Theorem 3) to inference using the βD in Theorem 6.

Theorem 6 (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

$$\max\{\operatorname{ess\,sup} f, \operatorname{ess\,sup} h, \operatorname{ess\,sup} g\} \leq M < \infty, \tag{3.103}$$

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

- For the two likelihood models $\{f(\cdot;\theta_f) : \theta_f \in \Theta_f\}$ and $\{h(\cdot;\theta_h) : \theta_h \in \Theta_h\}$ there exists $\epsilon > 0$ such that $f, h \in \mathcal{N}_\epsilon^{D_M}$ as defined in Definition 15.

Then

$$\begin{aligned}
&D_B^{(\beta)}(m_f^{(\beta)}(\cdot|x_{1:n})||m_h^{(\beta)}(\cdot|x_{1:n})) && (3.104)\\
&\quad\leq \frac{M^{\beta-1}}{\beta-1}\epsilon + \int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h \\
&D_B^{(\beta)}(m_h^{(\beta)}(\cdot|x_{1:n})||m_f^{(\beta)}(\cdot|x_{1:n})) && (3.105)\\
&\quad\leq \frac{M^{\beta-1}}{\beta-1}\epsilon + \int\int R(g||h(\cdot;\theta_h)||f(\cdot;\theta_f))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h.
\end{aligned}$$

where $R(g||f||h)$ and $R(g||h||f)$ were defined in Lemma 3 to be

$$R(g||f||h) = \int (g-f)\left(\frac{1}{\beta-1}h^{\beta-1} - \frac{1}{\beta-1}f^{\beta-1}\right)d\mu \tag{3.106}$$

$$R(g||h||f) = \int (g-h)\left(\frac{1}{\beta-1}f^{\beta-1} - \frac{1}{\beta-1}h^{\beta-1}\right)d\mu. \tag{3.107}$$

Theorem 5 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 $m_h^{(\beta)}(\cdot|x_{1:n})$ and $m_f^{(\beta)}(\cdot|x_{1:n})$ provided Condition 2 holds for data $x_{1:n}$ and priors $\pi_f(\theta_f)$ and $\pi_h(\theta_h)$ and the remainder terms

$$\int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h \tag{3.108}$$

$$\int\int R(g||h(\cdot;\theta_h)||f(\cdot;\theta_f))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h \tag{3.109}$$

are small. Similarly to the remainder term in Theorem 3 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 due to relevance and practicality of considering this neighbourhood in actual applications, see Section 3.4. Next we prove Theorem 6.

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 3 to show that

$$D_B^{(\beta)}(m_f^{(\beta)}(\cdot|x_{1:n})||m_h^{(\beta)}(\cdot|x_{1:n}))) \leq \int D_B^{(\beta)}(m_f^{(\beta)}(\cdot|x_{1:n})||h(\cdot;\theta_h))\pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h \tag{3.110}$$

$$\leq \int \left\{ \int D_B^{(\beta)}(f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f \right\} \pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h. \tag{3.111}$$

Now the three-point property associated with the βD (Lemma 3) gives us that

$$D_B^{(\beta)}(f||h) = D_B^{(\beta)}(g||h) - D_B^{(\beta)}(g||f) + R(g||f||h) \tag{3.112}$$

where $R(g||f||h)$ is defined in Eq. (3.106). Using this here provides

$$\begin{aligned}
&D_B^{(\beta)}(m_f^{(\beta)}(\cdot|x_{1:n})||m_h^{(\beta)}(\cdot|x_{1:n}))) && (3.113)\\
&\leq \int \left\{ \int D_B^{(\beta)}(f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f \right\} \pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h\\
&= \int \left\{ \int \left[D_B^{(\beta)}(g||h(\cdot;\theta_h)) - D_B^{(\beta)}(g||f(\cdot;\theta_f)) \right.\right.\\
&\quad \left.\left. + R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h) \right] \pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f \right\} \pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h && (3.114)\\
&= \int D_B^{(\beta)}(g||h(\cdot;\theta_h))\pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h - \int D_B^{(\beta)}(g||f(\cdot;\theta_f))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\\
&\quad + \int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h. && (3.115)
\end{aligned}$$

Now given the first part of Condition 2, Eq. 3.18, 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

$$\begin{aligned}
&D_B^{(\beta)}(m_f^{(\beta)}(\cdot|x_{1:n})||m_h^{(\beta)}(\cdot|x_{1:n})))\\
\leq&\int D_B^{(\beta)}(g||h(\cdot;\theta_h))\pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h-\int D_B^{(\beta)}(g||f(\cdot;\theta_f))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\\
&+\int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h\\
\leq&\int D_B^{(\beta)}(g||h(\cdot;\theta_f))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f-\int D_B^{(\beta)}(g||f(\cdot;\theta_f))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\\
&+\int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h \qquad (3.116)\\
=&\int\left(D_B^{(\beta)}(g||h(\cdot;\theta_f))-\int D_B^{(\beta)}(g||f(\cdot;\theta_f))\right)\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\\
&+\int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h. \qquad (3.117)
\end{aligned}$$

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 4, Eq. 3.73

$$\begin{aligned}
&D_B^{(\beta)}(m_f^{(\beta)}(\cdot|x_{1:n})||m_h^{(\beta)}(\cdot|x_{1:n})))\\
\leq&\int\left(D_B^{(\beta)}(g||h(\cdot;\theta_f))-\int D_B^{(\beta)}(g||f(\cdot;\theta_f))\right)\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\\
&+\int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h.\\
\leq&\int\frac{M^{\beta-1}}{\beta-1}\text{TVD}(h(\cdot;\theta_f),f(\cdot;\theta_f))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\\
&+\int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h. \qquad (3.118)
\end{aligned}$$

Which given the neighbourhood of likelihood models defined by $\mathcal{N}_\epsilon^{\text{TVD}}$ in Eq. (3.12)

$$\begin{aligned}
&D_B^{(\beta)}(m_f^{(\beta)}(\cdot|x_{1:n})||m_h^{(\beta)}(\cdot|x_{1:n})))\leq\int\frac{M^{\beta-1}}{\beta-1}\text{TVD}(h(\cdot;\theta_f),f(\cdot;\theta_f))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\\
&+\int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h.\\
\leq&\frac{M^{\beta-1}}{\beta-1}\epsilon+\int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h. \qquad (3.119)
\end{aligned}$$

We note that we could have instead considered $D_B^{(\beta)}(m_h^{(\beta)}(\cdot|x_{1:n})||m_f^{(\beta)}(\cdot|x_{1:n})))$, applied the corresponding version of the three-point property of Bregman divergences, 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

$$\begin{aligned}&D_B^{(\beta)}(m_h^{(\beta)}(\cdot|x_{1:n})||m_f^{(\beta)}(\cdot|x_{1:n})))\\ \leq&\frac{M^{\beta-1}}{\beta-1}\epsilon+\int\int R(g||h(\cdot;\theta_h)||f(\cdot;\theta_f))\pi_f^{(\beta)}(\theta_f|x_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|x_{1:n})d\theta_h.\end{aligned}\tag{3.120}$$

□

3.5 Experiments

The experiments in this next section serve to demonstrate the impact of the theorems proved in the chapter. 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.

3.5.1 Poisson stability

For the first time in this thesis we now 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 3 and 4 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, Poi(λ). 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.

$$\text{Poi}_{ZI}(y;\lambda,\rho) = (1-\rho)\,\text{Poi}(y;\lambda) + \rho\mathbb{I}_{y=0},\tag{3.121}$$

with $\rho \in (0, 1)$. Additionally, unmodelled heterogeneity in the data could be modelled by the addition of a second 'contaminating' Poisson component [see e.g. Wang et al., 1996; Leroux and Puterman, 1992; Dalrymple et al., 2003; Mufudza and Erol, 2016]

$$\mathrm{Poi}_{\mathrm{mix}}(y; \lambda_1, \lambda_2, \rho) = (1 - \rho)\,\mathrm{Poi}(y; \lambda_1) + \rho\,\mathrm{Poi}(y; \lambda_2), \tag{3.122}$$

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 note that over-dispersion in count data can also be dealt with using Negative-binomial regression [Lawless, 1987]. However, it is difficult to construct a TVD neighbourhood, $\mathcal{N}^{\epsilon}_{\mathrm{TVD}}$, from Definition 15 containing the Negative-Binomial and the Poisson and therefore we do not use this model to formally demonstrate the theorems of the previous section. Nevertheless, later in this section we observe empirically that the TVD-Bayes is still stable to the selection between the Poisson and the Negative-Binomial (see Figures 3.3 and 3.4.

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[1] [Long, 1990]

Data 2 GrouseTicks - the number of ticks on the heads of 403 red grouse chicks[2] [Elston et al., 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 $\mathcal{N}^{0.1}_{\mathrm{TVD}}$ neighbourhood of the standard Poisson likelihood a priori. For the GrouseTicks dataset we set $\hat{\rho} = 0.2$. For both datasets we consider priors on $\lambda \sim \mathcal{G}(2, 2)$ and $\rho \sim \mathrm{Unif}[0, \hat{\rho}]$.

Figure 3.1 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

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

[2] available in the '*lme4*' package in R

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.

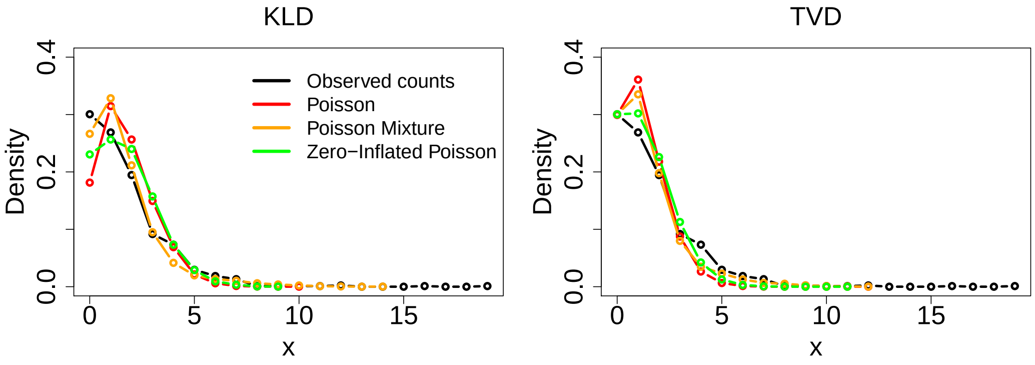

Figure 3.1: Posterior predictive mass functions for one exchangeable observation when fitting a Poisson likelihood (Poi), a two component mixture of Poissons likelihood (Poi Mix) and a zero-inflated Poisson likelihood (ZI Poi), constrained to fit within the neighbourhood $\mathcal{N}^{0.1}_{\text{TVD}}$, to the **BioChemist dataset**. **Left:** using Bayes' rule (KLD-Bayes) updating. **Right:** using updating aimed at minimising the TVD (TVD-Bayes).

These observations are reinforced by the estimates of the TVD between each of these predictive mass function which are presented in Table 3.1. 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, enforced by bounding $\hat{\rho} = 0.1$, demonstrates the result of Theorem 3. 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 diverge!

Table 3.1: 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 $\mathcal{N}^{0.1}_{\text{TVD}}$, 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	**TVD-Bayes**	Poi	Poi Mix	ZI Poi
Poi	-	0.1283	0.0748	Poi	-	0.0533	0.0589
Poi Mix	-	-	0.1317	Poi Mix	-	-	0.0684

Figure 3.2 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 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 3.2. 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 TVD between the Poisson and mixture of Poissons and the Poisson and the zero inflated Poisson are both less than 0.2, the a priori TVD been the models enforced by bounding $\hat{\rho} = 0.2$, again demonstrating the impact of Theorem 3. Additionally we note that stability under the TVD-Bayes still occurs for the GrouseTicks dataset even though all three models provide a worse approximation to the DGP than they did for the BioChemist dataset. This demonstrates the fact

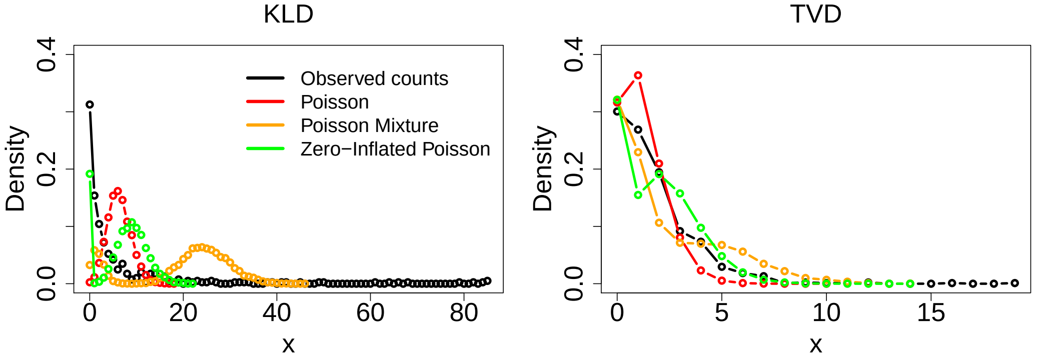

Figure 3.2: Posterior predictive mass functions for one exchangeable observation when fitting a Poisson likelihood (Poi), a two component mixture of Poissons likelihood (Poi Mix) and a zero-inflated Poisson likelihood (ZI Poi), constrained to fit within the neighbourhood $\mathcal{N}^{0.2}_{\text{TVD}}$, to the **GrouseTicks dataset**. **Left:** using Bayes' rule (KLD-Bayes) updating. **Right:** using updating aimed at minimising the TVD (TVD-Bayes).

that Theorem 3 holds independently of the DGP.

Table 3.2: 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 $\mathcal{N}^{0.2}_{\text{TVD}}$, 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

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 3.3 and 3.4 and Tables 3.3 and 3.4 show that the TVD-Bayes is still much more stable in terms of TVD than thee KLD-Bayes.

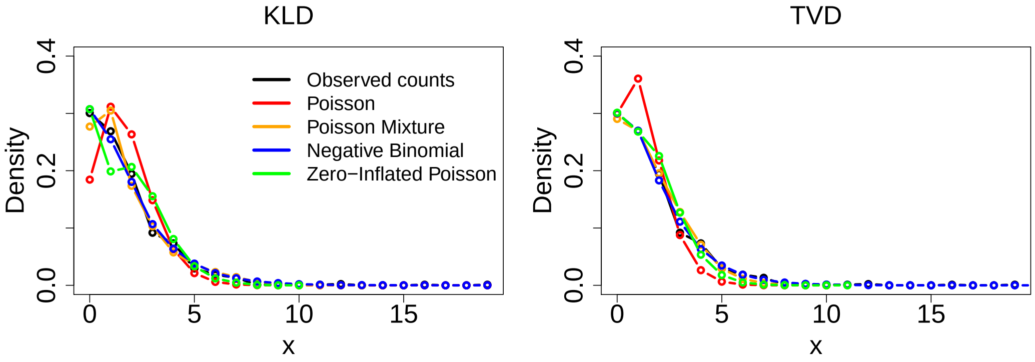

Figure 3.3: Posterior predictive mass functions for one exchangeable observation when fitting a Poisson likelihood (Poi), a two component mixture of Poissons likelihood (Poi Mix), a zero-inflated Poisson likelihood (ZI Poi) and a negative-binomial likelihood (NB), unconstrained a priori, to the **BioChemist dataset**. **Left:** using Bayes' rule (KLD-Bayes) updating. **Right:** using updating aimed at minimising the TVD (TVD-Bayes).

Table 3.3: 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

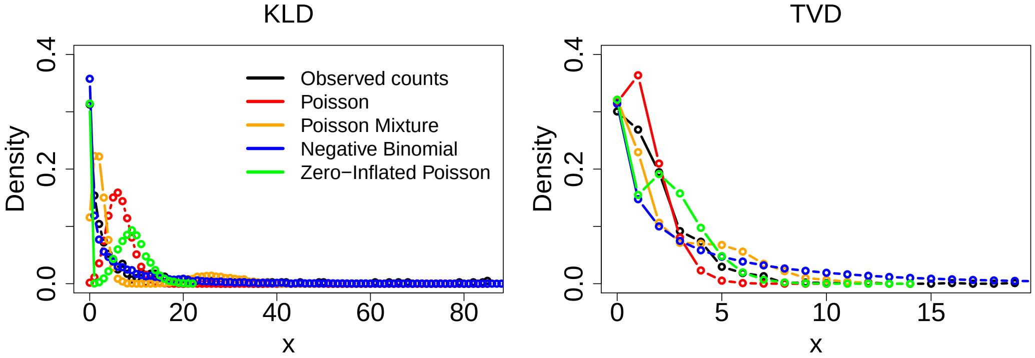

Figure 3.4: Posterior predictive mass functions for one exchangeable observation when fitting a Poisson likelihood (Poi), a two component mixture of Poissons likelihood (Poi Mix), a zero-inflated Poisson likelihood (ZI Poi) and a negative-binomial likelihood (NB), unconstrained a priori, to the **GrouseTicks dataset**. **Left:** using Bayes' rule (KLD-Bayes) updating. **Right:** using updating aimed at minimising the TVD (TVD-Bayes).

Table 3.4: 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

3.5.2 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 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. However, in principle there are a huge number of distributions sharing these properties that could have been used instead of the Gaussian.

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 3.5. 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 assumptions when using the βD-Bayes.

In order to do so 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

$$f(x;\theta_f) := \mathcal{N}\left(x;\mu_f,\sigma_f^2 \times \sigma^2_{adj}\right) \tag{3.123}$$

$$h(x;\theta_h) := \text{Student's} - t_\nu\left(x;\mu_h,\sigma_h^2\right) \tag{3.124}$$

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 $\sigma^2 = \sigma_f^2 = \sigma_h^2$. For $\nu = 5$ we find by optimisation that $\sigma^2_{adj} = 1.16$. In fact we can use the representation

$$\text{TVD}(g,f) = \int_{g>f}(g-f) = \int_{g>f} g - \int_{g>f} f \tag{3.125}$$

to estimate that this neighbourhood also corresponds to a $\mathcal{N}_\epsilon^{\text{TVD}}$ neighbourhood with $\epsilon = 0.043$ as defined in Eq. (3.12). Figure 3.5 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.

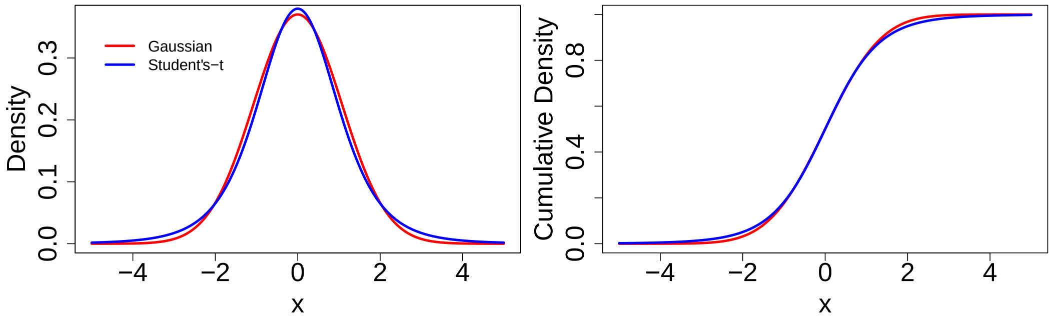

Figure 3.5: Probability density function (pdf) and cumulative density function (cdf) of a Gaussian $f(x;\theta_f) = \mathcal{N}\left(x;\mu_f,\sigma^2_{adj}\sigma^2_f\right)$ and a Student's-t $h(x;\theta_h) = t_\nu(x;\mu_h,\sigma^2_h)$ random variable, with $\mu = 0$ and $\sigma^2 = 1$. The parameters $\nu = 5$ and $\sigma^2_{adj} = 1.16$ where chosen such that the two densities have the same median and quartiles which also ensured that $\{f(x;\theta_f), h(x;\theta_h)\} \in \mathcal{N}^{\text{TVD}}_\epsilon$ (defined in (3.12)) 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 seems unreasonable.

A toy experiment

To investigate the stability of inferences across this neighbourhood $\mathcal{N}^{\text{TVD}}_\epsilon$ of likelihood models, we generated $n = 1000$ observations from the ϵ-contamination model introduced in Section Example 2.1 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 $(a_0 = 0.01, b_0 = 0.01, \mu_0 = 0, v_0 = 10)$. Figure 3.6 plots the parameter posterior and posterior predictive distributions for both models under both updating mechanisms.

The left hand side of Figure 3.6 demonstrates what most statistical practitioners expect when comparing the performance of a Gaussian and a Student's-t likelihood 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

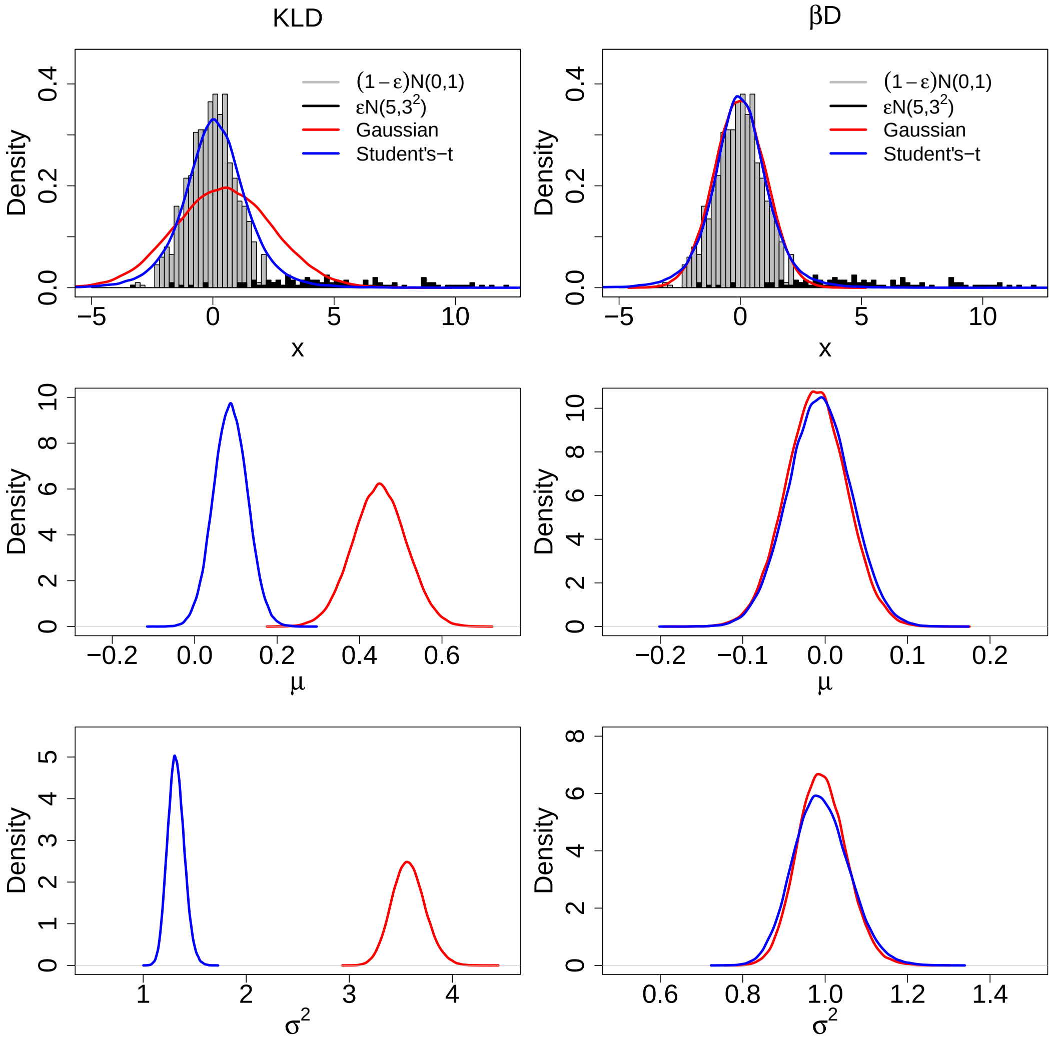

Figure 3.6: Posterior predictive distributions and parameter posterior distributions for $\left(\mu, \sigma^2\right)$ under Bayes' rule updating (KLD-Bayes) (**left**) and βD-Bayes (**right**) under the likelihood functions $f(x;\theta_f) = \mathcal{N}\left(x;\mu,\sigma^2_{adj}\sigma^2\right)$ (red) 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)

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 this $\mathcal{N}_\epsilon^{\text{TVD}}$ here, but we also argue that the βD predictive better captures the majority of the

Table 3.5: 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}

DGP better 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 3.5 presents the energy distance between the two predictives

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

$$Y = \boldsymbol{X}\boldsymbol{\theta}^T + \epsilon, \tag{3.126}$$

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

$$\epsilon \sim \mathcal{N}\left(0, \sigma^2 \times \sigma^2_{adj}\right) \tag{3.127}$$

$$\epsilon \sim t_\nu\left(0, \sigma^2\right), \tag{3.128}$$

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 requirements of buildings as a function of seven 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 four other power plant parameters.

- Concrete: 1030 observations seeking to understand the relationship between concrete's compression strength as a function of eight other features of the concrete.

- BostonHousing: 506 observations seeking to estimate the relationship between the median property value in neighbourhoods of Boston and thirteen 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 nothing about the stability in terms of the parameter estimates and the log-score.

Energy dataset: Figure 3.7 compares these six stability metrics for the four updating methods we consider above when 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 βD-loss is equivalent to the log-score for $\beta = 1$), the difference in this measure focuses less on the stability

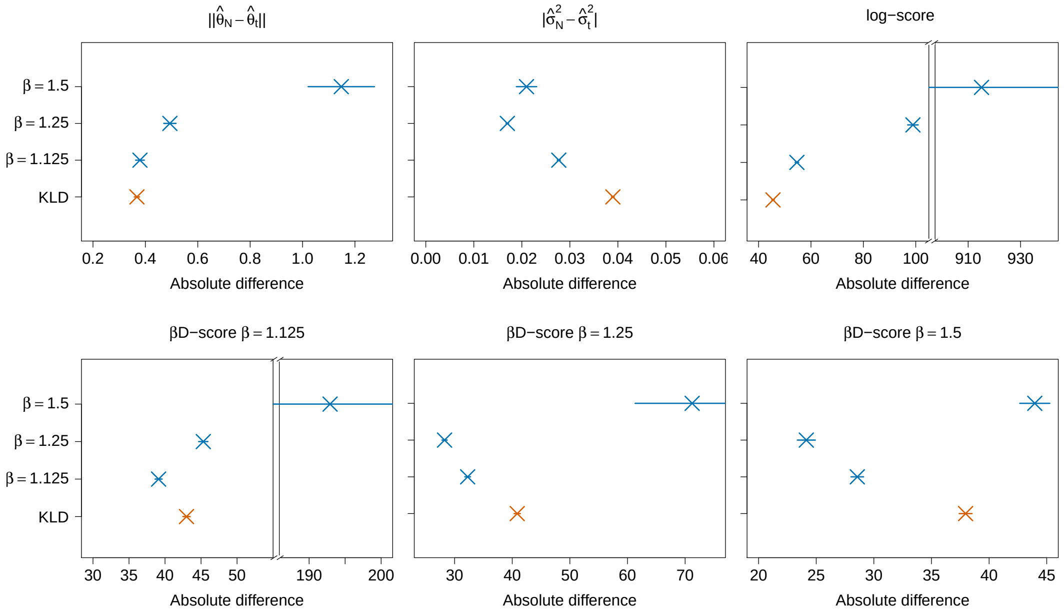

Figure 3.7: Plots comparing the stability of Bayes' rule and βD-Bayes for $\beta = 1.125, 1.25, 1.5$ inference for linear regression models under either Gaussian or Student's-t error distributions applied to the **Energy dataset**. **From left to right**: L_1 norm of the difference between the posterior 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.

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.

Power dataset: Figure 3.8 compares the 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.

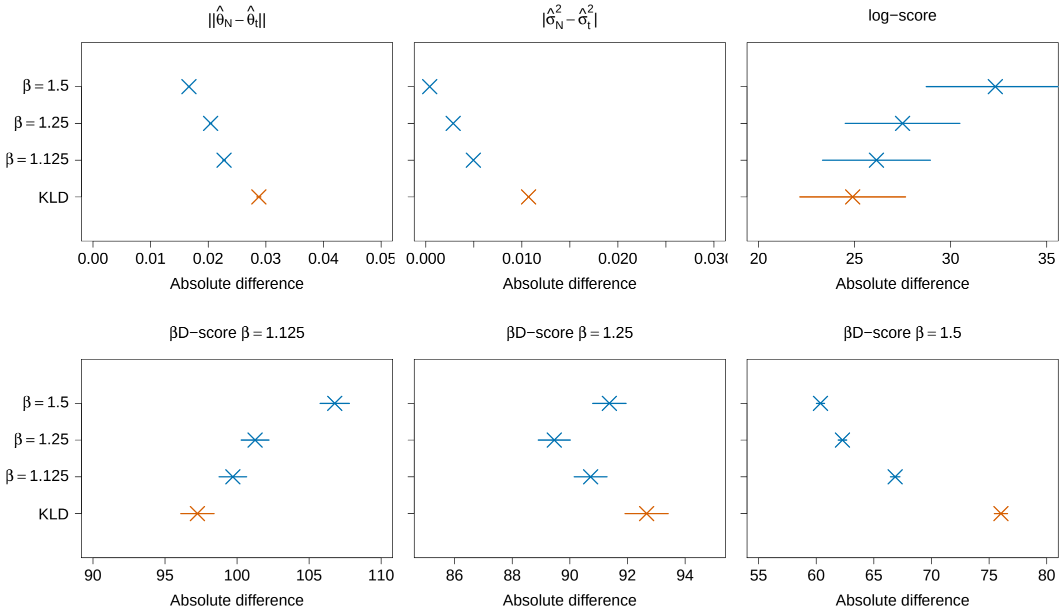

Figure 3.8: Plots comparing the stability of Bayes' rule and βD-Bayes for $\beta = 1.125, 1.25, 1.5$ inference for linear regression models under either Gaussian or Student's-t error distributions applied to the **Power dataset**. **From left to right**: L_1 norm of the difference between the posterior 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.

Concrete dataset: We observed something quite different for the Concrete dataset in Figure 3.9. 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!

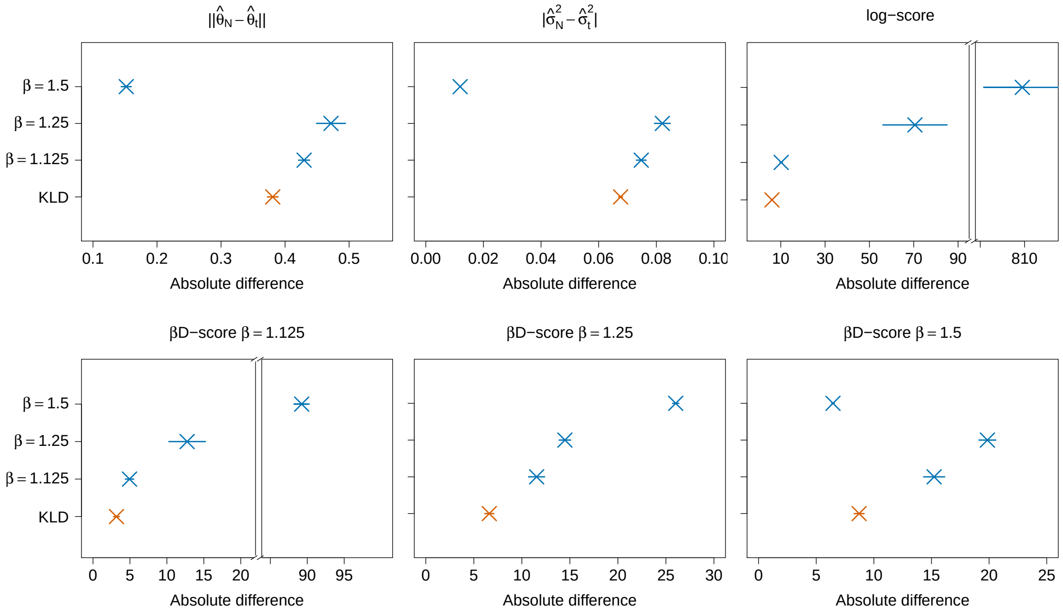

Figure 3.9: Plots comparing the stability of Bayes' rule and βD-Bayes for $\beta = 1.125$, 1.25, 1.5 inference for linear regression models under either Gaussian or Student's-t error distributions applied to the **Concrete dataset**. **From left to right**: L_1 norm of the difference between the posterior 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.

Boston Housing dataset: Lastly, Figure 3.10 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 3.9 and 3.10 demonstrate, similarly to the experiments of Chapter 2, 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.

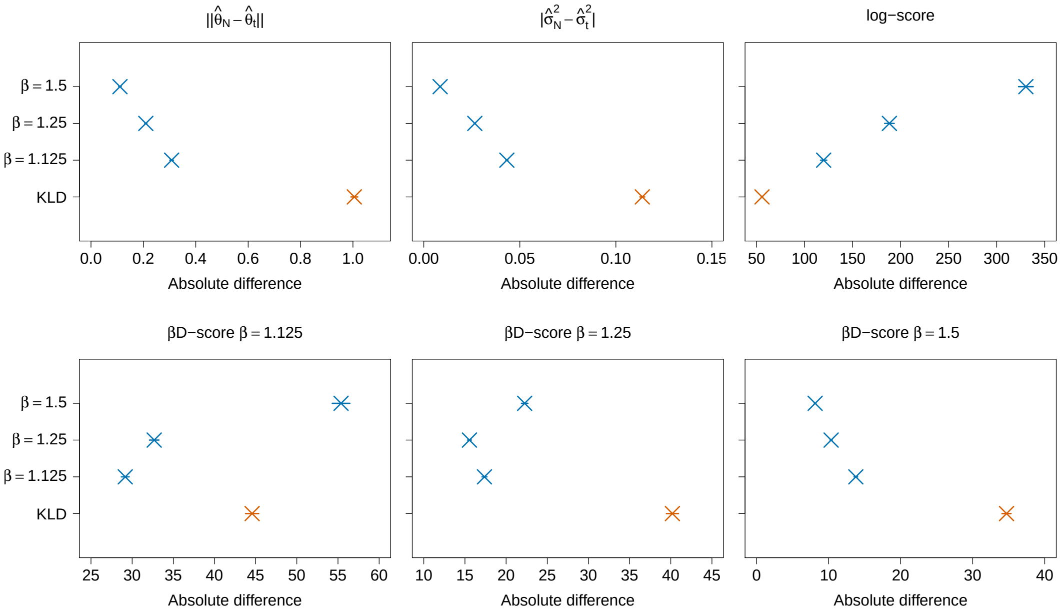

Figure 3.10: Plots comparing the stability of Bayes' rule and βD-Bayes for $\beta = 1.125,\ 1.25,\ 1.5$ inference for linear regression models under either Gaussian or Student's-t error distributions applied to the **BostonHousing dataset**. **From left to right**: L_1 norm of the difference between 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.

3.6 Further Work

The results of this Chapter demonstrate that inference aimed at minimising the TVD and HD (metrics) and also the βD, can be shown to be stable to certain 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 $\mathcal{M}$-OPEN world. These tell a DM exactly how far they need to go with their belief elicitation to be sure that any interpolation does not significantly effect the posterior conclusions. These compliment the results of Chapter 2, showing that these general Bayesian minimum divergence inferences are philosophically preferable both in a decision making capacity and in a belief elicitation capacity.

The next two chapters focus specifically on minimising the βD. Chapter 4 outlines a computationally convenient and philosophically interesting regime to conduct inference. Chapter 5 then applies this methodology to improve inferences from on-line change-point detection.

Chapter 4

Generalised Variational Inference

Chapters 2 and 3 presented a philosophically appealing way to consider conducting inference in the $\mathcal{M}$-OPEN world. This chapter embeds this within a wider framework of Bayesian analyses where we additionally consider the computational constraints that may also be present when conducting a modern $\mathcal{M}$-OPEN Bayesian analysis. In this chapter we view Bayesian inference as an optimisation problem over the space of densities. We use this insight to further generalise the definition of a principled Bayesian inference problem, containing Bayes' rule, GBI and VI as special cases. We then develop a further subclass of Bayesian inference problems to improve on the performance of VI. We call these generalised variational inference (GVI). We show that the modular formulation of GVI is able to solve 3 problems associated with traditional Bayesian and approximate Bayesian inference: tractability, robustness and over concentration. We show the advantages of GVI relative to other, less principled approximate Bayesian inference methods. The majority of this chapter appears in the article Knoblauch, Jewson, and Damoulas [2019] which has been submitted for publication.

The organisation of the chapter is as follows: Section 4.1 interprets Bayes' rule as the solution to an optimisation problem. Section 4.2 introduces variational inference (VI), a method to conduct 'approximate' posterior inference via optimisation. Here we observe that VI is doing constrained optimisation of the same objective function solved by Bayes' rule updating. We also outline the problems faced by VI approximate inference, namely over-concentration of posterior, and methods available in the literature to solve these. We are then motivated by the link between Bayes' rule and VI and the observed unsuitability of the VI objective function for modern inference to produce a generalised optimisation framework for Bayesian inference in Section 4.3. This is motivated axiomatically and made up of three interpretable components, a loss function defining the parameter of interest, a prior regularising divergence describing how the posterior quantifies uncertainty and a space of probability densities to be optimised over. While Chapters 2 and 3 have considered changing the loss function to achieve robustness, Section 4.4 considers changing the prior regularising divergence to solve the over-concentration problems associated with VI, we call this generalised variational inference (GVI). Section 4.5 demonstrates the importance of the axiom underpinning GVI by illustrating situations where the previously available solutions to the problems with VI produce undesirable or non-transparent inference. Section 4.6 interprets GVI as an approximation to the Bayes' rule posterior. Section 4.7 produces a black box optimisation algorithm to solve the GVI objective function and Section 4.8 shows that GVI can outperform the state of the art for both Bayesian Neural Networks and Deep Gaussian Processes.

4.1 Bayes' rule as optimisation

Before presenting VI, an approach to 'approximate' the Bayesian posterior relying on optimisation rather than sampling, we examine a little known optimisation view of Bayes' rule updating.

In developing GBI, Bissiri et al. [2016] shows that Bayes' rule can be written as an optimisation problem over the space of densities. Bayes' rule is recovered from the GBI updating rule when $\ell(\theta, \boldsymbol{x}) = -\sum_{i=1}^{n} \log f(x_i;\theta)$. As a result Eq. (1.17) shows that the traditional Bayesian posterior resulting from Bayes' rule is the solution to the following optimisation problem

$$\pi(\theta|\boldsymbol{x}) = \arg\min_{q\in\mathcal{P}} \left\{ \mathbb{E}_{q(\theta)}\left[-\sum_{i=1}^{n} \log\left(f(x_i;\theta)\right)\right] + \textsc{kld}(q(\theta)||\pi(\theta)) \right\} \tag{4.1}$$

where $\mathcal{P} = \left\{q(\theta) : \int q(\theta)d\theta = 1\right\}$. Note that by optimising over $\mathcal{P}$ we implicitly assume that the solution to Bayes' rule can be normalised. The proof of this can be found in Section 1.2. This interpretation of Bayes' rule can actually be traced back to Zellner [1988], who motivates the above objective function from an information theoretic standpoint and shows that Bayes' rule is the optimal information processing rule. Henceforth we refer to Eq. (4.1) as the traditional Bayesian objective function.

4.1.1 Interpreting the Bayesian objective function

The traditional Bayesian objective function presents a particularly transparent interpretation of Bayes' rule updating. Here we analyse the roles of the two terms in this objective function. The first term is

$$\mathbb{E}_{q(\theta)}\left[-\sum_{i=1}^{n}\log\left(f(x_i;\theta)\right)\right], \tag{4.2}$$

and this will be minimised at

$$q(\theta) = \mathbf{1}_{\left[\theta=\hat{\theta}_n\right]}, \qquad \hat{\theta}_n = \underset{\theta\in\Theta}{\arg\min}\sum_{i=1}^{n} -\log\left(f(x_i;\theta)\right). \tag{4.3}$$

This defines the parameter of interest under Bayes' rule as

$$\theta^* = \underset{\theta\in\Theta}{\arg\min}\int -\log\left(f(x;\theta)\right)dG(x) \tag{4.4}$$

which we know from Chapter 2 to be the parameter minimising the KLD between the DGP and the model.

The second term is

$$\text{KLD}(q(\theta)||\pi(\theta), \tag{4.5}$$

and this limits how far $q(\theta)$ can move from $\pi(\theta)$, allowing the prior to regularise the within-sample inference. This can be seen from the fact that KLD$(\mathbf{1}_{\left[\theta=\hat{\theta}\right]}||\pi(\theta))$ is undefined so this regularisation term ensures $\pi(\theta|\boldsymbol{x})$ does not degenerate and produces a posterior quantification of uncertainty.

4.2 Variational Inference (VI)

The view of Bayes' rule as optimising over the space of densities is not acknowledged by many. Bayesian inference is usually computed via conditional probability updates often requiring sampling procedures. Optimisation within Bayesian inference however, is usually associated with approximate inference procedures such as variational inference (VI) methods [Jordan et al., 1999; Beal et al., 2003]. Instead of attempting to sample from an intractable exact posterior $\pi(\theta|\boldsymbol{x})$, VI methods posit a family of tractable approximate posteriors $q(\theta;\kappa)$. This family is typically called the variational family. The hyperparameters of this variational family are then optimised to find the $\hat{q}(\theta) = q(\theta;\hat{\kappa}(\boldsymbol{x}))$ closest to the exact posterior. This distribution is then treated as an approximation to the exact posterior. Due to their computational convenience, it is common to assume that the variational family is a member of the exponential family [e.g. Beal et al., 2003]. That means we can write their densities in the form

$$q(\theta;\kappa) = h(\theta)\exp\left\{\eta(\kappa)^T T(\theta) - A(\eta(\kappa))\right\} \tag{4.6}$$

with $A(\eta(\kappa)) = \log\int h(\theta)\exp\left\{\eta(\kappa)^T T(\theta)d\theta\right\}$ and natural parameter space $\mathcal{N} = \{\eta(\kappa) : A(\eta(\kappa)) < \infty\}$. Henceforth we consider the variational family to be of this form.

4.2.1 Traditional VI

The closeness between the variational family $q(\theta;\kappa)$ and the exact posterior $\pi(\theta|\boldsymbol{x})$ is traditionally measured using the KLD [Jordan et al., 1999; Beal et al., 2003]. Minimising the KLD in this fashion provides a convenient objective function of the from

$$\begin{aligned}
\hat{q}(\theta) &= \arg\min_{q\in\mathcal{Q}} \{\mathrm{KLD}(q(\theta;\kappa)||\pi(\theta|\boldsymbol{x}))\} && (4.7)\\
&= \arg\min_{q\in\mathcal{Q}} \left\{\int q(\theta;\kappa)\log\frac{q(\theta;\kappa)}{\pi(\theta|\boldsymbol{x})}d\theta\right\}\\
&= \arg\min_{q\in\mathcal{Q}} \left\{\int -q(\theta;\kappa)\log\prod_{i=1}^{n} f(x_i;\theta) + q(\theta;\kappa)\log\frac{q(\theta;\kappa)}{\pi(\theta)} + q(\theta;\kappa)\log\pi(\boldsymbol{x})d\theta\right\}\\
&= \arg\min_{q\in\mathcal{Q}} \left\{\mathbb{E}_{q(\theta;\kappa)}\left[-\sum_{i=1}^{n}\log\left(f(x_i;\theta)\right)\right] + \mathrm{KLD}(q(\theta;\kappa)||\pi(\theta))\right\}, && (4.8)
\end{aligned}$$

where $\mathcal{Q} = \{q(\theta;\kappa) : \kappa \in \mathcal{K}\}$, $\mathcal{K} = \left\{\kappa \in \mathbb{R}^d; \int q(\theta;\kappa)d\theta = 1\right\}$ and the marginal likelihood $\pi(\boldsymbol{x})$ can be ignored as it does not depend on θ. The negative of Eq. (4.8) is traditionally known as the evidence lower bound (ELBO). This draws its name from the following identity [Jordan et al., 1999]

$$\begin{aligned}
&\mathbb{E}_{q(\theta)}\left[\sum_{i=1}^{n}\log\left(f(x_i;\theta)\right)\right] - \mathrm{KLD}(q(\theta)||\pi(\theta)) = \log\int\pi(\theta)\prod_{i=1}^{n} f(x_i;\theta)d\theta - \mathrm{KLD}(q(\theta)||\pi(\theta|\boldsymbol{x}))\\
&\Rightarrow \mathbb{E}_{q(\theta)}\left[\sum_{i=1}^{n}\log\left(f(x_i;\theta)\right)\right] - \mathrm{KLD}(q(\theta)||\pi(\theta)) \leq \log\int\pi(\theta)\prod_{i=1}^{n} f(x_i;\theta)d\theta \qquad (4.9)
\end{aligned}$$

By definition KLD$(q(\theta;\kappa)||\pi(\theta|\boldsymbol{x})) \geq 0$ and therefore the negative of the objective in Eq. (4.8) forms a lower bound on the log evidence

$$\log \pi(\boldsymbol{x}) = \log \int \pi(\theta) \prod_{i=1}^{n} f(x_i;\theta) d\theta. \tag{4.10}$$

This lower bound demonstrates that the approximate posterior $\hat{q}(\theta)$ maximises a lower bound for the evidence of the data under the model providing further justification for its use.

4.2.2 VI and Bayes' rule

Approaching the implementation of Bayes' rule as the optimisation in Eq. (4.1) illuminates striking similarities between Bayes' rule and traditional VI. In fact the objective function of Bayes' rule (Eq. (4.1)) is in fact identical to that of VI (Eq. (4.8)). The only difference between the two procedures is the class of densities over which the optimisation is performed. Bayes' rule optimises over the space of all densities finding the 'exact' minimal density for the objective function. Alternatively VI constrains the solution to live within the variational family. **VI is a constrained optimisation of the traditional Bayesian objective function for inference**. This demonstrates exactly where the tractability of VI comes from. Both Bayes' rule and VI are optimisation problems, but optimisation in VI is done over the real valued hyperparameters indexing the variational family rather than the space of infinite dimensional densities that are optimised over in Bayes' rule.

4.2.3 F-VI

VI performance has recently been largely criticised for tending to lead to over-concentration of marginal posteriors [see e.g. Turner and Sahani, 2011]. In large scale machine learning problems it is common for computational reasons to take the Gaussian mean-field variational family. The Gaussian mean-field variational family is simply the product of independent Gaussians for each parameter. When the actual posterior has dependency between parameters, the member of the mean-field family closest in KLD has considerably smaller marginal variances. This is demonstrated using 2-dimensional Gaussians in Figure 4.1. This plotted example is inspired by that of Bishop [2006]; Blei et al. [2017].

This phenomenon is attributed to the zero-forcing nature of minimising KLD$(q||\pi)$ associated with traditional VI. That is to say that, minimising KLD$(q||\pi)$ over q encourages the optimal $\hat{q}$ to have areas of density close to 0 where the density

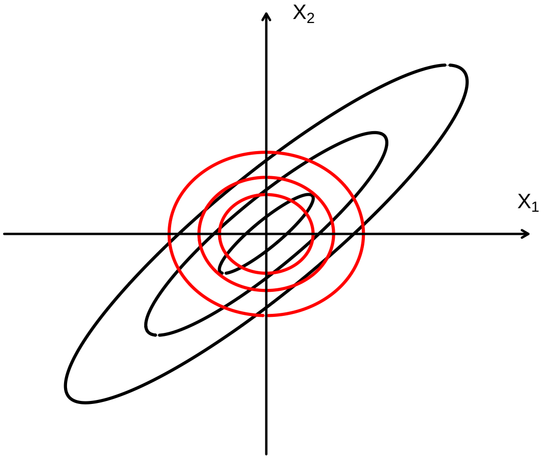

Figure 4.1: **Black**: Contours of a standard bivariate Gaussian with correlation $\rho = 0.9$ drawn at density $0.01, 0.1$ and 0.3. **Red**: corresponding contours of the independent bivariate Gaussian minimising the KLD associated with traditional VI

π is non-zero. Hence, the marginal variances of VI approximations are often too small.

Under-estimating marginal variances can be seen to damage the out-of-sample predictive performance of these algorithms. In recent years there has been an increasing volume of literature considering approximate inference minimising alternative divergences to the KLD between the variational family and the solution to Bayes' rule. These methods consider divergences with zero-avoiding properties in an attempt to achieve more conservative estimates of marginal uncertainty. These approaches can be separated into two categories: Global and Local divergence minimisation.

Global methods, similar to traditional VI posit a variational family for the whole posterior and construct an optimisation problem minimising some divergence between the approximate and exact posterior. The current literature on this approach is vast. For example: both Li and Turner [2016] and Saha et al. [2019] consider minimising the RÉNYI-αD; Regli and Silva [2018] minimise the scaled $\alpha\beta$-Divergence; Dieng et al. [2017] minimise χ-Divergences; Ambrogioni et al. [2018] minimise Wasserstein divergences; and Ranganath et al. [2016] produce the more general operator-VI.

Local methods, use the properties of exponential families to posit a variational family for each likelihood term and take inspiration from message-passing to estimate these 'locally' in an iterative manner. Expectation Propagation (EP) [Minka, 2001; Opper and Winther, 2000] locally minimises the reverse of the KLD used in traditional VI. Power-EP [Minka, 2004] extends this to locally minimising αD's while Black-Box αD minimisation [Hernandez-Lobato et al., 2016] has a similar objective function to Power-EP but fixes the estimate for each likelihood term to be the same across the dataset.

We combine both local and global divergence minimisation methods under the term F-VI

Definition 16 (F-VI methods)**.** F-VI methods produce approximate posterior $\hat{q} \in \mathcal{Q}$ by either locally or globally minimising divergence F between members of the variational family, $\mathcal{Q}$, and posterior $\pi(\theta|\boldsymbol{x})$,

$$\hat{q} = \arg\min_{q \in \mathcal{Q}} F(q(\theta)||\pi(\theta|\boldsymbol{x})), \tag{4.11}$$

where $F(q(\theta)||\pi(\theta|\boldsymbol{x})) \neq \text{KLD}(q(\theta)||\pi(\theta|\boldsymbol{x}))$

These methods are motivated as being able to produce more conservative estimates of marginal variance and as a result produce better test set predictive performance.

4.2.4 VI optimality

Now the previous sections present a paradox! Section 4.2.2 presented VI as a constrained optimisation of the traditional Bayesian objective function solved by Bayes' rule. Therefore, by definition VI must produce the optimal distribution within the constraining family.

Theorem 7 (Optimality of VI for the traditional Bayesian problem)**.** For fixed variational family $\mathcal{Q}$, VI minimising the KLD to the posterior provided by Bayes' rule provides the optimal approximation to this posterior measured using the objective function eliciting Bayes' rule in Eq.(4.1)

Proof. A sketch of this proof is as follows. Minimising the KLD between a member of the variational family $q \in \mathcal{Q}$ is equivalent to finding the optimal q according to the objective function in Eq. (4.8). This in turn is the same objective function which optimised over the space of all normalised probability densities provided Bayes' rule. A more formal proof is provided in Knoblauch et al. [2019] □

However, the alternative F-VI methods claim to produce more desirable approximate posteriors within the same variational family, $\mathcal{Q}$, motivated by the approximation of marginal variances and improved test set predictive performance. How can this be?

This paradox is resolved by noticing that VI produces the optimal posterior in the constrained family relative to the traditional Bayesian inference problem defined by Bayes' rule (Eq. (4.1)). F-VI methods must definitionally achieve no greater value for Eq. (4.1) than the VI approximation. Therefore, for these F-VI methods to be producing more desirable posteriors within the same constraining family, means the objective function associated with Bayes' rule must no longer reflect the inference problem the DM really wants to solve.

We established in Chapters 2 and 3 that while Bayes' rule is the correct thing to do in the $\mathcal{M}$-CLOSED world, where the DM has the time and introspection to exactly specify their model (and prior), this is no longer so clear in the $\mathcal{M}$-OPEN. Analogously, once the prior and model have been (correctly) specified, if there is not infinite time available for computation it appears that Bayes' rule no longer necessarily provides a suitable optimisation criteria for inference.

In some circumstance it appears that F-VI is able to implicitly provide a more desirable objective function for inference. However, this objective function is obscured by the formulation of F-VI as a posterior approximation. As a result of this, we are able to show two situations where the formulation of F-VI as a posterior approximation leads to opaque and undesirable inference in Sections 4.5 and 4.8.1

Instead, we take inspiration from the improved performance of F-VI methods and seek to generalise the optimisation problem in Eq. (4.1). We shall encode desirable qualities for the posterior directly into the objective function eliciting it, rather than an approximating divergence.

4.3 Generalising the Bayesian inference problem

Inspired by the optimality of VI relative to the traditional Bayesian inference problem and the realisation that the objective function defined by Bayes' rule is not necessarily suitable for modern Bayesian analyses, we derive a generalisation of the Bayesian inference problem. This is first defined and then we provide an axiomatic derivation of this form.

4.3.1 The Bayesian inference problem

Here, we argue that Bayesian inference should take the form $P(\ell_n, D, \Pi)$ given by

$$q^*(\theta) = \underset{q\in\Pi}{\arg\min}\left\{L(q|\boldsymbol{x}, \ell, D)\right\};$$
$$L(q|\boldsymbol{x}, \ell, D) = \mathbb{E}_{q(\theta)}\left[\ell(\theta, \boldsymbol{x})\right] + D\left(q(\theta)||\pi(\theta)\right), \tag{4.12}$$

where the arguments of the form $P(\ell_n, D, \Pi)$ are given by

- a **loss** ℓ_n defining the target parameter for inference relative to the sample distribution of observations, $\theta^* = \arg\min_{\theta\in\Theta}\int \ell(\theta, x)dG(x)$. This will often be additive over observations $\ell_n(\theta, \boldsymbol{x}) = \sum_{i=1}^n \ell(\theta, x_i)$ for some ℓ. The term $\mathbb{E}_{q(\theta)}\left[\ell(\theta, \boldsymbol{x})\right]$ will be minimised at $q(\theta) = \mathbf{1}_{[\theta=\hat{\theta}]}$ where $\hat{\theta} = \arg\min_{\theta\in\Theta}\ell_n(\theta, \boldsymbol{x})$.
- a **divergence** $D : \mathcal{P}(\Theta)\times\mathcal{P}(\Theta)\to\mathbb{R}_+$ regularising the posterior with respect to the prior π. D will determine how uncertainty around $\hat{\theta}$ is quantified by $q^*(\theta)$. As a result we term this the **uncertainty quantifier**.
- a set of **admissible probability distributions** for the posterior $\Pi\subseteq\mathcal{P}(\Theta)$ with $\mathcal{P}(\Theta) = \left\{q(\theta) : \int q(\theta) = 1\right\}$, over which the regularised expected loss is minimized.

Definition 17 (A Bayesian inference problem)**.** Any Bayesian inference method solving $P(\ell_n, D, \Pi)$ with admissible choices ℓ_n, $D(\cdot||\cdot)$ and Π. Theorems 8 and 9 prove Bayesian inference methods satisfy Axioms 1 – 4.

Before demonstrating the axiomatic construction of the form of $P(\ell_n, D, \Pi)$ we present traditional Bayes' rule inference as a special case of our framework. Firstly Eq. (4.1) demonstrates that the Bayes' rule posterior can be derived as the minimiser of an objective function over a space of densities, for Bayes' rule this is the space of all normalised probability densities $\mathcal{P}$. For the objective function, taking $D = \text{KLD}$ and $\ell(\theta, \boldsymbol{x}) = -\sum_{i=1}^n \log f(x_i;\theta)$ in Eq. (4.12) recovers the objective function in Eq. (4.1). As a result the Bayes' rule posterior is a solution to the Bayesian inference problem

$$P(-\sum_{i=1}^n \log(f(x_i;\theta)), \text{KLD}, \mathcal{P}(\Theta)). \tag{4.13}$$

We provide further examples of methods currently available in the literature satisfying our framework in Section 4.3.3. But first, we produce an axiomatic derivation of the form of $P(\ell_n, D, \Pi)$.

4.3.2 Axiomatic Derivation

This section imposes intuitive axioms on Bayesian inference which derive the form of $P(D, \ell_n, \Pi)$ in Eq. (4.12). While the objective function, Eq. (4.1), associated with Bayes' rule is underpinned by the axioms of conditional probability, and the VI objective function, Eq. (4.8), results from a posterior approximation, justifying the form of the objective Eq. (4.12) requires that we impose several axioms. We comment on how these axioms compare and contrast with the assumptions of Bissiri et al. [2016] below. For simplicity, $\boldsymbol{x} = x_{1:n}$ are treated as n independent draws, but the presented arguments extend to conditional independence structures.

Axiom 1 (Representation)**.** Bayesian inference infers posteriors q on Θ by

i. measuring how q fits a sample $\boldsymbol{x}$ via the expectation of a loss $\ell_n(\theta, \boldsymbol{x})$

ii. elicits uncertainty quantification about $\hat{\theta}$ via a divergence $D(q||\pi)$ between q and the prior π.

iii. optimise for q over the space of admissible probability distribution Π on Θ

This axiom formalizes Bayesian inference inspired by Zellner [1988]; Bissiri et al. [2016]. This implies that it is representable as a triplet $P(\ell_n, D, \Pi)$. Showing that $P(\ell_n, D, \Pi)$ takes the form in eq. (4.12) requires three more axioms.

Axiom 2 (Information difference)**.** $P(\ell_n, D, \Pi)$ produces different posteriors for $\boldsymbol{x} = x_{1:n}$ and $\tilde{\mathbf{x}} = x_{1:n+m}$ if and only if there is an information difference of $\tilde{\mathbf{x}}$ relative to $\boldsymbol{x}$, i.e. if $\ell_n(\theta, \boldsymbol{x}) \neq \ell_{n+m}(\theta, \tilde{\mathbf{x}})$.

Axiom 3 (Prior regularization)**.** q is regularized towards π by penalizing the divergence $D(q||\pi)$.

Axiom 4 (Translation Invariance)**.** For constant C and $\ell'_n = \ell_n + C$, $P(\ell'_n, D, \Pi) = P(\ell_n, D, \Pi)$.

Axiom 2 ensures that datasets containing different information about θ produces different posteriors. Axiom 3 says that $D(q||\pi)$ acts as a penalty. Axiom 4 enforces that inference is invariant to adding a constant to ℓ_n. Note that we do *not* want inference to be invariant to multiplications of ℓ_n by a constant. This, for example, would contradict Axiom 2 for additive ℓ_n.

Lemma 6 (Multiplicative Constants)**.** If Axiom 2 holds, ℓ_n is additive and $C \in \mathbb{N}$, $P(\ell_n, D, \Pi) \neq P(C \cdot \ell_n, D, \Pi)$.

Proof. If ℓ_n is additive then $\ell_n(\theta, \boldsymbol{x}) = \sum_{i=1}^n \ell(\theta, x_i)$ for some loss function ℓ and $\boldsymbol{x} = x_{1:n}$. For any $k \in \mathbb{N}$, write $\boldsymbol{x}(k) = x(k)_{1:k \times n}$ with $x(k)_i = x_{(i \mod n)+1}$, where ($a$ mod b) denotes the (integer) remainder of the division a/b. In words, $\boldsymbol{x}(k)$ copies the entries of $\boldsymbol{x}$ exactly k times. Now, simply note that $k \cdot \ell_n(\theta, \boldsymbol{x}) = \ell_{kn}(\theta, \boldsymbol{x}(k))$ to see that $P(D, \ell_{kn}, \Pi) = P(D, k \cdot \ell_n, \Pi) = P(D, \ell_n, \Pi)$ would violate Axiom 2 as $\ell_n(\theta, \boldsymbol{x}) \neq \ell_{kn}(\theta, \boldsymbol{x}(k))$ for $k > 1$ and any choice of D and Π. □

This proof shows that there is an equivalence between multiplying the loss function by a constant and a sample containing more observations. As a result multiplying the loss by a constant changes the amount of information in the sample and therefore results in a different posterior. On the other hand an additive constant cannot be interpreted in this way, the constant affects each observation equally. Therefore an additive constant represents no gain in information and must result in an equivalent posterior.

Finally, we motivate the form of $P(\ell_n, D, \Pi)$ in eq. (4.12). Since the additive nature of Eq. (4.1) and Eq. (4.8) both rely on log-additivity via $D = \text{KLD}$, we require more fundamental arguments for general D.

Theorem 8 (Form 1)**.** If Axiom 1 holds, $P(\ell_n, D, \Pi)$ can be written as

$$\arg\min_{q \in \Pi} \left\{ L(q|\boldsymbol{x}, \ell_n, D) \right\}, \tag{4.14}$$

for $L(q|\boldsymbol{x}, \ell_n, D) = f\left(\mathbb{E}_{q(\theta)}[\ell_n(\theta, \boldsymbol{x})], D(q||\pi)\right)$, where f is some function $f : \mathbb{R}^2 \to \mathbb{R}$.

Proof. This follows directly from Axiom 1: By (iii), Bayesian inference is an optimization over Π producing a posterior q. Moreover, by (i) this optimization depends on the expectation of the loss ℓ_n via q's expectation, i.e. via $\mathbb{E}_{q(\theta)}[\ell(\theta, \boldsymbol{x})]$. Further, by (ii) it also depends on the divergence D between prior and q, i.e. on $D(q||\pi)$. Hence, Bayesian inference is representable as $\arg\min_{q \in \Pi} \{L(q|\boldsymbol{x}, \ell_n, D)\}$ for $L(q|\boldsymbol{x}, \ell_n, D) = f\left(\mathbb{E}_{q(\theta)}[\ell_n(\theta, \boldsymbol{x})], D(q||\pi)\right)$ for a function f. □

Theorem 9 (Form 2)**.** For $P(\ell_n, D, \Pi)$ being $\arg\min_{q \in \Pi} \{L(q|\boldsymbol{x}, \ell_n, D)\}$ and $\circ$ an elementary operation on $\mathbb{R}$, $L(q|\boldsymbol{x}, \ell_n, D) = \mathbb{E}_{q(\theta)}\left[\ell_n(\theta, \boldsymbol{x})\right] \circ D(q||\pi)$ satisfies Axioms 3 and 4 only if $\circ = +$.

Proof. First rewrite

$$f\left(\mathbb{E}_{q(\theta)}[\ell_n(\theta, \boldsymbol{x})], D(q||\pi)\right) = \mathbb{E}_{q(\theta)}\left[\ell_n(\theta, \boldsymbol{x})\right] \circ D(q||\pi). \tag{4.15}$$

The elementary operations are addition, subtraction, multiplication and division. Consider the losses ℓ_n and $\ell'_n = \ell_n + C$ for $C \in \mathbb{R}$ a constant. It is straightforward to see that Axiom 4 will not hold in general if $\circ$ is multiplication, as

$$\begin{aligned} &\arg\min_q \left\{L(q|\boldsymbol{x}, \ell'_n, D)\right\} \\ &= \arg\min_q \left\{\mathbb{E}_{q(\theta)}\left[\ell_n(\theta, \boldsymbol{x}) + C\right] \cdot D(q||\pi)\right\} \\ &= \arg\min_q \left\{\mathbb{E}_{q(\theta)}\left[\ell_n(\theta, \boldsymbol{x})\right] D(q||\pi) + C \cdot D(q||\pi)\right\} \\ &\neq \arg\min_q \left\{\mathbb{E}_{q(\theta)}\left[\ell_n(\theta, \boldsymbol{x})\right] D(q||\pi)\right\} \\ &= \arg\min_q \left\{L(q|\boldsymbol{x}, \ell_n, D)\right\}, \end{aligned} \tag{4.16}$$

and similarly if $\circ$ is division. As C is a constant, it is however easy to show that if $\circ$ is addition or subtraction,

$$\begin{aligned} &\arg\min_q \left\{L(q|\boldsymbol{x}, \ell'_n, D)\right\} \\ &= \arg\min_q \left\{L(q|\boldsymbol{x}, \ell_n, D)\right\}. \end{aligned} \tag{4.17}$$

Since subtracting the prior regulariser is a direct and obvious violation of Axiom 3, it follows that addition is the only elementary operation on $\mathbb{R}$ satisfying both Axioms and the result follows. □

Comparison with Bissiri et al. [2016]

Next we compare our axiomatic derivation of the form of $P(\ell_n, D, \Pi)$ above to the assumptions used to justify the general Bayesian update in Bissiri et al. [2016]. Axiom 1 formalises the assertions of Bissiri et al. [2016] that posterior beliefs exist (iii) in the absence of a model for the data and that these can be defined through an optimisation problem using the expected loss (i) and a divergence (ii). Axiom 2 and part (i) of Axiom 1 combine to enforce the same behaviour of Assumption 3 of Bissiri et al. [2016]. Axiom 4 agrees exactly with Assumption 5 of Bissiri et al. [2016]. Assumption 4 of Bissiri et al. [2016] will hold under the generalisation of the Bayesian inference problem provided that the prior, $\pi(\theta)$, is a member of the admissible family of posterior distributions. This is an important consequence of using a divergence, which is by definition minimised to 0 when the distributions are equal, to regularise uncertainty quantification. It seems reasonable to assume that in general the prior is contained within the variational family. Otherwise, why would a DM take the time to elicit a prior distribution that was not admissible for the posterior? However if this were the case and the loss were constant the posterior becomes the member of the set of admissible posteriors closest in terms of prior regularising divergence.

A glaring omission form the discussion above is Assumption 1 of Bissiri et al. [2016], that GBI should be coherent. Here, this is taken to mean that first updating a prior with $x_{1:m}$ and then using this as the prior for $x_{m+1:n}$ results in the same posterior as had we updated the prior with $x_{1:n+m}$ in one go. It is this assumption which enforces $D = \text{KLD}$ for GBI. This assumption will in general be violated by GVI.

We give two reasons for not adopting this axiom within our framework. Firstly, we note that the solution to Eq. (4.12) using the KLD is only coherent when $\Pi = \mathcal{P}(\Theta)$, so restricting where the posterior lives breaks coherence even for $D = \text{KLD}$. Secondly demanding coherence must imply an extreme confidence in the specification of the model, an assumption associated with the $\mathcal{M}$-CLOSED world. For example consider a DM has a uni-modal model for the data and that they are updating their beliefs on-line. Further, consider that as the data accumulates it is clearly coming from a bimodal distribution. The demand for coherence prohibits the DM from changing their model. They must instead achieve the same posterior beliefs had they updated in one go, as they would updating one observation at a time. If a DM is sure that their model is well specified then this is reasonable. But in the $\mathcal{M}$-OPEN world this is no longer the case. Additionally coherence must prohibit any exploratory data analysis which is known by every statistician to be an important first stage in building a model. Assumption 2 is very closely linked to the desire for coherence and will again not be satisfied under the generalisation of the Bayesian inference problem when $D \neq \text{KLD}$ and $\Pi \neq \mathcal{P}(\Theta)$.

4.3.3 Special cases of the Bayesian inference problem

Next, we point out several popular forms of Bayesian inference the fit with our generalised framework. As we have already pointed out Eq. (4.1) shows that Bayes' rule solves

$$P(-\sum_{i=1}^{n} \log(f(x_i;\theta)), \text{KLD}, \mathcal{P}(\Theta)). \tag{4.18}$$

For $\mathcal{Q} = \{q(\theta|\kappa) : \kappa \in \mathcal{K}\}$ a variational family, the objective of

$$P(-\sum_{i=1}^{n} \log(f(x_i;\theta)), \text{KLD}, \mathcal{Q}) \tag{4.19}$$

is Eq. (4.8), the negative of the Evidence Lower Bound (ELBO) of VI. Further for loss function $\ell_n(\theta, \boldsymbol{x})$, GBI [Bissiri et al., 2016] is given by

$$P(\ell_n(\theta, \boldsymbol{x}), \text{KLD}, \mathcal{P}(\Theta)) \tag{4.20}$$

Table 4.1 below lists some special cases of the Bayesian inference problem already appearing in the literature.

Method	$\ell(\theta, x_i)$	D	Π
Standard Bayes	$-\log(f(x_i;\theta))$	KLD	$\mathcal{P}(\theta)$
Generalized Bayes[1]	any ℓ	$\frac{1}{w}$KLD, $w > 0$	$\mathcal{P}(\theta)$
Power Bayes[2]	$-\log(f(x_i;\theta))$	$\frac{1}{w}$KLD, $w > 0$	$\mathcal{P}(\theta)$
Divergence Bayes[3]	divergence-based ℓ	KLD	$\mathcal{P}(\theta)$
Standard VI	$-\log(f(x_i;\theta))$	KLD	$\mathcal{Q}$
Power VI[4]	$-\log(f(x_i;\theta))$	$\frac{1}{w}$KLD, $w > 0$	$\mathcal{Q}$
Regularized Bayes[5]	$-\log(f(x_i;\theta)) + \phi(\theta, x_i)$	KLD	$\mathcal{Q}$
(β-)VAE [6]	$-\log\left(f_{\boldsymbol{\zeta}}(x_i;\theta)\right)$	$\beta \cdot$ KLD, $\beta > 1$	$\mathcal{Q}$
Gibbs VI[7]	any ℓ	KLD	$\mathcal{Q}$
Generalized VI	any ℓ	any D	$\mathcal{Q}$

Table 4.1: $P(\ell_n, D, \mathcal{Q})$ and relation to some existing methods. All losses are additive and of the form $\ell_n(\theta, \boldsymbol{x}) = \sum_{i=1}^{n} \ell(\theta, x_i)$ for some $\ell(\theta, x_i)$. [1][Bissiri et al., 2016; Lyddon et al., 2018], [2][e.g. Holmes and Walker, 2017; Grünwald et al., 2017; Miller and Dunson, 2018], [3][e.g. Hooker and Vidyashankar, 2014; Ghosh and Basu, 2016; Futami et al., 2017; Jewson et al., 2018], [4][e.g. Yang et al., 2017; Huang et al., 2018] [5][Ganchev et al., 2010], but only if the regularizer can be written as $\mathbb{E}_{q(\theta)}\left[\phi(\theta, \boldsymbol{x})\right]$ as in [Zhu et al., 2014], [6][Kingma and Welling, 2013; Higgins et al., 2017], [7][e.g. Alquier et al., 2016; Futami et al., 2017]

Although it is common to consider the weight w as multiplying the loss function [Bissiri et al., 2016; Lyddon et al., 2018; Holmes and Walker, 2017; Grünwald et al., 2017; Miller and Dunson, 2018], it affects the uncertainty quantifying properties of the posterior rather than the limiting location (at least in finite parametric models). Therefore, exactly the same solution is obtained by multiplying the regulariser by $\frac{1}{w}$. We adopt this convention to maintain the transparent interpretability of each term in $P(\ell_n, D, \Pi)$.

We observe here that F-VI methods with $F \neq$KLD are not special cases of the Bayesian inference problem. A discussion of this and its implications takes place in Section 4.5

4.4 Generalised Variational Inference (GVI)

The principled formulation of the Bayesian inference objective function, $P(\ell_n, D, \Pi)$, now presents a DM with three components they can alter to achieve transparent effects on their inference. Two of these components define the form of the Bayesian objective function and we examine these here. We return to the situation motivating VI, where there are not the computational resources available to optimise over the space of all densities. As a result we consider optimising the objective function Eq. (4.12) over a constrained space of normalised densities $\mathcal{Q} \subset \Pi$. We call this subset of Bayesian inference problems generalised variational inference (GVI) problems.

Definition 18 (GVI inference problem)**.** Any Bayesian inference method solving $P(\ell_n, D, \mathcal{Q})$ with admissible choices ℓ_n, $D(\cdot||\cdot)$ and $\mathcal{Q} \subset \Pi$ is a Generalized Variational Inference (GVI) problem.

We defer the discussion of the variational family to Section 4.9.

4.4.1 The loss function

Chapters 2 and 3 have discussed in great depth why a DM may want to use an alternative loss function to the log-score in order to change the target parameter for inference, θ^*, to gain robustness to outliers. This is one of the components that can easily be tuned under GVI. We discuss this no further here. We simply note that although Bayes' rule provides the optimal objective function in the $\mathcal{M}$-CLOSED world, when the model is misspecified the DM may very reasonably seek to use an alternative objective function for their inference.

4.4.2 The prior regularising Divergence

Looking at the GVI objective function, Eq. (4.12), the prior regularising divergence is the only part of the optimisation which elicits uncertainty quantification. That is to say that without the prior regularising divergence the optimal posterior belief would be a point mass at the in-sample loss minimiser and we would recover frequentist inferential methods. As a result, changing the uncertainty quantifying properties of a posterior distribution must require a change to the prior regularising divergence. In fact, once we have decomposed the objective function of traditional VI into the GVI objective function, it is straightforward to see that it must be the KLD prior regularising term that causes the VI posteriors to over concentrate in the presence of unaccounted for correlation.

The large literature around the so called 'zero-avoiding' F-VI procedures suggest that practitioners find it attractive for posterior variances to be conservative, because this property can be shown to improve finite sample prediction. However, it is clear from Figure 4.1 that using the KLD prior regulariser in a GVI problem when optimising over a constrained space of densities neglecting the correlation between parameters does not respect this desire for conservative marginal variances. In fact it is well known that even when the optimisation is done over the whole space of probability distributions, Bayes' rule posteriors are known to over-concentrate when the model is misspecified, suggesting the regularising power of the KLD is not great enough [Holmes and Walker, 2017; Miller and Dunson, 2018; Grünwald, 2016].

Once again we stress that Bayes' rule provides the correct objective function to use when the model is correct and the DM is able to optimise over the space of all probability densities. However, there is no longer any such justification to do this when these assumptions are relaxed. In fact, the current literature on F-VI only goes to show that in practice DM's do wish to change their objective function in these circumstances.

In order to rectify this we propose using more general divergences than the KLD to regularise the prior. In particular we restrict our attention here to several members of the $\alpha\beta\gamma$D family. Specifically the: αD, RÉNYI-αD, βD, γD and $\frac{1}{w}$KLD where we down(up)-weight the KLD regulariser by taking $w > 1$ ($w < 1$).

To the best of our knowledge, we are first to consider inference with $D \neq \frac{1}{w}$KLD. This probably results from exact Bayesian inference – i.e. any problem $P(\ell_n, D, \Pi)$ for $\Pi = \mathcal{P}(\theta)$ – being coherent only if $D = \frac{1}{w}$KLD [Bissiri et al., 2016]. However we have already discussed departures from coherence in Section 4.3.2.

We consider the following toy example to investigate the inferences produced using different prior regularising divergences.

Bayesian Linear Regression (BLR) with correlated predictors

Consider the following simple Bayesian linear regression (BLR) example with two predictors and no intercept $X_i = (X_{i1}, X_{i2})$

$$\begin{aligned} \sigma^2 &\sim \mathcal{IG}(a_0, b_0) \\ \theta|\sigma^2 &\sim \mathcal{N}_2\left(\mu_0, \sigma^2 V_0\right) \\ y_i|\theta, \sigma^2 &\sim \mathcal{N}\left(X_i\theta, \sigma^2\right). \end{aligned} \tag{4.21}$$

This example is convenient as it provides a closed form conjugate exact posterior. Studying this exact posterior for $\theta = (\theta_1, \theta_2)$ reveals that if the two variables X_1

and X_2 are correlated, the corresponding exact Bayesian posterior will be strongly correlated, too. As a result we simulate

$$(X_1, X_2)^T \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix} \right). \tag{4.22}$$

We compare these exact posteriors with traditional VI and the different GVI methods introduced in the previous sections. In order to do so we consider the following variational family

$$\begin{aligned} \mathcal{Q} = \{q(\theta_1, \theta_2, \sigma^2) : \quad q(\theta_1, \theta_2, \sigma^2) &= q(\theta_1|\sigma^2, \boldsymbol{\kappa}_n) q(\theta_2|\sigma^2, \kappa_n) q(\sigma^2|\kappa_n), \kappa_n \in \boldsymbol{K}\} \\ q(\sigma^2|\kappa_n) &= \mathcal{IG}(\sigma^2|a_n, b_n) \\ q(\theta_1|\sigma^2, \kappa_n) &= \mathcal{N}\left(\theta_1|\mu_{1,n}, \sigma^2 v_{1,n}\right) \\ q(\theta_2|\sigma^2, \kappa_n) &= \mathcal{N}\left(\theta_2|\mu_{2,n}, \sigma^2 v_{2,n}\right), \end{aligned} \tag{4.23}$$

where $\kappa_n = (a_n, b_n, \mu_{1,n}, \mu_{2,n}, v_{1,n}, v_{2,n})^T$. Here the regression coefficients θ are constrained to be independent, mimicking the 'mean-field' family, but these coefficients are allowed to depend on the residual variance. Under this variational family and the log-score loss function, the objective functions for VI and GVI using the αD, RÉNYI-αD, βD and γD uncertainty quantifiers are available in closed form. In other words, both the uncertainty quantifier term as well as the expected loss term are available in closed form. Consequently, no sampling is required in order to find the GVI and VI posteriors.

Before considering any data Figure 4.2 plots the magnitude of the KLD, $D_A^{(\alpha)}$, $D_B^{(\beta)}$, $D_{AR}^{(\alpha)}$ and $D_G^{(\gamma)}$ between two members of $\mathcal{Q}$ for varying values of the divergence hyperparameters $\alpha = \beta = \gamma$. The KLD forms a central point for all five divergence, recovering the other four at $\alpha = \beta = \gamma = 1$. Figure 4.2 demonstrates that as we increase $\alpha = \beta = \gamma$ from 0, through the KLD, to 2 the magnitude of the $D_B^{(\beta)}$, $D_{AR}^{(\alpha)}$ and $D_G^{(\gamma)}$ is decreasing. The $D_A^{(\alpha)}$ alternatively, initially decrease but is then minimised at $\alpha = 0.5$ and then increases from there onwards. We refer to this plot throughout the next section to provide intuition about the size of the regularisation to the prior provided by different prior regularisers and their hyperparameters.

For the experiments, $n = 25$ observations are simulated from Eq. (4.22) with $\theta = (2, 3)$ and $\sigma^2 = 4$. The results for the different uncertainty quantifiers are depicted in Figs. 4.3-4.8. We summarize the most interesting results from these plots in the following three subsections.

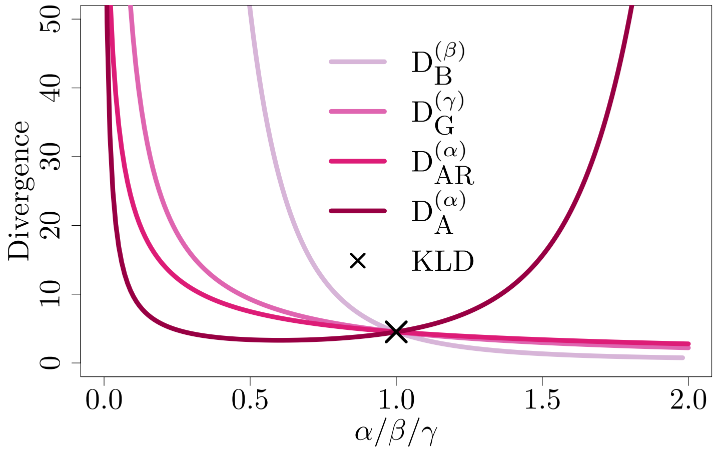

Figure 4.2: A comparison of the sizes of the KLD with the $D_A^{(\alpha)}$, $D_B^{(\beta)}$, $D_{AR}^{(\alpha)}$ and $D_G^{(\gamma)}$ between two bivariate NIG families with $a_n = 512$, $b_n = 543$, $\boldsymbol{\mu}_n = (2.5, 2.5)$, $\mathbf{V}_n = diag(0.3, 2)$ and $a_0 = 500$, $b_0 = 500$, $\mu_0 = (0, 0)$, $V_0 = diag(25, 2)$ for various values of the divergence hyperparameters.

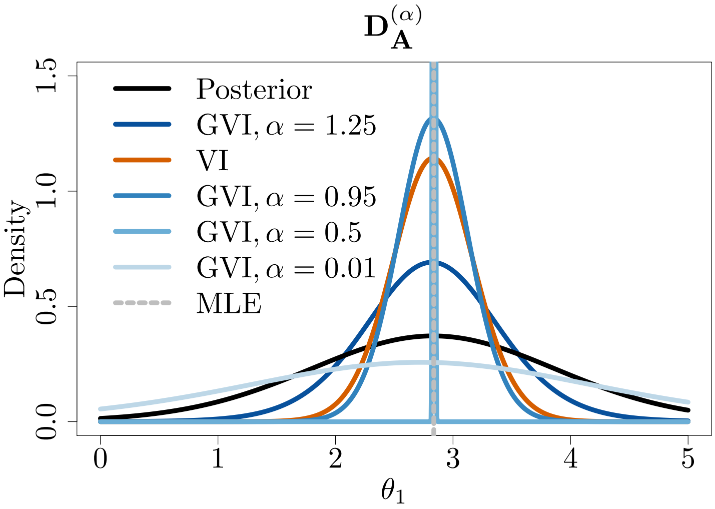

Figure 4.3: Marginal VI and GVI posteriors for the θ_1 coefficient of a Bayesian linear model under the $D_A^{(\alpha)}$ prior regulariser for different values of the divergence hyperparameters. The boundedness of the $D_A^{(\alpha)}$ causes GVI to severely over-concentrate if α is not carefully specified. Prior Specification: $\sigma^2 \sim \mathcal{IG}(20, 50)$, $\theta_1|\sigma^2 \sim \mathcal{N}(0, 25\sigma^2)$ and $\theta_2|\sigma^2 \sim \mathcal{N}(0, 25\sigma^2)$.

The boundedness of the (αD)

Of the alternative divergences to the KLD contained within the $\alpha\beta\gamma$D family [Cichocki and Amari, 2010], αD is arguably the most well known. Here we demonstrate that it is not necessarily suitable to quantify uncertainty in a Bayesian problem specified via $P(\ell_n, D_A^{(\alpha)}, \mathcal{Q})$. In particular, Figure 4.3 shows that the solutions to $P(\ell_n, D_A^{(\alpha)}, \mathcal{Q})$ can produce degenerate posteriors. For example, when $\alpha = 0.5$, $P(\ell_n, D_A^{(\alpha)}, \mathcal{Q})$ essentially collapses to the Maximum Likelihood Estimate. This is a consequence of the boundedness of αD for $\alpha \in (0, 1)$: One can show that $D_A^{(\alpha)}(q||\pi) \leq \frac{1}{\alpha(1-\alpha)}$. As α decreases from 1, this upper-bound initially decreases and as a result decreases the maximal penalty for uncertainty quantification far from the prior – this allows the optimisation to focus solely on minimising the in-sample loss. This phenomenon was depicted in Figure 4.2. However, Figure 4.2 also shows that the magnitude increases again as α approaches 0 and for $\alpha > 1$, where the divergence is no longer bounded. For these values of the hyperparameter, it is possible to achieve more conservative uncertainty quantification. In Figure 4.3 for example, $\alpha = 1.25$ and $\alpha = 0.01$ are able to achieve marginal variances that more closely correspond to the exact posterior. In spite of this, the αD stands as a cautionary tale: Without understanding the properties of the uncertainty quantifier D sufficiently well, GVI may well yield unsatisfactory posteriors.

Increasing the magnitude of the divergence results in posteriors with large variances

In this section, we summarize the impact that selecting one of the alternative divergences can have on the marginal variances of the solution to $P(\ell_n, D, \mathcal{Q})$.

Figure 4.2 provided some idea of how the magnitude of the uncertainty quantifier changes with the hyperparameter and Figure 4.4 illustrates the impact this has on the marginal variances of the resulting posteriors. The latter plot shows that βD, Rényi-αD and γD are able to produce more conservative posterior variance for $\beta, \alpha, \gamma < 1$ and less conservative posterior variance for $\beta, \alpha, \gamma > 1$. This is a manifestation of the posterior being penalized more heavily ($\beta, \alpha, \gamma < 1$) or less heavily ($\beta, \alpha, \gamma > 1$) for deviating from the prior than under the traditional VI. This is also illustrated by Figure 4.2 which shows that the magnitude of these divergences increases as the hyperparameters decrease below 1 (the value recovering the KLD) and decreases as the hyperparameters increase above 1.

As a result, by choosing the divergence and its hyperparameter appropriately, greater control can be exerted over the resulting posterior than is possible with standard VI. Specifically, it allows desirable properties for the posteriors (such as conservative uncertainty quantification) to be directly and transparently incorporated into the form $P(\ell_n, D, \mathcal{Q})$ via D.

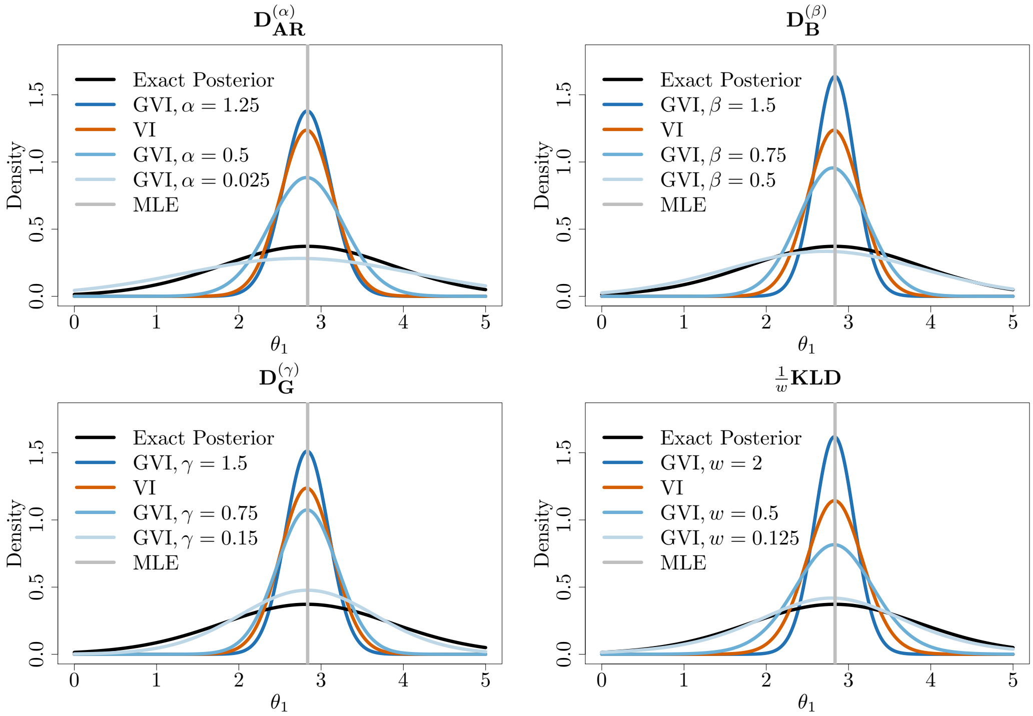

Figure 4.4: Marginal VI and GVI posteriors for the θ_1 coefficient of a Bayesian linear model under the $D_{AR}^{(\alpha)}$, $D_B^{(\beta)}$, $D_G^{(\gamma)}$ and $\frac{1}{w}$KLD prior regularisers for different values of the divergence hyperparameters. Correlated covariates cause dependency in the exact posterior of the coefficients θ, and as a result VI underestimates marginal variances. GVI has the flexibility to more accurately capture the exact marginal variances. Prior Specification: $\sigma^2 \sim \mathcal{IG}(20, 50)$, $\theta_1 \sim \mathcal{N}(0, 5^2)$ and $\theta_2 \sim \mathcal{N}(0, 5^2)$.

Robustness to the prior

In this section we compare the impact of changing the uncertainty quantifier on the posterior's sensitivity to the specification of the prior. Specifically, we consider and compare βD, Rényi-αD, γD and $\frac{1}{w}$KLD. When comparing $\frac{1}{w}$KLD with Rényi-αD and γD, we fixed $\alpha = \gamma = w$. Theorem 10, which for clarity we leave till the end of the chapter to state (Section 4.6), demonstrates that the objective function of the GVI problem $P(D_G^{(\gamma)}, \ell_n, \mathcal{Q})$ provides an upper bound on the objective function of the GVI problem $P(\frac{1}{\gamma}\text{KLD}, \ell_n, \mathcal{Q})$. This provides a connection between the value of γ and w and therefore we argue $\gamma = w$ provides a fair comparison. An analogous result for the α and Rényi-αD is available in Knoblauch et al. [2019]. The βD uncertainty quantifier required the selection of different values to ensure its availability in a closed form.

$\frac{1}{w}$**KLD** Firstly, Figure 4.5 examines how weighting the KLD impacts the solution to $P(\ell_n, \frac{1}{w}\text{KLD}, \mathcal{Q})$. Choosing $w < 1$ leads to posteriors that encourage larger variances, making them amenable to conservative uncertainty quantification. However, this comes at the price of making them *more* sensitive to the prior. Conversely $w > 1$ will result in posteriors that are less sensitive to the prior than standard VI. At the same time, they will also be more concentrated around the Maximum Likelihood Estimator. This makes the $\frac{1}{w}$KLD uncertainty quantifier unattractive: In essence, one has to choose between wider variances (at the expense of being robust to the prior) and prior robustness (at the expense of more concentrated posteriors). As we shall see, the necessity of performing this undesirable trade-off is *not* shared by the other (robust) divergences considered in this section.

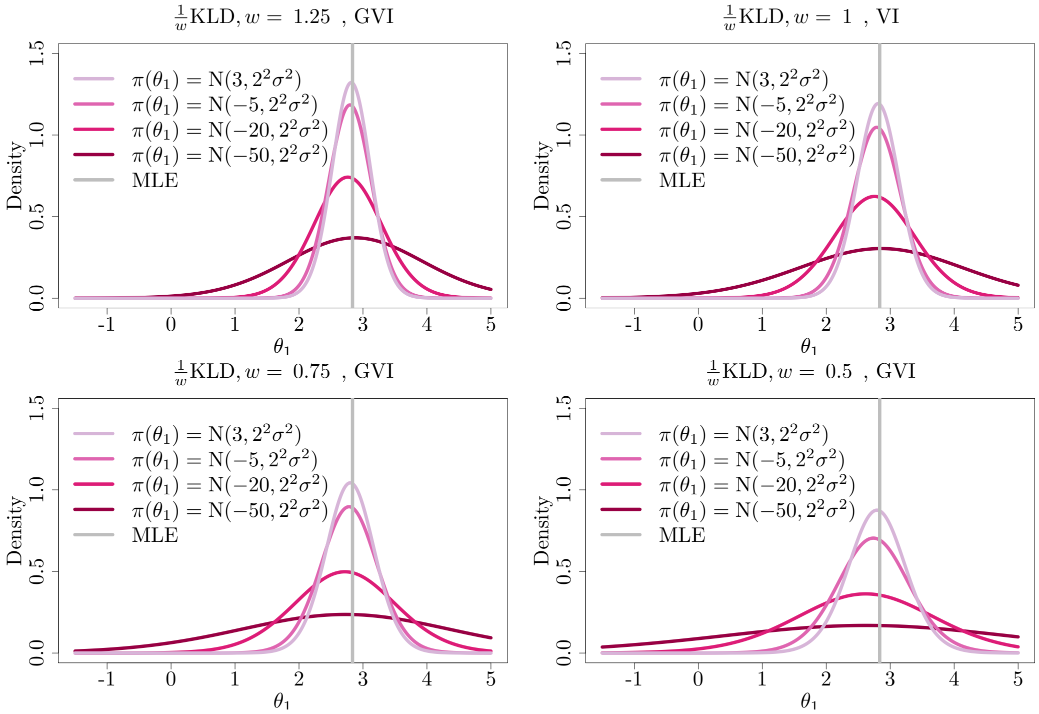

Figure 4.5: Marginal **VI** and **GVI** posteriors for the θ_1 coefficient of a Bayesian linear model under different prior specifications and the using the $\frac{1}{w}$KLD as the uncertainty quantifying divergence for several values of w. Prior specification: $\sigma^2 \sim \mathcal{IG}(3, 5)$.

Rényi-αD Figure 4.6 demonstrates the sensitivity of $P(\ell_n, D_{AR}^{(\alpha)}, \mathcal{Q})$ to prior specification. For $0 < \alpha < 1$, the RÉNYI-αD is able to provide more conservative marginal variances than standard VI while also being more robust to badly specified priors. That being said, when $\alpha > 1$ the RÉNYI-αD is more sensitive to the prior

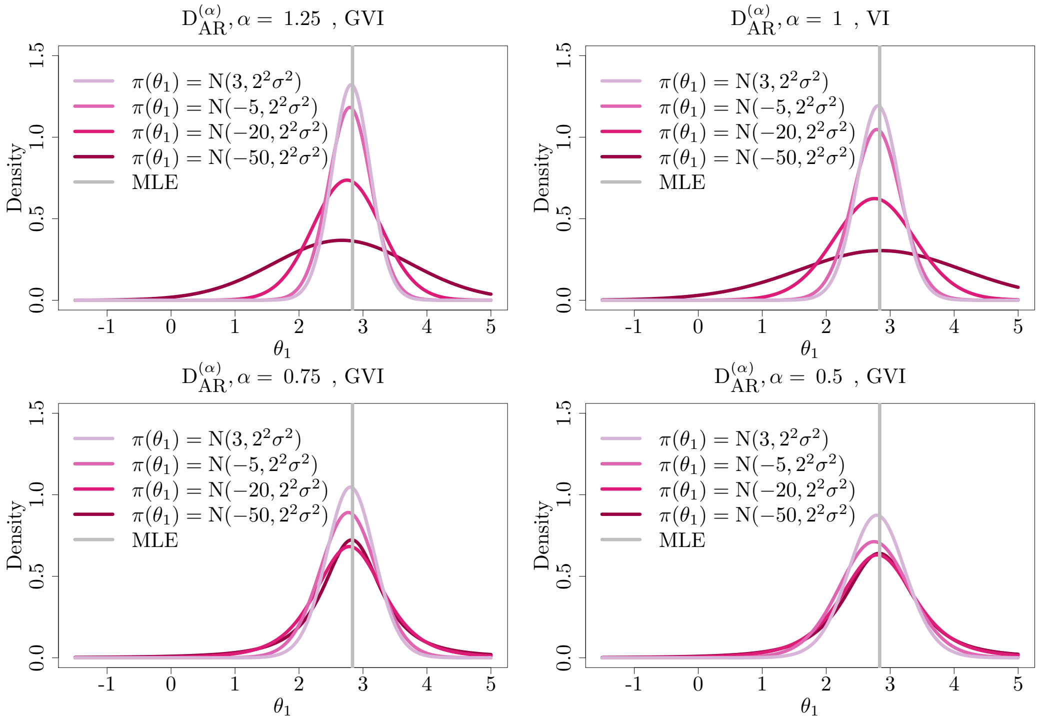

Figure 4.6: Marginal **VI** and **GVI** posteriors for the θ_1 coefficient of a Bayesian linear model under different prior specifications and the using the $D_{AR}^{(\alpha)}$ as the uncertainty quantifying divergence for several values of α. Prior specification: $\sigma^2 \sim \mathcal{IG}(3, 5)$.

than standard VI. This can be seen by examining the form of the Rényi-αD:

$$D_{AR}^{(\alpha)}(q(\theta)||\pi(\theta)) = \frac{1}{\alpha(\alpha-1)} \log \int q(\theta)^{\alpha} \pi(\theta)^{1-\alpha} d\theta \tag{4.24}$$

$$= \frac{1}{\alpha(\alpha-1)} \log \int \frac{q(\theta)^{\alpha}}{\pi(\theta)^{\alpha-1}} d\theta, \tag{4.25}$$

where we rearrange to ensure all of the powers are positive when $\alpha > 1$. There is now a ratio of densities in the Rényi-αD. This means that if $q(\theta)$ is large in an area where $\pi(\theta)$ is not, then a severe penalty is incurred. This limits how far $q(\theta)$ can move from the prior and thus results in lack of prior robustness.

β**D** Figure 4.7 demonstrates the sensitivity of $P(\ell_n, D_B^{(\beta)}, \mathcal{Q})$ to prior specification. The plot shows that $\beta > 1$ is able to achieve extreme robustness to the prior, while $\beta < 1$ causes extreme sensitivity to the prior. This phenomenon is a result of the fact that the βD decomposes into three integrals, one containing just the prior, one

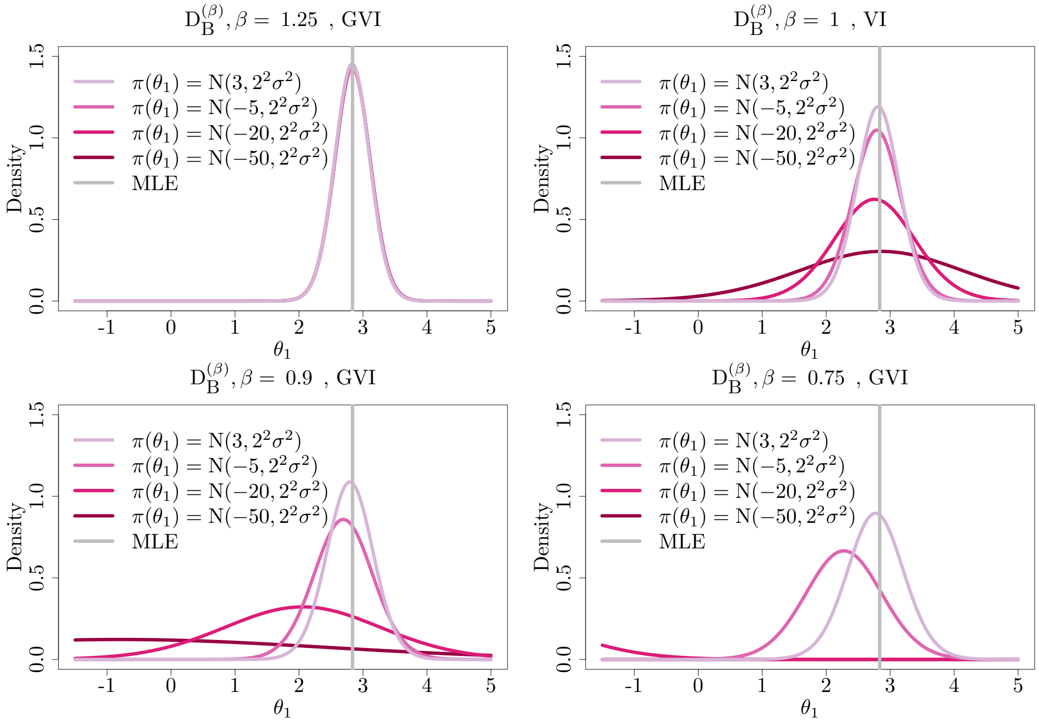

Figure 4.7: Marginal **VI** and **GVI** posteriors for the θ_1 coefficient of a Bayesian linear model under different prior specifications and using the $D_B^{(\beta)}$ as the uncertainty quantifying divergence for several values of β. Prior specification: $\sigma^2 \sim \mathcal{IG}(3, 5)$.

containing just $q(\theta)$ and one containing an interaction between them.

$$D_B^{(\beta)}(q(\theta)||\pi(\theta)) = \frac{1}{\beta}\int \pi(\theta)^\beta d\theta - \frac{1}{\beta-1}\int \pi(\theta)^{\beta-1}q(\theta)d\theta + \frac{1}{\beta(\beta-1)}\int q(\theta)^\beta d\theta.$$

The left hand integral depending only on the prior is constant in q and so we can ignore it. Now if $0 < \beta < 1$, then the signs of both of the remaining terms switch. Additionally, it is instructive to rewrite the middle term like $+\frac{1}{1-\beta}\int \frac{q(\theta)}{\pi(\theta)^{1-\beta}}d\theta$ with $1 - \beta > 0$. This shows that the prior now appears in the denominator of a fraction. The consequences of this are similar to the behaviour of the Rényi-αD for $\alpha > 1$, if $q(\theta)$ has density where the prior has little density then we are dividing a not so small number by a very small number and a huge penalty is incurred for this. As a result, the corresponding posterior will not be able to move away from the prior. In fact, notice that that two of the four posteriors for $\beta = 0.75$ favour the prior so much that there is virtually *no* mass at the Maximum Likelihood Estimate.

For $\beta > 1$ the opposite effect is observed. The prior no longer appears in the denominator of a fraction and therefore deviations from the prior are punished in a milder manner. In fact for $\beta = 1.25$ we observe almost prior invariant uncertainty quantification. We conjecture this is because the uncertainty quantification part of the Bayesian decision problem is largely controlled by right-hand term only involving $q(\theta)$. This integral is very large if the variance of $q(\theta)$ gets very small, which prevents it from converging to a point mass at the MLE as the GVI posterior using the αD did in Figure 4.3. Therefore, the βD appears to be able to to provide uncertainty quantification that is minimally impacted by the specification of the prior. This could have exciting implications within the objective Bayes literature [see, e.g. Jeffreys, 1961; Zellner, 1977; Bernardo, 1979; Berger and Bernardo, 1992; Jaynes, 2003; Berger, 2006] and is worthy of further investigation.

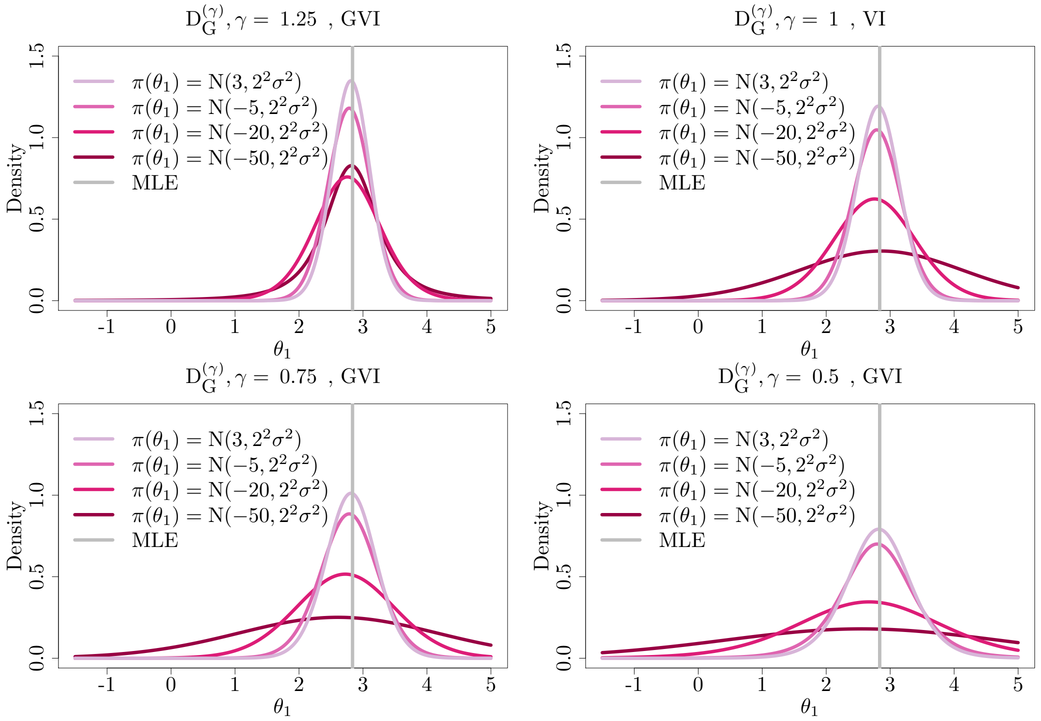

Figure 4.8: Marginal **VI** and **GVI** posteriors for the θ_1 coefficient of a Bayesian linear model under different prior specifications and the using the γD as the uncertainty quantifying divergence for several values of γ. Prior specification: $\sigma^2 \sim \mathcal{IG}(3, 5)$.

γD Lastly, Figure 4.8 demonstrates the sensitivity of $P(\ell_n, D_G^{(\gamma)}, \mathcal{Q})$ to prior specification. The γD with $\gamma > 1$ produces greater robustness to the prior than the $\frac{1}{w}$KLD uncertainty quantifier with $w > 1$. However, this robustness is not as extreme as was seen for the $D_B^{(\beta)}$. The reason for this is that although the γD consists of the exact same three terms as the βD, these terms are now logarithms. This means that the three integrals are combined multiplicatively (in the γD) rather than additively (in the βD), which makes the variation across γ much smoother than across β: Unlike for the βD, minimising the γD can no longer disregard any one term in order to minimise the others. For $\gamma < 1$ it appears as though the γD reacts similarly to the $\frac{1}{w}$KLD for $w < 1$.

Now that we have examined the power of GVI to produce more conservative estimates of marginal variance, we return to F-VI methods and examine the consequences of their failure to satisfy our axioms for Bayesian inference.

4.5 F-VI consequences

Unlike traditional VI, the F-VI techniques with $F \neq$ KLD introduced in Section 4.2.3 and other posterior approximation techniques such as integrated nested Laplace approximations (INLA) [Rue et al., 2009] do not fit into the format of a principled Bayesian GVI problem. This has the following consequences

(1) If $F \neq$ KLD, F-VI will violate Axioms 1–4.

(2) F-VI with a variational family $\mathcal{Q}$ gives provably suboptimal $\mathcal{Q}$-constrained approximations to its exact target $P(\ell_n, \text{KLD}, \mathcal{P}(\theta))$ relative to standard VI.

(3) F-VI conflates the effects of ℓ_n and D because it induces desirable properties for the posterior through F rather than through the clearly interpretable modularity of $P(\ell_n, D, \mathcal{Q})$.

Rather than building an axiomatically justified objective function to elicit desirable inferences, F-VI techniques build an approximation to the solution of the traditional Bayesian objective function. The starting point for these methods is the solution to Eq. (4.1), and then a member of the variational family is chosen to be close to this exact posterior by some criteria. Not only can this approximation be shown to provably perform worse than VI according to Eq. (4.1), but the resulting objective function for inference does not satisfy our axioms. Principally, the objective functions for F-VI will not satisfy Axiom 1. The consequences of this are that F-VI methods do not separate the target parameter for inference from the uncertainty quantifying properties of the posterior. As a result the uncertainty quantification and the target parameter are no longer two independent components that can be tuned, they interact! The example below demonstrates the dangers of not formulating a principled inference problem. Not separating the target parameter and the uncertainty quantifier cause F-VI to produce unexpected and/or unsatisfactory belief distributions. Behaviour of this type is also observed in one of the deep learning applications presented at the end of this chapter in Section 4.8 (see Section 4.8.1).

4.5.1 Multi-Modality caused by Label Switching

It is popular to demonstrate the impact of the *zero-forcing* or *zero-avoiding* nature of the divergence F used in F-VI by approximating a bimodal distribution with a uni-modal one [e.g. Minka, 2004; Hernandez-Lobato et al., 2016]. In fact, these situations provide an excellent example of the drawbacks of F-VI that we demonstrate here. A prime example of how bimodal posteriors are induced is the label switching phenomenon. This phenomenon occurs if the likelihood function is invariant to switching parameter labels. One straightforward example of this which is of great practical importance are Bayesian mixture models. Consequently, we use a Bayesian mixture model to investigate the differences between F-VI and GVI when faced with multi-modal posteriors. In particular, we conduct inference for the coefficients $\boldsymbol{\mu} = (\mu_1, \mu_2)$ in the model

$$f(x; \theta) = 0.5\mathcal{N}(x; \mu_1, \sigma^2) + 0.5\mathcal{N}(x; \mu_2, \sigma^2). \tag{4.26}$$

The bimodality in the posteriors for $\boldsymbol{\mu}$ is a consequence of the fact that $\boldsymbol{\mu} = (a, b)$ has exactly the same likelihood as $\boldsymbol{\mu} = (b, a)$. In order to compare the performance of VI, F-VI and GVI we generate $n = 100$ observations from 2 cases of this model.

case 1: $\boldsymbol{\mu} = (0, 1)$ and $\sigma^2 = 0.65^2$

case 2: $\boldsymbol{\mu} = (0, 2)$ and $\sigma^2 = 0.65^2$

We plot the exact and approximate posteriors in Figure 4.9. All variational posteriors come from a 'mean-field' Gaussian variational family, and all inference was based on the priors $\pi(\mu_1) = \pi(\mu_2) = \mathcal{N}(0, 2^2)$. Figure 4.9 shows the danger of not carefully constructing the inference problem under a multimodal posterior. The invariance to label switching in the likelihood means $-\log(f(x_i; \theta))$ is equally minimised at either mode of the exact posterior. The modular formulation of VI and GVI ensures that when q is constrained to be uni-modal then the optimal q still focuses on one of two equally good combinations of the parameter values minimising the loss. Predictively, the resulting parameter inference will therefore still perform well. Here, GVI uses the $D_{AR}^{(\alpha)}$ with $\alpha = 0.5$ and thus fits a larger posterior variance then VI. However F-VI is formulated as a posterior approximation rather than

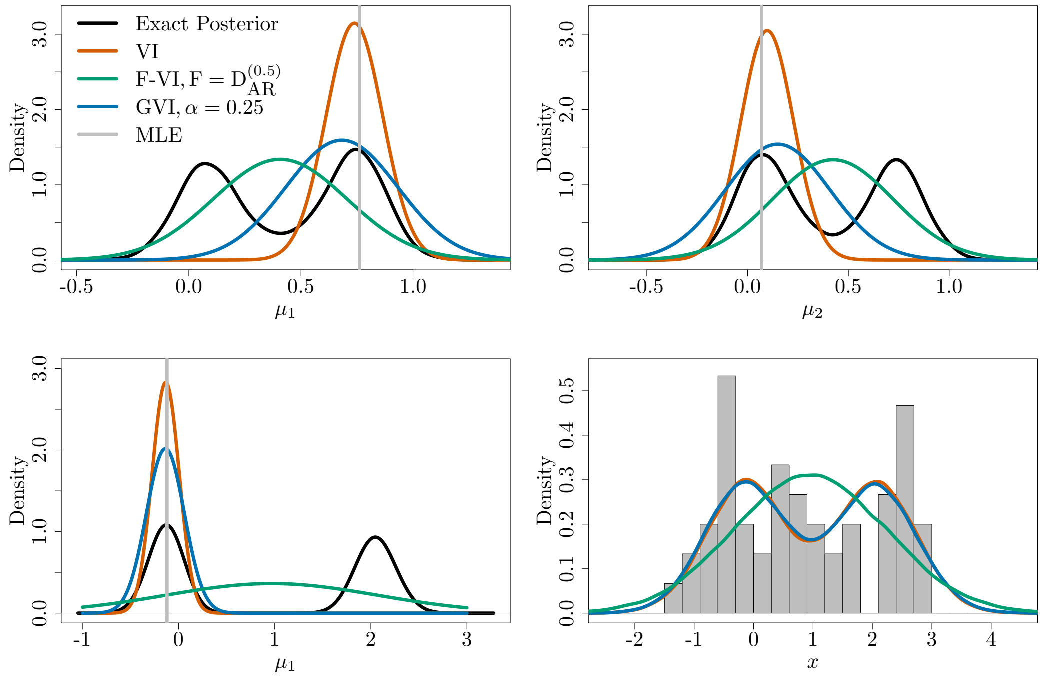

Figure 4.9: Marginal **Exact**, **VI**, $D_{AR}^{(\alpha)}$**-VI** [Li and Turner, 2016] and **GVI** using the $-\log(f(x_i;\theta))$ and $D_{AR}^{(\alpha)}$ prior regulariser posteriors for the coefficients $\theta = (\mu_1, \mu_2)$ of a 2-component mixture model. **Top:** Case 1. The exact posterior is bimodal as a result of label-switching. **VI** and **GVI** are able to concentrate on one of the combinations of parameters minimising the loss. In contrast, $D_{AR}^{(\alpha)}$**-VI** is smoothing between the two modes. Note in particular that the highest posterior mass is placed at a locally (least) likely combination of μ_1 and μ_2. **Bottom:** Case 2. Here the two components are separated further. The right hand plot shows the negative impact the $D_{AR}^{(\alpha)}$**-VI** posterior approximation has predictively.

a principled inference problem As a result, the optimal unimodal approximation focuses on approximating the exact posterior rather than minimising a regularised loss function. This causes F-VI to miss either local optimum and smooth between the two. In fact, it does the worst possible thing and concentrates on the values of μ_1 and μ_2 which locally maximize the loss/minimize the likelihood. This effect is exaggerated in case 2 where the 2 posterior modes are further separated. Here the negative effect that the F-VI approximation has on the predictive is clear.

4.6 GVI as a posterior approximation

Next, we demonstrate that although the GVI objective function is defined as an principled objective function for inference, we can still interpret its solution as approximating the Bayes' rule posterior if we wish.

One common interpretation given to VI is that its objective function,

$$L(q|\boldsymbol{x}, -\log f(x;\theta), \text{KLD}) \tag{4.27}$$

aka. the ELBO, forms a lower bound on the log-marginal-likelihood, see Eq. (4.9) Additionally the log-marginal-likelihood does not depend on θ and thus Eq. (4.9) shows that minimising $L(q|\boldsymbol{x}, -\log f(x;\theta), \text{KLD})$ is equivalent to minimising $\text{KLD}(q(\theta)||\pi(\theta|\boldsymbol{x}))$. Note that in this thesis we consider the objective to be minimised, while VI traditionally considers maximising the negation of this objective function. It is straightforward to rewrite Eq. (4.9) for the general loss function.

$$\begin{aligned} -L(q|\boldsymbol{x}, \ell(\theta,x), \text{KLD}) &= \log \int \exp\left(-\ell(\theta,x)\right)\pi(\theta)d\theta - \text{KLD}(q(\theta)||\pi^{\ell}(\theta|\boldsymbol{x})) \\ \Rightarrow -L(q|\boldsymbol{x}, \ell(\theta,x), \text{KLD}) &\leq \log \int \exp\left(-\ell(\theta,x)\right)\pi(\theta)d\theta. \end{aligned} \tag{4.28}$$

where $\pi^{\ell}(\theta|\boldsymbol{x}) \propto \pi(\theta)\exp\left(-\ell(\theta,x)\right)$ is the GBI posterior under loss function $\ell(\theta,x)$. We call the normaliser of this posterior,

$$\int \exp\left(-\ell(\theta,x)\right)\pi(\theta)d\theta \tag{4.29}$$

the marginal loss-likelihood.

Theorem 10 now proves an analogue to Eq. (4.28) using the γD prior regulariser. The interested reader can note that in the Appendix of Knoblauch et al. [2019] we report and prove similar results for the RÉNYI-αD and βD.

Theorem 10 (Lower bounding the marginal loss-likelihood using the $D_G^{(\gamma)}$ uncertainty quantifier)**.** The objective function, $L(q|\boldsymbol{x}, D_G^{(\gamma)}, \ell_n)$, associated with Bayesian problem, $P(D_G^{(\gamma)}, \ell_n, Q)$, can be used to lower bound the marginal loss-likelihood (normalising constant) of the GBI posterior in the following ways:

- If $0 < \gamma < 1$

$$\begin{aligned} L(q|D_G^{(\gamma)}, \ell_n) \geq \text{KLD}(q(\theta)||\pi^{\ell}(\theta|\boldsymbol{x})) - \log \int \pi(\theta)\exp(-\ell(\theta,x))d\theta \\ + \frac{1}{\gamma(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[q(\theta)^{\gamma-1}\right] - \mathbb{E}_{q(\theta)}\left[\log q(\theta)\right] \end{aligned} \tag{4.30}$$

where $\pi^{\ell}(\theta|\boldsymbol{x}) = \frac{\pi(\theta)\exp(-\ell_n(\theta,\boldsymbol{x}))}{\int \pi(\theta)\exp(-\ell_n(\theta,\boldsymbol{x}))d\theta}$.

- If $\gamma > 1$

$$L(q|D_G^{(\gamma)}, \ell_n) \geq \frac{1}{\gamma}\text{KLD}(q(\theta)||\pi^{\gamma\ell}(\theta|\boldsymbol{x})) - \frac{1}{\gamma}\log\int \pi(\theta)\exp(-\gamma\ell(\theta,x))d\theta$$
$$+\frac{1}{\gamma}\mathbb{E}_{q(\theta)}\left[\log\pi(\theta)\right] - \frac{1}{(\gamma-1)}\log\mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\gamma-1}\right] \qquad (4.31)$$

where $\pi^{\gamma\ell}(\theta|\boldsymbol{x}) = \frac{\pi(\theta)\exp(-\gamma\ell_n(\theta,\boldsymbol{x}))}{\int \pi(\theta)\exp(-\gamma\ell_n(\theta,\boldsymbol{x}))d\theta}$.

Proof. Firstly we note that the objective function associated with the Bayesian problem $P(D_G^{(\gamma)}, \ell_n, Q)$ can be simplified by removing the terms in the $D_G^{(\gamma)}$ that do not depend on $q(\theta)$

$$\arg\min_{q\in\mathcal{Q}}\left\{\mathbb{E}_{q(\theta)}\left[\ell(\theta,x)\right] + D_G^{(\gamma)}(q(\theta)||\pi(\theta))\right\} =$$
$$\arg\min_{q\in\mathcal{Q}}\left\{\mathbb{E}_{q(\theta)}\left[\ell(\theta,x)\right] + \frac{1}{\gamma(\gamma-1)}\log\mathbb{E}_{q(\theta)}\left[q(\theta)^{\gamma-1}\right] - \frac{1}{(\gamma-1)}\log\mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\gamma-1}\right]\right\}$$
$$=\arg\min_{q\in\mathcal{Q}}\left\{L(q|\boldsymbol{x}, \ell_n, D_G^{(\gamma)})\right\}$$

We have to consider two cases for γ as the positivity and negativity of $\gamma - 1$ affect the results that can be use.

Case 1) $0 < \gamma < 1$**:** The definition of the γD provides the following GVI objective function

$$\begin{aligned}&L(q|\boldsymbol{x}, D_G^{(\gamma)}, \ell_n)\\ &= \mathbb{E}_{q(\theta)}\left[\ell(\theta,x)\right] + \frac{1}{\gamma(\gamma-1)}\log\mathbb{E}_{q(\theta)}\left[q(\theta)^{\gamma-1}\right] - \frac{1}{(\gamma-1)}\log\mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\gamma-1}\right]\\ &= \mathbb{E}_{q(\theta)}\left[\ell(\theta,x)\right] + \frac{1}{\gamma(\gamma-1)}\log\mathbb{E}_{q(\theta)}\left[q(\theta)^{\gamma-1}\right] + \frac{1}{(1-\gamma)}\log\mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\gamma-1}\right].\end{aligned}$$

Jensen's inequality and the concavity of the natural logarithm applied to $\mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\gamma-1}\right]$ provides

$$\begin{aligned}
&L(q|\boldsymbol{x}, D_G^{(\gamma)}, \ell_n)\\
&= \mathbb{E}_{q(\theta)}\left[\ell(\theta, x)\right] + \frac{1}{\gamma(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[q(\theta)^{\gamma-1}\right] + \frac{1}{(1-\gamma)}\log \mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\gamma-1}\right].\\
&\geq \mathbb{E}_{q(\theta)}\left[\ell(\theta, x)\right] + \frac{1}{\gamma(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[q(\theta)^{\gamma-1}\right]\\
&\quad + \frac{1}{(1-\gamma)}\mathbb{E}_{q(\theta)}\left[(\gamma-1)\log \pi(\theta)\right] \qquad (4.32)\\
&= \mathbb{E}_{q(\theta)}\left[\ell(\theta, x)\right] + \frac{1}{\gamma(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[q(\theta)^{\gamma-1}\right] - \mathbb{E}_{q(\theta)}\left[\log \pi(\theta)\right]\\
&= \frac{1}{\gamma(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[q(\theta)^{\gamma-1}\right] - \mathbb{E}_{q(\theta)}\left[\log\left(\pi(\theta)\exp\left(-\ell(\theta, x)\right)\right)\right]. \qquad (4.33)
\end{aligned}$$

The final term above collected the term involving $\pi(\theta)$ and $\ell(\theta, x)$ to produce a term which looks like the GBI posterior. Next the normaliser of this, $\log \int \pi(\theta)\exp\left(-\ell(\theta, x)\right) d\theta$, is added and subtracted

$$\begin{aligned}
&L(q|\boldsymbol{x}, D_G^{(\gamma)}, \ell_n)\\
&\geq \frac{1}{\gamma(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[q(\theta)^{\gamma-1}\right] - \mathbb{E}_{q(\theta)}\left[\log\left(\pi(\theta)\exp\left(-\ell(\theta, x)\right)\right)\right]\\
&= \frac{1}{\gamma(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[q(\theta)^{\gamma-1}\right] - \mathbb{E}_{q(\theta)}\left[\log \frac{\pi(\theta)\exp\left(-\ell(\theta, x)\right)}{\int \pi(\theta)\exp\left(-\ell(\theta, x)\right) d\theta}\right]\\
&\quad - \log \int \pi(\theta)\exp\left(-\ell(\theta, x)\right) d\theta\\
&= \frac{1}{\gamma(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[q(\theta)^{\gamma-1}\right] + \text{KLD}(q(\theta)||\pi^{\ell}(\theta|x)) \qquad (4.34)\\
&\quad - \log \int \pi(\theta)\exp\left(-\ell(\theta, x)\right) d\theta - \mathbb{E}_{q(\theta)}\left[\log q(\theta)\right].
\end{aligned}$$

The final line above added and subtracted -$\mathbb{E}_{q(\theta)}\left[\log q(\theta)\right]$ in order to obtain the KLD$(q(\theta)||\pi^{\ell}(\theta|x))$ term. This gives Eq. (4.30).

Unfortunately, we cannot perform the same trick when $\gamma > 1$ as $\frac{1}{1-\gamma}$ is no longer positive and the inequality would be reversed.

Case 2) $\gamma > 1$: The definition of the γD provides the following GVI objective function where we multiply the second term by $1 = \frac{\pi(\theta)^{\gamma-1}}{\pi(\theta)^{\gamma-1}}$

$$
\begin{aligned}
&L(q|\boldsymbol{x}, D_G^{(\gamma)}, \ell_n)\\
&= \mathbb{E}_{q(\theta)}\left[\ell(\theta,x)\right] + \frac{1}{\gamma(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[q(\theta)^{\gamma-1}\right] - \frac{1}{(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\gamma-1}\right]\\
&= \mathbb{E}_{q(\theta)}\left[\ell(\theta,x)\right] + \frac{1}{\gamma(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[q(\theta)^{\gamma-1}\frac{\pi(\theta)^{\gamma-1}}{\pi(\theta)^{\gamma-1}}\right]\\
&-\frac{1}{(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\gamma-1}\right]. \qquad (4.35)
\end{aligned}
$$

Jensen's inequality and the concavity of the natural logarithm applied to $\mathbb{E}_{q(\theta)}\left[q(\theta)^{\gamma-1}\frac{\pi(\theta)^{\gamma-1}}{\pi(\theta)^{\gamma-1}}\right]$ provides

$$
\begin{aligned}
&L(q|\boldsymbol{x}, D_G^{(\gamma)}, \ell_n)\\
&= \mathbb{E}_{q(\theta)}\left[\ell(\theta,x)\right] + \frac{1}{\gamma(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[q(\theta)^{\gamma-1}\frac{\pi(\theta)^{\gamma-1}}{\pi(\theta)^{\gamma-1}}\right] - \frac{1}{(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\gamma-1}\right]\\
&\geq \mathbb{E}_{q(\theta)}\left[\ell(\theta,x)\right] + \frac{1}{\gamma}\mathbb{E}_{q(\theta)}\left[\log\frac{q(\theta)\pi(\theta)}{\pi(\theta)}\right] - \frac{1}{(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\gamma-1}\right]\\
&= \frac{1}{\gamma}\mathbb{E}_{q(\theta)}\left[\log\frac{q(\theta)}{\pi(\theta)\exp\left(-\gamma\ell(\theta,x)\right)}\right] + \frac{1}{\gamma}\mathbb{E}_{q(\theta)}\left[\log\pi(\theta)\right]\\
&-\frac{1}{(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\gamma-1}\right] \qquad (4.36)
\end{aligned}
$$

The final term above collected the term involving $\pi(\theta)$ and $\gamma\ell(\theta,x)$ to produce a term which looks like the GBI posterior. Next the normaliser of this, $\log\int\pi(\theta)\exp\left(-\gamma\ell(\theta,x)\right)d\theta$, is added and subtracted

$$
\begin{aligned}
&L(q|\boldsymbol{x}, D_G^{(\gamma)}, \ell_n)\\
&\geq \frac{1}{\gamma}\mathbb{E}_{q(\theta)}\left[\log\frac{q(\theta)}{\pi(\theta)\exp\left(-\gamma\ell(\theta,x)\right)}\right] + \frac{1}{\gamma}\mathbb{E}_{q(\theta)}\left[\log\pi(\theta)\right]\\
&-\frac{1}{(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\gamma-1}\right] \qquad (4.37)\\
&= \frac{1}{\gamma}\mathbb{E}_{q(\theta)}\left[\log\frac{q(\theta)}{\frac{\pi(\theta)\exp(-\gamma\ell(\theta,x))}{\int\pi(\theta)\exp(-\gamma\ell(\theta,x))d\theta}}\right] - \frac{1}{\gamma}\int\pi(\theta)\exp\left(-\gamma\ell(\theta,x)\right)d\theta\\
&+\frac{1}{\gamma}\mathbb{E}_{q(\theta)}\left[\log\pi(\theta)\right] - \frac{1}{(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\gamma-1}\right]\\
&= \frac{1}{\gamma}\text{KLD}(q(\theta)||\pi^{\gamma\ell}(\theta|x)) - \frac{1}{\gamma}\int\pi(\theta)\exp\left(-\gamma\ell(\theta,x)\right)d\theta\\
&+\frac{1}{\gamma}\mathbb{E}_{q(\theta)}\left[\log\pi(\theta)\right] - \frac{1}{(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\gamma-1}\right] \qquad (4.38)
\end{aligned}
$$

Where the final line simply uses the definition of the KLD. This gives Eq. (4.31).

□

We term this result 'Lower bounding the marginal loss-likelihood using the $D_G^{(\gamma)}$ uncertainty quantifier' because we can easily bring the log-marginal loss-likelihood in Eq. (4.30) and (4.31) to the LHS and the GVI objective function to the RHS and get something similar to Eq. (4.28). However we cannot necessarily be sure that the final two terms of Eq. (4.30) and (4.31)) will be positive. And thus we cannot be sure that the GVI objective function itself will lower-bound the marginal loss-likelihood. Instead, we present the results as we have done as they provide an interpretation of GVI as a posterior approximation. They provide a lower bound on an objective function that is to be minimised. As a result minimising the GVI objective function will be making the lower-bound small.

These lower bounds can be split into three terms: The negative log-marginal loss-likelihood which is independent of $q(\theta)$; the KLD between $q(\theta)$ and the general Bayesian posterior with loss ℓ and calibration weight w, termed $\pi^{w\ell}$ here for simplicity, and then an adjustment term.

Interpretation: Eq. (4.30) shows that when $\gamma \in (0,1)$ minimising $L(q|D_G^{(\gamma)}, \ell_n)$ trades off minimizing $\text{KLD}(q||\pi^{\ell})$ with minimizing adjustment term

$$T_{\text{G}}^{(0,1)}(q) = \frac{1}{\gamma(\gamma-1)} \log \mathbb{E}_{q(\theta)}\left[q(\theta)^{\gamma-1}\right] - \mathbb{E}_{q(\theta)}\left[\log q(\theta)\right]. \tag{4.39}$$

While $\text{KLD}(q||q^*)$ is the same target as in traditional VI applied to posterior π^{ℓ}, it is straightforward to show that the adjustment term encourages the solution to $P(D_G^{(\gamma)}, \ell_n, Q)$ with $0 < \gamma < 1$ to have greater variance than that of $P(\text{KLD}, \ell_n, Q)$ (i.e., VI), as evidenced by Figure 4.4. Specifically, one can rewrite

$$T_{\text{G}}^{(0,1)}(q) = -\frac{1}{\gamma} h_R^{(\gamma)}(q(\theta)) + h_{\text{KLD}}(q(\theta)), \tag{4.40}$$

where $h_{\text{KLD}}(q(\theta))$ is the Shannon entropy of $q(\theta)$ and $h_R^{(\gamma)}(q(\theta))$ is the Rényi entropy of $q(\theta)$ with parameter γ. Now Theorem 3 in Van Erven and Harremos [2014] can be extended to show that $h_R^{(\gamma)}(q(\theta))$ is decreasing in γ. Since it is also well-known that $\lim_{\gamma\to 1} h_R^{(\gamma)}(q(\theta)) = h_{\text{KLD}}(q(\theta))$, it follows that minimising $-\frac{1}{\gamma}h_R^{(\gamma)}(q(\theta)) + h_{\text{KLD}}(q(\theta))$ for $0 < \gamma < 1$ will make $h_R^{(\gamma)}(q(\theta))$ large – an effect that is achieved by increasing the variance of $q(\theta)$.

Alternatively Eq. (4.31) shows that when $\gamma \in (1, \infty)$ minimising $L(q|D_G^{(\gamma)}, \ell_n)$

is trading off making KLD$(q||\pi^{\gamma\ell})$ small with also making the adjustment term

$$T_{\mathrm{G}}^{(1,\infty)}(q) = \frac{1}{\gamma}\mathbb{E}_{q(\theta)}\left[\log \pi(\theta)\right] - \frac{1}{(\gamma-1)}\log \mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\gamma-1}\right]. \tag{4.41}$$

small. Minimising KLD$(q||q^*_\gamma)$ for $\gamma > 1$ will encourage the solution of $P(D_G^{(\gamma)}, \ell_n, Q)$ to be more concentrated than minimising KLD$(q||\pi^\ell)$. We can also show that the adjustment terms $T_{\mathrm{G}}^{(1,\infty)}(q)$ encourage shrinkage of the variance of $q(\theta)$ as evidenced by Figure 4.4. Jensen's inequality shows that for $\gamma > 1$

$$\frac{1}{\gamma-1}\log \mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\gamma-1}\right] \geq \mathbb{E}_{q(\theta)}\left[\log(\pi(\theta))\right] \geq \frac{1}{\gamma}\mathbb{E}_{q(\theta)}\left[\log(\pi(\theta))\right]. \tag{4.42}$$

As a result minimising $T_{\mathrm{G}}^{(1,\infty)}(q)$ will seek to make $\frac{1}{\gamma-1}\log \mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\gamma-1}\right]$ large. Fixing $\pi(\theta)$, maximising $\frac{1}{\gamma-1}\log \mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\beta-1}\right]$ plus $\frac{1}{\gamma}$ times the Rényi entropy of $q(\theta)$ is equivalent to minimising $D_G^{(\gamma)}(q(\theta)||\pi(\theta))$, and thus seeks $q(\theta)$ close to $\pi(\theta)$. The Rényi entropy term would have acted to increase the variance of $q(\theta)$ and therefore maximising $\frac{1}{\gamma-1}\log \mathbb{E}_{q(\theta)}\left[\pi(\theta)^{\gamma-1}\right]$ without adding the Rényi entropy will lead to shrinkage of the variance of $q(\theta)$.

These results are somewhat reassuring. However as we have mentioned, we prefer to treat GVI as a principled objective function for generating inference in its own right.

4.7 Black-Box GVI

In order to implement GVI beyond the simple linear regression example in Section 4.4.2, Stochastic Gradient Descent (SGD) was required. In most situations (see Knoblauch et al. [2019] Appendix-Section 5) the GVI objective function can be represented as an expectation over the variational family. In this situation the 'log-derivative-trick' can be used to write the gradient of the expectation as an expectation (with respect to the variational family) of a gradient which can then be estimated in an unbiased fashion by Monte-Carlo. Therefore SGD methods take steps in the direction of the estimated gradient until some stopping criteria is reached. In our experiments we specifically used the ADAM [Kingma and Ba, 2014] version of SGD.

A key part of a successful and efficient SGD optimiser is being able to reduce the variance of the Monte Carlo estimates of the gradient. In Knoblauch et al. [2019] we build upon the black-box variance reduction techniques of Ranganath et al. [2014] to produce a black-box implementation of GVI implementing control variates and antithetic variates to reduce variance[1].

4.8 Experiments

In this section we compare GVI to the state of the art variational inference procedures for two popular deep learning models, Bayesian Neural Network (BNN) regression and Deep Gaussian Process (GP) regression.

4.8.1 Bayesian Neural Network (BNN) regression

One major application of VI and F-VI techniques is to Bayesian Neural Network (BNN) regression [Neal, 2012]. BNN regression seeks to learn the relationship between some univariate[2] response Y and a set of p predictors $\boldsymbol{X}$, where the mean function of $Y|X$ is parametrised by a Neural Network with $L-1$ hidden layers.

$$\begin{aligned} \boldsymbol{h}_{il} &= f_l\left(\theta_l \cdot \boldsymbol{h}_{il-1}\right), \quad l = 1, \dots, L-1, \quad \boldsymbol{h}_{i0} = \boldsymbol{X}_i && (4.43)\\ M_{iL} &= f_L\left(\theta_{L-1} \cdot \boldsymbol{h}_{iL-1}\right) && (4.44)\\ Y_i &\sim \mathcal{N}\left(M_L\left(\theta, \boldsymbol{X}_i\right), \sigma^2\right) && (4.45) \end{aligned}$$

In order to facilitate comparisons with the current F-VI literature we follow the set-up of Hernandez-Lobato et al. [2016]; Li and Turner [2016]. In this case $M\left(\theta, \boldsymbol{X}\right)$ is a single hidden layer Neural Network with 50 nodes. The ReLU activation function, $f_l(\theta, \boldsymbol{h})$ =ReLU$(\theta, \boldsymbol{h}) = \max(0, \theta \cdot \boldsymbol{h})$, is used for each of these nodes. The parameter vector is $\theta = \{\theta_1, \dots, \theta_L\}$ with each θ_l containing the biases and weights for each layer of the network. Each member of θ has a standard Gaussian prior and the Gaussian mean-field family is optimised over $\mathcal{Q} = \left\{q(\theta; \boldsymbol{\mu}, \boldsymbol{\xi}^2) = \mathcal{N}_p\left(\boldsymbol{\mu}, \text{diag}\left(\boldsymbol{\xi}^2\right)\right)\right\}$. The residual variance, σ^2, is considered a nuisance parameter and is estimated by a procedure commonly referred to as Type-II maximum likelihood. This produces a point estimate for σ^2 by optimising the variational objective function over values for σ^2 as well as the variational family. This procedure is motivated by the fact that the objective functions of Hernandez-Lobato et al. [2016]; Li and Turner [2016] provide a lower bound to the marginal likelihood, the criteria Bayesian's would typically maximise to produce point estimates for nuisance parameters. We critically analyse the affects of this procedure in Section 4.8.1.

[1] I note that my co-authors on this paper derived and implemented these black-box variance reduction methods. However I implemented the ADAM SGD algorithm to produce both GVI and F-VI posterior approximation in the label switching case Figure 4.9.

[2] Y could in general have dimension greater than 1

The BNN's are applied to several datasets from the UCI repository [Lichman et al., 2013], with typical benchmark settings [as in Hernandez-Lobato et al., 2016; Li and Turner, 2016]. We compare F-VI using the αD [Hernandez-Lobato et al., 2016] and the RÉNYI-αD [Li and Turner, 2016][3] with VI and GVI using the $D_{AR}^{(\alpha)}$ prior regulariser and the log-score loss function. The methods are evaluated based on both the posterior expected log-score and the root-mean-squared error (RMSE) averaged over 50 splits with 90% training and 10% test data. The RMSE provides some idea of how well located the posteriors are while the posterior expected log-score additionally evaluates how well the posterior captures the finite sample uncertainty. See the Appendix of Knoblauch et al. [2019] for more information on the set-up of these experiments. We note that the $D_{AR}^{(\alpha)}(q||\pi)$ with $\alpha = 0$ corresponds to $\text{KLD}(\pi||q)$ (see Definition 13), the direction of the KLD normally associated with expectation propagation (EP), the opposite to that considered by VI and generally considered to have much greater zero-avoiding behaviour. Figure 4.10 presents the comparisons of these inference procedures.

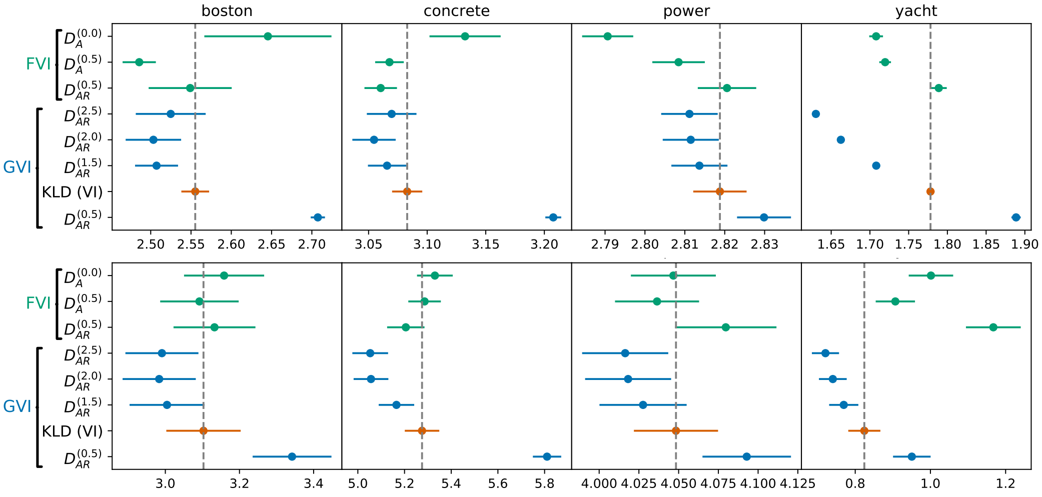

Figure 4.10: Comparing test set performance on the BNN between **F-VI**, **GVI** with alternative choices for D, and **VI**. **Top row**: Negative test log likelihoods. **Bottom row**: Test RMSE. The lower the better.

Firstly, remember that setting the prior regularising divergence as $D = D_{AR}^{(\alpha)}$ with $\alpha \in (0, 1)$ causes the GVI posterior to be less concentrated than VI. Conversely, GVI with $D = D_{AR}^{(\alpha)}$ provides more concentrated posteriors if $\alpha > 1$. Figure 4.10 shows that GVI's test performance is a banana shaped curve in α: Over-concentration relative to VI is an advantage for test set prediction, but concen-

[3] We report results with our parameterization of αD which is equivalent to using the one in [Li and Turner, 2016] for $1 - \alpha$

trating too far affects performance adversely. In fact the posteriors produced by $P(\ell_n, D_{AR}^{(\alpha)}, \mathcal{Q})$ for all settings of $\alpha > 1$ uniformly beat VI and for all settings of $\alpha < 1$ are beaten by VI. This suggests that for BNNs, one should select D to concentrate slightly *more* than standard VI, not less. This agrees with results such as [Bowman et al., 2015; Rossi et al., 2019] which observe that BNNs are generally over-parametrised functions for the DGP and thus the likelihood functions need up-weighting (prior down-weighting) to allow sufficient informations to be learnt about all parameters. The fact that performance eventually decrease with too greater concentration agrees with the well-known fact the Bayesian methods generally out perform frequentist MLE's on test sets because they provide some quantification of finite sample uncertainty.

Figure 4.10 additionally shows that the F-VI methods also outperform VI. However, F-VI does so using the αD and RÉNYI-αD approximating divergences with ($\alpha < 1$) chosen for their 'zero-avoiding' properties to increase posterior variances. The results of this section thus appear incoherent. GVI is able to improve on the performance of VI for BNNs by producing more concentrated posteriors, while F-VI improves on the performance of VI by producing less concentrated posterior. This is investigates further next in Section 4.8.1.

Hyperparameter optimisation

Following the apparent inconsistent performance of GVI and F-VI methods on the BNN experiments above, we provide a further investigation into these results. In order to investigate the levels of over and under concentration of both GVI and F-VI relative to traditional VI we consider plotting the variational parameter posteriors. Rather than do so for each of the many BNN parameters, we investigate the variational posterior for the mean function of the response given predictor values $\boldsymbol{x}$, $M(\theta, \boldsymbol{x})$ (given in Eq. (4.44)). This provides a summary of the posterior distribution for all of the parameters.

Fig 4.11 shows the posterior for the mean function, $q(M(\theta, x_i)|\boldsymbol{x})$, of the response given the predictors of three observations from the test set of one of the test/train splits of the data. Here we can see that the 'zero-avoiding' properties of the αD and RÉNYI-αD for $\alpha < 1$ ($\alpha = 1$ was the KLD associated with VI) when used in F-VI. These methods produce posteriors with marginally greater variances than were produced under VI. So here the divergence F is having the desired impact. Additionally, note how the GVI techniques with prior regularisers that are stronger than those used in VI ($\alpha < 1$), are also able to produce more conservative posterior variances, and are able to do so to a greater extent than the F-VI methods.

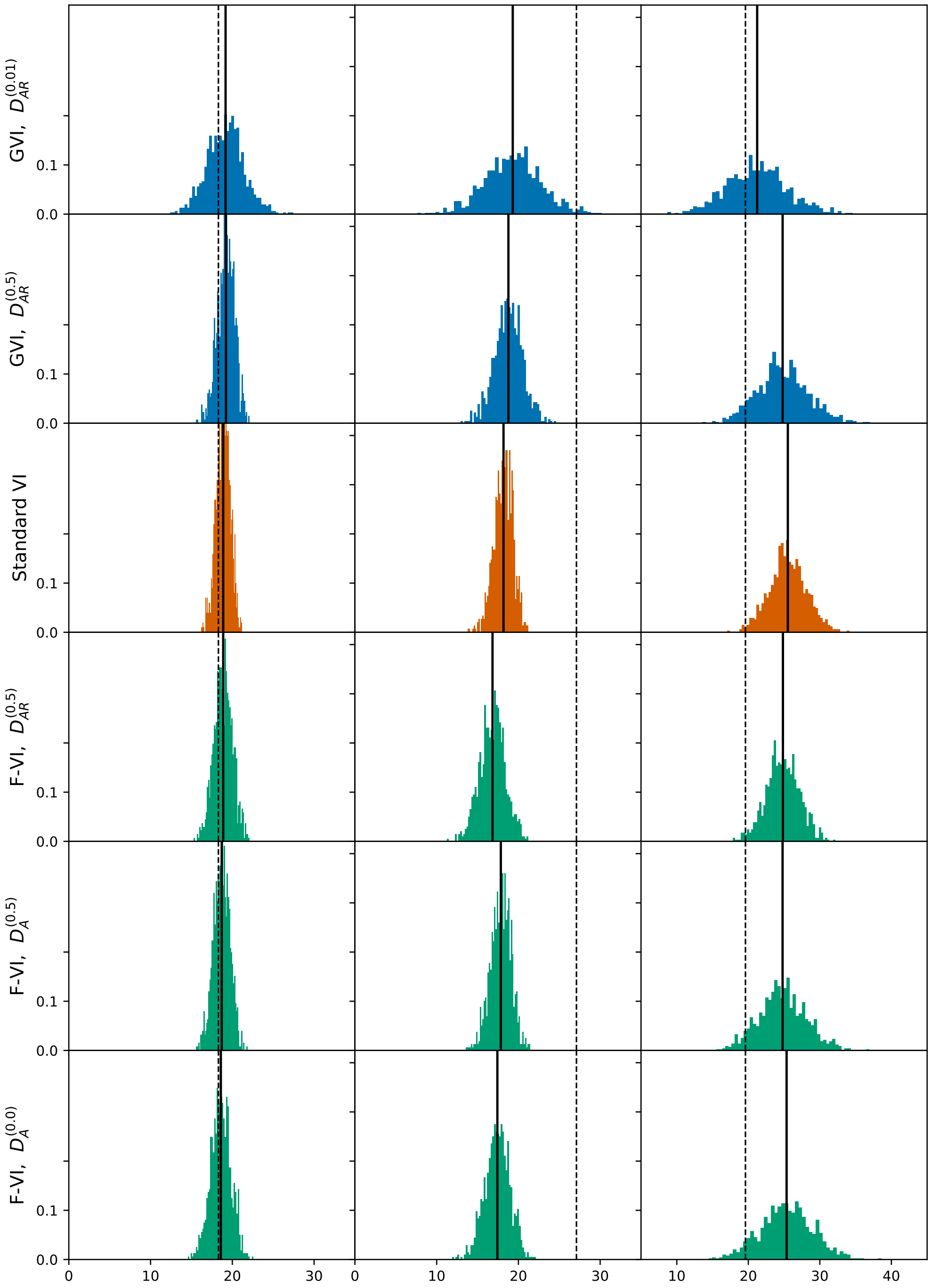

Figure 4.11: Variational posterior $q(M(\theta, x_i)|\boldsymbol{x})$ over M, the estimated mean of the response variable, for **VI**, the $D_A^{(\alpha)}$**-VI** method of Hernandez-Lobato et al. [2016], the $D_{AR}^{(\alpha)}$**-VI** method of Li and Turner [2016] and **GVI** for $D = D_{AR}^{(\alpha)}$ on three test points on the Boston data sets; based on 1,000 samples each. One can see that the posteriors over θ inherit the zero-avoiding properties of their approximating divergence as expected. Thus, they produce flatter variances. Note that the **GVI** methods with more conservative uncertainty quantification also provide flatter posterior variances over θ.

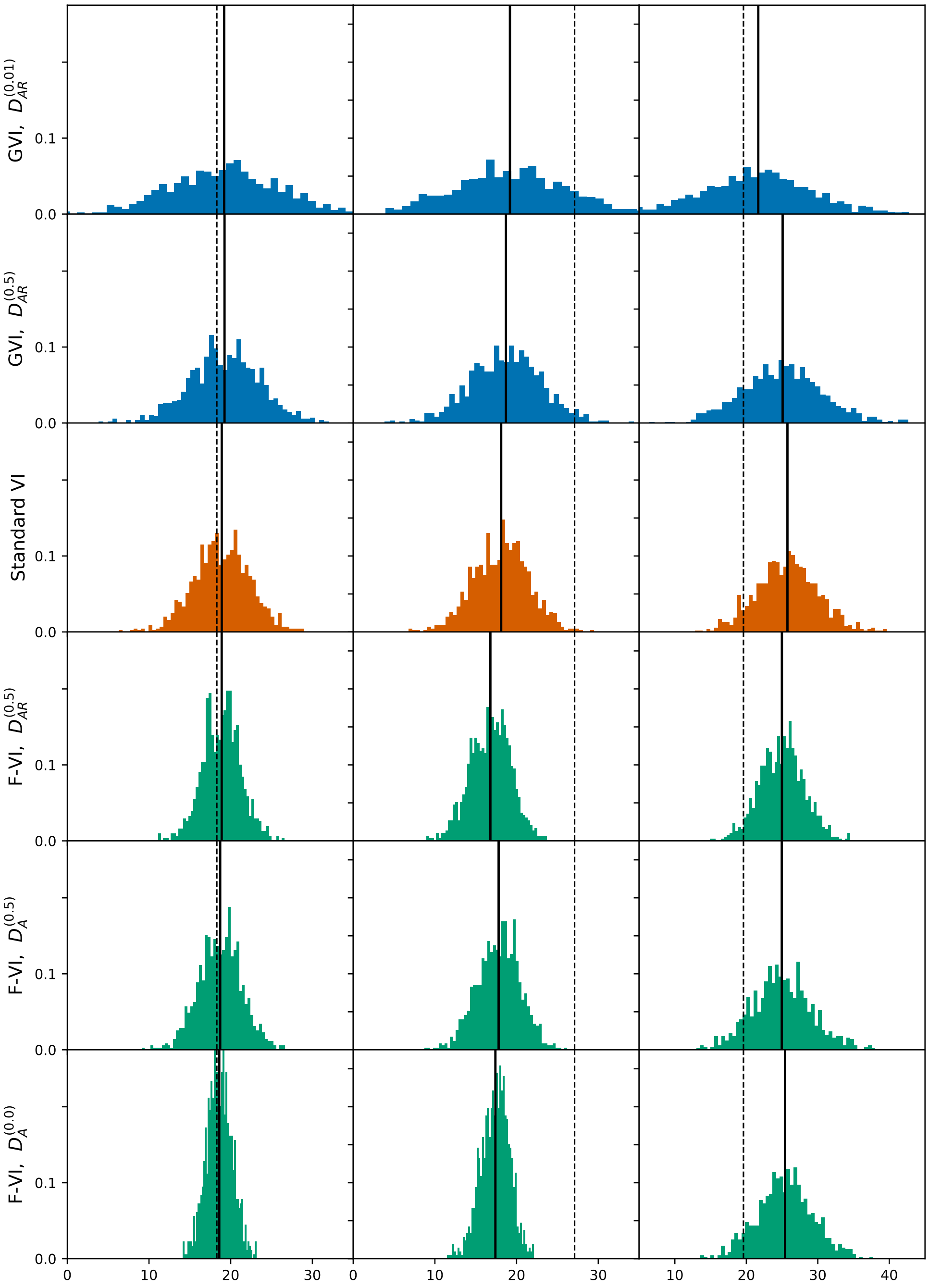

Figure 4.12: Posterior predictives $q(y|\boldsymbol{x})$ for **VI**, the $D_A^{(\alpha)}$**-VI** method of Hernandez-Lobato et al. [2016], the $D_{AR}^{(\alpha)}$**-VI** method of Li and Turner [2016] and **GVI** for $D = D_{AR}^{(\alpha)}$ on three test points on the Boston data sets. Notice that relative to standard **VI**, all **F-VI** posterior predictives are *more* contracted. In contrast, **GVI** with a more conservative uncertainty quantifier does what one would expect zero-avoiding **F-VI** methods to do. Thus, while **F-VI** may provide flatter marginal variances in the (variational) posterior for θ, this does not translate into the predictive.

In Figure 4.11 the posteriors appear to be behaving in the way we expected them to. Therefore, to investigate the inconsistent performance observed in Figure 4.10 we investigate the posterior predictive distributions for the same three test points. These are plotted in Figure 4.12. The posterior predictive combines the uncertainty in the posterior mean for each of the three data points (Figure 4.11) with the point estimate of the residual variance, $\hat{\sigma}^2$ estimated by the Type II maximum likelihood approach explained in Section 4.8.1. Although the F-VI methods fitted larger posterior variance than were estimated under VI for the mean function, the posterior predictive distributions have much smaller variances than those estimated by VI. This is strange as Bayesian analyses should propagate an increase in posterior parameter uncertainty to the posterior predictive distributions.

Table 4.2: Comparing the value of $\widehat{\sigma}^2$ for different $\mathcal{Q}$-constrained posterior inference methods (GVI with Rényi's α-divergence uncertainty quantifier and the F-VI methods of [Li and Turner, 2016] and [Hernandez-Lobato et al., 2016]). For F-VI methods, $\hat{\sigma}^2$ produces a substitution effect because it directly affects the target about which uncertainty is quantified. For GVI methods, uncertainty quantification and loss are additively separated, which prevents this substitution effect.

	GVI				F-VI		
	$D = D_{AR}^{(1.25)}$	**VI**	$D = D_{AR}^{(0.5)}$	$D = D_{A}^{(0.01)}$	$D_{AR}^{(0.5)}$	$D_{A}^{(0.5)}$	$D_{A}^{(0.0)}$
$\widehat{\sigma}^2$	9.225	**10.797**	16.811	39.533	4.016	5.856	0.911

The decrease in the variance of the predictive distributions, however, is explained when looking at the point estimates for the predictive residual variances under the different methods displayed in Table 4.2. The F-VI methods in the right hand table, using zero-avoiding divergence in their posterior approximation, estimate drastically smaller values for the residual variance than VI does. In particular comparing the $D_A^{(\alpha)}$ for $\alpha = 0.5$ and $\alpha = 0$ shows that the more zero-avoiding the approximating divergence is, the greater the decrease in the residual variance estimate! This acts to counter the small expansion in posterior variance for the mean parameters (Figure 4.11) and results in the F-VI posterior producing more **concentrated** predictive distributions. The desire for conservative estimates of marginal uncertainty must result from a desire to produce predictive distributions that do not underestimate the predictive variance. A desire for conservative future predictions. However, the selection of a zero-avoiding approximating divergence has the opposite effect of decreasing the predictive variance!

We contend that these difficulties are in fact a direct consequence of the unprincipled nature of the F-VI objective functions, more specifically its failure to adhere to Axiom 1. F-VI methods attempt to approximate the exact posterior using a divergence which elicits a single desirable behaviour, in this case a zero-avoidance. However, treating the residual variance as a hyperparameter means that the value of $\hat{\sigma}^2$ will impact shape of the exact posterior, the target of the approximation. Specifically if $\hat{\sigma^2}$ is decreased then the exact posterior variances for the biases and weights will decrease. These smaller variance posterior must then become desirable under F-VI because they are then easier to approximate with a zero-avoiding divergence (there are fewer areas of positive mass to avoid putting 0's). Therefore F-VI decrease the residual variance to achieve a easier to approximate posteriors.

Figure 4.12 and Table 4.2 shows that under GVI there is still an interaction between the level of uncertainty quantification and the residual variance estimate variance $\hat{\sigma}^2$. However in juxtaposition to what was seen for F-VI, the relationship in GVI is consistent with the desired impact. Table 4.2 shows that when a divergence that elicits conservation estimates of marginal posterior variance is chosen ($D_{AR}^{(\alpha)}$ with $\alpha < 1$) then a conservative estimate of residual variance is also produced. On the other hand, if a divergence eliciting more confident estimates of marginal posterior variance is used ($D_{AR}^{(\alpha)}$ with $\alpha > 1$) then more confident estimates of residual variance is produced. This transparency and consistency of desired outcomes is a result of how GVI separates the loss minimisation (posterior location) from the uncertainty quantification. There is no prior over σ^2 and as a result σ^2 does not appear in the uncertainty quantifier. The parameter σ^2 is optimised based only on the loss function part of the GVI objective function taking into account the uncertainty quantifying properties of the posteriors for the biases and weights. If the biases and weights have large posterior variance, the residual variance is increased in order to minimise the expected loss. On the other hand, if the biases and weights have smaller posterior variance the opposite happens, again in order to minimise the posterior expected loss.

4.8.2 Deep Gaussian Process (GP) regression

Deep Gaussian Processes (GPs) [Damianou and Lawrence, 2013] provide another application area where VI and F-VI have been implemented. Deep GPs can be formulated as follows:

$$p(f_l|\Theta_l) = \mathcal{GP}(f_l; \mathbf{0}, \boldsymbol{K}_l), \quad l = 1, \ldots, L \tag{4.46}$$

$$p(\boldsymbol{h}_l|f_l, \boldsymbol{h}_{l-1}, \sigma_l^2) = \prod_{i=1}^{n} \mathcal{N}(h_{l,i}; f_l(h_{l-1,i}, \sigma_l^2)), \quad h_{0,i} = x_i \tag{4.47}$$

$$p(\boldsymbol{y}|f_L, \boldsymbol{h}_{L-1}, \sigma_L^2) = \prod_{i=1}^{n} \mathcal{N}(y_i; f_L(h_{L-1,i}, \sigma_L^2)) \tag{4.48}$$

A GP prior is placed on the function corresponding to each layer. Specifying, that the function evaluation for any finite set of d-dimensional inputs $(x_1, \ldots, x_n)$ is Multivariate Gaussian with mean function $\mathbf{0}$ and covariance function $\boldsymbol{K} = k(x_i, x_j)$. The inputs to the first layer are the predictor variables for the regression $h_{0,i} = x_i$, the outputs of each of the $l = 1, \ldots, L-1$ layers are hidden variable $h_{l,i}$, corresponding to noisy evaluations of the GP function in that layer evaluated at inputs corresponding to the outputs of the previous layer $h_{l-1,i}$. The outputs of layer L are the observed regression response.

Here the performance of GVI is compared with that of VI using the variational families of Salimbeni and Deisenroth [2017] that outperformed competing F-VI methods [Bui et al., 2016]. These introduce inducing points, or pseudo-observations $\boldsymbol{Z}^l$ for each layer l, and define the GP prior for the observed data conditional on these. This sparse, inducing points framework provides tractability within each layer [Matthews et al., 2016], and the exact model is used, conditional on the inducing points, for the variational posterior, maintaining correlations between layers[4].

The use of the exact posterior for the variational family suggests that uncertainty quantification may be adequately dealt with by the KLD regulariser in VI. As a result, here we investigate the impact changing the loss function can have. In order to robustify the Deep GP against model misspecification we consider the local γD loss function (Eq. (2.33) from Section 2.6) instead of the βD which we have largely considered in previous chapters. The γD loss can be conveniently stored in log-form and is therefore desirable for computational stability. This is not the case for the βD loss as it can be negative. A full derivation of the GVI algorithms applied to Deep GPs can be found in Knoblauch [2019]. Similarly to the BNN examples we compare methods based on RMSE and negative log-likelihood on a 10% testing set over 50 repeats, under the same setting as Salimbeni and Deisenroth [2017].

Figure 4.13 shows that GVI with the γD loss is able to outperform VI no matter the number of GP layers used. Unsurprisingly, choosing γ close to 1 is

[4] My co-authors were responsible for deriving and implementing this Deep GPs GVI algorithm

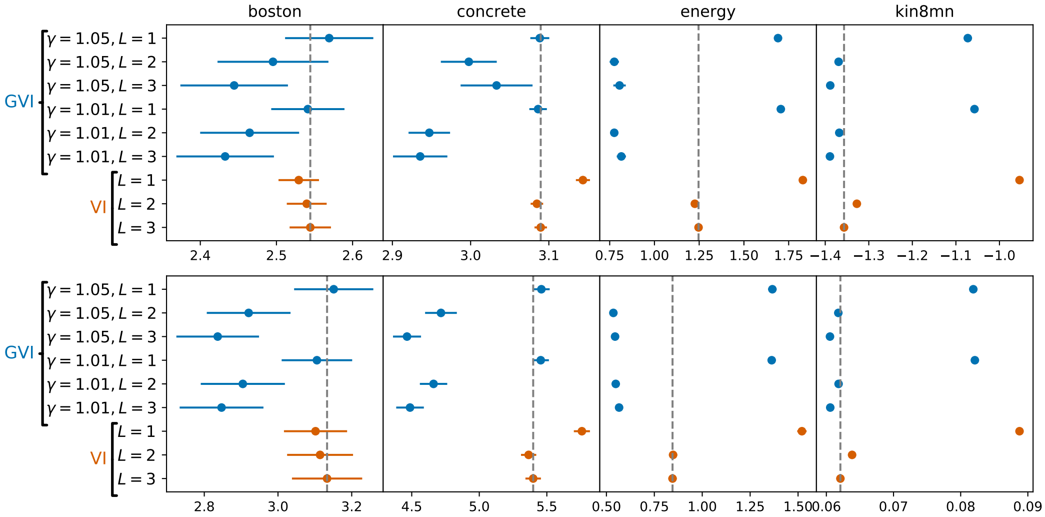

Figure 4.13: Comparing performance of **GVI** with γD-loss function and KLD prior regulariser and **VI** for DGPs with L layers. **Top row**: Negative test log likelihoods. **Bottom row**: Test RMSE. The lower the better.

desirable. We are evaluating the test set performance on the log-score (corresponding to $\gamma = 1$). However, the improved performance when taking γ marginally above 1 shows that providing robustness to extreme misspecifications allows the deep GPs to provide better test set generalisation.

4.9 Further work

This chapter has developed a generalised view of Bayesian inference as an optimisation problem depending on three inputs, a loss function defining the target parameter, a divergence eliciting uncertainty quantification and a space of probability densities to be optimised over. This was inspired by the optimisation perspective provided for Bayes' rule by Bissiri et al. [2016], which in turn allowed us to view traditional VI as constrained optimisation. GVI is shown to be underpinned by principled axioms which lead it to produce more transparent posterior inferences than several approximate inference alternatives to VI. We demonstrated the efficacy of GVI on two deep-learning problems.

Our takeaway message is as follows, in this modern data rich world we may no longer be able to exactly specify a belief model for the data or have the computational power to optimise for posterior beliefs over the space of all densities. When this is the case, one must resort to approximating beliefs and constraining the space of admissible posterior densities. In this paradigm we argue that embedding the fact that approximations have to be made in the objective function is the best way to proceed and that GVI provides the transparent and principled tools to do this.

The next chapter considers an applied problem where moderate dimensional observations arrive sequentially in real-time. The on-line nature of this analysis requires us to relax the assumption of infinite time to optimise over the space of all probability densities for our inference. Instead we tailor a specific case of GVI algorithms to this problem allowing us to produce fast and accurate Bayesian inference. The on-line setting also provides us with an opportunity to to consider setting the divergence hyperparameters discussed so far in this thesis in a prequential [Dawid, 1984] (predictive sequential) manner. We provide some initial work towards this in the next chapter.

Chapter 5

Robust Bayesian On-line Changepoint Detection

Chapters 2, 3 and 4 have proposed new methodology for conducting parametric model inference under the $\mathcal{M}$-OPEN world assumption, proved several philosophical and practical benefits of doing so, and considered novel inference methods to access these posteriors. This final chapter applies and extends these techniques to the challenging problem of Bayesian on-line changepoint detection (BOCPD). This example showcases the importance of applying robust and computationally convenient methodologies in the modern, high-dimensional, $\mathcal{M}$-OPEN world. This work has been published in Knoblauch, Jewson, and Damoulas [2018].

An outline of this chapter is as follows: Section 5.1 introduces the state of the art in BOCPD algorithms inferring a run-length posterior over the time since the last changepoint. Section 5.2 demonstrates the problems these algorithms face in the $\mathcal{M}$-OPEN world. Section 5.3 identifies that these posteriors are simply special cases of the general Bayesian posterior introduced in Chapter 1. Section 5.4 then proposes using the βD to robustify these posteriors producing robust BOCPD (RBOCPD). In so doing: Section 5.4.2 proves the extent to which the βD can robustify the run-length posterior; Section 5.4.4 derives a fast and accurate GVI (Chapter 4) algorithm for RBOCPD ensuring it is suitable for on-line implementation; Section 5.4.6 proposes methods to initialise and update the value of the divergence hyperparameter β. Lastly, Section 5.5 implements the RBOCPD algorithm. Firstly, on a canonical example form the literature and then to robustify the analysis of air pollution data from the City of London.

5.1 Bayesian Online Changepoint Detection (BOCPD)

Modelling non-stationary time series with changepoints (CPs) is popular [Khaleghi and Ryabko, 2012; Zhang et al., 2011; Lin et al., 2017] and important in a wide variety of research fields, including genetics [Caron et al., 2012; Grzegorczyk and Husmeier, 2009; Stimberg et al., 2011], finance [Kummerfeld and Danks, 2013], oceanography [Killick et al., 2010], brain imaging and cognition [Fox and Dunson, 2012; Huang and Paulus, 2016], cybersecurity [Polunchenko et al., 2012] and robotics [Alvarez et al., 2010; Konidaris et al., 2010]. In fact, CP detection has been identified as one of the major challenges for modern, big data applications (National Research Council, 2013)

For streaming data, a particularly important subclass are Bayesian Online Changepoint Detection (BOCPD) methods that can process data sequentially [Adams and MacKay, 2007; Fearnhead and Liu, 2007a; Turner et al., 2009; Xuan and Murphy, 2007; Wilson et al., 2010; Saatçi et al., 2010; Caron et al., 2012; Niekum et al., 2014; Turner et al., 2013; Ruggieri and Antonellis, 2016; Knoblauch and Damoulas, 2018] while providing full probabilistic uncertainty quantification. Consider a multivariate stream of data arriving at discrete time points $\{y_t\}_{t=1}^{\infty} = \{y_1, y_2, \ldots\}$, BOCPD wishes to produces inference about the location of changes in sample distribution of the observations in real time, that is as soon as possible after they occur. However in order to operate on-line any BOCPD algorithm needs to be computationally efficient.

We start by introducing the current state of the art in BOCPD algorithms and proceed to explain why we believe these methods are not suitable for the analysis of complex high-dimensional datasets. Instead we propose and apply one of the techniques developed in this thesis, GVI using the βD-loss function, in an attempt to resolve these issues.

5.1.1 The algorithm

We focus our attention on the BOCPD algorithm of Knoblauch and Damoulas [2018] which unifies the algorithms of Adams and MacKay [2007] and Fearnhead and Liu [2007b] in order to provide BOCPD with model selection. For simplicity of the illustration below we omit the model selection component of the algorithm. However, this is reintroduced and applied within the RBOCPD [Knoblauch et al., 2018] in the experiments below.

BOCPD exploits the product partition model (PPM) [Barry and Hartigan, 1992], assuming independence of parameters conditional on the CPs and independence of observations conditional on these parameters. These algorithms get their on-line efficiency by introducing a random variable called the run-length r_t for every time point t. The variable r_t is the time since the last CP. If $r_t = 0$ then y_t was generated by a different regime than all of the previous observations. While, if $r_t = l$ then y_t was generated by the same regime as the previous l observations. Now assume $r_t = l$ when a new observation y_{t+1} arrives. The run-length will either:

- increase by 1, $r_{t+1} = r_t + 1$, indicating that $\boldsymbol{y}_{t+1}$ is consistent with the previous l observations

- or shrink to 0, $r_{t+1} = 0$, indicating that $\boldsymbol{y}_{t+1}$ is different form the previous l observations and is thus the first observations from a new regime.

So for each run-length we only need to consider two possible alternatives after the observation at the next time point. In order to produce a full Bayesian quantification of uncertainty, BOCPD produces a posterior distribution over the run-length r_{t+1} given the observations $\boldsymbol{y}_{1:(t+1)}$. Thus, at each time point we have a discrete distribution over the time since the last CP. Adams and MacKay [2007] provide the following recursive Bayesian updating equations for the run-length posterior:

Growth probability

$$\begin{aligned} &\pi\left(r_t = l+1|\boldsymbol{y}_{1:t}\right) = \\ &\quad \frac{\pi_0\left(r_t = l+1|r_{t-1} = l\right) P\left(r_{t-1} = l, \boldsymbol{y}_{1:(t-1)}\right) P\left(y_t|r_t = l+1, \boldsymbol{y}_{1:(t-1)}\right)}{P\left(\boldsymbol{y}_{1:(t)}\right)} \end{aligned} \tag{5.1}$$

CP probability

$$\pi\left(r_t = 0|\boldsymbol{y}_{1:t}\right) = \sum_{i=0}^{t-1} \frac{\pi_0\left(r_t = 0|r_{t-1} = i\right) P\left(r_{t-1} = i, \boldsymbol{y}_{1:(t-1)}\right) P\left(y_t|r_t = 0, \boldsymbol{y}_{1:(t-1)}\right)}{P\left(\boldsymbol{y}_{1:(t)}\right)} \tag{5.2}$$

where the r_t's have a discrete finite support distribution and thus normalising their posteriors is straightforward $P\left(\boldsymbol{y}_{1:t}\right) = \sum_{i=0}^{t} P\left(r_t = i, \boldsymbol{y}_{1:t}\right)$. The distribution $P\left(r_{t-1} = i, \boldsymbol{y}_{1:(t-1)}\right)$ is the joint distribution of observations and parameters stored from previous time points. The distribution $\pi_0\left(r_t|r_{t-1}\right)$ is the run-length prior of the form

$$\pi_0\left(r_t|r_{t-1}\right) = \begin{cases} 1 - H\left(r_{t-1}+1\right) & \text{if } r_t = r_{t-1}+1 \\ H\left(r_{t-1}+1\right) & \text{if } r_t = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5.3}$$

$H(\cdot)$ is often called the hazard function of the run-length distribution and here represents the probability of a changepoint at time t given that the run length is r_t. The run-length distribution is often assumed to follow a geometric distribution which has the property of emitting a constant hazard function $H(\cdot) = \frac{1}{\lambda}$ for some $\lambda \in \mathbb{R}$ which is considered a fixed prior hyperparameter. The density $P\left(y_t|r_t = l+1, \boldsymbol{y}_{1:(t-1)}\right)$ is the predictive density of observation y_t given the previous $t-1$ observations and last CP l time points ago. Following the PPM assumption of independence before and after a CP,

$$\begin{aligned} P\left(y_t|r_t = l+1, \boldsymbol{y}_{1:(t-1)}\right) &= P\left(y_t|\boldsymbol{y}_{(t-l):(t-1)}\right) \\ &= \int f(y_t;\theta, \boldsymbol{y}_{(t-l):(t-1)})\pi(\theta|\boldsymbol{y}_{(t-l):(t-1)})d\theta \end{aligned} \tag{5.4}$$

where $f(y_t;\theta,\boldsymbol{y}_{(t-l):(t-1)})$ is a parametrised likelihood for the observations. BOCPD algorithms tend to restrict themselves to conjugate families in order to preserve computational efficiency when calculating the posterior $\pi(\theta|\boldsymbol{y}_{(t-l):(t-1)})$. In this case we can write the posterior predictive density as $P(y_t;\kappa(r_t,\boldsymbol{y}_{1:(t-1)}))$ where $\kappa(r_t,\boldsymbol{y}_{1:(t-1)})$ are the hyperparameters of the conjugate family associated with observations $\boldsymbol{y}_{(t-l):(t-1)}$ and the conjugate prior for θ. As a result all that needs to be stored to conduct run-length posterior inference at time t is the joint distribution $P\left(r_{t-1} = i, \boldsymbol{y}_{1:(t-1)}\right)$ for all i and the posterior hyperparametes $\kappa(r_t,\boldsymbol{y}_{1:(t-1)})$, which can be easily updated when the next observation arrives. Fixed time complexity can be achieved in BOCPD by truncating this CP posterior to only keep the the M most likely run-length for each t [Adams and MacKay, 2007].

As well as producing posterior distributions over the run lengths and thus the CP locations, Fearnhead and Liu [2007b] derive a recursive formula for the MAP segmentation in order to provide point estimates of the CP locations. Define $\text{MAP}_1 = 1$ then

$$\text{MAP}_t = \max_r \{p(\boldsymbol{y}_{1:t}, r_t = r)\text{MAP}_{t-r-1}\} \tag{5.5}$$

The MAP segmentation given r_t^* is then $S_t = S_{t-r_t^*-1}\bigcup\{t - r_t^*\}$. Where $S_0 = \emptyset$ and $t^{'} \in S_t$ means there was a CP at $t^{'} < t$.

Additionally, to the formulation of of Adams and MacKay [2007] explained above, Knoblauch and Damoulas [2018] are also able to produce inference on which of a set of likelihood models $m_t \in \{m_1, \ldots, m_M\}$ is the most suitable for each run length segment. Therefore, allowing the model to change from one segment to the next. This is particularly useful in Knoblauch and Damoulas [2018] as it allows them to select between regressors for their spatially structured BVAR models. Implementing this simply uses the one-step-ahead predictives to accumulate evidence for a given model in an analogous way to the run-length. As a result, implementing the robustification we propose in this chapter requires no further work if we additionally consider model selection. We exclude this for brevity from the exposition above, but note that it is used to analyse the 'air-pollution' data in Section 5.5.3.

5.1.2 The time-series model

As we mentioned above in order to maintain computational efficiency BOCPD algorithms restrict themselves to conjugate models and priors. One particular family suitable for time series analysis is the conjugate Bayesian linear model[1] with Gaussian likelihood and Normal-Inverse-Gamma (NIG) prior.

$$Y_i|X_i, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}\left(X_i\boldsymbol{\beta}, \sigma^2\right), \text{ for } i = 1, \ldots, n \tag{5.6}$$

$$\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}_p\left(\boldsymbol{\mu}_0, \sigma^2\boldsymbol{V}_0\right) \tag{5.7}$$

$$\sigma^2 \sim \mathcal{IG}\left(a_0, b_0\right), \tag{5.8}$$

where $(a_0, b_0, \boldsymbol{\mu}_n, \boldsymbol{V}_0)$ are prior hyperparameters. The posterior resulting from Bayes' rule is also in the NIG family

$$\pi(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{x}) = \mathcal{IG}\left(\sigma^2; a_n, b_n\right)\mathcal{N}_p\left(\boldsymbol{\beta}; \boldsymbol{\mu}_n, \sigma^2\boldsymbol{V}_n\right) \tag{5.9}$$

where parameters $(a_n, b_n, \boldsymbol{\mu}_n, \boldsymbol{V}_n)$ are given by closed form updating equations (e.g. see Banerjee [2008]). This family is particularly attractive as it contains the traditional auto-regressive (AR) and vector-auto-regressive (VAR) processes, where previous observations become predictors for the next. Further this model allows for the incorporation of multivariate responses by considering each dimension of a response as a new observation and correctly populating the design matrix with zeros to indicate which parameters correspond to which dimensions. The key assumption when extending this model to multiple dimensions is that the errors are uncorrelated across dimensions and have constant variance. Throughout this chapter we assume

[1] The conjugate Bayesian linear model is excellently reviewed by Banerjee [2008]

this model unless otherwise specified.

5.2 Problems in high-dimensions

In the $\mathcal{M}$-closed world the above BOCPD algorithm can be effective at providing efficient inference in an online setting. However in the high-dimensional $\mathcal{M}$-open world associated with big data applications the currently available CP detection technology can be shown to be insufficiently robust. For our motivating demonstration of this we take the London air pollution dataset analysed by Knoblauch and Damoulas [2018]. This consists of readings of Nitrogen Oxide (NOX) levels across 29 stations in London taken at 15 minute intervals over the course of a year. Before the analysis the quarter hourly measurements were averaged over 24 hrs to provide one observation per day, and weekly seasonality is accounted for by subtracting week-day averages for each station. Figure 5.1 plots these time series at 4 of the locations. Visualising medium to high dimensional data is tricky and it is difficult to see from these plots whether the congestion charge had any affect on the levels of pollution or not.

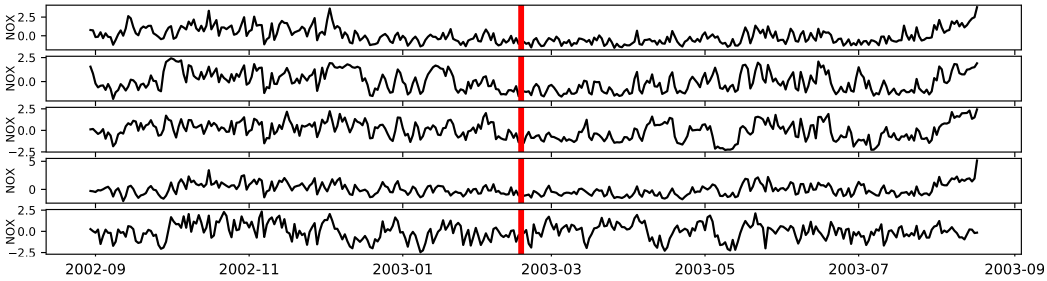

Figure 5.1: Air pollution data: NOX pollution levels across the year at 4 of the 29 sites in London. Quarter hourly readings are averaged to produce daily pollution levels and week-day variations are taken out by standardisation. Also marked is the congestion charge introduction, 17/02/2003 (solid vertical line).

Knoblauch and Damoulas [2018] analyse this data using Bayesian vector auto-regressions (BVAR) using a neighbourhood defined by the spatial structure of the observations to sparsify the predictor space. Observations can only be affected temporally by observations at its nearest neighbours. They demonstrate that their method finds a CP around the time that the congestion charge was introduced in London. This indicated that the congestion charge may very well have had the desired effect, something of interest to policy makers at the London Transport

Authority. However further inspection of this analysis reveals that BOCPD also finds a further 11 other CPs within a space of a year.

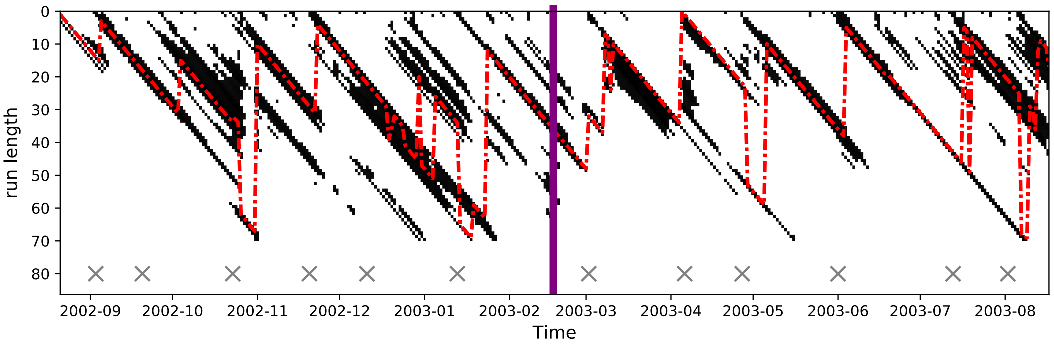

Figure 5.2: Air pollution data: Most likely run-lengths at each time point t for **standard** BOCPD run-length posterior. Also marked are the congestion charge introduction, 17/02/2003 (solid vertical line) and the retrospective MAP segmentations (crosses).

Figure 5.2 plots the posterior for the run-length, r_t, over the year in consideration. The grey-scale provides some idea of the distribution of the posterior mass across different run-lengths, darker colours correspond to more posterior mass. The red line shows the posterior MAP at each time points and the grey crosses show the CPs associated with the MAP segmentation (Eq. (5.5)). The shape of these posteriors, and in particular their MAP, demonstrates a key component of the run-length random variable: the run-length either grows by one at each time point, indicating this observation is consistent with the current segment, or the run length jumps to another value indicating that the observations comes from a different segment. The thick purple line indicated the introduction of the congestion charge and Figure 5.2 shows that Knoblauch and Damoulas [2018] found evidence of a changepoint just after its introduction. However, finding so many additional changepoints calls into question the validity of the conclusion that the introduction of the congestion charge had a lasting impact on the underlying dynamics of air pollution in London.

Although Knoblauch and Damoulas [2018] prove a theorem demonstrating that BVARs have the flexibility to capture a wide variety of stationary time series processes, we have strong reason to suspect that many of these CPs are false positive resulting from model misspecifications. These BVAR models were mainly chosen for their computational convenience in this on-line setting and the data is likely to be more complex. Section 2.7.5 has already discussed the increased risk of model misspecifications and outliers in such moderate to high-dimensional settings. For example foggy days, Christmas and school holidays are known to produce large irregularities in the air pollution levels recorded, and these have not been explicitly accounted for in the model. These changes to the system are not the types of changes we desire to detect. These constitute short, temporary changes in the underlying behaviour of the system followed by a return to the system's previous state. As a result we feel these can reasonably be defined as outlying segments. We are really interested in doing inference for long-term changes to the underlying process generating pollution in London. Further to this, BOCPD algorithms are particularly vulnerable to outliers. These algorithms declare CPs if the posterior predictive computed from $\boldsymbol{y}_{1:t}$ at time t has low density for the value of the observation $\boldsymbol{y}_{t+1}$ at time $t+1$. Naturally, this leads to a high false CP discovery rate in the presence of outliers and as they run on-line, pre-processing is not an option.

The literature on robust on-line CP detection so far is sparse and covers limited settings without Bayesian uncertainty quantification [e.g. Pollak, 2010; Cao and Xie, 2017; Fearnhead and Rigaill, 2017]. For example, the method in Fearnhead and Rigaill [2017] only produces point estimates and is limited to fitting a piecewise constant function to univariate data.

Here we aim to develop a procedure which is able to produce full and principled Bayesian uncertainty quantification, largely maintains the computational convenience of using conjugate Gaussian models, but also produces inference which is automatically robust to the misspecifications in the tails of this model.

5.3 General Bayesian Online Changepoint Detection

Eq. (5.1) and (5.2) provide the recursive formulation of the run-length posterior. We can expand the recursion in Eq. (5.1) by repeatedly substituting in the form of the joint distributions given by previous iterations to produce the more familiar representation of the run-length posterior as prior times likelihood. This more familiar representation lets us interpret the posterior as a special case of the generalised Bayesian posterior [Bissiri et al., 2016]. Setting $k = t - l - 1$ the run-legth posterior can be written as

$$\begin{aligned}\pi(r_t = l+1|\boldsymbol{y}_{1:t}) &\propto \overbrace{P(r_k = 0|\boldsymbol{y}_{1:k})\prod_{i=k}^{t}\pi_0(r_i|r_{i-1})}^{\text{prior}}\overbrace{\prod_{i=k}^{t}P(y_i|r_i = i-k, \boldsymbol{y}_{k:(i-1)})}^{\text{likelihood}}\\ &= P(r_k = 0|\boldsymbol{y}_{1:k})\prod_{i=k}^{t}\pi_0(r_i|r_{i-1})\exp\left(-\sum_{i=k}^{t} -\log P(y_i|r_i = i-k, \boldsymbol{y}_{k:(i-1)})\right).\end{aligned} \tag{5.10}$$

This demonstrates that analogously to Bayes' rule for traditional parameter updating (Eq. (1.20), Section 1.2.1), that the log-score is being applied to the one-step-ahead predictive densities in order to produce the run-length posterior.

Additionally this reveals the connection between BOCPD and Bayesian model selection using the marginal likelihood, as pointed out in Knoblauch and Damoulas [2018]. The run-length score function can equivalently be written as

$$-\sum_{i=k}^{t} \log P(y_i|r_i = i-k, \boldsymbol{y}_{k:(i-1)}) = -\log \prod_{i=k}^{t} P(y_i|r_i = i-k, \boldsymbol{y}_{k:(i-1)}) \tag{5.11}$$

$$= -\log P\left(\boldsymbol{y}_{1:t}, r_t = t-k\right) \tag{5.12}$$

where $P\left(\boldsymbol{y}_{k:t}, r_t = t-k\right)$ is the marginal likelihood for the observations $y_{k:t}$. Bernardo and Smith [2001] discussed this as a way of scoring a model and it also appears in the Bayes-factor ratio [Kass and Raftery, 1995]. Here is it corresponds to the likelihood (or evidence) in the data for a certain run-length. Viewed in an off-line sense the run length posterior of BOCPD is a special case of a Bayesian posterior over different models for the data.

5.4 Robust Bayesian Online Changepoint Detection

We have discussed at length in previous chapters why the log-score is not necessarily suitable for $\mathcal{M}$-OPEN inference. Chapter 2 demonstrated how inference minimising the log-score was very sensitive to outliers. This lead it to produce inference that provided limited performance guarantees when used to calculate expected utilities in a decision problem. Further, the experiments demonstrated that increasing the variance to deal with these outliers lead to less precise parameter and predictive inferences. Additionally, Chapter 3 identified that inference using the log-score was only stable to a very small equivalence class of likelihood models. Chapters 2 and 3 have posited several alternatives allowing parametric model inference to be conducted in a more robust fashion. In order to robustify the BOCPD run-length posterior we consider the βD loss function with parameter β_{rl}. This was mainly motivated by the βD's locality. Updating a density estimate on-line is computationally intensive and may perform poorly for small sample sizes. Using this produces robustified recursive Bayesian updating equations for the run-length posterior:

Robust growth probability

$$\pi^{(\beta_{rl})}\left(r_t = l+1|\boldsymbol{y}_{1:t}\right) = \frac{\pi_0\left(r_t = l+1|r_{t-1} = l\right)P^{(\beta_{rl})}\left(r_{t-1} = l, \boldsymbol{y}_{1:(t-1)}\right)\exp\left(-\ell^{(\beta_{rl})}\left(r_t = l+1, y_t\right)\right)}{P^{(\beta)}\left(\boldsymbol{y}_{1:(t)}\right)} \tag{5.13}$$

Robust CP probability

$$\pi^{(\beta_{rl})}\left(r_t = 0|\boldsymbol{y}_{1:t}\right) = \sum_{i=0}^{t-1}\frac{\pi_0\left(r_t = 0|r_{t-1} = i\right)P^{(\beta_{rl})}\left(r_{t-1} = i, \boldsymbol{y}_{1:(t-1)}\right)\exp\left(-\ell^{(\beta_{rl})}\left(r_t = 0, y_t\right)\right)}{P^{(\beta_{rl})}\left(\boldsymbol{y}_{1:(t)}\right)} \tag{5.14}$$

where

$$\ell^{(\beta_{rl})}\left(r_t = i, y_t\right) = -\frac{1}{\beta_{rl}-1}P(y_t|r_t = i, \boldsymbol{y}_{1:(t-1)})^{\beta_{rl}-1} + \frac{1}{\beta_{rl}}\int P(z|r_i, \boldsymbol{y}_{1:(t-1)})^{\beta_{rl}}dz. \tag{5.15}$$

The robust run-length posterior is still discrete and has finite support so can easily be normalised using $P^{(\beta)}\left(\boldsymbol{y}_{1:t}\right) = \sum_{i=0}^{t} P^{(\beta)}\left(r_t = i, \boldsymbol{y}_{1:t}\right)$.

5.4.1 Robustify the model

Section 2.2 provided several philosophical arguments for why the model is important for any statistical analyses and should not be abandoned or changed for fears of a lack of robustness, we expand upon these here. In the BOCPD case the BVAR model and particularly the spatial neighbourhoods introduced in Knoblauch and Damoulas [2018] impart important structure on the analysis. Additionally, Gaussian errors were assumed in order to make the algorithm computationally viable. One standard method from the robust statistics literature [e.g. Berger et al., 1994; O'Hagan, 1979] to guard against outliers is to robustify the model with heavy tails, for example a Student's-t distribution. We identify several philosophical and practical drawbacks of doing this relative to changing the loss function: I) switching from a Gaussian to a Student's-t likelihood breaks conjugacy, as a result either MCMC or variational methods will be required to approximate this posterior, we demonstrate in Section 5.4.4 that using the βD-loss function with a Gaussian likelihood can become more computationally convenient than switching to a Student's-t likelihood. II) Using a Student's-t likelihood corresponds to modelling the outliers rather than 'ignoring' them for the analysis. Predictive variances will still be increased as a result of outliers. III) Using the Student's-t likelihood only works for problems requiring symmetric continuous error distributions, while the βD-loss can be applied to achieve robustness much more generally, to count data for example. Lastly, IV) we now show that using a Student's-t likelihood is not sufficient to guarantee robustness in the run-length inference.

The algorithm of Fearnhead and Rigaill [2017] is robust because hyperparameters enforce that a single outlier is insufficient for declaring a CP. Analogously, we investigate conditions under which a single (outlying) observation $\boldsymbol{y}_{t+1}$ is able to force a CP in BOCPD. An intuitive way of achieving this is by studying the odds of $r_{t+1} \in \{0, r+1\}$ conditional on $r_t = r$. Under BOCPD we have that

$$\frac{\pi\left(r_{t+1} = r+1 | \boldsymbol{y}_{1:(t+1)}, r_t = r\right)}{\pi\left(r_{t+1} = 0 | \boldsymbol{y}_{1:(t+1)}, r_t = r\right)} = \frac{\cancel{P(r_t = r, \boldsymbol{y}_{1:t})} \cdot (1 - H(r+1))P(y_{t+1}|r_{t+1} = r+1, \boldsymbol{y}_{1:t})}{\cancel{P(r_t = r, \boldsymbol{y}_{1:t})} \cdot H(r+1)P(y_{t+1}|r_{t+1} = 0, \boldsymbol{y}_{1:t})}. \tag{5.16}$$

Taking a closer look at Eq. (5.16), if $\boldsymbol{y}_{t+1}$ is an outlier with low density under $P(y_{t+1}|r_{t+1} = r+1, \boldsymbol{y}_{1:t})$, the odds will move in favour of a CP provided that the prior is sufficiently uninformative to make $P(y_{t+1}|r_{t+1} = 0, \boldsymbol{y}_{1:t}) > P(y_{t+1}|r_{t+1} = r+1, \boldsymbol{y}_{1:t})$. In fact, even very small differences have a substantial impact on the odds. This is why using the Student's t error for the BLR model with standard Bayes will not provide robust run-length posteriors: While an outlying observation $\boldsymbol{y}_{t+1}$ will have greater density $P(y_{t+1}|r_{t+1} = r+1, \boldsymbol{y}_{1:t})$ under a Student's t error model than under a normal error model, $P(y_{t+1}|r_{t+1} = 0, \boldsymbol{y}_{1:t})$ (the density under the prior) will also be larger under the Student's t error model. As a result, changing the tails of the model only has a very limited effect on the ratio in Eq. (5.16). In fact, the perhaps unintuitive consequence is that Student's t error models will yield CP inference that very closely resembles that of the corresponding normal model. A range of numerical examples in the Appendix of Knoblauch et al. [2018] illustrate this surprising fact.

5.4.2 Quantifiable robustness

In contrast to the observations above, CP inference robustified via the βD does not suffer from this phenomenon. Under RBOCPD the corresponding odds are

$$\frac{\pi^{(\beta)}\left(r_{t+1} = r+1 | \boldsymbol{y}_{1:(t+1)}, r_t = r\right)}{\pi^{(\beta)}\left(r_{t+1} = 0 | \boldsymbol{y}_{1:(t+1)}, r_t = r\right)} = \frac{(1 - H(r+1))\exp\left(-\ell^{(\beta)}\left(r_t + 1 = r+1, y_{t+1}\right)\right)}{H(r+1)\exp\left(-\ell^{(\beta)}\left(r_t + 1 = 0, y_{t+1}\right)\right)}. \tag{5.17}$$

Theorem 11 then provides very mild conditions for the βD robustified run-length posterior under a conjugate BLR model ensuring that there is a combination of hyperparameters such that these odds (and thus the MAP of the run length posterior) never favour a CP after *any* single outlying observation $\boldsymbol{y}_{t+1}$. These results demonstrate that the RBOCPD requires a greater build up of evidence in favour of a CP before declaring one than the original BOCPD. This result is analogous to the robustness guarantee of Fearnhead and Rigaill [2017].

Theorem 11. If the likelihood model for BOCPD is the conjugate Bayesian Linear Regression (BLR) with $\boldsymbol{\mu} \in \mathbb{R}^p$ and priors a_0, b_0, μ_0, Σ_0; and if the posterior predictive's variance determinant is larger than $|V|_{\text{min}} > 0$, then one can choose any $(\beta_{\text{rl}}, H(r_t, r_{t+1})) \in S\left(p, \beta_{\text{rl}}, a_0, b_0, \mu_0, \Sigma_0, |V|_{\text{min}}\right)$ to guarantee that

$$\frac{\pi^{(\beta)}\left(r_{t+1} = r+1 | \boldsymbol{y}_{1:(t+1)}, r_t = r\right)}{\pi^{(\beta)}\left(r_{t+1} = 0 | \boldsymbol{y}_{1:(t+1)}, r_t = r\right)} \geq 1, \tag{5.18}$$

for all y_{t+1}. The set $S\left(p, \beta_{\text{rl}}, a_0, b_0, \mu_0, \Sigma_0, |V|_{\text{min}}\right)$ is defined by an inequality given in (5.36).

Proof. This proof looks at the run length posterior parameterised by β_{rl}, however to ease notation we refer to $\beta_{\text{rl}}= \beta$ throughout. First we condition on the event that $r_t = r$, then after one time step either $r_{t+1} = r+1$ or $r_{t+1} = 0$ and the odds ratio under the βD-bayes is

$$\begin{aligned} &\frac{\exp\left(-\ell^{(\beta)}\left(r_t + 1 = r+1, y_{t+1}\right)\right)}{\exp\left(-\ell^{(\beta)}\left(r_t + 1 = 0, y_{t+1}\right)\right)} \\ &= \exp\left(\frac{1}{\beta-1}\left(p(\boldsymbol{y}_{t+1}|\boldsymbol{y}_{(t-r):t})^{\beta-1} - p(\boldsymbol{y}_{t+1}|\boldsymbol{y}_0)^{\beta-1}\right) - \frac{1}{\beta}\int p(\boldsymbol{z}|\boldsymbol{y}_{(t-r):t})^{\beta} - p(\boldsymbol{z}|\boldsymbol{y}_0)^{\beta} d\boldsymbol{z}\right). \end{aligned} \tag{5.19}$$

This proof first seeks a lower bound for this ratio. A lower bound on $\frac{1}{\beta-1}p(\boldsymbol{y}_{t+1}|\boldsymbol{y}_{1:t})^{\beta-1}$ is 0, while the maximal value of $\frac{1}{\beta-1}p(\boldsymbol{y}_{t+1}|x_0)^{\beta-1}$ will occur at the prior mode. Under the conjugate BLR the one-step-ahead predictive distributions are multivariate t-distribution. For the multivariate t-distribution prior predictive with NIG hyperparameters a_0, b_0, μ_0, $\boldsymbol{\Sigma}_0$ of dimensions p the prior mode has density

$$p(\boldsymbol{\mu}_0|\nu_0, \boldsymbol{\mu}_0, \boldsymbol{V}_0, p) = \frac{\Gamma((\nu_0+p)/2)}{\Gamma(\nu_0/2)\nu_0^{p/2}\pi^{p/2}\left|\boldsymbol{V}_0\right|^{1/2}}\left[1 + \frac{1}{\nu_0}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_0)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_0)\right]^{-(\nu_0+p)/2} \tag{5.20}$$

$$= \frac{\Gamma((\nu_0+p)/2)}{\Gamma(\nu_0/2)\nu_0^{p/2}\pi^{p/2}\left|\boldsymbol{V}_0\right|^{1/2}} \tag{5.21}$$

$$= \frac{\Gamma(a_0 + p/2)}{\Gamma(a_0)\left(2b_0\pi\right)^{p/2}\left|\boldsymbol{I} + \boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^T\right|^{1/2}}. \tag{5.22}$$

As a result the only term in the lower bound of Eq. (5.19) that does not solely depend on the prior hyperparameter is $\frac{1}{\beta}\int p(\boldsymbol{z}|\boldsymbol{y}_{1:t})^\beta d\boldsymbol{z}$. This term appears in the negative and thus to lower bound Eq. (5.19), an upper bound for $\frac{1}{\beta}\int p(\boldsymbol{z}|\boldsymbol{y}_{1:t})^\beta d\boldsymbol{z}$ must be found. The multivariate t-distribution can be integrated as

$$\begin{aligned}
&\frac{1}{\beta}\int \mathrm{MVSt}_\nu(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{V})^\beta d\boldsymbol{z} \\
&= \frac{\Gamma((\nu+p)/2)^\beta\Gamma(((\beta-1)\nu+(\beta-1)p+\nu)/2)}{\Gamma(\nu/2)^\beta\Gamma(((\beta-1)\nu+(\beta-1)p+\nu+p)/2)}\frac{1}{(\beta(\nu\pi)^{((\beta-1)p)/2}|\boldsymbol{V}|^{(\beta-1)/2}} \\
&= \frac{\Gamma((\nu+p)/2)^{\beta-1}\Gamma((\nu+p)/2)\Gamma(((\beta-1)\nu+(\beta-1)p+\nu)/2)}{\Gamma(\nu/2)^{\beta-1}\Gamma(\nu/2)\Gamma(((\beta-1)\nu+(\beta-1)p+\nu+p)/2)}\frac{1}{\beta(\pi\nu)^{((\beta-1)p)/2}|\boldsymbol{V}|^{(\beta-1)/2}}
\end{aligned} \tag{5.23}$$

Given that $\frac{\Gamma(x+\frac{p}{2})}{\Gamma(x)}$ is increasing in x and as $\beta \geq 1$ and $\nu \geq 0$ then $((\beta-1)\nu+(\beta-1)p+\nu)/2) \geq \nu/2$ which implies

$$\frac{\Gamma((\nu+p)/2)\Gamma(((\beta-1)\nu+(\beta-1)p+\nu)/2)}{\Gamma(\nu/2)\Gamma(((\beta-1)\nu+(\beta-1)p+\nu+p)/2)} \leq 1. \tag{5.24}$$

This in turn provides the following inequality

$$\begin{aligned}
&\frac{1}{\beta}\int \mathrm{MVSt}_\nu(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{V})^\beta d\boldsymbol{z} \\
&= \frac{\Gamma((\nu+p)/2)^{\beta-1}\Gamma((\nu+p)/2)\Gamma(((\beta-1)\nu+(\beta-1)p+\nu)/2)}{\Gamma(\nu/2)^{\beta-1}\Gamma(\nu/2)\Gamma(((\beta-1)\nu+(\beta-1)p+\nu+p)/2)}\frac{1}{\beta(\pi\nu)^{((\beta-1)p)/2}|\boldsymbol{V}|^{(\beta-1)/2}} \\
&\leq \frac{\Gamma((\nu+p)/2)^{\beta-1}}{\Gamma(\nu/2)^{\beta-1}}\frac{1}{\beta(\pi\nu)^{((\beta-1)p)/2}|\boldsymbol{V}|^{(\beta-1)/2}}.
\end{aligned} \tag{5.25}$$

Now employing the well-known result using Stirling's formula to bound the gamma function

$$(2\pi)^{1/2}x^{x-1/2}\exp(-x) \leq \Gamma(x) \leq (2\pi)^{1/2}x^{x-1/2}\exp(1/(12x)-x) \tag{5.26}$$

we can therefore rewrite the ratio of gamma functions leaving

$$\begin{aligned}
&\frac{1}{\beta}\int \mathrm{MVSt}-t_\nu(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{V})^\beta d\boldsymbol{z} \leq \frac{\Gamma((\nu+p)/2)^{\beta-1}}{\Gamma(\nu/2)^{\beta-1}}\frac{1}{\beta(\pi\nu)^{((\beta-1)p)/2}|\boldsymbol{V}|^{(\beta-1)/2}} \\
&\leq \frac{\left(\sqrt{2\pi}((\nu+p)/2)^{(\nu+p-1)/2)}\exp(-(\nu+p)/2+1/6(\nu+p))\right)^{\beta-1}}{\left(\sqrt{2\pi}(\nu/2)^{(\nu-1)/2}\exp(-\nu/2)\right)^{\beta-1}\beta(\pi\nu)^{((\beta-1)p)/2}|\boldsymbol{V}|^{(\beta-1)/2}} \\
&= ((1+\frac{p}{\nu})^{(\beta-1)(\nu+p-1)/2)}\exp((\beta-1)(1/(6(\nu+p))-p/2))\frac{1}{\beta(\pi)^{((\beta-1)p)/2}|\boldsymbol{V}|^{(\beta-1)/2}}.
\end{aligned} \tag{5.27}$$

Clearly $\exp\left((\beta-1)(1/(6(\nu+p))-p/2)\right)$ is decreasing in ν for all p and to demonstrate when $((1+\frac{p}{\nu})^{(\beta-1)(\nu+p-1)/2)}$ is decreasing in ν we examine its derivative

$$
\begin{aligned}
w &= \left(1+\frac{p}{\nu}\right)^{(\beta-1)(\nu+p-1)/2} && (5.28)\\
&= \exp\left(((\beta-1)(\nu+p-1)/2)\log\left(\left(1+\frac{p}{\nu}\right)\right)\right) && (5.29)\\
\frac{dw}{d\nu} &= \frac{\beta-1}{2}\left(\log\left(1+\frac{p}{\nu}\right)-(\nu+p-1)\frac{\frac{p}{\nu^2}}{1+\frac{p}{\nu}}\right)\left(1+\frac{p}{\nu}\right)^{(\beta-1)(\nu+p-1)/2)}. && (5.30)
\end{aligned}
$$

The sign of $\frac{dw}{d\nu}$ is dictated by $\left(\log\left(1+\frac{p}{\nu}\right)-(\nu+p-1)\frac{\frac{p}{\nu^2}}{1+\frac{p}{\nu}}\right)$, which can be demonstrated to be positive always if $p=1$ and negative always if $p>1$.

Case 1: when $p>1$, $\frac{1}{\beta}\int p(\boldsymbol{z}|\boldsymbol{y}_{1:t})^{\beta}d\boldsymbol{z}$ is decreasing in ν and thus we can upper bound it by substituting the smallest value of ν. Here we bound ν above 1 in order to enforce that the mean of the predictive t-distribution exists. Under the KLD posterior it is clear that a_0 rises as more data is seen and while we do not have closed forms associated with the variational approximation to the βD posterior (see Section 5.4.4) we expect this to be the case here. As more data is seen the finite sampling uncertainty, represented by ν in the NIG case, should be decreasing. Therefore provided a_0 is set such that $2a_0>1$, then this lower bound should never be violated.

Case 2: when $p=1$, Stirling's formula has failed to provide a decreasing upper bound for $\frac{1}{\beta}\int p(\boldsymbol{z}|\boldsymbol{y}_{1:t})^{\beta}dz$. However in the univariate case

$$
\begin{aligned}
\frac{1}{\beta}\int \mathrm{St}_{\nu}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{V})^{\beta}d\boldsymbol{z} &\leq \frac{\Gamma((\nu+1)/2)^{\beta-1}}{\Gamma(\nu/2)^{\beta-1}}\frac{1}{\beta(\nu\,|\boldsymbol{V}|)^{(\beta-1)/2}\pi^{(\beta-1)/2}} && (5.31)\\
&\leq \frac{1}{\beta\,|\boldsymbol{V}|^{(\beta-1)/2}\,\pi^{(\beta-1)/2}} && (5.32)
\end{aligned}
$$

Where $p=1$ is substituted into the bound from equation (5.25) and the inequality comes from that fact that $\frac{\Gamma((x+1)/2)}{\Gamma(x/2)}\leq\sqrt{x}$. This bound conveniently does not depend on the degrees of freedom ν at all.

We can therefore lower bound

$$
\frac{\exp\left(-\ell^{(\beta)}\left(r_{t+1}=r+1,y_{t+1}\right)\right)}{\exp\left(-\ell^{(\beta)}\left(r_{t+1}=0,y_{t+1}\right)\right)} \qquad (5.33)
$$

as

$$\begin{cases} \exp\left\{-\frac{1}{\beta-1}\left(\frac{\Gamma(a_0+1/2)}{\Gamma(a_0)(2b_0\pi)^{1/2}|\boldsymbol{I}+\boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^T|^{1/2}}\right)^{\beta-1} - \frac{1}{\beta|\boldsymbol{V}|^{(\beta-1)/2}\pi^{(\beta-1)/2}}+\right. \\ \left.\frac{\Gamma(a_0+1/2)^\beta\Gamma((\beta-1)a_0+(\beta-1)/2+a_0)}{\Gamma(a_0)^\beta\Gamma((\beta-1)a_0+(\beta-1)/2+a_0+1/2)}\frac{1}{\beta(2\pi b_0)^{(\beta-1)/2}|\boldsymbol{I}+\boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^T|^{(\beta-1)/2}}\right\} & \text{if } p=1 \\ \exp\left\{-\frac{1}{\beta-1}\left(\frac{\Gamma(a_0+p/2)}{\Gamma(a_0)(2b_0\pi)^{p/2}|\boldsymbol{I}+\boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^T|^{1/2}}\right)^{\beta-1} +\right. \\ \frac{\Gamma(a_0+p/2)^\beta\Gamma((\beta-1)a_0+(\beta-1)p/2+a_0)}{\Gamma(a_0)^\beta\Gamma((\beta-1)a_0+(\beta-1)p/2+a_0+p/2)}\frac{1}{\beta(2\pi b_0)^{((\beta-1)p)/2}|\boldsymbol{I}+\boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^T|^{(\beta-1)/2}}- \\ \left.((1+p)^{(\beta-1)p/2)}\exp((\beta-1)(1/(6(1+p))-p/2))\frac{1}{\beta(\pi)^{((\beta-1)p)/2}|\boldsymbol{V}|^{(\beta-1)/2}}\right\} & \text{if } p>1 \end{cases} \tag{5.34}$$

Now fixing $p, a_0, b_0, \mu_0, \Sigma_0$ and $|V|_{\min}$, we are interested in the values of β and $H(r_t, r_{t+1})$ which ensure that

$$\frac{\pi^{(\beta)}\left(r_{t+1}=r+1|\boldsymbol{y}_{1:(t+1)}, r_t=r\right)}{\pi^{(\beta)}\left(r_{t+1}=0|\boldsymbol{y}_{1:(t+1)}, r_t=r\right)} \geq 1 \tag{5.35}$$

We demonstrate this for $p > 1$ but it is straightforward to see that it extends to when $p = 1$. Rearranging the inequality in equation (5.34) gives us that (5.35) holds providing

$$\frac{1}{|\boldsymbol{V}|^{(\beta-1)/2}} \leq \left(\frac{\Gamma(a_0+p/2)^{\beta-1}}{\Gamma(a_0)^{\beta-1}\,(2b_0\pi)^{(\beta-1)p/2}\,|\boldsymbol{I}+\boldsymbol{X}\boldsymbol{\Sigma}_0\boldsymbol{X}^T|^{(\beta-1)/2}}\right. \tag{5.36}$$
$$\frac{\left(\left(\frac{\Gamma(a_0+p/2)\Gamma((\beta-1)a_0+(\beta-1)p/2+a_0)}{\Gamma(a_0)\Gamma((\beta-1)a_0+(\beta-1)p/2+a_0+p/2)}\frac{1}{\beta}-\frac{1}{\beta-1}\right)+\log\left(\frac{1-H(r_t,r_{t+1}))}{H(r_t,r_{t+1})}\right)\right)\beta(\pi)^{((\beta-1)p)/2}}{((1+\frac{p}{2a_0})^{\alpha(2a_0+p-1)/2)}\exp((\beta-1)(1/(6(2a_0+p))-p/2))}$$

We define the set defined by inequality (5.36) as $S\left(p, \beta, a_0, b_0, \mu_0, \boldsymbol{\Sigma}_0, |\boldsymbol{V}|_{\min}\right) = \{(\beta, H(r_t, r_{t+1})) : (\beta, H(r_t, r_{t+1}))$ satisfy (5.36) for $p, \beta, a_0, b_0, \mu_0, \boldsymbol{\Sigma}_0, |\boldsymbol{V}|_{\min}\}$. As a result we can see that for fixed of $a_0, b_0, \mu_0, \Sigma_0$ and $|\boldsymbol{V}| \geq |\boldsymbol{V}|_{\min}$ it is always possible to choose values of β and $H(r_t, r_{t+1})$ such that this holds. To see this consider fixing β, the the upper bound is simply increasing in $\log\left(\frac{1-H(r_t,r_{t+1})}{H(r_t,r_{t+1})}\right)$ which takes values in $\mathbb{R}$ and thus can be set large enough so that the inequality holds. □

Theorem 11 says that one can bound the odds for a CP independently of $\boldsymbol{y}_{t+1}$. The requirement for a lower bound $|V|_{\min}$ results from the integral term in Eq. (5.15), which dominates βD-inference if $|V|$ is extremely small. In practice, this is not restrictive: E.g. for $p = 5$, $h(r) = \frac{1}{\lambda}$, $a_0 = 3, b_0 = 5, \Sigma_0 = \text{diag}(100, 5)$ used in Figure 5.7, Theorem 11 holds for $(\beta_{\text{rl}}, \lambda) = (1.15, 100)$ used for inference if

$|V|_{\min} \geq 8.12 \times 10^{-6}$.

These results, combined with the fact that a Student's-t likelihood was not sufficient to guarantee robustness for standard BOCPD, further the conclusions from Chapter 2 that a lack of robustness is not the fault of the model but of the loss function used for inference.

5.4.3 Robust parameter posterior

Additionally to using the βD to robustify the run-length posterior, we also use the βD loss function to robustify the parameter posterior $\pi(\theta|\boldsymbol{y}_{(t-l):(t-1)})$ used to calculate the one-step-ahead predictive densities in Eq. (5.4), as we did in Chapter 2. This introduces β_p that need not be the same as β_{rl}. We define this robust parameter posterior $\pi^{(\beta_p)}(\theta|\boldsymbol{y}_{(t-l):(t-1)})$ and refer to it as the βD-Bayes posterior. We call the combination of βD run-length inference and βD parameter inference as Robust Bayesian On-line Changepoint Detection (RBOCPD).

5.4.4 Quasi-conjugacy

The standard BOCPD algorithms restricted inference to conjugate models in order to guarantee the computational efficiency of the algorithm [Adams and MacKay, 2007; Knoblauch and Damoulas, 2018]. However, using the βD to robustify the parameter posterior removes the conjugacy property of the Bayesian updating. In Chapter 2 and previously in the literature [Ghosh and Basu, 2016; Jewson et al., 2018], MCMC has been used to sample from the βD-Bayes posteriors. However, it is not easy to scale these for the on-line inference. While we acknowledge the existence of sequential Monte Carlo algorithms [Del Moral et al., 2006] that can be tailored to online inference, we outline the following drawbacks. Any sampling regime would first need to sample from the parameter posterior $\pi^{(\beta_p)}(\theta|\boldsymbol{y})$, use this sample to estimate the posterior predictive $P^{(\beta_p)}(y^*|\boldsymbol{y})$ and then require further Monte Carlo methods (e.g. importance sampling) to estimate the integral term $\frac{1}{\beta+1}\int P^{(\beta_p)}(z|\boldsymbol{y})^{\beta+1}dz$ associated with the βD loss function applied to the posterior predictive. This will not only be slow but may also introduce two sources of added variance to the on-line algorithm at each time-point.

To resolve this we use a special subset of GVI algorithms. For exponential family model and corresponding conjugate prior,

$$f(y;\theta) = h(y)\exp\left(\eta(\theta)^T T(y) - A(\eta(\theta))\right) \tag{5.37}$$

$$\pi(\theta;\nu_0,\mathcal{X}_0) = g(\mathcal{X}_0,\nu_0)\exp\left(\nu_0\eta(\theta)^T\mathcal{X}_0 - \nu_0 A(\eta(\theta))\right) \tag{5.38}$$

we posit the conjugate prior family for the variational family when using the βD-loss.

$$\mathcal{Q} = \left\{q(\theta;\nu,\mathcal{X}) = g(\mathcal{X},\nu)\exp\left(\nu\eta(\theta)^T\mathcal{X} - \nu A(\eta(\theta))\right) : g(\nu,\mathcal{X}) < \infty\right\} \tag{5.39}$$

We note here we have used the traditional exponential family representation for the likelihood rather than for the prior as we did when introducing GVI in Chapter 4. This is to keep with convention. This variational family is well motivated by the fact that as $\beta \to 1$, βD $\to$ KLD and therefore the exact posterior, $\pi^{(\beta_p)}(\theta|\boldsymbol{y}_{(t-l):(t-1)})$, will be contained within the variational family as $\beta_p \to 1$. In order to gain robustness we take $\beta_p > 1$. However we expect some smoothness in the behaviour of the βD-posteriors for β_p not too far from 1.

We investigate this assumption and what it means to be 'too far from 1' in Section 5.4.5. For this reason we consider this a well motivated constraining family in contrast to many applications using the ad-hoc 'mean-field' approximation. The assumption of smoothness in the βD-Bayes posteriors also motivates $D =$ KLD as the prior regulariser. If the constraining family provides a good approximation to the exact solution then the KLD and its associated coherence properties are desirable. Putting all this together we can write the GVI problem as follows

$$\widehat{q}^{\beta_p}(\theta;\boldsymbol{y}_{(t-l):(t-1)}) := \arg\min_{q\in Q}\left\{\mathbb{E}_{q(\theta)}\left[\sum_{i=k}^{t-1}\ell^{(\beta_p)}(\theta,y_i)\right] + \text{KLD}(q(\theta)||\pi(\theta))\right\}. \tag{5.40}$$

We term this procedure of forcing the posterior to be in the conjugate prior when using different model based losses to the log-score, quasi-conjugacy. Quasi-conjugate updating with the βD has several other convenient properties. It ensures that the predictive distribution is available in closed form and we can further show broad conditions under which the GVI objective function (or ELBO) is also available in closed form. This result is presented in Theorem 12

Theorem 12. The GVI objective under the βD-loss function with parameter β_p and the KLD to the prior as in Eq. (5.40) of an exponential family likelihood model $f(\boldsymbol{y};\theta)$, conjugate prior $\pi(\theta;\nu_0,\mathcal{X}_0)$ and variational family $q(\theta;\nu_n,\mathcal{X}_n)$ within the same conjugate family,

$$f(\boldsymbol{y};\theta) = h(y)\exp\left(\eta(\theta)^T T(y) - A(\eta(\theta))\right) \tag{5.41}$$

$$\pi(\theta;\nu_0,\mathcal{X}_0) = g(\mathcal{X}_0,\nu_0)\exp\left(\nu_0\eta(\theta)^T\mathcal{X}_0 - \nu_0 A(\eta(\theta))\right) \tag{5.42}$$

$$q(\theta;\nu_n,\mathcal{X}_n) = g(\mathcal{X}_n,\nu_n)\exp\left(\nu_n\eta(\theta)^T\mathcal{X}_n - \nu_n A(\eta(\theta))\right), \tag{5.43}$$

is analytically available if and only if the following three quantities have closed form

$$\mathbb{E}_{q(\theta;\nu_n,\mathcal{X}_n)}\left[\eta(\theta)\right],\ \mathbb{E}_{q(\theta;\nu_n,\mathcal{X}_n)}\left[\log g(\eta(\theta))\right],\ \int A(z)^{\beta_{\mathrm{p}}}\left[h\left(\frac{\beta_{\mathrm{p}}T(z)+\nu_m\mathcal{X}_m}{\beta_{\mathrm{p}}+\nu_m},\beta_{\mathrm{p}}+\nu_m\right)\right]^{-1}dz.$$

for all values of $(\nu_n, \mathcal{X}_n)$ and y such that

$$\left(\frac{(\beta_p-1)T(y)+\nu_n\mathcal{X}_n}{\beta_p-1+\nu_n},\beta_p-1+\nu_n\right)\in\mathcal{N} \tag{5.44}$$

$$\left(\frac{\beta_pT(y)+\nu_n\mathcal{X}_n}{\beta_p+\nu_n},\beta_p+\nu_n\right)\in\mathcal{N} \tag{5.45}$$

where $\mathcal{N}$ is the natural parameter space of the conjugate prior family is defined as

$$\mathcal{N}=\left\{(\nu,\mathcal{X}):\int\exp\left(\nu\eta(\theta)^T\mathcal{X}-\nu A(\eta(\theta))\right)d\theta<\infty\right\} \tag{5.46}$$

Proof. For ease of notation, we use $\beta_{\mathrm{p}}=\beta$. We note that

$$A(\eta(\theta))=\log\left(\int h(y)\exp\left(\eta(\theta)^TT(y)\right)dy\right) \tag{5.47}$$

$$g(\mathcal{X}_i,\nu_i)=\left(\int\exp\left(\nu_i\eta(\theta)^T\mathcal{X}_i-\nu_iA(\eta(\theta))\,d\theta\right)^{-1} \tag{5.48}$$

The resulting GVI objective function is as follows

$$\begin{aligned}&L\left(q|\boldsymbol{y},\ell^{(\beta)},\mathrm{KLD}\right)=\\&\sum_{i=1}^{n}-\mathbb{E}_q\left[\ell^{(\beta)}(y_i;\theta)\right]-\mathrm{KLD}\left(q\left(\theta|\nu_n,\mathcal{X}_n\right),\pi\left(\theta|\nu_0,\mathcal{X}_0\right)\right).\end{aligned} \tag{5.49}$$

The βD-loss function can be expanded to give

$$\begin{aligned}-\ell^{(\beta)}(y;\theta)=&\frac{1}{\beta-1}\left(h(y)\exp\left(\eta(\theta)^TT(y)-A(\eta(\theta))\right)\right)^{\beta-1}-\\&\frac{1}{\beta}\int\left(h(y)\exp\left(\eta(\theta)^TT(y)-A(\eta(\theta))\right)\right)^{\beta}dz\\=&\frac{1}{\beta-1}\exp\left((\beta-1)\eta(\theta)^TT(y)-(\beta-1)A(\eta(\theta))\right)h(y)^{\beta-1}-\\&\frac{1}{\beta}\int\exp\left(\beta\eta(\theta)^TT(z)-\beta A(\eta(\theta))\right)h(z)^{\beta}dz.\end{aligned} \tag{5.50}$$

Therefore $L\left(q|\boldsymbol{y}, \ell^{(\beta)}, \text{KLD}\right)$ has three integrals that need evaluating

$$
\begin{aligned}
B_1 =& \sum_{i=1}^{n} \int \frac{h(y_i)^{\beta-1}}{\beta-1} \exp\left((\beta-1)\left\{\eta(\theta)^T T(y_i) - A(\eta(\theta))\right\}\right) q(\theta|\nu_n, \mathcal{X}_n) d\theta \quad (5.51)\\
B_2 =& \frac{n}{\beta} \int \left\{ \int h(z)^{\beta} \exp\left(\beta\eta(\theta)^T T(z) - \beta A(\eta(\theta))\right) dz \right\} q(\theta|\nu_n, \mathcal{X}_n) d\theta \quad (5.52)\\
B_3 =& \text{KLD}\left(q\left(\theta|\nu_n, \mathcal{X}_n\right) || \pi\left(\theta|\nu_0, \mathcal{X}_0\right)\right). \quad (5.53)
\end{aligned}
$$

Now firstly for the term B_1 in equation (5.51)

$$
\begin{aligned}
B_1 =& \sum_{i=1}^{n} \int \frac{h(y_i)^{\beta-1}}{\beta-1} \exp\left((\beta-1)\eta(\theta)^T T(y_i) - (\beta-1)A(\eta(\theta))\right)\\
& g(\mathcal{X}_n, \nu_n) \exp\left(\nu_n \eta(\theta)^T \mathcal{X}_n - \nu_n A(\eta(\theta))\right) d\theta\\
=& \sum_{i=1}^{n} \frac{h(y_i)^{\beta-1}}{\beta-1} g(\mathcal{X}_n, \nu_n)\\
& \int \exp\left(\eta(\theta)^T \left((\beta-1)T(y_i) + \nu_n \mathcal{X}_n\right) - (\beta-1+\nu_n)A(\eta(\theta))\right) d\theta\\
=& \sum_{i=1}^{n} \frac{h(y_i)^{\beta-1}}{\beta-1} g(\mathcal{X}_n, \nu_n) \frac{1}{g(\frac{(\beta-1)T(x_i)+\nu_n\mathcal{X}_n}{\beta-1+\nu_n}, \beta-1+\nu_n)}. \quad (5.54)
\end{aligned}
$$

Where we know that $g(\frac{(\beta-1)T(y_i)+\nu_n\mathcal{X}_n}{\beta-1+\nu_n}, \beta-1+\nu_n)$ is integrable and closed form as it represents the normalising constant of the same exponential family as the prior and the variational posterior, provided $\left(\frac{(\beta-1)T(y_i)+\nu_n\mathcal{X}_n}{\beta-1+\nu_n}, \beta-1+\nu_n\right) \in \mathcal{N}$.

Next we look at B_2 in equation (5.52). The whole integral is the product of two densities which must be positive and in order for the $L\left(q|\boldsymbol{y}, \ell^{(\beta)}, \text{KLD}\right)$ to be defined it must also be integrable. Therefore we can use Fubini's theorem to switch the order of integration

$$
\begin{aligned}
B_2 =& \frac{n}{\beta} \int \left\{ \int \exp\left(\beta\eta(\theta)^T T(z) - \beta A(\eta(\theta))\right) q(\theta|\nu_n, \mathcal{X}_n) d\theta \right\} h(z)^{\beta} dz\\
=& \frac{n g(\mathcal{X}_n, \nu_n)}{\beta} \int \left\{ \int \exp\left(\eta(\theta)^T (\beta T(z) + \nu_n \mathcal{X}_n) - (\beta+\nu_n)A(\eta(\theta))\right) d\theta \right\} h(z)^{\beta} dz\\
=& \frac{n g(\mathcal{X}_n, \nu_n)}{\beta} \int \frac{h(z)^{\beta}}{g(\frac{\beta T(z)+\nu_n\mathcal{X}_n}{\beta+\nu_n}, \beta+\nu_n)} dz. \quad (5.55)
\end{aligned}
$$

once again provided that $\left(\frac{\beta T(z)+\nu_n\mathcal{X}_n}{\beta+\nu_n}, \beta+\nu_n\right) \in \mathcal{N}$ then $g(\frac{\beta T(z)+\nu_n\mathcal{X}_n}{\beta+\nu_n}, \beta+\nu_n)$ is the normalising constant of the same exponential family as the prior and the variational posterior and is thus closed form.

Lastly we look at B_3 in equation (5.53)

$$
\begin{aligned}
B_3 = &\int q(\theta|\nu_n, \mathcal{X}_n) \log \frac{g(\mathcal{X}_n, \nu_n) \exp\left(\nu_n \eta(\theta)^T \mathcal{X}_n - \nu_n A(\eta(\theta))\right)}{g(\mathcal{X}_0, \nu_0) \exp\left(\nu_0 \eta(\theta)^T \mathcal{X}_0 - \nu_0 A(\eta(\theta))\right)} \\
=&\log \frac{g(\mathcal{X}_n, \nu_n)}{g(\mathcal{X}_0, \nu_0)} \int q(\theta|\nu_n, \mathcal{X}_n) \left\{ \left(\eta(\theta)^T \left(\nu_n \mathcal{X}_n - \nu_0 \mathcal{X}_0\right)\right) - \left(\nu_n - \nu_0\right) A(\eta(\theta)) \right\} \\
=&\log \frac{g(\mathcal{X}_n, \nu_n)}{g(\mathcal{X}_0, \nu_0)} \left\{ \left(\nu_n - \nu_0\right) \lambda_n + \left((\mu_n)^T \left(\nu_n \mathcal{X}_n - \nu_0 \mathcal{X}_0\right)\right) \right\}, \qquad (5.56)
\end{aligned}
$$

where $\mu_n = \mathbb{E}_{q(\theta;\nu_n,\mathcal{X}_n)}\left[\eta(\theta)\right]$ and $\lambda_n = \mathbb{E}_{q(\theta;\nu_n,\mathcal{X}_n)}\left[A(\eta(\theta))\right]$ are defined to be available in closed form by the conditions of the theorem. As a result we get that

$$
\begin{aligned}
L\left(q|\boldsymbol{y}, \ell^{(\beta)}, \text{KLD}\right) =& B_1 - B_2 - B_3 \\
=& \sum_{i=1}^{n} \frac{1}{\beta - 1} h(y_i)^{\beta-1} g(\mathcal{X}_n, \nu_n) \frac{1}{g\left(\frac{(\beta-1)T(y_i) + \nu_n \mathcal{X}_n}{\beta - 1 + \nu_n}, \beta - 1 + \nu_n\right)} \\
&- \frac{n}{\beta} g(\mathcal{X}_n, \nu_n) \int \frac{h(z)^{\beta}}{g\left(\frac{\beta T(z) + \nu_n \mathcal{X}_n}{\beta + \nu_n}, \beta + \nu_n\right)} dz. \qquad (5.57) \\
&- \log \frac{g(\mathcal{X}_n, \nu_n)}{g(\mathcal{X}_0, \nu_0)} \left\{ \left((\mu_n)^T \left(\nu_n \mathcal{X}_n - \nu_0 \mathcal{X}_0\right)\right) - \left(\nu_n - \nu_0\right) \lambda_n \right\}.
\end{aligned}
$$

□

The conditions of Theorem 12 are met by many exponential models, e.g. the Normal-Inverse-Gamma, the Exponential-Gamma, and the Gamma-Gamma. This closed form objective function allows us to use simple, off-the-shelf gradient descent technology and not have to worry about the sampling or black box approaches referred to in Section 4.7. Our quasi-conjugate algorithm has similarities with the algorithm of Ghahramani and Beal [2001] who demonstrate that the optimal variational posterior, factorised between parameters and missing variables, is a member of the conjugate prior family in conjugate-exponential hidden variables. While our algorithm emits a closed form objective function, Ghahramani and Beal [2001] produce closed form, iterative updating equations.

5.4.5 The accuracy of the quasi-conjugate Bayesian update

The quasi-conjugate GVI inference algorithm explained above was motivated by assuming a smoothness in the βD-Bayes posterior when moving β away from 1. We investigate this empirically for the BLR conjugate family that will be used in the RBOCPD algorithm.

Firstly Figure 5.3 plots the bivariate posterior marginals for (μ_0, μ_1, σ^2) of a uni-variate, BVAR of lag $L = 1$. This is equivalent to a BLR with two predictors, the intercept and the previous observation with coefficients μ_0 and μ_1 respectively. The red contours correspond to the exact posterior, $\pi^{(\beta_p)}(\theta|\boldsymbol{y}_{(t-l):(t-1)})$, produced by smoothing 95,000 Hamiltonian Monte Carlo samples produced using *stan* [Carpenter et al., 2016], and the blue is the GVI approximation to this, $\widehat{q}^{\beta_\text{p}}(\theta; \boldsymbol{y}_{(t-l):(t-1)})$. This shows that for $\beta_p = 1.25$, quasi-conjugate GVI provides a near perfect approximation to the exact βD-posterior.

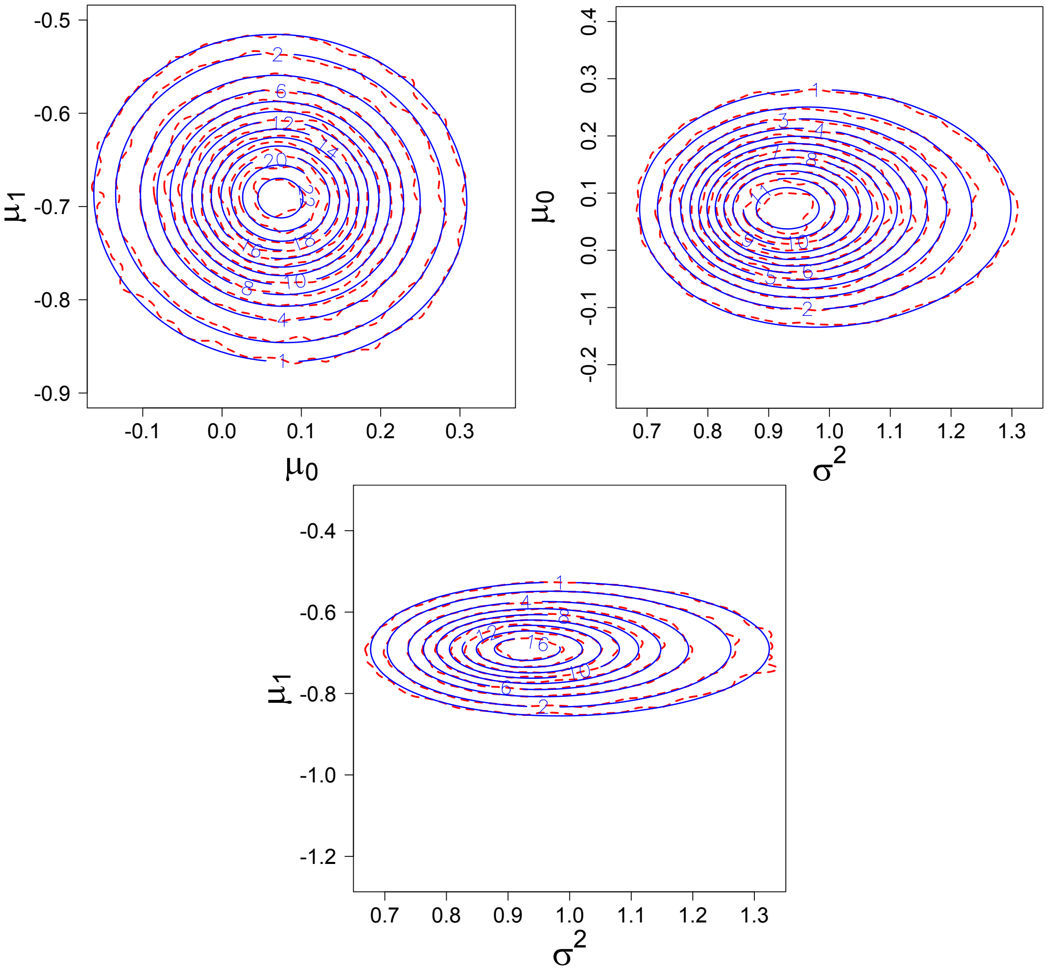

Figure 5.3: Exemplary contour plots of bivariate marginals for the approximation $\widehat{q}^{\beta_\text{p}}(\theta; \boldsymbol{y}_{(t-l):(t-1)})$ of Eq. (5.40) (dashed) and the target $\pi^{(\beta_p)}(\theta|\boldsymbol{y}_{(t-l):(t-1)})$ (solid) estimated and smoothed from $95,000$ Hamiltonian Monte Carlo samples for the βD-Bayes posterior for a BLR with $d = 1$, two regressors, and $\beta_\text{p} = 1.25$.

Next we investigate this further and in higher dimensions for both the response variable and for the parameters. Yao et al. [2018a] take inspiration from Pareto-Smoothed-Importance-Sampling [Vehtari et al., 2015] to produce a metric estimating the difference $\hat{k}$ between an exact posterior, say $\pi^{(\beta_p)}(\theta|\boldsymbol{y}_{(t-l):(t-1)})$, and a variational approximation, say $\widehat{q}^{\beta_\text{P}}(\theta; \boldsymbol{y}_{(t-l):(t-1)})$, relative to some posterior expectation. Pareto-Smoothed-Importance-Sampling [Vehtari et al., 2015] attempts to improve the quality of importance sampling estimates by fitting a generalised Pareto distribution to the tail of importance weights of the form

$$\omega_{\pi,q}(\theta) := \frac{\pi^{(\beta_p)}(\theta|\boldsymbol{y}_{(t-l):(t-1)})}{\widehat{q}^{\beta_\text{P}}(\theta; \boldsymbol{y}_{(t-l):(t-1)})}. \tag{5.58}$$

While this was originally used to address the bias-variance tradeoff in importance sampling, Yao et al. [2018a] are also able to interpret the shape parameter of the fitted generalised Pareto distribution, $\hat{k}$, as an estimate of the following goodness of fit criteria:

$$k := \left\{k^* > 0 : \mathbb{E}_{\theta \sim \widehat{q}^{\beta_\text{P}}}\left[\omega_{\pi,q}(\theta)^{\frac{1}{k^*}}\right] < \infty\right\} \tag{5.59}$$

Clearly if $\pi^{(\beta_p)}(\theta|\boldsymbol{y}_{(t-l):(t-1)}) = \widehat{q}^{\beta_\text{P}}(\theta; \boldsymbol{y}_{(t-l):(t-1)})$ this holds for all $k^* > 0$ and the closer the two distributions are togther the smaller the k for which this will be the case. In fact, $1/k$ is the maximal parameter α such that the $D_{AR}^{(\alpha)}(\pi^{(\beta_p)}||\widehat{q}^{\beta_\text{P}})$ is finite, where $D_{AR}^{(\alpha)}$ is defined in Eq. (1.29) in Section 1.3.1. Empirical studies lead Yao et al. [2018a] to conclude that if $\hat{k}$ is less than 0.5 then $\widehat{q}^{\beta_\text{P}}$ is 'close enough' to $\pi^{(\beta_p)}$. The posterior expectation of interest in BOCPD algorithms is the posterior predictive distribution. Therefore, we implement the method of Yao et al. [2018a] to help investigate the quality of the VI approximation to the one-step-ahead posterior predictive central to the performance of BOCPD.

In order to do so, we simulated data from BVAR models of dimension $d = 5, 10, 15, 25$, with lag length $L = 1$ and no interaction between dimensions. Each time series is thus modelled as $X_{t,j} \sim \mathcal{N}(\theta_{1,j}X_{t-1,j} + \theta_{0,j}, \sigma^2)$, $t = 1, \dots, T$ and $j = 1, \dots, d$, θ are the model parameters and the residual variance, σ^2, is shared across dimensions. Stacking dimensions creates a BLR model as in Eq. (5.6). Jointly modelling this results in a multivariate Gaussian likelihood function with $2d + 1$ parameters. Therefore increasing d increases the dimension of both the observations space and the parameter space. For each d we simulated fixed coefficients θ from Unif$[-0.5, 0.5]$ and then 50 replicates of datasets of size $T = 200$. Figure 5.4 plots the estimated values of $\hat{k}$, averaged across the 50 replicates of the data for different values of β_p whilst varying the dimension.

For β_p close to 1 we see that the variational family is very accurate for all dimensions. This corroborates our assumption of smoothness as β_p increases from 1. When $\beta_p = 1$ the exact posterior is contained within the variational family, no mat-

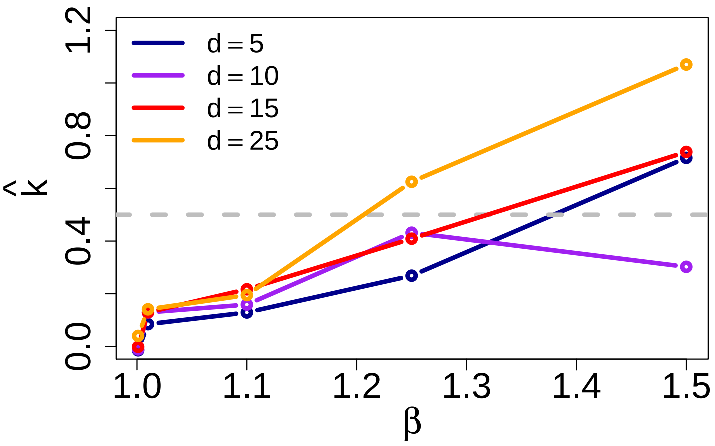

Figure 5.4: Plot of the estimate $\hat{k}$ [Yao et al., 2018a], quantifying the accuracy of the variational approximation $\widehat{q}^{\beta_\text{p}}$ to the exact posterior distribution $\pi^{(\beta_p)}$, for different values of β_p applied to a BVAR model with response dimension d and predictor dimension $2d$ for $d = 5$, 10, 15 and 25. The grey dotted line depicts the threshold of $\hat{k} < 0.5$ demonstrating the values of β_p for which $\widehat{q}^{\beta_\text{p}}$ can be considered ‘close enough’ to $\pi^{(\beta_p)}$ [Yao et al., 2018a].

ter what the dimension, and taking β_p slightly above 1 the variational family appears to provide an excellent approximation to the exact posterior. As β_p increases the estimated value of $\hat{k}$ also increases and thus the approximation appears to get worse. However we note that for quite a wide range of β_p the estimated $\hat{k}$ is less than 0.5, the threshold specified by Yao et al. [2018a] suggesting the approximation was sufficiently accurate. Figure 5.4 also shows that as the dimension of the parameter space increase, β_p needs to be taken increasingly small to ensure a good approximation of the exact posterior by the variational family. An explanation for this is as follows: As d increases, the magnitude of $f(\boldsymbol{y}_t; \theta, \boldsymbol{y}_{1:(t-1)})$ decreases rapidly. Hence, β_p needs to decrease as d increases to prevent the βD-Bayes inference from being dominated by the integral in Eq. (2.13) and disregarding $\boldsymbol{y}_t$, as was discussed in Section 2.7.5. When this happens the inference is no longer learning through the conjugate likelihood f, but is learning through the integral term and thus the accuracy of the conjugate variational approximation to the exact βD posterior degrades.

We additionally note that these values of β ensuring that the variational approximation is sufficiently accurate are exactly the values of β we want to choose for inference anyway. As was discussed in Section 2.7.5 when the integral term in Eq. (2.13) dominates the Bayesian inferece is ignoring the data. While we want robust learning, we still want at least the majority of the data to guide the inference. This is also reflected in our experiments in Section 5.5. Here we initialize $\beta_{\text{p}} = 0.05$ and $\beta_{\text{p}} = 0.005$ for $d = 1$ and $d = 29$, respectively. However, as Figures 5.3 and 5.4 illustrate, the approximation is still excellent for values of β_{p} that are much larger than that.

We discussed at the end of Section 2.7.4 that setting a calibration weight w using the method of Lyddon et al. [2018] could help mitigate the drop in efficiency associated with setting β too large. We conjecture that setting w may also be able to improve the quality of the GVI approximation to the exact posterior for large β. However, the method to set w of Lyddon et al. [2018] requires a full pass through the data and is therefore not suitable to be applied in the on-line setting.

5.4.6 Setting β_p and β_{rl}

While the log-score was parameterless, robustifying the run-length and parameter posteriors introduces two further hyperparameters into the algorithm. $\beta = 1$ recovers the log-score and increasing β away from 1 will buy increasing robustness to outliers. However taking β too far above 1 can down-weight the data too much and lead to nonsensical inferences (see Section 2.7.4). As a result both β_p and β_{rl} need to be selected carefully.

Initalisation

Firstly for initialisation we fix $\beta_I = \beta_p = \beta_{rl}$. It is common in Bayesian analyses to select hyperparameters by reverse engineering them from the prior predictive distributions [see e.g. Gelman et al., 2017; Gabry et al., 2019]. The same can be done for β_I using the influence curves introduced in Section 2.6.7 (Figure 2.2). Under the KLD these influence curves are increasing as an observation x moves away from the posterior mean. Under the βD the influence curves initially increase, mimicking the KLD, but then reach some point of maximum influence x^*_β and decrease after that. As $\beta \to 1$, $x^*_\beta \to \infty$. The location x^*_β marks the point at which the influence observations have on the posterior starts to decrease as they move away from the posterior mean, rather than increase. That is to say that after x^*_β, observations are increasingly considered as outliers relative to the current inferences. As a result, if we can specify a threshold, τ_x, for the prior predictive after which we desire to treat observations as outliers, we can initialise $\beta_I = \left\{\beta : x^*_\beta = \tau_x\right\}$. When the dimension of the observation space is small, τ_x could be taken to be some number of standard deviations from the posterior mean. As the dimension of the observation space increases it may be more useful to exploit the result of Hall et al. [2005], that we expect data to arrive at a Mahalanobis distance of square-root the dimension of the observation space.

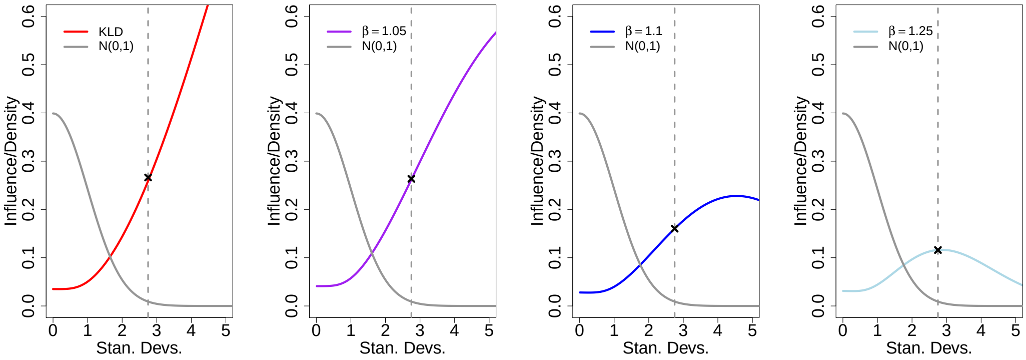

Figure 5.5: Visualisation for the initialisation of β. A threshold of $\sigma = 2.75$ standard deviations from the mean is chosen as the cut off to start declaring outliers. Under the KLD-Bayes ($\beta = 1$) the influence curve is always increasing. Increasing to $\beta = 1.05$ makes the βD-Bayes influence curve concave, but the maximum point is much larger than $\sigma = 2.75$. Increasing again $\beta = 1.1$ brings the maximal point closer to $\sigma = 2.75$ and increasing further to $\beta = 1.25$ makes the maximal influence point at $\sigma = 2.75$. Thus $\beta = 1.25$ would be chosen as the initial value for β given a $\sigma = 2.75$ threshold for outliers.

Figure 5.5 provides a pictorial representation of this initialisation process. These show the influence plots from Section 2.6.7 for increasing values of β moving from left to right. An outlier threshold of $\sigma = 2.75$ standard deviations from the mean is plotted as a vertical dotted grey line along with the data generating Gaussian density demonstrating that observations greater than this threshold can reasonably be considered outliers. Under the KLD-Bayes ($\beta = 1$) the influence curve on the far left is always increasing. Increasing β to $\beta = 1.05$ makes the βD-Bayes influence curve concave, but the maximum point is much larger than $\sigma = 2.75$. Increasing β again to $\beta = 1.1$ brings the maximal point closer to $\sigma = 2.75$ and increasing further to $\beta = 1.25$ makes the maximal influence point at $\sigma = 2.75$. For an outlier threshold of 2.75 standard deviations from the mean the initial value of β is chosen at $\beta_I = 1.25$

On-line updating

The on-line nature of RBOCPD allows us to additionally update our initial value of β as the algorithm progresses which can help guard against a poorly specified prior (or threshold). In order to do this we must first define a *meta-loss-function*, measuring how close the model predictions at each time point $\hat{Y}_t = \mathbb{E}\left[Y_t|\boldsymbol{y}_{1:(t-1)}, \beta_p, \beta_{rl}\right]$

are to the observed values y_t, $L(\hat{Y}_t, y_t)$. We note that $\hat{Y}_t$ does not depend on model parameter θ or run-length r_t as these have been integrated out, but it does depend on the $\boldsymbol{\beta} = (\beta_p, \beta_{rl})$ used to produce those posteriors. One sensible example for this and the one used in Knoblauch et al. [2018] is the absolute loss $L(\hat{Y}_t, y_t) = \left|\hat{Y}_t - y_t\right|$. Given such a loss function, after the observation of $\boldsymbol{y}_{1:t}$ we can find $\hat{\boldsymbol{\beta}}_t = \arg\min_\beta \sum_{i=1}^t L(\hat{Y}_i, y_i)$ and use this to update the parameter and run-length posteriors when we observe y_{t+1}. It is important that $\hat{\boldsymbol{\beta}}_t$ is used to update beliefs using observation y_{t+1} and not y_t to ensure the data is not used twice. The resulting procedure is very similar to that of Caron et al. [2012] who optimise prior hyperparameters.

5.5 Demonstrations

In order to get the important arguments of my contribution to this project I have presented a robustfied version of Adams and MacKay [2007] algorithm which corresponds to a simplification of the algorithm used in Knoblauch et al. [2018]. The experiments in this section use the full version of this algorithm. In addition to the algorithm presented above Knoblauch et al. [2018] implement:

- Model/Variable selection: Use the recursions in Knoblauch and Damoulas [2018] to not only produce a robustified run length posterior, but also provide a posterior for the model associated with each segment, allowing for structural and temporal selection within the BVAR framework

- Optimisation of prior hyperparameters: Similarly to Knoblauch and Damoulas [2018] the hyperperparameters of the prior for each model are optimised using an on-line gradient descent approach of Caron et al. [2012]. This is similar to the on-line optimisation of the loss hyperparameters (β_p, β_{rl}) discussed in Section 5.4.6

- Stochastic optimisation of the variational parameters: In order to achieve a trade-off between scalability and accuracy a combination of full optimisation and stochastic optimisation is used to estimate the variational hyperparameters. For small run length $r_t < W$ full optimisation is done every m observations, while when $r_t > W$ and we have enough observations to expect each new observation to have stable impact on the variational parameters, stochastic optimisation using a subset of the observations is used, for more information on this procedure see Knoblauch et al. [2018]

I derived the robust recursions, the variational inference regime and proved Theorems 11 and 12, adapting the space and time complexity of this algorithm was the work of my collaborator, as was implementing this full algorithm.

5.5.1 The *well-log* dataset

We first start with a canonical real-world example from the CP detection literature. The well-log data set was first studied in Ruanaidh et al. [1996] and has become a benchmark data set for univariate CP detection. However, except in Fearnhead and Rigaill [2017] its outliers have been removed before CP detection algorithms are run [e.g. Adams and MacKay, 2007; Levy-leduc and Harchaoui, 2008; Ruggieri and Antonellis, 2016]. Instead, we run both the standard BOCPD and our robustified RBOCPD on the full dataset without removing the outliers to more accurately reflect a true on-line analysis. Similarly to previous analyses we consider a model class $\mathcal{M}$ containing one BLR model of form $y_t = \mu + \varepsilon_t$.

The top panel of Figure 5.6 plots the observed data. The data appears to be well modelled by independent segments coming from different Gaussian distributions. However, many of these segments appear to have small numbers of observations which differ massively from those either side. Fearnhead and Rigaill [2017] states that these segments are "where the probe misfunctions". We argue that these observations are too few in number to be considered separate segments and are instead outliers. The bottom 2 panels of Figure 5.6 demonstrate the run-length posteriors provided by RBOCPD and BOCPD respectively. The bold blue and red lines represent the MAP of the run-length at each time point and the grey-scale represents the spread of the posterior mass. The bottommost plot shows that the standard BOCPD is very sensitive to the outlying segments. The MAP drops to zero 145 times, so declaring CPs online based on the run-length distribution's maximum [see e.g. Saatçi et al., 2010] yields a False Discovery Rate (FDR) $> 90\%$ compared with those found when outliers were removed [e.g. Adams and MacKay, 2007; Levy-leduc and Harchaoui, 2008; Ruggieri and Antonellis, 2016]. This problem persists even with non-parametric, Gaussian Process, models [p. 186, Turner, 2012]. Even using Maximum A Posteriori (MAP) segmentation Fearnhead and Liu [2007a], standard BOCPD mislabels 8 outliers as CPs, resulting in a FDR $> 40\%$. These are plotted as red dotted lines over the data on the top of Figure 5.6. This shows that even when the data is viewed completely the standard BOCPD struggles to differentiate outliers from true changes.

In contrast, the MAP segmentation of the RBOCPD using the βD, plotted in thick blue, does not mislabel any outliers and still finds what appear to be 'true'

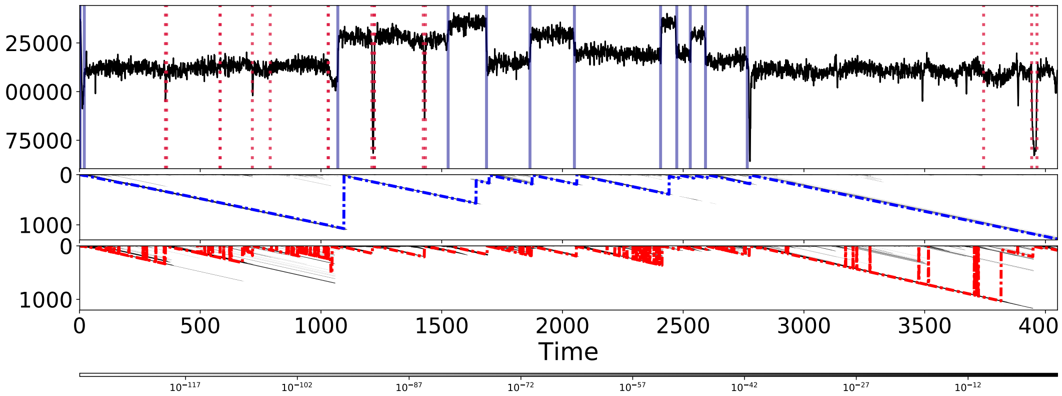

Figure 5.6: Maximum A Posteriori (MAP) segmentation and run-length distributions of the well-log data. Robust segmentation depicted using solid lines, CPs additionally declared under standard BOCPD with dashed lines. The corresponding run-length distributions for robust (middle) and standard (bottom) BOCPD are shown in greyscale. The most likely run-lengths are dashed.

changes in the underlying mean. Moreover, and in accordance with Theorem 11, the middle panel shows that its run-length distribution's maximum never drops to zero in response to outliers. This demonstrates that in a univariate setting the RBOCPD is able to be robust to outlying segments of data but still be sensitive to underling changes in the DGP.

Further, a natural by-product of the robust segmentation is a reduction in squared (absolute) prediction error by 10% (6%) compared to the standard BOCPD. The RBOCPD has more computational overhead than standard BOCPD, but still needs less than 0.5 seconds per observation using a 3.1 GHz Intel i7 and 16GB RAM.

Not only does robust BOCPD's segmentation in Figure 5.6 match that in Fearnhead and Rigaill [2017], but it also offers three additional on-line outputs: Firstly, it produces probabilistic (rather than point) forecasts and parameter inference. Secondly, it self-regulates its robustness via $\boldsymbol{\beta}$. Thirdly, it can compare multiple models and produce model posteriors (see Section 5.5.3). Further, unlike Fearnhead and Rigaill [2017], it is not restricted to fitting univariate data with piecewise constant functions.

5.5.2 Moderate dimensional simulated example

Next we consider a 5-dimensional simulated example. This serves to motivate the need for robust methods for moderate to high dimensional applications. Here five auto-regressions of lag $L = 1$ were simulated and jointly modelled as a BVAR. The fifth dimension was injected with additive Student's-t noise with degrees of freedom $\nu = 4$. There are two true CPs at time $t = 200$ and $t = 400$. Figure 5.7 plots the five time series stacked on top of each other. The CPs declared in the MAP segmentation of the standard BOCPD are plotted as dotted red lines while the equivalent MAP segmentation for the RBOCPD is plotted as solid blue lines. In addition to finding the two actual CPs, the standard BOCPD algorithm declares 11 further CPs in the MAP segmentation as a result of the Student's-t contamination. In contrast the RBOCPD with $(\beta_p, \beta_{rl}) = (1.1, 1.25)$ finds only the CPs corresponding to actual changes in the DGP. Although this example is artificial, it illustrates two important points mentioned in Section 2.7.5 about high-dimensional M-OPEN statistics. Firstly, when inspecting Figure 5.7 it is certainly not obvious that the spurious CPs that have been detected are a result of outliers, visualising data in more than one or two dimensions can be tricky and as a result spotting outliers is not straightforward. Secondly, it only took one dimension of the process to be misspecified to force the declaration of many spurious CPs. As the dimension of the problem goes up, the DM is required to make more and more beliefs statements, which in turn increases the likelihood that at least some of these are misspecified in some regard. These effects will only worsen as the dimension of the problem increases.

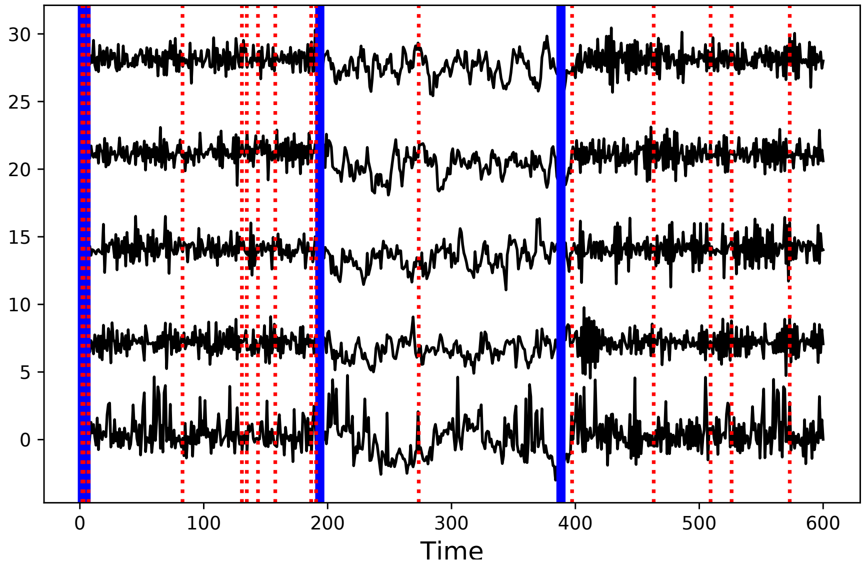

Figure 5.7: Maximum A Posteriori (MAP) CPs of standard BOCPD and shown as dashed vertical lines and RBOCPD show in soldi lines. True CPs at $t = 200, 400$. In **high dimensions** it becomes increasingly likely that the model's tails are misspecified in at least one dimension.

5.5.3 The London air pollution dataset

To illustrate the issues surrounding high-dimensional robustness further, we return to the London air pollution example. We showed in Section 5.2 that when this dataset was analysed by Knoblauch and Damoulas [2018] using traditional Bayesian updating, 12 CPs were found in the space of a year. One of these corresponded to the introduction of the congestion charge, but the number of CPs found questioned the reliability of this analysis. The top two panels of Figure 5.8 correspond to this analysis. The lower of the two shows the run length posterior with its MAP at each time point in red and its MAP segmentation marked with grey crosses. This plot is the same as Figure 5.2 shown in Section 5.2. Additionally, the top plot is the corresponding model posterior for this analysis. Knoblauch and Damoulas [2018] augmented previous BOCPD algorithms with a model universe, allowing inference to be conducted on the model as well as the parameters. In order to analyse the air pollution data they considered a model universe consisting of BVAR models with different lag lengths and different spatially structured neighbourhoods for the predictors. The top plot of Figure 5.8 shows how the posterior mass for three of the most likely BVAR models changes throughout the time window of the data. The model plotted in blue is clearly favoured after the introduction of the congestion charge with a combination of the yellow and green models favoured before those. See Knoblauch and Damoulas [2018] for more information on these specific models.

As we have mentioned before, we think it is possible that many of the declared CPs under BOCPD result from model misspecification. We therefore, seek to robustify the CP detection to investigate this further. Previous robust on-line methods [e.g. Pollak, 2010; Cao and Xie, 2017; Fearnhead and Rigaill, 2017] cannot be applied to this problem because they assume univariate data or do not allow for dependent observations.

Contrary to BOCPD, the bottom two panels of Figure 5.8 shows that RBOCPD finds only one CP close to the introduction of the congestion charge. This suggests that some of the CPs declared under standard BOCPD resulted from small perturbations from the underlying model. Standard BOCPD was highly sensitive to these while RBOCPD was able to ignore them. However, the fact that RBOCPD was still able to detect a CP around the time of the introduction of the congestion charge demonstrates that while being robust to outlying segments, the βD is still able to detect true changes to the underlying dynamics of pollution. In fact, the CP found by RBOCPD occurs before the introduction of the congestion charge where BOCPD found one after. It seems more likely to us that a change in dynamics would happen before the charge is introduced, as people anticipate the change in policy, rather than

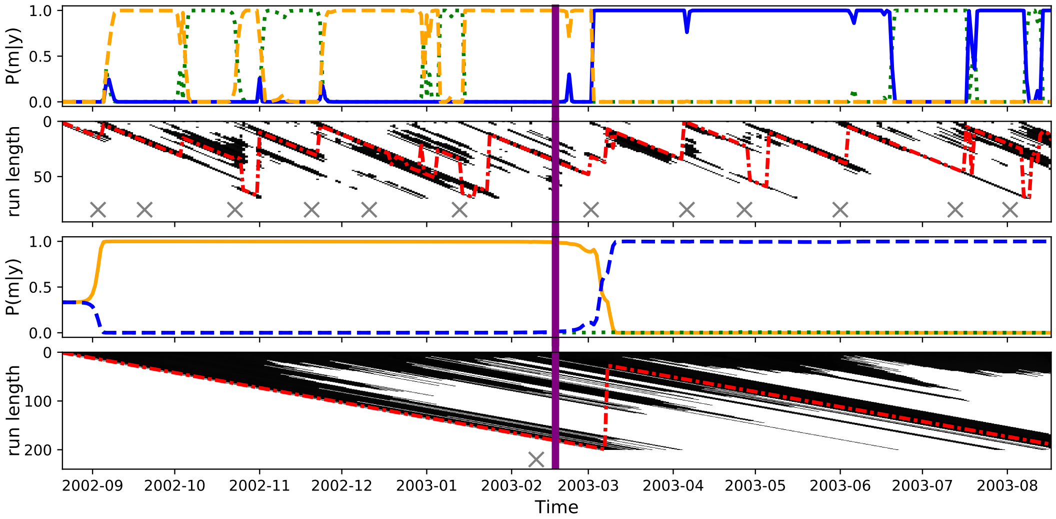

Figure 5.8: Air pollution data: On-line model posteriors for three of the most likely BVAR models (solid, dashed, dotted) and run-length posteriors (plotted in greyscale) with most likely run-lengths dashed for standard BOCPD (top two panels) and the RBOCPD (bottom two panels). Also marked are the congestion charge introduction, 17/02/2003 (solid vertical line) and the MAP segmentations (crosses)

after. Further, RBOCPD increases the average one-step-ahead predictive likelihood by 10% compared to standard BOCPD.

Figure 5.8 also demonstrates the robustness-efficiency trade-off experienced while using the βD to robustify the run-length posterior. Although the RBOCPD detects a CP close to the congestion charge in the MAP segmentation, looking at the run-length posterior it takes a number of observations to realise that a CP had occurred. In order to be robust to outliers, a greater build up of evidence is required for the run-length posterior to favour a CP. Therefore there is some delay in spotting 'actual' changes in the underlying system.

5.6 Conclusion

In this chapter I have shown how GVI can be used to solve an important and difficult problem in modern statistics. Using the robust model based βD loss function and the well motivated quasi-conjugate constraining family, GVI was able to produce robust CP and parameter inference in a computationally efficient manner that allows for on-line processing of data. While CP detection is a particularly salient example of unaddressed heterogeneity and outliers leading to poor inference, the capabilities of GVI presented extend far beyond this setting. With an ever increasing interest in the field of machine learning to efficiently and reliably quantify uncertainty, robust probabilistic inference will only become more relevant. This chapter presents a particularly striking demonstration of the inferential power that can be unlocked through divergence-based General Bayesian inference.

Chapter 6

Conclusion

This thesis has proposed several modifications to the applications of Bayesian analysis to modern, high-dimensional problems. Chapter 2 provides a DM with the ability to produce the most useful inference for their given problem by changing the target parameter of their inference through the loss function. Specific attention is given to tailoring inference to produce accurate estimates of expected utilities. Chapter 3 simplifies the task of belief elicitation for the DM. Providing a method to update beliefs that is both implementable and provides stability across an interpretable neighbourhood of probability models. Chapter 4 addresses the issue of computation in a Bayesian analysis, allowing a DM to tailor the inferential optimisation they solve to produce fast and desirable inferences. Lastly, Chapter 5 provides a practical demonstration of the methods proposed in this thesis and how they can be used to extract useful inferences from real-world data when traditional Bayesian updating fails. Several interesting areas of further work are outlined below.

Selecting the Divergence (hyperparameter): More work is required to advise on the selection of the divergence used for the updating. In general we believe the βD and γD will be more appealing for practical applications and this is reflected in Chapters 4 and 5. However, methods to choose between these two and further, strategies beyond those discussed in Sections 2.7.5 and 5.4.6 to select the divergence hyperparameters are very important in order for these methods to become practically useful tools. One promising avenue here is to consider methods additional to the influence curves of Section 2.6.7 to more clearly articulate the impact of choosing a certain divergence and its hyperparameter. Additionally, it would be interesting to explore whether techniques similar to Williamson et al. [2015] can exploit the concurrence, or lack there of, of analysis using different divergences to further help

DM's improve their belief specifications.

Theoretical Results: Theoretical results provide a further option to more clearly articulate the impact of the subjective selection of each divergence. The stability results in Chapter 3 are both interesting a novel and further inspection of these could lead to methods to help choose hyperparameters. For example results similar to that of the insensitivity of the γD to linear ϵ-contamination [Hung et al., 2018] would be useful to illustrate which model misspecifications these different divergences are robust to. Despite the results in Chapter 3, the theoretical analysis of these methods is far from complete and is certainly an area for further research. In order to convince practitioners to move away from Bayes' rule rigorous guarantees on the performance of alternative methods must be provided.

Computation: Another issue not fully explored in the thesis is how to tailor computational algorithms to the inferences we describe here. We discussed in Section 2.7.5 that even for simple models not minimising the KLD breaks the conjugacy property of Bayes' rule updating. Although Chapter 5 presented our quasi-conjugate posterior approximation which was specifically tailored to the βD (or γD), we have not experimented with 'exact' inference schemes. There exists a vast literature on optimising MCMC algorithms to sample from traditional Bayesian posteriors and in order to fully take advantage of the subjectivity this paper allows a statistician, a whole new class of computational algorithms tailored to different divergences may be required. In addition several of the divergences mentioned in Chapter 2 require a density estimate of the underlying process. In Chapter 3 we assumed a consistent estimate of this existed and further research into effectively doing this for complex high dimensional datasets can only improve the performances of these methods for real world problems.

Real-world examples: Further experimentation to that considered in Chapters 4 and 5, with complex real world data sets is also required to analyse how this robustness-efficiency trade-off associated with the selected divergence manifests itself in practice.

Mixture Models: Mixture models provide a particularly interesting application domain for the methodology in this paper. Particularly as their construction is amenable to the TVD neighbourhoods discussed in Chapter 3, i.e. if we know that 3 mixture components account for 95% of the observations but are unsure about the rest, this describes a TVD neighbourhood of size 0.05. However, there are several challenges to implementing these robust minimum divergence methods with mixture models. Firstly, for the βD and γD loss function we need to be able to calculate the term $\int f(z;\theta)^{\beta}dz$. Now if $f(z;\theta)$ is a mixture model then $f(z;\theta)^{\beta}$ will not be easily available in closed form and therefore we must result to numerical methods to calculate this integral. This integral is potentially evaluated for many values of θ and thus numerically approximating this presents a computational challenge.

Additionally, we have been made aware of the literature of Woodard et al. [2009a,b] who demonstrate problems with simulated and parallel tempering algorithms for multi-modal targets. The problems are caused by the fact that in cases when the modes are non-symmetric, the tempered targets do not preserve the regional/modal weights. That is to say, that the rate at which the tempered modes gain mass is connected to their variance in the original target. While this may seem unrelated, applying the βD or γD loss functions to a mixture model likelihood raises a multi-modal target to a power almost always less than 1. This phenomenon above could well cause problems with finite sample efficiency in the $\mathcal{M}$-CLOSED world and require some investigation of exactly what the βD minimising mixture model approximation might look like in the $\mathcal{M}$-OPEN world. If such problems exist one could appeal to the solutions in the statistical simulation literature [e.g. Tawn et al., 2018].

The variational/constraining family: Chapter 2 focused on the loss function, one input into the GVI loss function and Chapter 4 focussed on the uncertainty quantifying divergence. The third component of our generalised representation of Bayesian inference is the admissible set of densities for posterior. Chapter 4 stuck to default variational families, however we feel that viewing these as a constraining family rather than an approximating family opens up several interesting area of research.

Can we define principled constraining families? Are there subjective judgements that DMs cannot input into their prior or model but could go into the constraining family? The posterior regularisation of Ganchev et al. [2010] provides one example. Another situation could be the label switching scenario depicted in Figure 4.9. Here, the DM would surely prefer univariate posteriors even when the exact posterior is bivariate. This scenario is usual solved by post processing as by encoding an ordering of coefficients is difficult a priori [Marin and Robert, 2007]. A further example of this could be the ‘quasi-conjugate’ setting introduced in Section 5.4.4 where the posterior is forced into the same family as the prior.

Bernstein von Mises theorems tell us that the majority of posterior distributions are going to be reasonably approximated by a Gaussian (though we note that notable exceptions to apply) and as a result the poor performance of approximate inference methods comes from unaccounted for correlation. There is clearly a trade-off between capturing correlation and computational efficiency, every correlation considered is another variational hyperparameter. Methods such as Papaspiliopoulos and Rossell [2017] have considered this trade-off in linear regression scenarios and propose ways to optimise it.

Lastly GVI's increased flexibility to define its prior regulariser allows for the incorporation of penalised optimisation of hyperparameters. The VI objective function is often optimised for hyperparameters but the KLD prior regulairser associated with the KLD posterior approximation provides no room to penalise point estimates. These can now simply be added on to the GVI objective function, $D(q(\boldsymbol{\theta})||\pi(\boldsymbol{\theta})) + R(\lambda)$

Prior robustness: A further unexplored part of Chapter 4 is the prior robustness that can be gained by considering the prior regularising divergence to be any of the Rényi-αD, βD and γD. This was observed and compared empirically within the chapter but not thoroughly analysed. Further work could look at the theoretical properties of this and whether it could solve some of the undesirable results associated with Bayes' rule for misspecified priors [e.g. Gustafson and Wasserman, 1995]. In particular, it appeared as though the βD prior regularising divergence was able to produce posterior uncertainty quantification that was almost completely invariant to the prior. Much research has gone into 'objective-Bayes' and the specifications of prior distributions providing objective posterior uncertainty quantification [see, e.g. Jeffreys, 1961; Zellner, 1977; Bernardo, 1979; Berger and Bernardo, 1992; Jaynes, 2003; Berger, 2006]. Although the objectivity and uncertainty quantification associated with the GVI and the βD prior regulariser is not yet sufficiently well-understood, we believe it could be a very promising avenue for research in this area. Additionally methods to set the hyperparameter associated with these prior divergences have not been fully explored.

Robust model selection We believe the most exciting area for further work stemming from Chapter 5 comes from the connection between the BOCPD run-length posterior and the marginal-likelihood. The likelihood in the standard BOCPD algorithm corresponded the the marginal-likelihood of the data given the last CP which is the same quantity used in Bayes' Factor model selection [Kass and Raftery, 1995]. The lack of robustness in the standard BOCPD algorithm suggests that Bayes'

Factor model selection will have the same deficiencies.

Model selection with scores other than the log-score have been attempted before Dawid et al. [2015]; Shao et al. [2019] but these have been mainly motivated in the setting of an unnormalised prior rather than robustness to outliers. We believe applying the βD loss could prove powerful here. Specifically Gaussian graphical modelling, where models selection is done on models indexed by their conditional independence statements, could provide a salient example. Here Gaussianity is assumed because it has the convenient property that zeros in its precision matrix, $\Lambda = \Sigma^{-1}$, correspond to statements of conditional independence, rather than because the assumption of Gaussian observations is reasonable. Gaussian distributions are well-known to be non-robust [e.g. O'Hagan, 1979; Berger et al., 1994] and it is straightforward to construct an example where one observation forces the traditional Bayes factor to declare dependence between two dimensions when in fact there is none.

Dawid et al. [2015] prove model selection consistency for the general scoring function when used in a prequential (predictive sequential) manner [Dawid, 1984]. Under the prequential approach the focus of statistical inference is to make sequential probability forecasts about future observations rather than learning about model parameters. A convenient property of the log-score is that the same score is calculated if the data is viewed prequentially or together, this will not be the case for other scores [Fong and Holmes, 2019]. In the BOCPD setting above the prequential score naturally arises in the likelihood, however this may not be the case in wider model selection applications.

Bibliography

Ian S Abramson. On bandwidth variation in kernel estimates-a square root law. *The annals of Statistics*, pages 1217–1223, 1982.

Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Pierre Alquier, James Ridgway, and Nicolas Chopin. On the properties of variational approximations of gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414, 2016.

Mauricio Alvarez, Jan R Peters, Neil D Lawrence, and Bernhard Schölkopf. Switched latent force models for movement segmentation. In *Advances in neural information processing systems*, pages 55–63, 2010.

S Amari. Differential-geometrical methods in statistic. 1985.

Luca Ambrogioni, Umut Güçlü, Yağmur Güçlütürk, Max Hinne, Marcel AJ van Gerven, and Eric Maris. Wasserstein variational inference. In *Advances in Neural Information Processing Systems*, pages 2473–2482, 2018.

Pubh7440 Notes By Sudipto Banerjee. Bayesian linear model: Gory details. *Dowloaded from http://www. biostat. umn. edu/˜ ph7440*, 2008.

Andrew R Barron. *The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions.* Department of Statistics, University of Illinois, 1988.

Daniel Barry and John A Hartigan. Product partition models for change point problems. *The Annals of Statistics*, pages 260–279, 1992.

Ayanendranath Basu and Bruce G Lindsay. Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics*, 46(4):683–705, 1994.

Ayanendranath Basu, Ian R Harris, Nils L Hjort, and MC Jones. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3): 549–559, 1998.

Ayanendranath Basu, Hiroyuki Shioya, and Chanseok Park. *Statistical inference: the minimum distance approach.* CRC Press, 2011.

Matthew James Beal et al. *Variational algorithms for approximate Bayesian inference.* University of London London, 2003.

Rudolf Beran. Minimum hellinger distance estimates for parametric models. *The Annals of Statistics*, pages 445–463, 1977.

James O. Berger. The case for objective Bayesian analysis. *Bayesian analysis*, 1(3): 385–402, 2006.

James O Berger and José M Bernardo. On the development of the reference prior method. *Bayesian statistics*, 4(4):35–60, 1992.

James O Berger, Elías Moreno, Luis Raul Pericchi, M Jesús Bayarri, José M Bernardo, Juan A Cano, Julián De la Horra, Jacinto Martín, David Ríos-Insúa, Bruno Betrò, et al. An overview of robust bayesian analysis. *Test*, 3(1):5–124, 1994.

Robert H Berk et al. Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, 37(1):51–58, 1966.

Jose M. Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):113–128, 1979.

José M Bernardo and Adrian FM Smith. Bayesian theory, 2001.

Jacob Bernoulli. Positiones arithmeticae de seriebus infinitis earumque summa finita (treatise on infinite series), 1689.

Alexandros Beskos, Natesh Pillai, Gareth Roberts, Jesus-Maria Sanz-Serna, and Andrew Stuart. Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, 19(5A):1501–1534, 2013.

Olivier Binette. A note on reverse pinsker inequalities. *arXiv preprint arXiv:1805.05135*, 2019.

Christopher M Bishop. *Pattern recognition and machine learning.* springer, 2006.

PG Bissiri, CC Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

George EP Box. Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)*, pages 383–430, 1980.

Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

Thang Bui, Daniel Hernández-Lobato, Jose Hernandez-Lobato, Yingzhen Li, and Richard Turner. Deep gaussian processes for regression using approximate expectation propagation. In *International conference on machine learning*, pages 1472–1481, 2016.

Yang Cao and Yao Xie. Robust sequential change-point detection by convex optimization. In *Information Theory (ISIT), 2017 IEEE International Symposium on*, pages 1287–1291. IEEE, 2017.

François Caron, Arnaud Doucet, and Raphael Gottardo. On-line changepoint detection and parameter estimation with application to genomic data. *Statistics and Computing*, 22(2):579–595, 2012.

Bob Carpenter, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Michael A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 20, 2016.

Gilles Celeux, Jack Jewson, Julie Josse, Jean-Michel Marin, and Christian P Robert. Some discussions on the read paper" beyond subjective and objective in statistics" by a. gelman and c. hennig. *arXiv preprint arXiv:1705.03727*, 2017.

Herman Chernoff et al. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4): 493–507, 1952.

Victor Chernozhukov and Han Hong. An mcmc approach to classical estimation. *Journal of Econometrics*, 115(2):293–346, 2003.

Andrzej Cichocki and Shun-ichi Amari. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

Andrzej Cichocki, Sergio Cruces, and Shun-ichi Amari. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13(1):134–170, 2011.

Noel Cressie and Timothy RC Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 440–464, 1984.

Imre Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.*, 8:85–108, 1964.

Michelle L Dalrymple, Irene Lena Hudson, and Rodney Philip Kinvig Ford. Finite mixture, zero-inflated poisson and hurdle models with application to sids. *Computational Statistics & Data Analysis*, 41(3-4):491–504, 2003.

Andreas Damianou and Neil Lawrence. Deep gaussian processes. In *Artificial Intelligence and Statistics*, pages 207–215, 2013.

A Philip Dawid. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, pages 278–292, 1984.

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

A Philip Dawid, Monica Musio, et al. Bayesian model selection based on proper scoring rules. *Bayesian analysis*, 10(2):479–499, 2015.

A Philip Dawid, Monica Musio, and Laura Ventura. Minimum scoring rule inference. *Scandinavian Journal of Statistics*, 43(1):123–138, 2016.

Bruno De Finetti. Funzione caratteristica di un fenomeno aleatorio. 1931.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3): 411–436, 2006.

Arthur P Dempster. A subjectivist look at robustness. *Bull. Internat. Statist. Inst*, 46:349–374, 1975.

Luc Devroye and Laszlo Gyorfi. *Nonparametric density estimation: the L1 view*, volume 119. John Wiley & Sons Incorporated, 1985.

Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via chi upper bound minimization. In *Advances in Neural Information Processing Systems*, pages 2732–2741, 2017.

Shinto Eguchi et al. A differential geometric approach to statistical inference on the basis of contrast functionals. *Hiroshima mathematical journal*, 15(2):341–391, 1985.

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

Vassiliy A Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158, 1969.

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

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

Paul Fearnhead and Guillem Rigaill. Changepoint detection in the presence of outliers. *Journal of the American Statistical Association*, (just-accepted), 2017.

Peter Filzmoser, Ricardo Maronna, and Mark Werner. Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3):1694–1711, 2008.

Edwin Fong and Chris Holmes. On the marginal likelihood and cross-validation. *arXiv preprint arXiv:1905.08737*, 2019.

Emily Fox and David B Dunson. Multiresolution Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 737–745, 2012.

Hironori Fujisawa and Shinto Eguchi. Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis*, 99(9):2053–2081, 2008.

Futoshi Futami, Issei Sato, and Masashi Sugiyama. Variational inference based on robust divergences. *arXiv preprint arXiv:1710.06595*, 2017.

Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. Visualization in bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2):389–402, 2019.

Kuzman Ganchev, Jennifer Gillenwater, Ben Taskar, et al. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11 (Jul):2001–2049, 2010.

Andrew Gelman and Christian Hennig. Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2015.

Andrew Gelman, Daniel Simpson, and Michael Betancourt. The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10):555, 2017.

Zoubin Ghahramani and Matthew J Beal. Propagation algorithms for variational bayesian learning. In *Advances in neural information processing systems*, pages 507–513, 2001.

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

Abhik Ghosh and Ayanendranath Basu. General robust bayes pseudo-posterior: Exponential convergence results with applications. *arXiv preprint arXiv:1708.09692*, 2017.

Itzhak Gilboa and David Schmeidler. Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*, 18(2):141–153, 1989.

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

Michael Goldstein. Bayes linear analysis. *Wiley StatsRef: Statistics Reference Online*, 1999.

Michael Goldstein and David Wooff. *Bayes linear statistics: Theory and methods*, volume 716. John Wiley & Sons, 2007.

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

Irving John Good. Rational decisions. *Journal of the Royal Statistical Society, Series B*, 14:107–114, 1952.

Luca Greco, Walter Racugno, and Laura Ventura. Robust likelihood functions in bayesian inference. *Journal of Statistical Planning and Inference*, 138(5):1258–1270, 2008.

Peter Grünwald. Safe probability. *arXiv preprint arXiv:1604.01785*, 2016.

Peter Grünwald, Thijs Van Ommen, et al. Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12 (4):1069–1103, 2017.

Peter D Grünwald and A Philip Dawid. Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory. *Annals of Statistics*, pages 1367–1433, 2004.

Marco Grzegorczyk and Dirk Husmeier. Non-stationary continuous dynamic Bayesian networks. In *Advances in Neural Information Processing Systems*, pages 682–690, 2009.

Paul Gustafson and Larry Wasserman. Local sensitivity diagnostics for bayesian inference. *The Annals of Statistics*, 23(6):2153–2167, 1995.

Peter Hall, JS Marron, and Amnon Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427–444, 2005.

Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 114. John Wiley & Sons, 2011.

Bruce E Hansen. Nonparametric conditional density estimation. *Unpublished manuscript*, 2004.

Lars Peter Hansen and Thomas J Sargent. Acknowledging misspecification in macroeconomic theory. *Review of Economic Dynamics*, 4(3):519–535, 2001a.

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

Ernst Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 136:210–271, 1909.

Jose Hernandez-Lobato, Yingzhen Li, Mark Rowland, Thang Bui, Daniel Hernández-Lobato, and Richard Turner. Black-box alpha divergence minimization. In *International Conference on Machine Learning*, pages 1511–1520, 2016.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, volume 3, 2017.

Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401, 1999.

Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

CC Holmes and SG Walker. Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, 104(2):497–503, 2017.

Giles Hooker and Anand N Vidyashankar. Bayesian model robustness via disparities. *Test*, 23(3):556–584, 2014.

Chin-Wei Huang, Shawn Tan, Alexandre Lacoste, and Aaron C Courville. Improving explorability in variational inference with annealed variational objectives. In *Advances in Neural Information Processing Systems*, pages 9701–9711, 2018.

He Huang and Martin Paulus. Learning under uncertainty: a comparison between rw and Bayesian approach. In *Advances in Neural Information Processing Systems*, pages 2730–2738, 2016.

Peter J Huber. *Robust statistics*. Springer, 2011.

Peter J Huber and EM Ronchetti. Robust statistics, series in probability and mathematical statistics, 1981.

Hung Hung, Zhi-Yu Jou, and Su-Yun Huang. Robust mislabel logistic regression without modeling mislabel probabilities. *Biometrics*, 74(1):145–154, 2018.

Jenq-Neng Hwang, Shyh-Rong Lay, and Alan Lippman. Nonparametric multivariate density estimation: a comparative study. *IEEE Transactions on Signal Processing*, 42(10):2795–2810, 1994.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

Edwin T. Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

H. Jeffreys. Theory of probability: Oxford Univ. *Press (earlier editions 1939, 1948)*, 1961.

Jack Jewson, Jim Smith, and Chris Holmes. Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018.

MC Jones, Nils Lid Hjort, Ian R Harris, and Ayanendranath Basu. A comparison of related density-based minimum divergence estimators. *Biometrika*, 88(3):865–873, 2001.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

Azadeh Khaleghi and Daniil Ryabko. Locating changes in highly dependent data with unknown number of change points. In *Advances in Neural Information Processing Systems*, pages 3086–3094, 2012.

Rebecca Killick, Idris A Eckley, Kevin Ewans, and Philip Jonathan. Detection of changes in variance of oceanographic time-series using changepoint analysis. *Ocean Engineering*, 37(13):1120–1126, 2010.

Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Jeremias Knoblauch. Robust deep gaussian processes. *arXiv preprint arXiv:1904.02303*, 2019.

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

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Doubly robust Bayesian inference for non-stationary streaming data using β-divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 64–75, 2018.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Generalized variational inference. *arXiv preprint arXiv:1904.02063*, 2019.

George Konidaris, Scott Kuindersma, Roderic Grupen, and Andrew G Barto. Constructing skill trees for reinforcement learning agents from demonstration trajectories. In *Advances in neural information processing systems*, pages 1162–1170, 2010.

Arun Kumar Kuchibhotla and Ayanendranath Basu. A general set up for minimum disparity estimation. *Statistics & Probability Letters*, 96:68–74, 2015.

Arun Kumar Kuchibhotla and Ayanendranath Basu. On the asymptotics of minimum disparity estimation. *TEST*, pages 1–22, 2016.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Erich Kummerfeld and David Danks. Tracking time-varying graphical structure. In *Advances in neural information processing systems*, pages 1205–1213, 2013.

Sebastian Kurtek and Karthik Bharath. Bayesian sensitivity analysis with the fisher–rao metric. *Biometrika*, 102(3):601–616, 2015.

Jerald F Lawless. Negative binomial and mixed poisson regression. *Canadian Journal of Statistics*, 15(3):209–225, 1987.

David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

Brian G Leroux and Martin L Puterman. Maximum-penalized-likelihood estimation for independent and markov-dependent mixture models. *Biometrics*, pages 545–558, 1992.

Céline Levy-leduc and Zaïd Harchaoui. Catching change-points with lasso. In *Advances in Neural Information Processing Systems*, pages 617–624, 2008.

Meng Li and David B Dunson. A framework for probabilistic inferences from imperfect models. *arXiv preprint arXiv:1611.01241*, 2016.

Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081, 2016.

Moshe Lichman et al. Uci machine learning repository, 2013.

F Liese and I Vajda. Convex statistical distances, volume 95 of teubner-texte zur mathematik [teubner texts in mathematics]. *BSB BG Teubner Verlagsgesellschaft, Leipzig*, 1987.

Kevin Lin, James L Sharpnack, Alessandro Rinaldo, and Ryan J Tibshirani. A sharp error analysis for the fused Lasso, with application to approximate changepoint screening. In *Advances in Neural Information Processing Systems*, pages 6887–6896, 2017.

Bruce G Lindsay. Efficiency versus robustness: the case for minimum hellinger distance and related methods. *The annals of statistics*, pages 1081–1114, 1994.

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

Luo Lu, Hui Jiang, and Wing H Wong. Multivariate density estimation by bayesian sequential partitioning. *Journal of the American Statistical Association*, 108(504): 1402–1410, 2013.

SP Lyddon, CC Holmes, and SG Walker. General bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 2018.

Jean-Michel Marin and Christian Robert. *Bayesian core: a practical approach to computational Bayesian statistics.* Springer Science & Business Media, 2007.

Alexander G de G Matthews, James Hensman, Richard Turner, and Zoubin Ghahramani. On sparse variational methods and the kullback-leibler divergence between stochastic processes. *Journal of Machine Learning Research*, 51:231–239, 2016.

Minami Mihoko and Shinto Eguchi. Robust blind source separation by beta divergence. *Neural computation*, 14(8):1859–1886, 2002.

Jeffrey W Miller and David B Dunson. Robust bayesian inference via coarsening. *Journal of the American Statistical Association*, pages 1–13, 2018.

Thomas Minka. Power ep. Technical report, Technical report, Microsoft Research, Cambridge, 2004.

Thomas P Minka. Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001.

Tetsuzo Morimoto. Markov processes and the h-theorem. *Journal of the Physical Society of Japan*, 18(3):328–331, 1963.

Chipo Mufudza and Hamza Erol. Poisson mixture regression models for heart disease prediction. *Computational and mathematical methods in medicine*, 2016, 2016.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Scott Niekum, Sarah Osentoski, Christopher G Atkeson, and Andrew G Barto. CHAMP: Changepoint detection using approximate model parameters. Technical report, (No. CMU-RI-TR-14-10) Carnegie-Mellon University Pittsburgh PA Robotics Institute, 2014.

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

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

Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons, 2006.

Manfred Opper and Ole Winther. Gaussian processes for classification: Mean-field algorithms. *Neural computation*, 12(11):2655–2684, 2000.

Art Owen. Empirical likelihood for linear models. *The Annals of Statistics*, pages 1725–1747, 1991.

Omiros Papaspiliopoulos and D Rossell. Bayesian block-diagonal variable selection and model averaging. *Biometrika*, 104(2):343–359, 2017.

Fengchun Peng and Dipak K Dey. Bayesian analysis of outlier problems using divergence measures. *Canadian Journal of Statistics*, 23(2):199–213, 1995.

MS Pinkser. Information and information stability of random variables and processes, 1964.

Moshe Pollak. A robust changepoint detection method. *Sequential Analysis*, 29(2): 146–161, 2010.

Aleksey S Polunchenko, Alexander G Tartakovsky, and Nitis Mukhopadhyay. Nearly optimal change-point detection with an application to cybersecurity. *Sequential Analysis*, 31(3):409–435, 2012.

Parikshit Ram and Alexander G Gray. Density estimation trees. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 627–635. ACM, 2011.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

Rajesh Ranganath, Dustin Tran, Jaan Altosaar, and David Blei. Operator variational inference. In *Advances in Neural Information Processing Systems*, pages 496–504, 2016.

Jean-Baptiste Regli and Ricardo Silva. Alpha-beta divergence for variational inference. *arXiv preprint arXiv:1805.01045*, 2018.

Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

M Rosenblatt et al. On the maximal deviation of k-dimensional density estimates. *The Annals of Probability*, 4(6):1009–1015, 1976.

David Rossell. A framework for posterior consistency in model selection. *arXiv preprint arXiv:1806.04071*, 2018.

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

Simone Rossi, Sebastien Marmin, and Maurizio Filippone. Walsh-hadamard variational inference for bayesian deep learning. *arXiv preprint arXiv:1905.11248*, 2019.

Ó Ruanaidh, JK Joseph, and William J Fitzgerald. Numerical Bayesian methods applied to signal processing. 1996.

Donald B Rubin. The bayesian bootstrap. *The annals of statistics*, pages 130–134, 1981.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2): 319–392, 2009.

Eric Ruggieri and Marcus Antonellis. An exact approach to Bayesian sequential change point detection. *Computational Statistics & Data Analysis*, 97:71–86, 2016.

Yunus Saatçi, Ryan D Turner, and Carl E Rasmussen. Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 927–934, 2010.

Abhijoy Saha, Karthik Bharath, and Sebastian Kurtek. A geometric variational approach to bayesian inference. *Journal of the American Statistical Association*, (just-accepted):1–26, 2019.

Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep gaussian processes. In *Advances in Neural Information Processing Systems*, pages 4588–4599, 2017.

Igal Sason and Sergio Verdu. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

Leonard J Savage. *The foundations of statistics.* Courier Corporation, 1972.

Mark Schmidt, Glenn Fung, and Rmer Rosales. Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *European Conference on Machine Learning*, pages 286–297. Springer, 2007.

Stephane Shao, Pierre E Jacob, Jie Ding, and Vahid Tarokh. Bayesian model comparison with the hyvärinen score: computation and consistency. *Journal of the American Statistical Association*, pages 1–24, 2019.

Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

Jim Q Smith. *Bayesian decision analysis: principles and practice.* Cambridge University Press, 2010.

Jim Q Smith and Fabio Rigat. Isoseparation and robustness in finite parameter bayesian inference. *Annals of the Institute of Statistical Mathematics*, 64:495–519, 2012.

JQ Smith. Bayesian approximations and the hellinger metric. Unpublished manuscript, October 1995.

JQ Smith. Local robustness of bayesian parametric inference and observed likelihoods. 2007.

Florian Stimberg, Manfred Opper, Guido Sanguinetti, and Andreas Ruttor. Inference in continuous-time change-point models. In *Advances in Neural Information Processing Systems*, pages 2717–2725, 2011.

Nicholas Syring and Ryan Martin. Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486, 2019.

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

Roy N Tamura and Dennis D Boos. Minimum hellinger distance estimation for multivariate location and covariance. *Journal of the American Statistical Association*, 81(393):223–229, 1986.

Nicholas G Tawn, Gareth O Roberts, and Jeffrey S Rosenthal. Weight-preserving simulated tempering. *Statistics and Computing*, pages 1–15, 2018.

R E Turner and M Sahani. Two problems with variational expectation maximisation for time-series models. In *Bayesian time series models*. Cambridge University Press, 2011.

Ryan Turner, Yunus Saatci, and Carl Edward Rasmussen. Adaptive sequential Bayesian change point detection. In *Temporal Segmentation Workshop at NIPS*, 2009.

Ryan D Turner, Steven Bottone, and Clay J Stanek. Online variational approximations to non-exponential family change point models: with application to radar tracking. In *Advances in Neural Information Processing Systems*, pages 306–314, 2013.

Ryan Darby Turner. *Gaussian processes for state space models and change point detection.* PhD thesis, University of Cambridge, 2012.

Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2015.

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

Peiming Wang, Martin L Puterman, Iain Cockburn, and Nhu Le. Mixed poisson regression models with covariate dependent rates. *Biometrics*, pages 381–400, 1996.

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

Peter Whittle and Peter R Whittle. *Risk-sensitive optimal control*, volume 20. Wiley New York, 1990.

Daniel Williamson, Michael Goldstein, et al. Posterior belief assessment: Extracting meaningful subjective judgements from bayesian analyses with complex statistical models. *Bayesian Analysis*, 10(4):877–908, 2015.

Robert C Wilson, Matthew R Nassar, and Joshua I Gold. Bayesian online learning of the hazard rate in change-point problems. *Neural computation*, 22(9):2452–2476, 2010.

Robert L Winkler and Allan H Murphy. Evaluation of subjective precipitation probability forecasts. In *Proceedings of the first national conference on statistical meteorology*, pages 148–157. American Meteorological Society Boston, 1968.

Dawn Woodard, Scott Schmidler, Mark Huber, et al. Sufficient conditions for torpid mixing of parallel and simulated tempering. *Electronic Journal of Probability*, 14: 780–804, 2009a.

Dawn B Woodard, Scott C Schmidler, Mark Huber, et al. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *The Annals of Applied Probability*, 19(2):617–640, 2009b.

Xiang Xuan and Kevin Murphy. Modeling changing dependency structure in multivariate time series. In *Proceedings of the 24th international conference on Machine learning*, pages 1055–1062. ACM, 2007.

Yun Yang, Debdeep Pati, and Anirban Bhattacharya. alpha-variational inference with statistical guarantees. *arXiv preprint arXiv:1710.03266*, 2017.

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Yes, but did it work?: Evaluating variational inference. *arXiv preprint arXiv:1802.02538*, 2018a.

Yuling Yao, Aki Vehtari, Daniel Simpson, Andrew Gelman, et al. Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3):917–1007, 2018b.

Arnold Zellner. Maximal data information prior distributions. *New developments in the applications of Bayesian methods*, pages 211–232, 1977.

Arnold Zellner. Optimal information processing and bayes's theorem. *The American Statistician*, 42(4):278–280, 1988.

XianXing Zhang, Lawrence Carin, and David B Dunson. Hierarchical topic modeling for analysis of time-evolving personal choices. In *Advances in Neural Information Processing Systems*, pages 1395–1403, 2011.

Fang Zhou, Q Claire, and Ross D King. Predicting the geographical origin of music. In *2014 IEEE International Conference on Data Mining*, pages 1115–1120. IEEE, 2014.

Jun Zhu, Ning Chen, and Eric P Xing. Bayesian inference with posterior regularization and applications to infinite latent svms. *The Journal of Machine Learning Research*, 15(1):1799–1847, 2014.

Chapter 7

Appendix

7.1 Extending Theorems 3, 4, 5 and 6 to non-equal parameter spaces

Here we extend the results of Chapter 3 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 $x_{1:n} \sim g(\cdot)$ 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 $\theta^*_{f\backslash h}$ and $\theta^*_{h\backslash f}$ such that the posteriors $\pi_f^D(\theta_f|X_{1:n})$ and $\pi_h^D(\theta_h|X_{1:n})$ have concentrated to ensure

$$\begin{aligned}
&\int_{\Theta_f} D(g, h(\cdot; \left\{\theta_{U,f}, \theta^*_{h\backslash f}\right\}))\pi_f^D(\left\{\theta_{U,f}, \theta_{f\backslash h}\right\}|x_{1:n})d\theta_{U,f}d\theta_{f\backslash h}\\
\geq &\int_{\Theta_h} D(g, h(\cdot; \left\{\theta_{U,h}, \theta_{h\backslash f}\right\}))\pi_h^D(\left\{\theta_{U,h}, \theta_{h\backslash f}\right\}|x_{1:n})d\theta_{U,h}d\theta_{h\backslash f} \qquad (7.1)\\
&\int_{\Theta_h} D(g, f(\cdot; \left\{\theta_{U,h}, \theta^*_{f\backslash h}\right\}))\pi_h^D(\left\{\theta_{U,h}, \theta_{h\backslash f}\right\}|x_{1:n})d\theta_{U,h}d\theta_{h\backslash f}\\
\geq &\int_{\Theta_f} D(g, f(\cdot; \left\{\theta_{U,f}, \theta_{f\backslash h}\right\}))\pi_f^D(\left\{\theta_{U,f}, \theta_{f\backslash h}\right\}|x_{1:n})d\theta_{U,f}d\theta_{f\backslash h}. \qquad (7.2)
\end{aligned}$$

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

We require the introduction of $\theta^*_{f\backslash h}$ and $\theta^*_{h\backslash f}$ 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 3, 4, 5 and 6.

Theorem 13 (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 $x_{1:n}$ such that Condition 3 holds for divergence $D_M(\cdot,\cdot)$
- For the two likelihood models $\{f(\cdot;\theta_f):\theta_f\in\Theta_f\}$ and $\{h(\cdot;\theta_h):\theta_h\in\Theta_h\}$ there exists $\epsilon>0$ such that $f,h\in\mathcal{N}_{\epsilon}^{D_M}$ as defined in Definition 15.

Then for $m_f^{D_M}$ and $m_h^{D_M}$ as defined in Eq. (3.7)

$$D_M(m_f^{D_M}(\cdot|x_{1:n}),m_h^{D_M}(\cdot|x_{1:n})))\leq R^{D_M}(g,f,h,x_{1:n})+\epsilon, \tag{7.3}$$

where

$$\begin{aligned}&R^{D_M}(g,f,h,x_{1:n}):= \qquad (7.4)\\&2\min\left\{\int\left(D_M(g(\cdot),f(\cdot;\theta_f))\right)\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f,\int\left(D_M(g(\cdot),h(\cdot;\theta_h))\right)\pi_h^{D_M}(\theta_h|x_{1:n})d\theta_h\right\}.\end{aligned}$$

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

$$\psi(\mathbb{E}_X[\rho(X)])\leq\mathbb{E}_X[\psi(\rho(X))]. \tag{7.5}$$

Consider applying this with θ_f as the random variable of interest with distribution $\pi_f^{D_M}(\theta_f|x_{1:n})$, $\rho(\theta)=f(y;\theta)$ 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 the Bayesian predictive distribution is defined,

$$m_f^{D_M}(z|x_{1:n})=\mathbb{E}_{\pi_f^{D_M}(\theta_f|x_{1:n})}[f(z;\theta_f)]=\int f(z;\theta_f)\pi_f^{D_M}(\theta_f|x_{1:n})d\theta_f<\infty,\quad\forall z \tag{7.6}$$

and that

$$\mathbb{E}_{\pi_f^{D_M}(\theta_f|x_{1:n})}[D_M(h(\cdot),f(\cdot;\theta_f))] = \int D_M(h(\cdot),f(\cdot;\theta_f))\pi_f^{D_M}(\theta_f|x_{1:n})d\theta < \infty. \quad (7.7)$$

We note that by symmetry we could exchange f for h above. Therefore, by the convexity of $D_M(\cdot,\cdot)$, Jensen's inequality can be applied as described above, first to $m_h^{D_M}(\cdot|x_{1:n})$ and then to $m_f^{D_M}(\cdot|x_{1:n})$. Therefore,

$$D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n}))) \leq \int D_M(m_f^{D_M}(\cdot|x_{1:n}), h(\cdot;\theta_h))\pi_h^{D_M}(\theta_h|X_{1:n})d\theta_h \quad (7.8)$$

$$\leq \int \left\{ \int D_M(f(\cdot;\theta_f),h(\cdot;\theta_h))\pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f \right\} \pi_h^{D_M}(\theta_h|X_{1:n})d\theta_h. \quad (7.9)$$

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

$$D_M(f,h) \leq D_M(f,g) + D_M(g,h) = D_M(g,f) + D_M(g,h), \quad (7.10)$$

which can be used to show that

$$\begin{aligned}
&D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n}))) \\
&\quad\leq \int \left\{ \int D_M(f(\cdot;\theta_f),h(\cdot;\theta_h))\pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f \right\} \pi_h^{D_M}(\theta_h|X_{1:n})d\theta_h \\
&\quad\leq \int \left\{ \int D_M(g,f(\cdot;\theta_f)) + D_M(g,h(\cdot;\theta_h))\pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f \right\} \pi_h^{D_M}(\theta_h|X_{1:n})d\theta_h \quad (7.11)\\
&\quad= \int D_M(g,f(\cdot;\theta_f))\pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f + \int D_M(g,h(\cdot;\theta_h))\pi_h^{D_M}(\theta_h|X_{1:n})d\theta_h. \quad (7.12)
\end{aligned}$$

Now we decompose the parameter for each model into the part shared by the two likelihood models and what is left over. We therefore consider

$$\theta_f = \left\{\theta_{U,f}, \theta_{f\backslash h}\right\} \quad (7.13)$$

$$\theta_h = \left\{\theta_{U,h}, \theta_{h\backslash f}\right\}. \quad (7.14)$$

As a result, we can equivalently write

$$\begin{aligned}
&D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n})))\\
&\leq \int D_M(g, f(\cdot;\theta_f))\pi_f^{D_M}(\theta_f|X_{1:n})d\theta_f + \int D_M(g, h(\cdot;\theta_h))\pi_h^{D_M}(\theta_h|X_{1:n})d\theta_h\\
&= \int D_M(g, f(\cdot;\left\{\theta_{U,f},\theta_{f\backslash h}\right\}))\pi_f^{D_M}(\left\{\theta_{U,f},\theta_{f\backslash h}\right\}|X_{1:n})d\theta_{U,f}d\theta_{f\backslash h} \qquad (7.15)\\
&\quad + \int D_M(g, h(\cdot;\left\{\theta_{U,h},\theta_{h\backslash f}\right\}))\pi_h^{D_M}(\left\{\theta_{U,h},\theta_{h\backslash f}\right\}|X_{1:n})d\theta_{U,h}d\theta_{h\backslash f}.
\end{aligned}$$

Now given the first part of Condition 3, equation (7.1)

$$\begin{aligned}
&D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n})))\\
&\leq \int D_M(g, f(\cdot;\left\{\theta_{U,f},\theta_{f\backslash h}\right\}))\pi_f^{D_M}(\left\{\theta_{U,f},\theta_{f\backslash h}\right\}|X_{1:n})d\theta_{U,f}d\theta_{f\backslash h}\\
&\quad + \int D_M(g, h(\cdot;\left\{\theta_{U,h},\theta_{h\backslash f}\right\}))\pi_h^{D_M}(\left\{\theta_{U,h},\theta_{h\backslash f}\right\}|X_{1:n})d\theta_{U,h}d\theta_{h\backslash f}\\
&\leq \int D_M(g, f(\cdot;\left\{\theta_{U,f},\theta_{f\backslash h}\right\}))\pi_f^{D_M}(\left\{\theta_{U,f},\theta_{f\backslash h}\right\}|X_{1:n})d\theta_{U,f}d\theta_{f\backslash h}\\
&\quad + \int D_M(g, h(\cdot;\left\{\theta_{U,f},\theta^*_{h\backslash f}\right\}))\pi_f^{D_M}(\left\{\theta_{U,f},\theta_{f\backslash h}\right\}|X_{1:n})d\theta_{U,f}d\theta_{f\backslash h} \qquad (7.16)\\
&= \int \Big(D_M(g, f(\cdot;\left\{\theta_{U,f},\theta_{f\backslash h}\right\})) + D_M(g, h(\cdot;\left\{\theta_{U,f},\theta^*_{h\backslash f}\right\}))\Big)\\
&\qquad \pi_f^{D_M}(\left\{\theta_{U,f},\theta_{f\backslash h}\right\}|X_{1:n})d\theta_{U,f}d\theta_{f\backslash h}. \qquad (7.17)
\end{aligned}$$

We can add and subtract $D_M(f(\cdot;\left\{\theta_{U,f},\theta_{f\backslash h}\right\}), h(\cdot;\left\{\theta_{U,f},\theta^*_{h\backslash f}\right\}))$ inside the integral to give

$$\begin{aligned}
&D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n})))\\
&\leq \int \Big(D_M(g, f(\cdot;\left\{\theta_{U,f},\theta_{f\backslash h}\right\})) + D_M(g, h(\cdot;\left\{\theta_{U,f},\theta^*_{h\backslash f}\right\}))\\
&\quad - D_M(f(\cdot;\left\{\theta_{U,f},\theta_{f\backslash h}\right\}), h(\cdot;\left\{\theta_{U,f},\theta^*_{h\backslash f}\right\})) \qquad (7.18)\\
&\quad + D_M(f(\cdot;\left\{\theta_{U,f},\theta_{f\backslash h}\right\}), h(\cdot;\left\{\theta_{U,f},\theta^*_{h\backslash f}\right\}))\Big)\,\pi_f^{D_M}(\left\{\theta_{U,f},\theta_{f\backslash h}\right\}|X_{1:n})d\theta_{U,f}d\theta_{f\backslash h}.
\end{aligned}$$

Finally applying the triangle inequality once more gives us that

$$D_M(g,f) + D_M(f,h) \geq D_M(g,h) \Rightarrow D_M(g,f) \geq D_M(g,h) - D_M(f,h) \qquad (7.19)$$

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

$$\begin{aligned}
&D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n})))\\
\leq&\int \Big(D_M(g, f(\cdot; \{\theta_{U,f}, \theta_{f\setminus h}\})) + D_M(g, h(\cdot; \left\{\theta_{U,f}, \theta^*_{h\setminus f}\right\}))\\
&-D_M(f(\cdot; \{\theta_{U,f}, \theta_{f\setminus h}\}), h(\cdot; \left\{\theta_{U,f}, \theta^*_{h\setminus f}\right\}))\\
&+D_M(f(\cdot; \{\theta_{U,f}, \theta_{f\setminus h}\}), h(\cdot; \left\{\theta_{U,f}, \theta^*_{h\setminus f}\right\}))\Big)\, \pi_f^{D_M}(\{\theta_{U,f}, \theta_{f\setminus h}\}\,|X_{1:n}) d\theta_{U,f} d\theta_{f\setminus h}.\\
\leq& \int \left(2D_M(g, f(\cdot; \{\theta_{U,f}, \theta_{f\setminus h}\})) + \epsilon\right) \pi_f^{D_M}(\{\theta_{U,f}, \theta_{f\setminus h}\}\,|X_{1:n}) d\theta_f \qquad (7.20)\\
=& \,2\int \left(D_M(g, f(\cdot; \theta_f))\right) \pi_f^{D_M}(\theta_f|X_{1:n}) d\theta_f + \epsilon. \qquad (7.21)
\end{aligned}$$

We note that we could have applied the second part of Condition 3, equation (7.2), to exchange $\theta_f = \{\theta_{U,f}, \theta_{f\setminus h}\}$ for $\left\{\theta_{U,h}, \theta^*_{f\setminus h}\right\}$ in line (7.16) and the triangle inequality also gives us that

$$D_M(g, h) + D_M(f, h) \geq D_M(g, f) \Rightarrow D_M(g, h) \geq D_M(g, f) - D_M(f, h). \qquad (7.22)$$

Which, in turn can be used to show that

$$D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n}))) \leq 2\int \left(D_M(g, h(\cdot; \theta_h))\right) \pi_h^{D_M}(\theta_h|X_{1:n}) d\theta_h + \epsilon \qquad (7.23)$$

and thus

$$D_M(m_f^{D_M}(\cdot|x_{1:n}), m_h^{D_M}(\cdot|x_{1:n}))) \leq R^{D_M}(g, f, h, x_{1:n}) + \epsilon. \qquad (7.24)$$

where $R^{D_M}(g, f, h, x_{1:n})$ is defined in Eq. (7.4). □

Theorem 14 (Limiting predictive stability using divergence metrics)**.** Consider the following conditions:

- Divergence $D_M(\cdot, \cdot)$ satisfies M1 and M2 from Condition 1
- For the two likelihood models $\{f(\cdot; \theta_f) : \theta_f \in \Theta_f\}$ and $\{h(\cdot; \theta_h) : \theta_h \in \Theta_h\}$ there exists $\epsilon > 0$ such that $f, h \in \mathcal{N}_\epsilon^{D_M}$ as defined in Definition 15.

Then

$$\left|D_M(g, f(\cdot; \hat{\theta}_f^{D_M})) - D_M(g, h(\cdot; \hat{\theta}_h^{D_M}))\right| \leq \epsilon \qquad (7.25)$$

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

Proof. First we decompose

$$\hat{\theta}_f^{D_M} = \left\{\hat{\theta}_{U,f}^{D_M}, \hat{\theta}_{f\backslash h}^{D_M}\right\} \tag{7.26}$$

$$\hat{\theta}_h^{D_M} = \left\{\hat{\theta}_{U,h}^{D_M}, \hat{\theta}_{h\backslash f}^{D_M}\right\} \tag{7.27}$$

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 $\forall$ $\theta_U \in \Theta_U$, $\theta_{f\backslash h} \in \Theta_{f\backslash h}$ and $\theta_{h\backslash f} \in \Theta_{h\backslash f}$

$$\begin{aligned}
D_M(g, f(\cdot; \{\theta_U, \theta_{f\backslash h}\})) &\leq D_M(h(\cdot; \{\theta_U, \theta_{h\backslash f}\}), f(\cdot; \{\theta_U, \theta_{f\backslash h}\})) \qquad (7.28)\\
&\quad + D_M(g, h(\cdot; \{\theta_U, \theta_{h\backslash f}\}))\\
&\leq \epsilon + D_M(g, h(\cdot; \{\theta_U, \theta_{h\backslash f}\})) \qquad (7.29)\\
D_M(g, h(\cdot; \{\theta_U, \theta_{h\backslash f}\})) &\leq D_M(h(\cdot; \{\theta_U, \theta_{h\backslash f}\}), f(\cdot; \{\theta_U, \theta_{f\backslash h}\})) \qquad (7.30)\\
&\quad + D_M(g, f(\cdot; \{\theta_U, \theta_{f\backslash h}\}))\\
&\leq \epsilon + D_M(g, f(\cdot; \{\theta_U, \theta_{f\backslash h}\})). \qquad (7.31)
\end{aligned}$$

Now 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 we have

$$\begin{aligned}
D_M(g, f(\cdot; \left\{\hat{\theta}_{U,f}^{D_M}, \hat{\theta}_{f\backslash h}^{D_M}\right\})) &\leq D_M(g, f(\cdot; \left\{\hat{\theta}_{U,h}^{D_M}, \hat{\theta}_{f\backslash h}^{D_M}\right\})) \qquad (7.32)\\
&\leq \epsilon + D_M(g, h(\cdot; \left\{\hat{\theta}_{U,h}^{D_M}, \hat{\theta}_{h\backslash f}^{D_M}\right\})) \qquad (7.33)\\
D_M(g, h(\cdot; \left\{\hat{\theta}_{U,h}^{D_M}, \hat{\theta}_{h\backslash f}^{D_M}\right\})) &\leq D_M(g, h(\cdot; \left\{\hat{\theta}_{U,f}^{D_M}, \hat{\theta}_{h\backslash f}^{D_M}\right\})) \qquad (7.34)\\
&\leq \epsilon + D_M(g, f(\cdot; \left\{\hat{\theta}_{U,f}^{D_M}, \hat{\theta}_{f\backslash h}^{D_M}\right\})) \qquad (7.35)\\
\Rightarrow \left|D_M(g, h(\cdot; \hat{\theta}_h^{D_M})) - D_M(g, f(\cdot; \hat{\theta}_f^{D_M}))\right| &\leq \epsilon. \qquad (7.36)
\end{aligned}$$

□

Theorem 15 (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 $\{h(\cdot;\theta_h) : \theta_h \in \Theta_h\}$,

data generating process g such that

$$\max\{\operatorname{ess\,sup} f, \operatorname{ess\,sup} h, \operatorname{ess\,sup} g\} \leq M < \infty \tag{7.37}$$

- For the two likelihood models $\{f(\cdot;\theta_f) : \theta_f \in \Theta_f\}$ and $\{h(\cdot;\theta_h) : \theta_h \in \Theta_h\}$ there exists $\epsilon > 0$ such that $f, h \in \mathcal{N}_\epsilon^{D_M}$ as defined in Definition 15.

Then

$$\left| D_B^{(\beta)}(g||f(\cdot;\hat{\theta}_h^{(\beta)})) - D_B^{(\beta)}(g||h(\cdot;\hat{\theta}_f^{(\beta)}))\right| \leq \frac{M^{\beta-1}}{\beta-1}\epsilon \tag{7.38}$$

where $\hat{\theta}_f^{(\beta)} = \arg\min_\theta D_B^{(\beta)}(g||f(\cdot;\theta))$ and $\hat{\theta}_h^{(\beta)} = \arg\min_\theta D_B^{(\beta)}(g||h(\cdot;\theta))$.

Proof. First we decompose

$$\hat{\theta}_f^{(\beta)} = \left\{\hat{\theta}_{U,f}^{(\beta)}, \hat{\theta}_{f\backslash h}^{(\beta)}\right\} \tag{7.39}$$

$$\hat{\theta}_h^{(\beta)} = \left\{\hat{\theta}_{U,h}^{(\beta)}, \hat{\theta}_{h\backslash f}^{(\beta)}\right\} \tag{7.40}$$

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

Firstly, by 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 $\theta_{f\backslash h} \in \Theta_{f\backslash h}$ and $\theta_{h\backslash f} \in \Theta_{h\backslash f}$

$$D_B^{(\beta)}(g, f(\cdot;\left\{\hat{\theta}_{U,f}^{(\beta)}, \hat{\theta}_{f\backslash h}^{(\beta)}\right\})) \leq D_B^{(\beta)}(g, f(\cdot;\left\{\hat{\theta}_{U,h}^{(\beta)}, \theta_{f\backslash h}\right\})) \tag{7.41}$$

$$D_B^{(\beta)}(g, h(\cdot;\left\{\hat{\theta}_{U,h}^{(\beta)}, \hat{\theta}_{h\backslash f}^{(\beta)}\right\})) \leq D_B^{(\beta)}(g, h(\cdot;\left\{\hat{\theta}_{U,f}^{(\beta)}, \theta_{h\backslash f}\right\})). \tag{7.42}$$

Next, using the triangle type inequality proven in Lemma 4 and the definition of $\mathcal{N}_\epsilon^{\text{TVD}}$ shows that

$$\begin{aligned}
&D_B^{(\beta)}(g, f(\cdot;\left\{\hat{\theta}_{U,h}^{(\beta)}, \theta_{f\backslash h}\right\}))\\
\leq&\frac{M^{\beta-1}}{\beta-1}\text{TVD}(f(\cdot;\left\{\hat{\theta}_{U,h}^{(\beta)}, \theta_{f\backslash h}\right\}), h(\cdot;\left\{\hat{\theta}_{U,h}^{(\beta)}, \hat{\theta}_{h\backslash f}^{(\beta)}\right\})) + D_B^{(\beta)}(g, h(\cdot;\left\{\hat{\theta}_{U,h}^{(\beta)}, \hat{\theta}_{h\backslash f}^{(\beta)}\right\}))\\
\leq&\frac{M^{\beta-1}}{\beta-1}\epsilon + D_B^{(\beta)}(g, h(\cdot;\left\{\hat{\theta}_{U,h}^{(\beta)}, \hat{\theta}_{h\backslash f}^{(\beta)}\right\})) \qquad (7.43)\\
&D_B^{(\beta)}(g, h(\cdot;\left\{\hat{\theta}_{U,f}^{(\beta)}, \theta_{h\backslash f}\right\}))\\
\leq&\frac{M^{\beta-1}}{\beta-1}\text{TVD}(f(\cdot;\left\{\hat{\theta}_{U,f}^{(\beta)}, \hat{\theta}_{f\backslash h}^{(\beta)}\right\}), h(\cdot;\left\{\hat{\theta}_{U,f}^{(\beta)}, \theta_{h\backslash f}\right\})) + D_B^{(\beta)}(g, f(\cdot;\left\{\hat{\theta}_{U,f}^{(\beta)}, \hat{\theta}_{f\backslash h}^{(\beta)}\right\}))\\
\leq&\frac{M^{\beta-1}}{\beta-1}\epsilon + D_B^{(\beta)}(g, f(\cdot;\left\{\hat{\theta}_{U,f}^{(\beta)}, \hat{\theta}_{f\backslash h}^{(\beta)}\right\})). \qquad (7.44)
\end{aligned}$$

Combining these two inequalities results in

$$\Rightarrow \left| D_B^{(\beta)}(g, h(\cdot;\hat{\theta}_h^{(\beta)})) - D_B^{(\beta)}(g, f(\cdot;\hat{\theta}_f^{(\beta)})) \right| \leq \frac{M^{\beta-1}}{\beta-1}\epsilon \tag{7.45}$$

□

Theorem 16 (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

  $$\max\{\operatorname{ess\,sup} f, \operatorname{ess\,sup} h, \operatorname{ess\,sup} g\} \leq M < \infty, \tag{7.46}$$

  and priors $\pi(\theta_f)$ and $\pi(\theta_h)$ and data $x_{1:n}$ such that Condition 3 holds for divergence $D(\cdot,\cdot) = D_B^{(\beta)}(\cdot||\cdot)$
- For the two likelihood models $\{f(\cdot;\theta_f) : \theta_f \in \Theta_f\}$ and $\{h(\cdot;\theta_h) : \theta_h \in \Theta_h\}$ there exists $\epsilon > 0$ such that $f, h \in \mathcal{N}_\epsilon^{D_M}$ as defined in Definition 15.

Then

$$\begin{aligned} &D_B^{(\beta)}(m_f^{(\beta)}(\cdot|x_{1:n})||m_h^{(\beta)}(\cdot|x_{1:n})) \qquad (7.47)\\ &\quad\leq \frac{M^{\beta-1}}{\beta-1}\epsilon + \int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h\\ &D_B^{(\beta)}(m_h^{(\beta)}(\cdot|x_{1:n})||m_f^{(\beta)}(\cdot|x_{1:n})) \qquad (7.48)\\ &\quad\leq \frac{M^{\beta-1}}{\beta-1}\epsilon + \int\int R(g||h(\cdot;\theta_h)||f(\cdot;\theta_f))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h. \end{aligned}$$

where $R(g||f||h)$ and $R(g||h||f)$ were defined in Lemma 3 to be

$$R(g||f||h) = \int (g-f)\left(\frac{1}{\beta-1}h^{\beta-1} - \frac{1}{\beta-1}f^{\beta-1}\right)d\mu \tag{7.49}$$

$$R(g||h||f) = \int (g-h)\left(\frac{1}{\beta-1}f^{\beta-1} - \frac{1}{\beta-1}h^{\beta-1}\right)d\mu. \tag{7.50}$$

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 Theorems 3 and 13 to show that

$$D_B^{(\beta)}(m_f^{(\beta)}(\cdot|x_{1:n})||m_h^{(\beta)}(\cdot|x_{1:n}))) \leq \int D_B^{(\beta)}(f(y|X_{1:n})||h(\cdot;\theta_h))\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h \tag{7.51}$$

$$\leq \int \left\{ \int D_B^{(\beta)}(f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f \right\} \pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h. \tag{7.52}$$

Now the three-point property associated with the βD (Lemma 3) gives us that

$$D_B^{(\beta)}(f||h) = D_B^{(\beta)}(g||h) - D_B^{(\beta)}(g||f) + R(g||f||h) \tag{7.53}$$

where $R(g||f||h)$ is defined in Eq. (7.49) and using this here provides

$$\begin{aligned}
&D_B^{(\beta)}(m_f^{(\beta)}(\cdot|x_{1:n})||m_h^{(\beta)}(\cdot|x_{1:n}))) && (7.54)\\
&\leq \int \left\{ \int D_B^{(\beta)}(f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f \right\} \pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h \\
&= \int \left\{ \int \left[D_B^{(\beta)}(g||h(\cdot;\theta_h)) - D_B^{(\beta)}(g||f(\cdot;\theta_f)) \right.\right. \\
&\quad \left.\left. + R(g, f(\cdot;\theta_f), h(\cdot;\theta_h)\right] \pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f \right\} \pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h && (7.55)\\
&= \int D_B^{(\beta)}(g||h(\cdot;\theta_h))\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h - \int D_B^{(\beta)}(g||f(\cdot;\theta_f))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f \\
&\quad + \int \int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f \pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h. && (7.56)
\end{aligned}$$

Now we decompose the parameter for each model into the part shared by the two models and what is left over. We therefore consider

$$\theta_f = \left\{\theta_{U,f}, \theta_{f\backslash h}\right\} \tag{7.57}$$

$$\theta_h = \left\{\theta_{U,h}, \theta_{h\backslash f}\right\}. \tag{7.58}$$

We can then equivalently write

$$\begin{aligned}
&D_B^{(\beta)}(m_f^{(\beta)}(\cdot|x_{1:n})||m_h^{(\beta)}(\cdot|x_{1:n})))\\
\leq &\int D_B^{(\beta)}(g||h(\cdot;\theta_h))\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h - \int D_B^{(\beta)}(g||f(\cdot;\theta_f))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\\
&+\int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h \qquad (7.59)\\
= &\int D_B^{(\beta)}(g||h(\cdot;\{\theta_{U,h},\theta_{h\setminus f}\}))\pi_h^{(\beta)}(\{\theta_{U,h},\theta_{h\setminus f}\}|X_{1:n})d\theta_{U,h}d\theta_{h\setminus f}\\
&-\int D_B^{(\beta)}(g||f(\cdot;\{\theta_{U,f},\theta_{f\setminus h}\}))\pi_f^{(\beta)}(\{\theta_{U,f},\theta_{f\setminus h}\}|X_{1:n})d\theta_{U,f}d\theta_{f\setminus h}\\
&+\int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h. \qquad (7.60)
\end{aligned}$$

Now given the first part of Condition 3, Eq. (7.1), applied for the $D = D_B^{(\beta)}$ allows us to exchange $\pi_h^{(\beta)}(\left\{\theta_{U,h},\theta^*_{h\setminus f}\right\}|X_{1:n})$ for $\pi_f^{(\beta)}(\{\theta_{U,f},\theta_{f\setminus h}\}|X_{1:n})$ in the first integral

$$\begin{aligned}
&D_B^{(\beta)}(m_f^{(\beta)}(\cdot|x_{1:n})||m_h^{(\beta)}(\cdot|x_{1:n})))\\
\leq &\int D_B^{(\beta)}(g||h(\cdot;\{\theta_{U,h},\theta_{h\setminus f}\}))\pi_h^{(\beta)}(\{\theta_{U,h},\theta_{h\setminus f}\}|X_{1:n})d\theta_{U,h}d\theta_{h\setminus f}\\
&-\int D_B^{(\beta)}(g||f(\cdot;\{\theta_{U,f},\theta_{f\setminus h}\}))\pi_f^{(\beta)}(\{\theta_{U,f},\theta_{f\setminus h}\}|X_{1:n})d\theta_{U,f}d\theta_{f\setminus h}\\
&+\int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h\\
\leq &\int D_B^{(\beta)}(g||h(\cdot;\left\{\theta_{U,f},\theta^*_{h\setminus f}\right\}))\pi_f^{(\beta)}(\{\theta_{U,f},\theta_{f\setminus h}\}|X_{1:n})d\theta_{U,f}d\theta_{f\setminus h}\\
&-\int D_B^{(\beta)}(g||f(\cdot;\{\theta_{U,f},\theta_{f\setminus h}\}))\pi_f^{(\beta)}(\{\theta_{U,f},\theta_{f\setminus h}\}|X_{1:n})d\theta_{U,f}d\theta_{f\setminus h}\\
&+\int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h \qquad (7.61)\\
= &\int \left(D_B^{(\beta)}(g||h(\cdot;\left\{\theta_{U,f},\theta^*_{h\setminus f}\right\})) - \int D_B^{(\beta)}(g||f(\cdot;\{\theta_{U,f},\theta_{f\setminus h}\}))\right)\\
&\pi_f^{(\beta)}(\{\theta_{U,f},\theta_{f\setminus h}\}|X_{1:n})d\theta_{U,f}d\theta_{f\setminus h}\\
&+\int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h. \qquad (7.62)
\end{aligned}$$

The last line above 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 4, Eq. (3.73)

$$
\begin{aligned}
&D_B^{(\beta)}(m_f^{(\beta)}(\cdot|x_{1:n})||m_h^{(\beta)}(\cdot|x_{1:n})))\\
\leq&\int\left(D_B^{(\beta)}(g||h(\cdot;\left\{\theta_{U,f},\theta^*_{h\backslash f}\right\}))-\int D_B^{(\beta)}(g||f(\cdot;\left\{\theta_{U,f},\theta_{f\backslash h}\right\}))\right)\\
&\pi_f^{(\beta)}(\left\{\theta_{U,f},\theta_{f\backslash h}\right\}|X_{1:n})d\theta_{U,f}d\theta_{f\backslash h}\\
&+\int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h.\\
\leq&\int\frac{M^{\beta-1}}{\beta-1}\text{TVD}(h(\cdot;\left\{\theta_{U,f},\theta^*_{h\backslash f}\right\}),f(\cdot;\left\{\theta_{U,f},\theta_{f\backslash h}\right\}))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_{U,f}d\theta_{f\backslash h}\\
&+\int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h. \qquad (7.63)
\end{aligned}
$$

Which given the neighbourhood of likelihood models defined by $\mathcal{N}_\epsilon^{\text{TVD}}$ in Eq. (3.12) can be rewritten as

$$
\begin{aligned}
&D_B^{(\beta)}(m_f^{(\beta)}(\cdot|x_{1:n})||m_h^{(\beta)}(\cdot|x_{1:n})))\\
\leq&\int\frac{M^{\beta-1}}{\beta-1}\text{TVD}(h(\cdot;\left\{\theta_{U,f},\theta^*_{h\backslash f}\right\}),f(\cdot;\left\{\theta_{U,f},\theta_{f\backslash h}\right\}))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_{U,f}d\theta_{f\backslash h}\\
&+\int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h.\\
\leq&\frac{M^{\beta-1}}{\beta-1}\epsilon+\int\int R(g||f(\cdot;\theta_f)||h(\cdot;\theta_h))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h. \qquad (7.64)
\end{aligned}
$$

We note that we could have instead considered $D_B^{(\beta)}(m_h^{(\beta)}(\cdot|x_{1:n})||m_f^{(\beta)}(\cdot|x_{1:n})))$, applied the corresponding version for the three-point property of βD divergences, 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, to show that

$$
\begin{aligned}
&D_B^{(\beta)}(m_h^{(\beta)}(\cdot|x_{1:n})||m_f^{(\beta)}(\cdot|x_{1:n})))\\
\leq&\frac{M^{\beta-1}}{\beta-1}\epsilon+\int\int R(g||h(\cdot;\theta_h)||f(\cdot;\theta_f))\pi_f^{(\beta)}(\theta_f|X_{1:n})d\theta_f\pi_h^{(\beta)}(\theta_h|X_{1:n})d\theta_h. \qquad (7.65)
\end{aligned}
$$

□