Article

A Parametric Bayesian Approach in Density Ratio Estimation

Abdolnasser Sadeghkhani ^{1,*} , Yingwei Peng ² and Chunfang Devon Lin ³¹ Department of Mathematics & Statistics, Brock University, St. Catharines, ON L2S 3A1, Canada² Departments of Public Health Sciences, Queen's University, Kingston, ON K7L 3N6, Canada; yingwei.peng@queensu.ca³ Department of Mathematics & Statistics, Queen's University, Kingston, ON K7L 3N6, Canada; devon.lin@queensu.ca

* Correspondence: asadeghkhani@brocku.ca

Received: 7 March 2019; Accepted: 26 March 2019; Published: 30 March 2019

Abstract: This paper is concerned with estimating the ratio of two distributions with different parameters and common supports. We consider a Bayesian approach based on the log–Huber loss function, which is resistant to outliers and useful for finding robust M-estimators. We propose two different types of Bayesian density ratio estimators and compare their performance in terms of frequentist risk function. Some applications, such as classification and divergence function estimation, are addressed.

Keywords: Bayes estimator; Bregman divergence; density ratio; exponential family; log–Huber loss

1. Introduction

The problem of estimating the ratio of two densities appears in many areas of statistical and computer science. Density ratio estimation (DRE) is widely considered to be the most important factor in machine learning and information theory (Sugiyama et al. [1], and Kanamori et al. [2]). In a series of papers, Sugiyama et al. [3,4] developed DRE in different statistical data analysis problems. Some useful applications of DRE are as follows: non-stationary adaptation (Sugiyama and Müller [5] and Quiñonero-Candela et al. [6]), variable selection (Suzuki et al. [7]), dimension reduction (Suzuki and Sugiyama, [8]), conditional density estimation (Sugiyama et al. [9]), outlier detection (Hido et al. [10]), and posterior density estimation (Thomas et al. [11]), among others. This paper addresses cases where the density belongs to the parametric distributions (e.g., exponential family of distributions). Parametric methods are usually favorable thanks to the existence of closed forms (mostly simple and explicit formulas) and hence help enhance computational efficiency.

Several estimators of the ratio of two densities have recently been proposed. One of the simplest approaches to estimate the density ratio p/q , where p and q are two probability density (or mass) functions (PDFs or PMFs), is called “plug-in”, and the ratio of the estimated densities is computed. Alternatively, one can estimate the ratio of two densities directly. Several approaches have recently been explored to estimate the ratio, including the moment matching approach (Gretton et al. [12]) and the density matching approach (Sugiyama et al. [13]). There are other works, such as Nguyen [14] and Deledalle [15], which studied the application of DRE in estimating Kullback–Leibler (KL) divergence (or more generally the α -divergence function, also known as f -divergence) and vice versa. The main objective of this paper is to address Bayesian parametric estimation with some commonly used loss functions for the ratio of p/q and to compare the proposed estimators with other estimators in the literature.

The remainder of the paper is organized as follows. In Section 2, we discuss the DRE methodology and introduce some useful definitions and related examples. Section 3 discusses how to find a Bayesian DRE under several loss functions for any arbitrary prior density on the parameter, and provides some interesting examples from the exponential families. In Section 4 we study some of the DRE applications. Some numerical illustrations for considering the efficiency of the proposed DREs are given in Section 5. Finally, we make some concluding remarks in Section 6.

2. Density Ratio Estimation for the Exponential Distribution Family

Let $X|\eta$ and $Y|\gamma$ be conditionally independent multivariate random variables

$$p(x|\eta) = h(x) \exp \left\{ \eta^\top s(x) - \kappa(\eta) \right\}, \quad (1)$$

$$q(y|\gamma) = h(y) \exp \left\{ \gamma^\top s(y) - \kappa(\gamma) \right\}, \quad (2)$$

where $\eta, \gamma \in \mathbb{R}^d$ are natural parameters, $s(\cdot)$ is a sufficient statistic, and $\kappa(\cdot) = \log c(\cdot)$ is the log-normalizer, which ensures that the distribution integrates to one, that is, $c(a) = \int h(x) \exp \{a^\top s(x)\} dx$, for $a = \eta$ and γ .

Consider the problem of estimating the density ratio

$$r(t; \eta, \gamma) = \frac{p(t|\eta)}{q(t|\gamma)}, \quad (3)$$

which is obviously proportional to $\exp \{(\eta - \gamma)^\top s(t)\}$. So, one can merge two natural parameters η and γ into one single parameter $\beta = \eta - \gamma$ (it can be a vector), and write the density ratio in (3) as follows

$$r(t; \theta) = \exp \left\{ \alpha + \beta^\top s(t) \right\}, \quad (4)$$

where $\alpha = \kappa(\gamma) - \kappa(\eta)$, and $\theta = (\alpha, \beta^\top)$ are parameters of interest. Note that since $r(t; \theta)$ itself belongs to the exponential family, the normalization term α can be considered as $-\log N(\beta)$, where $N(\beta) = \int q(t) \exp \{ \beta^\top s(t) \} dt$, which guarantees $\int q(t|\gamma) r(t; \theta) dt = 1$ and hence $q(t)r(t; \theta)$ becomes a valid PDF (PMF).

For instance, suppose two normal densities $p(t|\mu_1, \sigma_1^2) = N(t|\mu_1, \sigma_1^2)$ and $q(t|\mu_2, \sigma_2^2) = N(t|\mu_2, \sigma_2^2)$. They correspond to (1) and (2) respectively. One can easily verify that $\eta = (\frac{\mu_1}{\sigma_1^2}, \frac{-1}{2\sigma_1^2})^\top$, $\gamma = (\frac{\mu_2}{\sigma_2^2}, \frac{-1}{2\sigma_2^2})^\top$, $s(t) = (t, t^2)^\top$, and $\kappa(\eta) = \frac{\mu_1^2}{2\sigma_1^2} + \log \sigma_1^2$, $\kappa(\gamma) = \frac{\mu_2^2}{2\sigma_2^2} + \log \sigma_2^2$, and so according to (4), we have $\alpha = \log \frac{\sigma_1}{\sigma_2} + \frac{\mu_1^2}{2\sigma_1^2} - \frac{\mu_2^2}{2\sigma_2^2}$ and $\beta = \left(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}, \frac{1}{2\sigma_1^2} - \frac{1}{2\sigma_2^2} \right)^\top$.

Another interesting point is the relationship between DRE and the probabilistic classification problem (Sugiyama et al. [16]). Suppose $Z = \{0, 1\}$ is a binary response variable which shows whether an observation t is drawn from either $p(t)$ (say $Z = 0$) or $q(t)$ (say $Z = 1$). In other words $p(t) = \mathbb{P}(t|Z = 0)$ and $q(t) = \mathbb{P}(t|Z = 1)$. The standard logistic regression model is:

$$\mathbb{P}(Z = 1|t) = \frac{\exp(\beta_0 + \beta^\top t)}{1 + \exp(\beta_0 + \beta^\top t)},$$

and consequently, $\mathbb{P}(Z = 0|t) = \frac{1}{1 + \exp(\beta_0 + \beta^\top t)}$ and

$$\frac{\mathbb{P}(Z = 1|t)}{\mathbb{P}(Z = 0|t)} = \exp(\beta_0 + \beta^\top t). \quad (5)$$

Letting $\pi = \mathbb{P}(Z = 1) = \int \mathbb{P}(Z = 1 | t) p(t) dt$, and applying Bayes' rule, we have

$$p(t) = \frac{\exp(\beta_0 + \beta^\top t)}{(1 + \exp(\beta_0 + \beta^\top t))\pi},$$

$$q(t) = \frac{p(t)}{(1 + \exp(\beta_0 + \beta^\top t))(1 - \pi)},$$

and hence

$$\frac{p(t)}{q(t)} = \frac{1 - \pi}{\pi} \exp(\beta_0 + \beta^\top t) = \exp\left(\alpha + \log \frac{1 - \pi}{\pi} + \beta^\top t\right),$$

which indeed has the form of $\exp(\alpha + \beta^\top t)$ with $\alpha = \beta_0 + \log \frac{1 - \pi}{\pi}$ and hence has a density ratio model in (4).

However if our response $Z = \{0, 1, \dots, m - 1\}$, we can extend Equation (5) to the multinomial logit model, and we have

$$\frac{\mathbb{P}(Z = k | t)}{\mathbb{P}(Z = 0 | t)} = \exp(\beta_{0k} + \beta_k^\top t), \quad k = 1, \dots, m - 1. \quad (6)$$

Analogously, letting $\mathbb{P}_k(t) = \mathbb{P}(t | Z = k)$ and $\pi_k = \mathbb{P}(Z = k)$ for $k = 1, \dots, m - 1$, we have $\frac{\mathbb{P}_k(t)}{\mathbb{P}_0(t)} = \exp(\alpha_k + \beta_k^\top t)$, with $\alpha_k = \beta_{0k} + \log \frac{1 - \sum_{k=1}^{m-1} \pi_k}{\pi_k}$.

In fact, a conceptually simple solution to the DRE is to separately estimate both densities and calculate the ratio. This can be done by replacing the estimators of the parameters into each density. This approach is known as a plug-in density ratio estimation, defined below:

$$\hat{r}_{plug}(t) = \frac{p(t | \hat{\eta}(t))}{q(t | \hat{\gamma}(t))}, \quad (7)$$

where $\hat{\eta}(\cdot)$ and $\hat{\gamma}(\cdot)$ are estimates of parameters of $p_\eta(\cdot)$ and $q_\gamma(\cdot)$ based on a sample of size n_p and n_q , respectively.

3. Bayesian DRE

Consider the loss function $\ell(\theta, \delta(t))$ and its average under long-term (repeated) use of $\delta(t)$. In estimating θ it is called frequentist risk and given by

$$R(\theta, \delta) = E^{T|\theta}[\ell(\theta, \delta)] \quad (8)$$

$$= \int_t \ell(\theta, \delta(t)) P_\theta(dt), \quad (9)$$

where $\mathbb{E}^{T|\theta}(\cdot)$ is the expectation with respect to the arbitrary measurable cumulative density function (CDF) $T \sim P(t | \theta)$. Given any prior distribution $\pi(\cdot)$, with the CDF $\Pi(\cdot)$, one can define the *integrated risk (Bayes risk)*, which is the frequentist risk averaged over the values of θ according to prior distribution $\pi(\theta)$ and posterior distribution $\pi(\theta | t)$.

$$\mathbb{E}^{\theta|t}[R(\delta(t), \theta)] = \int_{\Theta} \int_{\mathcal{T}} \ell(\theta, \delta(t)) P_\theta(dt) \Pi(d\theta) dt. \quad (10)$$

Finally, the Bayes estimator is the minimizer of the Bayes risk (10). It can be shown that the minimizer of the above expression also minimizes the posterior risk function $\mathbb{E}^{\theta|t}[\ell(\theta, \delta(t))]$, and hence is the Bayes estimator (see Lehman and Casella [17]).

Next, we will address the log-Huber loss and reason beyond choosing such a loss function, in order to study the efficiency of the DRE.

3.1. log-Huber's Robust Loss

One of the weakness of a squared error loss function z^2 , with $z = \delta - \theta$, is that it can over-emphasize outliers. Other loss functions can be used to avoid this issue. For instance, one can use the absolute loss function $|z|$, but since it is not a differentiable function at the origin, Huber [18] proposed the following function instead:

$$h(z) = \begin{cases} z^2 & |z| \leq 1, \\ 2|z| - 1 & |z| \geq 1, \end{cases} \quad (11)$$

which is less sensitive to outliers and more robust and equivalent to the squared error loss (L_2 loss) for errors that are smaller than 1, and absolute error loss (L_1 loss) for larger errors (see Figure 1). This loss borrows the advantages of L_1 and L_2 losses, and does not have their disadvantage. That is, it is not sensitive to outliers (as opposed to L_2) and it is differentiable everywhere (as opposed to L_1). In practice, optimizing the Huber loss (and consequently, log-Huber) is much faster because it enables us to use standard smooth optimization methods (e.g., quasi-Newton) instead of linear programming. Thanks to its robustness, the Huber loss is widely used in robust regression problems (e.g., Huber and Ronchetti [19], Chapter 7).

In the context of DRE, we propose taking $z = \log \frac{\hat{r}(t)}{r(t, \theta)}$ in (11), which yields the log-Huber loss function $\ell(\hat{r}, r) = (\log \frac{\hat{r}}{r})^2$, for $e^{-1} \leq \frac{\hat{r}}{r} \leq e$ (and $\ell(\hat{r}, r) = |\log \frac{\hat{r}}{r}|$, otherwise).

The corresponding frequentist risk function to the log-Huber loss function is given by

$$R(\hat{r}, \theta) = \begin{cases} \iint (\log \frac{\hat{r}}{r})^2 dP(x) dP(y) & e^{-1} \leq \frac{\hat{r}}{r} \leq e, \\ 2 \iint (|\log \frac{\hat{r}}{r}| - 1) dP(x) dP(y) & 0 < \frac{\hat{r}}{r} \leq e^{-1}, \frac{\hat{r}}{r} \geq e, \end{cases} \quad (12)$$

where $\theta = (\eta, \gamma)$.

Figure 1. Comparison of Huber, L_2 (least square), and L_1 (absolute error) loss functions in (11).

3.1.1. Log- L_2 Loss Function

The log- L_2 loss $(\log z)^2$, with $z = \frac{\hat{r}}{r}$, puts a small weight whenever the density ratio estimator $\hat{r}$ and the true density ratio r are close, and puts proportionally more weight when they are significantly different. Lemma 2 provides the Bayesian DRE of the density ratio r in (3). An interesting representation of Bayesian DRE is the Bayesian DRE that can be expressed in terms of the plug-in DRE in (7).

Theorem 1. For any PDF p and q belonging to (1) and (2), respectively, the Bayesian DRE of under log- L_2 loss $L(\hat{r}, r) = (\log \frac{\hat{r}}{r})^2$, and prior distribution $\pi(\theta)$, for $\theta = (\eta, \gamma)$, is given by

$$\hat{r}_\pi(t) = \hat{r}_{plug}(t) H(t), \quad (13)$$

where $\hat{r}_{plug}$ is obtained by replacing the posterior expectations of unknown parameters for given t :

$$\hat{r}_{plug}(t) = \frac{\exp\{\langle s(t), \mathbb{E}^{\eta|X=t,Y=t} \eta \rangle - \log \mathbb{E}^{\eta|X=t,Y=t} \log c(\eta)\}}{\exp\{\langle s(t), \mathbb{E}^{\gamma|X=t,Y=t} \gamma \rangle - \log \mathbb{E}^{\gamma|X=t,Y=t} \log c(\gamma)\}} \quad (14)$$

the correction factor is given by

$$H(t) = \frac{\exp\{\log \mathbb{E}^{\eta|X=t,Y=t} c(\eta) - \mathbb{E}^{\eta|X=t,Y=t} \log c(\eta)\}}{\exp\{\log \mathbb{E}^{\gamma|X=t,Y=t} c(\gamma) - \mathbb{E}^{\gamma|X=t,Y=t} \log c(\gamma)\}} \quad (15)$$

Proof. The Bayesian DRE $\hat{r}_{\pi}(t)$ is the minimizer of the posterior risk and

$$\frac{\partial}{\partial \log \hat{r}_{\pi}} \mathbb{E}^{\eta|X=t,Y=t} (\log \hat{r}_{\pi} - \log r)^2 = 0$$

implies $\log \hat{r}_{\pi}(t) = \mathbb{E}^{\eta|X=t,Y=t} \log r_{\pi}$, or equivalently,

$$\hat{r}_{\pi}(t) = \exp \left\{ \mathbb{E}^{\eta|X=t,Y=t} \log r(t) \right\},$$

with $\frac{\partial^2}{\partial \log \hat{r}_{\pi}^2} \mathbb{E}^{\eta|X=t,Y=t} (\log \hat{r} - \log r)^2 \geq 0$. Therefore

$$\begin{aligned} \hat{r}_{\pi}(t) &= \exp \left\{ \mathbb{E}^{\eta|X=t} \log p_{\eta}(t) - \mathbb{E}^{\gamma|Y=t} \log q_{\gamma}(t) \right\} \\ &= \frac{\exp \left\{ \mathbb{E}^{\eta|X=t} \log p_{\eta}(t) \right\}}{\exp \left\{ \mathbb{E}^{\gamma|Y=t} \log q_{\gamma}(t) \right\}} \\ &= \frac{h_1(t) \exp \left\{ \hat{\eta}^{\top}(x) s_1(t) - \log c_1(\hat{\eta}(t)) \right\}}{h_2(t) \exp \left\{ \hat{\gamma}^{\top}(t) s_2(t) - \log c_2(\hat{\gamma}(t)) \right\}} \\ &\quad \times \frac{\exp \left\{ c_1(\hat{\eta}_1(t)) - \log \widehat{c_1(\eta_1)}(t) \right\}}{\exp \left\{ c_2(\hat{\eta}_2(t)) - \log \widehat{c_2(\eta_2)}(t) \right\}} \\ &= \hat{r}_{plug}(t) H(t). \end{aligned}$$

This completes the proof. $\square$

In Table 1 we explore the correction factor $H(\cdot)$ in (15) for certain densities which belong to the exponential family. Let the likelihood functions below, with sub indices $i = 1, 2$ representing the underlying distributions, be drawn from p and q according to (1) and (2), respectively. Note that posing constraints on the hyper-parameters leads to having $H(t)$ as a constant function in t . In fact, $H(t) = 1$ induces the Bayesian DRE $\hat{r}_{\pi}$ to coincide with the plug-in DRE $\hat{r}_{plug}$.

Note that the notation $\psi(\alpha)$ below is for a “digamma” function, and is given by $\frac{\partial \Gamma(\delta)/\partial \delta}{\Gamma(\delta)}$. Also, in Tables 1 and 2, *Gam*, *P*, *Pa*, *W*, *Bin*, *Bet*, χ^2 , *Ray*, *U*, and *Ge* stand for gamma, Poisson, Pareto, Weibull, binomial, beta, chi-squared, Rayleigh, uniform, and geometric distributions, respectively.

Table 1. Correction factor $H(t)$ and conditions when $H(t) = 1$ associated with log- L_2 loss. *Bet* beta distribution; *Bin*: binary distribution; *Gam*: gamma distribution; *Ge*: geometric distribution; *IG*: inverse gamma conjugate prior distribution; *P*: Poisson distribution; *Pa*: Pareto distribution; *W*: Weibull distribution.

Density	Conjugate Prior	Correction Factor $H(t)$	Condition for $H = 1$
$Gam(\alpha_i, \beta_i)$	$\beta_i \sim Gam(a_i, b_i)$	$(\alpha_1 + a_1)^{\alpha_1} (\alpha_2 + a_2)^{\alpha_2} \exp \{-\alpha_1 \psi(\alpha_1 + a_1) - \alpha_2 \psi(\alpha_2 + a_2)\}$	$\alpha_1 = \alpha_2, a_1 = a_2$
$N(\theta_i, \sigma_i^2)$	$\theta_i \sim N(\mu_i, \tau_i^2)$	$\exp \{1/2(\frac{1}{b_1(1+b_1)} - \frac{1}{b_2(1+b_2)})\}$	$\frac{b_1}{b_2} = \frac{1+b_1}{1+b_2}$, where $b_i = \sigma_i^2 / \tau_i^2$
$N(\theta_i, \sigma_i^2)$	$\sigma_i^2 \sim IG(\alpha_i, \beta_i)$	$\sqrt{\frac{\alpha_1-1/2}{\alpha_2-1/2}} \exp \{1/2(\psi(\alpha_2 + 1/2) - \psi(\alpha_1 + 1/2))\}$	$\alpha_1 = \alpha_2$
$Pa(a_i, b_i)$	$b_i \sim Gam(\alpha_i, \beta_i)$	$\frac{\alpha_1+1}{\alpha_2+1} \exp \{\frac{\alpha_1-\alpha_2}{\alpha_1\alpha_2} + \psi(\alpha_2) - \psi(\alpha_1)\}$	$\alpha_1 = \alpha_2$
$W(\lambda_i, k_i)$	$\lambda_i \sim IG(\alpha_i, \beta_i)$	$\alpha_1^{k_1} / \alpha_2^{k_2} \exp \{k_2 \psi(\alpha_2 + 1) - k_1 \psi(\alpha_1 + 1)\}$	$k_1 = k_2, \alpha_1 = \alpha_2$
$Bin(n_i, p_i)$	$p_i \sim Bet(\alpha_i, \beta_i)$	$\exp \{\psi(\beta_2 + n - t) - \psi(\beta_1 + n_1 - t) + \psi(\alpha_1 + \beta_1 + n_1) - \psi(\alpha_2 + \beta_2 + n_2)\}$	$\alpha_1 = \alpha_2, \beta_1 = \beta_2, n_1 = n_2$
$P(\lambda_i)$	$\lambda_i \sim Gam(\alpha_i, \beta_i)$	1	$\forall \alpha_i, \beta_i$, and λ_i
$Ge(p_i)$	$p_i \sim Bet(\alpha_i, \beta_i)$	$\exp \{\psi(\beta_1 + t) - \psi(\alpha_1 + 1) + \psi(\beta_2 + t) - \psi(\alpha_2 + 1)\} \frac{\beta_2+t}{\beta_1+t} \frac{\alpha_1+1}{\alpha_2+1}$	$\alpha_1 = \alpha_2, \beta_1 = \beta_2$

Table 2. KL divergence between p and q . χ^2 : chi-squared distribution; U : uniform distribution; Ray : Rayleigh distribution.

Density	KL
$N_p(\theta_i, \Sigma_i)$	$\frac{1}{2} [(\theta_1 - \theta_2)' \Sigma_2^{-1} (\theta_1 - \theta_2) + tr(\Sigma_2^{-1} \Sigma_1) - p - \log \frac{det(\Sigma_1)}{det(\Sigma_2)}]$
$N(\theta_i, \sigma_i^2)$	$\frac{1}{2\sigma_2^2} [(\theta_1 - \theta_2)^2 + \sigma_1^2 - \sigma_2^2] - \frac{1}{2} \log \frac{\sigma_1^2}{\sigma_2^2}$
$Bet(a_i, b_i)$	$\log \frac{Bet(a_2, b_2)}{Bet(a_1, b_1)} + \psi(a_1)(a_1 - a_2) + \psi(b_1)(b_1 - b_2) + \psi(a_1 + b_1)(a_2 - a_1 + b_2 - b_1)$
$Gam(k_i, \lambda_i)$	$(\frac{\lambda_1}{\lambda_2} - 1)k_1 + (k_1 - k_2)(\log \lambda_1 + \psi(k_1)) - \log \frac{\Gamma(k_1)\lambda_1^{k_1}}{\Gamma(k_2)\lambda_2^{k_2}}$
$\chi^2(k_i)$	$\log \frac{\Gamma(\frac{k_2}{2})}{\Gamma(\frac{k_1}{2})} + \frac{1}{2} \psi(\frac{k_2}{2})(k_1 - k_2)$
$\log-N(\theta_i, \sigma_i^2)$	$\frac{1}{2\sigma_2^2} [(\theta_1 - \theta_2)^2 + \sigma_1^2 - \sigma_2^2] - \log \frac{\sigma_1^2}{\sigma_2^2}$
$Ray(\sigma_i)$	$2 \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 - \sigma_2^2}{\sigma_2^2}$
$Pa(m_i, \alpha_i)$	$\log(\frac{m_1}{m_2})^2 - \log \frac{a_2}{a_1} + \frac{a_2}{a_1} - 1$
$U(0, \theta_i)$	$\log(\frac{\theta_2}{\theta_1})$ Provided $\theta_2 > \theta_1$

Definition 1. In the exponential distribution family in (1), Bregman divergence associated with a real-valued strictly convex and differentiable function $c(\cdot)$ is defined by

$$B(\eta, \gamma) = \kappa(\eta) - \kappa(\gamma) - \langle \eta - \gamma, \nabla c(\gamma) \rangle, \quad (16)$$

where $\langle \cdot, \cdot \rangle$ is the inner product and it can be shown that for all regular exponential families (see Brown [20]), $\kappa(\cdot)$ is strictly convex, and furthermore $B(\eta, \gamma) = \log \frac{c(\eta)}{c(\gamma)} - \langle \eta - \gamma, \mathbb{E}^q(s(T)) \rangle$.

Lemma 1. The Kullback–Leibler (KL) divergence between p_η and q_γ in model (1) is equal to the Bregman divergence between natural parameters. That is:

$$KL(p(\cdot|\eta), q(\cdot|\gamma)) = B(\eta, \gamma),$$

where the KL divergence (Kullback–Leibler, [21]), between densities f_1 and f_2 , $KL(f_1, f_2) = \mathbb{E}^{f_1} \log f_1 / f_2$.

Proof. See Nielsen and Nock [22]. $\square$

Lemma 2. Consider $\mathbb{E}^p (\log \hat{r}(t) - \log r(t))^2$ as a loss function where the random variable T follows from p as in (1) and the prior function $\pi(\theta)$ on $\theta = (\eta, \gamma)$, then the Bayesian DRE of $r(t; \theta)$ is given by

$$\hat{r}_\pi(t) = \exp \{ \mathbb{E} (KL(p, q) \mid X = t, Y = t) \} , \quad (17)$$

and in addition for the natural exponential family model (1) it can also be expressed as

$$\hat{r}_\pi(t) = \exp \{ \mathbb{E} (B(\eta, \gamma) \mid X = t, Y = t) \} . \quad (18)$$

Proof. For simplicity, assume notations $\hat{r}_\pi(t)$ and $r(t; \eta, \gamma)$ as $\hat{r}$ and r respectively.

The Bayesian estimator $\hat{r}$ in estimating r is the minimizer of $\mathbb{E}(\ell(\hat{r}, r) \mid X = t, Y = t)$, with $\ell(\hat{r}, r) = \mathbb{E}^p(\log \frac{\hat{r}}{r})^2$, and therefore $\frac{\partial}{\partial \log \hat{r}} \mathbb{E}(\mathbb{E}^p(\log \hat{r} - \log r)^2 \mid X = t, Y = t) = 0$ implies $\log \hat{r}_\pi(t) = \mathbb{E}(\mathbb{E}^p \log r \mid X = t, Y = t) = \mathbb{E}(KL(p, q) \mid X = t, Y = t)$. Applying Lemma 1 and the fact that $\frac{\partial^2}{\partial \log \hat{r}^2} \mathbb{E}(\mathbb{E}^p(\log \hat{r} - \log r)^2 \mid X = t, Y = t) \geq 0$, completes the proof. $\square$

3.1.2. Log- L_1 Loss Function

For some larger errors in loss function (12), one needs to consider the log- L_1 loss function $\ell(\tilde{r}, r) = 2|\log \frac{\tilde{r}}{r}| - 2$. Let $\tilde{r}$ be the corresponding Bayesian DRE. As in Theorem 1 we can also express $\tilde{r}_\pi(t)$ in terms of the product of the correction factor and the plug-in DRE. That is, under log- L_1 loss $L(\tilde{r}, r) = |\log \frac{\tilde{r}}{r}|$, and prior distribution $\pi(\theta)$, for $\theta = (\eta, \gamma)$, we have

$$\tilde{r}_\pi(t) = \tilde{r}_{plug}(t) H'(t) , \quad (19)$$

where $\tilde{r}_{plug}(t)$ and $H'(t)$ are obtained in the same fashion in (14) and (15) except applying median $\mathbb{M}(\cdot \mid X = t, Y = t)$ instead of $\mathbb{E}(\cdot \mid X = t, Y = t)$. Notice that for instance the results in Table 1 hold wherever the posterior densities turn out to be symmetric about their means (or medians).

Similar calculation to Lemma 2, with loss function $\mathbb{E}^p |\log \hat{r}(t) - \log r(t)|$, suggests

$$\tilde{r}_\pi(t) = \exp \{ \mathbb{M}(B(\eta, \gamma) \mid X = t, Y = t) \} .$$

Similar to (17) and (18), we can write $\tilde{r}_\pi(t) = \exp \{ \mathbb{M}(KL(p, q) \mid X = t, Y = t) \}$, whose expectations are replaced by medians.

3.2. Examples of the Bayesian DRE and Some Applications

For instance, we consider the problem of finding Bayesian DRE under log- L_1 and log- L_2 losses for two different normal densities.

Example 1. Let $X|\theta_1 \sim \mathbb{N}(\theta_1, \sigma_1^2)$ be independent of $Y|\theta_2 \sim \mathbb{N}(\theta_2, \sigma_2^2)$, with the known variances and independent prior distributions $\theta_i \sim \mathbb{N}(\xi_i, \tau_i^2)$ for $i = 1, 2$. Hence, the posterior densities are given by

$$\theta_1 \mid X = t \sim \mathbb{N} \left(\frac{\xi_1 \tau_1^2 + \tau_1^2 t}{\tau_1^2 + \sigma_1^2}, \frac{\tau_1^2 \sigma_1^2}{\tau_1^2 + \sigma_1^2} \right) ,$$

$$\theta_2 \mid Y = t \sim \mathbb{N} \left(\frac{\xi_2 \tau_2^2 + \tau_2^2 t}{\tau_2^2 + \sigma_2^2}, \frac{\tau_2^2 \sigma_2^2}{\tau_2^2 + \sigma_2^2} \right) ,$$

yielding the Bayesian DRE under either loss functions $L(\hat{r}, r) = (\log \frac{\hat{r}}{r})^2$ or $L(\hat{r}, r) = |\log \frac{\hat{r}}{r}|^2$, as below:

$$\tilde{r}_\pi(t) = \hat{r}_\pi(t) = \frac{\sigma_2}{\sigma_1} \exp \left(\frac{\sigma_1^2 - \sigma_1^2 + \left(\frac{\xi_1 \sigma_1^2 + t \tau_1^2}{\sigma_1^2 + \tau_1^2} - \frac{\xi_2 \sigma_2^2 + t \tau_2^2}{\sigma_2^2 + \tau_2^2} \right)^2}{2\sigma_2^2} \right). \quad (20)$$

If we assume the multivariate case, that is, $X|\theta_1 \sim \mathbb{N}_p(\theta_1, \Sigma_1)$ and $Y|\theta_2 \sim \mathbb{N}_p(\theta_2, \Sigma_2)$, where both covariance matrices are known, since the conjugate prior for the mean is p -variate normal, $\theta_i \sim \mathbb{N}_p(\xi_i, V_i)$ for $i = 1, 2$, hence the posterior is p -variate normal. The results are analogous to the univariate case and $\theta_1|t \sim \mathbb{N}_p(W_1(\Sigma_1^{-1}t + V_1^{-1}\xi_1), W_1)$ and $\theta_2|t \sim \mathbb{N}_p(W_2(\Sigma_2^{-1}t + V_2^{-1}\xi_2), W_2)$ with $W_i = (\Sigma_i^{-1} + V_i^{-1})^{-1}$. Therefore the Bayesian DRE for the ratio of two multivariate normal densities equals to

$$\frac{1}{2} \left[(W_1(\Sigma_1^{-1}t + V_1^{-1}\xi_1) - W_2(\Sigma_2^{-1}t + V_2^{-1}\xi_2))' \Sigma_2^{-1} (W_1(\Sigma_1^{-1}t + V_1^{-1}\xi_1) - W_2(\Sigma_2^{-1}t + V_2^{-1}\xi_2)) + \text{tr}(\Sigma_2^{-1}\Sigma_1) - \frac{1}{2} \log \frac{\det(\Sigma_1)}{\det(\Sigma_2)} \right].$$

As we saw in the previous section, the key point is to find the KL divergence between two densities p and q (or equivalently, Bregman divergence in the cases of distributions belonging to exponential family), and we have a closed form for the divergence. Table 2 represents the KL divergence between p and q .

Example 2. Let $X|\sigma_1^2 \sim \text{Ray}(\sigma_1^2)$ and $Y|\sigma_2^2 \sim \text{Ray}(\sigma_2^2)$ be two independent random variables from a Rayleigh distribution with the PDF $p(x|\sigma^2) = \frac{x}{\sigma^2} \exp(-\frac{x^2}{2\sigma^2})$ for $x > 0$ and $\sigma^2 > 0$. The KL divergence between $p(x|\sigma_1^2)$ and $p(y|\sigma_2^2)$ is given in Table 2. Assuming $\lambda_i = 2\sigma_i^2$, and choosing the inverse gamma conjugate prior distribution (IG) with parameters α and β , with the PDF $\pi(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{-(\alpha+1)} \exp(-\frac{\beta}{\lambda})$ for $\lambda > 0$, yields the posterior distribution as follows:

$$\lambda_i | t \sim \text{IG}(1 + \alpha_i, \beta_i + t^2). \quad (21)$$

Therefore, the Bayesian DRE for the ratio of two Rayleigh densities is the exponential function of the conditional expectation of their KL loss divergence function. That is,

$$\exp \left(\mathbb{E}^{\lambda_2|t} \lambda_2 - \mathbb{E}^{\lambda_1|t} \log \lambda_1 + \mathbb{E}^{\lambda_1|t} \lambda_1 \mathbb{E}^{\lambda_2|t} \lambda_2^{-1} - 1 \right).$$

By making use of properties of $\lambda \sim \text{IG}(\alpha, \beta)$, such as $\mathbb{E}(\log \lambda) = \log \beta - \psi(\alpha)$, where $\psi(\alpha)$ is a “digamma” function, the Bayesian DRE under loss function $\mathbb{E}^p (\log \hat{r}(t) - \log r(t))^2$ is given by

$$\hat{r}_\pi(t) = \frac{\beta_2 + t^2}{\beta_1 + t^2} \exp \left(\frac{\alpha_2 - \alpha_1}{\alpha_1 \alpha_2} + \frac{\alpha_1}{1 + \alpha_2} \frac{\beta_1 + t^2}{\beta_2 + t^2} - \psi(\alpha_2) + \psi(\alpha_1) - 1 \right).$$

Since the median of an inverse gamma distribution does not have a closed form, we cannot express an explicit formula for the Bayesian DRE under loss function $\mathbb{E}^p |\log \hat{r}(t) - \log r(t)|$, and consequently for the log-Huber loss function, and they must be calculated iteratively.

The following remarks are some clarifications related to Table 2 and the obtained Bayes estimators in the previous examples in Section 3.2.

Remark 1. The Bayes DRE $\hat{r}_\pi(t)$ is connected to samples X and Y via the posterior density $\eta | X = t$ and $\gamma | Y = t$.

Remark 2. From Table 2 it can be seen that the KL divergence between two log-normal PDFs is the same as in normal distributions, since it is known that KL is invariant under parameter transformations.

Example 3. Suppose $X \sim \text{Uniform}(0, \theta_1)$ is independent of $Y \sim \text{Uniform}(0, \theta_2)$, with $\theta_1 < \theta_2$. It is easy to check that the Pareto distribution is conjugate to a uniform. Hence, assume that $\pi(\theta_i) = \alpha_i \beta_i^{\alpha_i} \theta_i^{-(\alpha_i+1)}$ for $\theta \geq \beta_i$ and $i = 1, 2$. Therefore,

$$\theta_i | t \sim \text{Pareto}(1 + \alpha_i, \max(\beta_i, t)).$$

Moreover, assuming p and q are the PDFs of X and Y , respectively, $\text{KL}(p, q) = \log \theta_2 - \log \theta_1$ (see Table 2). Employing the fact that the transformation $\log(z/b)$ has an exponential distribution with mean of a , given $Z \sim \text{Pareto}(a, b)$, and hence $\mathbb{E}^Z \log Z = a + \log b$, after some calculation we have the Bayes DRE associated with loss function $\mathbb{E}^p (\log \hat{r}(t) - \log r(t))^2$:

$$\hat{r}_\pi(t) = \frac{\max(\beta_2, t)}{\max(\beta_1, t)} + \alpha_2 - \alpha_1.$$

Similar to Example 2, the Bayes DRE under $\mathbb{E}^p |\log \hat{r}(t) - \log r(t)|$ must be calculated numerically.

4. Other Applications

Here, we discuss some of other applications of the DRE method.

1 Estimating the α -divergence function between two probability densities:

A discrepancy measure between densities p_η and q_γ applicable to the class of Ali-Silvey [23] distance, also known as α -divergence (Csiszàr, [24]), is given by

$$\ell_\alpha(p(\cdot | \eta), q(\cdot | \gamma)) = \int_{\mathbb{R}^d} h_\alpha \left(\frac{p(t | \eta)}{q(t | \gamma)} \right) dP(t | \gamma), \quad (22)$$

where

$$h_\alpha(z) = \begin{cases} \frac{4}{1-\alpha^2} (1 - z^{-(1+\alpha)/2}) & \text{for } |\alpha| \leq 1, \\ -\log(z)/z & \text{for } \alpha = -1, \\ \log(z) & \text{for } \alpha = 1. \end{cases}$$

Note that some of the notable divergence functions, such as Kullback–Leibler, reverse Kullback–Leibler (RKL), and Hellinger divergence functions correspond to $\alpha = 1, -1$, and 0 , respectively, and belong to this class. So, if $r(t)$ is estimated by $\hat{r}_\pi(t)$ under log- L_2 (or $\tilde{r}_\pi(t)$ under log- L_1 losses), then applying the Monte Carlo approximations method, the α -divergence is also estimated by

$$d_\alpha(t) = 1/n \sum_{i=1}^n h_\alpha(\hat{r}_\pi(t_i)), \quad (23)$$

where t_i are drawn from $p(\cdot | \eta)$. Note that there are other works in order to estimate the α -divergence loss function (e.g., Póczos and Schneider [25]), but our estimator in (23) is based on the Bayesian parametric method based on the DRE. In the next section we show the performance of the proposal estimator d_α .

2 Plug-in-type estimation of the density ratio under KL loss function:

Consider the plug-in density estimator, say $p_{\hat{\eta}}(t)$, for estimating p_{η} in the exponential family (1), based on the KL loss. We have

$$\begin{aligned} KL(p(t|\eta), p(t|\hat{\eta})) &= \int \log \frac{p(\eta|t)}{p(\hat{\eta}|t)} P(t|\eta) dt \\ &= \int p(t|\hat{\eta}) \log \frac{\exp\{\langle \eta, s(t) \rangle - c(\eta)\}}{\exp\{\langle \hat{\eta}(t), s(t) \rangle - c(\hat{\eta}(t))\}} dt \\ &= \langle \eta - \hat{\eta}(t), \mathbb{E}^p s(T) \rangle - (c(\eta) - c(\hat{\eta}(t))) \\ &= \langle \eta - \hat{\eta}(t), \frac{\partial c(\eta)}{\partial \eta} \rangle - (c(\eta) - c(\hat{\eta}(t))). \end{aligned} \quad (24)$$

Next, finding the plug-in density estimator $p(t, \hat{\eta}(t))$ that minimizes KL loss is equivalent to finding the point estimator $\hat{\eta}(t)$ which minimizes the posterior expectation associated with the loss function in (24). Therefore, $\frac{\partial}{\partial \hat{\eta}} \mathbb{E}^{\eta|t} KL(p(t|\eta), p(t|\hat{\eta})) = 0$ implies $c_1(\hat{\eta}) = \mathbb{E}^{\eta|t} \frac{c(\eta)}{\eta}$, and hence the Bayes estimator of η is given by

$$\hat{\eta}_1(t) = c_1^{-1} \left(\mathbb{E}^{\eta|t} \frac{c_1(\eta)}{\eta} \right). \quad (25)$$

Similar arguments can be applied to $q_{\hat{\gamma}}(t)$ for estimating q_{γ} in the exponential family (2), and we have:

$$\hat{\gamma}_2(y) = c_2^{-1} \left(\mathbb{E}^{\gamma|t} \frac{c_2(\gamma)}{\gamma} \right). \quad (26)$$

By setting the case when both densities follow the identical distribution from (1) (e.g., the ratio of two normal or two Poisson, etc.), substituting the Bayes estimators obtained in (25) and (26) into the plug-in estimator of $r(t)$ gives

$$\hat{r}(t) = \exp\{(\hat{\eta}_1(t) - \hat{\eta}_2(t))^{\top} s(t) - (c_1(\hat{\eta}_1(t)) - c_2(\hat{\eta}_2(t)))\}. \quad (27)$$

Note that $\tilde{r}(t)$ can be obtained similarly by replacing posterior medians $\tilde{\eta}_i$ instead of posterior expectations $\hat{\eta}_i$ for $i = 1, 2$ above.

5. Numerical Illustrations

We conclude this section with some numerical illustrations of log-Huber risk performance of the Bayesian and plug-in when both p and q are two normal models (belong to model (1) and (2)) with a common location parameter θ . We show that the performance of plug-in and Bayes are quite similar and by selecting the hyper-parameters these two density ratio estimators coincide and hence have the same frequentist risk. We start with comparing risk performance under the log- L_2 first and then extend them to log-Huber loss.

Figure 2 exhibits the frequentist risk performance of the plug-in DRE $\hat{r}_{plug}$ and the Bayes DRE $\hat{r}_{\pi}$ under log- L_2 for all possible values of θ using Theorem 1.

Figure 2. Risk function of the Bayes and plug-in density ratio estimation (DRE) under L_2 loss in normal models $p = \mathcal{N}(\theta, 1)$, $q = \mathcal{N}(\theta, 2)$ and corresponding hyper-parameters $\xi = 1$, $\xi = 2 = 0$, $\tau_1 = \tau_2 = 1$ corresponding to Example 1.

Figure 3 shows the changing frequentist risk performance of the plug-in DRE $\hat{r}_{plug}$ and the Bayes DRE $\hat{r}_\pi$ under log- L_2 when the ratio of variances σ_1^2/σ_2^2 varies (means are known) using Theorem 1. It can be seen the performance of plug-in estimation is better for larger ratios of variances.

Finally, Figures 4 and 5 depict the risk functions of the Bayes and the plug-in DRE under the log-Huber loss over possible ranges of the mean and variance, respectively, in Example 1. It is clear from Figure 5, in contrast to Figure 3 (risk functions of Bayes DRE under log- L_2 and plug-in DRE estimator), that the Bayes estimator DRE has a lower risk with respect to the plug-in DRE, and hence is more robust under the variation of variances.

Figure 3. Risk function of the Bayes and plug-in DRE under log- L_2 loss in normal models $p = \mathcal{N}(1, \sigma_1^2)$, $q = \mathcal{N}(1, \sigma_2^2)$ and hyper-parameters $\xi = 1$, $\xi = 2 = 0$, $\tau_1 = \tau_2 = 1$ in Example 1.

Figure 4. Risk function of the Bayes and plug-in DRE under log-Huber loss in normal models $p = \mathcal{N}(\theta, 1)$, $q = \mathcal{N}(\theta, 2)$ and corresponding hyper-parameters $\xi = 1$, $\xi = 2 = 0$, $\tau_1 = \tau_2 = 1$ for the whole possible range of θ in Example 1.

Figure 5. Risk function of the Bayes and plug-in DRE under log-Huber loss in normal models $p = \mathcal{N}(1, \sigma_1^2)$, $q = \mathcal{N}(1, \sigma_2^2)$ and corresponding hyper-parameters $\xi = 1$, $\xi = 2 = 0$, $\tau_1 = \tau_2 = 1$, for the whole possible range of σ^2 in Example 1.

6. Concluding Remarks

Estimation of the ratio of two or more densities has received widespread attention in recent years. Until now, most of the methods have been concentrated on solving this problem via nonparametric approaches. In this paper, we focused on a parametric Bayesian approach, when distributions come from the canonical form of the exponential family. We applied the log-Huber loss function to investigate the utility of the Bayesian and plug-in DREs. Our results confirm that Bayesian DRE along with the plug-in DRE (based on posterior expectations) perform similarly under log-Huber loss functions with the possibility of being exactly equal when the correction factor $H = 1$. This is a somewhat different result from a non-parametric point of view, which often happens when the plug-in estimators perform poorly as opposed to empirical non-parametric Bayesian methods which typically include stochastic processes such as the Gaussian process and the Dirichlet process. There are instances (see Krnjajic et al. [26] and the references therein) where for certain types of count data, the nonparametric Bayesian methodology provides enhanced flexibility to fit the data, provides rich posterior inferences, and provides finer predictive inference under a set of carefully selected criteria. However, there is an apparent major drawback. These processes have an infinite number of dimensions, and thus naive algorithmic approaches to compute posteriors are generally infeasible. Finally, the application to estimate the α -divergence between two PDFs was discussed.

Author Contributions: Methodology, calculations and writing the draft, A.S.; adding critical and constructive comments as well as editorial advice, Y.P. and C.D.L.

Funding: This research received no external funding.

Acknowledgments: The authors thank Eric Marchand (Université de Sherbrooke) for his useful suggestions. We also want to thank the referees for their valuable comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sugiyama, M.; Suzuki, T.; Kanamori, T. *Density Ratio Estimation in Machine Learning*; Cambridge University Press: Cambridge, UK, 2012.
2. Kanamori, T.; Hido, S.; Sugiyama, M. A Least-squares Approach to Direct Importance Estimation *J. Mach. Learn. Res.* **2009**, *10*, 1391–1445.
3. Sugiyama, M.; Kanamori, T.; Suzuki, T.; Hido, S.; Sese, J.; Takeuchi, I.; Wang, L. A density-ratio framework for statistical data processing. *IPSJ Trans. Comput. Vis. Appl.* **2009**, *1*, 183–208. [[CrossRef](#)]
4. Sugiyama, M.; Yamada, M.; Bunau, P.V.; Suzuki, T.; Kanamori, T.; Kawanabe, M. Direct density-ratio estimation with dimensionality reduction via least-squares hetero-distributional subspace search. *Neural Netw.* **2011**, *24*, 183–198. [[CrossRef](#)] [[PubMed](#)]
5. Sugiyama, M.; Müller, K.R. Input-dependent estimation of generalization error under covariate shift. *Stat. Decis.* **2005**, *23*, 249–279. [[CrossRef](#)]

6. Quiñonero-Candela, J.; Sugiyama, M.; Schwaighofer, A.; Lawrence, N. *Dataset Shift in Machine Learning*; MIT Press: Cambridge, MA, USA, 2009.
7. Suzuki, T.; Sugiyama, M.; Kanamori, T.; Sese, J. Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinform.* **2009**, *10*, S52. [[CrossRef](#)] [[PubMed](#)]
8. Suzuki, T.; Sugiyama, M. Sufficient dimension reduction via squared-loss mutual information estimation. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Chia Laguna, Italy, 13–15 May 2010; pp. 804–811.
9. Sugiyama, M.; Hara, S. von Büna, P.; Suzuki, T.; Kanamori, T.; Kawanabe, M. Direct density ratio estimation with dimensionality reduction. In Proceedings of the SIAM International Conference on Data Mining, Columbus, OH, USA, 29 April–1 May 2010.
10. Hido, S.; Tsuboi, Y.; Kashima, H.; Sugiyama, M.; Kanamori, T. Inlier-Based Outlier Detection via Direct Density Ratio Estimation. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 223–232.
11. Thomas, O.; Dutta, R.; Corander, J.; Kaski, S.; Gutmann, M.U. Likelihood-free inference by ratio estimation. *arXiv* **2016**, arXiv:1611.10242.
12. Gretton, A.; Smola, A.; Huang, J.; Schmittfull, M.; Borgwardt, K.; Schölkopf, B. Covariate shift by kernel mean matching. *Dataset Shift Mach. Learn.* **2009**, *3*, 5.
13. Sugiyama, M.; Suzuki, T.; Nakajima, S.; Kashima, H.; von Bunau, P.; Kawanabe, M. Direct importance estimation for covariate shift adaptation. *Ann. Inst. Stat. Math.* **2008**, *60*, 699–746. [[CrossRef](#)]
14. Nguyen, X.; Wainwright, M.J.; Jordan, M.I. Estimating divergence functional and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theory* **2010**, *56*, 5847–5861. [[CrossRef](#)]
15. Deledalle, C.A. Estimation of Kullback-Leibler losses for noisy recovery problems within the exponential family. *Electron. J. Stat.* **2017**, *11*, 3141–3164. [[CrossRef](#)]
16. Sugiyama, M.; Suzuki, T.; Kanamori, T. Density-ratio matching under the Bregman divergence: A unified framework of density-ratio estimation. *Ann. Inst. Stat. Math.* **2012**, *64*, 1009–1044. [[CrossRef](#)]
17. Lehman, E.L.; Casella, G. *Theory of Point Estimation*, 2nd ed.; Springer Texts in Statistics; Springer: New York, NY, USA, 1998.
18. Huber, P. Robust estimation of a location parameter. *Ann. Math. Stat.* **1964**, *53*, 73–101. [[CrossRef](#)]
19. Huber, P.J.; Ronchetti, E.M. *Robust Statistics*; John Wiley & Sons: Hoboken, NJ, USA, 2009; Volume 2.
20. Brown, L.D. *Fundamentals of Exponential Families*; IMS: Hayward, CA, USA, 1986.
21. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [[CrossRef](#)]
22. Nielsen, F.; Nock, R. Entropies and cross-entropies of exponential families. In Proceedings of the 2010 IEEE International Conference on Image Processing, Hong Kong, China, 26–29 September 2010; pp. 3621–3624.
23. Ali, S.M.; Silvey, S.D. A general class of coefficients of divergence of one distribution from another. *J. R. Stat. Soc. Ser. B* **1966**, *28*, 131–142. [[CrossRef](#)]
24. Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Stud. Sci. Math. Hung.* **1967**, *2*, 299–318.
25. Póczos, B.; Schneider, J. On the estimation of alpha-divergences. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 11–13 April 2011; pp. 609–617.
26. Krnjajic, M.; Kottas, A.; Draper, D. Parametric and nonparametric Bayesian model specification: A case study involving models for count data. *Comput. Stat. Data Anal.* **2008**, *52*, 2110–2128. [[CrossRef](#)]