Received: 28 April 2021 | Accepted: 25 January 2022

DOI: 10.1111/rssb.12500

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY Series B Statistical Methodology

Robust generalised Bayesian inference for intractable likelihoods

Takuo Matsubara[1,2] | **Jeremias Knoblauch**[2,3] | **François-Xavier Briol**[2,3] | **Chris J. Oates**[1,2]

[1]Newcastle University, Newcastle upon Tyne, UK

[2]The Alan Turing Institute, London, UK

[3]University College London, London, UK

Correspondence
Takuo Matsubara, Newcastle University, Newcastle upon Tyne, UK.
Email: tmatsubara@turing.ac.uk

Abstract

Generalised Bayesian inference updates prior beliefs using a loss function, rather than a likelihood, and can therefore be used to confer robustness against possible mis-specification of the likelihood. Here we consider generalised Bayesian inference with a Stein discrepancy as a loss function, motivated by applications in which the likelihood contains an intractable normalisation constant. In this context, the Stein discrepancy circumvents evaluation of the normalisation constant and produces generalised posteriors that are either closed form or accessible using the standard Markov chain Monte Carlo. On a theoretical level, we show consistency, asymptotic normality, and bias-robustness of the generalised posterior, highlighting how these properties are impacted by the choice of Stein discrepancy. Then, we provide numerical experiments on a range of intractable distributions, including applications to kernel-based exponential family models and non-Gaussian graphical models.

KEYWORDS

intractable likelihood, kernel methods, robust statistics, Stein's method

1 | INTRODUCTION

A considerable proportion of statistical modelling deviates from the idealised approach of fine-tuned, expertly crafted descriptions of real-world phenomena, in favour of default models fitted to a large dataset. If the default model is a good approximation to the data-generating mechanism this strategy can be successful, but things can quickly go awry if the default model is misspecified. Generalised Bayesian updating (Bissiri et al., 2016), and in particular using divergence-based loss functions (Jewson et al., 2018), has been shown to mitigate some of the risks involved when working with a model that is misspecified. Unlike other robust modelling strategies, these methods do *not* change the statistical model. Instead, they change how the model's parameters are scored, affecting how 'good' parameter values are discerned from 'bad' ones. This is a key practical advantage, as it implies that such strategies do not require precise knowledge about how the model is misspecified. This paper considers generalised Bayesian inference in the context of intractable likelihood. An *intractable likelihood*, in this paper, takes the form $p_\theta(x) = q(x, \theta)/Z(\theta)$, where $q(x, \theta)$ is an analytically tractable function and $Z(\theta)$ is an *intractable* normalising constant, each depending on the value of the unknown parameter θ of interest. Classical Bayesian posteriors resulting from intractable likelihood models are sometimes called *doubly intractable*, due to the computational difficulties they entail (Murray et al., 2006). For example, standard Markov chain Monte Carlo (MCMC) methods cannot be used in this setting, since they typically require explicit evaluation of the likelihood. Doubly intractable posteriors appear in many important statistical applications, including spatial models (Besag, 1974, 1986; Diggle, 1990), exponential random graph models (Park & Haran, 2018), models for gene expression (Jiang et al., 2021), and hidden Potts models for satellite data (Moores et al., 2020).

This paper proposes the first generalised Bayesian approach to inference for models that involve an intractable likelihood. To achieve this, we propose to employ a loss function based on a *Stein discrepancy* (Gorham & Mackey, 2015). As such, this research can be thought of as a Bayesian alternative to the minimum Stein discrepancy estimators of Barp et al. (2019). The methodology is developed for a particular Stein discrepancy called *kernel* Stein discrepancy (KSD), and we call the resulting generalised Bayesian approach *KSD-Bayes*. It is shown in this paper that KSD-Bayes (a) provides robustness to misspecified likelihoods; (b) produces a generalised posterior that is tractable for standard MCMC, or even closed-form when an appropriate conjugate prior (which we identify) is used together with an exponential family likelihood; (c) satisfies several desirable theoretical properties, including a Bernstein-von Mises result which holds irrespective of whether the likelihood is correctly specified. These results appear to represent a compelling case for the use of KSD-Bayes as an alternative to standard Bayesian inference with intractable likelihood. However, KSD-Bayes is no panacea and caution must be taken to avoid certain pathologies of KSD-Bayes, which we highlight in Section 3.5.

The paper is structured as follows: Section 2 contains necessary background on generalised Bayesian inference, Stein discrepancy, and robustness in the Bayesian context. Section 3 presents the KSD-Bayes methodology, including conjugacy of the generalised posterior under an exponential family likelihood. Section 4 elucidates the robustness and asymptotic properties of KSD-Bayes. Guidance for practical application of KSD-Bayes is contained in Section 5. The experimental results and empirical assessments are outlined in Section 6, and we draw our conclusions in Section 7. Code to reproduce all results in this paper can be downloaded from: https://github.com/takuomatsubara/KSD-Bayes.

2 | BACKGROUND

First we provide a short summary of generalised Bayesian inference and Stein discrepancies, putting in place a standing assumption on the domains in which data and parameters are contained:

Standing Assumptions 1: The topological space $\mathcal{X}$, in which the data are contained, is locally compact and Hausdorff. The set $\Theta \subseteq \mathbb{R}^p$, in which parameters are contained, is Borel.

2.1 | Notation

Measure theoretic notation: For a locally compact Hausdorff space such as $\mathcal{X}$, we let $\mathcal{P}(\mathcal{X})$ denote the set of all Borel probability measures on $\mathcal{X}$. A point mass at x is denoted $\delta_x \in \mathcal{P}(\mathcal{X})$. If $\mathcal{X}$ is equipped with a reference measure, then we abuse notation by writing $p \in \mathcal{P}(\mathcal{X})$ to indicate that the distribution with p.d.f. p is an element of $\mathcal{P}(\mathcal{X})$. For $\mathbb{P} \in \mathcal{P}(\mathcal{X})$, we occasionally overload notation by denoting by $L^q(\mathcal{X}, \mathbb{P})$ both the set of functions $f : \mathcal{X} \to \mathbb{R}$ for which $\|f\|_{L^q(\mathcal{X},\mathbb{P})} := (\int_{\mathcal{X}} |f|^q \mathrm{d}\mathbb{P})^{1/q} < \infty$ and the normed space in which two elements $f, g \in L^q(\mathcal{X}, \mathbb{P})$ are identified if they are $\mathbb{P}$-almost everywhere equal. If $\mathbb{P}$ is a Lebesgue measure, we simply write $L^q(\mathcal{X})$ instead of $L^q(\mathcal{X}, \mathbb{P})$. Let $\mathcal{P}_S(\mathbb{R}^d)$ be the set of all Borel probability measures $\mathbb{P}$ supported on $\mathbb{R}^d$, admitting an everywhere positive p.d.f. p and continuous partial derivatives $x \mapsto (\partial/\partial x_{(i)})p(x)$.

Real analytic notation: The Euclidean norm on $\mathbb{R}^d$ is denoted $\|\cdot\|_2$. The set of continuous functions $f : \mathcal{X} \to \mathbb{R}$ is denoted $C(\mathcal{X})$. We denote by $C_b^1(\mathbb{R}^d)$ the set of functions $f : \mathbb{R}^d \to \mathbb{R}$ such that both f and the partial derivatives $x \mapsto (\partial/\partial x_{(i)})f(x)$ are bounded and continuous on $\mathbb{R}^d$. We also denote by $C_b^{1,1}(\mathbb{R}^d \times \mathbb{R}^d)$ the set of bivariate functions $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that both f and the partial derivatives $(x, x') \mapsto (\partial/\partial x_{(i)})(\partial/\partial x'_{(j)})f(x, x')$ are bounded and continuous on $\mathbb{R}^d \times \mathbb{R}^d$. For an arbitrary set $\mathcal{S}(\mathcal{X})$ of functions $f : \mathcal{X} \to \mathbb{R}$, denote by $\mathcal{S}(\mathcal{X}; \mathbb{R}^k)$ the set of $\mathbb{R}^k$-valued functions whose components belong to $\mathcal{S}(\mathcal{X})$. Let ∇ and $\nabla\cdot$ be the gradient and the divergence operators in $\mathbb{R}^d$. For functions with multiple arguments, we sometimes use subscripts to indicate the argument to which the operator is applied (e.g. $\nabla_x f(x, y)$). For f an $\mathbb{R}^d$-valued function, $[\nabla f(x)]_{(I,j)} := (\partial/\partial x_{(i)})f_{(j)}(x)$ and $\nabla \cdot f(x) := \sum_{i=1}^d (\partial/\partial x_{(i)})f_{(i)}(x)$. For f an $\mathbb{R}^{d\times d}$-valued function, $[\nabla f(x)]_{(i,j,k)} := (\partial/\partial x_{(i)})f_{(j,k)}(x)$ and $[\nabla \cdot f(x)]_{(i)} := \sum_{j=1}^d (\partial/\partial x_{(j)})f_{(i,j)}(x)$.

2.2 | Generalised Bayesian inference

Consider a dataset consisting of independent random variables $\{x_i\}_{i=1}^n$ generated from $\mathbb{P} \in \mathcal{P}(\mathcal{X})$, together with a statistical model $\mathbb{P}_\theta \in \mathcal{P}(\mathcal{X})$ for the data, with p.d.f. p_θ, indexed by a parameter of interest $\theta \in \Theta$. The Bayesian statistician elicits a prior $\pi \in \mathcal{P}(\Theta)$, which may reflect a priori belief about the parameter $\theta \in \Theta$, and determines their *a posteriori* belief according to

$$\pi_n(\theta) \propto \pi(\theta) \prod_{i=1}^n p_\theta(x_i) = \pi(\theta) \exp\left\{ \sum_{i=1}^n \log p_\theta(x_i) \right\}. \tag{1}$$

In the *M-closed* setting there exists $\theta_0 \in \Theta$ for which $\mathbb{P} = \mathbb{P}_{\theta_0}$, and the Bayesian update is optimal from an information-theoretic perspective (see Williams, 1980; Zellner, 1988). Optimal processing of information is a desirable property, but in applications the assumption of adequate prior and

model specification is often violated. This has inspired several lines of research, including (but not limited to) strategies for the robust specification of prior belief (Berger et al., 1994), the so-called *safe Bayes* approach (Grünwald, 2011, 2012; Grünwald & van Ommen, 2017; de Heide et al., 2020), *power posteriors* (e.g. Holmes & Walker, 2017), *coarsened posteriors* (Miller & Dunson, 2019), *bagged posteriors* (Huggins & Miller, 2020), *ρ-posteriors* (Baraud & Birg"e, 2020) and Bayesian inference based on scoring rules (Giummolè et al., 2019). A particularly versatile approach to robustness, which encompasses most of the above, is *generalised Bayesian inference* (Bissiri et al., 2016) (see also the earlier work of Chernozhukov & Hong, 2003). This approach constructs a distribution, denoted π_n^L, using a *loss function* $L_n : \Theta \to \mathbb{R}$, which may be data-dependent, and a scaling parameter $\beta > 0$, according to

$$\pi_n^L(\theta) \propto \pi(\theta) \exp\left\{-\beta n L_n(\theta)\right\}. \tag{2}$$

The so-called *generalised posterior* π_n^L coincides with the Bayesian posterior π_n when $\beta = 1$ and the loss function is the negative average log-likelihood; $L_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log p_\theta(x_i)$. As discussed in Bissiri et al. (2016); Knoblauch et al. (2019), generalised Bayesian inference admits an optimisation-centric interpretation:

$$\pi_n^L = \underset{\rho \in \mathcal{P}(\Theta)}{\arg\min} \left\{\beta n \ \mathbb{E}_{\theta\sim\rho}\left[L_n(\theta)\right] + \mathrm{KL}(\rho\|\pi)\right\} \tag{3}$$

where $\mathrm{KL}(\rho\|\pi)$ denotes the Kullbac–Leibler (KL) divergence between two distributions $\rho, \pi \in \mathcal{P}(\Theta)$. This perspective reveals that the standard Bayesian posterior is an implicit commitment to a particular loss function– the negative log-likelihood– and that the weighting constant β controls the influence of this loss relative to the prior π. In particular, under mild conditions $L_n(\theta) \xrightarrow{\text{a.s.}} \mathrm{KL}(\mathbb{P}\|\mathbb{P}_\theta) + C$ as $n \to \infty$, for a constant C independent of θ, which reveals that standard Bayesian posterior concentrates around the value of θ that minimizes the KL divergence between the data-generating distribution $\mathbb{P}$ and the model $\mathbb{P}_\theta$. Outside of the M-closed setting such concentration is problematic, often leading to over-confident predictions (Bernardo & Smith, 2009).

The use of alternative, divergence-based loss functions has been demonstrated to mitigate the negative consequences of a misspecified statistical model, as pioneered in the work on α- and β-divergences in Hooker and Vidyashankar (2014) and Ghosh and Basu (2016) and extended to γ-divergence in Nakagawa and Hashimoto (2020). See also Baraud and Birg"e (2020). The properties of the divergence, including any potentially undesirable pathologies associated with it, determine the properties of the generalised posterior (Jewson et al., 2018; Knoblauch et al., 2019). These compelling theoretical results have led to considerable interest in generalised Bayesian inference with divergence-based loss functions, yet the divergences that have been considered to-date cannot be computed in the important setting of intractable likelihood.

2.3 | Stein discrepancy

In an independent line of research, *Stein discrepancies* were proposed in Gorham and Mackey (2015) to provide statistical divergences that are both computable and capable of providing various forms of distributional convergence control. The approach is based on the method of Stein

(1972), which requires the identification of a linear operator $\mathcal{S}_{\mathbb{Q}} : \mathcal{H} \to L^1(\mathcal{X}, \mathbb{Q})$, depending on a probability distribution $\mathbb{Q} \in \mathcal{P}(\mathcal{X})$ and acting on a Banach space $\mathcal{H}$, such that

$$\mathbb{E}_{X\sim\mathbb{Q}}[\mathcal{S}_{\mathbb{Q}}[h](X)] = 0 \quad \forall h \in \mathcal{H}. \tag{4}$$

Such an operator $\mathcal{S}_{\mathbb{Q}}$ is called a *Stein operator* and $\mathcal{H}$ is called a *Stein set*. Given a distribution $\mathbb{Q} \in \mathcal{P}(\mathcal{X})$, there are infinitely many operators $\mathcal{S}_{\mathbb{Q}}$ satisfying (4). A convenient example is the *Langevin Stein operator* (Gorham & Mackey, 2015), defined for $\mathcal{X} = \mathbb{R}^d$, $\mathbb{Q} \in \mathcal{P}_{\mathrm{S}}(\mathbb{R}^d)$ and a Banach space $\mathcal{H}$ of differentiable functions $h : \mathbb{R}^d \to \mathbb{R}^d$, as

$$\mathcal{S}_{\mathbb{Q}}[h](x) = h(x) \cdot \nabla \log q(x) + \nabla \cdot h(x) \tag{5}$$

where q is the p.d.f. of $\mathbb{Q}$. Under suitable regularity conditions on $\nabla \log q$ and $\mathcal{H}$, the Langevin Stein operator satisfies Equation (4); see Gorham and Mackey (2015 Proposition 1). Given $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{X})$ and a Stein operator $\mathcal{S}_{\mathbb{Q}} : \mathcal{H} \to L^1(\mathcal{X}, \mathbb{Q})$ whose image is contained in $L^1(\mathcal{X}, \mathbb{P})$, the *Stein discrepancy* is defined as

$$\mathrm{SD}(\mathbb{Q}\|\mathbb{P}) := \sup_{\|h\|_{\mathcal{H}}\leq 1} \left|\mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}[h](X)\right] - \mathbb{E}_{X\sim\mathbb{Q}}\left[\mathcal{S}_{\mathbb{Q}}[h](X)\right]\right| = \sup_{\|h\|_{\mathcal{H}}\leq 1} \left|\mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}[h](X)\right]\right|, \tag{6}$$

where the last equality follows directly from Equation (4). Under mild assumptions, Stein discrepancy defines a statistical divergence between two probability distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{P}(\mathcal{X})$, meaning that $\mathrm{SD}(\mathbb{Q}\|\mathbb{P}) \geq 0$ with equality if and only if $\mathbb{P} = \mathbb{Q}$; see Proposition 1 and Theorem 2 in Barp et al. (2019). Under slightly stronger assumptions, Stein discrepancy provides convergence control, meaning that a sequence $(\mathbb{Q}_n)_{n=1}^{\infty} \subset \mathcal{P}(\mathcal{X})$ converges in a specified sense to $\mathbb{Q}$ whenever $\mathrm{SD}(\mathbb{Q}\|\mathbb{Q}_n) \to 0$; see Gorham and Mackey (2015 Theorem 2, Proposition 3) and Gorham and Mackey (2017) Theorem 8, Proposition 9). An important property of Stein discrepancies that we exploit in this work is that, unlike other divergences, Stein discrepancies can often be computed with an un-normalised representation of $\mathbb{Q}$. For example, the Stein operators in Equation (5) depend on $\mathbb{Q}$ only through $\nabla \log q$, which can be computed when q is provided in a form that involves an intractable normalisation constant. The suitability of Stein discrepancy for use in generalised Bayesian inference has not previously been considered, and this is our focus next.

3 | METHODOLOGY

Highly structured data, or data belong to a high-dimensional domain $\mathcal{X}$, are often associated with an intractable likelihood. Moreover, the difficulty of modelling such data means that models will typically be misspecified. Thus there is a pressing need for Bayesian methods that are both robust and compatible with intractable likelihood. To this end, in Section 3.1 we introduce *SD-Bayes*, a generalised Bayesian procedure with a loss function based on Stein discrepancy. There are numerous Stein discrepancies that can be considered, and in Section 3.2 we focus in detail on KSD due to the possibility of performing fully conjugate inference in the context of exponential family models, as described in Section 3.3. Non-conjugate inference and its computational cost are discussed in Section 3.4. However, all statistical divergences have their pathologies, and one must bear in mind the pathologies of KSD when using KSD-Bayes; see the discussion in Section 3.5.

3.1 | SD-Bayes

Suppose we are given a prior p.d.f. $\pi \in \mathcal{P}(\Theta)$ and a statistical model $\{\mathbb{P}_\theta | \theta \in \Theta\} \subset \mathcal{P}(\mathcal{X})$. Let $\{x_i\}_{i=1}^n$ be independent observations generated from $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ and let $\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ be the empirical measure associated to this dataset. In this context, the SD-Bayes generalised posterior can now be defined:

Definition 1 (SD-Bayes). For each $\theta \in \Theta$, select a Stein operator $\mathcal{S}_{\mathbb{P}_\theta}$ and denote the associated Stein discrepancy $\mathrm{SD}(\mathbb{P}_\theta \| \cdot)$. Let $\beta \in (0, \infty)$. Then the SD-Bayes generalised posterior is defined as

$$\pi_n^D(\theta) \propto \pi(\theta) \exp\left\{-\beta n \mathrm{SD}^2(\mathbb{P}_\theta \| \mathbb{P}_n)\right\} \tag{7}$$

where $\theta \in \Theta$.

Here the 'D' superscript stands for *discrepancy*. Comparing (7) to (2) confirms that SD-Bayes is a generalised Bayesian method with loss function $L_n(\theta) = \mathrm{SD}^2(\mathbb{P}_\theta \| \mathbb{P}_n)$. There is an arbitrariness to using squared discrepancy, as opposed to another power of the discrepancy, but this choice turns out to be appropriate for the discrepancies considered in Section 3.2, ensuring that fluctuations of $L_n(\theta)$ about its expectation are $\mathcal{O}(n^{-1/2})$, analogous to the standard Bayesian loss, and permitting tractable computation (Section 3.3) and analysis (Section 4). A discussion of how the weight β should be selected is deferred until after our theoretical analysis, in Section 5.

3.2 | KSD-Bayes

Compared to other Stein discrepancies, KSDs are attractive because they enable the supremum in Equation (6) to be explicitly computed. To define KSD, we require the concept of a (matrix-valued) *kernel* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$; the precise definition is contained in Appendix A. For our purposes in the main text, it suffices to point out that any kernel K has a uniquely associated Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}^d$, called a *vector-valued reproducing kernel Hilbert space* (v-RKHS). This v-RKHS constitutes the Stein set in KSD, and we therefore denote this v-RKHS as $\mathcal{H}$. The associated norm and inner product will respectively be denoted $\|\cdot\|_{\mathcal{H}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{H}}$.

Let $\mathcal{S}_{\mathbb{Q}}$ be a Stein operator and denote the action of $\mathcal{S}_{\mathbb{Q}}$ on both the first and second argument[1] of a kernel K as $\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}}K$. The following result is a generalisation of the original construction of KSD (Chwialkowski et al., 2016; Liu et al., 2016) to general Stein operators.

Assumption 1 Let $\mathcal{H}$ be a v-RKHS with kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$. For $\mathbb{Q} \in \mathcal{P}(\mathcal{X})$, let $\mathcal{S}_{\mathbb{Q}}$ be a Stein operator with domain $\mathcal{H}$. For each fixed $x \in \mathcal{X}$, we assume $h \mapsto \mathcal{S}_{\mathbb{Q}}[h](x)$ is a continuous linear functional on $\mathcal{H}$. Further, we assume that $\mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}}K(X,X)\right] < \infty$.

Proposition 1 (Closed Form of Stein Discrepancy). *Under Assumption* 1, *we have*

$$\mathrm{SD}^2(\mathbb{Q}\|\mathbb{P}) = \mathrm{KSD}^2(\mathbb{Q}\|\mathbb{P}) := \mathbb{E}_{X,X'\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}}K(X,X')\right]$$

where X and X' are independent.

[1] More precisely, denoting the j-th column of $K(x,x') \in \mathbb{R}^{d\times d}$ by $K_{\_j}(x,x') \in \mathbb{R}^d$, we define $\mathcal{S}_{\mathbb{Q}}K(x,x') := [\mathcal{S}_{\mathbb{Q}}K_{\_1}(x,x'), \ldots, \mathcal{S}_{\mathbb{Q}}K_{\_d}(x,x')] \in \mathbb{R}^d$ where $\mathcal{S}_{\mathbb{Q}}K_{\_j}(x,x') := \mathcal{S}_{\mathbb{Q}}[K_{\_j}(\cdot,x')](x)$ is an action of $\mathcal{S}_{\mathbb{Q}}$ for the $\mathbb{R}^d$-valued function $K_{\_j}(\cdot,x')$ at each $x' \in \mathcal{X}$. We further define $\mathcal{S}_{\mathbb{Q}}\mathcal{S}_{\mathbb{Q}}K(x,x') := \mathcal{S}_{\mathbb{Q}}[\ \mathcal{S}_{\mathbb{Q}}K(x,\cdot)\ ](x')$ as an action of $\mathcal{S}_{\mathbb{Q}}$ for the $\mathbb{R}^d$-valued function $\mathcal{S}_{\mathbb{Q}}K(x,\cdot)$ at each $x \in \mathcal{X}$.

The proof is in Appendix B.1. Note that it is straightforward to verify the assumption that $h \mapsto \mathcal{S}_{\mathbb{Q}}[h](x)$ is a continuous linear functional for each fixed $x \in \mathcal{X}$ once the form of $\mathcal{S}_{\mathbb{Q}}$ is specified; see Appendix B.1.2. KSD is attractive for SD-Bayes since it enables the generalised posterior in Definition 1 to be explicitly computed:

$$\mathrm{KSD}^2(\mathbb{P}_\theta \| \mathbb{P}_n) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(x_i, x_j). \tag{8}$$

The resulting generalised posterior will be referred to as *KSD-Bayes* in the sequel. The explicit form of $\mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K$ depends on $\mathcal{S}_{\mathbb{P}_\theta}$. The case of $\mathcal{X} = \mathbb{R}^d$ and the Langevin Stein operator in Equation (5) is given by

$$\begin{aligned} \mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(x, x') = {} & \nabla \log p_\theta(x) \cdot K(x, x') \nabla \log p_\theta(x') + \nabla_x \cdot \left(\nabla_{x'} \cdot K(x, x') \right) \\ & + \nabla \log p_\theta(x) \cdot \left(\nabla_{x'} \cdot K(x, x') \right) + \nabla \log p_\theta(x') \cdot \left(\nabla_x \cdot K(x, x') \right) \end{aligned} \tag{9}$$

where p_θ is a p.d.f. for $\mathbb{P}_\theta \in \mathcal{P}_{\mathrm{S}}(\mathbb{R}^d)$. Clearly, this expression is straightforward to evaluate[2] whenever we have access to derivatives of the kernel and the log density. If the derivatives are analytically tedious, the expression above is amenable to the use of automatic differentiation tools (Baydin et al., 2018).

Whether KSD-Bayes is reasonable or not hinges crucially on whether KSD is a meaningful way to quantify the difference between the discrete distribution $\mathbb{P}_n$ and the parametric model $\mathbb{P}_\theta$. Sufficient conditions for convergence control have been established for the Langevin Stein operator, under which the convergence of $\mathrm{KSD}(\mathbb{P}_\theta \| \mathbb{P}_n)$ implies the weak convergence of $\mathbb{P}_n$ to $\mathbb{P}_\theta$ (Gorham & Mackey, 2017, Theorem 8). This provides some preliminary assurance that KSD-Bayes may work; we present formal theoretical guarantees in Section 4. These theoretical results motivate specific choices of K for use in KSD-Bayes, which we discuss in Section 5.

3.3 | Conjugate inference for exponential family models

The generalised posterior can be exactly computed in the case of an natural exponential family model when a conjugate prior is used. Let $\eta : \Theta \to \mathbb{R}^k$ and $t : \mathcal{X} \to \mathbb{R}^k$ be any sufficient statistic for some $k \in \mathbb{N}$ and let $a : \Theta \to \mathbb{R}$ and $b : \mathcal{X} \to \mathbb{R}$. An exponential family model has p.m.f. or p.d.f. (with respect to an appropriate reference measure on $\mathcal{X}$) of the form

$$p_\theta(x) = \exp(\eta(\theta) \cdot t(x) - a(\theta) + b(x)). \tag{10}$$

This includes a wide range of distributions with an intractable normalisation constant $\exp(a(\theta))$, used in statistical applications such as random graph estimation (Yang et al., 2015), spin glass

[2]For maximum clarity, the vector calculus notation is expanded as follows:

$$\begin{aligned} \mathcal{S}_{\mathbb{P}_\theta} \mathcal{S}_{\mathbb{P}_\theta} K(x, x') = {} & \sum_{i,j=1}^{d} \frac{\partial}{\partial x_{(i)}} \log p_\theta(x) \left[K(x, x') \right]_{(i,j)} \frac{\partial}{\partial x_{(j)}} \log p_\theta(x) + \frac{\partial^2}{\partial x_{(i)} \partial x'_{(j)}} \left[K(x, x') \right]_{(i,j)} \\ & + \frac{\partial}{\partial x_{(i)}} \log p_\theta(x) \frac{\partial}{\partial x'_{(j)}} \left[K(x, x') \right]_{(i,j)} + \frac{\partial}{\partial x'_{(j)}} \log p_\theta(x') \frac{\partial}{\partial x_{(i)}} \left[K(x, x') \right]_{(i,j)} \end{aligned}$$

models (Besag, 1974) and the kernel exponential family model (Canu & Smola, 2006). The model in Equation (10) is called *natural* when the canonical parametrisation $\eta(\theta) = \theta$ is employed.

Proposition 2 *Consider $\mathcal{X} = \mathbb{R}^d$ and the Langevin Stein operator $\mathcal{S}_{\mathbb{P}_\theta}$ in Equation* (5)*, where $\mathbb{P}_\theta$ is the exponential family in Equation* (10)*, and a kernel $K \in C_b^{1,1}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^{d\times d})$. Assuming the prior has a p.d.f. π, the KSD-Bayes generalised posterior has a p.d.f.*

$$\pi_n^D(\theta) \propto \pi(\theta) \exp\left(-\beta n\{\eta(\theta) \cdot \Lambda_n \eta(\theta) + \eta(\theta) \cdot \nu_n\}\right),$$

where $\Lambda_n \in \mathbb{R}^{k\times k}$ and $\nu_n \in \mathbb{R}^k$ are defined as

$$\Lambda_n := \frac{1}{n^2}\sum_{i,j=1}^n \nabla t(x_i) \cdot K(x_i, x_j) \nabla t(x_j),$$

$$\nu_n := \frac{1}{n^2}\sum_{i,j=1}^n \nabla t(x_i) \cdot (\nabla_{x_j} \cdot K(x_i, x_j)) + \nabla t(x_j) \cdot (\nabla_{x_i} \cdot K(x_i, x_j)) + 2\nabla t(x_i) \cdot K(x_i, x_j) \nabla b(x_j).$$

For a natural exponential family we have $\eta(\theta) = \theta$, and the prior $\pi(\theta) \propto \exp(-\frac{1}{2}(\theta - \mu) \cdot \Sigma^{-1}(\theta - \mu))$ leads to a generalised posterior

$$\pi_n^D(\theta) \propto \exp\left(-\frac{1}{2}(\theta - \mu_n) \cdot \Sigma_n^{-1}(\theta - \mu_n)\right),$$

where $\Sigma_n^{-1} := \Sigma^{-1} + 2\beta n \Lambda_n$ and $\mu_n := \Sigma_n^{-1}(\Sigma^{-1}\mu - \nu_n)$.

The proof is in Appendix B.2. That the Gaussian distribution will be conjugate in KSD-Bayes, even in the presence of intractable likelihood, is remarkable and notably different from the classical Bayesian case, albeit at a $\mathcal{O}(n^2)$ computational cost. Strategies to further reduce this computational cost are discussed in Section 3.4. It is well known that certain minimum discrepancy estimators, such as the *score matching estimator* (Hyvärinen, 2005) and the *minimum KSD estimator* (Barp et al., 2019), have closed forms in the case of an exponential family models; it is similar reasoning that has led us to Proposition 2.

3.4 | Non-conjugate inference and computational cost

To access the generalised posterior in the non-conjugate case, existing MCMC algorithms for *tractable* likelihood can be used[3]. The per-iteration computational cost appears to be $\mathcal{O}(n^2)$ since, for each state θ visited along the sample path, the KSD in Equation (8) must be evaluated. However, various strategies enable this computational cost to be mitigated. For concreteness of the discussion that follows, we consider the Langevin Stein operator, for which

$$(8) \overset{+C}{=} \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \left\{\begin{array}{c} \nabla \log p_\theta(x_i) \cdot K(x_i, x_j) \nabla \log p_\theta(x_j) + \nabla \log p_\theta(x_i) \cdot \nabla_{x_j} \cdot K(x_i, x_j) \\ +\nabla \log p_\theta(x_j) \cdot \nabla_{x_i} \cdot K(x_i, x_j) \end{array}\right\}$$

where the equality holds up to a θ-independent constant.

[3]For example, the Gaussian form of the data-dependent term in Proposition 2 suggests that elliptical slice sampling may work well when the natural parametrisation of the exponential family is employed (Murray et al., 2010).

Memoisation: The above expression depends on θ only through the terms $\{\nabla \log p_\theta(x_i)\}_{i=1}^n$, of which there are $\mathcal{O}(n)$, while all other terms involving K, of which there are $\mathcal{O}(n^2)$, can be computed once and memoised. The double summation still necessitates $\mathcal{O}(n^2)$ computational cost but this operation is *embarrassingly parallel*.

Finite rank kernel: Computational cost can be reduced from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$ using a finite rank kernel. A useful and important example is the rank one kernel $K(x, x') = I_d$, which reduces (8) to

$$(8) \overset{+C}{=} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla \log p_\theta(x_i) \right\|^2$$

and is closely related to divergences used in *score matching* (Hyvärinen, 2005). Random finite rank approximations of the kernel can also considered in this context (Huggins & Mackey, 2018).

Stochastic approximation: The construction of low-cost unbiased estimators for (8) is straight-forward via sampling *mini-batches* from the dataset. This enables a variety of exact and approximate algorithms for posterior approximation to be exploited (e.g. Ma et al., 2015). Alternatively, Huggins and Mackey (2018); Gorham et al. (2020) argued for stochastic approximations of KSD that could be used.

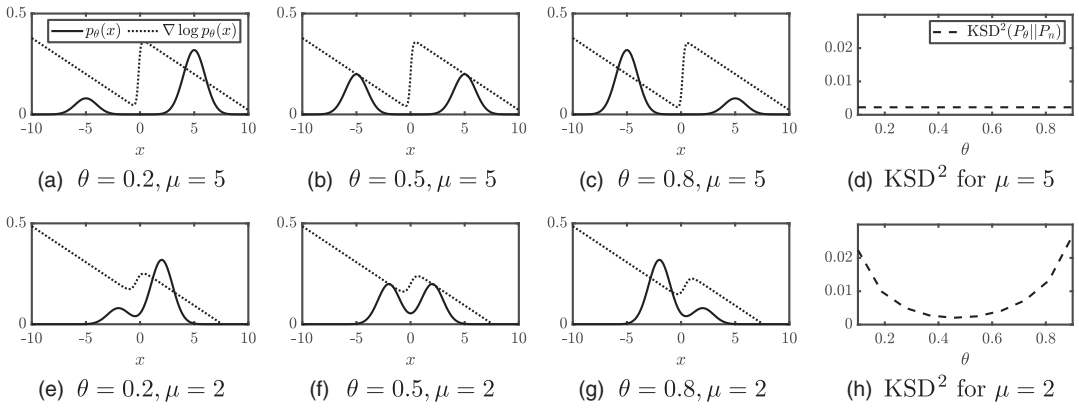
3.5 | Limitations of KSD-Bayes

A divergence $\mathrm{D}(\mathbb{Q}\|\mathbb{P})$ induces an information geometry (Amari, 1997), encoding a particular sense in which $\mathbb{Q}$ can be considered to differ from $\mathbb{P}$. As such, all divergence exhibit *pathologies*, meaning that certain characteristics that distinguish $\mathbb{Q}$ from $\mathbb{P}$ are less easily detected. A documented pathology of gradient-based discrepancies, including the Langevin KSD, is their insensitivity to the existence of high-probability regions which are well-separated; see (Gorham et al. 2019 Section 5.1) and Wenliang (2020). To see this, consider a Gaussian mixture model

$$p_\theta(x) = \frac{\theta}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right) + \frac{(1-\theta)}{\sqrt{2\pi}} \exp\left(-\frac{(x+\mu)^2}{2}\right) \tag{11}$$

where $\theta \in [0, 1]$ specifies the mixture ratio and $\mu \in \mathbb{R}$ controls the separation between the two components. If the two components are well-separated, that is, $\mu \gg 1$, the gradient $\nabla \log p_\theta$ becomes insensitive to θ and hence a gradient-based divergence such as KSD will be insensitive to θ, as demonstrated in Figure 1. For this reason, caution is warranted when gradient-based discrepancies are used. However, in practice direct inspection of the dataset and knowledge of how $\mathbb{P}_\theta$ is parametrised can be used to ascertain whether either distribution is multi-modal. Our applications in Section 6 are not expected to be multi-modal (with the exception of the kernel exponential family in Section 6.3 which was selected to demonstrate the insensitivity to mixing proportions of KSD-Bayes).

A second limitation of KSD-Bayes is non-invariance to a change of coordinates in the dataset. This is a limitation of loss-based estimators in general. In Section 5.1 we recommend a data-adaptive choice of kernel, which serves to provide approximate invariance to affine transformations of the dataset. As usual in statistical analyses, we recommend *post-hoc* assessment of the sensitivity of inferences to perturbations of the dataset.



(a) $\theta = 0.2, \mu = 5$ (b) $\theta = 0.5, \mu = 5$ (c) $\theta = 0.8, \mu = 5$ (d) KSD^2 for $\mu = 5$

(e) $\theta = 0.2, \mu = 2$ (f) $\theta = 0.5, \mu = 2$ (g) $\theta = 0.8, \mu = 2$ (h) KSD^2 for $\mu = 2$

FIGURE 1 Illustrating the insensitivity to mixture proportions of KSD. anels (a-c,e-g) display the density function $p_\theta(x)$ from Equation (11) together with the gradient $\nabla \log p_\theta(x)$, the latter rescaled to fit onto the same plot. Panels (d,h) display the discrepancy $\mathrm{KSD}^2\left(\mathbb{P}_\theta \| \mathbb{P}_n\right)$, where $\mathbb{P}_n$ is an empirical distribution of $n = 1000$ samples from the model with $\theta = 0.5$

A third limitation of KSD-Bayes is the loss of efficiency that can occur in settings where the data are high-dimensional. Sliced versions of KSD have been proposed to address the curse of dimension for KSD (Gong et al., 2021), but to limit scope we do not consider the combination of slicediscrepanciess and KSD-Bayes in this work.

Despite these limitations, KSD-Bayes represents a flexible and effective procedure for generalised Bayesian inference in the context of an intractable likelihood. Our attention turns next to theoretical analysis of KSD-Bayes.

4 | THEORETICAL ASSESSMENT

This section contains a comprehensive theoretical treatment of KSD-Bayes. The main results are *posterior consistency* and a *Bernstei–von Mises* theorem in Section 4.2, and *global bias-robustness* of the generalised posterior in Section 4.3. In obtaining these results we have developed novel intermediate results concerning an important V-statistic estimator for KSD; these are anticipated to be of independent interest, so we present these in Section 4.1 of the main text. Note that all theory is valid for the misspecified regime where $\mathbb{P}$ need not be an element of $\{\mathbb{P}_\theta : \theta \in \Theta\}$. Moreover, the results in Sections 4.1 and 4.2 hold for general data domains $\mathcal{X}$. For the entirety of this section we set $\beta = 1$, with all results for $\beta \neq 1$ immediately recovered by replacing K with βK. The results of this section motivate a specific choice for β that is described in Section 5.

Standing Assumptions 2: The dataset $\{x_i\}_{i=1}^n$ consists of independent samples generated from $\mathbb{P} \in \mathcal{P}(\mathcal{X})$, with empirical distribution denoted $\mathbb{P}_n := (1/n)\sum_{i=1}^n \delta_{x_i}$. The set $\Theta \subseteq \mathbb{R}^p$ is open, convex and bounded[4]. Assumption 1 holds with $\mathbb{Q} = \mathbb{P}_\theta$ for every $\theta \in \Theta$.

Notation: For shorthand, let ∂^1, ∂^2 and ∂^3 denote the partial derivatives $(\partial/\partial\theta_{(h)})$, $(\partial^2/\partial\theta_{(h)}\partial\theta_{(k)})$ and $(\partial^3/\partial\theta_{(h)}\partial\theta_{(k)}\partial\theta_{(l)})$ for $h, k, l \in \{1, \ldots, p\}$, where to reduce notation the

[4]The assumption that Θ is bounded is used only to simplify the statement of our results. For the case where Θ is not bounded, it is sufficient for Assumptions 2 and 3 to hold on an open, convex and bounded subset $U \subset \Theta$. Then it can be verified that Lemmas 2 and 3 hold on the bounded subset U, and that all the other results hold on Θ.

indices (h, k, l) are left implicit. The gradient and Hessian operators are $[\nabla_\theta]_{(h)} = (\partial/\partial\theta_{(h)})$ and $[\nabla^2_\theta]_{(h,k)} = (\partial^2/\partial\theta_{(h)}\partial\theta_{(k)})$.

4.1 | Minimum KSD estimators

First, we present novel analysis of the V-statistic $\mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P}_n)$. A related U-statistic estimator of KSD was analysed in Barp et al. (2019) but this is only an estimate of $\mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P})$, rendering it unsuitable for generalised Bayesian inference, which requires losses to be lower-bounded (Jewson et al., 2018). Furthermore, our results for the V-statistic do not depend on a specific form of $\mathcal{S}_{\mathbb{P}_\theta}$, in contrast to Barp et al. (2019) who considered the *diffusion* Stein operator, and may hence be of independent interest.

Despite the bias present in a V-statistic, our standing assumptions are sufficient to derive the following consistency result:

Lemma 1 (a.s. Pointwise Convergence). *For each* $\theta \in \Theta$,

$$\mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P}_n) - \mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P}) \overset{a.s.}{\to} 0.$$

The proof is contained in Appendix B.3.1. If we impose further regularity, we can obtain a uniform convergence result. It will be convenient to introduce a collection of assumptions that are indexed by $r_{\max} \in \{0, 1, 2, \dots\}$, as follows:

Assumption 2 ($r_{\max}$). For all integers $0 \le r \le r_{\max}$, the following conditions hold:

1. the map $\theta \mapsto \partial^r \mathcal{S}_{\mathbb{P}_\theta}[h](x)$ exists and is continuous, for all $h \in \mathcal{H}$ and $x \in \mathcal{X}$;
2. the map $h \mapsto (\partial^r \mathcal{S}_{\mathbb{P}_\theta})[h](x)$ is a continuous linear functional on $\mathcal{H}$, for each $x \in \mathcal{X}$;
3. $\mathbb{E}_{X\sim\mathbb{P}}[\sup_{\theta\in\Theta}((\partial^r \mathcal{S}_{\mathbb{P}_\theta})(\partial^r \mathcal{S}_{\mathbb{P}_\theta})K(X,X))] < \infty$,

where $(\partial^0 \mathcal{S}_{\mathbb{P}_\theta}) := \mathcal{S}_{\mathbb{P}_\theta}$; note that (2) with $r = 0$ is implied from Standing Assumption 2.

In the expression above, the first and second $(\partial^r \mathcal{S}_{\mathbb{P}_\theta})$ are applied, respectively, to the first and second argument of K, as with $\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x,x)$. These assumptions become concrete when considering a specific Stein operator; the case of the Langevin Stein operator is presented in Appendix B.3.5.

Lemma 2 (a.s. Uniform Convergence). *Suppose Assumption* 2 ($r_{\max} = 1$) *holds. Then*

$$\sup_{\theta\in\Theta} \left|\mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P}_n) - \mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P})\right| \overset{a.s.}{\to} 0.$$

The proof is contained in Appendix B.3.2.

Our next results concern consistency and asymptotic normality of the estimator θ_n that minimises the V-statistic in Equation (8).

Assumption 3 There exist minimisers θ_n of $\mathrm{KSD}(\mathbb{P}_\theta\|\mathbb{P}_n)$ for all sufficiently large $n \in \mathbb{N}$, and there exists a unique θ_* s.t. $\mathrm{KSD}(\mathbb{P}_{\theta_*}\|\mathbb{P}) < \inf_{\{\theta\in\Theta:\|\theta-\theta_*\|_2\ge\epsilon\}} \mathrm{KSD}(\mathbb{P}_\theta\|\mathbb{P})$ for any $\epsilon > 0$.

Lemma 3 (Strong Consistency). *Suppose Assumptions* 2 ($r_{\max} = 1$) *and* 3 *hold. Then*

$$\theta_n \overset{a.s.}{\to} \theta_*.$$

The proof is contained in Appendix B.3.3. For the well-specified case where $\exists\theta_0$ s.t. $\mathbb{P}_{\theta_0} = \mathbb{P}$, the uniqueness of θ_* holds automatically if KSD is a proper divergence, that is, $\mathrm{KSD}(\mathbb{P}\|\mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q}$. For example, if the preconditions of (Barp et al. 2019 Proposition 1) are satisfied and the parametrisation $\theta \mapsto \mathbb{P}_\theta$ is injective, the minimum is uniquely attained.

Let $H_* := \nabla^2_\theta \mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P})|_{\theta=\theta_*}$ and $J_* := \mathbb{E}_{X\sim\mathbb{P}}[S(X,\theta_*)S(X,\theta_*)^\top]$, where we define the column vector $S(x,\theta) := \mathbb{E}_{X\sim\mathbb{P}}[\nabla_\theta(\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x,X))]$. Asymptotic normality of θ_n can be established if further regularity is imposed:

Lemma 4 (Asymptotic Normality). *Suppose Assumptions* 2 ($r_{\max} = 3$) *and* 3 *hold. If* H_* *is non-singular*,

$$\sqrt{n}\,(\theta_n - \theta_*) \xrightarrow{d} \mathcal{N}(0, H_*^{-1}J_*H_*^{-1})$$

where $\xrightarrow{d}$ *denotes the convergence in distribution.*

The proof is contained in Appendix B.3.4. These preliminaries on minimum KSD estimation are required for our main results on KSD-Bayes, presented next.

4.2 | Posterior consistency and Bernstein–von mises

Armed with the technical results of Section 4.1, we can now establish consistency of KSD-Bayes and a Bernstein–von Mises result. Our consistency result requires a *prior mass condition*, similar to that of Cherief-Abdellatif and Alquier (2020):

Assumption 4 The prior is assumed to

1. admit a p.d.f. π that is continuous at θ_*, with $\pi(\theta_*) > 0$;
2. satisfy $\int_{B_n(\alpha_1)} \pi(\theta)\mathrm{d}\theta \geq e^{-\alpha_2\sqrt{n}}$ for some constants $\alpha_1, \alpha_2 > 0$,

where we define $B_n(\alpha_1) := \{\theta \in \Theta : |\mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P}) - \mathrm{KSD}^2(\mathbb{P}_{\theta_*}\|\mathbb{P})| \leq \alpha_1/\sqrt{n}\}$.

Assumption 4 specifies the amount of prior mass in a neighbourhood around the population-optimal value θ_* that is required. This is not a strong assumption and Appendix B.7 demonstrates how each of Assumptions 2 to 4 can be verified in the case of an exponential family model.

Theorem 1 (Posterior Consistency). *Suppose Assumptions* 3 *and* 4 *hold. Let* $\sigma(\theta) := \mathbb{E}_{X\sim\mathbb{P}}\left[\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(X,X)\right]$. *Then, for all* $\delta \in (0,1]$,

$$\mathbb{P}\left(\left|\int_\Theta \mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P})\pi_n^D(\theta)\mathrm{d}\theta - \mathrm{KSD}^2(\mathbb{P}_{\theta_*}\|\mathbb{P})\right| > \delta\right) \leq \frac{\alpha_1 + \alpha_2 + 8\sup_{\theta\in\Theta}\sigma(\theta)}{\delta\sqrt{n}}$$

where the probability is with respect to realisations of the dataset $\{x_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathbb{P}$.

The proof is contained in Appendix B.4.

Next, we derive a Bernstein–von Mises result. The pioneering work of Hooker and Vidyashankar (2014) and Ghosh and Basu (2016) established Bernstein–von Mises results for generalised posteriors defined by α- and β-divergences. Unfortunately, the form of KSD is rather

different and different theoretical tools are required to tackle it. Miller (2021) introduced a general approach to deriving Bernstein–von Mises results for generalised posteriors, demonstrating how the assumptions can be verified for several additive loss functions L_n. Our proof builds on Miller (2021), demonstrating that the required assumptions can also be satisfied by the non-additive KSD loss function in Equation (8).

Theorem 2 (Bernstein–von Mises). *Suppose Assumption* 2 ($r_{\max} = 3$), 3, *and part* (1) *of* 4 *hold. Let* $\hat{\pi}_n^D$ *the p.d.f. of the random variable* $\sqrt{n}(\theta - \theta_n)$ *for* $\theta \sim \pi_n^D$, *viewed as a p.d.f. on* $\mathbb{R}^p$. *If* H_* *is nonsingular*,

$$\int_{\mathbb{R}^p} \left| \hat{\pi}_n^D(\theta) - \frac{1}{\det(2\pi H_*^{-1})^{1/2}} \exp\left(-\frac{1}{2}\theta \cdot H_* \theta\right) \right| \mathrm{d}\theta \overset{a.s.}{\to} 0,$$

where the a.s. convergence is with respect to realisations of the dataset $\{x_i\}_{i=1}^n$.

The proof is contained in Appendix B.5. These positive results are encouraging, as they indicate the limitations of KSD-Bayes described in Section 3.5 are at worst a finite sample size effect. However, we note that the asymptotic precision matrix H_* from Theorem 2 differs to the precision matrix $H_* J_*^{-1} H_*$ of the minimum KSD estimator from Lemma 4; this is analogous to fact that Bayesian credible sets can have asymptotically incorrect frequentist coverage if the statistical model is misspecified (Kleijn & van der Vaart, 2012; Müller, 2013). This point will be addressed in Section 5.2.

Remark 1 The analysis in Sections 4.1 and 4.2 covers general domains $\mathcal{X}$ and Stein operators $\mathcal{S}_{\mathbb{P}}$. Henceforth, in the main text we restrict attention to $\mathcal{X} = \mathbb{R}^d$, but the case of a discrete domain $\mathcal{X}$, and the identification of an appropriate Stein operator in this context, are discussed in Appendix D.5.

4.3 | Global Bias-robustness of KSD-Bayes

An important property of KSD-Bayes is that, through a suitable choice of kernel, the generalised posterior can be made robust to contamination in the dataset. This robustness will now be rigorously established.

Consider the *ε-contamination model* $\mathbb{P}_{n,\epsilon,y} = (1-\epsilon)\mathbb{P}_n + \epsilon\delta_y$, where $y \in \mathcal{X}$ and $\epsilon \in [0,1]$ (see Huber & Ronchetti, 2009). In other words, the datum y is considered to be contaminating the dataset $\{x_i\}_{i=1}^n$. Robustness in the generalised Bayesian setting has been considered in Hooker and Vidyashankar (2014), Ghosh and Basu (2016) and Nakagawa and Hashimoto (2020). In what follows we write $L_n(\theta) = L(\theta; \mathbb{P}_n)$ to make explicit the dependence of the loss function L_n on the dataset $\mathbb{P}_n$. Following Ghosh and Basu (2016), we consider a generalised posterior based on a (contaminated) loss$L(\theta; \mathbb{P}_{n,\epsilon,y})$ with density $\pi_n^L(\theta; \mathbb{P}_{n,\epsilon,y})$, and define the *posterior influence function*

$$\mathrm{PIF}(y, \theta, \mathbb{P}_n) := \frac{\mathrm{d}}{\mathrm{d}\epsilon} \pi_n^L(\theta; \mathbb{P}_{n,\epsilon,y})|_{\epsilon=0}. \tag{12}$$

Here the notation $\pi_n^L(\theta; \mathbb{P}_{n,\epsilon,y})$ emphasises the dependence of the generalised posterior on the (contaminated) dataset $\mathbb{P}_{n,\epsilon,y}$. A generalised posterior π_n^L is called *globally bias-robust* if $\sup_{\theta\in\Theta} \sup_{y\in\mathcal{X}} |\mathrm{PIF}(y, \theta, \mathbb{P}_n)| < \infty$, meaning that the sensitivity of the generalised posterior to the

contaminant y is limited. The following lemma provides general sufficient conditions for global bias-robustness to hold:

Lemma 5 *Let π_n^L be a generalised Bayes posterior for a fixed $n \in \mathbb{N}$ with a loss $L(\theta; \mathbb{P}_n)$ and a prior π. Suppose $L(\theta; \mathbb{P}_n)$ is lower-bounded and $\pi(\theta)$ is upper-bounded over $\theta \in \Theta$, for any $\mathbb{P}_n$. Denote $\mathrm{D}L(y, \theta, \mathbb{P}_n) := (\mathrm{d}/\mathrm{d}\epsilon)L(\theta; \mathbb{P}_{n,\epsilon,y})|_{\epsilon=0}$. Then π_n^L is globally bias-robust if, for any $\mathbb{P}_n$,*

1. $\sup_{\theta\in\Theta} \sup_{y\in\mathcal{X}} |\mathrm{D}L(y, \theta, \mathbb{P}_n)|\, \pi(\theta) < \infty$, *and*
2. $\int_\Theta \sup_{y\in\mathcal{X}} |\mathrm{D}L(y, \theta, \mathbb{P}_n)|\, \pi(\theta)\mathrm{d}\theta < \infty$.

The proof is contained in Appendix B.6.1. Note that standard Bayesian inference does not satisfy the conditions of Lemma 5 in general. Indeed, when $L(\theta; \mathbb{P}_n)$ is the negative log likelihood, $\mathrm{D}L(y, \theta, \mathbb{P}_n) = \log p_\theta(y) - \sum_{i=1}^n \log p_\theta(x_i)$, and the term $\log p_\theta(y)$ can be unbounded over $y \in \mathcal{X}$. This can occur even if the statistical model is not heavy-tailed, e.g. for a normal location model p_θ on $\mathcal{X} = \mathbb{R}^d$. In contrast, the kernel K in KSD-Bayes provides a degree of freedom which can be leveraged to ensure that the conditions of Lemma 5 *are* satisfied; the specific form of $\mathrm{D}L(y, \theta, \mathbb{P}_n)$ for KSD-Bayes is derived in Appendix B.6.2. This enables us to derive sufficient conditions on K for global bias-robustness of KSD-Bayes, which we now present.

Theorem 3 (Globally Bias-Robust). *For each $\theta \in \Theta$, let $\mathbb{P}_\theta \in \mathcal{P}_{\mathrm{S}}(\mathbb{R}^d)$ and let $\mathcal{S}_{\mathbb{P}_\theta}$ denote the Langevin Stein operator in Equation* (5). *Let $K \in C_b^{1,1}(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^{d\times d})$. Suppose that π is bounded over Θ. If there exists a function $\gamma : \Theta \to \mathbb{R}$ such that*

$$\sup_{y\in\mathbb{R}^d} (\nabla_y \log p_\theta(y) \cdot K(y, y) \nabla_y \log p_\theta(y)) \leq \gamma(\theta) \tag{13}$$

and, in addition, $\sup_{\theta\in\Theta} |\pi(\theta)\gamma(\theta)| < \infty$ and $\int_\Theta \pi(\theta)\gamma(\theta)\mathrm{d}\theta < \infty$, then KSD-Bayes is globally bias-robust.

The proof is contained in Appendix B.6.3. The preconditions of Theorem 3 can be satisfied through an appropriate choice of kernel K; see Section 5.1. A comparison of KSD-Bayes to existing robust generalised Bayesian methodologies for tractable likelihood can be found in Appendix D.4. The difference in performance of robust and non-robust instances of KSD-Bayes is explored in detail in Section 6.

5 | DEFAULT SETTINGS FOR KSD-BAYES

The previous section considered β to be fixed, but an appropriate selection of β is essential to ensure the generalised posterior is calibrated. The choice of β is closely related to the choice of a Stein operator $\mathcal{S}_{\mathbb{P}_\theta}$ and kernel K; the purpose of this section is to recommend how these quantities are selected. If the recommendations of this section are followed, then KSD-Bayes has no remaining degrees of freedom to be specified.

5.1 | Default settings for $\mathcal{S}_{\mathbb{P}_\theta}$ and K

For Euclidean domains $\mathcal{X} = \mathbb{R}^d$, we advocate the default use of the Langevin Stein operator $\mathcal{S}_{\mathbb{P}_\theta}$ in Equation (5) and a kernel of the form

$$K(x,x') = \frac{M(x)M(x')^\top}{\left(1+(x-x')^\top\Sigma^{-1}(x-x')\right)^\gamma}, \tag{14}$$

where Σ is a positive definite matrix, $\gamma \in (0,1)$ is a constant, and $M \in C_b^1(\mathbb{R}^d;\mathbb{R}^{d\times d})$ will be called a matrix-valued *weighting function*.[5] For $M(x) = I_d$, (14) is called an *inverse multi-quadratic* (IMQ) kernel. The IMQ kernel and the Langevin Stein operator have appealing properties in the context of KSD. First, under mild conditions on $\mathbb{P}$, $\mathrm{KSD}(\mathbb{P}\|\mathbb{P}_n) \to 0$ implies that $\mathbb{P}_n$ converges weakly to $\mathbb{P}$ (Chen et al., 2019, Theorem 4). This convergence control ensures that small values of $\mathrm{KSD}(\mathbb{P}_\theta\|\mathbb{P}_n)$ imply similarity between $\mathbb{P}_\theta$ and $\mathbb{P}_n$ in the topology of weak convergence, so that minimising KSD is meaningful.[6] Secondly, and on a more practical level, the combination of Stein operator and IMQ kernel, with $\gamma = 1/2$, was found to work well in previous studies (Chen et al., 2019; Riabiz et al., 2021); we therefore also recommend $\gamma = 1/2$ as a default. The weighting function $M(x)$ facilitates an efficiency-robustness trade-off: If global bias robustness is *not* required then we recommend setting $M(x) = I_d$ as a default, which enjoys the aforementioned properties of KSD. If global bias-robustness *is* required then we recommend selecting $M(x)$ such that the supremum in Equation (13) exists and the preconditions of Theorem 3 are satisfied; see the worked examples in Section 6 and the further discussion in Appendix D.3.

The theoretical analysis of Section 4 assumed that K is fixed, but in our experiments we follow standard practice in the kernel methods community and recommend a data-adaptive choice of the matrix Σ. All experiments we report used the ℓ_1-regularised sample covariance matrix estimator of Ollila and Raninen (2019). The sensitivity of KSD-Bayes to the choice of kernel parameters is investigated in Appendix D.1.

5.2 | Default setting for β

For a simple normal location model, as described in Section 6.1, and in a well-specified setting, the asymptotic variance of the KSD-Bayes posterior with $\beta = 1$ is never smaller than that of the standard posterior. This provides a heuristic motivation for the default $\beta = 1$. However, in a misspecified setting smaller values of β are needed to avoid over-confidence in the generalised posterior, taking misspecification into account; see the recent review of Wu and Martin (2020). Here we aim to pick β such that the scale of the asymptotic precision matrix of the generalised posterior (H_*; Theorem 2) matches that of the minimum KSD point estimator ($H_* J_*^{-1} H_*$; Lemma 4), an approach proposed in Lyddon et al. (2019). This ensures the scale of the generalised posterior matches the scale of the sampling distribution of a closely related estimator whose frequentist properties can be analysed when the statistical model is misspecified. Since $\mathbb{P}$ is unknown, estimators of H_* and J_* are required. We propose the following default for β:

$$\beta = \min(1, \beta_n) \quad \text{where} \quad \beta_n = \frac{\mathrm{tr}(H_n J_n^{-1} H_n)}{\mathrm{tr}(H_n)}, \tag{15}$$

where the matrix H_* is approximated using $H_n := \nabla_\theta^2 \mathrm{KSD}^2(\mathbb{P}_\theta\|\mathbb{P}_n)|_{\theta=\theta_n}$, and the matrix J_* is approximated using

[5] The use of a non-constant weighting function is equivalent to replacing the Langevin Stein operator with a *diffusion* Stein operator whose *diffusion matrix* is $M(x)$; see Gorham et al. (2019).

[6] Note that other common kernels (e.g. Gaussian or Matérn kernels) fail to provide convergence control (Gorham & Mackey, 2017, Theorem 6).

$$J_n := \frac{1}{n}\sum_{i=1}^{n} S_n(x_i,\theta_n)S_n(x_i,\theta_n)^\top, \quad S_n(x,\theta) := \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta(\mathcal{S}_{\mathbb{P}_\theta}\mathcal{S}_{\mathbb{P}_\theta}K(x,x_i)).$$

The minimum of $\beta = 1$ and $\beta = \beta_n$ taken in Equation (15) provides a safeguard against selecting a value of β that over-shrinks the posterior covariance matrix — a phenomenon that we observed for the experiments reported in Sections 6.2–6.4, due to poor quality of the approximations H_n and J_n when n is small. The above expressions are derived for the exponential family model in Appendix B.7.

This completes our methodological and theoretical development, and next we turn to empirical performance assessment.

6 | EMPIRICAL ASSESSMENT

In this section, four distinct experiments are presented. The first experiment, in Section 6.1, concerns a normal location model, allowing the standard posterior and our generalised posterior to be compared and confirming our robustness results are meaningful. Section 6.2 presents a two-dimensional precision estimation problem, where standard Bayesian computation is challenging but computation with KSD-Bayes is trivial. Then, Section 6.3 presents a 25-dimensional kernel exponential family model, and Section 6.4 presents a 66-dimensional exponential graphical model. The kernel exponential family model allows us to explore a multi-modal dataset and to understand the potential limitations of KSD-Bayes in that context (c.f. Section 3.5). For all experiments, the default settings of Section 5 were used. An example of KSD-Bayes applied to a discrete dataset is presented in Appendix D.5.

6.1 | Normal location model

For expositional purposes we first consider fitting a normal location model $\mathbb{P}_\theta = \mathcal{N}(\theta, 1)$ to a dataset $\{x_i\}_{i=1}^n$. Our aim was to illustrate the robustness properties of KSD-Bayes, and we therefore generated the dataset using a contaminated data-generating model where, for each index $i = 1, \ldots, n$ independently, with probability $1-\epsilon$ the datum x_i was drawn from $\mathbb{P}_\theta$ with 'true' parameter $\theta = 1$, otherwise x_i was drawn from $\mathbb{P}_y = \mathcal{N}(y, 1)$, so that y and ϵ control, respectively, the nature and extent of the contamination in the dataset. The task is to make inferences for θ based on a contaminated dataset of size $n = 100$. The prior on θ was $\mathcal{N}(0, 1)$.

The standard Bayesian posterior is depicted in the leftmost panels of Figure 2, for varying ϵ (top row) and varying y (bottom row). Straightforward calculation shows that the expected posterior mean is $\frac{n}{n+1}[\theta + \epsilon(y-\theta)]$, which increases linearly as either y or ϵ are increased, with the other fixed. This behaviour is evident in the leftmost panels of Figure 2. The generalised posterior from KSD-Bayes is depicted in the central panels of Figure 2. This generalised posterior is slightly less sensitive to contamination compared to the standard posterior. Moreover, the variance slightly increases whenever either ϵ or y are increased, as a result of estimating β (c.f. Section 5.2). In the rightmost panels of Figure 2 we display the robust generalised posterior using the weighting function $M(x) = (1+x^2)^{-1/2}$, intended to bound the influence of large values in the dataset. This choice of $M(x)$ vanishes just fast enough as $|x| \to \infty$ to ensure that the bias-robustness conditions of Theorem 3 are satisfied; see Appendix D.3. The effect is clear from the bottom right panel of Figure 2, where even for $y = 20$ (and ϵ fixed to a small value, $\epsilon = 0.1$) the robust generalised

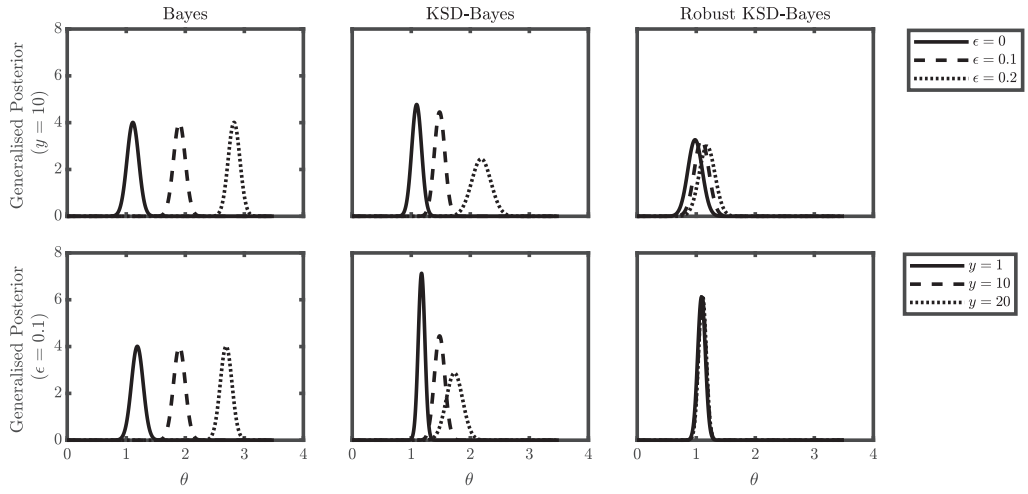


FIGURE 2 Posteriors and generalised posteriors for the normal location model. The true parameter value is $\theta = 1$, while a proportion ϵ of the data were contaminated by noise of the form $\mathcal{N}(y, 1)$. In the top row $y = 10$ is fixed and $\epsilon \in \{0, 0.1, 0.2\}$ are considered, while in the bottom row $\epsilon = 0.1$ is fixed and $y \in \{1, 10, 20\}$ are considered

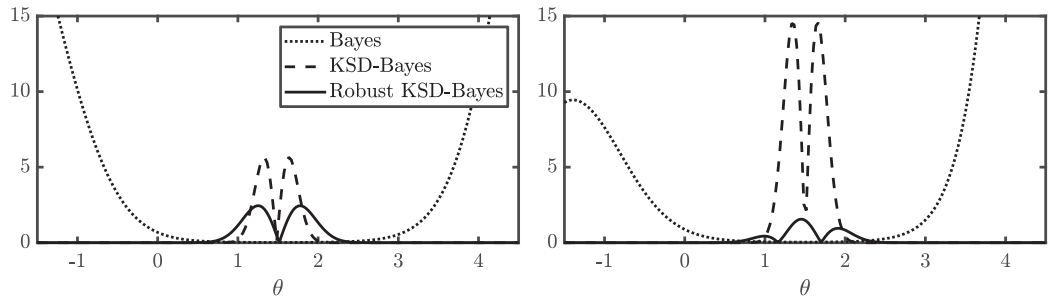


FIGURE 3 Posterior influence function for the normal location model

posterior remains centred close to the true value $\theta = 1$. While our theoretical results relate to y and do not guarantee robustness when ϵ is increased, the top right panel in Figure 2 suggests that the robust generalised posterior is indeed robust in this regime as well. Figure 3 displays the posterior influence function (12) for this normal location model. This reveals that the standard Bayesian posterior is not bias-robust, since the tails of the posterior are highly sensitive to the contaminant y. In contrast, the tails of the generalised posterior are insensitive to the contaminant. This appears to be the case for both weighting functions, despite only one weighting function satisfying the conditions of Theorem 3.

6.2 | Precision parameters in an intractable likelihood model

Our second experiment is a toy model due to Liu et al. (2019); an exponential family model $p_\theta(x) = \exp(\theta \cdot t(x) - a(\theta) + b(x))$ where $\theta \in \mathbb{R}^2$ are parameters to be inferred and $x \in \mathbb{R}^5$. The model specification is completed with

$$t(x) = (\tanh(x_{(4)}), \tanh(x_{(5)})), \quad b(x) = -0.5\sum_{i=1}^{5} x_{(i)}^2 + 0.6x_{(1)}x_{(2)} + 0.2\sum_{i=3}^{5} x_{(1)}x_{(i)}.$$

Despite the apparent simplicity of this model, the term $a(\theta)$, which determines the normalisation constant, is analytically intractable and exact simulation from this data-generating model is not straightforward (excluding the case $\theta = 0$). In sharp contrast, the generalised posterior produced by KSD-Bayes is available in closed form for this model. Our aim here was to assess robustness of the generalised posterior, focusing on the setting where y is fixed and ϵ is increased, since this is the regime for which our theoretical results do *not* hold. A dataset of size $n = 500$ was generated from the model $\mathbb{P}_\theta$ with true parameter $\theta = (0, 0)$, so that $\mathbb{P}_\theta$ has the form $\mathcal{N}(0, \Sigma)$ and can be exactly sampled. Each datum x_i was, with probability ϵ, shifted to $x_i + y$ where $y = (10, \ldots, 10)$. The prior on θ was $\mathcal{N}(0, 10^2 I)$.

The left column in Figure 4 displays the standard posterior,[7] which is seen to be sensitive to contamination in the dataset, in much the same way observed for the normal location model in Section 6.1. The generalised posterior with $M(x) = I_d$ is depicted in the middle column of Figure 4, and is seen to be *more* sensitive to contamination compared to the standard Bayesian posterior, in that the mean moves further from 0 as ϵ is increased. Finally, in the right column of Figure 4 we display the robust generalised posterior obtained with weighting function

$$M(x) = \mathrm{diag}\left(\left(1 + x_{(1)}^2 + \ldots + x_{(5)}^2\right)^{-1/2}, \left(1 + x_{(1)}^2 + x_{(2)}^2\right)^{-1/2}, \ldots, \left(1 + x_{(1)}^2 + x_{(5)}^2\right)^{-1/2}\right),$$

which ensures the criteria for bias-robustness in Theorem 3 are satisfied. From the figure, we observe that the robust generalised posterior remains centred close to the data-generating value $\theta = 0$, even for the largest contamination proportion considered ($\epsilon = 0.2$), with a variance that increases as ϵ is increased. At $\epsilon = 0$, the spread of the robust generalised posterior is almost twice that of the standard posterior, which reflects the trade-off between robustness and efficiency.

6.3 | Robust nonparametric density estimation

Our third experiment concerns density estimation using the kernel exponential family, and explores the performance of KSD-Bayes when the dataset is multi-modal (c.f. Section 3.5). Let q denote a reference p.d.f. on $\mathbb{R}^d$, and let $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a reproducing kernel. The *kernel exponential family* model (Canu & Smola, 2006)

$$p_\theta(x) \propto q(x) \exp(\langle f, \kappa(\cdot, x)\rangle_{\mathcal{H}(\kappa)}) \tag{16}$$

is parametrised by f, an element of the RKHS $\mathcal{H}(\kappa)$. The implicit normalisation constant of Equation (16), if it exists, is typically an intractable function of f. There appears to be no Bayesian or generalised Bayesian treatment of Equation (16) in the literature, which may be due to

[7] To obtain these results, the intractable normalisation constant was approximated using a numerical cubature method. To do this, we recognise that $p_\theta(x) = \mathcal{N}(x; 0, \Sigma) r_\theta(x)/C_\theta$ where $r_\theta(x) = \exp(\theta_1 \tanh(x_4) + \theta_2 \tanh(x_5))$. Then $C_\theta = \int r_\theta(x) \mathrm{d}\mathcal{N}(x; 0, \Sigma)$, which was approximated using (polynomial order 10) Gauss-Hermite cubature in 2D.

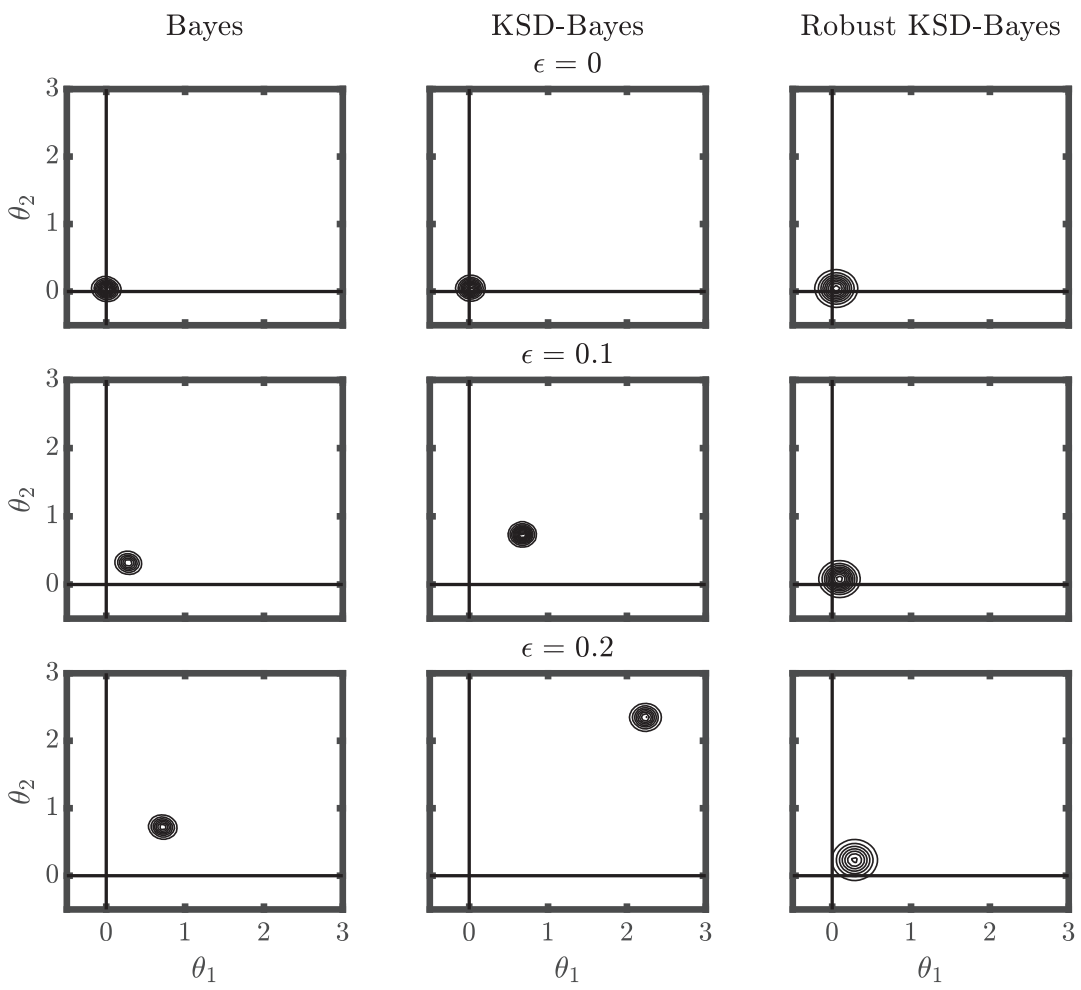


FIGURE 4 Posteriors and generalised posteriors for the Liu et al. (2019) model. The true parameter value is $\theta = 0$, while a proportion ϵ of the data were contaminated by being shifted by an amount $y = (10, 10)$

intractability of the likelihood. As the theory in this paper is finite-dimensional, we consider a finite-rank approximation of elements in $\mathcal{H}(\kappa)$ of the form $f(x) = \sum_{i=1}^{p} \theta_{(i)} \phi_{(i)}(x)$, with coefficients $\theta_{(i)} \in \mathbb{R}$ and basis functions $\phi_{(i)} \in \mathcal{H}(\kappa)$, where we will take θ to be $p = 25$ dimensional. Finite rank approximations have previously been considered for frequentist learning of kernel exponential families in Strathmann et al. (2015) and Sutherland et al. (2018). In our case, the finite rank approximation ensures that any prior we induce on f via a prior on the coefficients $\theta_{(i)}$ will be supported on $\mathcal{H}(\kappa)$. If one is interested in a well-defined limit as $p \to \infty$ then one will need to ensure a.s. convergence of the sum in this limit. If the ϕ_i are orthonormal in $\mathcal{H}(\kappa)$, and if the $\theta_{(i)}$ are a priori independent, then $\mathbb{E}[\|f\|^2_{\mathcal{H}(\kappa)}] = \sum_{i=1}^{p} \mathbb{E}[\theta^2_{(i)}]$ so a sufficient condition, for example, is $\mathbb{E}[\theta^2_{(i)}] = O(n^{-1-\delta})$ for some $\delta > 0$.

Our interest is in the performance of KSD-Bayes applied to a multi-modal dataset, and to explore these we considered the *galaxy data* of Postman et al. (1986) and Roeder (1990), comprising $n = 82$ velocities in km/sec of galaxies from 6 well-separated conic sections of a survey

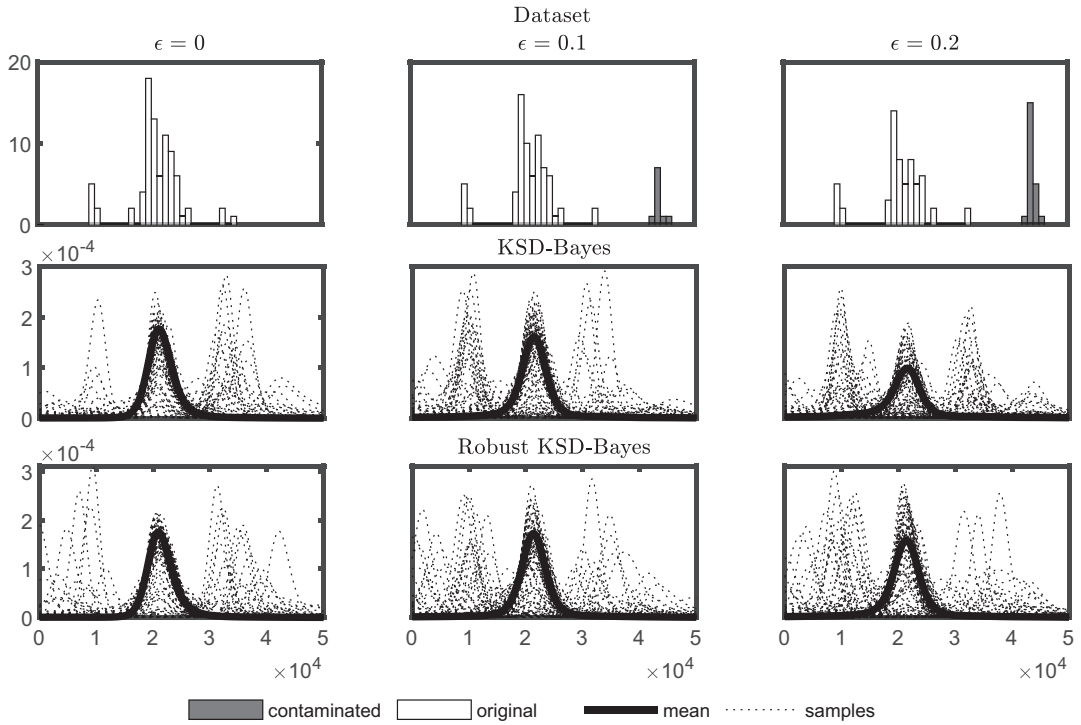


FIGURE 5 Generalised posteriors for the kernel exponential family model. A proportion ϵ of the data (top row) were contaminated. Samples from the generalised posteriors correspond to probability density functions, shown as dotted curves

of the *Corona Borealis*. The data were whitened prior to computation, but results are reported with the original scale restored. For the kernel exponential family we use $q(x) = \mathcal{N}(0, 3^2)$ and the kernel $\kappa(x, y) = \exp(-(x-y)^2/2)$, which ensures that (16) is normalisable due to Proposition 2 of Wenliang et al. (2019). For basis functions we use $\phi_{(i+1)}(x) = (x^i/\sqrt{i!})\exp(-x^2/2)$, $i = 0, \ldots, 24$, which are orthonormal in $\mathcal{H}(\kappa)$ (Steinwart et al., 2006). For our prior we let $\theta_{(i)} \sim \mathcal{N}(0, 10^2 i^{-1.1})$, which is weakly informative within the constraint of having a well-defined $p \to \infty$ limit. Our contamination model replaces a proportion ϵ of the dataset with values independently drawn from $\mathcal{N}(y, 0.1^2)$, with $y = 5$, shown as black bars in the top row of Figure 5.

The generalised posterior with $M(x) = 1$ is displayed in the second row of Figure 5, with the bottom row presenting a robust generalised posterior based on the weighting function $M(x) = (1 + x^2)^{-1/2}$, which ensures the conditions of Theorem 3 are satisfied. The results we present are for fixed y and increasing ϵ, since this regime is *not* covered by Theorem 3. The generalised posterior mean is a uni-modal density, which we attribute to the insensitivity of KSD to mixture proportions discussed in Section 3.5, but multi-modal densities are evident in sampled output. Our results indicate that the robust weighting function reduces sensitivity to contamination in the dataset (note how the mass in the central mode of the generalised posterior decreases when $\epsilon = 0.2$, when the identity weighting function is used). Whether this insensitivity of KSD to well-separated regions in the dataset is desirable or not will depend on the application, but in this case, it happens to be beneficial.

6.4 | Network inference with exponential graphical models

Our final example concerns an exponential graphical model, representing negative conditional relationships among a collection of random variables $W = (W_1, \ldots, W_d)$, described in (Yang et al. 2015 Section 2.5). The likelihood function is

$$p_{W|\theta}(w|\theta) \propto \exp\left(-\sum_i \theta_{(i)} w_{(i)} - \sum_{i<j} \theta_{(i,j)} w_{(i)} w_{(j)}\right), \tag{17}$$

where $w \in (0, \infty)^d$ and $\theta_{(i)} > 0, \theta_{(i,j)} \geq 0$. The total number of parameters is $p = d(d+1)/2$. Simulation from this model is challenging and the normalisation constant is an intractable integral, so in what follows a standard Bayesian analysis is not attempted. Our aim was to fit (17) to a protein kinase dataset, mimicking an experiment presented by Yu et al. (2016) in the score-matching context. This dataset, originating in Sachs et al. (2005), consists of quantitative measurements of $d = 11$ phosphorylated proteins and phospholipids, simultaneously measured from single cells using a fluorescence-activated cell sorter, so the parameter θ is 66-dimensional. Nine stimulatory or inhibitory interventional conditions were combined to give a total of 7,466 cells in the dataset. The data were square-root transformed and samples containing values greater than 10 standard deviations from their mean were judged to be *bona fide* outliers and were removed. The remaining dataset of size $n = 7{,}449$ was normalised to have unit standard deviation. In most cases the measurement reflects the activation state of the kinases, and scientific interest lies in the mechanisms that underpin their interaction.[8] These mechanisms are often summarised as a *protein signalling network*, whose nodes are the d proteins and whose edges correspond to the pairs of proteins that interact. An important statistical challenge is to *estimate* a protein signalling network from such a dataset (Oates, 2013). However, it is known that existing approaches to *network inference* are non-robust, in a general sense, with community challenges regularly highlighting the different conclusions drawn by different estimators applied to an identical dataset (Hill et al., 2016). Our interest is in whether networks estimated using KSD-Bayes are robust.

For our experiment the variables $w_{(i)}$ were re-parametrised as $x_{(i)} := \log(w_{(i)})$, in order that they are unconstrained and $\mathbb{P}_\theta \in \mathcal{P}_S(\mathbb{R}^d)$. For the contamination model, a proportion ϵ of the data were replaced with the fixed value $y = (10, \ldots, 10) \in \mathbb{R}^d$. Parameters were a priori independent with $\theta_{(i)} \sim \mathcal{N}_T(0, 1)$, $\theta_{(i,j)} \sim \mathcal{N}_T(0, 1)$, where $\mathcal{N}_T$ is the Gaussian distribution truncated to the positive orthant of $\mathbb{R}^p$. This prior is conjugate to the likelihood, as explained in Section 3.3, and allows the generalised posterior to be exactly computed. Generalised posteriors were produced both without and with the exponential weighting function $[M(x)]_{(i,i)} = \exp(-x_{(i)})$, the latter aiming to reduce sensitivity to large values in the dataset and coinciding with the identity weighting function at $x = 0$. From these, protein signalling networks were estimated using the s most significant edges, defined as the s largest values of $\overline{\theta}_{(i,j)}/\sigma_{(i,j)}$, where the generalised posterior marginal for $\theta_{(i,j)}$ is $\mathcal{N}_T(\overline{\theta}_{(i,j)}, \sigma^2_{(i,j)})$. Results are shown in Figure 6; to optimise visualisation we report results for $s = 5$, though for other values of s similar conclusions hold. It is interesting to observe little agreement between the networks returned when the identity weighting function is used, which may reflect the difficulty of the network inference task. Reduced sensitivity to ϵ was observed when the exponential weighting function was used. In Figure 6 we report the number of edges that are

[8]There is no scientific basis to expect only negative conditional dependencies in the dataset; in this sense the model is likely to be misspecified. Our interest is in assessing the robustness properties of KSD-Bayes only, and no scientific conclusions will be drawn using this model.

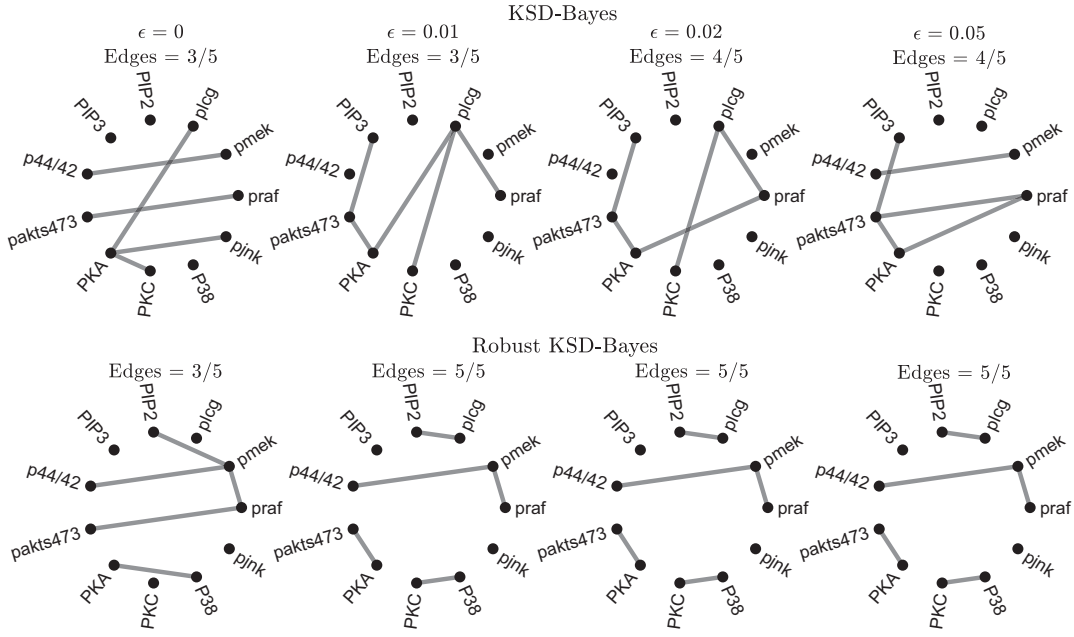


FIGURE 6 Exponential graphical model; estimated protein signalling networks as a function of the proportion ϵ of contamination in the dataset

consistent with the network reported in (Sachs et al. 2005 Figure 3A); the use of the exponential weighting function resulted in more edges being consistent with this benchmark network.

7 | CONCLUSION

There is little existing literature concerning robust Bayesian inference in the setting of intractable likelihood. Existing approaches to Bayesian inference for intractable likelihood fall into three categories: (1) likelihood-free methods such as *approximate Bayesian computation* and *Bayesian synthetic likelihood*; Beaumont et al., 2002; Cherief-Abdellatif & Alquier, 2020; Frazier, 2020; Marin et al., 2012; Price et al., 2018; Tavaré et al., 1997), (2) auxiliary variable MCMC (such as the exchange algorithm and pseudo-marginal MCMC; Andrieu & Roberts, 2009; M⊘ller et al., 2006; Murray et al., 2006; Park & Haran, 2018), and (3) approximate likelihood methods (such as *pseudo-likelihood* and *composite likelihood*; Besag, 1974; Dryden et al., 2002; Eidsvik et al., 2014), which are of course also applicable beyond the Bayesian context. Both (1) and (2) rely on either the ability to (exactly or approximately) simulate from the generative model, or the ability to unbiasedly estimate the data likelihood, whilst (3) represents a collection of approaches that are tailored to particular statistical models. These algorithms aim to approximate the standard Bayesian posterior, and do not attempt to confer robustness in situations where the model is misspecified.

This paper proposed KSD-Bayes, a generalised Bayesian procedure for likelihoods that involve an intractable normalisation constant. KSD-Bayes provides robust generalised Bayesian inference in this context, including a theoretical guarantee of global bias-robustness over Θ. Moreover, and unlike existing Bayesian approaches to intractable likelihood, the generalised posterior can

be approximated by standard sampling methods without additional levels of algorithmic complexity, even admitting conjugate analysis for the exponential family model. From a theoretical perspective, the soundness of KSD-Bayes, in terms of consistency and asymptotic normality of the generalised posterior, was established.

Although KSD-Bayes has several appealing features, it is not a panacea for intractable likelihood. The generalised posterior is not invariant to transformations of the dataset and, as discussed in Section 3.5, KSD can suffer from insensitivity to mixture proportions, which limits its applicability to models and datasets that are not 'too multi-modal'. The selection of β remains an open problem for generalised Bayesian inference, and further regularisation may be required when the parameter θ is high-dimensional relative to the size n of the dataset. These are challenging issues for future work. In addition, our experiments focused on continuous data, though our theory was general. The empirical performance of KSD-Bayes for discrete data remains to be assessed.

ACKNOWLEDGEMENTS

TM was supported EPSRC grant EP/N510129/1 at the Alan Turing Institute, UK. JK was funded by EPSRC grant EP/L016710/1, the Facebook Fellowship Programme, as well as the Biometrika Fellowship. FXB and CJO were supported by the Lloyd's Register Foundation programme on data-centric engineering at The Alan Turing Institute under the EPSRC grant EP/N510129/1. The authors thank the Associate Editor and three Reviewers for detailed feedback that led to an improved manuscript, and Oscar Key for pointing out an indexing error in an earlier version of the manuscript.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

ORCID

Takuo Matsubara https://orcid.org/0000-0003-2177-3056
François-Xavier Briol https://orcid.org/0000-0002-0181-2559

REFERENCES

Amari, S. (1997) Information geometry. *Contemporary Mathematics*, 203, 81–96.

Andrieu, C. & Roberts, G.O. (2009) The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2), 697–725.

Baraud, Y. & Birgé, L. (2020) Robust Bayes-like estimation: Rho-Bayes estimation. *The Annals of Statistics*, 48(6), 3699–3720.

Barp, A., Briol, F.-X., Duncan, A., Girolami, M. & Mackey, L. (2019) Minimum Stein discrepancy estimators. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*.

Baydin, A.G., Pearlmutter, B.A., Radul, A.A. & Siskind, J.M. (2018) Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153), 1–43.

Beaumont, M.A., Zhang, W. & Balding, D.J. (2002) Approximate Bayesian computation in population genetics. *Genetics*, 162(4), 2025–2035.

Berger, J., Moreno, E., Pericchi, L., Bayarri, M., Bernardo, J., Cano, J. et al. (1994) An overview of robust Bayesian analysis. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 3(1), 5–124.

Bernardo, J.M. & Smith, A.F. (2009) *Bayesian theory*. Hoboken: John Wiley & Sons.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2), 192–236.

Besag, J. (1986) On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(3), 259–302.

Bissiri, P.G., Holmes, C.C. & Walker, S.G. (2016) A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 78(5), 1103.

Canu, S. & Smola, A. (2006) Kernel methods and the exponential family. *Neurocomputing*, 69(7-9), 714–720.

Chen, W.Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L. et al. (2019) Stein point Markov chain Monte Carlo. In: *Proceedings of the 36th International Conference on Machine Learning*, pp. 1011–1021.

Cherief-Abdellatif, B.-E. & Alquier, P. (2020) MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. In: *Proceedings of the 2nd Symposium on Advances in Approximate Bayesian Inference*, pp. 1–21.

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

Chwialkowski, K., Strathmann, H. & Gretton, A. (2016) A kernel test of goodness of fit. In: *Proceedings of the 33rd International Conference on Machine Learning*, pp. 2606–2615.

Diggle, P.J. (1990) A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 153(3), 349–362.

Dryden, I., Ippoliti, L. & Romagnoli, L. (2002) Adjusted maximum likelihood and pseudo-likelihood estimation for noisy Gaussian Markov random fields. *Journal of Computational and Graphical Statistics*, 11(2), 370–388.

Eidsvik, J., Shaby, B.A., Reich, B.J., Wheeler, M. & Niemi, J. (2014) Estimation and prediction in spatial models with block composite likelihoods. *Journal of Computational and Graphical Statistics*, 23(2), 295–315.

Frazier, D.T. (2020) Robust and efficient approximate Bayesian computation: A minimum distance approach. *arXiv:2006.14126*.

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

Giummolè, F., Mameli, V., Ruli, E. & Ventura, L. (2019) Objective Bayesian inference with proper scoring rules. *Test*, 28(3), 728–755.

Gong, W., Li, Y. & Hernández-Lobato, J.M. (2021) Sliced kernelized Stein discrepancy. In: *Proceedings of the 9th International Conference on Learning Representations*.

Gorham, J. & Mackey, L. (2015) Measuring sample quality with Stein's method. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems*.

Gorham, J. & Mackey, L. (2017) Measuring sample quality with kernels. In: *Proceedings of the 34th International Conference on Machine Learning*, pages 1292–1301.

Gorham, J., Duncan, A.B., Vollmer, S.J. & Mackey, L. (2019) Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5), 2884–2928.

Gorham, J., Raj, A. & Mackey, L. (2020) Stochastic Stein discrepancies. In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Grünwald, P. (2011) Safe learning: Bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity. In: *Proceedings of the 24th Annual Conference on Learning Theory*, pages 397–420.

Grünwald, P. (2012) The safe Bayesian. In: *Proceedings of the 23rd International Conference on Algorithmic Learning Theory*, pages 169–183.

Grünwald, P. & van Ommen, T. (2017) Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4), 1069–1103.

de Heide, R., Kirichenko, A., Grunwald, P. & Mehta, N. (2020) Safe-bayesian generalized linear regression. Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, 108, 2623–2633,

Hill, S.M., Heiser, L.M., Cokelaer, T., Unger, M., Nesser, N.K., Carlin, D.E. et al. (2016) Inferring causal molecular networks: empirical assessment through a community-based effort. *Nature Methods*, 13(4), 310–318.

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

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

Huber, P.J. & Ronchetti, E.M. (2009) *Robust statistics*. Hoboken: Wiley.

Huggins, J.H. & Mackey, L. (2018) Random feature Stein discrepancies. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 1903–1913.

Huggins, J.H. & Miller, J.W. (2020) Robust inference and model criticism using bagged posteriors. *arXiv: 1912.07104*.

Hyvärinen, A. (2005) Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24), 695–709.

Jewson, J., Smith, J.Q. & Holmes, C. (2018) Principled Bayesian minimum divergence inference. *Entropy*, 20(6), 442.

Jiang, X., Li, Q. & Xiao, G. (2021) Bayesian modeling of spatial transcriptomics data via a modified Ising model. *arXiv:2104.13957*.

Kleijn, B.J. & van der Vaart, A.W. (2012) The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6, 354–381.

Knoblauch, J., Jewson, J. & Damoulas, T. (2019) Generalized variational inference: Three arguments for deriving new posteriors. *arXiv:1904.02063*.

Liu, Q., Lee, J. & Jordan, M. (2016) A kernelized Stein discrepancy for goodness-of-fit tests. In: *Proceedings of the 33rd International Conference on Machine Learning*, pages 276–284.

Liu, S., Kanamori, T., Jitkrittum, W. & Chen, Y. (2019) Fisher efficient inference of intractable models. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*.

Lyddon, S.P., Holmes, C.C. & Walker, S.G. (2019) General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2), 465–478.

M∅ller, J., Pettitt, A.N., Reeves, R. & Berthelsen, K.K. (2006) An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2), 451–458.

Ma, Y.-A., Chen, T. & Fox, E. (2015) A complete recipe for stochastic gradient MCMC. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 2917–2925.

Marin, J.-M., Pudlo, P., Robert, C.P. & Ryder, R. (2012) Approximate Bayesian computational methods. *Statistics and Computing*, 22(6), 1167–1180.

Miller, J.W. (2021) Asymptotic normality, concentration, and coverage of generalized posteriors. *Journal of Machine Learning Research*, 22(168), 1–53.

Miller, J.W. & Dunson, D.B. (2019) Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527), 1113–1125.

Moores, M., Nicholls, G., Pettitt, A. & Mengersen, K. (2020) Scalable Bayesian inference for the inverse temperature of a hidden Potts model. *Bayesian Analysis*, 15(1), 1–27.

Müller, U.K. (2013) Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, 81(5), 1805–1849.

Murray, I., Ghahramani, Z. & MacKay, D.J.C. (2006) MCMC for doubly-intractable distributions. In: *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*, pages 359–366.

Murray, I., Adams, R.P. & MacKay, D.J.C. (2010) Elliptical slice sampling. *The Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 9, 541–548.

Nakagawa, T. & Hashimoto, S. (2020) Robust Bayesian inference via-divergence. *Communications in Statistics - Theory and Methods*, 49(2), 343–360.

Oates, C.J. (2013) Bayesian inference for protein signalling networks. PhD thesis, University of Warwick.

Ollila, E. & Raninen E. (2019) Optimal shrinkage covariance matrix estimation under random sampling from elliptical distributions. *IEEE Transactions on Signal Processing*, 67(10), 2707–2719.

Park, J. & Haran, M. (2018) Bayesian inference in the presence of intractable normalizing functions. *Journal of the American Statistical Association*, 113(523), 1372–1390.

Postman, M., Huchra, J. & Geller, M. (1986) Probes of large-scale structure in the corona borealis region. *The Astronomical Journal*, 92, 1238–1247.

Price, L.F., Drovandi, C.C., Lee, A. & Nott, D.J. (2018) Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, 27(1), 1–11.

Riabiz, M., Chen, W., Cockayne, J., Swietach, P., Niederer, S.A., Mackey, L. et al. (2021) Optimal thinning of MCMC output. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, To appear.

Roeder, K. (1990) Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85(411), 617–624.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A. & Nolan, G.P. (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721), 523–529.

Stein, C. (1972) A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, Volume 2: Probability Theory.

Steinwart, I., Hush, D. & Scovel, C. (2006) An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Transactions on Information Theory*, 52(10), 4635–4643.

Strathmann, H., Sejdinovic, D., Livingstone, S., Szabo, Z. & Gretton, A. (2015) Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems*.

Sutherland, D.J., Strathmann, H., Arbel, M. & Gretton, A. (2018) Efficient and principled score estimation with Nyström kernel exponential families. In: *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pages 652–660.

Tavaré, S., Balding, D.J., Griffiths, R.C. & Donnelly, P. (1997) Inferring coalescence times from DNA sequence data. *Genetics*, 145(2), 505–518.

Wenliang, L.K. (2020) Blindness of score-based methods to isolated components and mixing proportions. *arXiv:2008.10087*.

Wenliang, L., Sutherland, D.J., Strathmann, H. & Gretton, A. (2019) Learning deep kernels for exponential family densities. In: *Proceedings of the 36th International Conference on Machine Learning*, pp. 6737–6746.

Williams, P.M. (1980) Bayesian conditionalisation and the principle of minimum information. *The British Journal for the Philosophy of Science*, 31(2), 131–144.

Wu, P.-S. & Martin, R. (2020) A comparison of learning rate selection methods in generalized Bayesian inference. *arXiv:2012.11349*.

Yang, E., Ravikumar, P., Allen, G.I. & Liu, Z. (2015) Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16(115), 3813–3847.

Yu, M., Kolar, M. & Gupta, V. (2016) Statistical inference for pairwise graphical models using score matching. In: *Proceedings of the 29th International Conference on Neural Information Processing Systems*.

Zellner, A. (1988) Optimal information processing and Bayes's theorem. *The American Statistician*, 42(4), 278–280.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Matsubara, T., Knoblauch, J., Briol, F.-X. & Oates, C.J. (2022) Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1–26. Available from: https://doi.org/10.1111/rssb.12500