Statistics and Computing (2023) 33:39
https://doi.org/10.1007/s11222-023-10205-7

ORIGINAL PAPER

Adaptation of the tuning parameter in general Bayesian inference with robust divergence

Shouto Yonekura[1] · **Shonosuke Sugasawa**[2,3]

Received: 8 July 2022 / Accepted: 8 January 2023 / Published online: 4 February 2023

Abstract
We introduce a novel methodology for robust Bayesian estimation with robust divergence (e.g., density power divergence or γ-divergence), indexed by tuning parameters. It is well known that the posterior density induced by robust divergence gives highly robust estimators against outliers if the tuning parameter is appropriately and carefully chosen. In a Bayesian framework, one way to find the optimal tuning parameter would be using evidence (marginal likelihood). However, we theoretically and numerically illustrate that evidence induced by the density power divergence does not work to select the optimal tuning parameter since robust divergence is not regarded as a statistical model. To overcome the problems, we treat the exponential of robust divergence as an unnormalisable statistical model, and we estimate the tuning parameter by minimising the Hyvarinen score. We also provide adaptive computational methods based on sequential Monte Carlo samplers, enabling us to obtain the optimal tuning parameter and samples from posterior distributions simultaneously. The empirical performance of the proposed method through simulations and an application to real data are also provided.

Keywords General Bayes · Robustness · Tuning parameter estimation · Density power divergence · Sequential Monte Carlo

1 Introduction

One well-known way to deal with outliers and model misspecification when conducting inference is to use robust divergences. Since the pioneering work of Basu et al. (1998) that proposed density power divergence as an extension of the standard likelihood, some variants of the divergence (e.g. Fujisawa and Eguchi 2008; Cichocki et al. 2011; Ghosh et al. 2017) and various statistical methods using robust divergences have been developed. Many robust divergences are indexed by a single tuning parameter that controls the robustness against outliers. If the tuning parameter is set to a smaller value than necessary, the resulting estimator may still be affected by outliers. On the other hand, using an unnecessarily large value for the tuning parameter leads to a loss of statistical efficiency (Basu et al. 1998). Despite the success of the theoretical analysis of properties of statistical methods based on robust divergences, how to adaptively estimate the tuning parameter from the data has often been ignored, with a few exceptions (Warwick and Jones 2005; Basak et al. 2021) that propose frequentist methods to select the optimal value via asymptotic mean squared errors. There is a growing body of literature on Bayesian approaches using robust divergence. For example, general theory has been considered by (Ghosh and Basu 2016; Jewson et al. 2018; Nakagawa and Hashimoto 2020) and some specific applications to statistical models such as linear regression (Hashimoto and Sugasawa 2020), change point detection (Knoblauch et al. 2018) and Bayesian filtering (Boustati et al. 2020). Nevertheless, a reasonable estimation strategy for the tuning parameter has not been carefully discussed. A natural consideration to find the best tuning parameter in the context of Bayesian statistics will be the use of model evidence or marginal likelihood. However, as we shall illustrate later, evidence is not useful for choosing the tuning parameter since the exponential of robust divergence cannot be directly interpreted as a normalised statistical model.

In this paper, this issue is addressed by taking advantage of ideas from statistical theories for unnormalised statistical models (Hyvärinen 2005) introducing the Hyvarinen

✉ Shonosuke Sugasawa
sugasawa@csis.u-tokyo.ac.jp

[1] Graduate School of Social Sciences, Chiba University, Chiba, Japan

[2] Center for Spatial Information Science, The University of Tokyo, Chiba, Japan

[3] Nospare Inc., Tokyo, Japan

score (H score), which is a finite version of Fisher divergence. Based upon the idea of Hyvärinen (2005), Dawid and Musio (2015), Shao et al. (2019) consider unnormalisable marginal likelihoods, with particular attention to model selection, where such unnormalisability is driven by the improper prior. Our main idea is to regard the exponential of robust divergence as unnormalisable models and employ a posterior distribution driven by the H-score, inspired by Dawid and Musio (2015), Shao et al. (2019), as an objective function of γ. Since our objective function cannot be computed analytically in general, we then take advantage of sequential Monte Carlo (SMC) samplers (Del Moral et al. 2006) within a Robbins-Monro stochastic gradient framework to approximate the objective function, to estimate the optimal tuning parameter and to obtain samples from posterior distributions.

Therefore, our work can be understood as an attempt to fill the current gap between the existing theory of such robust estimation and their practical implementation. Our proposed method has the following advantages over existing studies.

(i) Unlike existing methods (Warwick and Jones 2005; Basak et al. 2021), our proposed method does not require a pilot plot. To optimise a tuning parameter, it is necessary to determine a certain value as a pilot estimate. Thus, the estimates may be strongly dependent on the pilot estimate. In addition, such methods often estimate excessively large values of the tuning parameters, given the proportion of outliers in the data. In contrast, our algorithm is stable and statistically efficient since that does not require a pilot estimate.
(ii) Proposed methods in (Warwick and Jones 2005; Basak et al. 2021) require an analytical representation of the asymptotic variance that cannot be obtained in general. Compared to such methods, our proposed method does not require such expression, and therefore our method can be applied to rather complex statistical models, which seem difficult to be handled by the previous methods.
(iii) We take advantage of SMC samplers (Del Moral et al. 2006) in the generalised Bayesian framework (Bissiri et al. 2016) with a gradient-ascent approach to perform parameter inference for the tuning parameter and posterior sampling simultaneously. This is a unique favourable algorithmic characteristic compared to methods that estimate parameters by fixing tuning parameters or by estimating tuning parameters once and then estimating other objects of interest.

Recently, Jewson et al. (2021) has introduced a new Bayesian framework called *H-posterior* for unnormalisable statistical models based on Fisher divergence, and they have developed model selection criterion via the Laplace approximation of the marginal likelihood. The biggest difference from Jewson et al. (2021) is that we use a natural form of general posterior based on robust divergence, which is widely adopted in the literature (e.g. Ghosh and Basu 2016; Jewson et al. 2018), while the form of the posterior distribution in Jewson et al. (2021) is different from ours. As we mentioned, the main contribution of their work is the construction of model selection criteria of the BIC type through the Laplace approximation and the proof of their consistency. On the other hand, our research is about the estimation of tuning parameters and the inference of the posterior distribution, and the main objective is to propose an objective function and a computational method considered suitable for this purpose. In the framework of Generalised Bayesian, some methods that use variational Bayesian inference (Knoblauch et al. 2019; Frazier et al. 2021) have also been proposed. However, instead of computational speed, the approximate distribution obtained does not match the target distribution in the limit, and the method of estimating the tuning parameters is unclear. There is also a need to make some natural but somewhat stronger assumptions than our proposed method, such as that the target distribution is an exponential family.

The rest of the paper is organised as follows. In Sect. 2, we first set up the framework and then show theoretically and numerically that evidence induced by density power divergence to select the tuning parameter. Instead, we propose to estimate it based on the H-score (Hyvärinen 2005; Dawid and Musio 2015) and characterise its asymptotic behaviour. As mentioned earlier, our method involves functions for which it is difficult to obtain an analytic representation. Therefore, we develop an adaptive and efficient Markov chain Monte Carlo (MCMC) algorithm based on SMC samplers (Del Moral et al. 2006) in Sect. 3. Numerical applications of the developed methodology are provided in Sect. 4, then conclusions and directions for future research are provided in Sect. 5.

2 Bayesian inference with robust divergence

2.1 General posterior distribution

Suppose that we have d_y-dimensional $i.i.d.$ data $\{y_i\}_{i=1}^n \overset{i.i.d.}{\sim} G$ where G denotes the true distribution or the data-generating process. Also, assume that we have a statistical model $\{f_\theta : \theta \in \Theta\}$ where $\Theta \subseteq \mathbb{R}^d$ for some $d \geq 1$. We then write $y_{1:n}$ to denote $(y_1, \ldots, y_n)$ and let g denote the density of G with respect to dy. To make robust Bayesian inferences for θ, we use a potential function based on robust divergence instead of the standard likelihood function.

Here, we simply consider *the density power divergence* (Basu et al. 1998) but other ones with tuning parameters, such as γ-divergence (Fujisawa and Eguchi 2008; Nakagawa and Hashimoto 2020), α-divergence (Cichocki and Amari

2010), Hölder divergence (Nielsen et al. 2017), also can be used within our framework. Given a prior density $\pi(\theta)$ with respect to $d\theta$, we can define the corresponding posterior density

$$\Pi_\gamma(\theta \mid y_{1:n}) := \frac{\mathcal{L}_\gamma(y_{1:n};\theta)\pi(\theta)}{p_\gamma(y_{1:n})}, \tag{2.1}$$

where $\gamma \in (0, \infty]$, $p_\gamma(y_{1:n}) := \int_\Theta \mathcal{L}_\gamma(y_{1:n};\theta)\pi(\theta)d\theta$ and

$$\log \mathcal{L}_\gamma(y_{1:n};\theta) := \sum_{i=1}^{n} \log \mathcal{L}_\gamma(y_i;\theta),$$
$$\log \mathcal{L}_\gamma(y_i;\theta) := \frac{1}{\gamma} f_\theta(y_i)^\gamma - \frac{1}{1+\gamma}\int f_\theta(x)^{1+\gamma}dx. \tag{2.2}$$

Note that $p_\gamma(y_{1:n})$ is generally referred to as *evidence*. In many scenarios, robustified posterior densities such as (2.1) give much more accurate and stable inference against outliers and theoretical properties of the posterior have been investigated in (Ghosh and Basu 2016; Nakagawa and Hashimoto 2020). However, its performance depends critically on the choice of the tuning parameter γ in (2.1) (e.g. Ghosh and Basu 2016), which motivate us to find "best" γ to make inference successful. Notice that (2.1) can be seen as a special case of general Bayesian updating (Bissiri et al. 2016) with weight setting 1. As noted in Jewson et al. (2018), the density power divergence does not have any arbitrariness in the scale as a loss function, and one can set $\omega = 1$. Under the general framework of Bayesian updating, Corollary 1 of Fong and Holmes (2020) implies that evidence is still the unique coherent marginal score for Bayesian inference. Thus, from the viewpoint of Bayesian statistics, it appears to be natural to find the best γ based on evidence, but its property of it is unclear since $\mathcal{L}_\gamma(y_{1:n};\theta)$ is not a probability density of $y_{1:n}$. Furthermore, the tuning parameter γ cannot be interpreted as "model parameter" in this case. The following example highlights the problem of using $p_\gamma(y_{1:n})$ to find the best γ.

2.2 Failure of model evidence: motivating example

To see why evidence is not useful for estimating γ, we start with the following proposition for a rescaled $\log \mathcal{L}_\gamma(y_i;\theta)$.

Proposition 1 *Consider* $\log \mathcal{L}^{\mathbf{R}}_\gamma(y_i;\theta) := \log \mathcal{L}_\gamma(y_i;\theta) - \frac{1}{\gamma} + 1$. *Furthermore, assume that* $f(x) \le 1$ *for any* x. *Then* $\log \mathcal{L}^{\mathbf{R}}_\gamma(y_i;\theta)$ *is a monotonically increasing function of* γ.

Proof See "Appendix A". □

Since the term $-\frac{1}{\gamma}+1$ is eliminated when considering the posterior distribution (2.1), this rescaling is a non-essential modification in the method we shall propose later. The meaning of the rescaling is to ensure that $\log \mathcal{L}^{\mathbf{R}}_\gamma(y_i;\theta)$ converges to the log-likelihood as $\gamma \to 0$, so that $\log \mathcal{L}^{\mathbf{R}}_\gamma(y_i;\theta)$ can be regarded as a natural extension of the log-likelihood.

The important point here is that Proposition 1 implies that there are theoretically at least some situations where evidence is increasing monotonically for γ. Indeed, the following numerical example vividly illustrates such a situation. To see this numerically, we consider a simple but motivating example in which $\{y_i\}_{i=1}^{100} \overset{i.i.d.}{\sim} G = \mathcal{N}(1,1)$ and then randomly replace $\tau\%$ of $\{y_i\}_{i=1}^{100}$ by $y_i + 5$, where $0 \le \tau \le 100$ is called the contamination proportion. Here γ was determined by dividing equally $[0.01, 1]$ into 1,000 points. In other words, in the context of Bayesian model selection, this corresponds to choosing the model with the largest evidence as to the best model out of 1,000 models indexed by γ. With the choice $\tau = 10$, we then calculated 2000 Monte Carlo estimates of the model evidence $p_\gamma(y_{1:n})$ for each γ. The resulting $p_\gamma(y_{1:n})$ are shown in Fig. 1, which numerically shows that $p_\gamma(y_{1:n})$ is a monotonically increasing function of γ so that it does not have local maxima. This implies that one cannot estimate γ using $p_\gamma(y_{1:n})$. A similar phenomenon is also discussed in Jewson et al. (2021).

2.3 Estimation using H-score

To overcome the illustrated problem, we first treat $\mathcal{L}_\gamma(y_{1:n};\theta)$ as an unnormalised statistical model motivated by Hyvärinen (2005). Note that even with an unnormalisable model, the update in (2.1) can be considered as a valid belief update according to Bissiri et al. (2016). It should be noted that when $\log \mathcal{L}_\gamma(y_{1:n};\theta)$ is the density power divergence of the form (2.2), the normalising constant may not exist. For example, when f_θ is a normal distribution, $\log \mathcal{L}_\gamma(y_i;\theta)$ converges to a constant value under $|y_i| \to \infty$, so the integral of $\mathcal{L}_\gamma(y_{1:n};\theta)$ with respect to $y_{1:n}$ diverges. Recently, Jewson et al. (2021) has pointed out that the role of such unnormalisable models can be recognised in terms of relative probability.

For d_y dimensional observations y and twice differentiable density $p(\cdot)$, Hyvärinen (2005) defines the H-score as

$$\mathcal{H}(y, p) := \sum_{k=1}^{d_y}\left\{2\frac{\partial^2 \log p(y)}{\partial y_{(k)}^2} + \left(\frac{\partial \log p(y)}{\partial y_{(k)}}\right)^2\right\}.$$

We then select the optimal γ with the smallest leave-one-out H-score, defined as

$$\sum_{i=1}^{n} \mathcal{H}(y_i, p_\gamma(y_i|y_{-i})), \tag{2.3}$$

where $p_\gamma(y_i|y_{-i}) = \int \mathcal{L}_\gamma(y_i;\theta)\Pi_\gamma(\theta|y_{-i})d\theta$ and $y_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$. Note that Shao et al. (2019)

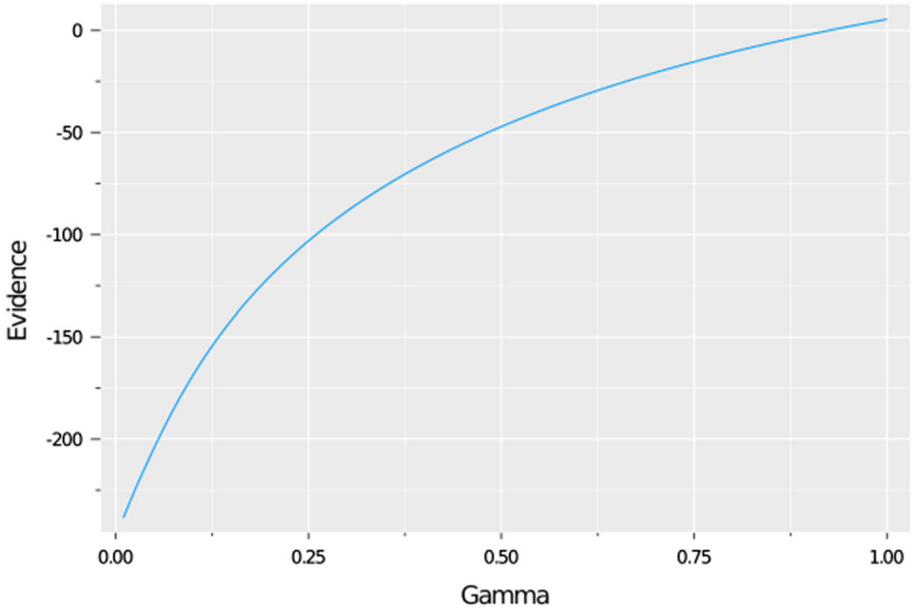


Fig. 1 Estimated $p_\gamma(y_{1:n})$. The Y-axis represents the value of $p_\gamma(y_{1:n})$, and the X-axis represents the value of γ

adoptes the H-score to define prequential score for state space models, and the criteria (2.3) can be seen as prequential score under *i.i.d.* settings. As shown in "Appendix B", under the assumptions stated in Shao et al. (2019), the leave-one-out H-score (2.3) can be rewritten as

$$\mathcal{H}_n(\gamma) := \sum_{i=1}^{n}\sum_{k=1}^{d_y}\left\{2\mathbb{E}_\gamma\left[\frac{\partial^2 \log \mathcal{L}_\gamma(y_i;\theta)}{\partial^2 y_{i_{(k)}}} + \left(\frac{\partial \log \mathcal{L}_\gamma(y_i;\theta)}{\partial y_{i_{(k)}}}\right)^2\right] - \left(\mathbb{E}_\gamma\left[\frac{\partial \log \mathcal{L}_\gamma(y_i;\theta)}{\partial y_{i_{(k)}}}\right]\right)^2\right\}, \tag{2.4}$$

where the expectation is with respect to the robustified posterior distribution (2.1). Then, we can estimate γ as follows

$$\hat{\gamma} := \arg\min_\gamma \mathcal{H}_n(\gamma). \tag{2.5}$$

As we shall discuss later, it can be shown that, under some conditions, $n^{-1}\mathcal{H}_n(\gamma)$ converges to the Fisher divergence, $\mathcal{J}(\gamma) := \int \left\|\nabla_y \log g(y_{1:n}) - \nabla_y \log p_\gamma(y_{1:n})\right\|^2 g(y_{1:n}) dy_{1:n}$ Therefore, $\mathcal{H}_n(\gamma)$ can be considered as an empirical approximation of the Fisher divergence $\mathcal{J}(\gamma)$ for the marginal distribution based on unnormaliseable models defined by robust divergence. An important point here is that the estimation by the H-score is independent of the normalisation constant. The following proposition is theoretical justification of selecting γ via (2.5).

Proposition 2 *Let* $\gamma^\star := \arg\min_\gamma \mathcal{J}(\gamma)$. *Then, under the conditions stated in "Appendix C", we have* $\hat{\gamma} \to \gamma^\star$ *w.p.1. as* $n \to \infty$.

Proof See "Appendix C". □

Remark 1 As we mentioned, the prequential version of $\mathcal{H}_n(\gamma)$ is also called the H-score in the context of Bayesian model selection (Shao et al. 2019; Dawid and Musio 2015). The main advantage of using the H-score in this context is that it will provide a consistent and coherent model selection criterion. Jewson et al. (2021) proposes a consistent model selection criterion that is similarly based on H-scores but with batch estimation. Although the prequential method is coherent, this comes with a very high computational cost, for every model, one must do posterior inference on all permutations of the data and increasing sample sizes. Here, we use a batch estimation approach to estimate γ, which avoids high computational costs. We also want to emphasise that, as we shall study later, such a batch approach will give rise to natural and efficient algorithms to estimate γ and posterior sampling.

Under the density power divergence (2.2), the first and second order derivatives of $\log \mathcal{L}_\gamma(y_i;\theta)$ are given by

$$\frac{\partial \log \mathcal{L}_\gamma(y_i;\theta)}{\partial y_i} = f_\theta(y_i)^{\gamma-1}\frac{\partial f_\theta(y_i)}{\partial y_i},$$
$$\frac{\partial^2 \log \mathcal{L}_\gamma(y_i;\theta)}{\partial y_i^2} = (\gamma-1) f_\theta(y_i)^{\gamma-2}\left(\frac{\partial f_\theta(y_i)}{\partial y_i}\right)^2 + f_\theta(y_i)^{\gamma-1}\frac{\partial^2 f_\theta(y_i)}{\partial y_i^2}.$$

These expressions do not include the integral term $\int f_\theta(x)^{1+\gamma}dx$, which makes the calculation of $\mathcal{H}_n(\gamma)$ much more straightforward in practice since the integral term often is a form of a complicated expression.

2.4 Numerical illustration of the H-score under normal distribution

We consider the same problem in the example in Sect. 2.2. For a normal distribution $\mathcal{N}(\mu, \sigma^2)$, the derivatives are as follows

$$\frac{\partial \log \mathcal{L}_\gamma(y_i;\theta)}{\partial y_i} = -\frac{\phi(y_i;\mu,\sigma^2)^\gamma (y_i-\mu)}{\sigma^2},$$
$$\frac{\partial^2 \log \mathcal{L}_\gamma(y_i;\theta)}{\partial y_i^2} = \frac{\phi(y_i;\mu,\sigma^2)^\gamma}{\sigma^4}\left\{\gamma(y_i-\mu)^2-\sigma^2\right\},$$

where $\phi(\cdot;\mu,\sigma^2)$ is the density function of $\mathcal{N}(\mu,\sigma^2)$. We calculated $\mathcal{H}_n(\gamma)$ in (2.4) for each γ, where posterior expectations were approximated by 2000 posterior samples of μ. The data were simulated in the same way as in Sect. 2.2. The results are shown in Fig. 2 when $\tau = 10$ (blue lines) and 30 (red lines). Our experiment shows numerically that $\mathcal{H}_n(\gamma)$ has a local minimum. Furthermore, it can be seen that in regions where γ is small, the posterior mean is relatively heavily influenced by outliers. In contrast, the posterior mean settles to a constant value in regions where γ is greater than the value that minimises $\mathcal{H}_n(\gamma)$. Uncertainty in the sense of CI becomes greater. This result would suggest that statistical inefficiencies occur in regions where γ is larger than is necessary.

3 Sequential Monte Carlo samplers

A natural way to obtain $\hat{\gamma}$ in (2.5) will be to use Robbins-Monro-type recursion.

$$\gamma_{t+1} = \gamma_t + \kappa_t \nabla_\gamma \mathcal{H}_t(\gamma_t), \tag{3.1}$$

where $\sum_t \kappa_t = \infty, \sum_t \kappa_t^2 < \infty$ but, in general, posterior sampling based on the Monte Carlo approximation will be required to evaluate $\nabla_\gamma \mathcal{H}_t(\gamma_t)$. That is, we need to construct an estimator of $\nabla_\gamma \mathcal{H}_t(\gamma_t)$. To do so, we first treat γ_t in (3.1) as the positive sequence such that $0 < \gamma_0 < \gamma_1 \cdots < \cdots \gamma_T$ where $0 \le t \le T$ is an artificial time index. Then (2.1) gives rise to the following tempering-like distributions on a common measurable space, say $(\Theta, \mathcal{B}(\Theta))$

$$\Pi_t(\theta \mid y_{1:n}) := \Pi_{\gamma_t}(\theta \mid y_{1:n}) \propto \mathcal{L}_{\gamma_t}(y_{1:n};\theta)\pi(\theta). \tag{3.2}$$

The meaning of tempering-like here is not the usual tempering. It means that a family of distributions is constructed in the same measurable space by a sequence of updated γ_t, gradually approaching the target distribution in the sense of an optimised γ, say $\gamma^\star$. Using this, we can define a sequence of distributions defined on product spaces $(\Theta^t, \mathcal{B}(\Theta^t)) := (\Theta^t := \prod_{i=1}^t \Theta)$

$$\tilde{\Pi}_t(\theta_{0:t} \mid y_{1:n}) := \Pi_t(\theta_t \mid y_{1:n}) \prod_{k=0}^{t-1} L_k(\theta_{k+1}, \theta_k), \tag{3.3}$$

where L_k is a transition kernel from Θ^{k+1} to Θ^k. Notice that $\tilde{\Pi}_t(\theta_{0:t} \mid y_{1:n})$ admits marginally $\Pi_t(\theta_t \mid y_{1:n})$. Also let $M_k(\theta_{k-1}, \theta_k)$ be a Π_k-reversible MCMC kernel. Then it is given by

$$L_k(\theta_{k+1}, \theta_k) = \frac{\Pi_k(\theta_{k-1} \mid y_{1:n}) M_k(\theta_{k-1}, \theta_k)}{\Pi_k(\theta_k \mid y_{1:n})}.$$

Since $L_{t-1} \otimes \Pi_t = \Pi_t \otimes M_t$ by construction, as the Radon-Nikodym derivative between them, one can derive unnormalised incremental weights as follows

$$\log w_t^j := \log\left(\frac{\mathcal{L}_{\gamma_t}(y_{1:n};\theta_{t-1}^j)}{\mathcal{L}_{\gamma_{t-1}}(y_{1:n};\theta_{t-1}^j)}\right) = \frac{-\frac{n}{1+\gamma_t}\int f_{\theta_{t-1}^j}(x)^{1+\gamma_t}dx + \frac{1}{\gamma_t}\sum_{i=1}^n f_{\theta_{t-1}^j}(y_i)^{\gamma_t}}{-\frac{n}{1+\gamma_{t-1}}\int f_{\theta_{t-1}^j}(x)^{1+\gamma_{t-1}}dx + \frac{1}{\gamma_{t-1}}\sum_{i=1}^n f_{\theta_{t-1}^j}(y_i)^{\gamma_{t-1}}}. \tag{3.4}$$

For a detailed discussion of the choice of L_k and M_k and how the weights are derived, see e.g. Del Moral et al. (2006), Dai et al. (2020). Then the SMC samplers (Del Moral et al. 2006) iterate the following steps. First, the normalised weights $W_t^j := (\sum_{k=1}^N w_t^k)^{-1} w_t^j$ are calculated for $j \in [1, N]$. Using these, one needs to sample ancestor indices $\left\{A_t^j\right\}_{j=1}^N$ from the categorical distribution induced by the normalised weights $\{W_t^j\}_{j=1}^N$, denoted by $\mathrm{C}\left(\{W_t^j\}_{j=1}^N\right)$. Finally, sample θ_t^j through $M_t(\theta_{t-1}^{A_t^j}, d\theta_t)$ for $j \in [1, N]$. We refer to Dai et al. (2020) for a number of recent advances in SMC samplers. As a result, we have the particle system $\{\theta_t^j, W_t^j\}_{j=1}^N$ that constructs an approximation $\sum_{j=1}^N W_t^j \delta_{\theta_t^j}(d\theta)$ of $\Pi_t(\theta \mid y_{1:n})$ at any time step t as required, where $\delta_z(dx)$ denotes the Dirac measure located at z. Using the system, an approximation $\nabla_\gamma \hat{\mathcal{H}}_t(\gamma_t)$ of $\nabla_\gamma \mathcal{H}_t(\gamma)$ can be obtained under appropriate regularity conditions that allow us to interchange differentiation with respect to γ and integration; see "Appendix D" for details. Therefore, an approximation of (3.1) will be

$$\gamma_{t+1} = \gamma_t + \kappa_t \nabla_\gamma \hat{\mathcal{H}}_t(\gamma_t). \tag{3.5}$$

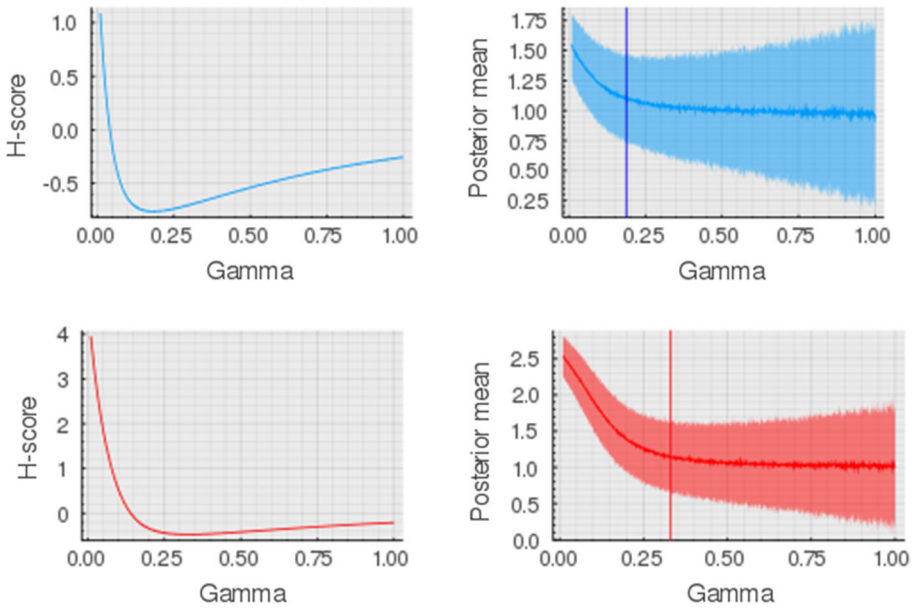


Fig. 2 The top left-hand plots the values of $\mathcal{H}_n(\gamma)$ on the Y-axis, and values of γ on the X-axis when $\tau = 10$. The minimum value of $\mathcal{H}_n(\gamma)$ was obtained when $\gamma = 0.1874$. The top right-hand figure plots the sample mean values of the posterior mean of μ on the Y-axis and the corresponding values of γ on the X-axis under the same setting. The vertical line represents the value of γ that minimises $\mathcal{H}_n(\gamma)$, and the thin ribbon line represents the 95% credible interval. The bottom left-hand and right-hand figures represent the same figures when $\tau = 30$ respectively. In this case, the minimum value of $\mathcal{H}_n(\gamma)$ was obtained when $\gamma = 0.3311$

Given the time steps $T > 0$ and initial $\gamma_0 > 0$, we update γ_t via (3.5) and iterate the SMC sampler T times. Our method can be algorithmically summarised as follows.

Algorithm 1

(i) Initialise the particles $\{\theta_0^j, W_0^j\}_{j=1}^N$, with $\theta_0^j \overset{i.i.d.}{\sim} \pi$, $\gamma_0 > 0$, $W_0^i = 1$ for $j \in [1, N]$.
(ii) Apply the iteration $\gamma_{t+1} = \gamma_t + \kappa_t \nabla_\gamma \hat{\mathcal{H}}_t(\gamma_t)$.
(iii) Calculate w_t^j in (3.4) and set $W_t^j = \frac{w_t^j}{\sum_{k=1}^N w_t^k}$ for $j \in [1, N]$.
(iv) Sample ancestor indices $\left\{A_t^j\right\}_{j=1}^N \sim \mathrm{C}\left(\{W_t^j\}_{j=1}^N\right)$.
(v) Sample particles $\theta_t^j \sim M_t(\theta_{t-1}^{A_t^j}, d\theta_t)$ for $j \in [1, N]$.
(vi) Obtain estimate of $\nabla_\gamma \mathcal{H}_{t+1}(\gamma_{t+1})$.

Remark 2 Since $\{w_t^j\}$ are independent of $\{\theta_t^i\}$ but dependent of $\{\theta_{t-1}^i\}$, the particles $\{\theta_t^i\}$ can be sampled after resampling in Algorithm 1. In addition, Algorithm 1 uses a simple multinomial resampling applied at each step. The variability of the Monte Carlo estimates can be further reduced by incorporating dynamic resampling via the use of effective sample size. See Del Moral et al. (2006); Dai et al. (2020) for details.

Theoretical guarantees of convergence of the Robbins-Monro algorithm usually require that $\nabla_\gamma \hat{\mathcal{H}}_t(\gamma_t)$ is unbiased. Even if $0 < \gamma_0 < \gamma_1 \cdots < \cdots \gamma_T$ is chosen adaptively, Beskos et al. (2016) shows that $\nabla_\gamma \hat{\mathcal{H}}_t(\gamma_t)$ is still a (weakly) consistent estimator, but not an unbiased estimator. Such unbiased estimation may be possible by using the recently developed MCMC with couplings in Algorithm 1, see Middleton et al. (2019), Jacob et al. (2020) for details. Instead of discussing convergence through such unbiased estimation, we shall discuss convergence through numerical experiments in the following sections.

We end this section by noting several advantages of the proposed method. First, Algorithm 1 enables us to estimate the tuning parameter and obtain posterior sampling simultaneously. This is a notable difference from existing methods, such as running MCMC or the EM algorithm with a fixed tuning parameter, for example, Fujisawa and Eguchi (2008); Ghosh and Basu (2016). We believe that it may be emphasised that by setting up a well-defined target function and using the stochastic gradient framework-based SMC samplers, it is possible to avoid the two-stage estimation that many previous studies have done in this context. We also emphasise that our proposed method has two notable advantages over existing methods: it does not require pilot plots, and it does not require an expression of the asymptotic variance of the model. Next, recall that as $\gamma \downarrow 0$, $\mathcal{L}_\gamma(y_{1:n}; \theta)$ converges to Kullback-Leibler divergence. Let $\gamma^\star$ be the value of converged γ in Algorithm 1. Then Algorithm 1 may be producing an approximated bridge between (multiplied by a prior distribution) Kullback-Leibler divergence and the target distribution $\mathcal{L}_{\gamma^\star}(y_{1:n}; \theta)$. Therefore, sampling from such tempering-like distributions induced by the density power divergence (3.2) could provide a beneficial tempering effect and a potential reduction in computational complexity, particularly when d is large (Neal 2001). Finally, Algorithm 1 will give rise to a natural way to construct adaptive MCMC kernels. Suppose that we use Metropolis-Hastings kernels based on a Gaussian random walk. Notice that before MCMC step in Algorithm 1, we have $\{\theta_{t-1}^j, W_{t-1}^j\}_{j=1}^N$ which approximates $\Pi_{t-1}(\theta \mid y_{1:n})$ so that estimates $\hat{\mu}_{t-1} := \sum_{j=1}^N W_{t-1}^j \theta_{t-1}^j$, $\hat{\Sigma}_{t-1} := \sum_{j=1}^N W_{t-1}^j (\theta_{t-1}^j - \hat{\mu}_{t-1})(\theta_{t-1}^j - \hat{\mu}_{t-1})^\top$ are available. Also Ghosh and Basu (2016, The-

orem 1) shows that $\Pi_{t-1}(\theta \mid y_{1:n})$ can be approximated by the Gaussian distribution. These will lead us to set $M_t(\theta_{t-1}^{A_t^j}, d\theta_t) = \theta_{t-1}^{A_t^j} + \xi_t$, $\xi_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 2.38 d^{-1/2} \hat{\Sigma}_{t-1})$, for instance. We note that this proposal is also can be considered as a consequence of the results from the optimal scaling analysis for random walk Metoropolis, see (Chopin 2002; Chopin and Papaspiliopoulos 2020, Chapter 17) and references therein for more details.

4 Numerical examples

To specify the schedule for the scaling parameters $\{\kappa_t\}$ in (3.5), we use the standard adaptive method termed ADAM, by Kingma and Ba (2014), known as stabilising the unnecessary numerical instability due to the choice of $\{\kappa_t\}$. Assume that after t steps we have $c_t := \nabla \hat{\mathcal{H}}_t(\gamma_t)$. ADAM applies the following iterative procedure,

$$
\begin{aligned}
&m_t = m_{t-1}\beta_1 + (1-\beta_1)c_t, \quad v_t = v_{t-1}\beta_2 + (1-\beta_2)c_t^2, \\
&\hat{m}_t = m_t/(1-\beta_1^t), \quad \hat{v}_t = v_t/(1-\beta_2^t), \\
&\gamma_{t+1} = \gamma_t - \alpha \hat{m}_t/(\sqrt{\hat{v}_t} + \epsilon),
\end{aligned}
$$

where $(\beta_1, \beta_2, \alpha, \epsilon)$ are the tuning parameters. The convergence properties of ADAM have been widely studied (Kingma and Ba 2014; Reddi et al. 2019). Following closely Kingma and Ba (2014), in all uses of ADAM below we set $(\beta_1, \beta_2, \alpha, \epsilon) = (0.9, 0.999, 0.003, 10^{-8})$. ADAM is nowadays a standard and very effective addition to the type of recursive inference algorithms we are considering here, even more so as for increasing dimension of unknown parameters. See the above references for more motivation and details.

4.1 Simulation studies

We here demonstrate the numerical performance of the proposed method. Throughout this study, we consider Gaussian models $\{f_\theta : \theta \in \Theta\} = \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown parameters. We first simulated data $\{y_i\}_{i=1}^{100} \overset{i.i.d.}{\sim} \mathcal{N}(1, 1)$ and then randomly replaced $\tau\%$ of $\{y_i\}_{i=1}^{100}$ by $y_i + 5$. This setting is commonly referred to as the M-open world (Bernardo and Smith 2009) in the sense that there is no $\theta^\star$ such that $g = f_{\theta^\star}$.

Experiment 1: convergence property

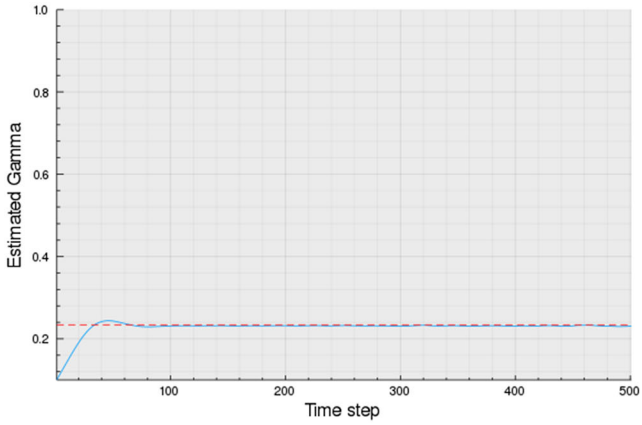
We investigate convergence behaviour of Algorithm 1. In this study, we set $(N, T, \gamma_0) = (2000, 500, 0.1)$ with 50 MCMC steps in Algorithm 1 to estimate γ. The results are shown in Fig. 3. As can be seen in the figure, our proposed method converges stably to the true value after about 100 iterations. Here, the true value was obtained by first approximating the H-score with MCMC in the same way as before and then using a grid search to find the γ that minimises it.

Experiment 2: comparison with methods using fixed values of γ

We next compare the performance of our proposed method with that of a non-adaptive method using a fixed value of γ. We set $\tau \in \{0, 10, 20, 30\}$, and for each case we computed the posterior distribution of μ using Algorithm 1 and the vanilla version of MCMC. For Algorithm 1, the tuning parameters were set to $(N, T, \gamma_0) = (2000, 300, 0.1)$ with 50 MCMC steps, and for vanilla MCMC, γ was set to $\gamma \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ in advance of the estimation with 100,000 MCMC steps. We used Metropolis-Hastings kernels based on a Gaussian random walk with $\mathcal{N}(0, 0.4)$ for the two cases under the uniform prior. Using the posterior samples obtained from Algorithm 1 and non-adaptive methods, we computed the posterior mean and 95% credible interval of μ. We ran 100 Monte Carlo experiments to calculate their (empirical) mean square error (MSE) and average 95% credible interval (ACI). The MSE is computed against the target value of 1, and the value is multiplied by 100. The results are given in Table1. Although it is a simple example, the results summarised in the table clearly show that the accuracy of the inference is improved by estimating γ from the data rather than simply fixing it in terms of MSE. In fact, the best γ among the five choices depends on the underlying contamination ratio that we do not know in practice. Hence, it is difficult to determine a suitable value of γ simply by looking at the data, while our method can automatically tune the value of γ from the data. It should also be noted that the importance of adaptive tuning of γ is reflected in the results of not only MSE but also ACI; that is, the interval length obtained from Algorithm 1 is narrow compared with the non-adaptive methods in all the four scenarios.

Experiment 3: comparison with Jewson et al. (2021)

We next compare the proposed method with the H-posterior proposed by Jewson et al. (2021) (denoted by JR hereafter), where the posterior of the model parameters (μ, σ^2) as well as γ can be obtained. To apply the JR method, we generated 1000 posterior samples after discarding the first 500 samples. We evaluate the performance of the inference of μ and σ by MSE (multiplied by 100), coverage probability (CP) and average length (AL) of 95% credible intervals. The results are shown in Table 2, where the average estimates of γ are also shown. Although both methods provide similar estimates of γ, the accuracy of point estimation of JR is slightly better than that of Algorithm 1. However, it is observed that JR tends to produce a short coverage length so that the CP



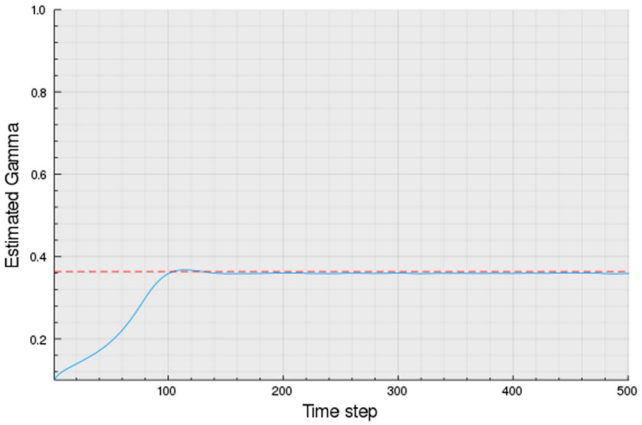


Fig. 3 Gaussian models experiment: Trajectories from execution of Algorithm 1. We used $N = 2000$ particles with 50 MCMC iterations and initial value $\gamma_0 = 0.1$. The left panel shows results when $\tau = 5$ and the right one shows when $\tau = 10$. The horizontal dashed lines in the plots show the true parameter $\gamma^\star = 0.2339$ for the left and $\gamma^\star = 0.3638$ for the right. The blue lines show the trajectory of $\hat{\gamma}$ estimated by Algorithm 1

Table 1 Empirical mean squared errors (MSE) and average 95% credible intervals (ACI) of Algorithm 1 and the non-adaptive method (the vanilla version of MCMC with fixed γ), based on 100 Monte Carlo experiments

	$\tau = 0$	$\tau = 10$	$\tau = 20$	$\tau = 30$
($\gamma = 0.1$)				
MSE	**3.66**	12.09	7.58	21.54
ACI	(0.62, 1.06)	(1.08, 1.57)	(1.00, 1.49)	(1.18, 1.71)
($\gamma = 0.3$)				
MSE	4.87	**4.08**	**2.27**	**2.71**
ACI	(0.56, 1.10)	(0.82, 1.44)	(0.69, 1.28)	(0.65, 1.27)
($\gamma = 0.5$)				
MSE	6.27	4.79	3.32	4.86
ACI	(0.49, 1.13)	(0.69, 1.45)	(0.61, 1.31)	(0.52, 1.24)
($\gamma = 0.7$)				
MSE	8.00	5.78	4.61	6.78
ACI	(0.41, 1.18)	(0.58, 1.51)	(0.55, 1.38)	(0.43, 1.29)
($\gamma = 0.9$)				
MSE	10.04	8.35	6.44	8.90
ACI	(0.33, 1.24)	(0.46, 1.59)	(0.47, 1.46)	(0.34, 1.35)
$\hat{\gamma}$	0.006	0.207	0.213	0.272
MSE	3.13	3.51	2.13	2.47
ACI	(0.66, 1.05)	(0.91, 1.46)	(0.76, 1.32)	(0.67, 1.28)

The best MSE value among different choices of γ is highlighted in bold. The bottom row shows estimated γ and the corresponding MSE and CI when estimated with our proposed method. The tuning parameters were set to $(N, T, \gamma_0) = (2000, 300, 0.1)$ with 50 MCMC steps

of the JR method is much smaller than the nominal level 95%. Accordingly, the average length is much smaller than those by Algorithm 1. This means that a direct application of the H-posterior by Jewson et al. (2021) may fail to capture the uncertainty of the posterior compared with the proposed method.

Experiment 4: comparison with tempering

Following Nakagawa and Hashimoto (2020), we compare the robustness to outliers for the two generalised posterior distributions. The first distribution is constructed in the same way as before, while the other is constructed using tempering. We specify a tempered posterior as $\Pi_{\phi_t}(\theta \mid y_{1:n}) \propto \mathcal{L}(y_{1:n}; \theta)^{\phi_t} \pi(\theta)$ where $0 = \phi_0 < \phi_1 \cdots < \cdots \phi_T = 1$. To construct the sequence, we divided the interval [0, 1] into 500 equal parts. We applied Algorithm 1 and the SMC sampler with the tempered posterior to test the robustness of the proposed method to data sets containing outliers. We used $N = 2000$ particles, 50 MCMC steps for both methods, and set $(T, \gamma_0) = (500, 0.1)$ for Algorithm 1. The prior and MCMC kernels were set as the previous experiment, and the density estimation results obtained from the estimation results are summarised in Fig. 4. The red line represents the posterior density estimate for μ when the data do not contain any outliers, and the blue line represents it when the data contain outliers. It is clear from the estimation results that our proposed method is robust even when the dataset contains outliers, while the SMC sampler with the tempered posterior is greatly affected by outliers, and the estimated posterior distributions are completely separated as a result.

4.2 Applications to real data

Newcomb data

We apply our methodology to Simon Newcomb's measurements of the speed of light data, motivated by applications in Stigler (1977), Basu et al. (1998), Basak et al. (2021). The data can be obtained from Andrew Gelman's webpage: http://www.stat.columbia.edu/~gelman/book/data/light.asc. The sam-

Table 2 Mean squared errors (MSE), coverage probability (CP) and average length (AL) of 95% credible intervals of μ and σ based on Algorithm 1 and the JR method

	τ	Mean($\hat{\gamma}$)	Var($\hat{\gamma}$)	MSE		CP		AL	
				μ	σ	μ	σ	μ	σ
Algorithm 1	5	0.194	0.001	3.62	5.23	97	84	0.57	0.53
	10	0.297	0.003	4.57	9.20	97	88	0.51	0.56
	15	0.377	0.009	1.06	12.60	92	82	0.61	0.67
JR	5	0.193	0.011	2.33	5.10	68	13	0.27	0.16
	10	0.230	0.009	2.46	6.25	64	10	0.28	0.16
	15	0.261	0.013	2.50	7.27	68	5	0.29	0.15

MSE is multiplied by 100. The tuning parameters in our algorithm were set to $(N, T, \gamma_0) = (2000, 300, 0.1)$ with 5 MCMC steps

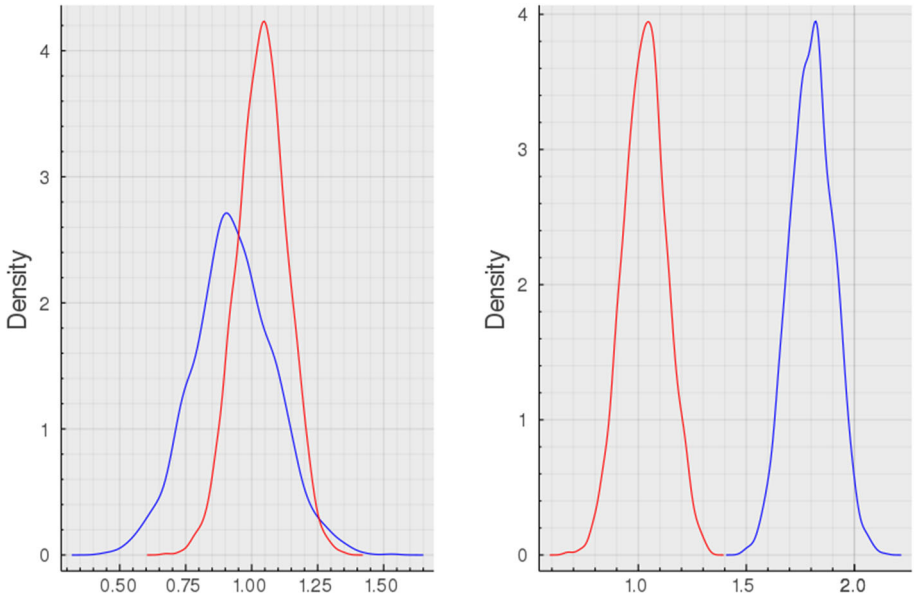


Fig. 4 The density estimation of μ estimated by Algorithm 1 (left) and SMC sampler with the tempered posterior (right). The blue line shows when $\tau = 20$ (contains 20% outliers) and the red one shows when $\tau = 0$ (contains 0% outliers) in both panels

ple size of the data set is 66 and contains two outliers, -44 and -2, illustrated in Fig. 5. We fitted a Gaussian distribution model $\{f_\theta : \theta \in \Theta\} = \mathcal{N}(\mu, \sigma^2)$ to the data and used Algorithm 1 to obtain the posterior distribution of the parameters (μ, σ). The tuning parameters (N, T, γ_0) in Algorithm 1 were set to (2000, 300, 0.1) with 50 MCMC iterations. The MCMC kernel was constructed as in the previous examples, and results are given in Fig. 6. The existing study (Basak et al. 2021) reported $\hat{\gamma} = 0.23$ for the same data set, which is very high compared to our estimate result of $\hat{\gamma} = 0.0855$. Since the method proposed in Basak et al. (2021) requires a pilot plot and the estimation results depend significantly on it, we believe our estimation results are more reasonable. In fact, it is unlikely that we will have to use a value of $\gamma = 0.23$ for a data set that contains only two outliers. As shown in Basu et al. (1998), the parameter estimates are almost the same when $\gamma = 0.0855$ and when $\gamma = 0.23$. However, from the point of view of statistical efficiency, it would be preferable to adopt the lower value of $\gamma = 0.0855$ if the estimates were the same. To confirm this, 100 bootstrap resamplings were performed on the data, and the posterior bootstrap mean of each parameter and the variance was calculated, reported in 3. For each re-sampled data, we compared the results when the posterior distribution was calculated while estimating γ with our method and when the posterior distribution was calculated using MCMC after estimating and fixing it with the method proposed in Basak et al. (2021). The numerical experiments show that although the means estimated parameters agree between the two methods, the variances are much smaller for our method, suggesting that overestimation of γ leads to statistical inefficiency.

Table 3 Mean and variance of the posterior means of the parameters from 100 bootstrap re-samplings

	Mean($\hat{\mu}$)	Var($\hat{\mu}$)	Mean($\hat{\sigma}$)	Var($\hat{\sigma}$)
Algorithm 1	27.559	0.0017	5.7605	0.0013
Basak et al. (2021)	27.674	0.4351	5.3976	0.5295

The first row shows the result when using the proposed method, and the second one shows when using the method studied in Basak et al. (2021)

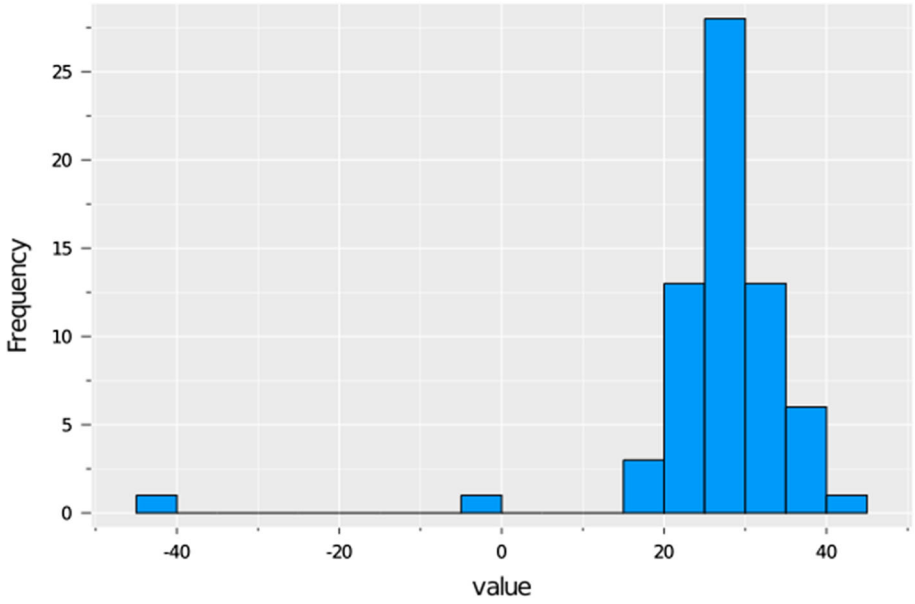


Fig. 5 The histogram of Simon Newcomb's measurements of the speed of light data

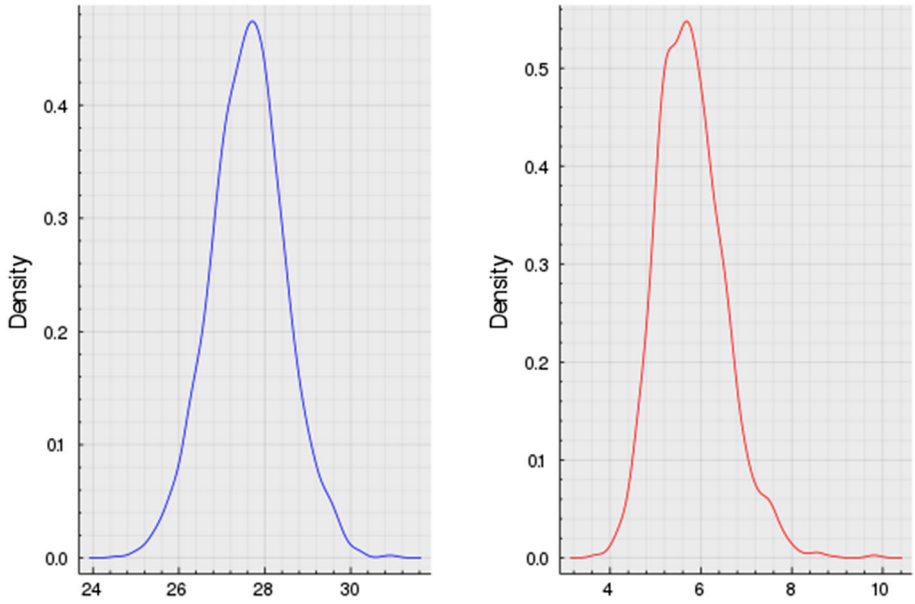


Fig. 6 The density estimation of μ (left) and σ (right) for Simon Newcomb's measurements of the speed of light data. The tuning parameters (N, T, γ_0) were set to $(2000, 300, 0.1)$ with 50 MCMC iterations. The mean value of the estimated μ is 27.6082 and σ is 5.7829 with $\hat{\gamma} = 0.0855$

Hertzsprung–Russell star cluster data

We use our methodology to perform linear regression models with normal errors, that is $y_i \sim \mathcal{N}(x_i^\top \beta, \sigma^2)$. Motivated by Basak et al. (2021), we fitted the regression model to the Hertzsprung-Russell star cluster data (Rousseeuw and Leroy 2005), without constants. The data set contains 47 observations on the logarithm of the effective temperature at the surface of the CYG OB1 star cluster (Te, covariates $\{x_i\}$) and the logarithm of its light intensity (L/L0, explained variables $\{y_i\}$). The data can be obtained from https://rdrr.io/cran/robustbase/man/starsCYG.html, and shown at Fig. 7. The tuning parameters (N, T, γ_0) in Algorithm 1 were set to $(2000, 300, 0.1)$ with 50 MCMC iterations, and we used the uniform prior for (β, σ). The MCMC kernel was constructed as in the previous examples, and results are given in Fig. 8. The corresponding OLS estimates were $(\hat{\beta}, \hat{\sigma}) = (1.1559, 0.7219)$. Whilst we obtained $\hat{\gamma} = 0.1165$, Basak et al. (2021) reported $\hat{\gamma} = 0.76$ for the same data set. Our numerical experiments and previous studies (Ghosh and Basu 2016; Nakagawa and Hashimoto 2020) will suggest that as the proportion of outliers in the data increases, the value of γ also tends to increase. Thus, such a large value of γ is not reasonable considering the proportion of outliers in the data (only four samples in the lower right part in Fig. 7), suggesting the superiority of our proposed method. Indeed, to confirm the suggested statistical inefficiency, the same experiments as in the previous section were carried out, and the

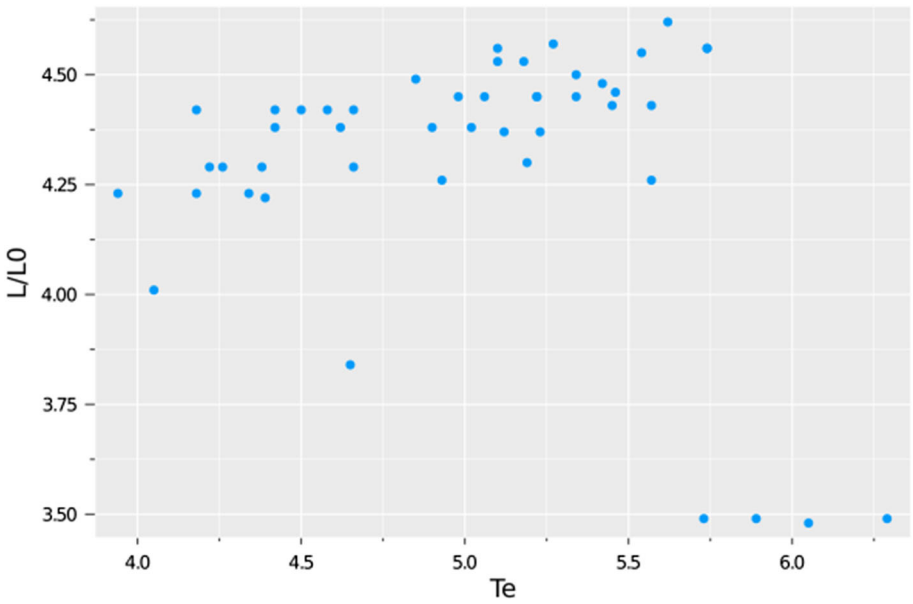


Fig. 7 The scatter plot of Hertzsprung-Russell star cluster data

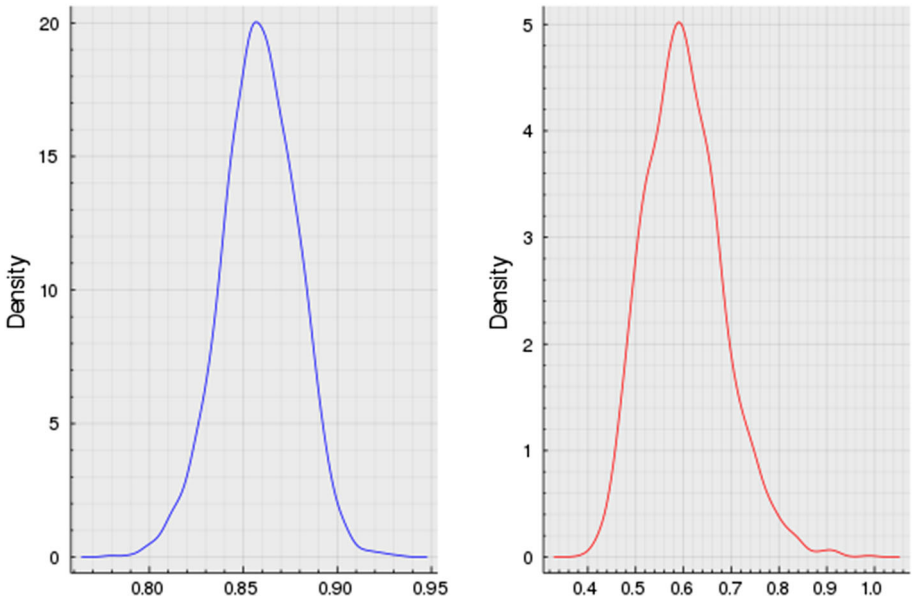


Fig. 8 The density estimation of β (left) and σ (right) for Hertzsprung-Russell star cluster data. The tuning parameters (N, T, γ_0) were set to $(2000, 300, 0.1)$ with 50 MCMC iterations. The mean value of the estimated β is 0.8586 and σ is 0.602 with $\hat{\gamma} = 0.1165$. The corresponding OLS estimates are $(\hat{\beta}, \hat{\sigma}) = (1.1559, 0.7219)$

Table 4 Mean and variance of the posterior means of the parameters from 100 bootstrap re-samplings

	Mean($\hat{\mu}$)	Var($\hat{\mu}$)	Mean($\hat{\sigma}$)	Var($\hat{\sigma}$)
Algorithm 1	0.8514	0.0002	0.6100	0.0037
Basak et al. (2021)	0.8763	0.0299	0.6536	0.0708

The first row shows the result when using the proposed method, and the second one shows when using the method studied in Basak et al. (2021)

results are summarised in Table 4. Although the results are not as striking as in the previous Newcomb data example, it would be possible to confirm that statistical inefficiencies occur in the estimation by the proposed method in Basak et al. (2021) in the regression model as well.

5 Concluding remarks

Our proposed method performs reasonably well and provides one of the few options, as far as we know, for routine robust Bayesian inference. To the best of our knowledge, this is the first attempt to propose both a theory and a computational algorithm to estimate the tuning parameters from data and to allow robust Bayesian estimation. We have shown numerically that a more efficient Bayesian estimation can be achieved by estimating the tuning parameter γ from the data. We have proposed an efficient sampling method using SMC samplers considering the sequence of γ as the temperature. Compared to existing studies (Warwick and Jones 2005; Basak et al. 2021), our method has specificity and use-

fulness in that we can estimate the tuning parameters and sample from the posterior distribution simultaneously, and pilot plots and the asymptotic variance formula are not necessary. In this paper, we have focused in particular on the case of the density power divergence, but we want to stress that our method is general enough in the sense that it can be applied to the Bayesian estimation of other robust divergence-induced models.

Furthermore, our framework opens up a number of routes for future research and insight, including those described below.

1. As we have noted, the integral term $\int f_\theta(x)^{1+\gamma}dx$ is eliminated in the H-score, while the computation of the posterior distribution is computationally expensive, so it may be better to consider the H-posterior studied in Jewson et al. (2021) in this respect. However, the robustness of the H-posterior has not yet been studied, and it would therefore be interesting in the future to investigate this point in more detail using the influence function.
2. Another direction of investigation involves the construction of an efficient MCMC kernel for posterior distributions derived from robust divergence such as (2.1). To make good inferences from data containing outliers, the posterior distribution induced by robust divergence is a model with a more or less heavy-tailed. As studied in Kamatani (2018), many standard MCMC algorithms are known to perform poorly in such cases, especially in higher dimensions. Therefore, studying MCMC algorithms within Algorithm 1 tailored to the posterior distribution induced by robust divergence would allow for more efficient Bayesian robust estimation.
3. In this study, we have not focused on time series data, but, as Shao et al. (2019) shows, the H-score can also be defined for models that are not independent, for example, state-space models. In fact, Boustati et al. (2020) proposes a method for Bayesian filtering of state-space models using robust divergence, but the tuning parameters need to be estimated before filtering, and in this sense, it is not online filtering. Algorithm 1 does batch estimation, but we believe that extending it to online estimation would allow robust filtering of the state-space model while estimating the tuning parameters online from the data.
4. This study has focused on estimation and computational methods proposed in the generalised Bayesian framework, particularly using robust divergence. On the other hand, methods using the Maximum Mean Discrepancy (Chérief-Abdellatif et al. 2020) and Kernel Stein Discrepancy (Matsubara et al. 2022) have also been proposed in recent years in the same generalised Bayesian framework, although not with the motivation of dealing with outliers. Both require adjustment of the hyperparameters of the kernel used, and it may be possible to estimate them using our proposed method and compare their performance. We have avoided comparing these potential alternative approaches because we believe this would obscure the main messages we have tried to convey within the numerical results section. Such a detailed numerical study can be the subject of future work.

Acknowledgements SY was supported by the Japan Society for the Promotion of Science (KAKENHI) under Grant Number 21K17713. SS was supported by the Japan Society for the Promotion of Science (KAKENHI) under Grant Number 21H00699.

Funding Open access funding provided by The University of Tokyo.

A Proof of Proposition 1

Recall that $\log \mathcal{L}_\gamma^{\mathbf{R}}(y_i;\theta) = \frac{1}{\gamma} f_\theta(y_i)^\gamma - \frac{1}{1+\gamma}\int f_\theta(x)^{1+\gamma}dx - \frac{1}{\gamma} + 1$. First, $\frac{1}{\gamma} f_\theta(y_i)^\gamma - \frac{1}{\gamma}$ is an increasing function of γ for arbitrary $f_\theta(y_i)$ and $\gamma > 0$. Furthermore, since we have assumed that $f_\theta(y_i) \leq 1$, it holds that $f_\theta(x)^{1+\gamma_1} \geq f_\theta(x)^{1+\gamma_2}$ for $\gamma_1 < \gamma_2$ and arbitrary x, so that $\frac{1}{1+\gamma}\int f_\theta(x)^{1+\gamma}dx$ is a decreasing function of γ. Hence, $\log \mathcal{L}_\gamma^{\mathbf{R}}(y_i;\theta)$ is increasing since it is a sum of two increasing functions.

B Derivation of (2.4)

For simplicity, we consider $d_y = 1$, but the extension to the general dimension is straightforward. Let $p(y|\theta)$ be a general model with parameter θ and $p(y) = \int p(y|\theta)\Pi(\theta)d\theta$ be the marginal likelihood under prior $\Pi(\theta)$. Under the assumptions stated in supplement (S6) in Shao et al. (2019), the following identify always holds

$$\sum_{k=1}^{d_y}\left\{2\frac{\partial^2 \log p(y)}{\partial y_{(k)}^2}+\left(\frac{\partial \log p(y)}{\partial y_{(k)}}\right)^2\right\}$$
$$=\sum_{k=1}^{d_y}\left\{\mathbb{E}\left[2\frac{\partial^2 \log p(y\mid\theta)}{\partial y_{(k)}^2}+\left(\frac{\partial \log p(y\mid\theta)}{\partial y_{(k)}}\right)^2\mid y\right]\right.$$
$$\left.-\left(\mathbb{E}\left[\frac{\partial \log p(y\mid\theta)}{\partial y_{(k)}}\mid y\right]\right)^2\right\},$$

where the expectation is taken with respect to the posterior distribution of θ given y. Using the above identity with $p(y)=p(y_i|y_{-i})$, $p(y|\theta)=p(y_i|\theta,y_{-i})$ and $\Pi(\theta)=\Pi(\theta|y_{-i})$, where $y_{-i}=(y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_n)$, we have

$$\sum_{k=1}^{d_y}\left\{2\frac{\partial^2 \log p(y_i\mid y_{-i})}{\partial y_{i(k)}^2}+\left(\frac{\partial \log p(y_i\mid y_{-i})}{\partial y_{i(k)}}\right)^2\right\}$$
$$=\sum_{k=1}^{d_y}\left\{\mathbb{E}\left[2\frac{\partial^2 \log p(y_i\mid y_{i-1},\theta)}{\partial y_{i(k)}^2}\right.\right.$$
$$\left.+\left(\frac{\partial \log p(y_i\mid y_{i-1},\theta)}{\partial y_{i(k)}}\right)^2\mid y_i,y_{i-1}\right]$$
$$\left.-\left(\mathbb{E}\left[\frac{\partial \log p(y_i\mid y_{i-1},\theta)}{\partial y_{i(k)}}\mid y_i,y_{i-1}\right]\right)^2\right\}.$$

Notice that, since we have assumed $i.i.d.$ observations, we have $p(y_i\mid y_{i-1},\theta)=p(y_i\mid\theta)$. Hence, the expression (2.4) follows by setting $p(y_i\mid\theta)=\mathcal{L}_\gamma(y_i;\theta)$.

C Proof of Proposition 2

The proof here is essentially the same as Shao et al. (2019) and Jewson et al. (2021), so we only provide an overview of the proof. Assume that, for simplicity, $d_y=1$. First one can show that $\mathcal{H}_n(\gamma)$ can be decomposed into the sum of conditional expectation terms of $\mathcal{H}(y_i,p_\gamma(y_i\mid\theta,y_{-i}))=\mathcal{H}(y_i,p_\gamma(y_i\mid\theta))$, and the sum of conditional variance terms of $\frac{\partial \log p_\gamma(y_i|\theta)}{\partial y_i}$, see Dawid and Musio (2015), Shao et al. (2019). Then under the assumptions stated in supplement of Shao et al. (2019), the variance term will converge at 0 w.p.1. Let $(\mathbf{B},\|\cdot\|)$ be the space of continuous real functions on the compact set of γ equipped with the sup-norm. Then, under the same assumptions, $\frac{1}{n}\sum_{i=1}^n\mathbb{E}_\gamma[\mathcal{H}(y_i,p_\gamma(y_i\mid\theta))]$ may take values in this space. Then strong law of large numbers on a separable Banach space (Azlarov and Volodin 1982; Beskos et al. 2009), may be applied to $\frac{1}{n}\sum_{i=1}^n\mathbb{E}_\gamma[\mathcal{H}(y_i,p_\gamma(y_i|\theta))]-\mathbb{E}_g[\mathcal{H}(y_1,p_\gamma(y_1\mid\theta))]$. Combining this result with (s10) in supplement of Shao et al. (2019) and integration by parts (Hyvärinen 2005; Dawid and Musio 2015) would yield $\limsup_\gamma\frac{1}{n}\mathcal{H}_n(\gamma)=\mathcal{J}(\gamma)$ w.p.1. The result follows under the assumption for identification such that $\mathcal{J}(\gamma)$ is only maximised at $\gamma^\star$, see A1. in Jewson et al. (2021) for instance.

D Derivatives of the H-score

In the following argument, we assume the exchangeability of integral and derivative without any remarks. For simplicity, we consider a univariate case, namely $d_y=1$. We define $\mathcal{D}_\gamma(y_{1:n};\theta):=\log\mathcal{L}_\gamma(y_{1:n};\theta)$ and $\mathcal{D}_\gamma(y_i;\theta):=\log\mathcal{L}_\gamma(y_i;\theta)$. The derivative of the H-score with respect to γ is expressed as

$$\frac{d}{d\gamma}\mathcal{H}_n(\gamma)$$
$$=2\sum_{i=1}^n\int\frac{d}{d\gamma}\left[\left\{\frac{\partial^2\mathcal{D}_\gamma(y_i;\theta)}{\partial^2 y_i}\right.\right.$$
$$\left.\left.+\left(\frac{\partial\mathcal{D}_\gamma(y_i;\theta)}{\partial y_i}\right)^2\right\}\Pi_\gamma(\theta\mid y_{1:n})\right]d\theta$$
$$-2\sum_{i=1}^n\mathbb{E}\left[\frac{\partial\mathcal{D}_\gamma(y_i;\theta)}{\partial y_i}\mid y_{1:n}\right]$$
$$\times\int\frac{d}{d\gamma}\left\{\frac{\partial\mathcal{D}_\gamma(y_i;\theta)}{\partial y_i}\Pi_\gamma(\theta\mid y_{1:n})\right\}d\theta,$$

which requires the computation of integral of the following form:

$$\int\frac{d}{d\gamma}\left\{C_\gamma^{(k)}(y_i;\theta)\Pi_\gamma(\theta\mid y_{1:n})\right\}d\theta$$
$$=\int\frac{d}{d\gamma}\left\{C_\gamma^{(k)}(y_i;\theta)\frac{e^{\mathcal{D}_\gamma(y_{1:n};\theta)}\pi(\theta)}{\int e^{\mathcal{D}_\gamma(y_{1:n};\theta)}\pi(\theta)d\theta}\right\}d\theta,\quad(D.1)$$

where

$$C_\gamma^{(1)}(y_i;\theta)=\frac{\partial^2\mathcal{D}_\gamma(y_i;\theta)}{\partial^2 y_i}+\left(\frac{\partial\mathcal{D}_\gamma(y_i;\theta)}{\partial y_i}\right)^2,$$
$$C_\gamma^{(2)}(y_i;\theta)=\frac{\partial\mathcal{D}_\gamma(y_i;\theta)}{\partial y_i},\quad(D.2)$$

and $\mathcal{D}_\gamma(y_{1:n};\theta) := \log \mathcal{L}_\gamma(y_{1:n};\theta)$. It follows that

$$\begin{aligned}
&\int \frac{d}{d\gamma}\left\{C_\gamma^{(k)}(y_i;\theta)\frac{e^{\mathcal{D}_\gamma(y_{1:n};\theta)}\pi(\theta)}{\int e^{\mathcal{D}_\gamma(y_{1:n};\theta)}\pi(\theta)d\theta}\right\}d\theta \\
&= \int \left\{\frac{d}{d\gamma}C_\gamma^{(k)}(y_i;\theta)\right\}\Pi_\gamma(\theta \mid y_{1:n})d\theta \\
&\quad + \int C_\gamma^{(k)}(y_i;\theta)\left\{\frac{d}{d\gamma}\mathcal{D}_\gamma(y_{1:n};\theta)\right\}\Pi_\gamma(\theta \mid y_{1:n})d\theta \\
&\quad - \int C_\gamma^{(k)}(y_i;\theta)\frac{e^{\mathcal{D}_\gamma(y_{1:n};\theta)}\pi(\theta)}{\left\{\int e^{\mathcal{D}_\gamma(y_{1:n};\theta)}\pi(\theta)d\theta\right\}^2} \\
&\quad \times \int \pi(\theta)e^{\mathcal{D}_\gamma(y_{1:n};\theta)}\left\{\frac{d}{d\gamma}\mathcal{D}_\gamma(y_{1:n};\theta)\right\}d\theta,
\end{aligned}$$

where the third term is further simplified to

$$\begin{aligned}
&\int C_\gamma^{(k)}(y_i;\theta)\Pi_\gamma(\theta \mid y_{1:n})d\theta \\
&\quad \times \int \left\{\frac{d}{d\gamma}\mathcal{D}_\gamma(y_{1:n};\theta)\right\}\Pi_\gamma(\theta \mid y_{1:n})d\theta.
\end{aligned}$$

Hence, the derivative (D.1) is expressed as

$$\begin{aligned}
&\mathbb{E}\left[\left\{\frac{d}{d\gamma}C_\gamma^{(k)}(y_i;\theta)\right\} + C_\gamma^{(k)}(y_i;\theta)\left\{\frac{d}{d\gamma}\mathcal{D}_\gamma(y_{1:n};\theta)\right\} \mid y_{1:n}\right] \\
&\quad - \mathbb{E}\left[C_\gamma^{(k)}(y_i;\theta) \mid y_{1:n}\right] \times \mathbb{E}\left[\frac{d}{d\gamma}\mathcal{D}_\gamma(y_{1:n};\theta) \mid y_{1:n}\right].
\end{aligned} \tag{D.3}$$

Finally, the derivative of the H-score is expressed as

$$\begin{aligned}
\frac{d}{d\gamma}\mathcal{H}_n(\gamma) = 2\sum_{i=1}^n &\mathbb{E}\left[\left\{\frac{d}{d\gamma}C_\gamma^{(1)}(y_i;\theta)\right\}\right. \\
&\left. + C_\gamma^{(1)}(y_i;\theta)\left\{\frac{d}{d\gamma}\mathcal{D}_\gamma(y_{1:n};\theta)\right\} \mid y_{1:n}\right] \\
&- 2\sum_{i=1}^n \mathbb{E}\left[C_\gamma^{(1)}(y_i;\theta) \mid y_{1:n}\right] \\
&\mathbb{E}\left[\frac{d}{d\gamma}\mathcal{D}_\gamma(y_{1:n};\theta) \mid y_{1:n}\right] \\
&- 2\sum_{i=1}^n \mathbb{E}\left[\frac{d}{dy_i}\mathcal{D}_\gamma(y_i;\theta) \mid y_{1:n}\right] \\
&\mathbb{E}\left[\left\{\frac{d}{d\gamma}C_\gamma^{(2)}(y_i;\theta)\right\}\right. \\
&\left. + C_\gamma^{(2)}(y_i;\theta)\left\{\frac{d}{d\gamma}\mathcal{D}_\gamma(y_{1:n};\theta)\right\} \mid y_{1:n}\right] \\
&+ 2\sum_{i=1}^n \mathbb{E}\left[\frac{d}{dy_i}\mathcal{D}_\gamma(y_i;\theta) \mid y_{1:n}\right] \\
&\mathbb{E}\left[C_\gamma^{(2)}(y_i;\theta) \mid y_{1:n}\right]\mathbb{E}\left[\frac{d}{d\gamma}\mathcal{D}_\gamma(y_{1:n};\theta) \mid y_{1:n}\right],
\end{aligned} \tag{D.4}$$

where $C^{(1)}$ and $C^{(2)}$ are defined in (D.2).

D.1 General case

Let $f(y_i;\theta)$ be a parametric density of y_i. The density power divergence (Basu et al. 1998) is

$$\mathcal{D}_\gamma(y_i;\theta) = \frac{1}{\gamma}f(y_i;\theta)^\gamma - \frac{1}{1+\gamma}\int f(t;\theta)^{1+\gamma}dt,$$

noting that the second term is irrelevant in the computation of the H-score since it does not depend on y_i. The detailed expressions of the quantities that appear in the derivative of the H-score in (D.4) are obtained as follows:

$$\begin{aligned}
C_\gamma^{(1)}(y_i;\theta) &= (\gamma-1)f(y_i;\theta)^{\gamma-2}f'(y_i;\theta)^2 \\
&\quad + f(y_i;\theta)^{\gamma-1}f''(y_i;\theta) \\
&\quad + f(y_i)^{2(\gamma-1)}f'(y_i)^2 \\
C_\gamma^{(2)}(y_i;\theta) &= f(y_i;\theta)^{\gamma-1}f'(y_i;\theta), \\
\frac{\partial}{\partial\gamma}\mathcal{D}_\gamma(y_{1:n};\theta) &= \frac{1}{\gamma^2}\sum_{i=1}^n f(y_i;\theta)^\gamma\{\gamma\log f(y_i;\theta) - 1\} \\
&\quad + \frac{n}{(1+\gamma)^2}\int f(t;\theta)^{1+\gamma}dt \\
&\quad - \frac{n}{1+\gamma}\int f(t;\theta)^{1+\gamma}\log f(t;\theta)dt \\
\frac{\partial}{\partial\gamma}C_\gamma^{(1)}(y_i;\theta) &= f(y_i;\theta)^{\gamma-2}f'(y_i;\theta)^2 \\
&\quad + \left\{C_\gamma^{(1)} + f(y_i;\theta)^{2(\gamma-1)}f'(y_i;\theta)\right\} \\
&\quad \log f(y_i;\theta) \\
\frac{\partial}{\partial\gamma}C_\gamma^{(2)}(y_i;\theta) &= C_\gamma^{(2)}(y_i;\theta)\log f(y_i;\theta)
\end{aligned}$$

D.2 Normal distribution case

When $y_i \sim \mathcal{N}(\mu,\sigma^2)$, the corresponding density power divergence is

$$\mathcal{D}_\gamma(y_i;\theta) = \gamma^{-1}\phi(y_i;\mu,\sigma^2)^\gamma - (2\pi\sigma^2)^{-\gamma/2}(1+\gamma)^{-3/2}.$$

The detailed expressions of quantities appeared in the derivative of the H-score in (D.4) are obtained as follows:

$$
\begin{aligned}
C_\gamma^{(1)}(y_i;\theta) &= \frac{1}{\sigma^4}\left[w_i\left\{\gamma(y_i-\mu)^2-\sigma^2\right\}+w_i^2(y_i-\mu)^2\right],\\
C_\gamma^{(2)}(y_i;\theta) &= \frac{d}{dy_i}\mathcal{D}_\gamma(y_i;\theta) = -\frac{w_i(y_i-\mu)}{\sigma^2},\\
\frac{d}{d\gamma}\mathcal{D}_\gamma(y_{1:n};\theta) &= \frac{1}{\gamma^2}\sum_{i=1}^{n} w_i\left\{\gamma\log\phi(y_i;\mu,\sigma^2)-1\right\}\\
&\quad+\frac{n}{2}(2\pi\sigma^2)^{-\gamma/2}(1+\gamma)^{-5/2}\\
&\quad\left\{(1+\gamma)\log(2\pi\sigma^2)+3\right\},\\
\frac{\partial}{\partial\gamma}C_\gamma^{(1)}(y_i;\theta) &= \frac{1}{\sigma^4}\Big[w_i\left\{\gamma(y_i-\mu)^2-\sigma^2\right\}\\
&\quad\log\phi(y_i;\mu,\sigma^2)+w_i(y_i-\mu)^2\\
&\quad+2w_i^2(y_i-\mu)^2\log\phi(y_i;\mu,\sigma^2)\Big],\\
\frac{\partial}{\partial\gamma}C_\gamma^{(2)}(y_i;\theta) &= -\frac{w_i(y_i-\mu)}{\sigma^2}\log\phi(y_i;\mu,\sigma^2),
\end{aligned}
$$

where $w_i = \phi(y_i;\mu,\sigma^2)^\gamma$. When $y_i \sim \mathcal{N}(x_i^\top\beta,\sigma^2)$, the derivative of the H score in the model is obtained by replacing μ with $x_i^\top\beta$.

References

Azlarov, T.A., Volodin, N.A.: Laws of large numbers for identically distributed banach-space valued random variables. Theory Probab. Appl. **26**(3), 573–580 (1982)

Basak, S., Basu, A., Jones, M.: On the optimal density power divergence tuning parameter. J. Appl. Stat. **48**(3), 536–556 (2021)

Basu, A., Harris, I.R., Hjort, N.L., Jones, M.: Robust and efficient estimation by minimising a density power divergence. Biometrika **85**(3), 549–559 (1998)

Bernardo, J.M., Smith, A.F.: Bayesian Theory, vol. 405. Wiley (2009)

Beskos, A., Jasra, A., Kantas, N., Thiery, A.: On the convergence of adaptive sequential monte Carlo methods. Ann. Appl. Probab. **26**(2), 1111–1146 (2016)

Beskos, A., Papaspiliopoulos, O., Roberts, G.: Monte Carlo maximum likelihood estimation for discretely observed diffusion processes. Ann. Stat. **37**(1), 223–245 (2009)

Bissiri, P.G., Holmes, C.C., Walker, S.G.: A general framework for updating belief distributions. J. R. Stat. Soc. Ser. B Stat Methodol. **78**(5), 1103 (2016)

Boustati, A., Akyildiz, Ö. D., Damoulas, T., Johansen, A.: Generalized Bayesian filtering via sequential Monte Carlo (2020). Preprint arXiv:2002.09998

Chérief-Abdellatif, B.-E., Alquier, P.: Mmd-bayes: robust Bayesian estimation via maximum mean discrepancy. In: Symposium on Advances in Approximate Bayesian Inference, pp. 1–21. PMLR (2020)

Chopin, N.: A sequential particle filter method for static models. Biometrika **89**(3), 539–552 (2002)

Chopin, N., Papaspiliopoulos, O.: An Introduction to Sequential Monte Carlo. Springer, Berlin (2020)

Cichocki, A., Amari, S.-I.: Families of alpha-beta-and gamma-divergences: flexible and robust measures of similarities. Entropy **12**(6), 1532–1568 (2010)

Cichocki, A., Cruces, S., Amari, S.: Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. Entropy **13**(1), 134–170 (2011)

Dai, C., Heng, J., Jacob, P.E., Whiteley, N.: An invitation to sequential monte carlo samplers (2020)

Dawid, A.P., Musio, M., et al.: Bayesian model selection based on proper scoring rules. Bayesian Anal. **10**(2), 479–499 (2015)

Del Moral, P., Doucet, A., Jasra, A.: Sequential Monte Carlo samplers. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **68**(3), 411–436 (2006)

Fong, E., Holmes, C.: On the marginal likelihood and cross-validation. Biometrika **107**(2), 489–496 (2020)

Frazier, D.T., Loaiza-Maya, R., Martin, G.M., Koo, B.: Loss-based variational bayes prediction (2021). Preprint arXiv:2104.14054

Fujisawa, H., Eguchi, S.: Robust parameter estimation with a small bias against heavy contamination. J. Multivar. Anal. **99**(9), 2053–2081 (2008)

Ghosh, A., Basu, A.: Robust bayes estimation using the density power divergence. Ann. Inst. Stat. Math. **68**(2), 413–437 (2016)

Ghosh, A., Harris, I.R., Maji, A., Basu, A., Pardo, L., et al.: A generalized divergence for statistical inference. Bernoulli **23**(4A), 2746–2783 (2017)

Hashimoto, S., Sugasawa, S.: Robust Bayesian regression with synthetic posterior distributions. Entropy **22**(6), 661 (2020)

Hyvärinen, A.: Estimation of non-normalized statistical models by score matching. J. Mach. Learn. Res. **6**(4) (2005)

Jacob, P.E., O'Leary, J., Atchadé, Y.F.: Unbiased Markov chain Monte Carlo methods with couplings. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **82**(3), 543–600 (2020)

Jewson, J., Rossell, D.: General Bayesian loss function selection and the use of improper models (2021). Preprint arXiv:2106.01214

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

Kamatani, K.: Efficient strategy for the Markov chain Monte Carlo in high-dimension with heavy-tailed target probability distribution. Bernoulli **24**(4B), 3711–3750 (2018)

Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014). Preprint arXiv:1412.6980

Knoblauch, J., Jewson, J., Damoulas, T.: Doubly robust bayesian inference for non-stationary streaming data with β-divergences (2018). Preprint arXiv:1806.02261

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

Matsubara, T., Knoblauch, J., Briol, F.-X., Oates, C., et al.: Robust generalised Bayesian inference for intractable likelihoods. J. R. Stat. Soc.: Ser. B (2022)

Middleton, L., Deligiannidis, G., Doucet, A., Jacob, P.E.: Unbiased smoothing using particle independent metropolis-hastings. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 2378–2387. PMLR (2019)

Nakagawa, T., Hashimoto, S.: Robust Bayesian inference via γ-divergence. Commun. Stat.-Theory Methods **49**(2), 343–360 (2020)

Neal, R.M.: Annealed importance sampling. Stat. Comput. **11**(2), 125–139 (2001)

Nielsen, F., Sun, K., Marchand-Maillet, S.: On hölder projective divergences. Entropy **19**(3), 122 (2017)

Reddi, S.J., Kale, S., Kumar, S.: On the convergence of adam and beyond (2019). Preprint arXiv:1904.09237

Rousseeuw, P.J., Leroy, A.M.: Robust Regression and Outlier Detection, vol. 589. Wiley (2005)

Shao, S., Jacob, P.E., Ding, J., Tarokh, V.: Bayesian model comparison with the hyvärinen score: computation and consistency. J. Am. Stat. Assoc. (2019)

Stigler, S.M.: Do robust estimators work with real data? Ann. Stat. 1055–1098 (1977)

Warwick, J., Jones, M.: Choosing a robustness tuning parameter. J. Stat. Comput. Simul. **75**(7), 581–588 (2005)