 doi:10.20944/preprints201903.0025.v1

Article

A Parametric Bayesian Approach in Density Ratio Estimation

Abdolnasser Sadeghkhani [1,*], Yingwei Peng [2] and C. Devon Lin [3]

[1] Department of Mathematics, Brock University, St. Catharines, ON, Canada; asadeghkhani@brocku.ca
[2] Departments of Public Health Sciences, Queen's University, ON, Canada; yingwei.peng@queensu.ca
[3] Department of Mathematics & Statistics, Queen's University, ON, Canada devon.lin@queensu.ca;
* Correspondence: asadeghkhani@brocku.ca;

Abstract: This paper considers 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 to find robust M-estimators. We propose two different types of Bayesian density ratio estimators and compare their performance in terms of Bayesian risk function with themselves as well as the usual plug–in density ratio estimators. 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. The density ratio estimation (DRE) is widely considered to be the most important factor in the machine learning and information theory. Sugiyama et al. in a series of papers (e.g. 2009, 2011) developed the DRE in different statistical data analysis problems. Some useful applications of the DRE are as follows: non–stationary adaptation (Sugiyama and Müller 2005 and Quiñonero-Candela et al., 2009), variable selection (Suzuki et al., 2009), dimension reduction (Suzuki and Sugiyama, 2010), conditional density estimation (Sugiyama et al., 2010), outlier detection (Hido et al., 2008) among others. This paper addresses cases when the density belongs to the parametric distributions (e.g., exponential family of distributions). Parametric methods are usually favourable thanks to existence of closed forms (mostly, simple and explicit formulas) and hence help to enhance computational efficiency.

Recently several estimators of the ratio of two densities have been proposed. One of the simplest approaches to estimate density ratio p/q, where p and q are two probability density (or mass) functions (PDF or PMF), is called "plug–in", which the ratio of the estimated densities is computed. Alternatively, one can estimate the ratio of two densities directly. Several approaches have been explored recently to estimate the ratio including moment matching approach (Gretton et al., 2009) and density matching approach (Sugiyama, 2008). There exists other work such as Nguyen (2010) and Deledalle (2017) which studied the application of DRE in estimating Kullback–Leibler (KL) divergence (or more generally α–divergence function, also known as f–divergence) and vice versa. The main objective of this paper is to address the 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 methodology of the DRE 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. Se. Some numerical illustrations for considering the efficiency of the proposed DRE's are given in Section 5. Finally, we make some concluding remarks in Section 6.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0025.v1

2. Density ratio estimation for exponential distribution family

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

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

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

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.

Consider the problem of estimating the density ratio

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

which obviously is 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

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

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 $N(\beta) = \int q(t) \exp\{\beta^\top s(t)\}\, dt$, that guarantees $\int q(t \mid \gamma) r(t;\theta)\, dt = 1$ and hence $q(t) r(t;\theta)$ becomes a valid PDF (PMF).

For instance, suppose two normal densities $p(t \mid \mu_1, \sigma_1^2) = \mathrm{N}(t \mid \mu_1, \sigma_1^2)$ and $q(t \mid \mu_2, \sigma_2^2) = \mathrm{N}(t \mid \mu_2, \sigma_2^2)$. They are corresponding 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 $\kappa(\eta) = \frac{\mu_1^2}{2\sigma_1^2} + \log \sigma_1^2$, $\gamma(\eta) = \frac{\mu_2^2}{2\sigma_2^2} + \log \sigma_2^2$, 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 probabilistic classification problem (Sugiyama et al. 2012 c) Suppose $Z = \{0, 1\}$ is a binary response variable which shows whether an observation t is drown from either $p(\cdot|\eta)$ (say $Z = 0$) or $q(\cdot|\gamma)$ (say $Z = 1$). In such a classification problem the logistic model has the form

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

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

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

Letting $\pi = \mathbb{P}(Z = 1) = \int \mathbb{P}(Z = 1 \mid t) p(t)\, dt$ and $p(t) = \mathbb{P}(t \mid Z = 0)$ and $q(t) = \mathbb{P}(t \mid Z = 1)$, we have

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0025.v1

and hence

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

which has indeed 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 $Z = \{0, 1, \ldots, m-1\}$, we can extend equation (5) to the multinomial logit model, and we have

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

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

In fact a conceptually simple solution to the DRE is to estimate separately each of two 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 in below

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

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 is called frequentist risk and given by

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

$$= \int_t \ell(\theta, \delta(t))\, P_\theta(dt), \tag{9}$$

where $\mathbb{E}^{T|\theta}(\cdot)$ is the expectation with respect to the arbitrary measurable cumulative density function (CDF) $T \sim P(t \mid \theta)$. Given any prior distribution π, it is also possible to 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 $\pi(\theta \mid t)$.

$$\mathbb{E}^{\theta|t}\left[R(\delta(t), \theta)\right] = \int_\Theta \int_\mathcal{T} \ell(\theta, \delta(t))\, \pi(\theta \mid t)\, dt\, d\theta. \tag{10}$$

Finally, the Bayes estimator is the minimizer of the Bayes risk (10). It can be shown that the minimizer of the above expression also minimize the posterior risk function and hence is the Bayes estimator (See Lehman and Casella 1998) .

Next, we will address the log–Huber loss and reasons 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 overly emphasize on outliers. As a result, other loss functions can be used to avoid this issue. For instance one

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0025.v1

can use the absolute loss function $|z|$, but since it is not a differentiable function at the origin, Huber (1964), proposed the following function instead

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

which is 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 advantages of L_1 and L_2 losses and does not have their disadvantage. That is, it is not sensitive to the outliers (as opposed to L_2) and it is it is everywhere differentiable (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 (such as quasi-Newton) instead of linear programming.
In the context of the DRE, we propose taking $z = \log \frac{\hat{r}(t)}{r(t,\theta)}$ in (11), yields the Log–Huber's loss function $\ell(\hat{r}, r) = \int (\log \frac{\hat{r}}{r})^2 \, dP(t)$

The corresponding frequentist risk function to the Log–Huber's loss function is given by

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

where $\theta = (\eta, \gamma)$ and hence the Bayes risk for the Bayesian DRE is given by $R_\pi(\hat{r}) = \mathbb{E}^{\theta|t} R(\hat{r}, \theta)$, which is a scalar and a function of t.

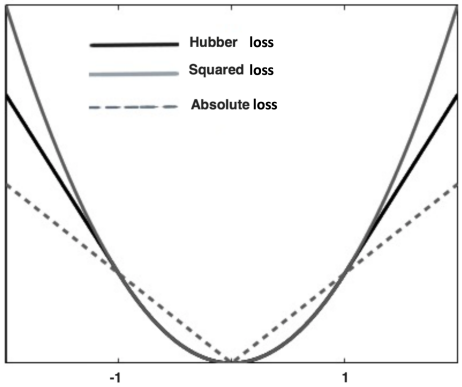

Figure 1. Comparison of Huber, L_2 (least square), and L_1 (absolute error) loss functions in (11)

Before discussing further on this topic, it is worthwhile to consider the specific cases of Log–L_2 and Log–L_1 respectively along with some useful definitions. Here are a definition and a lemma appeared in Nielsen and Nock (2010).

Definition 1. *Bregman divergence associated with a real valued strictly convex and differentiable function $c(\cdot)$, is defined by*

$$B\left(\eta, \gamma\right) = \kappa(\eta) - \kappa(\gamma) - \langle \eta - \gamma, \nabla c(\gamma) \rangle\,, \tag{13}$$

where $\langle \cdot, \cdot \rangle$ is the inner product and and other notations are similar to equation (1). It can be shown that for all regular exponential families (see, Brwon 1986) $\kappa(\cdot)$, 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) distance between p_η and q_γ in model (1) is equal to Bregman divergence between natural parameters. That is:*

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0025.v1

where the KL divergence (Kullback–Leibler, 1951), between densities f_1 and f_2, $KL(f_1, f_2) = \mathbb{E}^{f_1} \log f_1 / f_2$.

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 proportionately more weight when they are significantly different. Lemma 2 provides the Bayesian DRE of the density ratio r in (3).

Lemma 2. *For any measurable PDF $p(\cdot \mid \eta)$ and $q(\cdot \mid \gamma)$, the Bayesian DRE of $r(t; \theta)$ associated with log–L_2 loss function $(\log \frac{\hat{r}}{r})^2$ and prior distribution $\pi(\theta)$ on $\theta = (\eta, \gamma)$ is given by*

$$\hat{r}_\pi(t) = \exp\left\{ \mathbb{E}^{\theta \mid X=t, Y=t} \, KL\,(p, q) \right\}, \tag{14}$$

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

$$\hat{r}_\pi(t) = \exp\left\{ \mathbb{E}^{\theta \mid X=t, Y=t} \, B\,(\eta, \gamma) \right\}. \tag{15}$$

Note that $\mathbb{E}^{\theta \mid X=t, Y=t}(\cdot)$ represents the expectation associated with $\theta \mid X = t, Y = t$.

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 the posterior risk $\mathbb{E}^{\theta \mid t} \ell(\hat{r}, r)$, with $\ell(\hat{r}, r) = \mathbb{E}^p (\log \frac{\hat{r}}{r})^2$, and therefore, $\frac{\partial}{\partial \log \hat{r}} \mathbb{E}^{\theta \mid t} \mathbb{E}^p \left(\log \hat{r} - \log r\right)^2 = 0$ implies $\log \hat{r}_\pi(t) = \mathbb{E}^{\theta \mid t} \mathbb{E}^p \log r = \mathbb{E}^{\theta \mid t} KL(p, q)$. Applying Lemma 1 and the fact that $\frac{\partial^2}{\partial \log \hat{r}^2} \mathbb{E}^{\eta \mid t} \mathbb{E}^p \left(\log \hat{r} - \log r\right)^2 \geq 0$, completes the proof. □

An interesting alternative representation of Bayesian DRE from Lemma 2 is that is the Bayesian DRE can be expressed in terms of the plug–in DRE in (7).

Corollary 1. *For any PDF p and q belong 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*

$$\hat{r}_\pi(t) = \hat{r}_{plug}(t)\, H(t), \tag{16}$$

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

$$\hat{r}_{plug}(t) = \frac{\exp\{\langle s(t), \mathbb{E}^{\eta \mid X=t, Y=t} \eta \rangle - \log \mathbb{E}^{\eta \mid X=t, Y=t} \log c(\eta)\}}{\exp\{\langle s(t), \mathbb{E}^{\gamma \mid X=t, Y=t} \gamma \rangle - \log \mathbb{E}^{\gamma \mid X=t, Y=t} \log c(\gamma)\}\}}, \tag{17}$$

and the correction factor

$$H(t) = \frac{\exp\{\log \mathbb{E}^{\eta \mid X=t, Y=t} c(\eta) - \mathbb{E}^{\eta \mid X=t, Y=t} \log c(\eta)\}}{\exp\{\log \mathbb{E}^{\gamma \mid X=t, Y=t} c(\gamma) - \mathbb{E}^{\gamma \mid X=t, Y=t} \log c(\gamma)\}}, \tag{18}$$

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

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

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0025.v1

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

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

This completes the proof. □

Table 1 explores the correction factor $H(\cdot)$ for certain densities which belong to exponential family associated with log–L_2 loss. Let the likelihood functions in below with sub indices $i = 1, 2$ represent the underlying distributions are drawn from p and q accordingly to (1) and (2) respectively. It is worth noting that posing constraints on the hyper–parameters leads to have $H(t)$ as a constant function in t. In fact, $H(t) = 1$ induces the Bayesian DRE $\hat{r}_\pi$ coincides with the plug–in DRE $\hat{r}_{plug}$.

Note that the notation $\psi(\alpha)$ in below is for a "digamma" function and is given by $\frac{\partial \Gamma(\delta)/\partial \delta}{\Gamma(\delta)}$. Also, *Gam*, *P*, *Pa*, *W* and *Geo* stand for gamma, poisson, pareto, weibull and geometric distributions respectively.

Density	**Conjugate Prior**	**Correction factor** $H(t)$	**Condition for** $H = 1$
$Gam(\alpha_i, \beta_i)$	$\beta_i \sim Gam(a_i, b_i)$	$(\alpha_1 + a_1)^{\alpha_1} (\alpha_2 + a_2)^{\alpha_2} \exp\{-\alpha_1 \psi(\alpha_1 + a_1) - \alpha_2 \psi(\alpha_2 + a_2)\}$	$\alpha_1 = \alpha_2,\ a_1 = a_2$
$\mathbb{N}(\theta_i, \sigma_i^2)$	$\theta_i \sim \mathbb{N}(\mu_i, \tau_i^2)$	$\exp\{1/2(\frac{1}{b_1(1+b_1)} - \frac{1}{b_2(1+b_2)})\}$	$\frac{b_1}{b_2} = \frac{1+b_1}{1+b_2}$, where $b_i = \sigma_i^2/\tau_i^2$
$\mathbb{N}(\theta_i, \sigma_i^2)$	$\sigma_i^2 \sim IG(\alpha_i, \beta_i)$	$\sqrt{\frac{\alpha_1 - 1/2}{\alpha_2 - 1/2}} \exp\{1/2(\psi(\alpha_2 + 1/2) - \psi(\alpha_1 + 1/2))\}$	$\alpha_1 = \alpha_2$
$Pa(a_i, b_i)$	$b_i \sim Gam(\alpha_i, \beta_i)$	$\frac{\alpha_1 + 1}{\alpha_2 - 1} \exp\{\frac{\alpha_1 - \alpha_2}{\alpha_1 \alpha_2} + \psi(\alpha_2) - \psi(\alpha_1)\}$	$\alpha_1 = \alpha_2$
$W(\lambda_i, k_i)$	$\lambda_i \sim IG(\alpha_i, \beta_i)$	$\alpha_1^{k_1} / \alpha_2^{k_2} \exp\{k_2 \psi(\alpha_2 + 1) - k_1 \psi(\alpha_1 + 1)\}$	$k_1 = k_2, \alpha_1 = \alpha_2$
$Bin(n_i, p_i)$	$p_i \sim Bet(\alpha_i, \beta_i)$	$\frac{\alpha_2 + \beta_2 + n_2}{\alpha_1 + \beta_1 + n_1} \frac{\beta_1 + n_1 - t}{\alpha_2 + n_2 - t} \times$ $\exp\{\psi(\beta_2 + n - t) - \psi(\beta_1 + n_1 - t) + \psi(\alpha_1 + \beta_1 + n_1) - \psi(\alpha_2 + \beta_2 + n_2)\}$	$\alpha_1 = \alpha_2,\ \beta_1 = \beta_2,\ n_1 = n_2$
$P(\lambda_i)$	$\lambda_i \sim Gam(\alpha_i, \beta_i)$	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 1. Correction factor $H(t)$ and conditions when $H(t) = 1$ associated with log–L_2 loss

3.1.2. Log–L_1 loss function

For some larger errors in loss function (12), one needs to consider Log–L_1 loss function $\ell(\tilde{r}, r) = 2|\log \frac{r}{\tilde{r}}| - 2$ for all $\tilde{r} \leq r/e$, $\tilde{r} \geq r\, e$. Let $\tilde{r}$ be the corresponding Bayesian DRE. Similar calculation to Lemma 2 suggests

$$\tilde{r}_\pi(t) = \exp\left\{\mathbb{M}^{\theta|X=t,Y=t}\, B\,(\eta, \gamma)\right\},$$

where $\mathbb{M}^{\theta|t}(\cdot)$ is the median of the posterior density function $\theta \mid t$. Similar to Section 3.1.1, and equations (14) and (15), we can write $\tilde{r}_\pi(t) = \exp\left\{\mathbb{M}^{\theta|X=t,Y=t} KL(p, q)\right\}$ which expectations are replaced by medians.

Alike Corollary 1 we also can express $\tilde{r}_\pi(t)$ in terms of product of the correction factor and the plug–in DRE. That is, under Log–L_1 loss $L(\tilde{r}, r) = |\log \frac{\hat{r}}{r}|$, and prior distribution $\pi(\theta)$, for $\theta = (\eta, \gamma)$, we have

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0025.v1

where the $\tilde{r}_{plug}(t)$ and $H'(t)$ are obtained in the same fashion in (17) and (18) except applying median $\mathbb{M}^{\theta|X=t,Y=t}(\cdot)$ instead of $\mathbb{E}^{\eta|X=t,Y=t}(\cdot)$. Notice that for instance the results in Table 1 hold, wherever the posterior densities turn out to be symmetric about their means (or medians).

3.2. Examples of the Bayesian DRE and some applications

We consider commonly used families of distributions and study the corresponding Bayesian DREs. 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. The following table presents the KL divergence between p and q.

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

Table 2. KL divergence between p and q

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*

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

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

yields the Bayesian DRE as below.

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

If we assume the multivariate case, ie. $X|\theta_1 \sim \mathbb{N}_p(\theta_1, \Sigma_1)$ 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 univarite case and $\theta_1|t \sim$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0025.v1

$\mathbb{N}_p(W_1(\Sigma_1^{-1}t + V_1^{-1}\xi_1), W_1)$ *and* $\theta_2|t \sim \mathbb{N}_p(W_2(\Sigma_2^{-1}t + V_2^{-1}\xi_2), W_2)$ *with* $W_i = (\Sigma_i^{-1} + V_i^{-1})^{-i}$. *Therefore the Bayesian DRE for the ratio of two multivariate normal densities equals to*

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

Example 2. *Let* $X|\sigma_1^2 \sim Ray(\sigma_1^2)$ *and* $Y|\sigma_2^2 \sim Ray(\sigma_2^2)$ *be two independent random variables from 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 the 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* $\lambda > 0$, *which yields to have the posterior distribution as follows*

$$\lambda_i \mid t \sim IG(1 + \alpha_i, \beta_i + t^2)\,. \tag{21}$$

Therefore the Bayesian DRE for 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 $\lambda \sim IG(\alpha,\beta)$, such as $\mathbb{E}(\log\lambda) = \log\beta - \psi(\alpha)$, where $\psi(\alpha)$ is a "digamma" function, the Bayesian DRE under log–L_2 loss is give 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 log–L_1 loss function and consequently for the log–Huber loss and they must be calculated iteratively.

The following remarks are some clarification 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* $\eta \mid X = t$ *and* $\gamma \mid Y = t$.

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

Remark 3. *Equations (14) and (15) are applicable for not only exponential family but any distributions. The key point is that the nice representation of Bayesian DRE in (16) and (19) are correct based on models (1) and (2).*

Example 3. *Suppose,* $X \sim Uniform(0,\theta_1)$ *is independent of* $Y \sim Uniform(0,\theta_2)$, *with* $\theta_1 < \theta_2$. *It is easy to check Pareto distribution is conjugate to a uniform. Hence assume that* $\pi(\theta_i) = \alpha_i\beta_i^{\alpha_i}\theta_i^{-(\alpha_1+1)}$ *for* $\theta \geq \beta_i$ *and* $i = 1, 2$, *therefore*

$$\theta_i \mid t \sim Pareto(1 + \alpha_i, \max(\beta_i, t))$$

Moreover, assuming p *and* q *are the PDF's of* X *and* Y *respectively, therefore,* $KL(p,q) = \log\theta_2 - \log\theta_1$ *(see Table 2). Employing the fact that transformation* $\log(z/b)$ *has exponential distribution with mean of* a, *given* $Z \sim Pareto(a,b)$, *and hence* $\mathbb{E}^Z\log Z = a + \log b$, *after some calculation we have the Bayes DRE associated with log–*L_2

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0025.v1

4. Other applications

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

1. **Estimating α–divergence function between two probability densities:**
A discrepancy measure between densities p_η and q_γ applicable to the class of Ali-Silvey (1966) distance also known as α–divergence (Csiszàr, 1967) is given by

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

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 say, 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

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

where t_i are drown from $p(\cdot \mid \eta)$. It is worthwhile to note that there are other work in order to estimate the α–divergence loss function (eg. Póczos and Schneider 2010) but our estimator in (23), is based on the Bayesian parametric method based on the DRE. 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

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

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

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

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

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0025.v1

By setting the case when the both densities follow the identical distribution from (1) (for instance 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

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

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

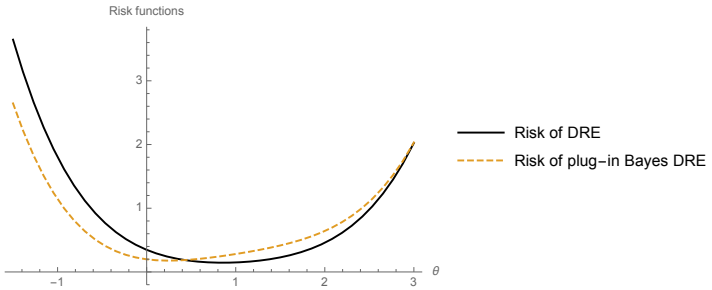

Figure 2. Risk function of the Bayes and plug–in DRE under L_2 loss in normal models $p = \mathbb{N}(\theta, 1)$, $q = \mathbb{N}(\theta, 2)$ and corresponding hyper–parameters $\xi = 1, \xi = 2 = 0, \tau_1 = \tau_2 = 1$ correspond to Example 1.

Figure 3 shows changing the frequentist risk performance of the plug–in DRE $\hat{r}_{plug}$ and the Bayes DRE $\hat{r}_\pi$ under log–L_2 when the variance σ^2 varies using Corollary 1.

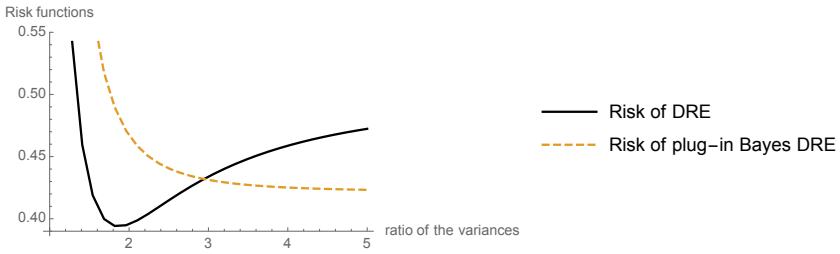

Figure 3. Risk function of the Bayes and plug–in DRE under log–L_2 loss in normal models $p = \mathbb{N}(1, \sigma^2)$, $q = \mathbb{N}(1, \sigma^2)$ and hyper–parameters $\xi = 1, \xi = 2 = 0, \tau_1 = \tau_2 = 1$ in Example 1.

Finally, the following graphical illustrations depict the risk function of the Bayes and plug–in DRE under log–Huber loss over possible ranges of the mean and variance respectively in Example 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2019 doi:10.20944/preprints201903.0025.v1

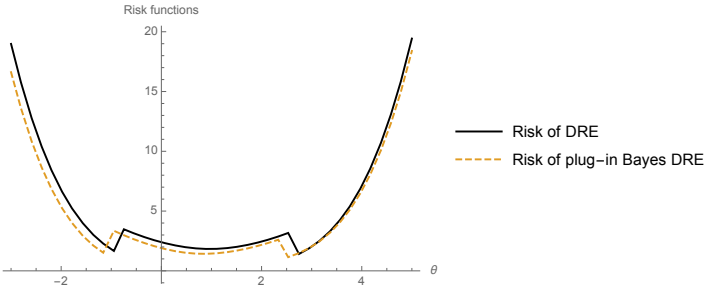

Figure 4. Risk function of the Bayes and plug–in DRE under log–Huber loss in normal models $p = \mathbb{N}(\theta, 1)$, $q = \mathbb{N}(\theta, 2)$ and corresponding hyper–parameters $\xi = 1$, $\xi = 2 = 0$, $\tau_1 = \tau_2 = 1$ for all possible range of μ in Example 1.

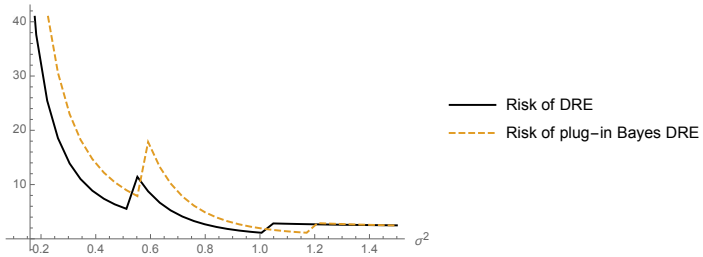

Figure 5. Risk function of the Bayes and plug–in DRE under log–Huber loss in normal models $p = \mathbb{N}(0, \sigma^2)$, $q = \mathbb{N}(1, \sigma^2)$ and corresponding hyper–parameters $\xi = 1$, $\xi = 2 = 0$, $\tau_1 = \tau_2 = 1$, for all possible range of σ^2 in Example 1.

6. Concluding remarks

Estimating the ratio of two or more densities has received a widespread attention in recent years. Till date, 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 DRE . 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 somehow different result from a non–parametric point of view which is often the time the plug–in estimators perform poorly as opposed to empirical non–parametric Bayesian methods typically include stochastic processes such as the Gaussian process and the Dirichlet process. There are instances (for example, see Krnjajic et al. 2008, and the references therein) that for certain type of count data, the nonparametric Bayesian methodology provides enhanced flexibility to fit the data, provide rich posterior inferences, and provide 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 thus naive algorithmic approaches to computing posteriors is generally infeasible. Finally, the application to estimating the α–divergence between two PDF's was discussed.

Acknowledgments: The authors thank Prof. Eric Marchand (Université de Sherbrooke) for his useful comments.

References

3. Ali, S. M. and Silvey, S. D. A general class of coefficients of divergence of one distribution from another. *J. Royal Stat. Soc. Ser. B* **1966**, *28*, 131–142.
4. Berger, J.O. *Statistical Decision Theory and Bayesian Analysis.* New York, Springer- Verlag, 1985.
5. Csiszàr, I. Information-type measures of difference of probability distributions and indirect observation. *Studia Sci. Math. Hungar* **1967,** *2*, 299–318.

6. Deledalle, C. A. Estimation of Kullback-Leibler losses for noisy recovery problems within the exponential family. *Electronic journal of statistics* **2017,** *11(2)*, 3141-3164.
7. Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., Schölkopf, B. Covariate shift by kernel mean matching. In J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, N. Lawrence (Eds.), Data set shift in machine learning. *Cambridge, MA, USA: MIT Press,* Chapter 8, 131-160.
8. Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.
9. Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., Kanamori, T. Inlier-Based Outlier Detection via Direct Density Ratio Estimation. *ICDM '08 Proceedings of the 2008 Eighth IEEE International Conference on Data Mining* **2008**, 223-232.
10. Huber, P. Robust estimation of a location parameter, *Annals of Mathematical Statistics* **1964** *53*, 73-101.
11. Kanamori, T., Hido, S., and Sugiyama, M. A Least-squares Approach to Direct Importance Estimation *The Journal of Machine Learning Research* **2009**, *10*, 1391-1445.
12. Krnjajic, M., Kottas, A., & Draper, D. Parametric and nonparametric Bayesian model specification: A case study involving models for count data. *Computational Statistics & Data Analysis* **2008**, *52*, 2110-2128.
13. Lehman, E.L. and Casella, G. *Theory of point estimation. Springer Texts in Statistics.* Springer-Verlag, New York, second edition, 1998.
14. Murphy, Kevin P. *Machine learning: a probabilistic perspective* **212,** MIT press.
15. Nguyen, X., Wainwright, M.J., Jordan, M.I. Estimating divergence functional and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory* **2010,** *56(11)*, 5847–5861
16. Póczos, B., Schneider, J. On the estimation of alpha-divergences. *In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* **2011,** 609-617.
17. Nielsen F., Nock. "Entropies and cross–entropies of exponential families," *IEEE International Conference on Image Processing, Hong Kong* **2010,** 3621-3624.
18. Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N. *Dataset shift in machine learning. Cambridge, MA, USA, MIT Press.*, 2009.
19. Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, *90* **2000,**, 227-244.
20. 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 Networks, 24* **2011,** 183-198.
21. Sugiyama, M., Kanamori, T., Suzuki, T., Hido, S., Sese, J., Takeuchi, I., and Wang, L. A density-ratio framework for statistical data processing. *IPSJ Transactions on Computer Vision and Applications* **2009,** 1, 183-208.
22. Sugiyama, M., Suzuki, T., Kanamori, T. *Density ratio estimation in machine learning.* Cambridge, UK, Cambridge University Press. 2012 a.
23. Sugiyama, Masashi and Kawanabe, Motoaki. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation.* The MIT Press. 2012 b.
24. Sugiyama, M., Müller, K. R. Input-dependent estimation of generalization error under covariate shift. *Statistics and Decisions* **2005,** *23*, 249-279.
25. Sugiyama, M., Krauledat, M., Müller, K. R. *Covariate shift adaptation by importance weighted cross-validation.* Journal of Machine Learning Research, *8* **2007,** 985-1005.
26. Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bunau, P., and Kawanabe, M. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics* **2008,** *60*, 699-746.
27. Sugiyama, M. Hara, S. von Bünau, P. Suzuki, T. Kanamori, T. Kawanabe, M. Direct density ratio estimation with dimensionality reduction. *SIAM International Conference on Data Mining*, **. 2010.**
28. Sugiyama, Mi, Suzuki, T., and Kanamori, T. Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics* **2012 c,** *64*, 1009–1044.