General Bayesian loss function selection and the use of improper models

Jack Jewson\textsuperscript{1,2} | David Rossell\textsuperscript{1,2}

\textsuperscript{1}Department of Business and Economics, Universitat Pompeu Fabra, Barcelona, Spain
\textsuperscript{2}Data Science Center, Barcelona School of Economics, Barcelona, Spain

\textbf{Correspondence}
Jack Jewson, Department of Business and Economics, Universitat Pompeu Fabra, Barcelona, Spain.
Email: jack.jewson@upf.edu

\textbf{Funding information}
Fundación BBVA, Grant/Award Number: Equipos de Investigacion Cientifica 2017; Ministerio de Ciencia e Innovación, Grant/Award Numbers: Plan Nacional PGC2018-101643-B-I00, Europa Excelencia grant EUR2020-112096, RYC-2015-18544

\textbf{Abstract}
Statisticians often face the choice between using probability models or a paradigm defined by minimising a loss function. Both approaches are useful and, if the loss can be re-cast into a proper probability model, there are many tools to decide which model or loss is more appropriate for the observed data, in the sense of explaining the data’s nature. However, when the loss leads to an improper model, there are no principled ways to guide this choice. We address this task by combining the Hyvärinen score, which naturally targets infinitesimal relative probabilities, and general Bayesian updating, which provides a unifying framework for inference on losses and models. Specifically we propose the $\mathcal{H}$-score, a general Bayesian selection criterion and prove that it consistently selects the (possibly improper) model closest to the data-generating truth in Fisher’s divergence. We also prove that an associated $\mathcal{H}$-posterior consistently learns optimal hyper-parameters featuring in loss functions, including a challenging tempering parameter in generalised Bayesian inference. As salient examples, we consider robust regression and non-parametric density estimation where popular loss functions define improper models for the data and hence cannot be dealt with using standard model selection tools. These examples illustrate advantages in robustness-efficiency.
trade-offs and enable Bayesian inference for kernel density estimation, opening a new avenue for Bayesian non-parametrics.

**KEYWORDS**
general Bayes, Hyvärinen score, improper models, kernel density estimation, loss functions, robust regression

1 | INTRODUCTION

A common task in Statistics is selecting which among a set of models is most appropriate for an observed data set \( y \). Tools to address this problem include a variety of penalised likelihood, shrinkage prior and Bayesian model selection methods. Under suitable conditions, these approaches consistently select the model closest to the data-generating truth in Kullback–Leibler divergence (e.g., see Rossell, 2021 and references therein for a recent discussion). However, many data analysis methods are not defined in terms of probability models but as minimising a given loss function, for example to gain robustness or flexibility. It is then no longer clear how to use the data to guide the choice of the most appropriate loss function, or its associated hyper-parameters. A key observation is that while the likelihood \( f_k(y; \theta_k) \) of a model \( k \) with parameters \( \theta_k \) always defines a loss function \( \ell_k(y; \theta_k) = -\log f_k(y; \theta_k) \) (Good, 1952), the converse is not true. The exponential of an arbitrary loss \( \ell_k(y; \theta_l) = \exp(-\ell_k(y; \theta_k)) \) may not integrate to a finite constant and therefore, defines an improper model on \( y \). For example, this occurs in robust regression with Tukey’s loss (Figure 1) and in kernel density estimation. There is also a growing literature following Basu et al. (1998) that applies the Tsallis score to define a robust loss function to fit any given probability model. Such robustification depends on a hyper-parameter that governs robustness-efficiency trade-offs and often leads to an improper model similar to Figure 1 (see Figure S1, and Yonekura and Sugasawa (2021) for a recent pre-print building on our improper model interpretation to address the hyper-parameter selection). In these scenarios traditional model selection tools are not applicable to choose the more appropriate loss. Neither are methods to evaluate predictive performance such as cross-validation, since they require specifying a loss or criterion to evaluate performance in the first place, and do not attain consistent model selection even in simple settings (Shao, 1997). Methods to tackle intractable but finite normalisation constants, such as approximate Bayesian computation (Beaumont et al., 2002; Robert, 2016), also do not apply since they require simulating from a proper model.

We propose methodology to evaluate how well each given loss \( \ell_k \) captures the unknown data-generating distribution \( g(y) \). The main idea is viewing \( \ell_k \) as defining a (possibly improper) model \( \exp\{-\ell_k(y, \theta_k)\} \), and then measuring how well it approximates \( g(y) \) via Fisher’s divergence. As we shall see, Fisher’s divergence and its related Hyvärinen score (Hyvärinen, 2005) do not depend on normalising constants, and in fact they allow for such constant to be infinite, hence giving a strategy to compare improper models. Note that one could conceivably define the likelihood in ways other than \( \exp\{-\ell_k(y, \theta_k)\} \). However, defining losses as negative log-likelihoods provides the only smooth, local, proper scoring rule (Bernardo, 1979), and is also the only transformation that leads to consistent parameter estimation for a certain general class of likelihoods (Bissiri & Walker, 2012). Further, it seems reasonable that the loss should be additive over independent pieces of information, and that the likelihood of an improper model should
factorise under such independence, and for both properties to hold one must take the exponent of the negative loss. Our framework consistently selects the best model in Fisher’s divergence, and in particular the (proper) data-generating model \( g(y) \) if it is under consideration. We also show how, after a model is chosen, one can learn important hyper-parameters such as the likelihood tempering in generalised Bayes and PAC-Bayes, robustness-efficiency trade-offs in regression and the level of smoothing in kernel density estimation. For clarity and space we focus on continuous real-valued \( y = (y_1, \ldots, y_n) \) with full support on \( \mathbb{R}^n \). Our ideas can be extended to other settings such as discrete or positive \( y \), for example, following Hyvärinen (2007). However, these are slightly less interesting in our context. Improper models cannot occur for discrete \( y \) with finite support, and one may log-transform a positive outcome and subsequently apply our methodology.

The use of probability models versus algorithms is one of the most fundamental, long-standing debates in Statistics. In an influential piece, Breiman (2001) argued that models are not realistic enough to represent reality in a useful manner, nor flexible enough to predict accurately complex real-world phenomena. Despite advances in flexible and non-parametric models, this view remains in the current era where predictive machine learning proliferates, and shows ample potential to tackle large and/or complex data. Their limitations notwithstanding, probability models remain a fundamental tool for research. Paraphrasing Efron (2020): ‘Abandoning mathematical models comes close to abandoning the historic scientific goal of understanding nature’. We agree with this view that there are many situations where models facilitate describing the phenomenon under study. We seek to bridge these two views by noting that loss functions define improper models, which also lead to natural interpretations in terms of relative probabilities, and proposing a strategy to learn which loss gives a better description of the process underlying the data.

Our strategy is to view \( f_k \) as expressing relative (as opposed to absolute) probabilities, for example \( f_k(y_0, \theta_k)/f_k(y_1, \theta_k) \) describes how much more likely it is to observe \( y \) near \( y_0 \) than near \( y_1 \). A convenient manner to describe such ratios is by comparing the gradient of \( \log f_k(y; \theta_k) \) to the gradient of the log data-generating density \( \log g(y) \). This can be achieved by minimising Fisher’s divergence

\[
D_F(g||f_k) := \frac{1}{2} \int \| \nabla_y \log g(y) - \nabla_y \log f_k(y; \theta_k) \|^2 g(y) dy,
\]

where \( \nabla_y \) is the gradient operator. Under certain minimal tail conditions, minimising Fisher’s divergence is equivalent to minimising the Hyvärinen score (Hyvärinen, 2005). The latter has
been used for models with intractable, but finite normalising constants (Hyvärinen, 2005) and more recently to define posterior distributions based on scoring rules (Giummolè et al., 2019) and to conduct Bayesian model selection using improper priors (Dawid & Musio, 2015; Shao et al., 2019). We consider for the first time its use to select between possibly improper models, and learn their associated hyper-parameters.

A possible alternative Fisher’s divergence proposed by Lyu (2009) is to use linear operators to define a generalised Fisher divergence. The operators do not require a finite normalisation constant, that is, they can be applied to improper models. Although interesting, the specific proposals in Lyu (2009) are a conditional mean operator for latent variable models and a marginalisation operator that requires proper conditionals, neither of which seems directly applicable to our setting. In fact, it is important to distinguish our framework from settings where, by combining proper conditional models, one defines an improper joint model. For example, intrinsic auto-regressive models in spatial Statistics have proper conditionals and an improper joint. Such models can be fit using a pseudo-likelihood (Besag, 1975) or the marginalisation operator of Lyu (2009), for instance. In our framework, neither the conditionals nor the joint of \( f_k(y; \theta_k) \) need be proper, for example, Tukey’s loss example in Figure 1.

It is also important to distinguish our framework with approaches designed for models with intractable, but finite, normalisation constants (i.e. proper models). Popular strategies include contrastive divergence (Hinton, 2002), minimum velocity learning (unpublished work by Movellan, 2007, see Wang et al., 2020b) and contrastive noise estimation (Gutmann & Hyvärinen, 2010). These methods define certain Monte Carlo dynamics to transition from observed samples from the data-generating \( g(y) \) into samples from \( f_k(y; \theta_k) \). Informally, if \( g(y) \) is close to \( f_k(y; \theta_k) \) then such dynamics have a small gradient, defining a divergence between these two distributions. These methods do not require evaluating the normalisation constant. However, the notion of sampling from \( f_k(y; \theta_k) \) requires it to be a proper probability model, and hence these diversences do not apply to improper models. Another interesting example for intractable normalisation constant are the kernel Stein discrepancy posteriors of Matsubara et al. (2022). However, Stein discrepancies are based on differences in expectations, and hence also require \( f_k \) to be proper. A further issue of kernel discrepancies is that they do not lead to coherent updating of beliefs, that is, the posterior \( \pi(\theta_k | y_1, y_2) \) obtained after observing \( (y_1, y_2) \) does not match the posterior based on observing \( y_2 \) and using \( \pi(\theta_k | y_1) \) as the prior.

The paper proceeds as follows. Section 2 reviews recent developments in Bayesian updating with loss functions, discusses our motivating examples and some failures of standard methodology. Section 3 explains how we interpret the inference provided by an improper model in terms of relative probabilities, and their relation to Fisher’s divergence and the \( H \)-score. It also outlines our methodology: the definition of an \( H \)-posterior, a \( \sqrt{n} \)-consistency result to learn parameters and hyper-parameters, and the definition of the integrated \( H \)-score and \( H \)-Bayes factors as a criterion to choose among possibly improper models. Section 4 gives consistency rates for \( H \)-Bayes factors, including important non-standard cases where optimal hyper-parameters lie at the boundary, as can happen when considering nested models. Section 5 applies our procedure to robust regression. Section 6 produces a Bayesian implementation of kernel density estimation, which cannot be tackled by standard Bayesian methods, since kernel densities define an improper model for the observed data. All proofs and some additional technical results are in Appendix S1. Code to reproduce the examples of Sections 5 and 6 can be found at https://github.com/jejewson/HyvarinenImproperModels.
2 | PROBLEM FORMULATION

We define the problem and notation and then provide the necessary foundations by reviewing general Bayesian updating, providing two motivating examples, identifying the shortcomings of standard approaches, and finally introducing Fisher's divergence and the Hyvärinen score.

Let $y = (y_1, \ldots, y_n)$ denote an observed continuous outcome, where $y_i \in \mathbb{R}$ are independent draws from an unknown data-generating distribution with density $g(\cdot)$. One is given a set of $M$ probability models and $L$ loss functions which, in general, may or may not include $g$. As usual each model $k = 1, \ldots, M$ is associated to a density $f_k(y; \theta_k, \kappa_k)$, where $\theta_k$ are parameters of interest and $\kappa_k$ are hyperparameters, $(\theta_k, \kappa_k) \in \Theta_k \times \Phi_k \subseteq \mathbb{R}^{d_k}$. Any such density defines a loss $\ell_k(y; \theta_k, \kappa_k) = -\log f_k(y; \theta_k, \kappa_k)$. Similarly, denote by $\ell_k(y; \theta_k, \kappa_k)$ for $k = M + 1, \ldots, M + L$ the given loss functions. For $k > M$, we refer to $f_k(y; \theta_k, \kappa_k) = \exp(-\ell_k(y; \theta_k, \kappa_k))$ as the (possibly improper) density associated to $\ell_k$. In general such $f_k$ need not integrate to a finite number with respect to $y$, that is, $f_k$ may define an improper model on $y$. Our goal is to choose which among $f_1, \ldots, f_{L+M}$ provides the best representation of $g$, in a sense made precise below.

2.1 | General Bayesian updating

In the frequentist paradigm it is natural to infer parameters by minimizing loss functions, a classical example being $M$-estimation (Huber & Ronchetti, 1981). Loss functions are also used in the PAC-Bayes paradigm, where one considers the posterior distribution on the parameters

$$\pi_k(\theta_k | y, \kappa_k) \propto \pi_k(\theta_k | \kappa_k) \exp\{-\ell_k(y; \theta_k, \kappa_k)\},$$

where $\pi_k(\theta_k | \kappa_k)$ is a given prior distribution and $\propto$ denotes ‘proportional to’. See Guedj (2019) for a review on PAC-Bayes, and Grünwald (2012) for the safe-Bayes paradigm, which can be seen as a particular case where $\ell_k$ is a tempered negative log-likelihood. At this stage we consider inference for $\theta_k$ for a given hyper-parameter $\kappa_k$, we discuss learning $\kappa_k$ later. See also (Giummolè et al., 2019) for a framework where the loss function is defined by scoring rules such as the Tsallis score and the Hyvärinen score. The latter gives rise to the $\mathcal{H}$-posterior discussed in Section 3.2, a critical component of our construction. As a key result supporting the interpretation of (2) as conditional probabilities akin to Bayes rule, Bissiri et al. (2016) showed that (2) leads to a coherent updating of beliefs, and referred to the framework as general Bayesian updating.

These results allow Bayesian inference on parameters $\theta_k$ based on loss functions. The properties of the general Bayesian posterior have been well-studied, for example under suitable regularity conditions Chernozhukov and Hong (2003) and Lyddon et al. (2019) showed that it is asymptotically normal. However, the emphasis of prior work is on inference for $\theta_k$. To our knowledge viewing $\exp\{-\ell_k(y; \theta_k, \kappa_k)\}$ as an improper density has not been considered, which is critical for interpretation and posterior predictive inference, nor has the problem of choosing which loss best represents the data.

2.2 | Motivating applications

We introduce two problems which, despite being classical, cannot be tackled with standard inference. We first consider robust regression where one contemplates a parametric model and a robust
loss, and wishes to assess which represents the data best. To our knowledge there are no solutions for this problem. We next consider learning the bandwidth in kernel density estimation, where the goal is predictive inference on future data. While there are many frequentist solutions, Bayesian methods are hampered by the associated loss defining an improper model for the observed data.

2.2.1 Robust regression with Tukey’s loss

Consider the linear regression of \( y_i \) on an \( p \)-dimensional vector \( x_i \),

\[
y_i = x_i^T \beta + \epsilon_i, \quad \text{with } E(\epsilon_i) = 0, V(\epsilon_i) = \sigma^2 \quad \text{for } i = 1, \ldots, n.
\]

Consider first a Gaussian model, denoted \( k = 1 \), so that \( f_1(y_i; \theta_1) = \mathcal{N}(y_i; x_i^T \beta, \sigma^2) \), where \( \theta_1 = \{ \beta, \sigma^2 \} \in \mathbb{R}^p \times \mathbb{R}_+ \), and there are no hyper-parameters (\( \kappa_1 = \emptyset \)). The negative log-likelihood gives the least-squares loss \( \ell_1(y; \theta_1) = \sum_{i=1}^n \ell_1(y_i; \theta_1) \), where

\[
\ell_1(y_i; \theta_1) = -\log f_1(y_i; \beta, \sigma^2) = \frac{1}{2} \log \left( 2\pi \sigma^2 \right) + \frac{(y_i - x_i^T \beta)^2}{2\sigma^2}.
\]  

(3)

Since the least-squares loss is non-robust to outliers, one may consider alternatives. A classical choice is Tukey’s loss (Beaton & Tukey, 1974), which we denote \( k = 2 \), given by

\[
\ell_2(y_i; \theta_2, \kappa_2) = \begin{cases} 
\frac{1}{2} \log \left( 2\pi \sigma^2 \right) + \frac{(y_i - x_i^T \beta)^2}{2\sigma^2} - \frac{(y_i - x_i^T \beta)^4}{2\sigma^4 \kappa_2^2} + \frac{(y_i - x_i^T \beta)^6}{6\sigma^6 \kappa_2^4}, & \text{if } |y_i - x_i^T \beta| \leq \kappa_2 \sigma \\
\frac{1}{2} \log \left( 2\pi \sigma^2 \right) + \frac{\kappa_2^2}{6}, & \text{otherwise}
\end{cases}
\]

(4)

where \( \theta_2 = \{ \beta, \sigma^2 \} \in \mathbb{R}^p \times \mathbb{R}_+ \) and \( \kappa_2 \in \mathbb{R}_+ \) is a cut-off hyper-parameter. Note that (4) is parametrised such that the units of the cut-off parameter \( \kappa_2 \) are SDs away from the mean.

The density \( f_2(y; \theta_2, \kappa_2) = \exp\left( -\sum_{i=1}^n \ell_2(y_i; \theta_2, \kappa_2) \right) \) integrates to infinity, defining an improper model. Figure 1 plots Tukey’s loss for several \( \kappa_2 \) and their corresponding densities. (4) is similar to (3) when \( |y_i - x_i^T \beta| \) is close to 0, while for large \( |y_i - x_i^T \beta| \) it becomes flat, bounding the influence of outliers. The Gaussian model is recovered when \( \kappa_2 = \infty \). As we shall see, such nested comparisons pose methodological challenges that motivated our developments. As a technical remark, in robust statistics \( \sigma^2 \) is typically estimated separately from \( \beta \), either as part of a two-stage procedure (see e.g. Chang et al., 2018) or using S-estimation (Rousseeuw & Yohai, 1984). Instead, our framework allows one to jointly estimate \( (\beta, \sigma^2) \).

We note that one can add more losses into our framework, for example those in (Basu et al., 1998; Black & Rangarajan, 1996) or Wang et al. (2020a). Also, our framework is not limited to linear regression. One may replace \( x_i^T \beta \) by a non-linear function, for example from a deep learning or Gaussian process regression.

Setting \( \kappa_2 \) in Tukey’s loss is related to the so-called robustness-efficiency trade-off, an issue that has been unsettled for at least 60 years; see Box (1953) and Tukey (1960). While \( \kappa_2 = \infty \) gives the most efficient parameter estimates if the data are Gaussian, they are least robust outliers. Decreasing \( \kappa_2 \) increases robustness, but can significantly reduce estimation efficiency if \( \kappa_2 \) is set too small. We address this issue by learning from the data whether or not to estimate parameters in a robust
fashion, that is, selecting between $\ell_1$ and $\ell_2$, and estimating $\kappa_2$ when $\ell_2$ is chosen. We remark that the estimated $\kappa_2$ does not attempt to provide an optimal robustness-efficiency trade-off in the sense of minimising estimation error under an arbitrary data-generating $g(y)$. Rather, it seeks to define a predictive distribution on $y$ that portrays its nature accurately, in terms of infinitesimal relative probabilities (Section 3). This said, our method attains a good robustness-efficiency trade-off in our examples. Also, per our model selection consistency results, if data are truly Gaussian (or nearly so) then our framework collapses to inference under the Gaussian model, leading to efficient estimates.

Despite the importance of the robustness-efficiency trade-off, we are aware of limited work setting $\kappa_2$ in a principled, data-driven manner. Rule-of-thumb methods are popular, for example setting $\kappa_2 = 4.6851$, gives approximately 95% asymptotic efficiency of L2 minimisation on the standard normal distribution of residuals’ (Belagiannis et al., 2015), setting $\kappa_2 = 1.547$ to obtain a breakdown of 1/2 (Rousseeuw & Yohai, 1984), or a balance of breakdown and efficiency (e.g. Riani et al., 2014). More formal approaches rely on estimating quantiles of the data (e.g. Sinova & Van Aelst, 2016), minimising an estimate of parameter mean squared error (see Li et al., 2021, who applied the method of Warwick & Jones, 2005 to Tukey’s loss), or minimising the maximum change in parameter estimates resulting from perturbing one observation (Li et al., 2021).

2.2.2 Non-parametric kernel density estimation

Suppose that $y_i \sim g$ independently for $i = 1, \ldots, n$ and one wishes to estimate $g$. The kernel density estimate at a given value $x$ is given by

$$
\hat{g}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x - y_i}{h} \right),
$$

where the kernel $K(\cdot)$ is a symmetric, finite variance probability density, and $h$ is the bandwidth parameter. For simplicity we focus on the Gaussian kernel $K(x) = N(x; 0, 1)$.

The bandwidth $h$ is an important parameter controlling the smoothness and accuracy of $\hat{g}_h$. Popular strategies to set the bandwidth are rule-of-thumb and plug-in methods (e.g. Silverman, 1986), cross-validation (Habbema & Van den Broek, 1974; Robert, 1976) and minimising an estimate of integrated square error (Bowman, 1984; Rudemo, 1982).

Unfortunately, standard Bayesian inference cannot be used to learn $h$ from data. The reason is that although (5) defines a proper probability distribution for a future observation $x$, a Bayesian framework requires a proper model for the observed data given the parameter. In our notation, the model likelihood $f(y; \theta, \kappa) \propto \prod_{i=1}^{n} \hat{g}_h(y_i)$ with $\theta = \emptyset$ and hyper-parameter $\kappa = \{h\}$ has an infinite normalising constant. To see this, note that

$$
\hat{g}_h(y_i) = \frac{1}{nh} \sum_{j=1}^{n} K \left( \frac{y_i - y_j}{h} \right) = \frac{1}{nh \sqrt{2\pi}} + \frac{1}{nh \sqrt{2\pi}} \sum_{j \neq i}^{n} \exp \left\{ -\frac{(y_i - y_j)^2}{2h^2} \right\},
$$

where the first term integrates to infinity with respect to $y_i$. Hence, (5) illustrates a situation where one has an ‘algorithm’ for producing a density estimate for future observations that defines an improper probability model for the observed data.
2.3 The failure of standard technology

As we discussed a main challenge is that standard tools are not, in general, applicable to compare improper models. Another challenge occurs when one wishes to estimate a hyperparameter $\kappa_k$ of a given improper model $k$, for example Tukey’s cut-off or the kernel bandwidth. For example, the general Bayesian might consider mimicking standard Bayes or marginal likelihood estimation by defining

$$\hat{\kappa}_k := \arg \max_{\kappa \in \mathcal{K}} \int \pi_k(\theta_k) \exp \{-\ell_k(y; \theta_k, \kappa_k)\} d\theta_k.$$  \hspace{1cm} (7)

Unfortunately, such procedure often produces degenerate estimates. For example, from (4), it is clear that for fixed $\theta_2 = \{\beta, \sigma^2\}$ Tukey’s loss is increasing in $\kappa_2$ and therefore (7) selects $\hat{\kappa}_2 = 0$ independently of the data. Similarly, in the kernel density estimation example (7) selects $h = 0$ (Habbema & Van den Broek, 1974; Robert, 1976).

2.4 Fisher’s divergence and the Hyvärinen score

Hyvärinen (2005) proposed a score matching approach for models with intractable normalising constants. Score matching minimises Fisher’s divergence to the data-generating $g(\cdot)$ in (1), that is

$$\theta_k^* := \arg \min_{\theta_k \in \Theta_k} D_F(g(\cdot)||f_k(\cdot; \theta_k)) = \arg \min_{\theta_k \in \Theta_k} \mathbb{E}_{z \sim g} \left[H(z; f_k(\cdot; \theta_k))\right],$$  \hspace{1cm} (8)

where the right-hand side follows from integration by parts under minimal tail conditions, and

$$H(z; f_k(\cdot; \theta_k)) := 2 \frac{\partial^2}{\partial z^2} \log f_k(z; \theta_k) + \left(\frac{\partial}{\partial z} \log f_k(z; \theta_k)\right)^2,$$  \hspace{1cm} (9)

is the Hyvärinen score ($H$-score). Importantly for our purposes, neither the definition of Fisher’s divergence neither the steps required to obtain the right-hand side of (8) require $f_k$ to be a proper density function. As discussed we focus on univariate $z \in \mathbb{R}$, but (9) can be extended to multivariate $z = (z_1, \ldots, z_d)$ (Hyvärinen, 2005) via

$$H(z; f_k(\cdot; \theta_k)) := 2 \sum_{j=1}^d \frac{\partial^2}{\partial z_j^2} \log f_k(z; \theta_k) + ||\nabla_z \log f_k(z; \theta_k)||^2.$$  \hspace{1cm} (10)

See also Lyu (2009) for further options on extending score-matching to multivariate settings.

Given $y_i \sim g$ independently across $i = 1, \ldots, n$ one can estimate $\theta_k^*$ by minimising

$$\hat{\theta}_k := \arg \min_{\theta_k \in \Theta_k} \frac{1}{n} \sum_{i=1}^n H(y_i; f_k(\cdot; \theta_k)).$$  \hspace{1cm} (11)

A critical feature for our purposes is that the $H$-score depends only on the first and second derivatives of $\log f_k$ and hence the normalising constant does not play a role, independently of whether it is finite or not. Hence, unlike methods designed for intractable but finite normalisation
constants, the $H$-score is applicable to improper models. The $H$-score enjoys desirable properties. For example, Dawid et al. (2016) proved that $\hat{\theta}_k$ is consistent and asymptotically normal, and (Dawid & Musio, 2015; Shao et al., 2019) showed that its prequential application leads to consistent Bayesian model selection under improper priors.

Another feature is that the $H$-score is computationally convenient in many common settings. Evaluating (10) has a linear cost in $n$ and in the outcome dimension $d$. There are however settings with many outcomes or parameters (e.g. deep learning) where evaluating the required derivatives can be cumbersome. In such settings one may for example use the denoising score matching framework of Vincent (2011), which avoids second-order derivatives, of the sliced score matching of Song et al. (2020), which only requires certain Hessian-vector products and provides a consistent and asymptotically normal estimator (under suitable conditions).

Closest to our work, Matsuda et al. (2019) proposed the score matching information criteria (SMIC) to select between models with intractable normalising constants. This criterion estimates Fisher's divergence by correcting the in-sample Hyvärinen score by an estimate of its asymptotic bias. We emphasise two main distinctions with our work. The first is the extension to improper models. Second, these authors used cross-validations and predictive criteria similar to the AIC, which do not lead to consistent model selection, whereas we focus on structural learning where one seeks guarantees on recovering the loss that best approximates the data-generating $g$.

3 | INFERENCE FOR IMPROPER MODELS

We now present our framework. Section 3.1 interprets an improper model in terms of relative probabilities and motivates Fisher's divergence as a criterion to fit such a model. Section 3.2 proposes using the $H$-score to define a general Bayesian posterior to learn hyper-parameters and choose among a collection of models, some or all of which may be improper. Section 3.3 proposes a Laplace approximation to the $H$-Bayes factors, which we use both in our theoretical treatment and examples. Finally, Section 3.4 argues for using the $H$-score within a two-step procedure: first selecting a model and estimating hyper-parameters using the $H$-score, then reverting to standard general Bayes to learn the parameters of interest.

3.1 | Inference through relative probabilities

Our goal is to select which model $f_k$ describes the data best, in terms of helping interpret the data-generating $g$. The main difficulty is that, since $f_k$ may be improper, it is unclear how to define ‘best’. Our strategy is to view $f_k$ as expressing relative probabilities, in contrast to the usual absolute probabilities. For example $f_k(y_0, \theta_k)/f_k(y_1, \theta_k)$ describes how much more likely is it to observe $y$ near $y_0$ than near $y_1$. As an illustration, consider Tukey’s loss in (4). For any pair $(y_0, y_1)$ such that $|y_0 - x^T \beta|, |y_1 - x^T \beta| < \kappa \sigma$ are small, Tukey’s loss is approximately equal to the squared loss, hence

$$\frac{f_2(y_0; \theta_2, \kappa_2)}{f_2(y_1; \theta_2, \kappa_2)} \approx \mathcal{N} \left( y_0; x^T \beta, \sigma^2 \right) \mathcal{N} \left( y_1; x^T \beta, \sigma^2 \right).$$ (12)
In contrast, for any pair such that $|y_0 - x^T \beta|, |y_1 - x^T \beta| > \kappa_2 \sigma$ we have $f_2(y_0, \theta_2, \kappa_2)/f_2(y_1, \theta_2, \kappa_2) = 1$. That is, Tukey’s loss induces relative beliefs that observations near the mode behave like Gaussian variables, while all faraway observations are equally likely. This encodes the notion that one does not know much about the tails beyond their being thick, which is difficult to express using a proper probability distribution.

We argue that Fisher’s divergence is well-suited to evaluate how closely the relative probabilities of any such $f_k$ approximate those from $g$. Assuming that the gradients of $g(y)$ and $f_k(y)$ are finite for all $y$, Fisher’s divergence in (1) can be expressed as

$$D_F(g||f_\theta) = \int \left\| \frac{\log (g(y+\epsilon)/g(y)) - \log (f_k(y+\epsilon; \theta)/f_k(y; \theta))}{\epsilon} \right\|_2^2 g(y) dy.$$  \hfill (13)

Therefore, minimising Fisher’s divergence (equivalently, the Hyvärinen score) targets a $f_k$ that approximates the relative probabilities of $g$ in an infinitesimal neighbourhood around $y$, in the quadratic error sense, on the average with respect to $g(y)$. This observation extends the usual motivation for the Hyvärinen score as a replacement of likelihood inference when the normalising constant is intractable to being a justifiable criteria to score improper models.

### 3.2 The $H$-score

We consider a general Bayesian framework where the loss is defined by applying the $H$-score to the density $f_k(y; \theta_k, \kappa_k) = \exp\{-\ell_k(y; \theta_k, \kappa_k)\}$, which gives the general posterior

$$\pi^H(\theta_k, \kappa_k | y) \propto \pi_k(\theta_k, \kappa_k) \exp \left\{ -\sum_{i=1}^{n} H(y_i; f_k(\cdot; \theta_k, \kappa_k)) \right\}.$$  \hfill (14)

We refer to (14) as the $H$-posterior, which is a particular case of the scoring rule posteriors of Giummolè et al. (2019). Note that (14) is different from the general Bayesian posterior directly associated to $\ell_k$ in (2). An important property of (14) is that it provides a consistent estimator for parameters $\theta_k$ and hyper-parameters $\kappa_k$. Specifically, Proposition 1 shows that, under regularity conditions, $\tilde{\eta}_k = (\tilde{\theta}_k, \tilde{\kappa}_k)$ maximising (14) recovers the optimal $\eta^*_k = (\theta^*_k, \kappa^*_k)$ according to Fisher’s divergence.

Proposition 1 requires mild regularity conditions A1–A3, discussed in Section A.1.2. Briefly, A1 requires continuous second derivatives of the Hyvärinen score, that it has a unique minimiser, and that its first derivative has finite variance. A2 requires that the Hyvärinen score is dominated by an integrable function, which can be easily seen to hold for Tukey’s loss, for example. Finally, A3 requires that the Hessian of the $H$-score is positive and finite around $\eta^*_k$.

Proposition 1 extends Hyvärinen (2005) (Corollary 3), who stated that $\tilde{\eta}_k$ converges to $\eta^*_k$ in probability, to also give a $1/\sqrt{n}$ convergence rate. Another difference is that Hyvärinen (2005) considered the well-specified case where $g(y) = f_k(y; \theta^*_k, \kappa^*_k)$, which in particular requires $f_k$ to be a proper model. Proposition 1 also extends Dawid et al. (2016, theorem 2), who proved asymptotic normality for $H$-score-based estimators. Said asymptotic normality does in general not hold when $\eta^*_k$ lies on the boundary of the parameter space, for example, for Tukey’s loss if the data are truly Gaussian then $\kappa^*_2 = \infty$. By Proposition 1, even if normality does not hold, one still
attains $\sqrt{n}$ consistency to estimate $\kappa_2^*$. In terms of technical conditions, Dawid et al. (2016) do not list them but refer to standard M-estimation theory assumptions, see theorem 5.23 in Van der Vaart (2000). These are similar to our assumptions and include differentiability and Lipschitz conditions in $\eta_k$, the existence of a second-order Taylor expansion around $\eta_k^*$, and a non-singular Hessian at $\eta_k^*$. The main differences are that we require twice differentiability in $y$ and $\eta_k$, and that we consider a compact parameter space to allow for boundary parameter values (e.g. $1/\kappa_2^* = 0$ for Tukey’s loss, after a re-parameterisation discussed in Section 4). Further, Proposition 1 extends previous results by explicitly considering improper models and the learning of their hyper-parameters.

As is standard under model mis-specification, the shape of the $H$-posterior does not match the frequentist distribution of the posterior mode $\hat{\eta}_k = (\hat{\theta}_k, \hat{\kappa}_k)$ (Giummolè et al., 2019), that is, it does not have valid frequentist coverage. Per Proposition 1 and Theorem 1 this is not a major issue for selecting a loss $k$ and hyper-parameters $\kappa_k$, which is our main focus. However, posterior inference on $\theta_k$ under the selected $(k, \kappa_k)$ should be properly calibrated, as we discuss in Section 3.4.

**Proposition 1.** Let $y = (y_1, \ldots, y_n) \sim g$, $\tilde{\eta}_k = (\tilde{\theta}_k, \tilde{\kappa}_k)$ maximise (14), and $\eta_k^* = (\theta_k^*, \kappa_k^*)$ minimise Fisher’s divergence from $f_k(\theta_k, \kappa_k)$ to $g$. Assume Conditions A1 and A2 in Section A.1.2. Then, as $n \to \infty$,

$$
\left\| \tilde{\eta}_k - \eta_k^* \right\|_2 = o_p(1),
$$

where $\left\| \cdot \right\|_2$ is the $L_2$-norm. Further, if Condition A3 also holds, then

$$
\left\| \tilde{\eta}_k - \eta_k^* \right\|_2 = O_p(1/\sqrt{n}).
$$

A consequence of Proposition 1 is that one can use (14) to learn tempering hyper-parameters. Specifically, suppose that one considers a family of losses $w_k \ell_k()$, where $w_k > 0$ is a tempering parameter. While $w_k$ does not affect the point estimate of $\theta_k$ given by (2), it plays an important role in driving the posterior uncertainty on $\theta_k$. Within our framework, one may define

$$
f_k(y; \theta_k, \kappa_k') = \exp\{-\ell_k'(y; \theta_k, \kappa_k')\} = \exp\{-w_k \ell_k'(y; \theta_k, \kappa_k)\},
$$

where $\kappa_k' = (\kappa_k, w_k)$ and $\ell_k'(y; \theta_k, \kappa_k') = w_k \ell_k(y; \theta_k, \kappa_k)$. By Proposition 1, one can consistently learn the Fisher-divergence optimal $\kappa_k'$, and in particular $w_k$. In contrast, in the general Bayes posterior (2) it is challenging to estimate such $w_k$. Current strategies to set $w_k$ are optimising an upper-bound on generalisation error in PAC-Bayes (Catoni, 2007), estimating $w_k$ via marginalisation similar to the ‘Safe Bayesian’ fractional likelihood approach of Grünwald (2012), or using information theoretic arguments to calibrate $w_k$ to match certain limiting sampling distributions (Holmes & Walker, 2017; Lyddon et al., 2019). These strategies essentially view $w_k$ as a tuning parameter. In contrast, in our framework $w_k$ is viewed as a parameter of interest that controls the dispersion of the improper model and affects its interpretation. See Section 6 for an illustration in kernel density estimation.

Recall that our main goal is model comparison. To this end, in analogy to the marginal likelihood in standard Bayesian model selection, we define the integrated $H$-score

$$
H_k(y) = \int \pi_k(\theta_k, \kappa_k) \exp \left\{ -\sum_{i=1}^n H(y_i; \tilde{f}_k(\cdot; \theta_k, \kappa_k)) \right\} \, d\theta_k d\kappa_k.
$$
Also, analogously to Bayes factors and posterior model probabilities, we define the $H$-Bayes factor as $B_{k_l}^{(H)} := \mathcal{H}_k(y)/\mathcal{H}_l(y)$ and

$$
\pi(k|y) = \frac{\mathcal{H}_k(y)\pi(k)}{\sum_l \mathcal{H}_l(y)\pi(l)} = \left(1 + \sum_{l \neq k} B_{lk}^{(H)} \frac{\pi(l)}{\pi(k)}\right)^{-1},
$$

where $\pi(k)$ is a given prior probability for each model $k$. In our examples we use uniform $\pi(k)$, since we focus on the comparison of a few models, but in high-dimensional settings it may be desirable to set $\pi(k)$ to favour simpler models.

We note that an interesting alternative strategy for model comparison, also based on the $\mathcal{H}$-score, is to extend the prequential framework of Dawid and Musio (2015) and Shao et al. (2019) designed for improper priors. Therein one could adopt a general Bayesian framework, replacing the likelihood by $f_k(x_i; \theta_k, \kappa_k) = \exp\{-\ell_k(x_i; \theta_k, \kappa_k)\}$. Prequential approaches enjoy desirable properties, such as consistency and leading to joint coherent inference on the model and parameter values. Unfortunately, prequential inference is computationally hard, particularly when considering several models. First, for each model inference needs to be updated $n$ times to calculate the one-step-ahead predictive distribution. Second, said updates are not permutation invariant, so one should in principle consider the $n!$ orderings of the data. Thus, while interesting, we leave such line of research for future work. We remark, however, that a salient feature in Dawid and Musio (2015) is that one may use improper priors. In contrast, our framework requires one to use proper priors. The reason is that otherwise one may suffer the usual Jeffreys-Lindley paradox in Bayesian model selection (Lindley, 1957). Specifically, if one considers two nested models and sets an improper prior under the larger one, then our $H$-Bayes factor in favour of the smaller model is infinite, that is, one selects the smaller model regardless of the data.

### 3.3 Laplace approximation and BIC-type criterion

There are many available strategies to compute or estimate integrals such as $\mathcal{H}_k(y)$ in (15) (see Llorente et al., 2020 for a review). For its speed and analytic tractability, we consider the Laplace approximation

$$
\tilde{\mathcal{H}}_k(y) := (2\pi)^{d_k} \pi_k(\tilde{\eta}_k) \exp\left\{-\sum_{i=1}^n H(y_i; f_k(\cdot; \tilde{\eta}_k))\right\} |A_k(\tilde{\eta}_j)|^{-\frac{1}{2}},
$$

with $\tilde{\eta}_k := \arg\min_{\eta_k} \sum_{i=1}^n H(y_i; f_k(\cdot; \eta_k)) - \log \pi_k(\eta_k),$

being the mode of the log $\mathcal{H}$-posterior, $A_k(\eta_k)$ its Hessian at $\eta_k$, and $\eta_k = (\theta_k, \kappa_k)$. We denote the corresponding Laplace approximate $\mathcal{H}$-Bayes-factor by $\tilde{B}_{k_l}^{(H)} := \mathcal{H}_k(y)/\mathcal{H}_l(y)$.

Computational tractability is important when one considers many models or the integrand is expensive to evaluate, for example, in our kernel density examples it requires $\mathcal{O}(n^2)$ operations. Further, the availability of a closed-form expression facilitates its theoretical study (Section 4). See Kass et al. (1990) for results on the validity of Laplace approximations. We do not undertake such a study, instead we prove our results directly for the approximation (17) that we actually use for inference.
Although we motivate our methodology from a Bayesian standpoint, we note that \( \bar{H}_k(y) \) can be viewed as Bayesian-inspired information criteria analogous to the BIC (Schwarz, 1978). Specifically, in regular settings where \( A_k \) is of order \( 1/n \), one could take the leading terms in \( \log \bar{H}_k(y) \) to obtain

\[
- \sum_{i=1}^{n} H(y_i; f_k(\cdot; \tilde{\eta}_k)) - \frac{d_k}{2} \log(n) + \log \pi_k(\tilde{\eta}_j),
\]

as a model selection criterion. This expression is analogous to the BIC, except for the log prior density term, which converges to a constant for any \( \pi_k \) bounded away from 0 and infinity. The log prior term can play a relevant role however when considering non-local priors where \( \pi_k(\eta_k) \) can be equal to 0 for certain \( \eta_k \), see Section 4.

3.4 Two-step inference and model averaging

As discussed the \( H \)-posterior (14) asymptotically recovers the parameters \( (\theta_k^*, \kappa_k^*) \) minimising Fisher’s divergence, whereas the general Bayesian posterior (2) recovers the parameters minimising the expected loss \( \tilde{\theta}_k = \arg\min_{\theta_k \in \Theta_k} \int \ell_k(y; \theta_k, \kappa_k)g(y)dy \), for a given \( \kappa_k \).

We adopt the pragmatic view that, while one may consider the \( H \)-posterior to choose a model \( k \) and learn the associated hyper-parameter \( \kappa_k \), after said choice one may want to obtain standard inference under the selected model. That is, one desires to learn

\[
\arg\min_{\theta_k \in \Theta_k} \int \ell_k(y; \theta_k, \kappa_k^*)g(y)dy.
\]

For example, suppose that the \( H \)-score selects a proper probability model (e.g. the Gaussian model). One may then wish that inference collapses to standard Bayesian inference under that model, rather than being based on the \( H \)-posterior in (14). This is easily achieved with a two-step procedure. First, one uses (16) to select \( \hat{k} \) and (14) to estimate \( \hat{\kappa}_k \). Second, given \( (\hat{k}, \hat{\kappa}_k) \) one uses the general Bayesian posterior (2) for \( \hat{\theta}_k \). A further alternative to selecting a single model is to mimic Bayesian model averaging (Hoeting et al., 1999), where the estimates under each model are weighted according to the posterior probabilities in (16).

As an important point for quantifying uncertainty on \( \theta_k \), if the data-generating \( g(y) \) is not contained in any of the considered models, the generalised posterior \( \pi_k(\theta_k|y, \kappa_k) \propto \exp\{-\ell_k(y; \theta_k, \kappa_k)\} \pi_k(\theta_k|\kappa_k) \) in (2) is miss-calibrated relative to the sampling distribution of its posterior mode \( \tilde{\theta}_k \). This issue is not specific to generalised posteriors, it also affects any standard Bayesian posterior when the model is mis-specified. When the posterior is asymptotically normal, it is possible to define a calibrated version that provides valid frequentist uncertainty quantification. Briefly, following Ribatet et al. (2012) and Giummolè et al. (2019), define the calibrated posterior as

\[
\pi^C_k(\theta_k|y, \kappa_k) = \pi_k(\tilde{\theta}_k + C(\theta_k - \tilde{\theta}_k)|y, \kappa_k),
\]

where \( C \) is any matrix satisfying \( C^T J_k C = J_k K_k^{-1} J_k \), \( J_k \) is the expected Hessian of \( -\ell_k \) and \( K_k \) the covariance of its gradient, evaluated at \( \tilde{\theta}_k \) (in practice, at its estimate \( \hat{\theta}_k \)). It is easy to check that \( \pi^C_k \) has the same mode as \( \pi_k \). Specifically, one may take \( C = J_k K_k^{-1} J_k \)
Then, the shape of the calibrated posterior \( \pi^C_k \) at \( \theta^*_k \) asymptotically matches the sampling distribution of \( \hat{\theta}_k \), and hence provides valid uncertainty quantification (Giummolè et al., 2019, theorem 3.1).

4 | CONSISTENCY OF \( H \)-SCORE MODEL SELECTION

We now state Theorem 1, our main result that (17) consistently selects the model closest in Fisher’s divergence to the data-generating \( g \). When several models attain the same minimum, as may happen when considering nested models, then (17) selects that of smallest dimension. The proof does not require that the Hyvärinen score is asymptotically normal, which holds under the conditions in theorem 2 of Dawid et al. (2016), but simply the \( \sqrt{n} \)-consistency proven in Proposition 1.

Theorem 1 mirrors standard results for Bayes factors (theorem 1 in Dawid, 2011), the main difference being that it involves Fisher’s rather than Kullback–Leibler divergence. As discussed after the theorem, when the optimal hyper-parameter occurs at the boundary the model selection consistency provided by Theorem 1 may not hold, unless one uses a suitable adjustment. Before stating the theorem, we interpret its implications. Part (i) considers a situation where one compares two models \( l \) and \( k \) such that the former is closer to \( g \) in Fisher’s divergence. Then \( B_{kl}^{(H)} \) converges to 0 at an exponential rate in \( n \). Part (ii) considers that both models attain the same Fisher’s divergence, for example, nested models such as Tukey’s and the Gaussian model when the data are truly Normal. Then, \( B_{kl}^{(H)} \) favours the smaller model at a polynomial rate in \( n \).

Dawid and Musio (2015) also proved model selection consistency for a prequential application of the \( H \)-score (which differs from our methodology), restricted to well-specified linear regression models. Shao et al. (2019) extended the results to other models but only considered non-nested comparisons, for nested cases only a conjecture is given. Similarly to us they require twice differentiability, Lipschitz conditions on the score and \( H \)-score functions, but other assumptions are stronger. For example, the posterior is assumed to concentrate on the optimal \( \eta^*_k \) (we prove such result) and must have certain bounded posterior expectations in supremum norm.

Theorem 1. Assume Conditions A1–A4, and let \( \eta^*_k = (\theta^*_k, \kappa^*_k) \) be as in Proposition 1.

(i) Suppose that \( \mathbb{E}_g[H(y; f_l(\cdot; \eta^*_l)) < \mathbb{E}_g[H(y; f_k(\cdot; \eta^*_k))] \). Then

\[
\frac{1}{n} \log B_{kl}^{(H)} = \mathbb{E}_g[H(y; f_l(\cdot; \eta^*_l))] - \mathbb{E}_g[H(y; f_k(\cdot; \eta^*_k))] + o_p(1).
\]

(ii) Suppose that \( \mathbb{E}_g[H(y; f_l(\cdot; \eta^*_l))] = \mathbb{E}_g[H(y; f_k(\cdot; \eta^*_k))] \). Then

\[
\log B_{kl}^{(H)} = \frac{d_l - d_k}{2} \log(n) + O_p(1).
\]

The result requires Conditions A1–A4 given and discussed in Section A.1.2. Conditions A1–A3 are mild and also summarised before Proposition 1, whereas A4 imposes a Lipschitz condition on the \( H \)-score and its Hessian. As an important remark, the slower rate in Part (ii) is due to Condition A1 that the prior density \( \pi_l(\eta^*_l) > 0 \) at the optimal \( \eta^*_l \), where \( l \) is the larger model. This defines a so-called local prior, in contrast to non-local priors (Johnson & Rossell, 2012; Rossell & Telesca, 2017) which place zero density at the value \( \kappa^*_{l \setminus k} \) where the more complicated model \( l \)
recover the simpler model \(k\). Since non-local priors violate A1, Corollary A.1 extends Theorem 1 to show that non-local priors attain faster rates in Part (ii), while maintaining the exponential rates in Part (i). We will demonstrate that this improvement can have non-negligible practical implications in Section 5.2.

Another practically relevant remark is that A3 requires a finite expected Hessian near the optimal \((\theta^*_k, \kappa^*_k)\), which can be problematic in certain settings. For example, in Tukey’s loss if data are truly Gaussian then \(\kappa^*_2 = \infty\), which leads to an infinite Hessian. For Tukey’s loss the problem can be avoided by re-parameterising \(\nu_2 = 1/\kappa_2^2\), for which the Hessian is finite, but more generally such a re-parameterisation may not be obvious or not exist. These cases provide a further use for non-local priors. By Corollary A.1, one may set a non-local prior that vanishes sufficiently fast at the boundary (basically, a faster rate than that at which the Hessian diverges) to attain model selection consistency.

As a final remark, we also prove that under extended Conditions A1–A7 the score matching information criterion of Matsuda et al. (2019) does not lead to consistent model selection for the nested setting in Part (ii), see Corollary A.1. This result is analogous to predictive criteria such as cross-validation or Akaike’s information criterion not leading to consistent model selection, see Shao (1997).

5 ROBUST REGRESSION WITH TUKEY’S LOSS

We revisit the robust regression in Section 2.2, where one considers a Gaussian model and the improper model defined by Tukey’s loss. Section 5.1 illustrates that when the data contain outliers, the \(H\)-score chooses Tukey’s model and learns its cut-off hyper-parameter in a manner that leads to robust estimation. Section 5.2 shows the opposite situation, where data are truly Gaussian, and the benefits of setting a non-local prior on Tukey’s cut-off hyper-parameter to improve the model selection consistency rate. Finally, Section 5.3 shows two gene expression datasets, one exhibiting Gaussian behaviour and the other thicker tails. We compare our results to the SMIC (Matsuda et al., 2019) which, despite not being designed to compare improper models, to our knowledge is the only existing criterion that can be used for this task.

The \(H\)-scores for the squared loss \((\ell_1)\) and Tukey’s loss \((\ell_2)\) (see Section A.2.1) are

\[
H_1(y; f(\cdot; x, \theta_1)) = \sum_{i=1}^{n} -\frac{2}{\sigma^2} + \frac{(y_i - x_i^T \beta)^2}{\sigma^4} \quad (20)
\]

\[
H_2(y; f(\cdot; x, \theta_2, \kappa_2)) = \sum_{|y_i - x_i^T \beta| \leq \kappa_2 \sigma} \left\{ \left( \frac{y_i - x_i^T \beta}{\sigma^2} - \frac{2(y_i - x_i^T \beta)^3}{\kappa_2^2 \sigma^4} + \frac{(y_i - x_i^T \beta)^5}{\kappa_2^4 \sigma^6} \right)^2 \right. \\
- 2 \left( \frac{1}{\sigma^2} - \frac{6(y_i - x_i^T \beta)^2}{\kappa_2^2 \sigma^4} + \frac{5(y_i - x_i^T \beta)^4}{\kappa_2^4 \sigma^6} \right) \right\}. \quad (21)
\]

Given our interest in learning \(\kappa_2\), it is important to remark that its role is to define the proportion of observations that should be viewed as outliers. Thus, too small \(\kappa_2\) leads to problems. First, minimising \(H_2\) in (21) has a trivial degenerate solution by setting \((\beta, \sigma, \kappa_2)\) where \(\kappa_2\) is so small that only one observation satisfies \(|y_i - x_i^T \beta| \leq \kappa_2 \sigma\), and \(y_i = x_i^T \beta\) for that observation. Such
solution is undesirable, as it views all observations but one as outliers. Fortunately, it is possible
to define a range of reasonable $\kappa_2$ using the notion of the breakdown point, that is, the number of
observations that can be perturbed without causing arbitrary changes to an estimator. Following
Rousseeuw and Yohai (1984), this leads to a constraint on $(\beta, \sigma, \kappa_2)$ such that
\begin{equation}
\frac{1}{\rho(\kappa_2, \kappa_2)} \sum_{i=1}^{n} \rho \left( \frac{y_i - x_i^T \hat{\beta}}{\sigma}, \kappa_2 \right) \leq \frac{n}{2} - p,
\end{equation}
where $\rho \left( \frac{y_i - x_i^T \hat{\beta}}{\sigma}, \kappa_2 \right) = \varepsilon_2(y_i; x_i, \theta_2, \kappa_2) - \frac{1}{2} \log(2\pi\sigma^2)$, $\varepsilon_2$ is as in (4), and $p = \dim(x_i)$. See Section A.4.2 for the derivation and further discussion.

Regarding priors, throughout we set $\sigma^2 \sim IG(2, 0.5)$ and $\beta|\sigma^2 \sim N(0, 5\sigma^2I)$ for the Gaussian and Tukey models, in the latter case truncated to satisfy (22). The idea is that these are
mildly informative priors, for example, the prior variance of $\sigma^2$ is infinite, but avoid degenerate
solutions by truncating using (22). For $\kappa_2$ we set an inverse gamma prior in terms of $\nu_2 = 1/\kappa_2^2$,
$\pi_2^{NL}(\nu_2) = IG(\nu_2; a_0, b_0)$. The inverse gamma is a non-local prior (i.e. has vanishing density at
$\nu_2 = 0$), which as discussed in Section 5.2 carries important benefits for model selection. Its prior
parameters $(a_0, b_0) = (4.35, 1.56)$ are set such that $P(\kappa_2 \in (1, 3)) = 0.95$. The reasoning is that one
assigns high prior probability to the cutoff being in a reasonable default range, between 1-σ and
3-σ. If data were truly Gaussian, 0.3% would lie outside the 3-σ region and 68.3% within 1-σ.
Hence, $\kappa_2 = 3$ would exclude clear outliers under the Gaussian, and $\kappa_2 = 1$ would keep most of
the Gaussian data. We remark that this is a mildly informative prior, when warranted by the data the
posterior can concentrate on $\kappa_2 \in [1, 3]$ outside this interval (e.g. for the DLD data in Section 5.3 the
posterior mode was $\hat{\kappa}_2 = 4.12$). See Section A.4.4 for further discussion.

Finally, to satisfy the differentiability conditions of Theorem 1 and enable the use of standard
second-order optimisation software, we implemented a differentiable approximation to the
indicator function in Tukey’s loss (see Section A.4.3).

### 5.1 The marginal $H$-score in $\kappa$

Our first example illustrates the properties of the $H$-score for calibrating the robustness-efficiency
trade-off, in a setting where the data contain outliers. We simulated $n = 500$ observations from
the data-generating $g(y) = 0.9N(y; 0, 1) + 0.1N(y; 5, 3)$. The goal is to estimate the parameters of
the larger component (uncontaminated data), in a manner that is robust to the presence of data
from the smaller component (outliers). We compare the estimation from the Gaussian model $f_1$, 
which is correctly specified for 90% of the data, with that of the robust improper model arising
from Tukey’s loss $f_2$ (where $x_i = 1$ contains only the intercept term).

A first question of interest is studying the ability of the $H$-score to learn the cutoff
hyper-parameter $\kappa_2$. To this end, we measured the evidence for different $\kappa_2$ provided by the
marginal $H$-score
\begin{equation}
\mathcal{H}_2(y; \kappa_2) = \int \pi_2(\theta_2) \exp \left( -\sum_{i=1}^{n} H(y_i; f_2(\cdot; \theta_2, \kappa_2)) \right) d\theta_2.
\end{equation}

The left panel in Figure 2 shows $\mathcal{H}_2(y; \kappa_2)$ (black line) for a grid $\kappa_2 \in \{1, 1.5, 2, \ldots, 10\}$, along
with an asymptotic approximation to the root mean squared error (RMSE) of $\hat{\beta}^{(\kappa_2)}$ motivated by
Warwick and Jones (2005) (see Section A.4.5). The marginal $H$-score is highest for $\hat{\kappa}_2 = 5$, that is,
this value is chosen as best approximating the data-generating g. The right panel of Figure 2 shows that Tukey’s model for $\hat{\kappa}_2 = 5$ provides a good description of the uncontaminated component of the data, in the sense of capturing the log-gradient of $g(y)$ around the mode, and excludes most outliers.

Interestingly, the RMSE associated to $\hat{\beta}(\kappa_2^2)$ at $\hat{\kappa}_2 = 5$ was close to optimal (Figure 2, grey line). A convenient feature of the asymptotic RMSE is that we can examine the decomposed effect of the bias (due to contamination) and the variance, see Section A.4.5 for details. For too small $\kappa_2$ the RMSE increases, since then there are more observations beyond the cutoff (viewed as outliers), increasing the variance of $\hat{\beta}(\kappa_2^2)$. If $\kappa_2$ is too large, then the contaminated observations lie within the cutoff, which increases the bias (recall that at $\kappa_2 = \infty$, $\hat{\beta}(\kappa_2^2)$ is the sample mean).

### 5.2 Non-local priors and model selection consistency

We demonstrate the selection consistency (Theorem 1) when data are truly Gaussian, and that setting a non-local prior on the cutoff hyper-parameter $\kappa_2$ speeds up this selection (Corollary A.1). We simulated 100 independent data sets of sizes $n = 100, 1000, 10,000$ and 100,000 from the data-generating $g(y_i) = \mathcal{N}(y_i; \mathbf{x}_i^T \beta, \sigma^2)$, where the first entry in $\mathbf{x}_i \in \mathbb{R}^6$ corresponds to the intercept and the remaining entries are Gaussian with unit variances and 0.5 pairwise covariances, $\beta = (0, 0.5, 1, 1.5, 0, 0)$ and $\sigma^2 = 1$.

Recall that Tukey’s loss collapses to the Gaussian model for $\kappa_2 = \infty$, and otherwise adds certain flexibility by allowing one to consider an improper model. If this extra flexibility is not needed, following Occam’s razor one wants to choose the Gaussian model. While Theorem 1 guarantees this to occur asymptotically, our experiments show that setting a local prior on $\kappa_2$ leads to poor performance, even for $n = 100,000$. Specifically, we compare our default non-local prior $\pi^{NLP}_2(v_2) = IG(v_2; 4.35, 1.56)$ to a (local) half-Gaussian prior $\pi^{LP}_2(v_2) \propto 1_{v_2 \geq 0} \mathcal{N}(v_2; 0, 1)$, where $v_2 = 1/\kappa_2^2$. By Corollary A.2, the log-$H$-Bayes factor under the non-local prior should favour the Gaussian model at least at a $\sqrt{n}$ rate, in contrast to the local prior’s log($n$) rate.

Figure 3 compares the SMIC with our integrated $H$-score under the local and non-local priors. Score differences are plotted such that negative values indicate correctly selecting the Gaussian model. Firstly, there is no evidence of SMIC being consistent as $n$ grows, even for $n = 100,000$ the wrong model was selected 11% of the time. The $H$-score under the local prior has a decreasing median in $n$, but exhibits heavy tails and even for $n = 100,000$ it also failed to select the Gaussian
model 11% of the time. Under the non-local prior, already for n = 1000 the correct decision was made 99% of the time. These experiments illustrate the benefits of non-local priors to penalise parameter values near the boundary (1/\( \kappa_2^2 \) = 0, in this example). Recall also that, as discussed in Section 4, in general in such situations a local prior need not even attain consistency.

5.3 | Real data sets

We considered two gene expression data sets from Rossell and Rubio (2018). In the first, the data are well-approximated by a Gaussian distribution, whereas the second exhibits thicker tails. In both examples we used the \( h \)-score to compare the Gaussian model (\( \ell_1 \)) and Tukey’s loss (\( \ell_2 \)).

5.3.1 | TGF-\( \beta \) data

The data set from Calon et al. (2012) concerns gene expression data for \( n = 262 \) colon cancer patients. Previous work (Rossell & Rubio, 2018; Rossell & Telesca, 2017) focused on selecting genes that have an effect on the expression levels of TGF-\( \beta \), a gene known to play an important role in colon cancer progression. Instead, we study the relation between TGF-\( \beta \) and the 7 genes (listed in Section A.5.1) that appear in the ‘TGF-\( \beta \) 1 pathway’ according to the KEGGREST package in R (Tenenbaum, 2016), so that \( p = 8 \) after including the intercept.

The top panels in Figure 4 summarises the results. The integrated \( h \)-score for the Gaussian model was \( \hat{H}_1(y) = 272.88 \) and that for Tukey’s loss \( \hat{H}_2(y) = 233.90 \), providing strong evidence for the Gaussian model. This is in agreement with the SMIC (SMIC\(_1\)(y) = −283.93 and SMIC\(_2\)(y) = −277.55, where minimisation is desired) and results in Rossell and Rubio (2018), who found evidence for Gaussian over (thicker) Laplace tails. The left panel shows the fitted densities which, in conjunction with the Q-Q Normal plots in Figure S3, show that the residual distribution is well-approximated by a Gaussian. The right panel shows the squared difference between the \( h \)-posterior mean parameter estimates of each \( \beta_j \) under Tukey’s model minus that under the
FIGURE 4  Top: TGF-β data, where the \( H \)-score selected the Gaussian model. Bottom: DLD data, where the \( H \)-score selected Tukey’s loss. Left: fitted Tukey-based density to the residuals. Right: Squared difference between the \( H \)-posterior mean estimates of each \( \beta_j \) under Tukey’s model minus that under the Gaussian (solid black line), and difference between their variances (estimated with \( B = 500 \) bootstrap re-samples).

Gaussian, and the differences between their sampling variances (estimated via bootstrap, dashed grey line). Both models returned very similar point estimates, but the Gaussian had smaller variance for all parameters. Altogether, these results strongly support that the Gaussian model should be selected over Tukey’s.

5.3.2  DLD dataset

We consider an RNA-sequencing data set from Yuan et al. (2016) measuring gene expression for \( n = 192 \) patients with different types of cancer. Rossell and Rubio (2018) studied the impact of 57 predictors on the expression of DLD, a gene that can perform several functions such as metabolism regulation. To illustrate our methodology, we selected the 15 variables with the five highest loadings in the first three principal components, and used the integrated \( H \)-score to choose between the Gaussian and Tukey’s loss. Section A.5.1 lists the selected variables.

The bottom panels of Figure 4 summarise the results. The \( H \)-score strongly supported Tukey’s loss (\( \tilde{H}_1(y) = 155.57 \) for the Gaussian model, \( \tilde{H}_2(y) = 783.94 \) for Tukey’s), with \( H \)-posterior mean estimate \( \tilde{k}_2 = 4.12 \). Indeed, the bottom left panel indicates that the residuals have thicker-than-Gaussian tails, see also the Q-Q Normal residual plot in Figure S3. The bottom right panel illustrates two things. Firstly, the estimated coefficients of 6 of the 16 predictors differ quite considerably between the Gaussian model and Tukey’s loss (solid line). Second, the latter often have smaller variance (estimated via bootstrap). Both observations
align with the presence of thicker-than-normal tails, which can cause parameter estimation biases and inflated variance. Notably, the $H$-score agrees with Rossell and Rubio (2018), who selected Laplace over Gaussian tails, but disagrees with the SMIC (SMIC$_1(\gamma_1;192) = -145.72$ and SMIC$_2(\gamma_1;192) = -141.67$, where minimisation is desired). We speculate that the SMIC results may have been affected by outliers, which could cause instability in the SMIC asymptotic bias estimation.

6 | KERNEL DENSITY ESTIMATION

We revisit the kernel density estimate $\hat{g}_h(\cdot)$ from Section 2.2.2. The associated loss is

$$\ell(y; y, h, w) = -w \log \hat{g}_h(y).$$

This loss has two hyper-parameters $\kappa = (h, w)$, where $w > 0$. As discussed in Section 3.2, standard Bayesian inference on the bandwidth $h$ is not possible, since the loss does not define a proper probability model for the observed data $y$. We also included a tempering hyper-parameter $w$ to illustrate how it can provide added flexibility and improve performance. A downside of allowing $w \neq 1$ is that the estimated density $\hat{g}_{h,w}(x) = \exp\{-\ell(x; y, h, w)\}$ for a future observation $x$ has a convoluted normalisation constant. In our examples we found it more convenient to approximate the required univariate integrals with numerical integration.

We note that while kernel density estimation is a well-studied area, to the best of our knowledge the incorporation of a tempering parameter $w$ has not been considered, and is hence a further innovation enabled by the $H$-score associated to (24) (derived in Section A.2.2).

Regarding the prior on $(h, w)$, we used the prior family $\pi(h^2, w) = IG(h^2; 2, b_0)\text{Exp}(w; \lambda_0)$, setting $(b_0, \lambda_0)$ to minimise the mean integrated squared error of $\hat{g}_{h,\theta}$ under truly Gaussian data. This results in a default $\pi(h^2, w)$ centred on $(h, w)$ values such that, if the data were truly Gaussian, the estimated density would be accurate, subject to the prior being minimally informative (e.g. $h^2$ has infinite prior variance). The elicited values were $b_0 = 0.024$ and $\lambda_0 = 0.725$. See Section A.4.7 for full details.

6.1 | Gaussian mixture implementations

We simulated data from four Gaussian mixture models considered by Marron and Wand (1992) (with two further examples provided in Section A.5.2). All scenarios consider a data-generating

$$g(y) = \sum_{j=1}^{J} m_j \mathcal{N}(y; \mu_j, \sigma_j),$$

with parameters chosen as follows.

- **Bimodal**: $J = 2$ components, $\mu_1 = -1.5, \mu_2 = 1.5, \sigma_1 = \sigma_2 = 1/2$ and $m_1 = m_2 = 0.5$.
- **Trimodal**: $J = 3$ components, $\mu_1 = -1.2, \mu_2 = 0, \mu_3 = 1.2, \sigma_1 = \sigma_3 = 3/5, \sigma_2 = 1/4, m_1 = m_3 = 0.45$, and $m_2 = 0.1$.
- **Claw**: $J = 6$ components, $\mu_1 = 0, \sigma_1 = 1$ and $m_1 = 0.5$ then for $j = 2, \ldots, 6$: $\mu_j = -2 + j/2$, $\sigma_j = 0.1$ and $m_j = 0.1$.
- **Skewed**: $J = 8$ components, for $j = 1, \ldots, 8$: $\sigma_j = (2/3)^{j-1}, \mu_j = 3(\sigma_j - 1)$ and $m_j = 1/8$. 
We compared the $H$-score estimated density when learning $h$ and $w$ from the data against the default kernel estimate in R, which sets the bandwidth using Silverman’s rule-of-thumb (Silverman, 1986), using unbiased cross-validation (Bowman, 1984; Rudemo, 1982) to set the bandwidth hyper-parameter, implemented in R’s kedd package (Guidoum, 2015), a Bayesian non-parametric density estimate using a Dirichlet Process mixture of Gaussians (DPMM), implemented in R’s dirichletprocess package (Ross & Markwick, 2018) using their default parameters and placing a prior on the Dirichlet Process concentration parameter, and a finite Bayesian Mixture model where we use the marginal likelihood to select the number of Gaussian mixture components, implemented in R’s mombf package (Rossell et al., 2021) (full details are provided in Section A.4.6). Since the data are truly generated by a finite Gaussian mixture, the last competitor serves as a benchmark for all methods. Additional comparisons against the $H$-score estimated density with fixed $w = 1$ are in Section A.5.2. We sampled $n = 1000$ observations from each simulation setting above and standardised the data to zero mean and unit variance, as recommended for the dirichletprocess package.

Figure 5 provides histograms of the standardised data and the estimated densities, while Table 1 shows Fisher’s divergence of the fitted models to the data-generating $g(y)$. In Table 1, the $H$-score estimate was always the best or second best, with significant improvements over other methods. Its performance overall was comparable to the finite mixture model, which assumes the correct data-generating mechanism. Among the other competitors the cross-validation estimate performed best. This is also apparent in Figure 5, where in the Claw data the DPMM and R’s kernel estimate missed the 5 modes, and to a lesser extent in the Trimodal and Skewed data.

![Figure 5](image_url)

**Figure 5** Density estimation for Gaussian mixture data. Histograms of observed data, standardised to 0 mean and unit variance, and estimated density by R’s density function with the bandwidth rule of Silverman (1986), unbiased cross-validated bandwidth estimation, DPMM, Bayesian mixtures with marginal likelihood selection of components, and the $H$-posterior estimate with tempering ($w \neq 1$). For Bayesian methods the density associated with MAP estimates is plotted. [Colour figure can be viewed at wileyonlinelibrary.com]
Table 1  Fisher’s divergence between the density estimates and the data-generating Gaussian mixtures in four simulation scenarios

| Scenario | KDE | KDE (ucv) | DPMM | Finite mixture model | $\mathcal{H}$-KDE ($w \neq 1$) |
|----------|-----|-----------|------|-----------------------|-------------------------------|
| Bimodal  | 1.03 | 0.37      | 0.13 | 0.10                  | **0.09**                      |
| Claw     | 13.77 | 6.09      | 15.25 | **2.17**              | 2.51                         |
| Trimodal | 0.26 | **0.12**  | 0.32 | 0.33                  | 0.18                         |
| Skewed   | 21.34 | 16.15     | 17.61 | **6.12**              | 9.51                         |

Note: The best performing method in each is highlighted in bold face.

sets. In the Claw data however, the cross-validation KDE exhibited a wiggly behaviour in the tails. Section A.5.2 shows that in most examples the $\mathcal{H}$-score estimate improved by learning the hyper-parameter $w$, relative to $w = 1$, suggesting that the tempering parameter can add useful flexibility.

7 | DISCUSSION

We considered a novel problem, that of selecting between probability models and possibly improper models defined via loss functions. Despite being non-standard, the latter can be naturally interpreted in terms of relative probabilities, and open an avenue to increase the flexibility at the disposal of the data analyst. Relative probabilities motivated our use of Fisher’s divergence and the Hyvärinen score, which led to an associated $\mathcal{H}$-posterior and $\mathcal{H}$-Bayes factors to estimate hyper-parameters and model selection, respectively. Fisher’s divergence satisfies two key desirata: (i) it can be applied to improper models, and (ii) the solution best approximates infinitesimal relative probabilities. Methods designed for intractable, but finite, normalisation constants do not satisfy (i). The generalised Fisher divergences of Lyu (2009) offer an interesting basis to extend our ideas, although their proposed marginalisation and the conditional expectation operators proposed do not seem directly applicable to our examples, and do not satisfy (ii). A further desideratum is being computationally tractable. The $\mathcal{H}$-score leads to a closed-form posterior in exponential family models (Hyvärinen, 2007), and here we used Laplace approximations for computational convenience, which we proved to recover the model closest to the data-generating truth in Fisher’s divergence.

A limitation worth mentioning of score-based methods, like Fisher’s divergence, is their inability to learn the mixture weights for an assumed mixture model when the components are well-separated (Wenliang & Kanagawa, 2020). This is not an issue in our Tukey and kernel density examples, which do not feature mixture weights to be learnt from data (in KDE each component receives a fixed $1/n$ weight). To illustrate this, in Section A.5.3 we extended the Tukey and KDE examples to a situation where the data-generating truth features two strongly-separated components. The performance of our methodology was essentially unaffected by the larger separation. We remark however that well-separated data would be problematic in settings such as mixture-model based regression, where one seeks to learn the mixture weights, for example.

We emphasise that whatever loss is defined as best depends on whether one uses the Hyvärinen score or some other model-fitting strategy, for example those discussed in the introduction.
We do not view the $H$-score as the only possible choice, but merely as a natural one leading to simple interpretations in terms of relative probabilities. We hope that the proposed approach is a solid first step in answering the question of how to choose loss functions for generalised Bayesian inference, and that it opens up interesting future work to tackle this important problem.

We illustrated how the method provides promising results in robust regression and kernel density estimation, providing a new strategy to address the robustness-efficiency trade-off for the former and enabling Bayesian inference in the latter. We remark however that Fisher divergence does not seek to optimise robustness nor efficiency. Rather, it defines a predictive criterion such that the chosen model best represents future observations, in terms of relative probabilities. This said, if the data-generating truth is contained in the considered models, then one selects the true model asymptotically and obtains optimal estimation. Fisher divergence was also shown by Lyu (2009) to possess robustness properties. It is equal to the derivative of the Kullback–Leibler divergence upon adding Gaussian noise with infinitesimal variance. Hence, Fisher divergence chooses a model that does not change much when noise is added to it. Interestingly, a referee pointed out that theorem 3.1 in Liu and Wang (2016) can be used to show that any Stein discrepancy is also the derivative of a Kullback–Leibler divergence, and hence shares this robustness property.

Our asymptotic results mimic those for Bayesian model selection in proper models. There is however an important issue when considering hyper-parameters whose optimal value lies at the boundary, where consistency may fail for standard priors and it may be necessary to use non-local priors. As future work, one could attempt to improve the finite sample performance of the $H$-score using weighting procedures to calibrate its sampling distribution, for example, as proposed by Dawid et al. (2016). Another interesting direction is to extend the robust regression examples to variable selection. Rossell and Rubio (2018) showed the importance of specifying a well-fitting error distribution to avoid non-negligible drops in statistical power, particularly in high-dimensional cases. Considering robust improper models such as that stemming from Tukey’s loss may improve the selection performance there.

We also provided a novel strategy to learn likelihood tempering hyper-parameters from the data, which could be combined with various tempered likelihood approaches (e.g. Grünwald, 2012; Holmes & Walker, 2017; Miller & Dunson, 2018) to estimate this value for standard likelihoods, or to learn the Tsallis score hyper-parameter when defining a robust version of any parametric model. Yet another interesting extension is to consider multivariate responses, for example in graphical or factor models where outliers can be hard to detect (e.g. Filzmoser et al., 2008; Hall et al., 2005) and result in significantly worse inference. In summary, we believe there are many settings where the consideration of improper models can lead to fruitful research avenues.

**ACKNOWLEDGEMENTS**

The authors would like to thank Piotr Zwiernik for insightful discussions about Tukey’s loss minimisation and the breakdown point. Jack Jewson and David Rossell were partially funded by the Ayudas Fundación BBVA a Equipos de Investigación Científica 2017 and Proyectos de Investigación Científica en Matemáticas 2021, and Government of Spain’s Plan Nacional PGC2018-101643-B-I00 grants. David Rossell was also partially funded by the Europa Excelencia grant EUR2020-112096 and Ramón y Cajal Fellowship RYC-2015-18544.
DATA AVAILABILITY STATEMENT
The data that support the findings of this study are openly available at https://github.com/jejewson/HyvarinenImproperModels.

ORCID
Jack Jewson https://orcid.org/0000-0003-2703-7180

REFERENCES
Basu, A., Harris, I.R., Hjort, N.L. & Jones, M. (1998) Robust and efficient estimation by minimising a density power divergence. Biometrika, 85(3), 549–559.
Beaton, A.E. & Tukey, J.W. (1974) The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics, 16(2), 147–185.
Beaumont, M.A., Zhang, W. & Balding, D.J. (2002) Approximate Bayesian computation in population genetics. Genetics, 162(4), 2025–2035.
Belagiannis, V., Rupprecht, C., Carneiro, G. & Navab, N. (2015) Robust optimization for deep regression. Proceedings of the IEEE international conference on computer vision, pp. 2830–2838.
Bernardo, J.M. (1979) Expected information as expected utility. The Annals of Statistics, 7(3), 686–690.
Besag, J. (1975) Statistical analysis of non-lattice data. Journal of the Royal Statistical Society: Series D (The Statistician), 24(3), 179–195.
Bissiri, P., Holmes, C. & Walker, S.G. (2016) A general framework for updating belief distributions. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 78(5), 1103–1130.
Bissiri, P.G. & Walker, S. (2012) On Bayesian learning via loss functions. Journal of Statistical Planning and Inference, 142(12), 3167–3173.
Black, M.J. & Rangarajan, A. (1996) On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. International Journal of Computer Vision, 19(1), 57–91.
Bowman, A.W. (1984) An alternative method of cross-validation for the smoothing of density estimates. Biometrika, 71(2), 353–360.
Box, G.E. (1953) Non-normality and tests on variances. Biometrika, 40(3/4), 318–335.
Breiman, L. (2001) Statistical modeling: the two cultures (with comments and a rejoinder by the author). Statistical Science, 16(3), 199–231.
Calon, A., Espinet, E., Palomo-Ponce, S., Tauriello, D.V., Iglesias, M., Céspedes, M.V. et al. (2012) Dependency of colorectal cancer on a TGF-β-driven program in stromal cells for metastasis initiation. Cancer Cell, 22(5), 571–584.
Catoni, O. (2007) Pac-Bayesian supervised classification: the thermodynamics of statistical learning. IMS.
Chang, L., Roberts, S. & Welsh, A. (2018) Robust lasso regression using Tukey’s biweight criterion. Technometrics, 60(1), 36–47.
Chernozhukov, V. & Hong, H. (2003) An MCMC approach to classical estimation. Journal of Econometrics, 115(2), 293–346.
Dawid, A. (2011) Posterior model probabilities. In: Philosophy of statistics. North Holland: Elsevier, pp. 607–630.
Dawid, A.P. & Musio, M. (2015) Bayesian model selection based on proper scoring rules. Bayesian Analysis, 10(2), 479–499.
Dawid, A.P., Musio, M. & Ventura, L. (2016) Minimum scoring rule inference. Scandinavian Journal of Statistics, 43(1), 123–138.
Efron, B. (2020) Prediction, estimation, and attribution. Journal of the American Statistical Association, 115(530), 636–655.
Filzmoser, P., Maronna, R. & Werner, M. (2008) Outlier identification in high dimensions. Computational Statistics & Data Analysis, 52(3), 1694–1711.
Giummolè, F., Mameli, V., Ruli, E. & Ventura, L. (2019) Objective Bayesian inference with proper scoring rules. Test, 28(3), 728–755.
Good, I.J. (1952) Rational decisions. Journal of the Royal Statistical Society, Series B, 14, 107–114.
Grünwald, P. (2012) The safe Bayesian. Proceedings of the international conference on algorithmic learning theory, Springer, pp. 169–183.
Guedj, B. (2019) A primer on PAC-Bayesian learning. arXiv preprint arXiv:1901.05353.

Guidoum, A.C. (2015) Kernel estimator and bandwidth selection for density and its derivatives. Houari Boumediene, Algeria: Department of Probabilities and Statistics, University of Science and Technology.

Gutmann, M. & Hyvärinen, A. (2010) Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. Proceedings of the 13th international conference on artificial intelligence and statistics, JMLR workshop and conference proceedings, pp. 297–304.

Habbema J.D. & Van den Broek K. (1974) A stepwise discriminant analysis program using density estimation.

Hall, P., Marron, J. & Neeman, A. (2005) Geometric representation of high dimension, low sample size data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(3), 427–444.

Hinton, G.E. (2002) Training products of experts by minimizing contrastive divergence. Neural Computation, 14(8), 1771–1800.

Hoeting, J.A., Madigan, D., Raftery, A.E. & Volinsky, C.T. (1999) Bayesian model averaging: a tutorial. Statistical Science, 14(4), 382–401.

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

Huber, P.J. & Ronchetti, E. (1981) Robust statistics, series in probability and mathematical statistics.

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

Hyvärinen, A. (2007) Some extensions of score matching. Computational Statistics & Data Analysis, 51(5), 2499–2512.

Johnson, V.E. & Rossell, D. (2012) Bayesian model selection in high-dimensional settings. Journal of the American Statistical Association, 107(498), 649–660.

Kass, R.E., Tierney, L. & Kadane, J.B. (1990) The validity of posterior expansions based on Laplace’s method.

In: Geisser, S., Hodges, J.S., Press, S.J. & Zellner, A. (Eds.) Essays in Honor of George Barnard. Amersterdam: North-Hollans, pp. 473–488.

Li, Q., Chen, H. & Zhu, F. (2021) Robust estimation for Poisson integer-valued GARCH models using a new hybrid loss. Journal of Systems Science and Complexity, 34, 1578–1596.

Lindley, D.V. (1957) A statistical paradox. Biometrika, 44(1/2), 187–192.

Liu, Q. & Wang, D. (2016) Stein variational gradient descent: a general purpose Bayesian inference algorithm. Advances in Neural Information Processing Systems, 29, 1–9.

Llorente, F., Martino, L., Delgado, D. & Lopez-Santiago, J. (2020) Marginal likelihood computation for model selection and hypothesis testing: an extensive review. arXiv 2005.08334, 1–91.

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

Lyu, S. (2009) Interpretation and generalization of score matching. Proceedings of the 25th conference on uncertainty in artificial intelligence, pp. 359–366.

Marron, J.S. & Wand, M.P. (1992) Exact mean integrated squared error. The Annals of Statistics, 20(2), 712–736.

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, 84(3), 997–1022.

Matsuda, T., Uehara, M. & Hyvarinen, A. (2019) Information criteria for non-normalized models. arXiv preprint arXiv:1905.05976.

Miller, J.W. & Dunson, D.B. (2018) Robust Bayesian inference via coarsening. Journal of the American Statistical Association, 114(527), 1113–1125.

Movellan, J.R. (2007) A minimum velocity approach to learning. Unpublished draft.

Riani, M., Cerioli, A. & Torti, F. (2014) On consistency factors and efficiency of robust S-estimators. Test, 23(2), 356–387.

Ribatet, M., Cooley, D. & Davison, A.C. (2012) Bayesian inference from composite likelihoods, with an application to spatial extremes. Statistica Sinica, 22(2), 813–845.

Robert, C.P. (2016) Approximate Bayesian computation: a survey on recent results. In: Monte Carlo and Quasi-Monte Carlo methods. Cham: Springer, pp. 185–205.

Robert, P. (1976) On the choice of smoothing parameters for Parzen estimators of probability density functions. IEEE Transactions on Computers, 25(11), 1175–1179.
Ross, G.J. & Markwick, D. (2018) Dirichlet process: an R package for fitting complex Bayesian nonparametric models.
Rossell, D. (2021) Concentration of posterior model probabilities and normalized L0 criteria. Bayesian Analysis, 17(2), 565–591.
Rossell, D., Cook, J.D., Telesca, D., Roebuck, P. & Rossell, M.D. (2021) Package mombf.
Rossell, D. & Rubio, F.J. (2018) Tractable Bayesian variable selection: beyond normality. Journal of the American Statistical Association, 113(524), 1742–1758.
Rossell, D. & Telesca, D. (2017) Nonlocal priors for high-dimensional estimation. Journal of the American Statistical Association, 112(517), 254–265.
Rousseeuw, P. & Yohai, V. (1984) Robust regression by means of S-estimators. In: Robust and nonlinear time series analysis. New York: Springer, pp. 256–272.
Rudemo, M. (1982) Empirical choice of histograms and kernel density estimators. Scandinavian Journal of Statistics, 9(2), 65–78.
Schwarz, G. (1978) Estimating the dimension of a model. The Annals of Statistics, 6(2), 461–464.
Shao, J. (1997) An asymptotic theory for linear model selection. Statistica Sinica, 7(2), 221–242.
Shao, S., Jacob, P.E., Ding, J. & Tarokh, V. (2019) Bayesian model comparison with the Hyvärinen score: computation and consistency. Journal of the American Statistical Association, 114(528), 1826–1837.
Silverman, B.W. (1986) Density estimation for statistics and data analysis, Vol. 26. Boca Raton: CRC Press.
Sinova, B. & Van Aelst, S. (2016) Tukey’s biweight loss function for fuzzy set-valued M-estimators of location. Proceedings of the international conference on soft methods in probability and statistics, Springer, pp. 447–454.
Song, Y., Garg, S., Shi, J. & Ermon, S. (2020) Sliced score matching: a scalable approach to density and score estimation. In: Uncertainty in artificial intelligence. PMLR, pp. 574–584.
Tenenbaum, D. (2016) Keggrest: client-side rest access to KEGG. R package version 1 1.
Tukey, J.W. (1960) A survey of sampling from contaminated distributions. In: Contributions to probability and statistics, pp. 448–485.
Van der Vaart, A.W. (2000) Asymptotic statistics, Vol. 3. Cambridge: Cambridge University Press.
Vincent, P. (2011) A connection between score matching and denoising autoencoders. Neural Computation, 23(7), 1661–1674.
Wang, L., Peng, B., Bradic, J., Li, R. & Wu, Y. (2020a) A tuning-free robust and efficient approach to high-dimensional regression. Journal of the American Statistical Association, 115(532), 1700–1714.
Wang, Z., Cheng, S., Yueru, L., Zhu, J. & Zhang, B. (2020b) A Wasserstein minimum velocity approach to learning unnormalized models. Proceedings of the international conference on artificial intelligence and statistics, PMLR, pp. 3728–3738.
Warwick, J. & Jones, M. (2005) Choosing a robustness tuning parameter. Journal of Statistical Computation and Simulation, 75(7), 581–588.
Wenliang, L.K. & Kanagawa, H. (2020) Blindness of score-based methods to isolated components and mixing proportions. arXiv preprint arXiv:2008.10087.
Yonekura, S. & Sugasawa, S. (2021) Adaptation of the tuning parameter in general Bayesian inference with robust divergence. arXiv preprint arXiv:2106.06902.
Yuan, T., Huang, X., Woodcock, M., Du, M., Dittmar, R., Wang, Y. et al. (2016) Plasma extracellular RNA profiles in healthy and cancer patients. Scientific Reports, 6(1), 1–11.

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: Jewson, J. & Rossell, D. (2022) General Bayesian loss function selection and the use of improper models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 84(5), 1640–1665. Available from: https://doi.org/10.1111/rssb.12553