Measuring Local Sensitivity in Bayesian Inference using a new class of metrics

Tabassom Sedighi[‡], Amin Hosseinian-Far[§] and Alireza Daneshkhah[*]

‡ Complex Systems Research Centre, School of Management, Cranfield University, UK

§ Department of Business Systems and Operations, The University of Northampton, Northampton, UK

* School of Computing, Electronics and Mathematics, Coventry University, UK

January 17, 2021

Abstract

The local sensitivity analysis is recognised for its computational simplicity, and potential use in multi-dimensional and complex problems. Unfortunately, its major drawback is its asymptotic behaviour where the prior to posterior convergence in terms of the standard metrics (and also computed by Fréchet derivative) used as a local sensitivity measure is not appropriate. The constructed local sensitivity measures do not converge to zero, and even diverge for the most multidimensional classes of prior distributions. Restricting the classes of priors or using other ϕ-divergence metrics have been proposed as the ways to resolve this issue which were not successful. We overcome this issue, by proposing a new flexible class of metrics so-called *credible metrics* whose asymptotic behaviour are far more promising and no restrictions is required to impose. Using these metrics, the stability of Bayesian inference to the structure of the prior distribution will

[*]The correspondance author; e-mail: ali.daneshkhah@coventry.ac.uk

be then investigated. Under appropriate condition, we present a uniform bound in a sense that a close credible metric a priori will give a close credible metric a posteriori. As a result, we do not get the sort of divergence based on other metrics. We finally show that the posterior predictive distributions are more stable and robust.

Keywords: *Bayesian robustness, Bayesian stability, credible metrics, local sensitivity analysis.*

1 Introduction

Robust Bayesian analysis is the study of sensitivity of Bayesian answers to uncertain inputs such as sampling model, prior distribution, or loss function, or any combination of them. There are several reasons to examine the robustness of Bayesian answers to the above misspecification: foundational motivation, practical Bayesian motivation, and acceptance of Bayesian analysis (see [14, 17, 12]).

Sensitivity analysis can be divided into two broad categories, global and local sensitivity. The common approach to assessing sensitivity is to measure the size of the class of posteriors (or perhaps just a particular posterior quantity) that arises from a specified class of priors. This is referred to as global sensitivity analysis ([6]). The global sensitivity analysis does not rely on perturbation lying within a given parametric family ([15]). Alternatively, an appropriate divergence measure is applied to first specify a neighbourhood system around each model. The bounds are then computed for the maximum deviation in the inference that could be obtained by a model in this neighbourhood. If this deviation is small then the model is considered ro be robust ([9]; [16].

The fact that global analyses often entail a large and complex computational problem ([7])) has led to the local sensitivity analyses, originally introduced by Gustafson and Wasserman []Gustafson-Wasserman95, and developed further Gustafson et al.[10]. The idea of a local analysis is to examine the rate at which the posterior changes, relative to the prior. In local sensitivity analysis, a chosen base prior distribution is perturbed using a finite parametrised modification. Hence, measures which are 'functionally close' to the chosen/elicited prior are considered and the behaviour of the posterior functional forms, under infinitesimal departures from the prior, are studied.

The local sensitivity analysis, as studied in the last chapter, is recognised for its computational simplicity, and its potential use in multi-dimensional and similar complex problems where global robustness investigation may be difficult ([5]). The major drawback of this approach is about the asymptotic behaviour. It is reasonable that in most cases the influence of prior distribution on the posterior quantities becomes less important as the sample size tends to infinity. In this article, we study asymptotic behaviour of the local sensitivity methods which measure the effect of infinitesimal perturbations of the prior on the posterior quantities ([1, 9]). We assume $\mathbf{x}_n = (x_1, \ldots, x_n)$, $n \geq 1$ is a random sample with observed sample densities $p(\mathbf{x}_n|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Let $\mathcal{P}$ be the set of all probability measures on the parameter space, Θ, and given a prior distribution $\pi(\boldsymbol{\theta})$, we denote $\pi(\boldsymbol{\theta} \mid \mathbf{x})$ as the corresponding posterior distribution. We denote $T : \mathcal{P} \to \mathcal{T}$ some quantity of interest. For instance, the whole posterior distribution can be derived by taking $T(\pi) = \pi(\boldsymbol{\theta} \mid \mathbf{x})$ and $\mathcal{T} = \mathcal{P}$. We denote the predictive distribution of a new observation $\mathbf{x}_n^*$ by, $p(\mathbf{x}_n|\mathbf{x}_n^*) = \int p(\mathbf{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{x}_n^*)d\boldsymbol{\theta}$.

Gustafson and Wasserman (1995) define the local sensitivity of a prior π in the direction of another prior ψ as

$$s(\pi, \psi) = \lim_{\epsilon \to 0} \frac{d_2(T(\psi_\epsilon), T(\pi))}{d_1(\psi_\epsilon, \pi)}, \tag{1}$$

where ψ_ϵ is the perturbed prior distribution, and d_1 and d_2 denote total variation distance unless otherwise stated. We denote the overall sensitivity by $s(\pi, \Gamma; \mathbf{x}_n) = \sup_{\psi \in \Gamma} s(\pi, \psi; \mathbf{x}_n)$, for some class of priors $\Gamma \subset \mathcal{P}$.

It was shown that under mild regularity conditions, $s(\pi, \mathcal{P})$ ($s(\pi, \Gamma)$ for many classes of Γ) increases at rate $n^{\frac{k}{2}}$ ([9]). Therefore, if we use this quantity as a diagnostic measure, the posterior distribution becomes increasingly sensitive to the chosen prior distribution as the sample size becomes very large. This is because, $\mathcal{P}$ consists of the unreasonable prior distributions (e.g., priors which put all the mass at one point, or priors that have very noisy behaviour at their tails). Restricting the class of priors to a subset Γ_R of $\mathcal{P}$ was then proposed by ([10]). However, they showed that the mentioned issue still remains as long as π is an interior point of Γ_R with respect to the density ratio metric. This is a very severe constraint on any prior family, but despite this, the type of divergence discussed above will still occur under this prior family constraint.

The parametric prior distributions as the restricted class was another solution to tackle the aforementioned issue proposed in ([9]). The corresponding Bayesian inference under any prior in this class is still rather unsatisfactory. Since, the local sensitivity measure will then depend on a prior lying in a particular parametric family which should be avoided. A similar asymptotic behaviour will be also observed even if other ϕ-divergence distances or the geometric perturbation are used for d_1 and d_2 given in Eq. (1).

In this paper, we examine prior to posterior convergence and the sensitivity issues mentioned above in terms of a new class of metrics called *credible metric*. Section 2 is dedicated to introduce this metric which its asymptotic behaviour is more desirable, and if it is used as a local sensitivity measure, does not require us to restrict ourselves to a particular class of priors. We also present some preliminary definitions, theorems and lemmas which will be used to study the asymptotic behaviour of this metric in Section 3. The predictive performance of this metric is investigated in Section 4. We show the computed metric in terms of the posterior predictive distributions would be be more stable and robust compared to the metric derived in terms of the posterior distributions.

2 A new class of metrics

In this section, we introduce a new class of metrics called *credible metrics*. We illustrate that the local sensitivity measures based on these metrics for the posterior distributions are more stable in the sense that they at least do not diverge as we obtain more data. We first present some notations and preliminarily results regarding the total variation distance which are required to introduce our new class of metrics.

We denote the total variation metric on probability distributions (Π, Ψ), defined over a common σ-algebra $\mathbb{C}$ on a parameter space Θ, as follows:

$$d(\Pi, \Psi) = \sup_{C \in \mathbb{C}} |\Pi(C) - \Psi(C)| \tag{2}$$

This metric can be also written in terms of the respective densities of π and ψ as follows

$$d(\Pi, \Psi) = \frac{1}{2} \int_{\theta \in \Theta} |\pi(\theta) - \psi(\theta)| d\theta \tag{3}$$

In addition to the above property, this metric is also invariant with respect to transformations $\eth$ in the following sense. If $\eth : \theta \to \theta'$, $\theta \in \Theta$ and $\theta' \in \Theta'$, is bijective and measurable and $(\Pi(\theta), \Psi(\theta))$ and $(\Pi(\theta'), \Psi(\theta'))$ are probability distributions defined on θ and $\theta' = \eth(\theta)$, then

$$d(\Pi(\theta), \Psi(\theta)) = d(\Pi(\theta'), \Psi(\theta'))$$

It can be also shown that for a fixed known family of sample distributions, the total variation distance between two predictive distributions is no larger than the distance between their prior distributions as discussed in ([5]). In other words,

$$d(\Pi(X), \Psi(X)) \leq d(\Pi(\theta), \Psi(\theta))$$

where $\Pi(X)$ and $\Psi(X)$ are probability distributions associated with the following density functions

$$\pi(x) = \int \pi(\theta) p(x \mid \theta) d\theta, \quad \text{and} \quad \psi(x) = \int \psi(\theta) p(x \mid \theta) d\theta$$

Despite all these nice properties, using the following example we verify the results reported in ([9]) that this distance cannot converge as $n \to \infty$.

Example 1 Suppose $X_1, X_2, ..., X_n$ is a random sample from a standard normal distribution, $N(\theta, 1)$. Let us define $S_i(n) = n^{-\frac{1}{2}} \sum_{j=n(i-1)+1}^{ni} X_j$. It can be easily concluded that for two different prior densities $p_j(\theta)$, $j = 1, 2$, and $\forall n > 0$, $p_j(\theta|X^{(n)} = x^{(n)}) = p_j(\theta|S_1(n) = s_1(n))$, where $x^{(n)} = \{x_1, x_2, ..., x_n\}$. We also have that

$$p_j(s_2(n)|x^{(n)}) = \int_{\theta\in\Theta} p_j(s_2(n)|\theta)p_j(\theta|x^{(n)})d\theta = \int_{\theta\in\Theta} p_j(s_2(n)|\theta)p_j(\theta|s_1(n))d\theta$$
$$= \int_{\theta'\in\Theta} p_j(s_2(n)|\theta')p_j(\theta'|s_1(n))d\theta'.$$

Since $p_j(s_2(n)|\theta') = p_j(s_2(1)|\theta) \sim N(\theta', 1)$, then $p_j(s_2(n)|x^{(n)}) = \int_{\theta'} p_j(s_2(1)|\theta')p_j(\theta'|s_1(n))d\theta'$, where $\theta' = n^{\frac{1}{2}}\theta$. Thus, in a sense, the problem of predictive densities does not appear to depend on the number of observations, n. In particular, $d(p_1(s_2(n)|x^{(n)}), p_2(s_2(n)|x^{(n)}))$ does not depend on n. It can then be concluded that for all $n > 0$,

$$d(p_1(\theta|x^{(n)}), p_2(\theta|x^{(n)})) \geq d(p_1(s_2(n)|x^{(n)}), p_2(s_2(n)|x^{(n)})) = d(p_1(x_2|x_1), p_2(x_2|x_1)).$$

This looks counter-intuitive, since we know whatever the prior for θ, in this circumstance, given $x^{(n)}$, $n^{-\frac{1}{2}}(\theta - \bar{x})$ tends to the standard normal density, the posterior densities will be close to one another, spiking near $\bar{x}$. On the other hand, we know that the variation metric is scale invariant, and we need to see the difference between the posterior densities appropriately magnified up onto the region to which θ converges. However, if we have some way of fixing the scale of the deviation, then this is not so. For example,

$$\begin{aligned}
d(p_1(x_{n+1}|x^{(n)}), p_2(x_{n+1}|x^{(n)})) &= \int |p(x_{n+1}|\theta)\{p_1(\theta|x^{(n)}) - p_2(\theta|x^{(n)})\}|d\theta \\
&\leq \int_{\theta\in B(\bar{x},\delta)} p(x_{n+1}|\theta)|p_1(\theta|x^{(n)}) - p_2(\theta|x^{(n)})|d\theta \\
&+ \int_{\theta\notin B(\bar{x},\delta)} p(x_{n+1}|\theta)(p_1(\theta|x^{(n)}) - p_2(\theta|x^{(n)}))d\theta \\
&\leq \sup_{\theta\in B(\bar{x},\delta)} \{p(x_{n+1}|\theta)\}\mu(B(\bar{x},\delta)) + \sup_{\theta\notin B(\bar{x},\delta)} \{p(x_{n+1}|\theta)\}2\eta(\delta) \\
&< (2\pi)^{-1}\{\mu(B(\bar{x},\delta)) + 2\eta(\delta)\} \to 0, \text{ as } n \to \infty,
\end{aligned}$$

where $B(\bar{x}, \delta)$ is an open ball with its centre at $\bar{x}$ and diameter δ, $\mu(B(\bar{x}, \delta))$ is its dominating measure, and $\eta(\delta) = \int_{\theta\notin B(\bar{x},\delta)} p_1(\theta|x^{(n)})d\theta = \int_{\theta\notin B(\bar{x},\delta)} p_2(\theta|x^{(n)})d\theta$.

Therefore, one step ahead, prediction of the next observation certainly converges. These predictions will be stable if prediction about θ are stable. So stability in terms of this (and in many other metrics) is consistent with ideas about Bayesian Sufficiency. Using posterior predictive distributions or making inference in terms of Bayesian predictive measures have been supported by several researchers ([3, 14]).

Before introducing the credible metric and study its asymptotic behaviour, we need to present some further notations and definitions.

Let $P|A$ and $Q|A$ denote respectively the conditional probability distributions associated with P and Q given an event $A \in \mathbb{C}$, $P(A) > 0$ and define

$$d_{A[P]}(P, Q) = d(P \mid A, Q \mid A), \tag{4}$$

where $\mathbb{C}$ is a common $\sigma-$ algebra on the parameter (or sample) space Θ. Note that this is a pseudometric (i.e. all the metric axioms hold other than $d_{A[p]}(P, Q) = 0 \Rightarrow P = Q$).

Call a set $\mathbb{A}$ of events P-conditioning, if $\Theta \subseteq \mathbb{A} \subseteq \mathbb{C}^+[P]$, where $\mathbb{C}^+[P] = \{C \in \mathbb{C} : P(C) > 0\}$. For any P-conditioning set $\mathbb{A}$ denote

$$d_{\mathbb{A}}(P, Q) = \sup\{d_{A[P]}(P, Q) : A[P] \in \mathbb{A}\}. \tag{5}$$

Finally denote by $\mathbb{P}(P)^+$ the set of all probability measures with the same support as P.

Lemma 2.1 *If $\mathbb{A}$ is P-conditioning then $d_{\mathbb{A}}(.,.)$ is a metric on $\mathbb{P}(P)^+$.*

Proof Let $P, Q, R \in \mathbb{P}(P)^+$. Since $d(.,.)$ is a metric on $\mathbb{P}(P)^+$, we can then conclude that $P \neq Q$ and $d_{\mathbb{A}}(P, Q) \geq d(P, Q) > 0$. Furthermore, since $d_{A[P]}(P, Q)$ is a pseudometric for all $A[P] \in \mathbb{A}$, we have both

$$d_{\mathbb{A}}(P, P) = \sup_{A[P] \in \mathbb{A}} \{d_{A[P]}(P, P)\} = 0, \quad \& \quad d_{\mathbb{A}}(P, Q) = d_{\mathbb{A}}(Q, P).$$

Finally, again since $d_{A[P]}(P,Q)$ is a pseudometric for all $A[P] \in \mathbb{A}$,

$$d_{\mathbb{A}}(P,Q) = \sup_{A[P]\in\mathbb{A}} \{d_{A[P]}(P,Q)\} \leq \sup_{A[P]\in\mathbb{A}} \{d_{A[P]}(P,R) + d_{A[P]}(R,Q)\}$$

$$\leq \sup_{A[P]\in\mathbb{A}} \{d_{A[P]}(P,R)\} + \sup_{A[P]\in\mathbb{A}} \{d_{A[P]}(R,Q)\} \leq d_{\mathbb{A}}(P,R) + d_{\mathbb{A}}(R,Q) \tag{6}$$

Note that this result does not rely on $d(.,.)$ being the variation metric. In particular, it works with the Hellinger metric as well.

To clarify the nature of this new class of metric, we make a few remarks in the lemmas below.

Lemma 2.2 *If P is discrete and $\mathbb{A}$ contains all two point sets $\{i,j\}$, then the $d_{\mathbb{A}}(P,Q)$ neighbourhoods of P are contained in DeRobertis density ratio spheres*

$$\wedge_s(p;\epsilon) = \{Q : \sup_{i,j} |\log(p_i) - \log(q_i) - \log(p_j) + \log(q_j)| \leq \epsilon\}, \tag{7}$$

Proof Suppose, without loss of generality that

$$\rho = \frac{p_i}{p_j} \geq \frac{q_i}{q_j}.$$

Then

$$d_{\{i,j\}}(P,Q) = \frac{p_i}{p_i+p_j} - \frac{q_i}{q_i+q_j} = \frac{p_iq_j - p_jq_i}{(p_i+p_j)(q_i+q_j)} = \frac{\rho}{1+\rho}\frac{c-1}{c+\rho},$$

where $c = \frac{p_iq_j}{p_jq_i} = \exp\{|\log p_i - \log q_i - \log p_j + \log q_j|\} \geq 1$.

Clearly $d_{\{i,j\}}(P,Q)$ is increasing in $c \geq 1$. Thus, for discrete variables, the topology defined by such a metric is at least as refined as topology defined by density ratio spheres.

Now assume that $\mathbb{A} = \mathbb{C}$. Note that, for any set $A \in \mathbb{C}$

$$\begin{aligned} 2d_A(P,Q) &= \int_A |\frac{p(\boldsymbol{\theta})}{P(A)} - \frac{q(\boldsymbol{\theta})}{Q(A)}|d\boldsymbol{\theta} = \frac{1}{P(A)}\int_A p(\boldsymbol{\theta})|\frac{P(A)}{Q(A)}\exp\{t(\boldsymbol{\theta})\} - 1|d\boldsymbol{\theta} \\ &\leq \sup_{\boldsymbol{\theta}\in A} |\frac{P(A)}{Q(A)}\exp\{t(\boldsymbol{\theta})\} - 1|, \end{aligned}$$

where $t(\boldsymbol{\theta}) = |\log p(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta})|$.

Now assume $|t(\boldsymbol{\theta})| \leq \tau$, and note that

$$e^{-\tau} \leq \frac{Q(A)}{P(A)} = \frac{\int_A q(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_A p(\boldsymbol{\theta})d\boldsymbol{\theta}} = \frac{\int_A \exp\{t(\boldsymbol{\theta})\}p(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_A p(\boldsymbol{\theta})d\boldsymbol{\theta}} \leq e^{\tau},$$

which implies

$$2d_A(P,Q) \leq \exp\{2\tau\} - 1.$$

Thus we have proved the following lemma.

Lemma 2.3 *Suppose probability measures P and Q have respective densities p and q with respect to the same dominating measures, and are strictly positive on their shared support. Then if, for all $\epsilon > 0$, there exist (small) values of $\tau(\epsilon) > 0$, and if $\boldsymbol{\theta} \in A$, $A \in \mathbb{C}$,*

$$|\log p(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta})| \leq \tau \quad \textit{then} \quad d_A(P,Q) \leq \epsilon.$$

It is clear therefore that although these metrics are much fiercer than the variation metric, the open sets around P are rich, provided that $\mathbb{A}$ does not contain sets which are too improbable. The following lemma present a partial converse of this result.

Lemma 2.4 *Suppose probability measures P and Q ($P \neq Q$) have respective continuous densities p and q with respect to the same dominating measure, non-zero on their shared support. For all $\tau > 0$, write*

$$A_U(\tau) = \{\boldsymbol{\theta} : \log p(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta}) \geq \tau\}, \tag{8}$$

$$A_L(\tau) = \{\boldsymbol{\theta} : \log p(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta}) \leq -\tau\}, \tag{9}$$

$$A_M(\tau) = \{\boldsymbol{\theta} : |\log p(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta})| < \tau\}. \tag{10}$$

Suppose there exists a value of $\eta > 0$ such that, for all $\tau < \eta$,

$$\min\{P(A_U(\tau)), P(A_L(\tau))\} > 0.$$

Then, for all $\epsilon > 0$, there exists a value $\tau > 0$ and a set $C(\tau) \subset \Theta$, $P(C) > 0$,

$$d_C(P,Q) \geq (1 - e^{-\tau}).$$

Proof First note that

$$P(A_U(\tau)) - Q(A_U(\tau)) = \int_{A_U(\tau)} (p(\theta) - q(\boldsymbol{\theta}))d\boldsymbol{\theta} \geq (1 - e^{-\tau})P(A_U), \quad \text{and}$$
$$Q(A_L(\tau)) - P(A_L(\tau)) = \int_{A_L(\tau)} (q(\boldsymbol{\theta}) - p(\boldsymbol{\theta}))d\boldsymbol{\theta} \geq (e^{\tau} - 1)P(A_L)$$

Now if $\mu(A_U(\tau)) = 0$, then

$$\begin{aligned} d(P, Q) &= \frac{1}{2}\int_{\boldsymbol{\theta}\in\Omega} |p(\boldsymbol{\theta}) - q(\boldsymbol{\theta})|d\boldsymbol{\theta} \\ &\leq \frac{1}{2}\int_{\boldsymbol{\theta}\in\Omega} (q(\boldsymbol{\theta}) - p(\boldsymbol{\theta}))d\boldsymbol{\theta} + \int_{A_M} |p(\boldsymbol{\theta}) - q(\boldsymbol{\theta})|d\boldsymbol{\theta} \leq \tau. \end{aligned}$$

Similarly, if $\mu(A_L(\tau)) = 0$

$$d(P, Q) \leq \frac{1}{2}\int_{\boldsymbol{\theta}\in\Omega} (p(\boldsymbol{\theta}) - q(\boldsymbol{\theta}))d\boldsymbol{\theta} + \int_{A_M} |p(\boldsymbol{\theta}) - q(\boldsymbol{\theta})|d\boldsymbol{\theta} \leq \tau.$$

Therefore, f$\forall\tau$, $\min\{\mu(A_U(\tau)), \mu(A_L(\tau))\} = 0$, then $P = Q$ is in contradiction to our hypothesis. Hence, provided τ is small enough, say $\delta < \eta$, we have $\min\{\mu(A_U(\tau)), \mu(A_L(\tau))\} > 0$, which, since p is strictly positive in turn implies $\min\{P(A_U(\tau)), P(A_L(\tau))\} > 0$.

It follows from the above that both A_U and A_L are such that $P(A_U(\tau)) - Q(A_U(\tau)) > 0$ and $P(A_L(\tau)) - Q(A_L(\tau)) > 0$.

As a result, when $P(A_U) - Q(A_U) \geq Q(A_L) - P(A_L)$, one can then choose any subset B_U of A_U such that $P(B_U) - Q(B_U) = Q(A_L) - P(A_L)$. This is clearly possible if P and Q are continuous. On the other hand if $P(A_L) - Q(A_L) \geq Q(A_U) - P(A_U)$, then one can choose any subset B_L of A_L such that $P(B_L) - Q(B_L) = Q(A_U) - P(A_U)$.

Therefore, under the conditions given above, we can construct subsets B_U, B_L such that

$$B_U = \{\boldsymbol{\theta} : \log p(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta}) \geq \tau\} \quad \text{and} \quad B_L = \{\boldsymbol{\theta} : \log p(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta}) \leq -\tau\},$$

where $P(C) = Q(C)$, and $C = B_U \cup B_L$. We can then write

$$\begin{aligned} 2d_C(P, Q) &= \int_C |\frac{p(\boldsymbol{\theta})}{P(C)} - \frac{q(\boldsymbol{\theta})}{Q(C)}|d\boldsymbol{\theta} = \frac{1}{P(C)}\{\int_{B_U} (p(\boldsymbol{\theta}) - q(\boldsymbol{\theta}))d\boldsymbol{\theta} - \int_{B_L} (p(\boldsymbol{\theta}) - q(\boldsymbol{\theta}))d\boldsymbol{\theta}\} \\ &= \frac{1}{Q(C)}\{Q(B_U)(e^{\tau} - 1) + Q(B_L)(1 - e^{-\tau})\} \geq (1 - e^{-\tau}) \end{aligned}$$

which implies $d_C(P,Q) \geq (1-e^{-\tau})$ as required.

It should be noted the metric developed above when $\mathbb{A} = \mathbb{C}$ is not really new, and it essentially demands that the log-densities of two distributions are close everywhere. Furthermore, this is not that practical, because it demands proportionate closeness in the tails of the density and it would be unrealistic to expect such levels of subjective certainty on the sets with very small probability. In the following lemma, we demonstrate that the sets that have large prior probability do not affect the topology of $d_{\mathbb{A}}(P,Q)$.

Lemma 2.5 *If $P(A) > c > 0$, then for all $\epsilon > 0$ there exits a δ such that if $d(P(A),Q(A)) < \delta$, then $d_A(P(A),Q(A)) < \epsilon$.*

Proof

$$\begin{aligned}
2d_A(P(A),Q(A)) &= \int_A |\frac{p(\theta)}{P(A)} - \frac{q(\boldsymbol{\theta})}{Q(A)}|d\boldsymbol{\theta} \\
&\leq \frac{1}{P(A)}\int_A |p(\boldsymbol{\theta}) - q(\boldsymbol{\theta})|d\boldsymbol{\theta} + |\frac{1}{P(A)} - \frac{1}{Q(A)}|\int_A |q(\boldsymbol{\theta})|d\boldsymbol{\theta} \\
&\leq \frac{1}{P(A)}\int_A |p(\boldsymbol{\theta}) - q(\boldsymbol{\theta})|d\boldsymbol{\theta} + \frac{|P(A)-Q(A)|}{P(A)Q(A)} \\
&\leq \frac{2\delta}{c} + \frac{\delta}{c(c-\delta)} = \frac{\delta(2c+1-\delta)}{c(c-\delta)},
\end{aligned}$$

which for fixed values of c is continuous at zero and equal to zero when $\delta = 0$.

It can be concluded that, when the limits are considered, we will gain nothing over the variation metric by including the sets with higher than a threshold probability. It is the distances associated with small sets A which might contribute to something new. However, when we learn through Bayes rule, typically, as our sample increases in size, the posterior densities associated with different priors will tend to concentrate around the same small open balls. It follows that there may be considerable gain by restricting our attention to the whole space together with small open balls. This provokes the following definition.

Definition 1 Call $d_{\mathbb{A}}(P,Q) = d^{\triangle|C}(P,Q)$ the (δ, C)-*credibility metric* if

$$\mathbb{A} = \{\Theta\} \cup \bigcup \{B(\boldsymbol{\theta}_0;\delta) : \boldsymbol{\theta}_0 \in C \subseteq \Theta, 0 < \delta \leq \triangle\}, \tag{11}$$

where $B(\boldsymbol{\theta}_0;\delta)$ is a Euclidean open ball with center at $\boldsymbol{\theta}_0$ and diameter δ, and $d(.,.)$ is the total variation metric.

By writing $d^{\triangle}(P,Q) = d^{\triangle|\Theta}(P,Q)$, we shall see that, provided that the space of densities we consider is smooth enough, this metric gives the sort of limiting results we require. Furthermore, the type of smoothness conditions we need to impose, seems relatively benign and plausible from a subject perspective.

Explicitly, we can write $d_{B(\boldsymbol{\theta}_0;\delta)}(P,Q) = d(P \mid B(\boldsymbol{\theta}_0;\delta), Q \mid B(\boldsymbol{\theta}_0;\delta))$. We show that, within a set $\mathbb{A}$ a sufficiently "small" ball $B(\boldsymbol{\theta}_0,\delta)$ is not active in $d_{\mathbb{A}}(P,Q)$, provided the log-densities of P and Q are defined and continuous at $\boldsymbol{\theta}_0$. So, P and Q can be very different in variation metric and still be closed under this conditional metric. All we require is that both are sufficiently smooth.

Lemma 2.6 *Suppose probability measures P and Q have respective densities p and q, and for all $\omega > 0$, there exists a $\delta(\boldsymbol{\theta}_0;\omega,p) > 0$ such that, for all $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0;\delta)$,*

$$|\log p(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}_0)| < \omega$$

and, for all $\omega > 0$, there exists a $\delta(\boldsymbol{\theta}_0;\omega,q) > 0$ such that, for all $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0;\delta)$,

$$|\log q(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta}_0)| < \omega,$$

then

$$\frac{1}{[\mu(B(\boldsymbol{\theta}_0;\delta))]} \int_{B(\boldsymbol{\theta}_0;\delta)} |\frac{p(\boldsymbol{\theta})}{p(\boldsymbol{\theta}_0)} - 1| d\boldsymbol{\theta} < (e^{\omega} - 1),$$
$$\frac{1}{[\mu(B(\boldsymbol{\theta}_0;\delta))]} \int_{B(\boldsymbol{\theta}_0;\delta)} |\frac{q(\boldsymbol{\theta})}{q(\boldsymbol{\theta}_0)} - 1| d\boldsymbol{\theta} < (e^{\omega} - 1),$$
$$e^{-\omega} < \frac{p(\boldsymbol{\theta}_0)\mu(B(\boldsymbol{\theta}_0;\delta))}{P(B(\boldsymbol{\theta}_0;\delta))} < e^{\omega}, \quad e^{-\omega} < \frac{q(\boldsymbol{\theta}_0)\mu(B(\boldsymbol{\theta}_0;\delta))}{Q(B(\boldsymbol{\theta}_0;\delta))} < e^{\omega},$$

where $\mu(B(\boldsymbol{\theta}_0;\delta))$ denote the dominating measure.

Proof . The first assertion follows, since for all $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0; \delta)$, it can be easily shown that

$$|\log p(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}_0)| < \omega \Leftrightarrow |\frac{p(\boldsymbol{\theta})}{p(\boldsymbol{\theta}_0)} - 1| < e^{\omega} - 1.$$

To prove the second assertion, note that

$$\left(\frac{P(B(\boldsymbol{\theta}_0; \delta))}{p(\boldsymbol{\theta}_0)\mu(B(\boldsymbol{\theta}_0; \delta))} - 1\right) = \frac{\int_B (\frac{p(\boldsymbol{\theta})}{p(\boldsymbol{\theta}_0)} - 1) d\boldsymbol{\theta}}{\mu(B(\boldsymbol{\theta}_0; \delta))},$$

and substituting the first result gives $e^{-\omega} - 1 < \left(\frac{P(B(\boldsymbol{\theta}_0;\delta))}{p(\boldsymbol{\theta}_0)\mu(B(\boldsymbol{\theta}_0;\delta))} - 1\right) < e^{\omega} - 1$, which rearranges to the given expression. The last two inequalities hold, simply by substituting q for p.

One immediate consequence of these inequalities is that they hold if and only if the corresponding conditions hold for the posterior distribution of a shared sampling model, and the log-likelihood is smooth and continuous at $\boldsymbol{\theta}_0$. Therefore,

$$\begin{aligned}
|\log p(\boldsymbol{\theta} \mid \mathbf{x}) - \log p(\boldsymbol{\theta}_0 \mid \mathbf{x})| &= |\log p(\boldsymbol{\theta}) + \log p(\mathbf{x} \mid \boldsymbol{\theta}) + \log \int p(\boldsymbol{\theta}) p(\mathbf{x} \mid \boldsymbol{\theta}) d\boldsymbol{\theta} \\
&\quad - \log \int p(\boldsymbol{\theta}) p(\mathbf{x} \mid \boldsymbol{\theta}) d\boldsymbol{\theta} - (\log p(\boldsymbol{\theta}_0) + \log p(\mathbf{x} \mid \boldsymbol{\theta}_0))| \\
&\leq |\log p(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}_0)| + |\log p(\mathbf{x} \mid \boldsymbol{\theta}) - \log p(\mathbf{x} \mid \boldsymbol{\theta}_0)|,
\end{aligned}$$

so that, for all $\omega' > 0$, provided δ is chosen small enough, for all $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0; \delta)$

$$|\log p(\mathbf{x} \mid \boldsymbol{\theta}) - \log p(\mathbf{x} \mid \boldsymbol{\theta}_0)| < \omega',$$

and we obtain analogous inequalities for the posterior densities in $B(\boldsymbol{\theta}_0; \delta)$. It means that, with a continuity condition on the likelihood, prior closeness with respect to this metric guarantees posterior closeness. We use this fact in the next section.

Theorem 2.7 *For all $\epsilon > 0$, if P and Q satisfy the continuity conditions above, there exists values of $\eta > 0$, such that if $\delta < \eta$ then $d_{B(\boldsymbol{\theta}_0;\delta)}(P, Q) < \epsilon$.*

Proof

$$2d_{B(\theta_0;\delta)}(P,Q) = 2d(P \mid B(\boldsymbol{\theta}_0;\delta), Q \mid B(\boldsymbol{\theta}_0;\delta))$$
$$= \int_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_0;\delta)} |\frac{p(\boldsymbol{\theta})}{P(B(\boldsymbol{\theta}_0;\delta))} - \frac{q(\boldsymbol{\theta})}{Q(B(\boldsymbol{\theta}_0;\delta))}|d\boldsymbol{\theta} \le A(\delta) + B(\delta) + C(\delta),$$

where, whenever $\log p$, $\log q$ are continuous,

$$A(\delta) = \int_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_0;\delta)} |\frac{p(\boldsymbol{\theta})}{P(B(\boldsymbol{\theta}_0;\delta))} - \frac{p(\boldsymbol{\theta}_0)}{P(B(\boldsymbol{\theta}_0;\delta))}|d\boldsymbol{\theta}$$
$$= \frac{p(\boldsymbol{\theta}_0)\mu(B(\boldsymbol{\theta}_0;\delta))}{P(B(\boldsymbol{\theta}_0;\delta))}\{[\frac{1}{\mu(B(\boldsymbol{\theta}_0;\delta))}]\int_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_0;\delta)} |\frac{p(\boldsymbol{\theta})}{p(\boldsymbol{\theta}_0)} - 1|d\boldsymbol{\theta}\},$$

which by the inequalities above $A(\delta) < e^{\omega(\boldsymbol{\theta}_0;\delta,p)}(e^{\omega(\boldsymbol{\theta}_0;\delta,p)} - 1)$. Similarly

$$C(\delta) = \int_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_0;\delta)} |\frac{q(\boldsymbol{\theta}_0)}{Q(B(\boldsymbol{\theta}_0;\delta))} - \frac{q(\boldsymbol{\theta})}{Q(B(\boldsymbol{\theta}_0;\delta))}|d\boldsymbol{\theta} < e^{\omega(\boldsymbol{\theta}_0;\delta,q)}(e^{\omega(\boldsymbol{\theta}_0;\delta,q)} - 1)$$

and

$$B(\delta) = \mu(B(\boldsymbol{\theta}_0;\delta))|\frac{p(\boldsymbol{\theta}_0)}{P(B(\boldsymbol{\theta}_0;\delta))} - \frac{q(\boldsymbol{\theta}_0)}{Q(B(\boldsymbol{\theta}_0;\delta))}|$$
$$\le |\frac{\mu(B(\boldsymbol{\theta}_0;\delta))p(\boldsymbol{\theta}_0)}{P(B(\boldsymbol{\theta}_0;\delta))} - 1| + |\frac{\mu(B(\boldsymbol{\theta}_0;\delta))q(\boldsymbol{\theta}_0)}{Q(B(\boldsymbol{\theta}_0;\delta))} - 1|,$$

which by the inequalities above, $B(\delta) \le (e^{\omega(\boldsymbol{\theta}_0;\delta,p)} - 1) + (e^{\omega(\boldsymbol{\theta}_0;\delta,q)} - 1)$. Thus, for a given $\boldsymbol{\theta}_0$, and P and Q, for all $\epsilon > 0$, there is a value of δ such that

$$d_{B(\boldsymbol{\theta}_0;\delta)}(P,Q) < \frac{1}{2}\{e^{\omega(\boldsymbol{\theta}_0;\delta,p)}(e^{\omega(\boldsymbol{\theta}_0;\delta,p)} - 1) + e^{\omega(\boldsymbol{\theta}_0;\delta,q)}(e^{\omega(\boldsymbol{\theta}_0;\delta,q)} - 1) + (e^{\omega(\boldsymbol{\theta}_0;\delta,p)} - 1)$$
$$+ (e^{\omega(\boldsymbol{\theta}_0;\delta,q)} - 1)\} = \frac{1}{2}\{(e^{2\omega(\boldsymbol{\theta}_0;\delta,p)} - 1) + (e^{2\omega(\boldsymbol{\theta}_0;\delta,q)} - 1)\} = \epsilon(\boldsymbol{\theta}_0, P, Q, \delta),$$

as required.

Corollary 2.8 *Suppose that P has a differentiable log density with derivative $D\log p(\boldsymbol{\theta})$, bounded by M for all $\boldsymbol{\theta}$, i.e. $|D\log p(\boldsymbol{\theta})| \le M$. Then, for all distributions Q with differentiable log-densities bounded by M and for all $\epsilon > 0$, there exists a value of η such that for all sets $B(\boldsymbol{\theta}_0;\delta)$, whenever $\delta < \eta$, $d_{B(\boldsymbol{\theta}_0;\delta)}(P,Q) < \epsilon$.*

Proof . This follows immediately from the lemma above, since if $D \log p(\theta)$, $D \log q(\boldsymbol{\theta})$ are bounded, they are then automatically uniformly continuous in $\boldsymbol{\theta}_0$.

It should be noticed that according to this corollary, we can write

$$d_{B(\boldsymbol{\theta}_0;\delta)}(P, Q) = d(P \mid B(\boldsymbol{\theta}_0;\delta), Q \mid B(\boldsymbol{\theta}_0;\delta)).$$

As outlinedd before, $B(\boldsymbol{\theta}_0;\delta)$ is a sufficiently small ball in $\mathbb{A}$ which is not active in (δ, C)-credibility metric, $d_{\mathbb{A}}(P, Q) = d^{\Delta|C}(P, Q)$, where $\mathbb{A}$ is defined in (11), and δ is defined in Corollary 1. This is very useful and particularly implies that two strictly positive unimodal bounded prior densities with sub-exponential tails will look locally similar in the sense of this metric. In Section 3, we will use this to relate the metric above to well-known results about the robust families of priors. We could also link this to the Gustafson's ideas to restrict the class of prior distributions into a parameterised class of priors (as also closely discussed in [5]).

In a practical setting, it would be challenging to assert the condition of this corollary, which makes strong statements about the tail behaviour of a prior density. Fortunately, the required uniform continuity for the convergence of our metric can be obtained, provided closeness for sets $B(\boldsymbol{\theta}_0;\delta)$ for which $p(\boldsymbol{\theta}_0) \geq c > 0$.

Corollary 2.9 *Suppose that P has a continuous bounded density p at all points $\boldsymbol{\theta}_0$, such that $p(\boldsymbol{\theta}_0) \geq c_p > 0$, and all distributions Q have a continuous bounded density q at all points $\boldsymbol{\theta}_0$, such that $q(\boldsymbol{\theta}_0) \geq c > 0$. Suppose the sets $D_p = \{\boldsymbol{\theta}_0 : p(\boldsymbol{\theta}_0) \geq c_p > 0\}$ and $D_q = \{\boldsymbol{\theta}_0 : q(\boldsymbol{\theta}_0) \geq c_q > 0\}$ are compact. Then, for all $\epsilon > 0$ there exists a value of η such that for all sets $B(\boldsymbol{\theta}_0 : \delta)$, $\boldsymbol{\theta}_0 \in D_p \cup D_q$, whenever $\delta < \eta$, $d_{B(\boldsymbol{\theta}_0;\delta)}(P, Q) < \epsilon$.*

Proof . The required uniform continuity is immediate from the compactness of the sets and the continuity and boundedness of p and q. Thus, for small open sets in a credibility set, with sufficient smoothness assumptions, we can expect all associated variation distances to be small a priori. Therefore, we may be able to assert densities which are close and do not wobble too much (see [5] for more details).

3 Sensitivity analysis using η-Credibility metrics

The usefulness of the credibility metrics arises from the following plausible observation.

Theorem 3.1 *Suppose P^* and Q^* are the posterior distributions associated with P and Q respectively after we observe that $\boldsymbol{\theta} \in B \in \mathbb{A}$. Then, if $\mathbb{A}$ is closed under intersection,*

$$d_{\mathbb{A}}(P^*, Q^*) \leq d_{\mathbb{A}}(P, Q).$$

Proof Since $\mathbb{A}$ is closed under intersection with B, $\mathbb{A} \mid B = \{A' \in \mathbb{A} : A' = A \cap B, A \in \mathbb{A}\} \subseteq \mathbb{A}$, and because $d_A(P^*, Q^*) = d_A(P \mid \{\boldsymbol{\theta} \in B\}, Q \mid \{\boldsymbol{\theta} \in B\}) = d_{A \cap B}(P, Q)$, the result is now immediate by definition.

This means that under an extended variation metric, learning about $\boldsymbol{\theta}$ directly cannot increase neighbourhoods: in particular the Fréchet derivative (used as the local sensitivity measure) always reduces as zero-one information about $\boldsymbol{\theta}$ arrives. This is in strong contrast to the use of the ordinary variation metric for which this is untrue in general [9]. In particular, if our experiment indicates that $\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0; \delta)$ and $\delta \to 0$ then, under the conditions of the corollaries 1 and 2, the Fréchet derivative does not diverge, and is bounded.

There are more problems here when we learn through a sample distribution. In 2004, Daneshkhah [5] reported that prior small credibility closeness gives rise to posterior credibility closeness with a likelihood continuous at all the relevant $\boldsymbol{\theta}_0$.

We next show that the variation distance between posterior cannot explode, if we use closed priors that equals with smooth priors.

Theorem 3.2 *Suppose for all $\gamma > 0$ there exists a value $\triangle$ such that, for all $\delta < \triangle$ and Q such that $d_{\mathbb{A}}(P, Q) < \eta$, $Q(B^c) < \gamma$, where $B = \bigcup_{i=1}^m B(\theta_i^0; \delta)$ and for all $\omega > 0$, and all $\{i : 1 \leq i \leq m\}$,*

$$|\log p(\theta_i) - \log p(\theta_i^0)| < \omega, \quad \textit{and} \quad |\log p(\boldsymbol{x} \mid \theta_i) - \log p(\boldsymbol{x} \mid \theta_i^0)| < \omega;$$

then for all $\epsilon > 0$, $d(P \mid \boldsymbol{x}, Q \mid \boldsymbol{x}) < \epsilon$.

Proof

$$
\begin{aligned}
d(P \mid \mathbf{x}, Q \mid \mathbf{x}) &= \int |p(\boldsymbol{\theta} \mid \mathbf{x}) - q(\boldsymbol{\theta} \mid \mathbf{x})| d\boldsymbol{\theta} \\
&\leq \sum_{i=1}^{m} \int_{B(\theta_i^0;\delta)} |p(\theta_i \mid \mathbf{x}) - q(\theta_i \mid \mathbf{x})| d\theta_i + \int_{B^c} |p(\boldsymbol{\theta} \mid \mathbf{x}) - q(\boldsymbol{\theta} \mid \mathbf{x})| d\boldsymbol{\theta} \\
&\leq \Sigma_{i=1}^{m} (I_i^1(\delta) + I_i^2(\delta) + I_i^3(\delta)) + \int_{B^c} p(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta} + \int_{B^c} q(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta} \\
&\leq \Sigma_{i=1}^{m} I_i^1(\delta) + \Sigma_{i=1}^{m} I_i^2(\delta) + \Sigma_{i=1}^{m} I_i^3(\delta) + 2\gamma,
\end{aligned}
$$

where

$$
\begin{aligned}
I_i^1(\delta) &= \int_{\theta_i \in B(\theta_i^0;\delta)} \left| \frac{p(\theta_i^0 \mid \mathbf{x})}{p(\theta_i \mid \mathbf{x})} - 1 \right| p(\theta_i \mid \mathbf{x}) d\theta_i \\
&\leq P(\theta_i \in B(\theta_i^0;\delta) \mid \mathbf{x})(\exp(2\omega) - 1).
\end{aligned}
$$

Moreover

$$
\left| \frac{p(\theta_i^0 \mid \mathbf{x})}{p(\theta_i \mid \mathbf{x})} - 1 \right| = |\exp[(\log p(\mathbf{x} \mid \theta_i^0) - \log p(\mathbf{x} \mid \theta_i)) - (\log p(\theta_i) - \log p(\theta_i^0))] - 1| \leq [e^{2\omega} - 1].
$$

Similarly, $I_i^3 \leq Q(\theta_i \in B(\theta_i^0;\delta))(\exp(2\omega) - 1)$. Finally

$$
\begin{aligned}
I_i^2(\delta) &= \int_{B(\theta_i^0;\delta)} |p(\theta_i \mid \mathbf{x}) - q(\theta_i \mid \mathbf{x})| d\theta_i = \mu(B(\theta_i^0;\delta)) |p(\theta_i^0 \mid \mathbf{x}) - q(\theta_i^0 \mid \mathbf{x})| \\
&= \frac{\mu(B(\theta_i^0;\delta)) p(\mathbf{x} \mid \theta_i^0) |p(\theta_i^0) M_i(q) - q(\theta_i^0) M_i(p)|}{M_i(p) M_i(q)}, \quad \text{where}
\end{aligned}
$$

$$
\begin{aligned}
M_i(p) &= \int_{B(\theta_i^0;\delta)} p(\mathbf{x} \mid \theta_i) p(\theta_i) d\theta_i, \quad M_i(q) = \int_{B(\theta_i^0;\delta)} p(\mathbf{x} \mid \theta_i) q(\theta_i) d\theta_i, \\
I_i^2(\delta) &= S(\theta_i^0, \delta, \mathbf{x}) \times T(\theta_i^0, \delta, \mathbf{x}), \quad S(\theta_i^0, \delta, \mathbf{x}) = \frac{p(\theta_i^0) \mu(B(\theta_i^0;\delta)) p(\mathbf{x} \mid \theta_i^0)}{\int_{B(\theta_i^0;\delta)} p(\mathbf{x} \mid \theta_i) p(\theta_i) d\theta_i}, \\
T(\theta_i^0, \delta, \mathbf{x}) &= \frac{\int_{B(\theta_i^0;\delta)} p(\mathbf{x} \mid \theta_i) q(\theta_i) |1 - \frac{p(\theta_i)}{p(\theta_i^0)} \frac{q(\theta_i^0)}{q(\theta_i)}| d\theta_i}{\int_{B(\theta_i^0;\delta)} p(\mathbf{x} \mid \theta_i) q(\theta_i) d\theta_i} \leq \sup_{\theta_i \in B(\theta_i^0;\delta)} \{|1 - \frac{p(\theta_i)}{p(\theta_i^0)} \frac{q(\theta_i^0)}{q(\theta_i)}|\} \\
&= \sup_{\theta_i | B(\theta_i^0;\delta)} \{|1 - \exp\{(\log p(\theta_i) - \log p(\theta_i^0)) - (\log q(\theta_i) - \log q(\theta_i^0))|\} \\
&\leq (e^{2\omega} - 1)
\end{aligned}
$$

Now, note that by hypothesis

$$\begin{aligned} S^{-1}(\theta_i^0, \delta, \mathbf{x}) &= \frac{\int_{B(\theta_i^0;\delta)} p(\mathbf{x} \mid \theta_i) p(\theta_i) d\theta_i}{p(\theta_i^0)\mu(B(\theta_i^0;\delta))p(\mathbf{x} \mid \theta_i^0)} \\ &= \frac{\int_{B(\theta_i^0;\delta)} \exp\{[\log p(x \mid \theta_i) - \log p(x \mid \theta_i^0)] - [\log p(\theta_i) - \log p(\theta_i^0)]\} d\theta_i}{\mu(B(\theta_i^0;\delta))} \\ &\geq \exp\{-2\omega\} \end{aligned}$$

Thus, $I_i^2(\delta) \leq e^{2\omega}\{e^{2\omega} - 1\}$. As a result,

$$d(P \mid \mathbf{x}, Q \mid \mathbf{x}) \leq (e^{2\omega} - 1)\{\Sigma_{i=1}^m [P(\theta_i \in B(\theta_i^0; \delta) \mid \mathbf{x}) + Q(\theta_i \in B(\theta_i^0; \delta) \mid \mathbf{x}) + e^{2\omega}]\} + 2\gamma$$

$$\leq (e^{2\omega} - 1)\{2m + me^{2\omega}\} + 2\gamma = \epsilon.$$

By hypothesis, the function on the right hand side of the inequality above can be made as small as we like by choosing $\triangle$ small enough when required.

Therefore, contrary to the assertion that it is necessary to restrict our class of prior distributions into a parametrised family of distributions, we can work with a general class of priors here. It is just needed to work with an appropriate extended variation metric presented above.

Example 2. Let the prior densities $p(\theta)$ and $q(\theta)$ be *Beta* distributions with the following density functions $p(\theta \mid \alpha) \propto \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1}$, $q(\theta \mid \beta) \propto \theta^{\beta_1 - 1}(1-\theta)^{\beta_2 - 1}$, and the corresponding posterior distributions for a sample drawn from a *Binomial* distribution with size n is given by $p_n(\theta \mid \alpha, x) \propto \theta^{(\alpha_1 + x) - 1}(1-\theta)^{(n+\alpha_2 - x)-1}$, $q_n(\theta \mid \beta, x) \propto \theta^{(\beta_1 + x)-1}(1-\theta)^{(n+\beta_2 - x)-1}$, where x is the number of successes observed in the sample. The posterior mean and and variance of the *Beta* distribution, $p_n(\theta \mid \alpha, x)$ are respectively given by

$$\theta_0 = \frac{(\alpha_1 + x)}{(\alpha_1 + \alpha_2) + n}, \quad \sigma_n^2 = \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2 + n)^2(\alpha_1 + \alpha_2 + n + 1)}$$

As discussed above, the conventional local sensitivity measures, as introduced in [1, 9, 11], do not converge for the large sample size. For example, the local sensitivity measure, defined in Eq. (1), under mild regularity conditions, increases at rate $n^{\frac{k}{2}}$ for many classes of prior distributions, where k is dimension of the parameter space. In this example, we illustrate that the asymptotic behaviour of the local sensitivity measure, developed in this paper using the credible metric is more promising, and will converge for the large sample size.

It is reasonable to consider δ as a function of σ_n^2, such as $\delta = (\sigma_n^2)^{0.4}$ or $\delta = (\sigma_n^2)^{0.45}$. Therefore,

$$\begin{aligned} I^1(\delta) &= P_{\theta_0|x}\left(\exp\{(\alpha_1+x-1)\log(\frac{\theta_0}{\theta}) + (n+\alpha_2-x-1)\log(\frac{1-\theta_0}{1-\theta})\}-1\right) \leq \\ P_{\theta_0|x} &\times \left(\exp\{(\alpha_1+x-1)\log(\frac{\theta_0}{\theta_0-\delta}) + (n+\alpha_2-x-1)\log(\frac{1-\theta_0}{1-\theta_0-\delta})\}-1\right) \end{aligned}$$

where $P_{\theta_0|x} = P(\theta \in (\theta_0-\delta, \theta_0+\delta) \mid x)$. Similarly

$$\begin{aligned} I^3(\delta) &= Q_{\theta_0|x}\left(\exp\{(\alpha_1+x-1)\log(\frac{\theta_0}{\theta}) + (n+\alpha_2-x-1)\log(\frac{1-\theta_0}{1-\theta})\}-1\right) \leq \\ Q_{\theta_0|x} &\times \left(\exp\{(\alpha_1+x-1)\log(\frac{\theta_0}{\theta_0-\delta}) + (n+\alpha_2-x-1)\log(\frac{1-\theta_0}{1-\theta_0-\delta})\}-1\right), \end{aligned}$$

where $Q_{\theta_0|x} = Q(\theta \in (\theta_0-\delta, \theta_0+\delta) \mid x)$, and

$$\begin{aligned} I^2(\delta) &\leq \left(\exp\{(\alpha_1+x-1)\log(\frac{\theta_0}{\theta}) + (n+\alpha_2-x-1)\log(\frac{1-\theta_0}{1-\theta})\}\right) \\ &\times \left(\exp\{(\alpha_1+x-1)\log(\frac{\theta_0}{\theta}) + (n+\alpha_2-x-1)\log(\frac{1-\theta_0}{1-\theta})\}-1\right). \end{aligned}$$

Using the Chebychev's inequality, we can show that $P(|\theta-\theta_0| > \delta) \leq (\frac{\sigma_n^2}{\delta^2})$, which for $\delta = (\sigma_n^2)^{0.45}$, it becomes as

$$P(|\theta-\theta_0| > \delta) \leq (\sigma_n^2)^{0.1}.$$

Therefore, as n increases, $d(P \mid x, Q \mid x)$ tends to zero. Figure 1 illustrates the distance bounds associated with $p(\theta) = Beta(4,6)$, $q(\theta) = Beta(6,8)$, $x = 7$ and for $\delta = (\sigma_n^2)^{0.4}$ and $\delta = (\sigma_n^2)^{0.45}$.

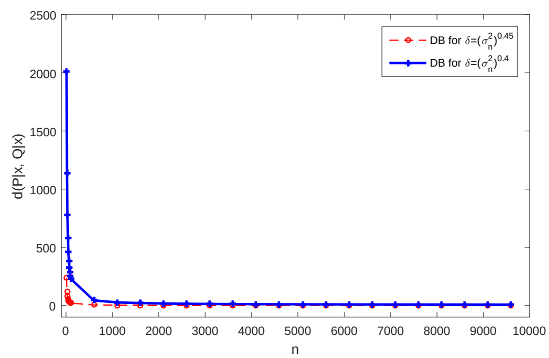

Figure 1: The distance bounds associated with $p(\theta) = Beta(4, 6)$, $q(\theta) = Beta(6, 8)$, and $\delta = (\sigma_n^2)^{0.4}$ and $\delta = (\sigma_n^2)^{0.45}$.

4 Credible Metrics Between Posterior predictive Distributions

Theorem 3 can be adapted for posterior predictive distributions. However, working with these distributions is quite difficult due to the complex computation, but by using them, we can avoid of the priors with unstable behaviours (the ones with too much wobble). We present similar results as given in the previous section for posterior predictive distributions.

First, we should show that $d(p(\mathbf{z} \mid \mathbf{x}), q(\mathbf{z} \mid \mathbf{x}))$ is a lower bound for $d(p(\boldsymbol{\theta} \mid \mathbf{x}), q(\boldsymbol{\theta} \mid \mathbf{x}))$. That means, $d(p(\mathbf{z} \mid \mathbf{x}), q(\mathbf{z} \mid \mathbf{x})) \leq d(p(\boldsymbol{\theta} \mid \mathbf{x}), q(\boldsymbol{\theta} \mid \mathbf{x}))$, where $p(\mathbf{z} \mid \mathbf{x}) = \int_{\boldsymbol{\theta}} p(\mathbf{z} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{x}) d\boldsymbol{\theta}$.

For this purpose, we use the total variation distance as follows,

$$
\begin{aligned}
d(p(\mathbf{z} \mid \mathbf{x}), q(\mathbf{z} \mid \mathbf{x})) &= \frac{1}{2} \int_{\mathbf{z}} |p(\mathbf{z} \mid \mathbf{x}) - q(\mathbf{z} \mid \mathbf{x})| d\mathbf{z} \\
&\leq \frac{1}{2} \int_{\mathbf{z}} \int_{\boldsymbol{\theta}} |p(\mathbf{z} \mid \boldsymbol{\theta})| |p(\boldsymbol{\theta} \mid \mathbf{x}) - q(\boldsymbol{\theta} \mid \mathbf{x})| d\boldsymbol{\theta} d\mathbf{z}
\end{aligned}
$$

It is trivial to show that (see also the assumptions mentioned in Theorem 4)

$$
\begin{aligned}
\frac{1}{2}\int_{\mathbf{z}}\int_{\boldsymbol{\theta}} |p(\mathbf{z}\mid\boldsymbol{\theta})||p(\boldsymbol{\theta}\mid\mathbf{x})-q(\boldsymbol{\theta}\mid\mathbf{x})|d\boldsymbol{\theta}d\mathbf{z} &= \\
\frac{1}{2}\int_{\boldsymbol{\theta}} |p(\boldsymbol{\theta}\mid\mathbf{x})-q(\boldsymbol{\theta}\mid\mathbf{x})|\{\int_{\mathbf{z}}|p(\mathbf{z}\mid\boldsymbol{\theta})|d\mathbf{z}\}d\boldsymbol{\theta} &= \\
\frac{1}{2}\int_{\boldsymbol{\theta}} |p(\boldsymbol{\theta}\mid\mathbf{x})-q(\boldsymbol{\theta}\mid\mathbf{x})|\,d\boldsymbol{\theta} &= \\
d(p(\boldsymbol{\theta}\mid\mathbf{x}),q(\boldsymbol{\theta}\mid\mathbf{x})).
\end{aligned}
$$

In the following theorem, we will show that as $n\to\infty$ (or equivalently as $\Delta\to 0$), $d(p(\mathbf{z}\mid\mathbf{x}),q(\mathbf{z}\mid\mathbf{x}))$ will then become very small (and bounded).

Theorem 4.1 *Suppose the likelihood function $p(\boldsymbol{x}\mid\boldsymbol{\theta})$ is bounded by M, and suppose that any prior distribution P has a differentiable log density with derivative $D\log p(\boldsymbol{\theta})$ bounded by N, i.e. there exists $N>0$ such that for all $\boldsymbol{\theta}$, $|D\log p(\boldsymbol{\theta})|\leq N$, and for any other arbitrary prior distribution, Q and for all $\gamma>0$ there exists $\Delta>0$ such that for all $\delta\leq\Delta$, $Q(B^c(\boldsymbol{\theta}_0(\boldsymbol{x});\delta))<\gamma$, where $\boldsymbol{\theta}_0(\boldsymbol{x})$ denote an estimation such as maximum likelihood. Then, for all $\epsilon>0$, there exists a $\Delta>0$ such that for all $\delta\leq\Delta$, $d(p(\boldsymbol{z}\mid\boldsymbol{x}),q(\boldsymbol{z}\mid\boldsymbol{x}))<\epsilon$.*

Proof We can write the equation below

$$
\begin{aligned}
|p(\mathbf{z}\mid\mathbf{x})-q(\mathbf{z}\mid\mathbf{x})| = &\int_{\boldsymbol{\theta}\in B(\boldsymbol{\theta}_0(\mathbf{x});\delta)} p(\mathbf{z}\mid\boldsymbol{\theta})|p(\boldsymbol{\theta}\mid\mathbf{x})-q(\boldsymbol{\theta}\mid\mathbf{x})|d\boldsymbol{\theta} \\
&+\int_{\boldsymbol{\theta}\notin B(\boldsymbol{\theta}_0(\mathbf{x});\delta)} p(\mathbf{z}\mid\boldsymbol{\theta})|p(\boldsymbol{\theta}\mid\mathbf{x})-q(\boldsymbol{\theta}\mid\mathbf{x})|d\boldsymbol{\theta},
\end{aligned}
$$

where $B(\boldsymbol{\theta}_0(\mathbf{x});\delta)$ is an open ball with its centre at $\boldsymbol{\theta}_0(\mathbf{x})$ and diameter δ. It can be easily concluded that $\int_{\boldsymbol{\theta}\notin B(\boldsymbol{\theta}_0(\mathbf{x});\delta)} p(\mathbf{z}\mid\boldsymbol{\theta})|p(\boldsymbol{\theta}\mid\mathbf{x})-q(\boldsymbol{\theta}\mid\mathbf{x})|d\boldsymbol{\theta}\leq 2M\gamma$. By the hypothesis in Corollary 1, we can say that for all $\omega>0$ there exists $\Delta>0$ such that for all $\delta<\Delta$, $|\log p(\boldsymbol{\theta})-\log q(\boldsymbol{\theta})|<\omega$, where p and q denote the densities associated with P and Q respectively. Therefore, by the results taken from Theorem 3, the following inequality can be achieved

$$\int_{\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0(\mathbf{x});\delta)} p(\mathbf{z} \mid \boldsymbol{\theta}) |p(\boldsymbol{\theta} \mid \mathbf{x}) - q(\boldsymbol{\theta} \mid \mathbf{x})| d\boldsymbol{\theta} \leq M e^{2\omega} \{e^{2\omega} - 1\}.$$

Therefore, $|p(\mathbf{z} \mid \mathbf{x}) - q(\mathbf{z} \mid \mathbf{x})| \leq M\{2\gamma + e^{2\omega}\{e^{2\omega} - 1\}\} = \epsilon$, as required.

5 Discussion

In this paper we present a new local sensitivity measure in terms of the credibility metrics. We have shown that these metrics asymptotically behave better. We have argued that the corresponding Fréchet derivative similar to the derivatives studied by Gustafson et al.[10] does not tend to zero. However, we do have uniform boundedness under appropriate conditions. That means a close credible metric a priori will give a close credible metric a posteriori. Therefore, we do not get the sort of divergence derived with the total variation metric as discussed in [9, 10].

It is important to investigate how the local sensitivity measure proposed in this paper is applicable to Bayesian networks. However, the proposed likelihood (multinomial distributions) and prior distribution (Dirichlet or product of Dirichlet's) for Bayesian networks with discrete variables would provide the conditions (especially, continuity condition) in the theorems and lemmas presented in this paper. Nevertheless, this still needs to be formally investigated.

Smith and Daneshkhah [15] developed new explicit total variation bounds on the posterior density as the function of closeness of the base prior to the approximating one (selected from a class of priors very similar to the one proposed in this paper) used and certain summary statistics of the calculated posterior density. It was illustrated that the approximating posterior density often converges to the base (or genuine) posterior as the number of sample point increases and the proposed bounds would allow them to identify when the posterior approximation might not.

Another inspiring work, which require further practical works, is related to investigate the asymptotic behaviour of the local sensitivity measures (or closeness distances), and compare it with the closeness distances reported in [15]. It should be noted that the local sensitivity measures introduced in this paper could be usually expressed in terms of difference between logarithms of the posterior densities. In many cases, this difference would ensure that the Hellinger distance (or total variation bounds as proposed in in [15, 18]), and thereby the corresponding local sensitivity measure will least be bounded for large enough sample sizes as shown in this paper.

The local sensitivity analysis, as studied in this paper and other relevant works, would be also very useful to answer the following questions, which are commonly raised when modelling multivariate data with complex dependency using Bayesian hierarchical models ([13]), Bayesian network, or Bayesian network pair-copula models ([2]). It is of great importance to investigate whether the network structure that is learned from data would be robust with respect to changes of the directionality of some specific arrows. Ross et al. [13] examine local sensitivity of the Bayesian hierarchical models by developing a new local sensitivity framework, known as ϵ-local sensitivity. The next important problem is to study whether the local conditional distribution/probability associated with the specified node would be robust with respect to the changes to its prior distribution or to the changes to the local conditional distribution of another node. However, this problem is addressed in [5, 15, 18], but there still exist areas for continued development. In particular, it is of great importance to examine the behaviour of the posterior distribution associated with the parameters of any node robust with respect to the changes to the prior distribution associated with the parameters of one specific node. Finally, would the quantities mentioned above be robust with respect to the changes in the independence assumptions as described in [4].

References

[1] Basu, S. 2000. Uniform stability of posteriors, *Statistics and Probability Letters* 46:53-58.

[2] Chatrabgoun, O., Hosseinian-Far, A., Chang, V., Stocks, N.G. and Daneshkhah, A., 2018. Approximating non-Gaussian Bayesian networks using minimum information vine model with applications in financial modelling. *Journal of computational science*, **24**, 266-276.

[3] Cowell, R. G. 1996. On compatible priors for Bayesian networks. *IEEE Trans. Pattern Anal. Machine Intelligence* 18:901-911.

[4] Daneshkhah, A. and Smith, J.Q., 2002. Multicausal Prior Families, Randomisation and Essential Graphs. In Probabilistic Graphical Models.

[5] Daneshkhah, A. 2004. *Estimation in Causal Graphical Models*, PhD Thesis, The University of Warwick.

[6] Daneshkhah, A., and Bedford, T. 2008. Sensitivity analysis of a reliability system using Gaussian processes. In *Advances In mathematical modelling for reliability*, (pp. 46-54), IOS Press.

[7] Daneshkhah, A., Hosseinian-Far, A., and Chatrabgoun, O. 2017. Sustainable maintenance strategy under uncertainty in the lifetime distribution of deteriorating assets. In *Strategic Engineering for Cloud Computing and Big Data Analytics* (pp. 29-50). Springer, Cham.

[8] Daneshkhah, A., and Smith, J. Q. 2002. Multicausal Prior Families, Randomisation and Essential Graphs. In *Probabilistic Graphical Models*.

[9] Gustafson, P., and L. Wasserman. 1995. Local sensitivity diagnostics for Bayesian inference. *The Annals of Statistics* 23:2153-2167.

[10] Gustafson, P., C. Srinivasan, and L. Wasserman. 1996. Local sensitivity. In *Bayesian Statistics 5*, ed. J. M. Bernardo et al, Oxford: Oxford University Press.

[11] Gomez-Deniz, E., and Calderin-Ojeda, E. 2010. A study of Bayesian local robustness with applications in actuarial statistics. *Journal of Applied Statistics*, **37(9)**, 1537-1546.

[12] Insua, D. R., Ruggeri, F., Alfaro, C., and Gomez, J. 2016. Robustness for adversarial risk analysis. In *Robustness analysis in decision aiding, optimization, and analytics* (pp. 39-58). Springer, Cham.

[13] Roos, M., Martins, T. G., Held, L., and Rue, H. 2015. Sensitivity analysis for Bayesian hierarchical models. *Bayesian Analysis*, **10(2)**, 321-349.

[14] Smith, J. Q. 2010. *Bayesian Decision Analysis: Principles and Practice*, Camb. Uni. Press.

[15] Smith, J. Q. and A. Daneshkhah. 2010. On the robustness of Bayesian networks to learning from non-conjugate sampling. *International Journal of Approximate Reasoning* 51:558–572.

[16] Smith, J. Q., and Rigat, F. 2008. Isoseparation and Robustness in Finite Parameter Bayesian Inference. *CRiSM Res Rep*, 07-22.

[17] Wasserman, L. (1992). Recent methodological advances in robust Bayesian inference. *Bayesian Statistics*, **4**, 483-502.

[18] Wright, S. K., and Smith, J. Q. 2018. Bayesian Networks, Total Variation and Robustness. arXiv preprint arXiv:1811.07179.