Imperial College London

IMPERIAL COLLEGE LONDON

DEPARTMENT OF MATHEMATICS

Assessing the impact of the prior in Bayesian Statistics through distances between probability measures.

Author:
Nikolaos Tsikouras

Supervisor:
Dr. Andrew B. Duncan

Submitted in partial fulfillment of the requirements for the MSc degree in Statistics of Imperial College London

September 2021

Abstract

It is known that as the sample size $n \to \infty$, the impact of the prior distribution in the resulting posterior diminishes. A practical question many researchers have however is what happens when n is fixed. The aim of this thesis is to inform the reader about some of the most common methods used to answering this question and propose a new method to answering the aforementioned question. In the first part of the thesis we introduce Stein's method and describe how this allowed to quantitatively measure the prior impact for univariate nested priors. The need to generalise the assessment of prior impact motivated us to use an already derived metric the Kernelized Stein Discrepancy (KSD), operating in the Reproducing Kernel Hilbert Space (RKHS), which is normally used in the context of goodness–of–fit tests. We extend the KSD to the Bayesian framework and derive a rate of change of the KSD with respect to the parameters in the prior parameter space which allows for a straightforward assessment of prior impact and sensitivity. By numerically comparing our metric results with the results from a common approach which provides bounds for the Wasserstein distance between two posteriors, we argue that we get similar behaviour but with the benefit of scalability to higher dimensions. Finally, we conclude with the limitations of our method as well as ways to overcome them in future work.

Acknowledgments

I would like to acknowledge the assistance of my supervisor Dr. Andrew B. Duncan. Through our meetings he would motivate, guide and support me. He has helped me shape into a better researcher. I would like to thank my family since without their support I would not have been in the position I am today. I want to express my sincere appreciation to the John S. Latsis Public Benefit Foundation which funded my MSc studies. Thank you to my friends Christos and Yorgos, who helped me proofread this thesis and whose company has helped me become a better person. Thanks to some new friends, Panayiota and Tom who made this past year a little less stressful. Last but not least, thank you to Lysandros who made the restless nights in the library more enjoyable.

Contents

List of Figures

Chapter 1

Introduction

1.1 Stating the Problem and the Motivation Behind the Thesis

Bayesian statistics is a popular approach for applying probability to statistical problems. It provides mathematical tools to update our beliefs about random events in light of seeing new data or evidence about those events. It is for this reason that there is increasing popularity of Bayesian Methods across many fields, including but not limited to, image processing [9], cyber security [59], econometrics [61].

One of the main things that make Bayesian inference different from the frequentist approach, is the use of prior knowledge into the statistical problem which is done by the so called *prior distribution*. The problem of defining a prior has long been an issue for statisticians and as of yet there is no universally approved way of picking one. Supposing one has defined a prior and has gathered data (likelihood), the Bayes theorem is used to calculate the *posterior distribution*. Typically the prior distribution does not depend on the observed data and it only expresses the practitioner's belief about the quantity of interest. Some common types of prior distributions are the following,

- Subjective priors. In this case the prior expresses the practitioner's personal beliefs, [5].
- Objective and informative priors. The practitioner may have information or data from prior experiments that compose a prior.
- Non-informative priors. This expresses ignorance to the parameter of interest and is generally dominated by the likelihood function.

The choice of the prior distribution can have a major impact on the accuracy of the results, thus picking a suitable prior is a crucial step in Bayesian analysis. Studies have demonstrated that informative priors can have a strong impact on the results, some examples include, [29, 58]. Surprisingly, even non-informative priors can have a (negative) effect on the final results, even with "larger" sample sizes, [14], some examples are, [36, 46]. Generally the more complex the model, the higher the need for an informative prior, thus using a non-informative prior may lead to not having enough information to produce accurate results,

[14]. However, this problem exists even in simpler problems and should thus be examined regardless, [14].

Because of this, and since there is no "rigorous" way of picking a prior, this is one of the main criticisms of the Bayesian approach. To help counter this criticism, every Bayesian analysis should include a *prior sensitivity analysis*, where the researchers examine, with different ways, the impact of the prior distributions on their respective posteriors. Typically, one would perturb the prior distribution's parameters or use a prior distribution from a different parametric family and calculate how far the new posteriors are, based on visual and statistical comparisons. Conducting sensitivity analysis allows the researcher to understand the impact of the prior and account for it before stating any significant results. We note here that having similar posterior results using different priors shows robustness, on the other hand having substantially different results with different priors is not a "bad" result, as it means that one is required to explain the results in a more careful way.

Additionally, having a method to quantitatively measure the impact of the prior distributions to the respective posteriors is of major importance to practitioners. It is well known that as the number of sample points tend to infinity the impact of the prior disappears, meaning that only data determine the posterior, [15, 16]. However, in practice gathering data might be expensive or difficult, thus the question researchers are interested in, is what happens at a finite sample size. Until recently, the question has not been tackled quantitatively because of a lack of adequate tools for measuring the said impact [20]. A new method using the famous *Stein's method* has been put forward in [38] allowing to answer the aforementioned question which we introduce more formally in Chapter 3.

1.2 Aim of the Thesis

The aim of this thesis is initially to provide the reader with an overview of the methods used up to date for measuring prior sensitivity using finite sample size, and showcase why Stein's method is an important tool for this problem. In addition, we wish to provide a way to conduct such an analysis for unnormalised distributions in order to have compatibility with MCMC methods. This is of major importance since in most applications we rarely have the opportunity to work with normalised densities and thus to compute the normalising constant one normally resorts to computationally expensive methods. Furthermore, we wish to conduct such analyses in high dimensions for both the parameter and the sample spaces. We note here that it would be desirable for this procedure to be done in a more rigorous and automatic way, so as to make sensitivity analysis quicker and more efficient.

1.3 Structure of the Thesis

The structure of this thesis is the following; Chapter 2 is a literature review where we discuss the previous ways of conducting sensitivity analysis as well as their limitations in practice. Next, in Chapter 3 we provide a detailed description of the two parts of Stein's method. In addition we look into how Stein's method links to the assessment of the impact of a prior distribution. Finally, we introduce the first Stein based approach to tackling prior sensitivity, and the disadvantages of the method. Chapter 4 includes some useful definitions and

introduces the Reproducing Kernel Hilbert Space (RKHS). In Chapter 5 we introduce the Kernelized Stein Discrepancy (KSD) and we give our contribution to this field by defining a new kernelized prior impact measure and a rate of change of that measure. In Chapter 6 we present how our method works through some simulations and numerical comparisons with the method mentioned in Chapter 3. Finally, in Chapter 7 we conclude this thesis by briefly summarising the main concepts, to discuss some limitations of our method with ways of overcoming them for future work.

Chapter 2

Existing Work on Prior Sensitivity

In this Chapter we present existing work on Bayesian prior sensitivity. This is by no means an exhaustive list, but it is a good starting point on what methods are being used, as hint the motivation for our metric. This Chapter includes only non–Stein based methods.

2.1 Non-Stein Based Methods For Prior Sensitivity and Prior Impact Assessment

2.1.1 Inbuilt Robustness

One approach to sensitivity analysis is to avoid the need for it by building robustness at the beginning rather than attempting to verify it at the end, [6]. It has been shown that using a prior distribution with flatter tails tends to produce results that are much more robust, examples are [10, 13]. Another possibility is to use nonparametric and infinite parametric Bayes procedures, [6], where a large class of nonparametric models is considered and given a prior (usually a Dirichlet process prior or a Gaussian process prior). This way the data will "force" the analysis to automatically adapt to the true model, examples are, [11, 37]. However caution must be exercised since it has been shown that nonparametric Bayesian procedures do not always satisfy consistency, [7, 16]. This approach can be problematic; as we have already mentioned the results can sometimes be surprising especially in more complicated models. Thus one should always conduct and report prior sensitivity analysis results regardless, so as to demonstrate robustness.

2.1.2 Common Methodology

The typical methodology for sensitivity analysis any article using Bayesian analysis should report is the following, [14]:

1. The researcher chooses a set of priors for the parameters to use for calculating the posterior model.

2. The model is calculated and convergence is reached for all parameters.
3. The researcher then comes up with another set of competing priors which are going to be used to examine robustness of the results.
4. The posteriors are calculated for the competing priors and compared via visual (if possible) and statistical comparisons.
5. The final model from (1) together with the results as well as the sensitivity analysis is presented, with comments on how robust, or not, the final model results are to different prior settings.

Next, another popular approach is to investigate *local* and *global* sensitivity to the prior, [6].

2.1.3 Local Sensitivity Approach

The main idea for local sensitivity is presented in [4]. Assuming there is a class of possible priors $\mathcal{P}$ and one is interested in a function of the posterior distribution, $\rho(\pi_0)$ where π_0 is the current prior in use and $\rho(\cdot)$ can be any function of the posterior, say the mean or even the posterior itself. Then one would like to investigate the rate of change of $\rho(\pi), \pi \in \mathcal{P}$ as we move infinitesimally from π_0. In parametric families of prior distributions, with say a set of parameters λ, $\mathcal{G} = \{\pi_\lambda, \lambda \in \Lambda\}$ this can be done by considering the derivative of $\rho(\pi_\lambda)$ with respect to the prior parameter λ evaluated at λ_0. This would indicate how sensitive is the prior to local changes in the prior parameter space around the prior of interest π_{λ_0}. However, this gets more complicated in higher dimensions as the derivative is *direction specific*. Therefore, when carrying this local sensitivity analysis in multiple dimensions it has been suggested in [4] to use the norm of the total derivative or the maximum values over all directions.

Along these lines, the authors in [45], introduce a prior sensitivity measure PS where they consider the derivative of the posterior mean with respect to the prior mean which is a trivial calculation. More specifically, the PS measure approximates the largest change of the posterior mean that can be induced by changing the prior mean by the multivariate analogue of one prior standard deviation, [45]. This method suffers from the fact that it requires the normalised posterior which in most practical applications is intractable or require expensive methods such as MCMC.

Another interesting approach is introduced in [51], where the authors define the formal $\epsilon-$local circular sensitivity and use it on Bayesian hierarchical models. More specifically, this $\epsilon-$local circular sensitivity set consists of ratios of the form,

$$S^c_{\lambda_0}(\epsilon) = \left\{ \frac{d(\rho(\pi_\lambda), \rho(\pi_{\lambda_0}))}{\epsilon}, \text{ for } \pi_\lambda \in G_{\lambda_0}(\epsilon) \right\}, \tag{2.1}$$

where $G_{\lambda_0}(\epsilon)$ is the grid of parameter values of the form,

$$G_{\lambda_0}(\epsilon) = \{\lambda : d(\pi_\lambda, \pi_{\lambda_0}) = \epsilon\},$$

where $\rho(\cdot)$ denotes the posterior distribution, $d(\cdot,\cdot)$ denotes a discrepancy between two densities and π_{λ_0} is the base prior density with λ_0 being the parameter values. Simply put, the $S^c_{\lambda_0}$ set consists of ratios where the numerator is given by the discrepancy of a posterior based on a base prior π_{λ_0}, and posteriors based on general priors π_λ that have discrepancy with the base prior exactly equal to ϵ while the denominator is ϵ. This "circular" approach examines every possible direction in the space of prior parameter values.

Circular sensitivity can be summarized by a single number, the authors in [51] use the worst-case sensitivity $S_{\lambda_0}(\epsilon)$ which is the maximum of the circular sensitivity $S^c_{\lambda_0}(\epsilon)$,

$$S_{\lambda_0}(\epsilon) = \max_{\lambda \in G_{\lambda_0}(\epsilon)} \{S^c_{\lambda_0}(\epsilon)\}. \tag{2.2}$$

The only input required for these sensitivity estimates is the prior distribution $\pi_{\lambda_0}(\theta)$ and the corresponding posterior density $\pi_{\lambda_0}(\theta|\mathbf{y})$. Having said that, if the sensitivity estimates in (2.1) and (2.2) are close to one, then the differences in posteriors and priors are comparable which is desirable. If however the ratios are larger than one, then this leads to *super-sensitivity*, [48] which is common in practice, [47, 63]. Again however, this approach suffers from tractability issues of the normalisation constant.

2.1.4 Global Sensitivity Approach

Global sensitivity on the other hand is concerned with a class of prior densities, [6], say Γ_π, which are compatible with the a priori information. The aim of these methods is to compute $\underline{\rho} = \inf_{\pi\in\Gamma_\pi} \rho(\pi)$ and $\overline{\rho} = \sup_{\pi\in\Gamma_\pi} \rho(\pi)$, without having to carry out the actual analysis for each prior in the class, [51]. Next, one reports the interval $(\underline{\rho}, \overline{\rho})$ as the range of possible answers and if this interval is small, then the result is considered to be robust; otherwise more data or other techniques ought to be used.

One of the most important decisions in global robustness is that of choosing a "proper" class of prior densities Γ. Berger et al. in [6] state some important properties for these classes, namely; they should be easy to handle computationally, they should be as easy as possible to interpret, the size of them should reflect the prior uncertainty, and finally they should be extendable to higher dimensions and adaptable in terms of allowing incorporation of constraints. In the overview of Berger et al. in [6] one can find most of the common classes used in practice. Although doing global robustness would be ideal, it suffers from the fact that it can only be used in simple models, [63].

2.2 Mean Observed Prior Effective Sample Size

One more metric for quantifying the impact of the prior was introduced in [50], where the authors defined the Effective Prior Sample Size (EPSS) which is an index that estimates the number of extra observations needed in order to transform the posterior based on a base prior into the posterior based on the prior of interest. Motivated by this idea, the authors in [32], introduced an extension to EPSS, the Mean Observed Prior Effective Sample Size (MOPESS), which is a set of steps that compares the two posteriors by measuring the relative impact of

the priors. In addition to that it can also specify which of the two priors has the largest effect on the posterior. Like most of the other prior impact measures, this too requires a specification of a base prior. Another disadvantage of this method is it requires sampling methods to be repeated multiple times. Thus the computational cost, especially for non–conjugate models, can be high, [23].

2.3 Neutrality Index

Another non-Stein based metric is the Neutrality index N introduced in [33]. It is an "absolute measure" for each prior which means that it does not require a specification of other priors. The Neutrality of a posterior distribution Θ with density $p(\theta; x)$ is essentially the probability of the posterior having a smaller value than the frequentist maximum likelihood estimate $\hat{\theta}_{MLE}$ of the parameter of interest,

$$N = \mathbb{P}(\Theta < \hat{\theta}_{MLE}) = \int_{\alpha}^{\hat{\theta}_{MLE}} p(\theta; x) d\theta,$$

where α is the lower bound of the posterior distribution's support. The closer N is to $1/2$, the smaller the impact of the prior. Once more, this approach suffers from requiring normalised densities as well as a high computational cost for higher dimensions. In addition, there exist cases where the MLE is at the boundaries of the parameter space and thus it can not be implemented, [23].

Recently, the authors in [38], introduced a Stein based method to tackle quantitatively the issue the problem of prior sensitivity. We are going to give an introduction to Stein's method and the method mentioned above, as well as the disadvantages of the method in the next Chapter.

Chapter 3

Stein Based Methods for Prior Sensitivity

3.1 Wasserstein-1 Distance

Before we jump into the main theorem that motivated this thesis, we are going to define the Wasserstein distance, also well known as earth mover's distance, and has started to be used increasingly in Statistics and Machine Learning [18, 35]. The Wasserstein distance provides a notion of measure of difference between probability measures, or in our case posterior probability distributions. The idea for the Wasserstein distance arose from the field of transportation theory, and is intuitively measuring the minimum cost to be paid in order to transform one distribution to another, [52]. The Wasserstein-1 (or simply Wasserstein) distance between two distributions, P_1 and P_2 is formally defined as, [38],

$$d_W(P_1, P_2) = \sup_{h \in \mathcal{H}} |\mathbb{E}_{X \sim P_1}[h(X)] - \mathbb{E}_{X \sim P_2}[h(X)]|, \tag{3.1}$$

where $\mathcal{H}$ is the class of functions belonging of the Lipschitz–1 $= \{g : \mathbb{R} \to \mathbb{R} : |g(y) - g(x)| \leq |y - x|\}$ class. One could opt for other distances such as the Kullback-Leibler divergence (KL divergence for short), but as we shall see in the next Section the Wasserstein distance has a nice link with Stein's Method, [39]. Basic properties of the Wasserstein distance include that it can take any value from 0 to infinity and it is insensitive to small changes of the two distributions, [60].

Generally, obtaining an exact expression of the Wasserstein distance between two posteriors is not an easy task. However, in the situation in which the two chosen priors lead to nested posteriors, the authors in [38], have provided sharp lower and upper bounds of it. We are also going to mention an extension to this theorem [23].

3.2 Stein's Method

Stein's method is a general method which allows one to obtain bounds on the distance between two probability distributions with respect to a probability metric. It was first introduced by Stein, [56], aiming to find how close is a distribution W to a well-understood

probability distribution Z (typically Normal or Poisson), [39]. In the first part of this Chapter, we give the outline of Stein's method and later we present the main result of [38] and its extension to the Bayesian framework. Stein's method has two parts, [21, 39],

> Part A: a way of converting the problem of bounding the error in the approximation of the two distributions W and Z into the problem of bounding the expectation of a function of W.
>
> Part B: a way of bounding the expectation in Part A. It is clear that this is conditional on W and the form of the function.

In more detail, say one is interested in a target probability distribution P with support $\mathcal{I}$. In Part A, we would first find a suitable linear operator $\mathcal{A} := \mathcal{A}_P$, called a *Stein operator*, and a "wide" class of functions $\mathcal{F}(\mathcal{A}) := \mathcal{F}(\mathcal{A}_\mathcal{P})$, called a *Stein class* such that for any other probability measure Q on $\mathcal{I}$ we have,

$$Q \sim P \text{ if and only if } \mathbb{E}_{X\sim Q}[\mathcal{A}_P f(X)] = 0, \text{ for all } f \in \mathcal{F}(\mathcal{A}_P),$$

where $Q \sim P$ means that Q has distribution P. The above equivalence is called a *Stein characterization* of P. Many times in practice we don't need the characterization of the class, [2], but only require a *Stein identity* for P namely,

$$\mathbb{E}_{X\sim P}[\mathcal{A}_P f(X)] = 0, \text{ for all } f \in \mathcal{F}(\mathcal{A}_P). \tag{3.2}$$

One of the most crucial steps in Stein's method is finding a suitable Stein operator which is tractable for the random variable in use, [39]. The different ways of choosing a Stein operator is not in the scope of this thesis, but most authors have used Stein operators which were differential operators, for more details the reader can look at [39].

For the rest of this thesis we are going to consider a Stein operator which is produced via the density approach, [17], while there are of course alternatives. Let P have an absolutely continuous probability density function p with respect to the Lebesgue measure on $\mathbb{R}$. Additionally, assume that p has interval support $\mathcal{I}$ with boundary points a, b, where $a, b \in \mathbb{R} \cup \{-\infty, +\infty\}$. Then, the Stein operator we are going to use is the following, [39].

Definition 3.2.1. Consider the Stein class for p to be the the collection $\mathcal{F}(p)$ of functions $f : \mathbb{R} \to \mathbb{R}$ such that (i) $x \mapsto f(x)p(x)$ is differentiable, (ii) $x \mapsto (f(x)p(x))'$ is integrable and (iii) $\lim_{x\to a} f(x)p(x) = \lim_{x\to b} f(x)p(x) = 0$. Then, the differential Stein operator is the following,

$$\mathcal{A}_P : f \mapsto \mathcal{A}_P f = \frac{(fp)'}{p} = f(\log p)' + f', \tag{3.3}$$

with the convention that $\mathcal{A}_P f(x) = 0$ outside of the support of $\mathcal{I}$. We note here that $\mathcal{A}_p f$ is also a valid notation and we will use these interchangeably.

Suppose now, we have a Stein operator and measures P and Q with densities $p(x)$ and $q(x)$ both having support $\mathbb{R}$. Let $\mathcal{A}_P$ be the Stein operator defined as in Definition (3.3), that is, $\mathcal{A}_P f(x) = f(x)\nabla_x \log p(x) + \nabla_x f(x)$ with a Stein class $\mathcal{F}(\mathcal{A}_P)$, with functions $f(x) \in \mathcal{F}(\mathcal{A}_P)$ being smooth and satisfying $\lim_{x\to a} f(x)p(x) = \lim_{x\to b} f(x)p(x) = 0$. Then we can prove that Stein's identity for P holds. The proof is done using integration by parts.

Proof.

$$\begin{aligned}
\mathbb{E}_{X\sim P}[\mathcal{A}_P f(x)] &= \int_a^b \left[f(x)\nabla_x \log p(x) + \nabla_x f(x)\right] p(x)dx \\
&= \left(\int_a^b f(x)\nabla_x p(x) + p(x)\nabla_x f(x)dx\right) \\
&= \int_a^b f(x)\nabla_x p(x)dx + p(x)f(x)\Big|_a^b - \int_a^b f(x)\nabla_x p(x)dx \\
&= p(x)f(x)\Big|_a^b = 0.
\end{aligned}$$

□

Then for any *Stein set*, [2], $\mathcal{F} \subset \mathcal{F}(\mathcal{A}_P)$, this motivates the definition of a dissimilarity measure called *Stein discrepancy* [26] as,

$$\mathbb{S}(Q, \mathcal{A}_P, \mathcal{F}) = \sup_{f\in\mathcal{F}} \|\mathbb{E}_{X\sim Q}[\mathcal{A}_P f(X)]\|, \tag{3.4}$$

for some norm $\|\cdot\|$. This discrepancy is intuitively finding the function f which maximises the violation of Stein's identity. By construction if $\mathbb{S}(Q, \mathcal{A}_P, \mathcal{F}) \neq 0$, then $Q \neq P$ and if $\mathcal{F}$ is "wide" enough then $\mathbb{S}(Q, \mathcal{A}_P, \mathcal{F}) = 0$ implies $Q = P$, [2]. If the Stein operator and the Stein set are properly chosen, then $\mathbb{S}(Q, \mathcal{A}_P, \mathcal{F})$ should capture the amount of dissimilarity between P and Q, while also ensuring tractability, [2]. To see why the first point can be achieved we are going to introduce the *Stein equation*.

Suppose $\mathcal{H}$ is a measure-determining class on $\mathcal{I}$ and for each $h \in \mathcal{H}$ one can find a solution $f = f_h \in \mathcal{F}(\mathcal{A}_P)$ of the Stein equation below,

$$h(x) - \mathbb{E}_{X\sim P}[h(X)] = \mathcal{A}_P f(x). \tag{3.5}$$

Then after taking expectations with respect to Q in both sides (assuming it is allowed), we get,

$$\mathbb{E}_{X\sim Q}[h(X)] - \mathbb{E}_{X\sim P}[h(X)] = \mathbb{E}_{X\sim Q}[\mathcal{A}_P f(W)]. \tag{3.6}$$

For this thesis, and in a large part of the literature on Stein's method, the distances used are known as Integral Probability Metrics (IPM), [44], which are defined as,

$$d_{\mathcal{H}}(P, Q) = \sup_{h\in\mathcal{H}} |\mathbb{E}_{X\sim P}[h(X)] - \mathbb{E}_{X\sim Q}[h(X)]|, \tag{3.7}$$

where $\mathcal{H}$ is a class of real-valued measurable functions for which the two expectations in (3.7) are finite.

From (3.7), using (3.6) we get,

$$d_{\mathcal{H}}(P,Q) = \sup_{h\in\mathcal{H}} |\mathbb{E}_{X\sim P}\left[h(X)\right] - \mathbb{E}_{X\sim Q}\left[h(X)\right]| \leq \sup_{f\in\mathcal{F}(\mathcal{H})} |\mathbb{E}_{X\sim Q}[\mathcal{A}_P f(X)]| = \mathbb{S}(Q,\mathcal{A}_P,\mathcal{F}(\mathcal{H})), \tag{3.8}$$

where $\mathcal{F}(\mathcal{H}) = \{f_h \mid h \in \mathcal{H}\}$ is the class of the solutions of (3.5). Stein in [56] found that using (3.8) provides a way to bound the distance between P and Q. This is because the problem of bounding $d_H(P,Q)$ has been transformed into the problem of bounding the Stein discrepancy $\sup_{f\in\mathcal{F}(\mathcal{H})} |\mathbb{E}_{X\sim Q}[\mathcal{A}_P f(X)]|$. From this point on, this is Part B of Stein's method, where a wide variety of approaches have been developed to handle this expectation.

Different choices of $\mathcal{H}$ give rise to different IPMs, including the Kolmogorov distance and the bounded Wasserstein distance. Since we are interested in the Wasserstein-1 distance we have defined in (3.1), we can derive it in this setting by picking $\mathcal{H} =$ Lipschitz–1 set.

Having said that, using (3.8) we get,

$$d_W(P,Q) = \sup_{h\in\mathcal{H}} |\mathbb{E}_{X\sim P}\left[h(X)\right] - \mathbb{E}_{X\sim Q}\left[h(X)\right]| \leq \sup_{f\in\mathcal{F}(\mathcal{H})} |\mathbb{E}_{X\sim Q}[\mathcal{A}_P f(X)]|, \tag{3.9}$$

where $\mathcal{F}(\mathcal{H}) = \{f_h \mid h \in \mathcal{H}\}$ is defined as above and $\mathcal{H}$ is the Lipschitz–1 class.

3.2.1 Stein Kernel

One of the most powerful tools in Stein's method and a key to a successful application of it, is the *Stein kernel*, [57].

Definition 3.2.2. Let $X \sim P$ be an absolutely continuous probability distribution with p.d.f. p, mean μ and Stein operator $\mathcal{A}_P$ defined as in Definition (3.3). Suppose that p has interval support with closure $[a,b]$. Then, let Id denote the identity function, the *Stein kernel* of P is the function $x \mapsto \tau_P(x)$ defined by,

$$\tau_P(x) = \mathcal{A}_P^{-1}(\mu - Id)(x) = \frac{1}{p(x)}\int_a^x (\mu - y)p(y)dy. \tag{3.10}$$

The random variable $\tau_P(X)$ is also called a Stein kernel for P.

From Definition 3.2.2, we can understand that the Stein kernel is a special function which if the Stein operator is applied to, returns the target density minus the point which it is applied. In one dimension the Stein kernel is a unique function, however in higher dimensions there may be many different functions that satisfy the property required. After defining

Stein's method we introduce a crucial theorem which was one of the first tools to quantitatively measure the impact of the prior distribution on the posterior by giving bounds on the Wasserstein distance between univariate continuous densities, [38].

3.2.2 Bounds on the Wasserstein Distance Between Univariate Continuous Densities.

In this subsection we are going to a present general result for two univariate continuous densities and later we are going to make it specific for the Bayesian framework. The theorem we are going to state works for nested distributions, however there are ways to relax this assumption, [23].

For both distributions we define the *Stein pairs* $(\mathcal{A}_i, \mathcal{F}_i)$ for $i = 1, 2$ to be the pair of Stein operator and Stein set for each distribution respectively. The following theorem provides computable and meaningful bounds on the Wasserstein distance $d_W(P_1, P_2)$ defined in (3.1), [38].

Theorem 3.2.1. For $i = 1, 2$ let P_i be a probability distribution with an absolutely continuous density p_i with support I_i with closure $\overline{I_i} = [a_i, b_i]$ for some $-\infty \leq a_i < b_i \leq +\infty$; suppose that $I_2 \subset I_1$ and let $X_i \sim P_i$ have finite means μ_i for $i = 1, 2$. Let $\mathcal{H} \equiv$ Lipschitz–1 functions on $\mathbb{R}$ and assume that $\pi_0(x) = \frac{p_2}{p_1}(x)$, defined on I_2, is differentiable on I_2 and satisfies,

1. $\mathbb{E}_{X\sim P_1}|(X - \mu_1)\pi_0(X)| < \infty$.
2. $\left(\pi_0(x) \int_{a_1}^{x} (h(y) - \mathbb{E}_{X\sim P_1}[h(X)])p_1(y)dy\right)'$ is integrable for all h in $\mathcal{H}$.
3. $\lim_{x\to a_2, b_2} \pi_0(x) \int_{a_1}^{x} (h(y) - \mathbb{E}_{X\sim P_1}[h(X)])p_1(y; x)dy = 0 \; \forall h \in \mathcal{H}$.

Then,

$$|\mathbb{E}_{X\sim P_1}[\pi_0'(X)\tau_1(X)]| \leq d_{\mathcal{W}}(P_1, P_2) \leq \mathbb{E}_{X\sim P_1}[|\pi_0'(X)|\tau_1(X)], \tag{3.11}$$

where τ_1 is the Stein kernel of P_1.

This result is able to quantify the difference between any two continuous univariate nested random variables. More specifically as we shall see in the next subsection it can quantify the difference between any two posteriors arising from two continuous univariate nested prior distribution p_1 and p_2 for any given sample size n, thus allowing the practitioner to make a more informed choice of prior distributions.

3.2.3 Influence of the Prior on Bayesian Statistics

We are going to adjust Theorem 3.2.1 to the Bayesian framework which allows the evaluation of the discrepancy between two posterior distributions resulting from two different priors. In

addition if one of the two priors is chosen to be the Uniform distribution then this methodology can be used to assess the impact of the other prior distribution on the posterior. Starting by fixing the notation, denote the likelihood of independent and identically distributed observations $X_1, \ldots, X_n$ by $\ell(x;\theta)$ coming from a parametric model with parameter of interest $\theta \in \mathbf{\Theta} \subseteq \mathbb{R}$.

Take two different (possibly improper) prior densities $p_1(\theta), p_2(\theta)$, which give the following two posterior distributions:

$$p_i(\theta; x) = \kappa_i(x)\ell(x;\theta)p_i(\theta), \quad i = 1, 2.$$

Denote by (Θ_1, P_1), (Θ_2, P_2) the pairs of random variables and cumulative distribution functions associated with the two densities respectively. The aim thus, becomes to find a bound to determine how close these two posterior distributions are. Suppose additionally that the two posterior distributions have finite means μ_1 and μ_2, have supports $I_i = (a_i, b_i)$ and that they have a Stein kernel defined as in Definition 3.2.2,

$$\tau_i(\theta; x) = \frac{1}{p_i(\theta; x)} \int_{a_i}^{\theta} (\mu_i - y)p_i(y; x)dy, \quad i = 1, 2.$$

Lets assume that $p_1(\theta; x)$ and $p_2(\theta; x)$ are nested, say $I_2 \subseteq I_1$, therefore $p_2(\theta; x)$ can be expressed as $\frac{\kappa_2(x)}{\kappa_1(x)}\rho(\theta)p_1(\theta; x)$, where $\kappa_i(x)$ are the normalizing constants and $\rho(\theta) = \frac{p_2(\theta)}{p_1(\theta)}$.

Therefore, under the same assumption as Theorem 3.2.1, we get that the Wasserstein distance between the two posterior distributions can be bounded by, [22],

$$|\mu_1 - \mu_2| = \frac{|\mathbb{E}_{X\sim\Theta_1}[\tau_1(X; x)\rho^{'}(X)]|}{\mathbb{E}_{X\sim\Theta_1}[\rho(X)]} \leq d_{\mathcal{W}}(P_1, P_2) \leq \frac{\mathbb{E}_{X\sim\Theta_1}[\tau_1(X; x)|\rho^{'}(X)|]}{\mathbb{E}_{X\sim\Theta_1}[\rho(X)]}, \tag{3.12}$$

where again τ_1 is the Stein kernel of P_1. Furthermore, in the special case where $p_1(\theta)$ is a uniform prior distribution and $p_2(\theta)$ is a general (possibly improper) prior density, then (3.12) can be rewritten as [38],

$$|\mu_1 - \mu_2| = |\mathbb{E}_{X\sim\Theta_2}[\tau_1(X)\rho_0(X)]| \leq d_W(P_1, P_2) \leq \mathbb{E}_{X\sim\Theta_2}[\tau_1(X)|\rho_0(X)|], \tag{3.13}$$

where $\rho_0(\theta) = \frac{p_2'(\theta)}{p_2(\theta)}$, and as we shall see is a very useful expression for computations.

By looking at the bounds in (3.12) and (3.13) we can see that they are close to each other as both the upper and lower bounds contain the same quantities. Additionally, (3.12) allows to quantitatively measure the discrepancy of the two posterior distributions with regard to the Wasserstein distance and thus provides a way of assessing the robustness of a base prior against competing priors. Furthermore, (3.13) is a way to quantify the impact of a prior by itself.

An interesting corollary of this theorem is that if we assume that $\rho(\cdot)$ is monotone then we can precisely quantify the impact of a prior because we have an exact expression of the Wasserstein distance between the posteriors,

$$d_W(P_1, P_2) = \frac{\mathbb{E}_{X\sim\Theta_1}[\tau_1(X;x)|\rho^{'}(X)|]}{\mathbb{E}_{X\sim\Theta_1}[\rho(X)]}.$$

3.2.4 Illustration in a Normal Model

To illustrate the importance of this methodology in practice we are going to consider the Normal likelihood, Normal prior setting and measure the impact of this prior. Let $X_1, \ldots, X_n$ be a random sample coming from a $N(\theta, \sigma^2)$, where σ^2 is known, and assume prior $p_1(\theta)$ to be an improper uniform distribution, and $p_2(\theta)$ to be a $N(\mu_0, \sigma_0^2)$. Using the bounds in (3.13) and noting that $\tau_1 = \frac{\sigma^2}{n}$ we have from [38],

$$\frac{\sigma^2}{n\sigma_0^2 + \sigma^2}|\bar{X} - \mu_0| \leq d_{\mathcal{W}}(P_1, P_2) \leq \frac{\sigma^2}{n\sigma^2 + \sigma^2}|\bar{X} - \mu_0| + \frac{\sqrt{2}}{\sqrt{\pi}}\frac{\sigma^3}{n\sigma_0\sqrt{\sigma_0^2 n + \sigma^2}}. \tag{3.14}$$

The above inequality allows us to quantify the impact of having a model with a Normal prior against having a model that depends solely on the data (Uniform prior). As expected, as $n \to \infty$ the Wasserstein distance converges to zero. Additionally, the distance increases as the prior mean gets further away from the data ($|\bar{X} - \mu_0|$ becomes large). Furthermore, as $\sigma_0 \to \infty$ the distance again converges to zero, which intuitively makes sense since the Normal prior converges to a Uniform improper prior. The two bounds in (3.14) differ by an $O(n^{-3/2})$ term, thus we have an exact expression for the Wasserstein distance with a $1/n$ precision, [38].

To help visualise the impact of the prior on the posterior we generated $N = 1000$ samples from a Normal distribution for each sample size $n = 10, 11, ..., 100$ with parameters $\theta = 0, \sigma^2 = 3$. For each of these samples we calculated the two bounds in (3.14), the difference of these two, as well as the true Wasserstein distance between them and calculated the average over all N iterations. Figure 3.1 shows all these values.

From the figure we can observe how fast the distance between the two bounds converges to 0 as n increases. Plots like this provide a nice visualisation of the impact of the chosen prior at any sample size, and can help practitioners understand from what sample size on the effect of the prior diminishes as well as make a more informed choice out of the existing priors in theory, [21]. We note here that we aim to have the distance between the two bounds as tight as possible in order to extract more information about the true Wasserstein distance.

Theorem 3.2.1 provides bounds that are meaningful and easy to calculate which do not depend on the normalizing constant. The main issue with this method is that in most of real life applications we can not have nested and univariate distributions. Additionally, although the Stein kernel is a powerful tool in Stein's method, recalling the definition of it we can see that it requires the computation of an integral, which in higher dimensions can be prohibitive as it would require super expensive methods such as MCMC, thus, it can not be evaluated in any useful way. However, the Stein operator requires just differentiation and this is easier to do in higher dimensions.

In addition, even the extension to this Theorem can be problematic. This is the Wasserstein Impact Measure (WIM), introduced in [23] and defined as,

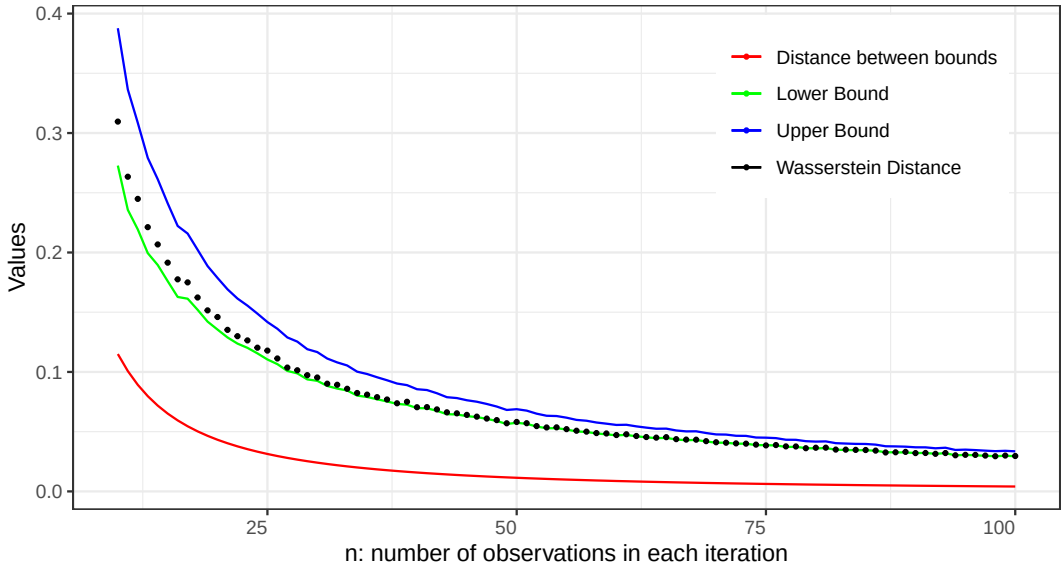

Figure 3.1: Figure shows the bounds of (3.14), the distance between the bounds as well as the true Wasserstein distance for N = 1000 iterations for sample sizes that range from 10 to 100 by steps of 1. The hyperparameters are $\mu_0 = 1$ and $\sigma_0^2 = 1$, and the likelihood parameters are $\theta = 0$ (unknown) and $\sigma^2 = 3$.

$$d_W(P_1, P_2) = \int_{a_1}^{b_1} \cdots \int_{a_m}^{b_m} |F_1(\theta_1, \ldots, \theta_m; x) - F_2(\theta_1, \ldots, \theta_m; x)| d_{\theta_1} \ldots d_{\theta_m}, \tag{3.15}$$

where a_j, b_j are the bounds of the support of $\theta_j, j = 1, \ldots, m$ and $F_i(\theta; x)$ it the cumulative distribution of the two posterior distributions P_1 and P_2. The issue with (3.15) is that, although it can be used irrespectively of the nested supports assumption and in high dimensions, it is common to work with complicated posterior cdfs which are intractable. Thus, again one would have to resort in MCMC methods to draw samples from **both** P_1 and P_2 and then use Monte Carlo integration. Therefore, it is clear that this presents a large computational burden.

The need to generalise this effectively to higher dimensions lead us to use the results from [42], where the authors introduce a new computable Stein discrepancy in the Reproducing Kernel Hilbert Space (RKHS). In the next Section we are going to introduce the RKHS and see how using the so called "kernel trick" is going to allow us to tackle our problem.

Chapter 4

Reproducing Kernel Hilbert Space (RKHS)

4.1 A Gentle Introduction to Hilbert Spaces

In this Section we are going to provide every useful definition of functional analysis before we define the Reproducing Kernel Hilbert Space and subsequently a kernel.

The first definition is that of the *norm*, [28].

Definition 4.1.1. Let $\mathcal{F}$ be a vector space over $\mathbb{R}$. A function $\|\cdot\| : \mathcal{F} \to [0, +\infty)$ is said to be a *norm* on $\mathcal{F}$ if,

1. $\|f\|_{\mathcal{F}} = 0$ if and only if $f = 0$ (norm separates points).
2. $\|\lambda f\|_{\mathcal{F}} = |\lambda| \, \|f\|_{\mathcal{F}} \forall \lambda \in \mathbb{R}, \forall f \in \mathcal{F}$ (positive homogeneity).
3. $\|f + g\|_{\mathcal{F}} \leq \|f\|_{\mathcal{F}} + \|g\|_{\mathcal{F}}, \forall f, g \in \mathcal{F}$ (triangle inequality).

An essential definition for a Hilbert space is the *complete space*, which essentially is a space for which the limit of every *Cauchy sequence* is inside the space. A formal definition of both the Cauchy sequence and the complete space is, [34],

Definition 4.1.2. A sequence $\{f_n\}_{n=1}^{\infty}$ of elements of a normed vector space $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ is said to be a *Cauchy sequence* if for every $\epsilon > 0$, there exists $N = N(\epsilon) \in \mathbb{N}$, such that for all $n, m \geq N, \|f_n - f\| < \epsilon$.

Definition 4.1.3. A space $\mathcal{X}$ is complete if every Cauchy sequence in $\mathcal{X}$ converges and the limit is in $\mathcal{X}$.

In order to study useful geometrical notions e.g., orthogonality, it is required to define the notion of *inner product*, [34].

Definition 4.1.4. Let $\mathcal{F}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ is said to be an *inner product* on $\mathcal{F}$ if,

1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{F}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{F}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{F}}$.

2. $\langle f, g\rangle_{\mathcal{F}} = \langle g, f\rangle_{\mathcal{F}}$.
3. $\langle f, f\rangle_{\mathcal{F}} \geq 0$ and $\langle f, f\rangle_{\mathcal{F}} = 0$ if and only if $f = 0$.

A vector space with an inner product is called an *inner product space*.

Finally, we define the *Hilbert space*, [34].

Definition 4.1.5. A *Hilbert space* is a complete inner product space.

4.2 Definition of the Reproducing Kernel Hilbert Space

Let $\mathcal{H}$ be a Hilbert space of functions mapping from some non-empty set $\mathcal{X}$ to $\mathbb{R}$. We write the inner product on $\mathcal{H}$ as $\langle f, g\rangle_{\mathcal{H}}$, and the associated norm as $\|f\|^2_{\mathcal{H}}$.

To define the RKHS it is essential to define the *Dirac functional*, [28].

Definition 4.2.1. Let $\mathcal{H}$ be a Hilbert Space of functions $f : \mathcal{X} \to \mathbb{R}$, defined on a non-empty set $\mathcal{X}$. For a fixed $x \in \mathcal{X}$, map $\delta_x : \mathcal{H} \to \mathbb{R}, \delta_x : f \to f(x)$ is called the *Dirac evaluation function* at x.

With the above definition we are ready to define a *Reproducing Kernel Hilbert Space*, [28].

Definition 4.2.2. A Hilbert Space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$, defined on a non-empty set $\mathcal{X}$ is said to be a *Reproducing Kernel Hilbert Space* (RKHS) if δ_x is continuous $\forall x \in \mathcal{X}$.

To define the reproducing kernel we first define a *positive definite* function and a *kernel*, [28].

Definition 4.2.3. A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is *positive definite* if $\forall n \geq 1$, $\forall(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n, \forall(x_1, \ldots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0.$$

Definition 4.2.4. Let $\mathcal{X}$ be a non-empty set. The function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is said to be a *kernel* if there exists a Hilbert Space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, y \in \mathcal{X}$,

$$k(x, y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{H}}.$$

Every inner product is a positive definite function, [28], thus every kernel is a positive definite function.

Now, we define a *reproducing kernel*, [28], which as we shall later going to see is going to crucially allow us to compute the supremum in (3.4) analytically in terms of the reproducing kernel.

Definition 4.2.5. Let $\mathcal{H}$ be a Hilbert space of $\mathbb{R}$- valued functions on a non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *reproducing kernel* of $\mathcal{H}$ if it satisfies,

- $\forall x \in \mathcal{X}, k(\cdot, x) \in \mathcal{H}$,
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$ (reproducing property).

In particular, for any $x, y \in \mathcal{X}$,

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y)\rangle_{\mathcal{H}}.$$

Finally, the Moore-Aronszajn theorem [8] states that every positive definite kernel k, is associated with a unique RKHS with reproducing kernel k.

4.3 Definition of the Vector-valued Reproducing Kernel Hilbert Space

More generally, we can make use of the vector-valued RKHS. The construction for it essentially follows the construction of the scalar RKHS, with the main difference that the reproducing kernel is now a matrix, [1]. Following the construction of the previous section we define the *vector-valued Dirac functional*, [3], and *vector-valued Reproducing Kernel Hilbert Space*, [1].

Definition 4.3.1. Let $\mathcal{H}$ be a Hilbert Space of functions $f : \mathcal{X} \to \mathbb{R}^d$, defined on a non-empty set $\mathcal{X}$. For a fixed $x \in \mathcal{X}$, map $\delta_x : \mathcal{H} \to \mathbb{R}^d, \delta_x : \boldsymbol{f} \to \boldsymbol{f}(\boldsymbol{x})$ is called the *Dirac evaluation function* at x.

Definition 4.3.2. A Hilbert Space $\mathcal{H}$ of functions $\boldsymbol{f} : \mathcal{X} \to \mathbb{R}^d$, defined on a non-empty set $\mathcal{X}$ is said to be a *vector-valued Reproducing Kernel Hilbert Space* (RKHS) if δ_x is continuous $\forall x \in \mathcal{X}$.

Now, we are going to extend Definition 4.2.4 to a *vector-valued kernel*, [1].

Definition 4.3.3. Let $\mathcal{X}$ be a non-empty set. The function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$ is said to be a *vector-valued kernel* if there exists a Hilbert Space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, y \in \mathcal{H}$,

$$k(x, y) = \langle \phi(x), \phi(y)\rangle_{\mathcal{H}}.$$

Finally, we define the *vector-valued reproducing kernel*, [1].

Definition 4.3.4. Let $\mathcal{H}$ be a Hilbert space of $\mathbb{R}^d$-valued functions on a non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$ is called a *vector-valued reproducing kernel* of $\mathcal{H}$ if it satisfies,

- $\forall x \in \mathcal{X}, \forall \mathbf{c} \in \mathbb{R}^d, k(\cdot, x)\boldsymbol{c} \in \mathcal{H}$,
- $\forall x \in \mathcal{X}, \forall \boldsymbol{f} \in \mathcal{H}, \langle \boldsymbol{f}, k(\cdot, x)\mathbf{c}\rangle_{\mathcal{H}} = \boldsymbol{f}(\boldsymbol{x})^\intercal \mathbf{c}$ (reproducing property).

In particular, for any $x, y \in \mathcal{X}$,

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}.$$

Similarly as before, there exists an extension to the Moore-Aronszajn theorem, [43], which states that a vector-valued reproducing kernel k is uniquely associated with a vector-valued RKHS.

Chapter 5

Kernelized Stein Discrepancy (KSD)

5.1 Introducing the Kernelized Stein Discrepancy

A new discrepancy statistic was proposed in [42], where it was used in the context of goodness–of–fit testing. The idea was motivated by Stein's method which we defined in Chapter 3 and the RKHS which we defined in Chapter 4, thus the name *Kernelized Stein Discrepancy* (KSD). KSD provides a convenient way to assess the distance between models with intractable normalization constants, [41]. It also allows us to generalise the assessment of prior impact to higher dimensions.

We have already seen Stein's method in the univariate case and now we are going to extend it in the multivariate case, [42]. We will denote $\nabla_x \boldsymbol{f}(x) = \left[\frac{\partial f_j(x)}{\partial x_i}\right]$ where $i = 1, \dots, d$ and $j = 1, \dots, d'$, be the gradient operator of function $\boldsymbol{f} = [f_1(x), \dots, f_{d'}(x)]$ which maps from $\mathcal{X}$ to $\mathbb{R}^d$.

Assume $\mathcal{X} \subset \mathbb{R}^d$. Let $p(x)$ be a continuous smooth density with support $\mathcal{X}$. The multivariate Stein operator of p is defined as, [42],

$$\mathcal{A}_p f(x) = \nabla_x \log p(x) f(x) + \nabla_x f(x),$$

where $f : \mathcal{X} \to \mathbb{R}$ is a smooth function that belongs in the Stein class of p, meaning it satisfies, $\lim_{\|x\| \to \infty} f(x)p(x) = 0$.

We want to emphasise here that both $\nabla_x \log p(x)$ and $\mathcal{A}_p$ are vector valued functions which map from $\mathcal{X}$ to $\mathbb{R}^d$. The above definitions can be extended even further for vector valued functions $\boldsymbol{f} = [f_1(x), \dots, f_{d'}(x)]$, by defining a $d \times d'$ matrix valued Stein operator as, [42],

$$\mathcal{A}_p \boldsymbol{f}(x) = \nabla_x \log p(x) \boldsymbol{f}(x)^\intercal + \nabla_x \boldsymbol{f}(x),$$

where each component of $\boldsymbol{f}$ belongs in the Stein class of p, meaning it is smooth and satisfies $\lim_{\|x\| \to \infty} f_i(x)p(x) = 0, \forall i = 1, \dots, d'$.

Similarly to the univariate case one can generalise the Stein identity for p in higher dimen-

sions, [42],

$$\mathbb{E}_{x\sim p}[\mathcal{A}_p \boldsymbol{f}(x)] = \mathbb{E}_{x\sim p}[\nabla_x \log p(x) \boldsymbol{f}(x)^\intercal + \nabla_x \boldsymbol{f}(x)] = 0, \tag{5.1}$$

for every $\boldsymbol{f}$ in the Stein class of p. We note that for the rest of this thesis for notational convenience we are going to be using $\mathbb{E}_{x\sim p}[\cdot]$ instead of $\mathbb{E}_{X\sim P}[\cdot]$.

We stress here again that $\mathcal{A}_p$ is a quantity that can be calculated in practice, even for high dimensions and for models with intractable normalization constants. This is because this operator depends on the unknown density p only via the score function $\nabla_x \log p(x)$.

As we have seen for the univariate case, using Stein's identity defined in (3.2) one can arrive in the Stein Discrepancy, [42], similarly one can arrive in the Stein Discrepancy for higher dimensions. However, in order to derive some of the following results it is more convenient to use the square of the typical Stein discrepancy, [26].

$$\mathbb{S}^2(P, Q) = \sup_{\boldsymbol{f}\in\mathcal{F}} \mathbb{E}_{x\sim q}[\mathcal{A}_p \boldsymbol{f}(x)]^2 = \sup_{\boldsymbol{f}\in\mathcal{F}} \mathbb{E}_{x\sim q}[\nabla_x \log p(x) \boldsymbol{f}(x)^\intercal + \nabla_x \boldsymbol{f}(x)]^2, \tag{5.2}$$

where the expectation is with respect to, q while the Stein operator is with respect to p and $\mathcal{F}$ is a set of smooth functions that satisfies (5.1) ensuring that $\mathbb{S}^2(P, Q) > 0$, when $p \neq q$. We note that in our simulations we are still going to report the actual Stein discrepancy and not the squared Stein discrepancy.

The problem with (5.2) is that usually it is computationally intractable. Therefore, the authors in [42] proposed $\mathcal{F}$ to be the unit ball of an RKHS $\mathcal{H}$ associated with a smooth positive definite kernel $k(x, x')$, with associated norm $\|\cdot\|_{\mathcal{H}}$ and inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. This framework leads to the squared *Kernelized Stein Discrepancy* which is easy to compute, [41].

Definition 5.1.1. The squared *Kernelized Stein Discrepancy* (KSD) $\mathbb{S}^2(P, Q)$ between distribution p and q is defined as,

$$\mathbb{S}^2(P, Q) = \sup_{\boldsymbol{f}\in\mathcal{H}} \{\mathbb{E}_{x\sim q}[\mathcal{A}_p \boldsymbol{f}(x)]^2 : \|\boldsymbol{f}\|_{\mathcal{H}} \leq 1\}. \tag{5.3}$$

We have already mentioned the reproducing property of the RKHS. In addition we also have a partial derivative reproducing property, $\nabla_x \boldsymbol{f}(x) = \langle \boldsymbol{f}(\cdot), \nabla_x k(x, \cdot)\rangle_{\mathcal{H}}$, [62], where the derivative operator is shifted to the kernel [41]. Thus we get [41],

$$\mathbb{E}_{x\sim q}[\mathcal{A}_p \boldsymbol{f}(x)] = \langle \boldsymbol{f}(\cdot), \mathbb{E}_{x\sim q}[\mathcal{A}_p k(\cdot, x)]\rangle_{\mathcal{H}}.$$

Therefore, our aim is to find the optimal $\boldsymbol{f}$ so as to compute,

$$\sup_{\boldsymbol{f}\in\mathcal{H}} \mathbb{E}_{x\sim q}[\mathcal{A}_p \boldsymbol{f}(x)] = \sup_{\boldsymbol{f}\in\mathcal{H}} \langle \boldsymbol{f}(\cdot), \mathbb{E}_{x\sim q}[\mathcal{A}_p k(\cdot, x)]\rangle_{\mathcal{H}}, \text{ where } \|\boldsymbol{f}\|_{\mathcal{H}} \leq 1.$$

For notational convenience we define $\beta_{q,p} = \mathbb{E}_{x\sim q}\mathcal{A}_p k(\cdot, x)$. Then, the maximum (it is obtained) value of this inner product is precisely $\|\beta_{q,p}\|_{\mathcal{H}}$, which follows from the Cauchy–Schwarz

inequality, since we are operating in the unit ball space. The value can be obtained by setting $\boldsymbol{f} = \beta_{q,p} / \|\beta_{q,p}\|_{\mathcal{H}}$ and thus we get,

$$\mathbb{S}^2(P,Q) = \sup_{\boldsymbol{f}\in\mathcal{H}} \mathbb{E}_{x\sim q}[\mathcal{A}_p \boldsymbol{f}(x)]^2 = \max_{\boldsymbol{f}\in\mathcal{H}} \mathbb{E}_{x\sim q}[\mathcal{A}_p \boldsymbol{f}(x)]^2 = \|\beta_{q,p}\|^2_{\mathcal{H}} = \|\mathbb{E}_{x\sim q}\mathcal{A}_p k(\cdot,x)\|^2_{\mathcal{H}}. \quad (5.4)$$

Lastly, we are going to present a final result, [40], which will be used to derive another form of the KSD, which is eventually used to derive the final, more computationally convenient form. The results states that, assuming Stein's identity holds for q we have,

$$\mathbb{E}_{x\sim q}[\mathcal{A}_p \boldsymbol{f}(x)] = \mathbb{E}_{x\sim q}[\mathcal{A}_p \boldsymbol{f}(x) - \mathcal{A}_q \boldsymbol{f}(x)] = \mathbb{E}_{x\sim q}[(\nabla_x \log p(x) - \nabla_x \log q(x))\boldsymbol{f}(x)^\intercal]. \quad (5.5)$$

Therefore, by notating $p = [p^1, \dots, p^d]$ and $q = [q^1, \dots, q^d]$ one can obtain another form of the squared KSD,

$$\begin{aligned}
&\mathbb{S}^2(P,Q) = \|\mathbb{E}_{x\sim q}\mathcal{A}_p k(\cdot,x)\|^2_{\mathcal{H}} = \|\mathbb{E}_{x\sim q}[(\nabla_x \log p(x) - \nabla_x \log q(x))k(\cdot,x)^\intercal]\|^2_{\mathcal{H}} \\
&= \sum_{i=1}^{d} \left\langle \mathbb{E}_{x\sim q}[(\nabla_x \log p^i(x) - \nabla_x \log q^i(x))k(\cdot,x)], \mathbb{E}_{x'\sim q}[(\nabla_{x'} \log p^i(x') - \nabla_{x'} \log q^i(x'))k(\cdot,x')] \right\rangle_{\mathcal{H}} \\
&= \mathbb{E}_{x\sim q}\mathbb{E}_{x'\sim q}[(\nabla_x \log p(x) - \nabla_x \log q(x))^\intercal \langle k(x,\cdot), k(x',\cdot)\rangle_{\mathcal{H}} (\nabla_{x'} \log p(x') - \nabla_x \log q(x'))] \\
&= \mathbb{E}_{x\sim q}\mathbb{E}_{x'\sim q}[(\nabla_x \log p(x) - \nabla_x \log q(x))^\intercal k(x,x')(\nabla_{x'} \log p(x') - \nabla_{x'} \log q(x'))]. \quad (5.6)
\end{aligned}$$

For completeness we state that the authors in [42], have proven that under some conditions, KSD is indeed a valid discrepancy measure, meaning $\mathbb{S}(P,Q) \geq 0$ and $\mathbb{S}(P,Q) = 0$ if and only if $P = Q$.

However, this KSD form requires the knowledge of both p and q, therefore in [42], the authors apply result (5.5) twice to derive a more convenient form that only requires knowledge from p.

Theorem 5.1.1. Assume p and q are smooth densities and that $k(x,x')$ is a function for which Stein's identity (5.1) for q holds. Define

$$\begin{aligned}
h_p(x,x') = &\nabla_x \log p(x)^\intercal k(x,x') \nabla_{x'} \log p(x') + \nabla_{x'} \log p(x')^\intercal \nabla_x k(x,x') \\
&+ \nabla_x \log p(x)^\intercal \nabla_{x'} k(x,x') + \text{trace}(\nabla_{x,x'} k(x,x')),
\end{aligned}$$

$$\text{then} \qquad \mathbb{S}^2(P,Q) = \mathbb{E}_{x,x'\sim q}[h_p(x,x')]. \quad (5.7)$$

The result in Theorem 5.1.1 is a tractable formula since it only involves gradients which are in general straightforward to compute even in higher dimensions. Therefore, it allows empirical evaluation of $\mathbb{S}^2(P,Q)$ either via the $U-$statistic, [42],

$$\hat{\mathbb{S}}^2_u(P,Q) = \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} h_p(x_i,x_j), \quad (5.8)$$

which is a minimum variance unbiased estimator for $\mathbb{S}_u^2(P,Q)$, [54], or the $V-$ statistic, [42],

$$\hat{\mathbb{S}}_v^2(P,Q) = \frac{1}{n^2}\sum_{i,j=1}^{n} h_p(x_i,x_j), \tag{5.9}$$

which has the advantage that is always non-negative [42].

In the next Section we are going to show our contribution to the problem of prior sensitivity analysis. We are going to derive a KSD formula for any two posteriors as well as a rate of change of the KSD between posteriors with priors belonging in the same parametric family.

5.2 Kernelized Stein Discrepancy for Prior Impact Assessment and Sensitivity Analysis

We are first going to set up some notation, let $\mathbf{X_1}, \ldots, \mathbf{X_n}$ be independent and identically distributed observations of a random variable and denote the likelihood by $\ell(\mathbf{x};\theta)$ from a parametric model with parameter of interest $\theta \in \mathbf{\Theta}$, which need not be a one dimensional quantity. Denote by $(\mathcal{A}_P, \mathcal{F}_P)$, $(\mathcal{A}_Q, \mathcal{F}_Q)$ the pairs of Stein operator and Stein class (as defined in the start of this Chapter) for each distribution respectively.

Now consider two different prior densities $p_1(\theta), p_2(\theta)$, which give the following to two posterior distributions:

$$P(\theta;\mathbf{x}) \propto \ell(\mathbf{x};\theta)p_1(\theta).$$
$$Q(\theta;\mathbf{x}) \propto \ell(\mathbf{x};\theta)p_2(\theta).$$

Consider,

$$\begin{aligned}(\mathcal{A}_P - \mathcal{A}_Q)\boldsymbol{f}(\theta)^\intercal = \nabla \log \ell(\mathbf{x};\theta)\boldsymbol{f}(\theta)^\intercal + \nabla \log p_1(\theta)\boldsymbol{f}(\theta)^\intercal - \nabla \log \ell(\mathbf{x};\theta)\boldsymbol{f}(\theta)^\intercal \\ - \nabla \log p_2(\theta)\boldsymbol{f}(\theta)^\intercal = (\nabla \log p_1 - \nabla \log p_2)(\theta)\boldsymbol{f}(\theta)^\intercal.\end{aligned} \tag{5.10}$$

The Bernstein von–Mises theorem, [12], says that, under appropriate conditions:

$$\mathbb{P}(\sqrt{n}(\theta - \theta_{MLE})|\mathbf{x} < t) \to \mathbb{P}(N(0, I^{-1}(\theta_0)) < t, \text{ as } n \to \infty,$$

for all $t \in \mathbb{R}$, where θ_0 is the "true" parameter and θ_{MLE} is the maximum likelihood estimator of θ given the data (in particular this is independent of the prior, and depends only the likelihood).

Note that the right hand side is independent of n. This suggests that, rather than comparing the empirical posteriors P and Q, we should compare the rescaled empirical posteriors, where

we shift the mean by θ_{MLE} (independent of the prior) and rescale the variance by n. To this end, we compare,

$$\tilde{P}(\theta; \mathbf{x}) \propto P\left(\theta; (\mathbf{x} - \theta_{MLE})/\sqrt{n}\right),$$

and

$$\tilde{Q}(\theta; \mathbf{x}) \propto Q(\theta; (\mathbf{x} - \theta_{MLE})/\sqrt{n}).$$

Note that,

$$\nabla \log \tilde{P}(\theta; x) = \frac{1}{\sqrt{n}} \nabla \log P\left(\theta; (x - \theta_{MLE})/\sqrt{n}\right).$$
$$\nabla \log \tilde{Q}(\theta; x) = \frac{1}{\sqrt{n}} \nabla \log Q\left(\theta; (x - \theta_{MLE})/\sqrt{n}\right).$$

with densities $\tilde{p}, \tilde{q}$. Now, given $\theta \sim P$ and $\theta \sim Q$, this implies that,

$$\theta_{MLE} + \sqrt{n}\theta \sim \tilde{P} \tag{5.11a}$$
$$\theta_{MLE} + \sqrt{n}\theta \sim \tilde{Q}. \tag{5.11b}$$

We therefore use KSD to compare $\tilde{P}$ and $\tilde{Q}$. In addition, it makes sense to rescale the kernel k by n, so that we use,

$$k_n(\theta, \theta') = \frac{1}{n} k((\theta - \theta_{MLE})/\sqrt{n}, (\theta' - \theta_{MLE})/\sqrt{n}).$$

Following exactly the same steps for (5.6) from the previous Section, and using (5.10) we obtain,

$$\mathbb{S}^2(\tilde{P}, \tilde{Q}) = \mathbb{E}_{x\sim\tilde{q}} \mathbb{E}_{x'\sim\tilde{q}} \Bigg[\frac{1}{\sqrt{n}} (\nabla_x \log p_1((x - x_{MLE})/\sqrt{n}) - \nabla_x \log p_2((x - x_{MLE})/\sqrt{n}))^\intercal k_n(x, x') \\ \frac{1}{\sqrt{n}} (\nabla \log p_1((x' - x_{MLE})/\sqrt{n}) - \nabla_x \log p_2((x' - x_{MLE})/\sqrt{n})) \Bigg]. \tag{5.12}$$

But substituting for $\tilde{P}, \tilde{Q}$ using (5.11a) and (5.11b) in (5.12) yields,

$$\mathbb{S}^2(P, Q) = \frac{1}{n^2} \mathbb{E}_{x\sim q} \mathbb{E}_{x'\sim q} [\nabla_x (\log p_1(x) - \log p_2(x))^\intercal k(x, x') \nabla_{x'} (\log p_1(x') - \log p_2(x'))]$$

$$= \frac{1}{n^2}\mathbb{E}_{x\sim q}\mathbb{E}_{x'\sim q}[\nabla_x \log(p_1/p_2)(x)^\intercal k(x,x')\nabla_{x'}\log(p_1/p_2)(x')]. \tag{5.13}$$

This is a similar estimator to (5.6) but rescaled accordingly. We note here that the substitution we did in (5.13) is mathematically correct since P and Q are empirical measures, that is a sum of Dirac, thus there is no need to change the volume.

Again, one can estimate this KSD by either the $U-$ statistic or the $V-$ statistic by, for instance generating $X_1, \ldots, X_N$ of particle samples for the model Q (or P, notice the symmetry). Here we are going to present the $V-$statistic since we want to take advantage of the fact that it always returns a non-negative value.

$$\hat{\mathbb{S}}_v^2(P,Q) = \frac{1}{n^2}\frac{1}{N^2}\sum_{i,j=1}^{N} \nabla \log(p_1/p_2)(X_i)^\intercal k(X_i,X_j)\nabla\log(p_1/p_2)(X_j) \tag{5.14}$$

Remark. Although the likelihood $\ell(\mathbf{x};\theta)$ terms cancel out, the data still impact this estimate via the expectations.

We note here that this is the squared KSD for the two posteriors, thus we would need to calculate and report the square root of it. This formula is fairly straightforward to compute in practice since the ingredients required are the gradient of the log of the ratio of the two priors and samples from any of the two posteriors (we would always pick the simplest one to sample from). In practice, one would need to calculate the KSD for a range of possible prior distributions and investigate any "suspicious" differences. If there exist relatively big differences between the base prior and any of the competing ones then this should be accounted when one reports the Bayesian results based on the base prior. Otherwise, one can be more confident that the Bayesian results are robust.

5.2.1 Advantages Over Score Matching

An alternative to the KSD is the Score Matching (SM) estimator, [30]. For a posterior distribution P, with probability density function p and given a particle sample size of N, the SM estimator is given by,

$$\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{2}|\nabla \log p(X_i)|^2 + \Delta \log p(X_i)\right],$$

where $\Delta \log p = \nabla^2 \log p$, is the Laplacian operator. One could have used the SM instead of the KSD since it also has the advantage of not depending on the normalization constant of p while also having just one sum instead of the two we have in formula (5.14).

The reasons however we chose not to work with the SM estimator is that first it requires two derivatives which can sometimes be expensive to compute. Apart from that, there exist probability distributions for which the second derivative does not exist, thus in a sense it is less robust. The most important disadvantage though is that plugging in the posteriors as in (5.10) we did not find an elegant way that would cancel the likelihood term.

This is crucial since for (5.13), which is independent of the likelihood, we would require to evaluate the likelihood only if an MCMC algorithm is required for particle sample generation. On the other hand using the SM estimator we would have to calculate the likelihood for every realization X_i. Therefore, if the likelihood is expensive to compute and the data size is large this can be a massive computational burden.

5.3 Rate of Change of the Kernelized Stein Discrepancy

Let's assume now that the prior distribution belongs to a parametric family of distributions which depend on a parameter δ, which need not be one–dimensional. Therefore, we can write $p(\theta;\delta)$ instead of $p(\theta)$, where δ belongs to the *parameter space*; the set of all possible values that the parameter δ can take. Now, since in practice knowledge about most models can not be accurate, the question we want to answer is, if we fix $\delta = \delta_0$ and perturb it, say $\delta = \delta_0 + \epsilon$, how significant will the impact be on the resulting posterior. This is essentially the local sensitivity approach we have seen in the Literature Review.

So, let $\mathbf{X_1}, \ldots, \mathbf{X_n}$ be independent and identically distributed observations of a random variable and denote the likelihood by $\ell(\mathbf{x};\theta)$ from a parametric model with parameter of interest $\theta \in \mathbf{\Theta}$, which again need not be univariate. This would then give posterior distributions, which depend on the parameter δ,

$$P_1(\theta;\mathbf{x}) = P(\theta;\mathbf{x},\delta_0+\epsilon) \propto \ell(\mathbf{x};\theta)p(\theta;\delta_0+\epsilon).$$
$$P_2(\theta;\mathbf{x}) = P(\theta;\mathbf{x},\delta_0) \propto \ell(\mathbf{x};\theta)p(\theta;\delta_0),$$

with densities p_1 and p_2. Thus, by using the first equality of (5.13) for these two posterior distributions we get,

$$\mathbb{S}^2(P_1,P_2) = \frac{1}{n^2}\mathbb{E}_{x\sim p_2}\mathbb{E}_{x'\sim p_2}[(\nabla \log p(x;\delta_0+\epsilon) - \nabla \log p(x;\delta_0))^\intercal k(x,x') \\ (\nabla \log p(x';\delta_0+\epsilon) - \nabla \log p(x';\delta_0))]$$

$$\frac{\mathbb{S}^2(P_1,P_2)}{(\epsilon)^2} = \frac{1}{n^2}\mathbb{E}_{x\sim p_2}\mathbb{E}_{x'\sim p_2}\left(\frac{\nabla \log p(x;\delta_0+\epsilon) - \nabla \log p(x;\delta_0)}{\epsilon}\right)^\intercal k(x,x') \\ \left(\frac{\nabla \log p(x';\delta_0+\epsilon) - \nabla \log p(x';\delta_0)}{\epsilon}\right).$$

Now, by taking $\epsilon \to 0$ we get the Rate Of Change (ROC) of the KSD between P_1 and P_2 with respect to the parameter δ evaluated at δ_0 to be:

$$ROC(P_1,P_2)^2 = \frac{1}{n^2}\mathbb{E}_{x\sim p_2}\mathbb{E}_{x'\sim p_2}\left(\nabla_\delta \nabla_x \log p(x;\delta_0)\right)^\intercal k(x,x')\left(\nabla_\delta \nabla_x \log p(x';\delta_0)\right). \quad (5.15)$$

Again, one can estimate this ROC by either the $U-$statistic or the $V-$statistic. We are going to be using the $V-$statistic to get a non-negative estimation. Then by simulating $X_1, \ldots, X_N$ of particle samples for the model P_2 (or P_1, notice the symmetry) for the posterior distribution with prior parameters δ_0 we get,

$$\hat{ROC}(P_1, P_2)^2 = \frac{1}{n^2}\frac{1}{N^2}\sum_{i,j=1}^{N} \left(\nabla_\delta \nabla_x \log p(X_j; \delta_0)\right)^\intercal k(X_i, X_j) \left(\nabla_\delta \nabla_x \log p(X_j; \delta_0)\right). \tag{5.16}$$

Note, that this is the squared rate of change of the posterior discrepancy with respect to the parameter δ at $\delta = \delta_0$, thus we would need to calculate and report the square root of it. This formula is very straightforward to evaluate in practice, since usually the priors are simple enough that one can trivially compute the gradient with respect to both parameters. If this rate of change is "small" for an "area" around the selected value of δ, then one can be more confident that the results of the Bayesian analysis are robust. This means that even if there are minor misspecifications in the prior parameter values this would not be having a major impact on the resulting posteriors.

We are going to present a pseudocode for the script we use to calculate the KSD and rate of change of KSD between two posteriors. First the KSD between posterior P and Q.

Algorithm 1 KSD

Require: Particle sample X, priors p_1, p_2
1: Set n to be the particle sample size X
2: Initialize empty $n \times n$ matrix k
3: Calculate r to be the gradient ratio of log of the two priors p_1, p_2 for each particle ▷ $n \times 1$ vector
4: **for** $i = 1, \ldots, n$ **do**
5: **for** $j = 1, \ldots n$ **do**
6: Calculate KSD values and store in $k_{i,j}$
7: **end for**
8: **end for**
9: **return** $\sqrt{\text{mean}(r^\intercal k r)}$

Next, the rate of change of the KSD for a parametric family of priors.

Algorithm 2 Rate of change of the KSD

Require: Particle sample X, prior p
1: Set n to be the particle sample size X
2: Initialize empty $n \times n$ matrix k
3: Calculate r to be the gradient with respect to the prior parameter and with respect to the parameter of interest of log of prior p for each particle ▷ $n \times 1$ vector
4: **for** $i = 1, \ldots, n$ **do**
5: **for** $j = 1, \ldots n$ **do**
6: Store KSD values in $k_{i,j}$
7: **end for**
8: **end for**
9: **return** $\sqrt{\text{mean}(r^\intercal k r)}$

Chapter 6

Numerical Results

6.1 Defining the Kernels

A common kernel which is used in practice is the Gaussian kernel, [53], since it tends to perform well, under smoothness assumptions, and should be taken into consideration especially when there is no added information about the data, [55]:

$$k_{Gau}(x, x') = \sigma_G^2 \exp\left(-\frac{\|x - x'\|}{2l^2}\right).$$

Another popular kernel used in the KSD setting is the Inverse Multiquadric (IMQ) kernel which was used in [27] so as to measure sample quality, because of its slow decay rate:

$$k_{IMQ}(x, x') = (c^2 + \|x - x'\|_2^2)^{\beta}, \text{ for some } \beta \in (-1, 0), c > 0.$$

Finally, another kernel which is especially used in NLP tasks is the Polynomial kernel, [24]:

$$k_{Poly}(x, x') = (x \cdot x' + c)^d, \text{ for some } c > 0, d \in \mathbb{N}$$

As we are later going to see, we can even use a combination of these kernels. This is mathematically justified since the sums of kernels and products of kernels are valid kernels [28].

In the following sections we are going to present the KSD as well as the rate of change of the KSD for different prior parameter values, different kernels, and using a particle sample size of 10^3 for every simulation, unless otherwise stated.

6.2 Illustration of the KSD in a Normal Model

Similar to Chapter 3 we are going to test our result in the Normal likelihood, Normal prior example. Assume we have n data coming from a $N(\theta, \sigma^2)$ distribution, where σ^2 is known.

Assume additionally, that we have two "competing" Normal priors for the mean $p_1(\theta)$ which is a $N(\theta_1, \sigma_1^2)$ and $p_2(\theta)$ which follows a $N(\theta_2, \sigma_2^2)$, which give posterior distributions $P(\theta; \mathbf{x})$ and $Q(\theta; \mathbf{x})$. For illustration purposes we are going to generate n samples from a $N(0, 1)$ distribution, which is going to be the likelihood of our model, note here that $\theta = 0$ is unknown. The kernel we are going to use is the Gaussian.

The Gaussian kernel hyperparameter σ_G^2 will be assigned to the value 1, since it is a common value in the literature and it is only a scale parameter thus the larger this value is the larger the KSD. As for the lengthscale l parameter we are going to find the optimal parameter using a Gradient Ascent algorithm (GA), where the goal is to maximise the KSD function. Alternatively, a common rule of thumb used in practice is to select l as the median of the squared Euclidean distance between pairs of sample points. We note that in our testing we tried both methods and got similar results, thus indicating robustness to the lengthscale parameter.

First we are going to calculate the KSD in the Normal likelihood setting for an "average" sample size of $n = 100$ between a base Normal prior with $\theta_1 = 0$ and $\sigma_1^2 = 1$, that is a prior is in total agreement with the data and competing Normal priors with parameters indicated by every possible pair in the grid $\{\theta_2 = [-4, 4]$ by $1,\ \sigma_2^2 = [1, 4]$ by $0.1\}$. Denote P as the posterior distribution coming from the base Normal prior distribution and Q the posterior distribution coming from any of the competing Normal priors. To calculate the KSD estimate we generate $N = 10^3$ Normal samples $X_1, \ldots, X_N$ from either of the two Normal posterior distributions, say Q, with any simple method and using Formula (5.14), we calculate,

$$\begin{aligned}\hat{\mathbb{S}}_v(P, Q) &= \sqrt{\frac{1}{n^2}\frac{1}{N^2}\sum_{i,j=1}^{N} \nabla \log(p_1/p_2)(X_i) k_{Gau}(X_i, X_j) \nabla \log(p_1/p_2)(X_j)} \\ &= \sqrt{\frac{1}{n^2}\frac{1}{N^2}\sum_{i,j=1}^{N} \left(\frac{X_i - \theta_2}{\sigma_2^2} - \frac{X_i - \theta_1}{\sigma_1^2}\right) k_{Gau}(X_i, X_j) \left(\frac{X_j - \theta_2}{\sigma_2^2} - \frac{X_j - \theta_1}{\sigma_1^2}\right)}.\end{aligned}$$

The resulting KSD values are shown as a contour plot in Figure 6.1. Figure 6.1 indicates that all of the competing priors give similar posterior results. This is true because the KSD values are all very close to 0. Thus, if these were the only competing a priori knowledge we could conclude that the Bayesian inference based on the base prior is robust

Additionally, in Figure 6.1 we observe symmetry around the true unknown mean 0, which is expected since we are working with a Normal model. We can also see that it matches our intuition since first, the KSD decays smoothly as the two prior distributions come close to each other. Additionally, when the prior mean gets further away from the true unknown mean, the KSD is higher when σ_2^2 is close to 1, since we have an average sample size and the competing prior indicates with "certainty" that the mean is far from 0, while as σ_2^2 increases the two priors start overlapping and thus the two resulting posterior distributions come closer.

This is also obvious in Figure 6.2 where we have plotted two cases. On the upper left panel we have plotted a $N(0, 1)$ prior (light red) and a $N(4, 1)$ prior (light blue), while on the upper right panel we have plotted the resulting posteriors. As expected, the impact is low but still existent, indicated by a "small" KSD value ($\approx 10^{-3}$). On the bottom left panel we have plotted a $N(0, 1)$ prior (light red) and a $N(4, 4)$ prior (light blue), while on the bottom right panel we

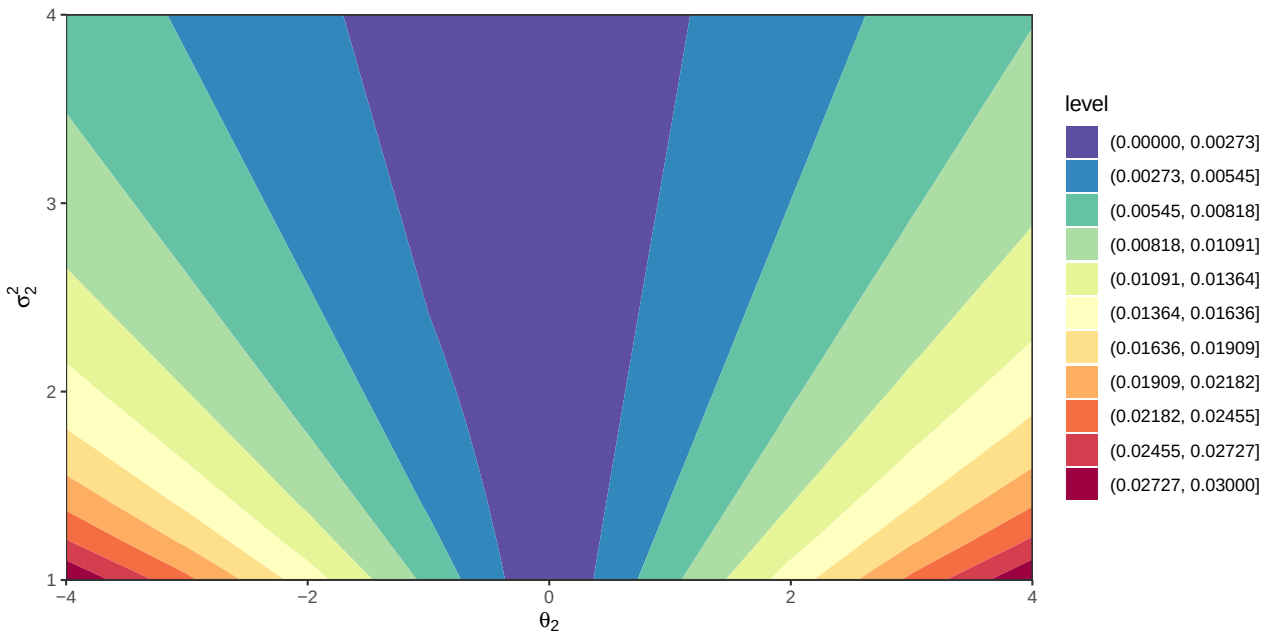

Figure 6.1: Figure showing the KSD between a $N(0,1)$ prior and Normal priors with parameters indicated by every possible pair in the grid $\{\theta_2 = [-4,4]$ by $1,\ \sigma_2 = [1,4]$ by $0.1\}$. The true underlying distribution comes from a $N(0,1)$ distribution, where μ is unknown.

have plotted the resulting posteriors. The two priors have started to overlap and the effects can also be seen on the resulting posteriors since they are closer than the previous scenario which is expected from the smaller KSD value ($\approx 10^{-4}$).

Next, since we are using prior distributions in the same parametric family we are going to calculate the rate of change of the KSD for the mean with respect to both prior parameters μ_0 and σ_0^2. The parameters we are going to consider lie in the space $\mathcal{C} = \{\theta_0 = [-10,10]$ by $1, \sigma_0 = [1,5]$ by $0.1\}$. We note here that we chose not to work with the total rate of change but instead decided to report ROC with respect to the parameters individually. This is done for demonstration purposes as we noticed that having one plot for each parameter is more intuitive. However, if one wants to report the total rate of change, then one of the most popular ways would be to report the maximum value over all directions, [4].

To calculate the ROC with respect to μ_0 we generate $N = 10^3$ Normal samples $X_1, \ldots, X_N$ from the resulting posterior for all prior parameter pairs in $\mathcal{C}$ and using Formula 5.16, we would calculate,

$$\begin{aligned} R\hat{O}C(P_1, P_2) &= \sqrt{\frac{1}{n^2}\frac{1}{N^2}\sum_{i,j=1}^{N} \left(\nabla_\theta \nabla_x \log p(X_j;\theta_0)\right) k_{Gau}(X_i, X_j) \left(\nabla_\theta \nabla_x \log p(X_j;\theta_0)\right)} \\ &= \sqrt{\frac{1}{n^2}\frac{1}{N^2}\sum_{i,j=1}^{N} \left(\frac{1}{\sigma} k_{Gau}(X_i, X_j) \frac{1}{\sigma}\right)}. \end{aligned}$$

Similarly, to calculate the ROC with respect to σ_0^2 we generate $N = 10^3$ Normal samples $X_1, \ldots, X_N$ from the resulting posterior for all prior parameter pairs in $\mathcal{C}$ and using Formula 5.16, we would calculate,

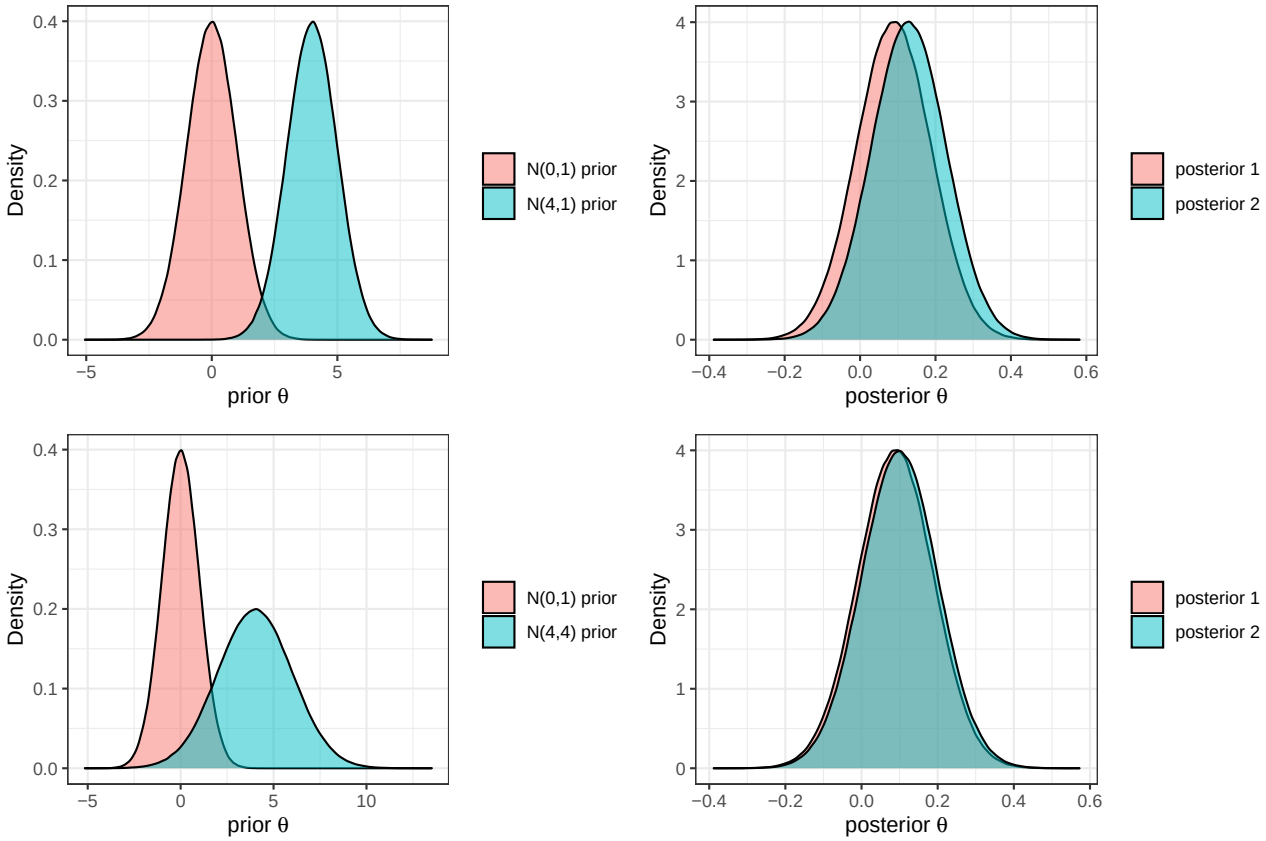

Figure 6.2: Figure showing the relationship between prior and posterior using Normal distributions. The top left panel shows a $N(0,1)$ prior (light red) and a $N(4,1)$ prior (light blue), while the top right panel shows the resulting posteriors. The bottom left panel shows a $N(0,1)$ prior (light red) and a $N(4,4)$ prior (light blue), while the top right panel shows the resulting posteriors.

$$R\hat{O}C(P_1, P_2) = \sqrt{\frac{1}{n^2}\frac{1}{N^2}\sum_{i,j=1}^{N}\left(\frac{2X_i - 2\theta_0}{\sigma_0^3}\right)k_{Gau}(X_i, X_j)\left(\frac{2X_j - 2\theta_0}{\sigma_0^3}\right)}.$$

The rate of change of the KSD with respect to σ_0^2 is shown on the top panel of Figure 6.3 while the rate of change of the KSD with respect to μ_0 is shown on the bottom panel of Figure 6.3.

Starting from the bottom panel of Figure 6.3, we can see that it matches our intuition, that is, the KSD has an almost constant rate of change as we perturb μ_0 while holding σ_0^2 constant. Furthermore the larger the value of σ_0^2, the smaller the rate of change of the KSD, since the priors start having more uncertainty, therefore tending to an uninformative prior, and thus the impact on the resulting rate of change of the KSD starts to reduce.

In the top panel of Figure 6.3, what we observe also matches our intuition. Given the prior mean is (fixed) close to the true unknown mean 0 the rate of change of the KSD when we perturb σ_0^2, is not very significant since the true unknown mean does not lie in the tails of the prior. On the other hand, when the prior mean is far from the true unknown mean, then the rate of change becomes significant, for small values of σ_0^2, since the data are not very informative and the prior distribution is extremely unfavorable. A small perturbation in the parameter space is going to translate to a relatively larger change in the posterior distributions. On the other hand, when σ_0^2 is larger the prior starts becoming less unfavorable

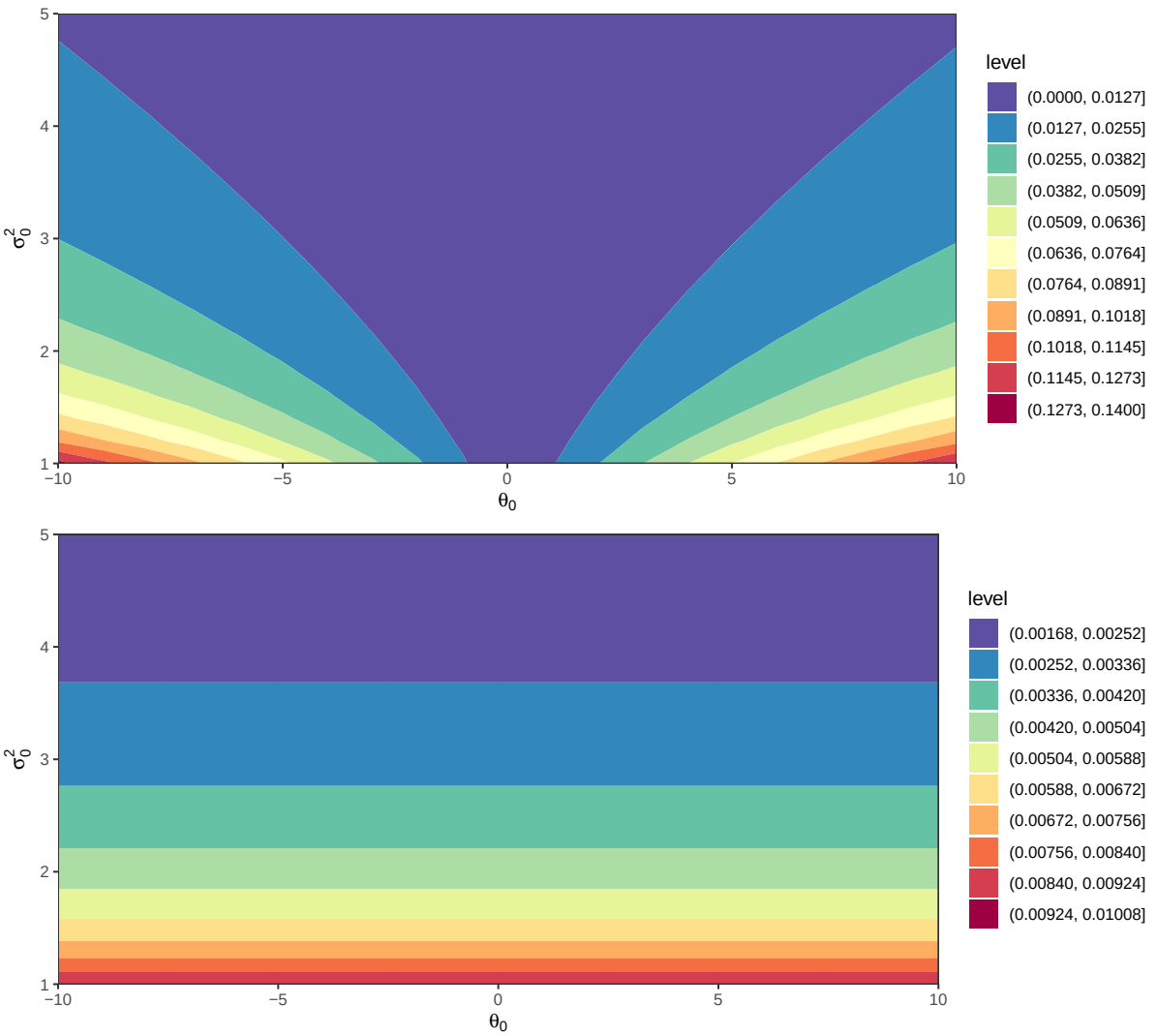

Figure 6.3: Figure showing the rate of change of the KSD with respect to σ_0^2 (upper panel) and θ_0 (lower panel) for priors with parameters indicated by every possible pair in the grid $\{\theta_0 = [-10, 10]$ by $1,\ \sigma_0 = [1, 5]$ by $0.1\}$. The true underlying distribution comes from a $N(0, 1)$ distribution, which μ is unknown.

and thus the rate of change becomes smaller.

To demonstrate this, we are going to pick two points on the upper panel of Figure 6.3, one is going to be in an area where there the rate of change of the KSD is small (blue), while the other is going to be in an area where the rate of change is high (red). Then, we are going to perturb σ_0^2 by the same value and visually confirm that the point in the red area will have a bigger impact on the posterior than the point in the blue area.

Figure 6.4 confirms what we mentioned before. In the upper left panel we have plotted a $N(3, 1)$ prior distribution (light red) and a $N(3, 2)$ prior distribution (light blue). These points belong in a smaller rate of change area in the upper panel of Figure 6.3, thus the resulting posteriors which are shown in the upper right panel of Figure 6.4 are not "far" from each other. In the bottom left panel of Figure 6.4 we have plotted a $N(10, 1)$ prior distribution (light red) and a $N(10, 2)$ prior distribution (light blue). These points belong in a high rate of change area in the upper panel of Figure 6.3, therefore the same perturbation in the parameter σ_0^2 has a higher impact on the resulting posteriors as is obvious in the bottom right panel of Figure 6.4.

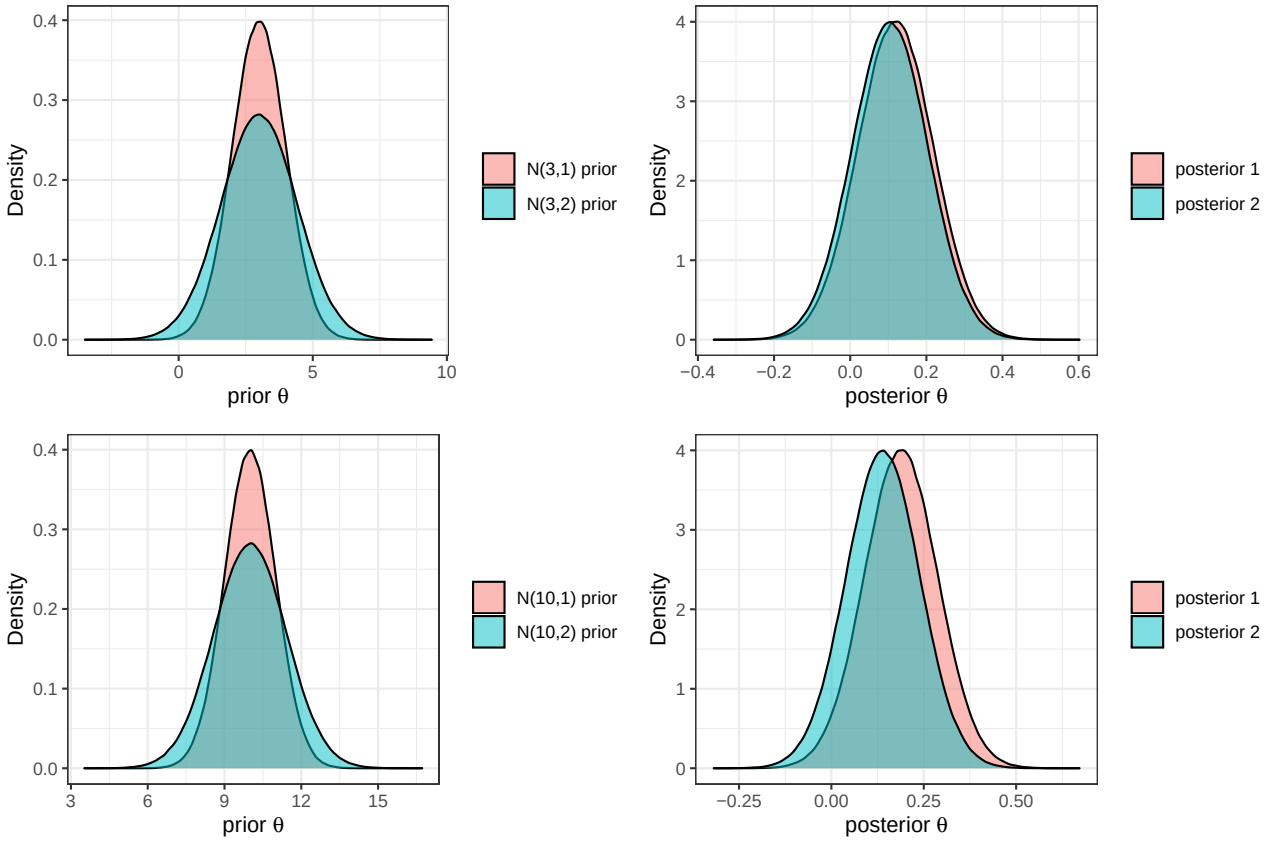

Figure 6.4: Figure confirming that points in high rate of change of the KSD areas impact the posterior more. The upper left panel shows a $N(3,1)$ prior distribution (light red) and a $N(3,2)$ (light blue) prior distribution while the upper right panel shows the resulting posteriors. The bottom left panel shows a $N(10,1)$ prior distribution (light red) and a $N(10,2)$ (light blue) prior distribution while the bottom right panel shows the resulting posteriors.

From Figure 6.3 we can conclude that if the prior of choice is a Normal distribution, then the parameter that has a larger impact is the variance σ_0^2. Thus, if the prior information for the mean is heavily against the actual mean, perturbing the variance would have considerable impact on the resulting posterior. This would indicate that the results are sensitive to the prior parameters. On the other hand if the prior information gives a mean fairly close to the actual mean, then the variance σ_0^2 starts having a minor impact as well, meaning that specifying it very accurately becomes less necessary.

6.3 Student–t Prior Versus Normal Prior

To demonstrate a major advantage of our method we are going to look at an example where the likelihood is again going to be Normal, where the parameter of interest is μ and the variance is known. Assume $\sigma^2 = 0.5$ and for the purpose of simulating $n = 100$ data points the mean would be set equal to $\mu = 2$. Furthermore, the kernel we are going to use is the polynomial kernel $k_{Poly}(x, x') = x \cdot x' + 1$. This time the two competing priors are going to be a Normal prior with $\mu_0 = 0$ and $\sigma_0^2 = 1$, resulting in posterior Q and a prior from a different parametric family, say a Student–t prior distribution. resulting in posterior P. A Student–t distribution with ν degrees of freedom has the following probability density function,

$$f(x;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \text{ for } x \in (-\infty,\infty),$$

where $\Gamma(\cdot)$ is the gamma function.

Since the Student–t distribution is not a conjugate prior of the Normal likelihood it does not have a closed form expression for the resulting posterior. Thus, one would require an expensive method such as a MCMC to calculate its posterior, as well as use any of the known metrics in the literature to statistically assess the impact of this prior against a Normal prior.

For the KSD however, we can sample from any of the two posteriors, thus we are going to pick the Normal posterior to sample from, since we can sample from it in a straightforward and not computationally demanding way. In addition to that, one would require to compute the gradient of the log of the priors' ratio which again is simple to compute. Thus, we generated $N = 10^3$ particles from Normal samples $X_1, \ldots, X_N$ from the resulting posterior for the $N(0,1)$ prior, and using Formula 5.16, we would calculate,

$$\begin{aligned}\hat{\mathbb{S}}_v(P,Q) &= \sqrt{\frac{1}{n^2}\frac{1}{N^2}\sum_{i,j=1}^{N}\nabla\log(p_1/p_2)(X_i)k_{Prod}(X_i,X_j)\nabla\log(p_1/p_2)(X_j)} \\ &= \sqrt{\frac{1}{n^2}\frac{1}{N^2}\sum_{i,j=1}^{N}\left(\frac{X_i-\mu_0}{\sigma_0^2}-\frac{(\nu+1)X_i}{X_i^2+\nu}\right)k_{Prod}(X_i,X_j)\left(\frac{X_j-\mu_0}{\sigma_0^2}-\frac{(\nu+1)X_j}{X_j^2+\nu}\right)}.\end{aligned}$$

To illustrate this, we present in Figure 6.5 the KSD between the $N(0,1)$ prior distribution and a central Student–t prior distribution with degrees of freedom $\nu = 1, 2, \ldots, 500$. As expected, as $\nu \to \infty$ the KSD converges to 0, since the Student–t prior distribution converges to $N(0,1)$.

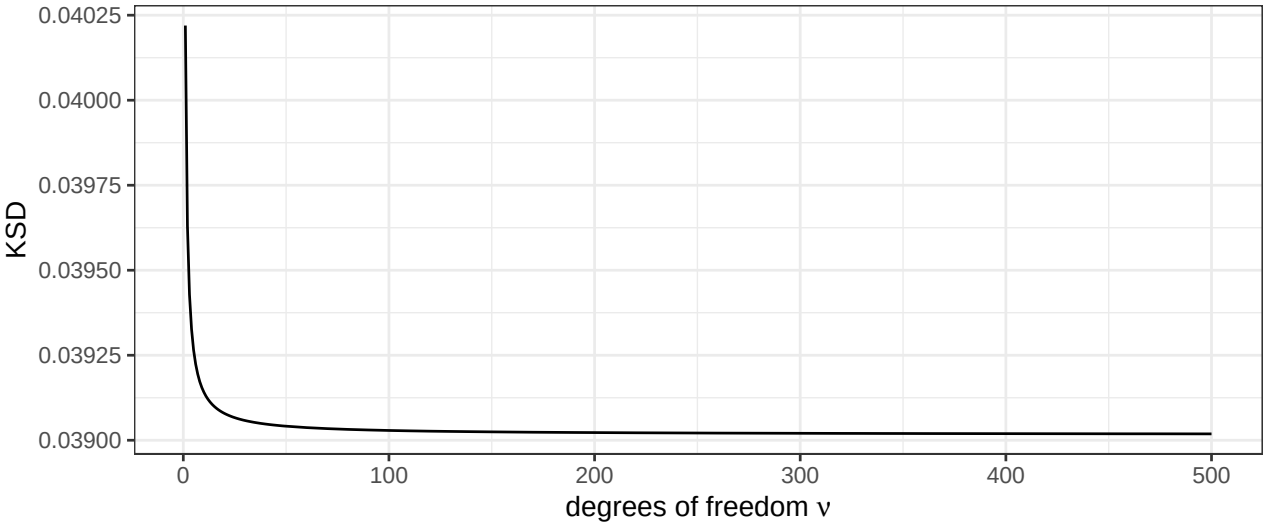

Figure 6.5: Figure showing the KSD between a $N(0,1)$ prior distribution and a Student–t prior distribution with degrees of freedom $\nu = 1, 2, \ldots, 500$. The real simulated data points come from a $N(2, 0.5)$ distribution.

6.4 Application in Variational Inference

Suppose that one wants to conduct a sequential updating of Bayesian model for unknown parameter θ. Assume that the prior distribution is $p(\theta)$, the data in hand is $X_1, \ldots, X_N$ and the model in use has already been updated on these data points, thus has a posterior distribution,

$$P(\theta; X_1, \ldots, X_N) \propto p(\theta)p(X_1, \ldots, X_N; \theta).$$

Now if the practitioners want to use a new data point X_{N+1}, to update the model they can use the old posterior as the prior and compute the new posterior as,

$$P_{new}(\theta; X_1, \ldots, X_N, X_{N+1}) \propto \underbrace{P(X_{N+1}; \theta)}_{\text{likelihood}} \underbrace{P(\theta; X_1, \ldots, X_N)}_{\text{new prior}}.$$

A typical problem is that $P(\theta; X_1, \ldots, X_N)$ is sometimes very expensive to compute, thus is one needs do a substantial amount of MCMC simulations to generate samples and marginalise out any possible nuisance variables from the model.

A common methodology is the so called variational inference, where the posterior distribution $P(\theta; X_1, \ldots, X_N)$ is replaced with an approximate model, say $Q(\theta)$. Therefore the approximate new posterior is,

$$\hat{P}_{new}(\theta; X_1, \ldots, X_N, X_{N+1}) \propto P(X_{N+1}; \theta)Q(\theta)$$

A possibility, there are of course alternatives, is the Laplacian approximation, which makes use of the first order Taylor series approximation. After picking an approximate model it is required to find a way to quantify how good of an approximation is $\hat{P}_{new}$ to P_{new}.

This can be quantified by our KSD metric by computing the discrepancy between the two posteriors mentioned above. If the KSD is high, this would indicate that the approximation is poor since the prior distribution has remained fixed, thus the only difference in the derivation is coming from the posteriors. Furthermore, one can compare different approximating models with this method by computing the KSD for any number of approximating models and finding which KSD is the lowest.

6.5 Application in Practice

The question one might have is how we can use this KSD value in practice. Reporting a KSD value by itself is not very useful, since it is merely an index and has no meaning when viewed in isolation. However, by reporting tables of KSD values or even graphs, between the main a priori belief (base prior distribution) and different competing prior distributions, one can gain meaningful insights about the robustness of the base prior distribution. Furthermore, if the competing prior is set to be a Uniform prior, one can measure the impact of the base prior on the given problem. To demonstrate how KSD can be use in practice, we are going to present a full analysis of an example.

For simplicity in the calculations we are going to assume that we can gather data coming from a Normal distribution with known variance $\sigma^2 = 2$. It is known that to get high quality data is difficult and expensive, therefore we would like to gather as little data as possible but additionally be confident that our Bayesian analysis results are going to be robust. The question then becomes, from what sample size on, does changing the prior distribution start having a lesser impact on the resulting posterior distribution.

For this problem we are going to work with the polynomial kernel $k_{Poly}(x, x') = x \cdot x' + 1$. The base prior distribution we are going to be working with is a $N(-5, 3)$. For demonstration purposes we are going to use just twelve priors. Six of which are going to be Normal distribution with parameters being every possible pair in the grid of values $\{\mu_0 = \{-8, -2\}, \sigma_0^2 = \{0.5, 1, 2\}\}$. Five of which are going to be central Student–t distributions with degrees of freedom $\nu = \{2, 5, 10, 30, 50\}$ and the final one is going to be an improper Uniform distribution. In order to show how this method works we conducted the following simulation study. We calculated the different KSD values as a function of the sample size $n = 100, 110, \ldots, 1000$ for $N = 100$ iterations. Figure 6.6 shows the average over all N replications of the different KSD values for the competing priors.

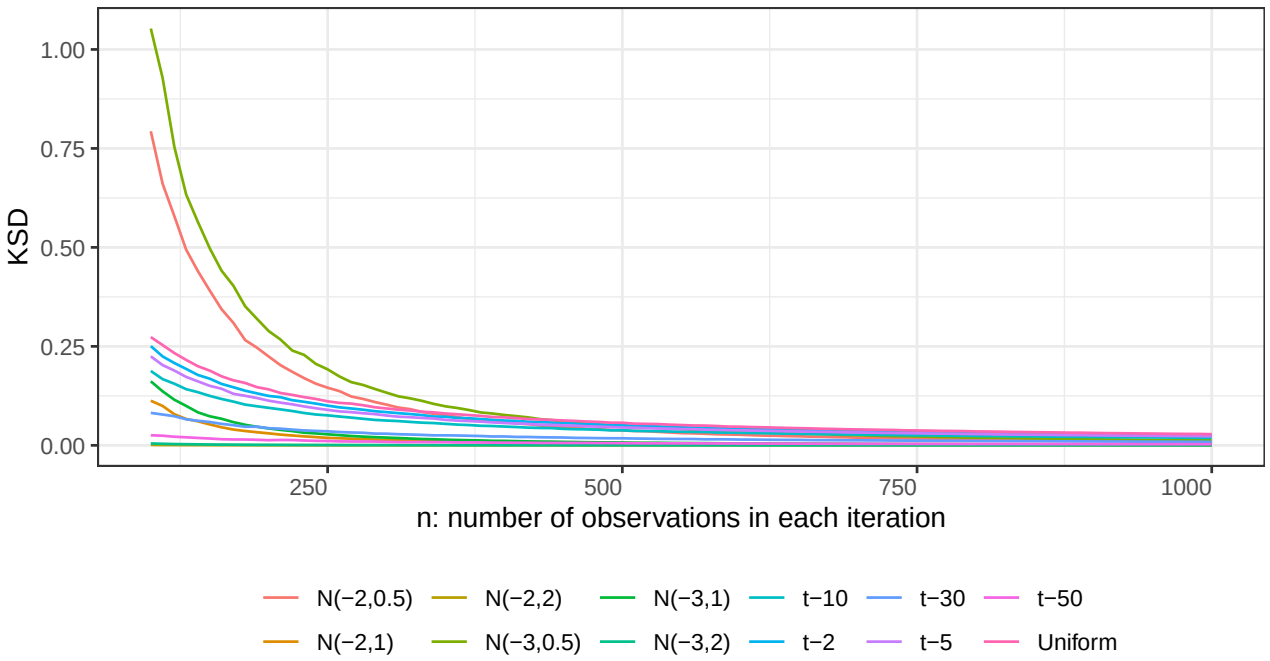

Figure 6.6: Figure showing the KSD between the base prior distribution $N(-5, 3)$ and 12 different competing prior distributions The real simulated data points come from a $N(3, 2)$.

From Figure 6.6 we can notice several interesting points. First as expected, as $n \to \infty$ the KSD converges to 0, therefore the resulting posteriors become less sensitive in the choice of prior distribution. However, in our hypothetical scenario as well as any real life application, we are interested at finite sample size n. We can see that around $n \approx 400$, the KSD values between the base prior distribution and the competing priors are getting close to each other and close to 0 as well. This would indicate that after this sample size on, our Bayesian analysis results can be considered robust. But even if we chose a value below 400, we could quantify the lack of robustness through by reporting the difference of the maximum and minimum KSD values. Additionally, in Figure 6.7 we have plotted the KSD of the base prior against just an improper Uniform in order to visually investigate the base prior impact against a posterior with only the data. Similarly, we can see how fast the impact of the prior goes to 0 as $n \to \infty$.

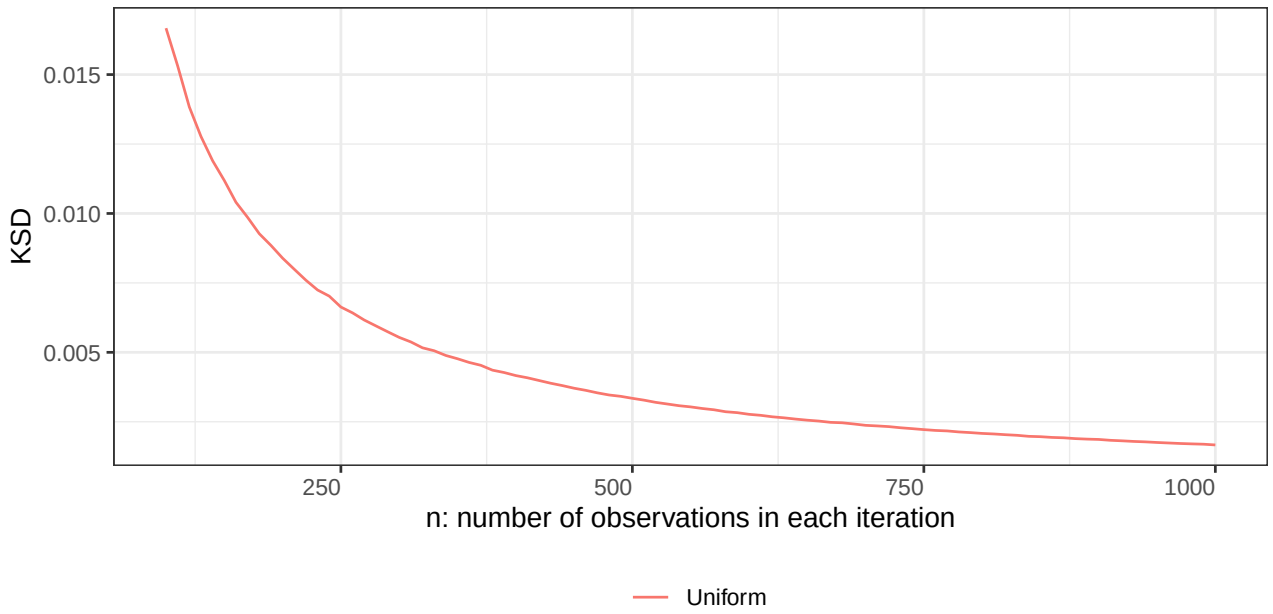

Figure 6.7: Figure showing the KSD between the base prior distribution $N(-5, 3)$ and a Uniform competing prior distribution. The real simulated data points come from a $N(3, 2)$.

Plots like these convey the required information in an understandable, concise manner. It can help practitioners inspect if any prior has a high influence and take it into consideration before presenting their results, without having to compute every posterior distribution. In addition, it allows to visually inspect from what sample size on the effect of a prior starts becoming unimportant. This has practical significance particularly when one has to choose between a more complicated prior distribution and a simpler, closed–form one.

6.6 Numerical Comparisons

In this Section we are going to investigate the performance of our method against the result from Theorem 3.2.1.

6.6.1 Rate of Decay of the KSD

In this subsection we are going to show some numerical comparisons between the KSD for different kernels, so as to demonstrate robustness of the method as well as visually inspect the decay of the KSD as a function of the sample size n.

Looking at the bound in (3.14) where we used Theorem 3.2.1, one can see that the bounds are of order n^{-1}, thus we have an exponential decay, which in a $\log_e$-$\log_e$ plot should give a straight line. To demonstrate that the KSD also has an exponential decay we are going to use the following example. Assume we have data coming from a Normal distribution where μ is unknown and $\sigma^2 = 1$. To demonstrate the rate of decay we will generate n samples, for different values of n using $\mu = 0$. Assume also that we have two prior distributions,

one is a Normal with $\mu_0 = 5$ and $\sigma_0^2 = 1$ and the other is an improper Uniform distribution. The kernels we will be using for the following experiment are: Gaussian (with lengthscale parameter l optimized by the Gradient Ascent algorithm), Gaussian times Polynomial (with $c = 1$, $d = 1$ and l chosen as before), IMQ (with $c = 1$, $\beta = -1/2$), Polynomial ($c = 1$, $d = 1$).

As previously we are going to generate $N = 100$ samples for each sample size $n = 100, 200, \ldots, 10000$, and using $N_p = 1000$ particles we will calculate the KSD between the two posterior distributions and report the average over all N iterations. The KSD will be calculated using Formula 5.14,

$$\hat{\mathbb{S}}_v(P, Q) = \sqrt{\frac{1}{n^2}\frac{1}{N_p^2}\sum_{i,j=1}^{N_p} \nabla \log(p_1/p_2)(X_i) k(X_i, X_j) \nabla \log(p_1/p_2)(X_j)}$$
$$= \sqrt{\frac{1}{n^2}\frac{1}{N_p^2}\sum_{i,j=1}^{N_p} \left(\frac{X_i - \mu_0}{\sigma_0^2}\right) k(X_i, X_j) \left(\frac{X_j - \mu_0}{\sigma_0^2}\right)}.$$

Using the same setting as before we are going to calculate the lower bound of (3.14) that is, $\frac{\sigma^2}{n\sigma_0^2 + \sigma^2}|\bar{X} - \mu_0|$ and report the average over all N iterations. We have plotted a $\log_e - \log_e$ plot of these values in Figure 6.8. Note that, for demonstration purposes, we have jittered the values for the Gaussian, IMQ, Polynomial kernels and the Lower Bound from (3.14) because they were on top of each other.

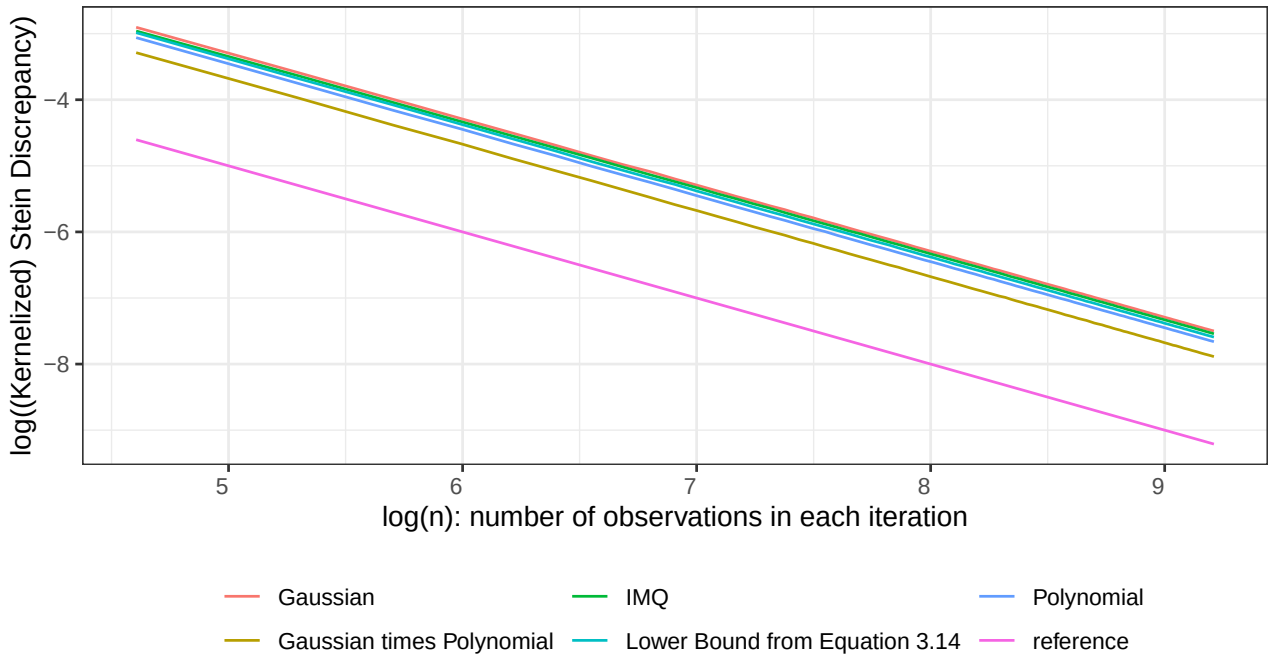

Figure 6.8: Figure showing the KSD values for several kernel and the Lower bounds of 3.14 for the Normal likelihood, Normal prior setting in a $\log_e - \log_e$ plot. There is also a 1/n reference line to compare the decays.

Two things can be concluded from Figure 6.8. First, the KSD has an exponential decay as well, since the $\log_e$ KSD values lie in a straight line being parallel to the $1/n$ reference line.

In addition, the method seems to be robust to the different kernels used as every kernel plotted here gives roughly the same value, which is a desirable property for this method since we would not want to have a method that would depend on the kernel. Interestingly, the Gaussian times Polynomial kernel returns a lower value than any of the kernels used.

6.6.2 Comparison with the Wasserstein Metric

In this subsection, which is motivated by [23] we are going to compare the bound of the Wasserstein distance from Theorem 3.2.1 to our KSD metric using the kernel $k_{Poly}(x, x') = x \cdot x' + 1$ and find out how close the upper and lower bounds are to the KSD as well as notice any differences or similarities so as to compare these two prior impact measures. The two examples we are going to look are priors for the success parameter of a Binomial distribution and priors for the rate parameter of a Poisson distribution.

6.6.2.1 Comparison on a Binomial Model

First we are going to consider the Binomial distribution $Bin(n, \theta)$, where $n \in \mathbb{N}$ is the number of trials and $\theta \in [0, 1]$ is the success parameter. The probability of getting exactly x successes is the probability mass function,

$$f(x; n, \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x},$$

for $x = 1, 2, \dots n$. The conjugate prior for the Binomial distribution is the Beta distribution $B(\alpha, \beta)$, where both $\alpha, \beta > 0$. Special cases of the $B(\alpha, \beta)$ include the Uniform prior ($\alpha = \beta = 1$), the Jeffreys prior ($\alpha = \beta = 1/2$) and the Haldane's prior $\alpha = \beta = 0$, [19]. Using Theorem 3.2.1 we can find lower and upper bounds of the Wasserstein distance between the posterior resulting from either Jeffreys or Haldane's prior against the posterior resulting from a Uniform prior. The detailed derivation for the two cases is presented in [22, 38].

First for the Jeffreys prior against a Uniform prior we have:

$$\frac{|\frac{n}{2} - x|}{(n+2)(n+1)} \leq d_W(P_1, P_2) \leq \frac{1}{n+2} \left(\sqrt{\frac{(x+1/2)(n-x+1/2)}{(n+2)(n+1)^2}} + \left| \frac{x+1/2}{n+1} - \frac{1}{2} \right| \right).$$

Next, for the Haldane's prior against a Uniform prior we have:

$$\frac{2|\frac{n}{2} - x|}{n(n+2)} \leq d_W(P_1, P_2) \leq \frac{2}{n+2} \left(\sqrt{\frac{x(n-x)}{n^2(n+1)}} + \left| \frac{x}{n} - \frac{1}{2} \right| \right).$$

We calculate the upper and lower bounds as well as the KSD values for $n = 100, 200$ and $\theta = 0.05, 0.1, \dots, 0.95$. For the upper and lower bounds we generated $N = 1000$ samples from the Binomial distribution using the different parameter pairs and we have plotted the average

of these in the top panel of Figures 6.9 for the Jeffreys prior and 6.10 for the Haldane's prior. As for the KSD, we have calculated $N = 1000$ posterior samples for the two different priors, for every pair of parameters mentioned above. Then, using $N_p = 10^3$ particles we calculate the KSD between the posterior resulting from a Uniform and the posterior resulting from either Jeffreys or Haldane's prior and report the average over all N iterations. To calculate the KSD between any two general Beta priors $B(\alpha_i, \beta_i), i = 1, 2$, we used Formula 5.14,

$$\begin{aligned}\hat{\mathbb{S}}_v(P, Q) &= \sqrt{\frac{1}{n^2}\frac{1}{N_p^2}\sum_{i,j=1}^{N_p} \nabla \log(p_1/p_2)(X_i) k_{Poly}(X_i, X_j) \nabla \log(p_1/p_2)(X_j)} \\ &= \sqrt{\frac{1}{n^2}\frac{1}{N_p^2}\sum_{i,j=1}^{N_p} \left(\frac{\alpha_1 - \alpha_2}{X_i} + \frac{\beta_2 - \beta_1}{1 - X_i}\right) k_{Poly}(X_i, X_j) \left(\frac{\alpha_1 - \alpha_2}{X_j} + \frac{\beta_2 - \beta_1}{1 - X_j}\right)}.\end{aligned}$$

The average of these values can be seen in the bottom panel of Figures 6.9 for the Jeffreys prior and 6.10 for the Haldane's prior.

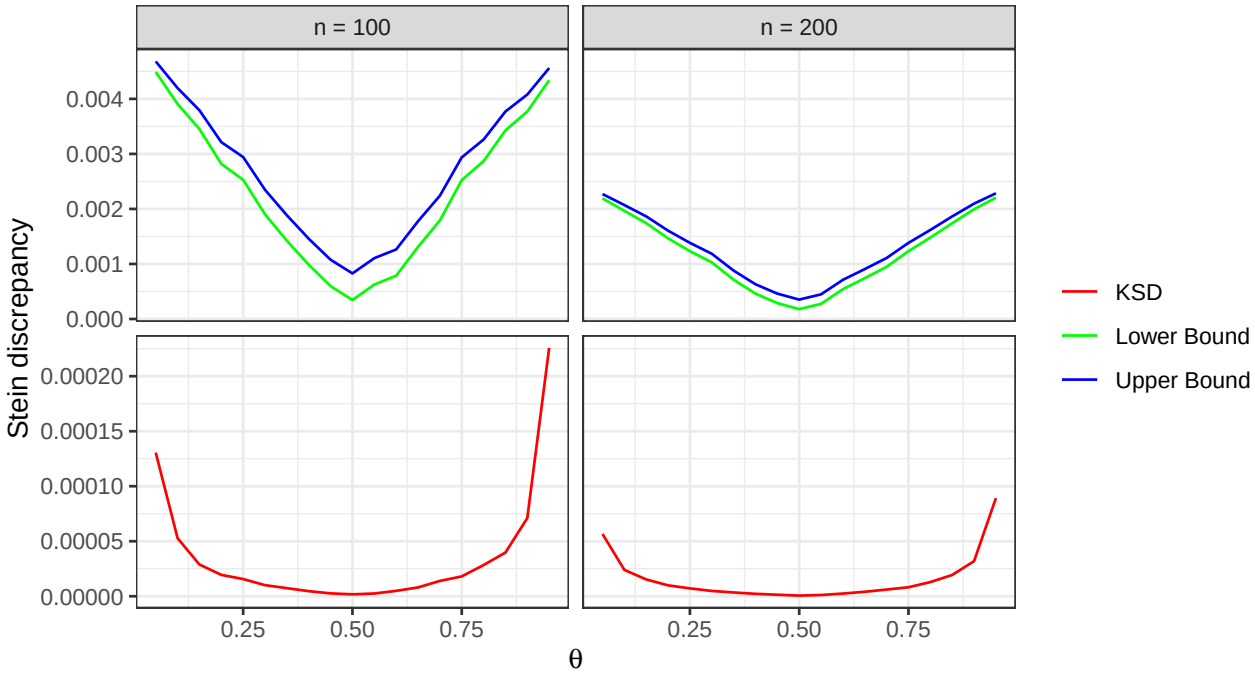

Figure 6.9: Figure showing the lower and upper bounds (top panel) as well as the KSD (bottom panel) for the Binomial distribution for posteriors based on Jeffreys prior against Uniform priors. The values for the unknown parameter used are $n = 100, 200$ and $\theta = 0.05, 0.1, \ldots, 0.95$.

As expected, both metrics have a similar pattern, they decrease towards the central value $\theta = 0.5$ and increase towards the edges for both priors and both sample sizes tested. This indicates that both metrics capture the dissimilarity of the two posteriors in a similar manner. We notice however that although both metrics have small values they are not in the same scale. More specifically, the KSD values have lower values than both the upper and lower bounds. Differences like this ought to be expected since the KSD is defined in the RKHS space while the Wasserstein distance is defined in the Lipschitz-1 class.

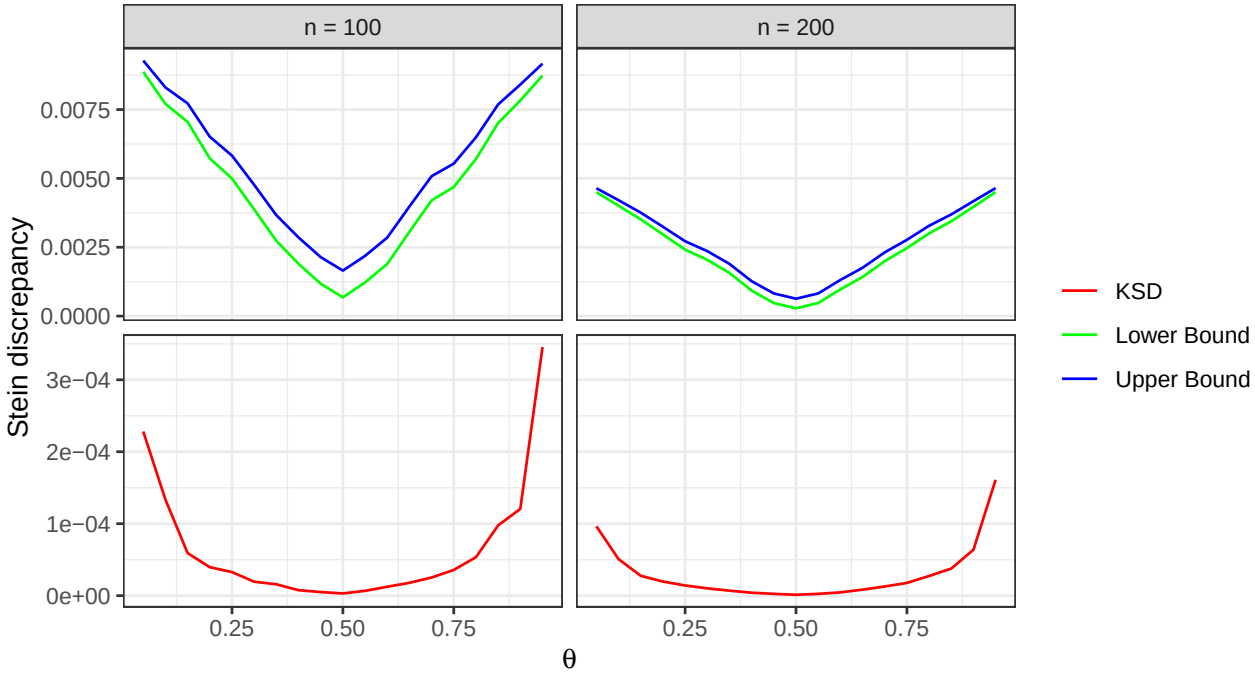

Figure 6.10: Figure showing the lower and upper bounds (top panel) as well as the KSD (bottom panel) for the Binomial distribution for posteriors based on Haldane's prior against Uniform priors. The values for the unknown parameter used are $n = 100, 200$ and $\theta = 0.05, 0.1, \ldots, 0.95$.

6.6.2.2 Comparison on a Poisson Model

The final distribution we are going to be considering is the Poisson $P(\lambda)$, with parameter $\lambda > 0$. The probability mass function is given by,

$$f(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!},$$

for $x = 0, 1, 2, \ldots$. The conjugate prior for the Poisson distribution is the Gamma distribution with shape and scale parameters α and β, where $\alpha, \beta > 0$. In this case Theorem 3.2.1 gives an exact formula for the Wasserstein distance when one of two scenarios hold $(1) : \alpha_1 < \alpha_2$ and $\beta_1 > \beta_2$ or $(2) : \alpha_1 > \alpha_2$ and $\beta_1 < \beta_2$, [23]. The detailed derivation is presented in [22],

$$d_W(P_1, P_2) = \frac{1}{n + \beta_1} \left| \alpha_2 - \alpha_1 - (\beta_2 - \beta_1) \frac{\alpha_2 + \sum_{i=1}^{n} x_i}{n + \beta_2} \right|,$$

where $x_i, i = 1, \ldots, n$ are the data in hand. For this experiment we are fixing the base prior to a Gamma with parameters $\alpha_1 = 4, \beta_1 = 0.5$ and the second prior to a Gamma with parameters α_2, β_2, where $\alpha_2 \in \{0.5, 3.5\}$ by 0.15 and $\beta_2 \in \{2, 4\}$ by 0.1. In addition we are using $n = 200$ as the sample size and $\lambda = 1, 5, 10$. Then, as before we generate $N = 1000$ random samples for every combination of parameter values $n, \lambda, \alpha_1, \beta_1, \alpha_2, \beta_2$, calculate the exact Wasserstein distance and compute the average over all N iterations. Similarly for the KSD, as before we generate $N = 1000$ posterior samples for every posterior resulting from the priors mentioned above. Then, using $N_p = 10^3$ particles we compute the KSD between the posterior resulting from the base prior and the posterior resulting from every prior mentioned

above and calculate the average KSD over all N iterations. To calculate the KSD between any two general Gamma priors $Gamma(\alpha_i, \beta_i), i = 1, 2$, we used Formula 5.14,

$$\hat{\mathbb{S}}_v(P, Q) = \sqrt{\frac{1}{n^2}\frac{1}{N_p^2}\sum_{i,j=1}^{N_p} \nabla \log(p_1/p_2)(X_i) k_{Poly}(X_i, X_j) \nabla \log(p_1/p_2)(X_j)}$$
$$= \sqrt{\frac{1}{n^2}\frac{1}{N_p^2}\sum_{i,j=1}^{N_p} \left(\frac{\alpha_1 - \alpha_2}{X_i} + \beta_2 - \beta_1\right) k_{Poly}(X_i, X_j) \left(\frac{\alpha_1 - \alpha_2}{X_j} + \beta_2 - \beta_1\right)}.$$

The difference between these values is plotted in Figure 6.11.

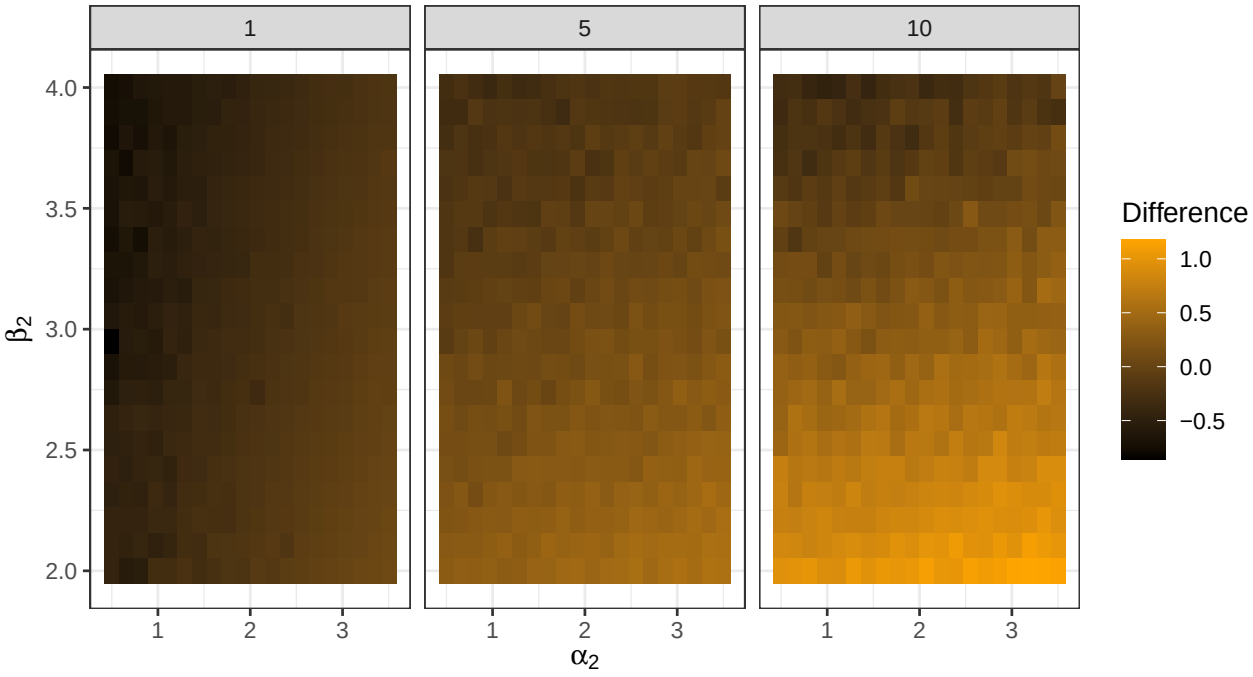

Figure 6.11: Figure showing the difference between the actual Wasserstein distance and the KSD for the Poisson distribution using $Gamma(4, 0.5)$ as a base prior and $Gamma(\alpha_2, \beta_2)$ as the second prior, where $\alpha_2 \in \{0.5, 3.5\}$ by 0.15 and $\beta_2 \in \{2, 4\}$ by 0.1. Additionally, the sample size used is $n = 200$ and the values for the unknown parameter used are $\lambda = 1, 5, 10$.

From Figure 6.11 we conclude that both metrics are quite close to each other with the median difference being 0.03. In addition as λ increases we start to notice some larger differences especially as the parameters for the second prior get further from the base prior parameters $\alpha_1 = 4, \beta = 0.5$. Again, differences like this ought to be expected because the metrics are defined in different spaces.

In this Section we carried out some numerical simulations to investigate the behaviour of our metric against the results from Theorem 3.2.1. We can be confident that in these examples, both methods return similar results and have similar behaviour.

Chapter 7

Conclusion

7.1 Summary

In this thesis we have investigated how Stein's method can be used in order to conduct sensitivity analysis and assess the impact of the prior distribution in Bayesian Statistics. Our first contribution was to establish a way to extend the Kernelized Stein Discrepancy introduced in [42], to the Bayesian framework. Our approach allows the usage of unnormalised posterior distribution which are common in modern application, thus making it compatible with MCMC techniques. Additionally, this method can be used on higher dimensions of the sample space as well as any number of dimensions in the parameter space, provided we can sample from one of the two posteriors we are comparing. Furthermore, we have introduced the rate of change of the Kernelized Stein Discrepancy with respect to the parameters of the prior parameter space, provided the priors belong in the same parametric family. The rate of change can be used to measure the sensitivity of the chosen prior when its parameters are perturbed.

Through a series of examples we demonstrated how this method can be used in practice. First, one can pick a base prior distribution and measure the KSD between any number of competing priors so as to check robustness and examine which, if any, priors give significantly different posterior results. In addition, if all the priors belong to the same parametric family, one can straightforwardly check if the base prior is sensitive to perturbation around the chosen prior parameter values. If that is the case, this would indicate that it is necessary for the prior parameters to be very accurately chosen in order to obtain accurate results.

We can confidently say that both the KSD index is an attractive alternative to other existing methods such as Mean Observed Prior Effective Sample Size (MOPESS), Neutrality, or the Wasserstein bounds from Theorem 3.2.1. The reasons for that is that (1) it does not require the specification of the normalisation constant for the posterior, unlike most methods. (2) It does not require the evaluation of the likelihood which is critical for big data scenarios. (3) It can be used in multiple dimensions both in the parameter space and in the sample space. (4) It has the flexibility of choosing any of the two competing posterior distributions to generate samples from. (5) Finally, it is easily explainable.

Additionally, the rate of change of the KSD is also an attractive alternative to Circular sensitivity or the PS measure for the same reasons mentioned above.

Finally, we note that our KSD index as well as the rate of change of the KSD have "universal usage", meaning that, they can be used in most of the problems in practice and this is also something that we could work on in the future; to apply our metric to various real data sets and demonstrate its usage.

7.2 Limitations

Although we have advertised that the two metrics can work without normalized posterior densities, they require samples from these unnormalized densities. In many cases, to generate samples from any of the two posterior densities would require implementing an MCMC algorithm which can get computationally expensive, but this is something that, as of now, no known method to us can solve.

Another disadvantage of our method is that the KSD metric is a comparative measure. When seen in isolation it does not provide much information about the discrepancy between the two posteriors under comparison, unlike the Neutrality index which is an absolute measure. In practice we would need to specify a base prior and calculate the KSD between it and a set of competing priors to get meaningful insights on the prior robustness and impact of certain prior distributions.

A final limitation of the KSD is that the computational cost grows at least quadratically in the particle sample size, since we are using the $V-$statistic which uses a double sum, as an estimate. Since these are asymptotic results, if a large particle sample size is required this can make it prohibitive. In addition, the base KSD especially with the Gaussian kernel, suffers from the curse of dimensionality, [25]. There are however ways to get over these issues which we are going to mention in the next Section.

7.3 Future Work

To solve the high computational burden introduced from high sample sizes, we need to use another estimator for the KSD which is not quadratic in time to estimate. There have been introduced several options in the literature, such as replacing the $U-$ statistic in (5.8) with a running average which is introduced in [42], or using the *Finite Set Stein Discrepancies (FSSD)* introduced in [31]. This essentially is a measure depending on a set of features, typically small, that are used to evaluate the *Stein witness function*. However, these methods require more investigation in our framework since the sliced Stein discrepancies have only been used for goodness–of–ft tests.

Another method we can use to reduce the complexity to linear is the *Random Features* introduced in [49]. The main idea of Random Features is to map the data into a low–dimensional Euclidean inner product space with "randomized feature map" in a way that the inner product between the pairs of of transformed points approximates their kernel evaluation. The authors in [49] propose Fourier Random Features which project the data into a line and then transforms the result with a sinusoid.

To solve the issue of dimensionality, the authors in [25] introduce the *sliced Stein discrep-*

ancies, where they project the score function as well as the test inputs in various "one dimensional slicing directions". This results in a different form of the Stein discrepancy which depends only one–dimensional inputs for the test functions. This method again would require further investigation in out framework, as it was also developed for goodness–of–fit tests

An interesting development to the KSD metric would be to use the Diffusion Kernelized Stein Discrepancy (DKSD) introduced in [3]. There the authors consider a vector valued RKHS which we defined in Section 4.3 mapping from some non empty set $\mathcal{X}$ to $\mathbb{R}^d$ and matrix-valued kernels $k : \mathcal{X} \times \mathcal{X} \to R^{d\times d}$ and use a *diffusion matrix* which is an arbitrary $m \in \mathbb{R}^{d\times d}$. Then the authors define the DKSD between distributions P and Q with densities p and q as,

$$DKSD(P, Q) = \mathbb{E}_{x,x'\sim q} k^0(x, x'),$$

where $k^0(x, x') = \frac{1}{p(x)p(x')} \nabla_x \nabla_{x'} \left(p(x)m(x)K(x, x')m(x')^\intercal p(x')\right)$. DKSD can be approximated with the $U-$statistic:

$$\hat{DKSD}(P, Q) = \frac{1}{n(n-1)} \sum_{i\neq j} k^0(X_i, X_j),$$

which they have proved has certain desirable properties.

In addition, we remind the reader that to solve some of the issues of the other metrics, it was required to define the Stein set $\mathcal{H}$ to be the RKHS. However, this is not the only interesting set we could have investigated. Another possibility would be to use the Fourier basis function which gives rise to the so called *Fourier features*. We could also look into a few of these basis and study them in detail since each one would have different properties.

Finally, in order for this work to be helpful to practitioners it would be useful to create an R package for some of the most common prior-likelihood settings, where it would automatically provide diagnostic plots of prior robustness and sensitivity analysis.

Bibliography

[1] Mauricio A Alvarez, Lorenzo Rosasco, and Neil D Lawrence. Kernels for vector-valued functions: A review. *arXiv preprint arXiv:1106.6251*, 2011. pages 20

[2] Andreas Anastasiou, Alessandro Barp, François-Xavier Briol, Bruno Ebner, Robert E Gaunt, Fatemeh Ghaderinezhad, Jackson Gorham, Arthur Gretton, Christophe Ley, Qiang Liu, et al. Stein's method meets statistics: A review of some recent developments. *arXiv preprint arXiv:2105.03481*, 2021. pages 11, 12

[3] Alessandro Barp, Francois-Xavier Briol, Andrew B Duncan, Mark Girolami, and Lester Mackey. Minimum stein discrepancy estimators. *arXiv preprint arXiv:1906.08283*, 2019. pages 20, 47

[4] Sanjib Basu, Sreenivasa Rao Jammalamadaka, and Wei Liu. Local posterior robustness with parametric priors: maximum and average sensitivity. In *Maximum Entropy and Bayesian Methods*, pages 97–106. Springer, 1996. pages 7, 32

[5] James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013. pages 3

[6] 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. pages 6, 7, 8

[7] L Mark Berliner and Steven N MacEachern. Examples of inconsistent bayes procedures based on observations on dynamical systems. *Statistics & probability letters*, 17(5):355–360, 1993. pages 6

[8] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011. pages 20

[9] Julian Besag. Digital image processing: Towards bayesian image analysis. *Journal of Applied statistics*, 16(3):395–407, 1989. pages 3

[10] George EP Box and George C Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011. pages 6

[11] Lawrence J Brunner and Albert Y Lo. Bayes methods for a symmetric unimodal density and its mode. *The Annals of Statistics*, pages 1550–1566, 1989. pages 6

[12] Anirban DasGupta. *Asymptotic theory of statistics and probability*. Springer Science & Business Media, 2008. pages 25

[13] A Philip Dawid. Posterior expectations for large observations. *Biometrika*, 60(3):664–667, 1973. pages 6

[14] Sarah Depaoli, Sonja D Winter, and Marieke Visser. The importance of prior sensitivity analysis in bayesian statistics: Demonstrations using an interactive shiny app. *Frontiers in Psychology*, 11, 2020. pages 3, 4, 6

[15] Persi Diaconis and David Freedman. On inconsistent bayes estimates of location. *The Annals of Statistics*, pages 68–87, 1986. pages 4

[16] Persi Diaconis and David Freedman. On the consistency of bayes estimates. *The Annals of Statistics*, pages 1–26, 1986. pages 4, 6

[17] Persi Diaconis, Charles Stein, Susan Holmes, and Gesine Reinert. Use of exchangeable pairs in the analysis of simulations. In *Stein's Method*, pages 1–25. Institute of Mathematical Statistics, 2004. pages 11

[18] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso Poggio. Learning with a wasserstein loss. *arXiv preprint arXiv:1506.05439*, 2015. pages 10

[19] Seymour Geisser. On prior distributions for binary trials. *The American Statistician*, 38(4):244–247, 1984. pages 41

[20] Fatemeh Ghaderinezhad. New insights into the impact of the choice of the prior for the success parameter of binomial distributions. In *7th Annual International Conference on Computational Mathematics, Computational Geometry & Statistics (CMCGS 2018)*, 2018. pages 4

[21] Fatemeh Ghaderinezhad and Christophe Ley. On the impact of the choice of the prior in bayesian statistics. In *Bayesian Inference on Complicated Data*. IntechOpen, 2019. pages 11, 16

[22] Fatemeh Ghaderinezhad and Christophe Ley. Quantification of the impact of priors in bayesian statistics via stein's method. *Statistics & Probability Letters*, 146:206–212, 2019. pages 15, 41, 43

[23] Fatemeh Ghaderinezhad, Christophe Ley, and Ben Serrien. The wasserstein impact measure (wim): a generally applicable, practical tool for quantifying prior impact in bayesian statistics. *arXiv preprint arXiv:2010.12522*, 2020. pages 9, 10, 14, 16, 41, 43

[24] Yoav Goldberg and Michael Elhadad. splitsvm: fast, space-efficient, non-heuristic, polynomial kernel computation for nlp applications. In *Proceedings of ACL-08: HLT, Short Papers*, pages 237–240, 2008. pages 30

[25] Wenbo Gong, Yingzhen Li, and José Miguel Hernández-Lobato. Sliced kernelized stein discrepancy. *arXiv preprint arXiv:2006.16531*, 2020. pages 46

[26] Jackson Gorham. *Measuring sample quality with Stein's method*. Stanford University, 2017. pages 12, 23

[27] Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR, 2017. pages 30

[28] Arthur Gretton. Introduction to rkhs, and some simple kernel algorithms. *Adv. Top. Mach. Learn. Lecture Conducted from University College London*, 16:5–3, 2013. pages 18, 19, 30

[29] Daniel F Heitjan, Mengye Guo, Riju Ray, E Paul Wileyto, Leonard H Epstein, and Caryn Lerman. Identification of pharmacogenetic markers in smoking cessation therapy. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 147(6):712–719, 2008. pages 3

[30] Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005. pages 27

[31] Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of-fit test. *arXiv preprint arXiv:1705.07673*, 2017. pages 46

[32] David E Jones, Robert N Trangucci, and Yang Chen. Quantifying observed prior impact. *Bayesian Analysis*, 1(1):1–28, 2021. pages 8

[33] Jouni Kerman. Neutral noninformative and informative conjugate beta and gamma prior distributions. *Electronic Journal of Statistics*, 5:1450–1470, 2011. pages 9

[34] Joel Klipfel. A brief introduction to hilbert space and quantum logic. 2009. pages 18, 19

[35] Soheil Kolouri, Gustavo K Rohde, and Heiko Hoffmann. Sliced wasserstein distance for learning gaussian mixture models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3427–3436, 2018. pages 10

[36] Paul C Lambert, Alex J Sutton, Paul R Burton, Keith R Abrams, and David R Jones. How vague is vague? a simulation study of the impact of the use of vague prior distributions in mcmc using winbugs. *Statistics in medicine*, 24(15):2401–2428, 2005. pages 3

[37] Tom Leonard. Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(2):113–132, 1978. pages 6

[38] Christophe Ley, Gesine Reinert, and Yvik Swan. Distances between nested densities and a measure of the impact of the prior in bayesian statistics. *The Annals of Applied Probability*, 27(1):216–241, 2017. pages 4, 9, 10, 11, 14, 15, 16, 41

[39] Christophe Ley, Gesine Reinert, and Yvik Swan. Stein's method for comparison of univariate distributions. *Probability Surveys*, 14:1–52, 2017. pages 10, 11

[40] Christophe Ley and Yvik Swan. Stein's density approach and information inequalities. *Electronic Communications in Probability*, 18:1–14, 2013. pages 24

[41] Qiang Liu. A short introduction to kernelized stein discrepancy. pages 22, 23

[42] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR, 2016. pages 17, 22, 23, 24, 25, 45, 46

[43] Charles A Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural computation*, 17(1):177–204, 2005. pages 21

[44] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. pages 12

[45] Ulrich K Müller. Measuring prior sensitivity and prior informativeness in large bayesian models. *Journal of Monetary Economics*, 59(6):581–597, 2012. pages 7

[46] Ranjini Natarajan and Charles E McCulloch. Gibbs sampling with diffuse proper priors: A valid approach to data-driven inference? *Journal of Computational and Graphical Statistics*, 7(3):267–277, 1998. pages 3

[47] CJ Pérez, Jacinto Martín, and María Jesús Rufo. Mcmc-based local parametric sensitivity estimations. *Computational Statistics & Data Analysis*, 51(2):823–835, 2006. pages 8

[48] M Plummer. Local sensitivity in bayesian graphical models. *URL http://www-ice. iarc. fr/˜ martyn/papers/sensitivity. ps*, 327, 2001. pages 8

[49] Ali Rahimi, Benjamin Recht, et al. Random features for large-scale kernel machines. In *NIPS*, volume 3, page 5. Citeseer, 2007. pages 46

[50] Matthew Reimherr, Xiao-Li Meng, and Dan L Nicolae. Being an informed bayesian: Assessing prior informativeness and prior likelihood conflict. *arXiv preprint arXiv:1406.5958*, 2014. pages 8

[51] Małgorzata Roos, Thiago G Martins, Leonhard Held, and Håvard Rue. Sensitivity analysis for bayesian hierarchical models. *Bayesian Analysis*, 10(2):321–349, 2015. pages 7, 8

[52] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000. pages 10

[53] Bernhard Schölkopf, Alexander J Smola, Francis Bach, et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002. pages 30

[54] Robert J Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009. pages 25

[55] Alex J Smola, Bernhard Schölkopf, and Klaus-Robert Müller. The connection between regularization operators and support vector kernels. *Neural networks*, 11(4):637–649, 1998. pages 30

[56] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*, pages 583–602. University of California Press, 1972. pages 10, 13

[57] Charles Stein. Approximate computation of expectations. IMS, 1986. pages 13

[58] Rens van de Schoot, Marit Sijbrandij, Sarah Depaoli, Sonja D Winter, Miranda Olff, and Nancy E Van Loey. Bayesian ptsd-trajectory analysis with informed priors based on a systematic literature search and expert elicitation. *Multivariate Behavioral Research*, 53(2):267–291, 2018. pages 3

[59] Peng Xie, Jason H Li, Xinming Ou, Peng Liu, and Renato Levy. Using bayesian networks for cyber security analysis. In *2010 IEEE/IFIP International Conference on Dependable Systems & Networks (DSN)*, pages 211–220. IEEE, 2010. pages 3

[60] Yunan Yang, Björn Engquist, Junzhe Sun, and Brittany F Hamfeldt. Application of optimal transport and the quadratic wasserstein metric to full-waveform inversion. *Geophysics*, 83(1):R43–R62, 2018. pages 10

[61] Arnold Zellner. Applications of bayesian analysis in econometrics. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2):23–34, 1983. pages 3

[62] Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of computational and Applied Mathematics*, 220(1-2):456–463, 2008. pages 23

[63] Hongtu Zhu, Joseph G Ibrahim, and Niansheng Tang. Bayesian influence analysis: a geometric approach. *Biometrika*, 98(2):307–323, 2011. pages 8