Electronic Journal of Statistics
Vol. 7 (2013) 1587–1606
ISSN: 1935-7524
DOI: 10.1214/13-EJS817

Normalized estimating equation for robust parameter estimation

Hironori Fujisawa*

The Institute of Statistical Mathematics
Tachikawa, Tokyo 190-8562, Japan
e-mail: fujisawa@ism.ac.jp

Abstract: Robust parameter estimation has been discussed as a method for reducing a bias caused by outliers. An estimating equation using a weighted score function is often used. A typical estimating equation is non-normalized, but this paper considers a normalized estimating equation, which is corrected to ensure that the mean of the weight is one. In robust parameter estimation, it is important to control the difference between the target parameter and the limit of the robust estimator, which is referred to as the latent bias in this paper. The latent bias is usually discussed in terms of influence function and breakdown point. It is illustrated by some examples that the latent bias can be close to zero for the normalized estimating equation even if the proportion of outliers is not small, but not close to zero for the non-normalized estimating equation. Furthermore, this behavior of the normalized estimating equation can be proved under mild conditions. The asymptotic normality of the robust estimator is also presented and then it is shown that the outliers are naturally ignored with an appropriate proportion of outliers from the viewpoint of asymptotic variance. The results can be extended to the regression case. The behaviors of the latent bias and mean squared error are investigated by numerical studies.

AMS 2000 subject classifications: Primary 62F35.
Keywords and phrases: Latent bias, normalized estimating equation, robust parameter estimation, weighted score function.

Received April 2013.

Contents

*This work was supported by Grant-in-Aid for Scientific Research of the Ministry of Education, Culture, Sports, Science and Technology.

1. Introduction

Maximum likelihood estimation is a typical form of parameter estimation. However, if an outlier is present in observations, then it often causes a severe bias. To overcome this problem, many methods for robust parameter estimation against outliers have been proposed (Hampel et al., 1986; Maronna, Martin and Yohai, 2006; Huber and Ronchetti, 2009).

Let the parametric density be denoted by $f(x;\theta) = f_\theta(x)$. Let the log-likelihood and score functions be denoted by $l(x;\theta) = \log f(x;\theta)$ and $s(x;\theta) = (\partial/\partial\theta)l(x;\theta)$, respectively. Then the maximum likelihood estimator is a root of the estimating equation given by $\sum_{i=1}^n s(x_i;\theta) = 0$, where $x_1,\dots,x_n$ are the observations. To weaken the adverse effect of outlier, an estimating equation using a weighted score function, $\sum_{i=1}^n w(x_i;\theta)s(x_i;\theta) = 0$, can be considered for robust parameter estimation (Field and Smith, 1994). The weight $w(x_i;\theta)$ is small when x_i is an outlier. However, the bias-correction is necessary to ensure Fisher consistency. The bias-corrected estimating equation is given by

$$\frac{1}{n}\sum_{i=1}^n w(x_i;\theta)s(x_i;\theta) = \mathrm{E}_{f_\theta}\left[w(x;\theta)s(x;\theta)\right].$$

Recently, the density power weight $w(x;\theta) = f(x;\theta)^\gamma$ $(\gamma > 0)$ has been discussed for robust parameter estimation, because $f(x_i;\theta)$ is close to zero when x_i is an outlier. Basu et al. (1998) proposed this type of estimating equation and discussed the corresponding divergence. The divergence with $\gamma = 1$ is the same as the L_2-divergence, which is a well-known divergence to generate a strong robust estimator (Scott, 2001). The divergence limits to the KL-divergence as γ goes to zero. The tuning parameter γ controls the trade-off between bias and variance. The divergence was applied to independent component analysis (Minami and Eguchi, 2002), a mixture of independent component analysis models (Mollah, Minami and Eguchi, 2006), Gaussian graphical models (Miyamura and Kano, 2006), model selection (Mattheou, Lee and Karagrigoriou, 2009) and kernel principal component analysis (Huang, Yeh and Eguchi, 2009). A general weight $\xi(l(x;\theta))$, including the density power weight, was also discussed by Eguchi and Kano (2001) and Murata et al. (2004):

$$\frac{1}{n}\sum_{i=1}^n \xi(l(x_i;\theta))s(x_i;\theta) = \mathrm{E}_{f_\theta}\left[\xi(l(x;\theta))s(x;\theta)\right]. \tag{1.1}$$

It is easy to construct the corresponding divergence, which belongs to a class of Bregman divergence.

On the estimating equation with the weight $\xi(l(x;\theta))$, the weight is not always normalized; that is, the mean of weight is typically not one, more precisely, $(1/n)\sum_{i=1}^{n}\xi(l(x;\theta)) \neq 1$. In this paper, this type of estimating equation is called a non-normalized estimating equation. Let us consider another estimating equation by replacing the weights $\xi(l(x_i;\theta))$ and $\xi(l(x;\theta))$ by

$$w(x_i;\theta) = \frac{\xi(l(x_i;\theta))}{(1/n)\sum_{i=1}^{n}\xi(l(x_i;\theta))} \text{ and } W(x;\theta) = \frac{\xi(l(x;\theta))}{\mathrm{E}_{f_\theta}\left[\xi(l(x;\theta))\right]},$$

so that $(1/n)\sum_{i=1}^{n} w(x_i;\theta) = 1$ and $\mathrm{E}_{f_\theta}\left[W(x;\theta)\right] = 1$. A normalized estimating equation is defined as

$$\frac{\sum_{i=1}^{n}\xi(l(x_i;\theta))s(x_i;\theta)}{\sum_{i=1}^{n}\xi(l(x_i;\theta))} = \frac{\mathrm{E}_{f_\theta}\left[\xi(l(x;\theta))s(x;\theta)\right]}{\mathrm{E}_{f_\theta}\left[\xi(l(x;\theta))\right]}. \tag{1.2}$$

The normalized estimating equation with the density power weight was proposed by Windham (1995). Jones et al. (2001) constructed the corresponding divergence, which was further explored by Fujisawa and Eguchi (2008). The divergence is related to Tsallis entropy (Tsallis, 1988, 2009; Ferrari and Yang, 2010; Ferrari and La Vecchia, 2012; Eguchi and Kato, 2010; Eguchi, Komori and Kato, 2011). Additionally, two types of divergences were further discussed by Cichocki and Amari (2010) and Eguchi and Kato (2010) and applied to vector quantization by Villmann and Haase (2011).

Here, we remark the difference between the non-normalized and normalized estimating equations. It is easy to construct the corresponding divergence for the non-normalized estimating equation, as described already, but not easy (often impossible) for the normalized estimating equation except for the density power weight. For this reason, there would have been no discussion about other weights except for the density power weight on the normalized estimating equation. The non-existence of the corresponding divergence for the estimating equation was discussed in the framework of generalized estimating equation (McCullagh and Nelder, 1983).

In robust parameter estimation, it is important to control the difference between the target parameter and the limit of the robust estimator, which is referred to as the latent bias in this paper. The latent bias is usually discussed in terms of influence function and breakdown point. A distinguishing feature of the normalized estimating equation with the density power weight is that the latent bias becomes arbitrarily small when the occurrence probability of the outlier becomes arbitrarily small in a certain sense, even if the proportion of outliers is not small (Fujisawa and Eguchi, 2008). It should be noted that this favorable property was proved by using a specific property of the corresponding divergence. In this paper, this result is extended to a normalized estimating equation with any weight, which enables us to use various weights, including the logistic weight (Eguchi and Kano, 2001; Murata et al., 2004; Takenouchi

and Eguchi, 2004). The approach of the proof is different from the divergence-based one, because we cannot use a convenient property of divergence. It is further illustrated by some examples that the latent bias can be close to zero for the normalized estimating equation, but not always for the non-normalized estimating equation.

This paper is organized as follows. The non-normalized and normalized estimating equations are described in Section 2. The corresponding estimators can be regarded as M-estimators. In Section 3, the latent bias is discussed for non-normalized and normalized estimating equations and it is shown that the latent bias can be arbitrarily small for a normalized estimating equation even if the proportion of outliers is not small. Asymptotic properties of the robust estimators are presented in Section 4. These results are extended to the regression case in Section 5. Numerical examples are illustrated in Section 6. Some discussions are given in Section 7.

2. Estimating equation

The non-normalized estimating equation given by (1.1) can be expressed as

$$\sum_{i=1}^{n} \psi_U(x_i;\theta) = 0,$$

where

$$\psi_U(x;\theta) = \xi(l(x;\theta))s(x;\theta) - \mathrm{E}_{f_\theta}\left[\xi(l(x;\theta))s(x;\theta)\right]. \tag{2.1}$$

The robust estimator $\hat{\theta}_U$ is defined as a root of this estimating equation, which is an M-estimator. To weaken an adverse effect of outlier, we assume that the weight $\xi(l(x;\theta))$ is close to zero for an outlier x, more precisely,

$$\lim_{a\to-\infty} \xi(a) = 0, \tag{2.2}$$

because $f(x;\theta)$ is close to zero for an outlier x and $l(x;\theta) = \log f(x;\theta)$ goes to minus infinity as $f(x;\theta)$ approaches zero. Suppose that the function $\xi(a)$ and the density function $f(x;\theta)$ satisfy some conditions, including differentiability, integrability, and so on, which are described in the subsequent sections. Various properties, including asymptotic properties and robustness of the estimator and test, can be easily obtained by the theory of M-estimator (Maronna, Martin and Yohai, 2006; Heritier and Ronchetti, 1994).

Some types of weights have been discussed. One is the density power weight, $\xi(l(x;\theta)) = \exp(\gamma l(x;\theta)) = f(x;\theta)^\gamma$, as described already in Section 1. The density power weight for an outlier x^* decreases with increasing γ. The logistic weight, $\xi(l(x;\theta)) = \exp(l(x;\theta))/(\exp(l(x;\theta)) + \eta) = f(x;\theta)/(f(x;\theta) + \eta)$, was considered by Eguchi and Kano (2001) and used in a divergence related to the boosting (Takenouchi and Eguchi, 2004). The tuning parameter η was referred to as the value of saturation in multilayer perceptron models in neural networks.

The logistic weight decreases with increasing η and it is essentially proportional to $f(x;\theta)$ for a sufficiently large η. The threshold type of weight, $\xi(l(x;\theta)) = \min\{f(x;\theta), c\}$, can also be used and a similar type of weight was applied to a logistic model by Croux and Haesbroeck (2003).

The normalized estimating equation given by (1.2) can be rewritten as

$$\sum_{i=1}^{n} \psi_N(x_i;\theta) = 0,$$

where

$$\psi_N(x;\theta) = \xi(l(x;\theta))s(x;\theta)\mathrm{E}_{f_\theta}\left[\xi(l(x;\theta))\right] - \xi(l(x;\theta))\mathrm{E}_{f_\theta}\left[\xi(l(x;\theta))s(x;\theta)\right]. \quad (2.3)$$

The robust estimator $\hat{\theta}_N$ is defined as a root of this estimating equation, which is also an M-estimator.

3. Latent bias

Let $f(x) = f(x;\theta^*)$ be the target density. Let $\delta(x)$ be the contamination density related to outliers. Suppose that the observations are drawn from the underlying density given by

$$g(x) = (1-\varepsilon)f(x) + \varepsilon\delta(x),$$

where ε is the proportion of outliers. Let $\hat{\theta}_\psi$ be the estimator defined as a root of the estimating equation $\sum_{i=1}^{n}\psi(x_i;\theta) = 0$. We assume the Fisher consistency. Let θ^*_ψ be the limit of $\hat{\theta}_\psi$. The bias caused by contamination can be expressed as $\theta^*_\psi - \theta^*$, which is hereafter referred to as the latent bias.

The latent bias $\theta^*_\psi - \theta^*$ can be approximated to $\varepsilon\mathrm{IF}_\psi(x^*;\theta^*)$ if $\delta(x)$ is the dirac function at x^* and ε is sufficiently small, where $\mathrm{IF}_\psi(x^*;\theta^*)$ is the influence function, given by $\mathrm{IF}_\psi(x^*;\theta^*) = -\{\mathrm{E}_{f_{\theta^*}}[(\partial\psi/\partial\theta'(x;\theta^*)]\}^{-1}\psi(x^*;\theta^*)$ (Huber and Ronchetti, 2009). It is favorable that the influence function $\mathrm{IF}_\psi(x;\theta)$ approaches zero as $|x|$ goes to infinity, because the latent bias can be approximated to zero for a large value of $|x|$. The function $\psi(x;\theta)$ is said to be redescending when the function $\psi(x;\theta)$ approaches zero as $|x|$ goes to infinity, which implies the above favorable property of the influence function from the formula of IF_ψ.

Consider the simple case where the target density is an exponential distribution with mean one. Figure 1 shows the influence functions for the non-normalized and normalized estimating equations with density power weight ($\gamma = 1$). The influence function for the normalized estimating equation is redescending because it is easily shown that $\xi(l(x;\theta))$ and $\xi(l(x;\theta))s(x;\theta)$ on the formula (2.3) approach zero as x goes to infinity. This property is generalized in what follows. However, the influence function for the non-normalized estimating equation is not redescending from the formula (2.1), because the bias-correction term for ψ_U, $\mathrm{E}_{f_\theta}\left[\xi(l(x;\theta))s(x;\theta)\right]$, is not zero.

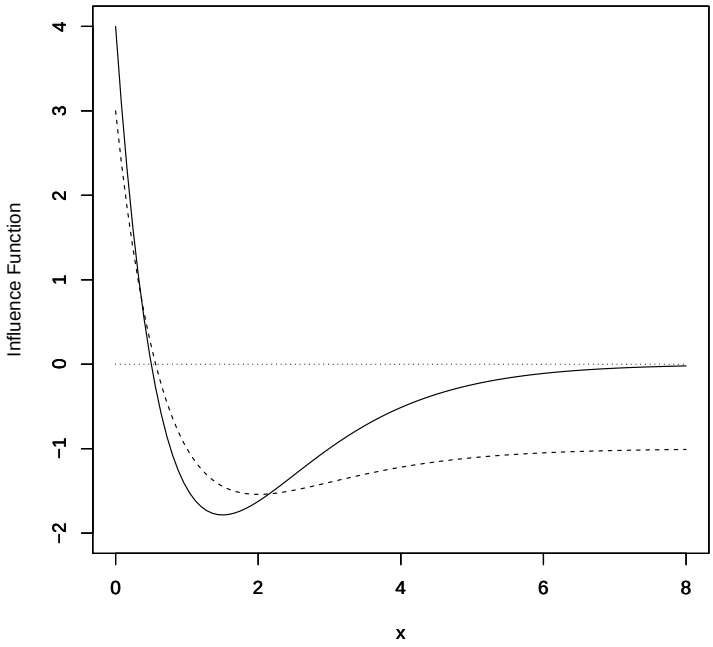


FIG 1. *Influence function for the mean parameter of the exponential distribution. The true mean is one. The tuning parameter γ is one. The solid and dotted lines correspond to the normalized and non-normalized estimating equations, respectively.*

Suppose that

$$\mathrm{E}_\delta\left[\xi(l(x;\theta))\right] \approx 0, \qquad \mathrm{E}_\delta\left[\xi(l(x;\theta))s(x;\theta)\right] \approx 0, \tag{3.1}$$

in the neighborhood of $\theta = \theta^*$. These conditions hold for various combinations of weights and distributions. Here we consider the simple case where δ is the dirac function at a sufficiently large x^*. We can suppose that $f(x^*;\theta) \approx 0$ because x^* can be regarded as an outlier. The first condition becomes $\xi(l(x^*;\theta)) \approx 0$. This immediately follows from the property (2.2). The second condition becomes $\xi(l(x^*;\theta))s(x^*;\theta) \approx 0$. This holds for various combinations of weights and distributions. (Some examples are given in Appendix A). Consequently, from the formula (2.3), the condition (3.1) implies the redescending property of ψ_N. Therefore, we see that the condition (3.1) is more general than the redescending property of ψ_N, because the redescending property is considered under a restricted situation that θ is the true parameter θ^* and δ is the dirac function at x^*. Under the condition (3.1), we obtain a stronger property than usual, as described later.

Let us consider the limit of the normalized estimating equation, given by

$$\frac{\mathrm{E}_g\left[\xi(l(x;\theta))s(x;\theta)\right]}{\mathrm{E}_g\left[\xi(l(x;\theta))\right]} = \frac{\mathrm{E}_{f_\theta}\left[\xi(l(x;\theta))s(x;\theta)\right]}{\mathrm{E}_{f_\theta}\left[\xi(l(x;\theta))\right]}. \tag{3.2}$$

We see that

$$\begin{aligned} \mathrm{E}_g\left[\xi(l(x;\theta))s(x;\theta)\right] &= (1-\varepsilon)\mathrm{E}_f\left[\xi(l(x;\theta))s(x;\theta)\right] + \varepsilon\mathrm{E}_\delta\left[\xi(l(x;\theta))s(x;\theta)\right], \\ \mathrm{E}_g\left[\xi(l(x;\theta))\right] &= (1-\varepsilon)\mathrm{E}_f\left[\xi(l(x;\theta))\right] + \varepsilon\mathrm{E}_\delta\left[\xi(l(x;\theta))\right]. \end{aligned}$$

From the condition (3.1), the normalized estimating equation (3.2) is roughly expressed as

$$\frac{\mathrm{E}_{f_{\theta^*}}\left[\xi(l(x;\theta))s(x;\theta)\right]}{\mathrm{E}_{f_{\theta^*}}\left[\xi(l(x;\theta))\right]} \approx \frac{\mathrm{E}_{f_\theta}\left[\xi(l(x;\theta))s(x;\theta)\right]}{\mathrm{E}_{f_\theta}\left[\xi(l(x;\theta))\right]}. \tag{3.3}$$

Note that the proportion of outliers, ε, vanishes. If this approximation is replaced by equality, then θ^* is a root. Let the limit of the normalized estimating equation be denoted by

$$\lambda_g(\theta) = \mathrm{E}_g\left[\psi_N(x;\theta)\right] = 0.$$

Let θ_N^* be the root of $\lambda_g(\theta) = 0$. The formula (3.3) implies that $\theta_N^* \approx \theta^*$, which shows the possibility that the latent bias can be close to zero even if the proportion of outliers is not small.

Let us give a clear statement of the above discussion. Before that, we calculate the differential of $\lambda_{f_{\theta^*}}(\theta)$ at $\theta = \theta^*$. It follows from straightforward but lengthy calculations that

$$\begin{aligned} \frac{\partial\lambda_{f_{\theta^*}}}{\partial\theta'}(\theta^*) &= \int \xi(l(x;\theta^*))s(x;\theta^*)f(x;\theta^*)dx \int \xi(l(x;\theta^*))s(x;\theta^*)'f(x;\theta^*)dx \\ &\quad - \int \xi(l(x;\theta^*))f(x;\theta^*)dx \int \xi(l(x;\theta^*))s(x;\theta^*)s(x;\theta^*)'f(x;\theta^*)dx. \end{aligned} \tag{3.4}$$

This is non-positive by Cauchy-Schwartz's inequality and usually negative definite for various weights and density functions. The following theorem can be shown from this assumption and implicit function theorem. The proof is given in Appendix B.

Theorem 3.1. *Suppose that $\partial\lambda_{f_{\theta^*}}/\partial\theta'(\theta^*)$ is negative definite. Let $B_\nu(a)$ be the ball region with center a and radius ν. Assume that for any sufficiently small $\nu > 0$, $\lambda_\delta(\lambda_{f_{\theta^*}}^{-1}(\tau)) \in B_{\nu(1-\varepsilon)/\varepsilon}(0)$ for $\tau \in B_\nu(0)$. Then, there exists a root θ_N^* of $\lambda_g(\theta) = 0$ such that $\theta_N^* \in B_\nu(\theta^*)$ and $\theta_N^*(f_{\theta^*}) = \theta^*$.*

In the conclusion of Theorem 3.1, $\theta_N^* \in B_\nu(\theta^*)$ and $\theta_N^*(f_{\theta^*}) = \theta^*$ say that the latent bias can be arbitrarily small and Fisher consistency holds, respectively. Next we compare the assumption of Theorem 3.1 with the condition (3.1). We see that

$$\begin{aligned} \lambda_\delta(\theta) &= \mathrm{E}_\delta[\psi_N(\theta)] \\ &= \mathrm{E}_\delta\left[\xi(l(x;\theta))s(x;\theta)\right]\mathrm{E}_{f_\theta}\left[\xi(l(x;\theta))\right] - \mathrm{E}_\delta\left[\xi(l(x;\theta))\right]\mathrm{E}_{f_\theta}\left[\xi(l(x;\theta))s(x;\theta)\right]. \end{aligned}$$

The condition (3.1) implies that $\lambda_\delta(\theta) \approx 0$ in the neighborhood of $\theta = \theta^*$. Note that $\lambda_{f_{\theta^*}}^{-1}(B_\nu(0))$ is the neighborhood of $\theta = \theta^*$ for a sufficiently small $\nu > 0$, becasue $\lambda_{f_{\theta^*}}(\theta^*) = 0$ and $\partial\lambda_{f_{\theta^*}}/\partial\theta'(\theta^*)$ is negative definite. Hence, we have $\lambda_\delta(\theta) \approx 0$ for $\theta \in \lambda_{f_{\theta^*}}^{-1}(B_\nu(0))$. This corresponds to the assumption of Theorem 3.1.

4. Asymptotic property

The M-estimator $\hat{\theta}_\psi$, which is a root of the estimating equation $\sum_{i=1}^n \psi(x;\theta) = 0$, has consistency and asymptotic normality under mild conditions (van der Vaart, 1998). We can use the following theorem to obtain the asymptotic properties of $\hat{\theta}_U$ and $\hat{\theta}_N$.

Theorem 4.1 (Theorems 5.41 and 5.42 of van der Vaart (1998))**.** *Suppose that $x_1, \ldots, x_n$ are randomly drawn from the underlying density g. We assume: (a) The function $\psi(x;\theta)$ is twice continuously differentiable with respect to θ for any x. (b) There exists a root θ^*_ψ of $E_g[\psi(x;\theta)] = 0$. (c) $E_g[\|\psi(x;\theta^*_\psi)\|^2] < \infty$. (d) $E_g[\partial\psi/\partial\theta'(\theta^*_\psi)]$ exists and is nonsingular. (e) The second-order differentials of $\psi(x;\theta)$ with respect to θ are dominated by a fixed integrable function $\ddot{\psi}(x)$ in a neighborhood of $\theta = \theta^*_\psi$. Then there exists a sequence of roots, $\{\hat{\theta}_n\}_{n=1}^\infty$, such that*

$$\begin{aligned} &(i) \qquad \hat{\theta}_n \xrightarrow{P} \theta^*_\psi, \\ &(ii) \qquad \sqrt{n}\left(\hat{\theta}_n - \theta^*_\psi\right) \xrightarrow{d} N\left(0, \tau_g^2(\theta^*_\psi)\right), \end{aligned}$$

where

$$\begin{aligned} &\tau_g^2(\theta) = J_g(\theta)^{-1} K_g(\theta) \{J_g(\theta)'\}^{-1}, \quad J_g(\theta) = E_g\left[\partial\psi/\partial\theta'(\theta)\right], \\ &K_g(\theta) = E_g\left[\psi(\theta)\psi(\theta)'\right]. \end{aligned}$$

There are many assumptions in the above theorem. The assumptions (a), (c) and (e) are very easy to verify, but the assumptions (b) and (d) are necessary to verify. When the estimating equation is non-normalized or normalized with the density power weight, the assumptions (b) and (d) are easy to verify because the corresponding divergence exists (Basu et al., 1998; Jones et al., 2001). Consider the case of the normalized estimating equation under the aforementioned conditions. The assumption (b) directly follows from Theorem 3.1. The assumption (d) can be verified by adding an extra condition like (3.1) (Appendix C). Therefore, all the assumptions hold for both non-normalized and normalized estimating equations. Furthermore, the asymptotic variance for the normalized estimating equation can be expressed as follows. The derivation is given in Appendix D.

Theorem 4.2. *Consider a normalized estimating equation. Assume the same conditions as in Theorem 3.1, and the continuity of $J_f(\theta)$, $J_\delta(\theta)$, $K_f(\theta)$ and $K_\delta(\theta)$, and that $J_\delta(\theta^*) \approx O$ and $K_\delta(\theta^*) \approx O$. It then holds that*

$$\tau_g^2(\theta^*_N) \approx \frac{1}{1-\varepsilon}\tau_f^2(\theta^*).$$

This theorem implies that the outliers are naturally ignored with the appropriate proportion of outliers. Suppose that the sample size of outlier is m. The proportion of outliers is expressed as $\varepsilon = m/n$. The asymptotic variance of the

robust estimator based on the $n-m$ observations without outliers is given by $\tau_f^2(\theta^*)/(n-m) = \tau_f^2(\theta^*)/n(1-\varepsilon) \approx \tau_g^2(\theta_N^*)/n$, which is the asymptotic variance of the robust estimator based on all the observations.

5. Regression case

The robust parameter estimation for the regression case has been discussed in Cantoni and Ronchetti (2001), Copt and Heritier (2007) and Croux, Gijbels and Prosdocimi (2012). The content in the previous sections can be extended to the regression case in a similar manner. Let x and y be the explanatory and response variables, respectively. Let $f(y|x;\theta)$ be the parametric conditional density of y given x. Let the log-likelihood and score functions be denoted by $l(y|x;\theta) = \log f(y|x;\theta)$ and $s(y|x;\theta) = (\partial/\partial\theta)l(y|x;\theta)$, respectively. The estimating equation for the maximum likelihood estimator is $\sum_{i=1}^n s(y_i|x_i;\theta) = 0$. The downweighted estimating equation is $\sum_{i=1}^n \xi(l(y_i|x_i;\theta))s(y_i|x_i;\theta) = 0$ and then the bias-corrected (non-normalized) estimating equation is given by

$$\sum_{i=1}^n \xi(l(y_i|x_i;\theta))s(y_i|x_i;\theta) = \sum_{i=1}^n \mathrm{E}\left[\xi(l(y|x_i;\theta))s(y|x_i;\theta)|f(y|x_i;\theta)\right].$$

The normalized estimating equation is expressed as

$$\frac{\sum_{i=1}^n \xi(l(y_i|x_i;\theta))s(y_i|x_i;\theta)}{\sum_{i=1}^n \xi(l(y_i|x_i;\theta))} = \frac{\sum_{i=1}^n \mathrm{E}\left[\xi(l(y|x_i;\theta))s(y|x_i;\theta)|f(y|x_i;\theta)\right]}{\sum_{i=1}^n \mathrm{E}\left[\xi(l(y|x_i;\theta))|f(y|x_i;\theta)\right]}.$$

Let $z = (x, y)$ and $z_i = (x_i, y_i)$. They can also be expressed as

$$\sum_{i=1}^n \psi_U(z_i;\theta) = 0, \qquad \sum_{i,j=1}^n \psi_N(z_i, z_j;\theta) = 0,$$

where

$$\begin{aligned}
\psi_U(z_i) &= \xi(l(y_i|x_i;\theta))s(y_i|x_i;\theta) - \mathrm{E}\left[\xi(l(y|x_i;\theta))s(y|x_i;\theta)|f(y|x_i)\right], \\
\psi_N(z_i, z_j) &= \xi(l(y_i|x_i;\theta))s(y_i|x_i;\theta)\mathrm{E}\left[\xi(l(y|x_j;\theta))|f(y|x_j)\right] \\
&\quad -\xi(l(y_i|x_i;\theta))\mathrm{E}\left[\xi(l(y|x_j;\theta))s(y|x_j;\theta)|f(y|x_j)\right].
\end{aligned}$$

It was shown in Section 3 that the latent bias can become arbitrarily small for a normalized estimating equation under some conditions. The same result holds for the regression case under some conditions similar to in Section 3. Let the underlying density of x and y given x be denoted by $g(x)$ and $g(y|x) = (1-\varepsilon)f(y|x;\theta^*) + \varepsilon\delta(y|x)$. The necessary conditions for the regression case can be obtained by replacing $f(x;\theta^*)$ and $\delta(x)$ by $f(y|x;\theta^*)g(x)$ and $\delta(y|x;\theta^*)g(x)$ in Section 3.

Note that a non-normalized estimating equation is based on the U-statistic but a normalized estimating equation is based on the V-statistic by replacing $\psi_N(z_i, z_j)$ by $(\psi_N(z_i, z_j) + \psi_N(z_j, z_i))/2$. The asymptotic properties of the U-

and V-statistics are investigated (Serfling, 1980; Lee, 1990). Hence, we expect that under additional conditions, the asymptotic distributions of the robust estimators derived from the non-normalized and normalized estimating equations can be obtained by suitable extensions of the proof for the i.i.d. case.

The weight $\xi(l(y|x;\theta))$ can downweight the score function only when y is an outlier. The robustness against an outlier of x is not incorporated on the above estimating equations. This is possible by replacing $\xi(l(y|x;\theta))$ by $\xi(l(y|x;\theta))w(x)$, where $w(x)$ is small when x is an outlier. This idea is frequently adopted in robust parameter estimation.

6. Numerical examples

The latent bias and mean squared error were investigated when the target distribution was normal with mean zero and variance one, the contamination distribution was normal with mean five and variance one, and the parametric model was normal with mean μ and variance σ^2. Two types of weights were used; one was the density power weight and the other was the logistic weight. The proportion of outliers was set to be $\varepsilon = 0.05, 0.2$. The root of the estimating equation was obtained through an iterative algorithm (Appendix E).

6.1. Latent bias

Figure 2 illustrates the latent bias in the case of the density power weight. For the mean parameter, two latent biases for the normalized and non-normalized estimating equations are closer to zero as the tuning parameter γ becomes larger. They are almost the same when $\varepsilon = 0.05$, but slightly different around $\gamma = 0.4$ when $\varepsilon = 0.2$. For the standard deviation parameter, as the tuning parameter γ becomes larger, the latent bias for the normalized estimating equation is closer to zero. In contrast, the latent bias for the non-normalized estimating equation presents a different behavior. It can not be close to zero and is farther from zero after it has attained the minimum. Additionally, when $\varepsilon = 0.2$, the minimum value is larger than 0.2, which is very large in comparison to the target parameter $\sigma = 1$. Figure 3 depicts the latent bias in the case of the logistic weight. The behaviors were similar to those in the case of density power weight, except that the latent bias for the non-normalized estimating equation can be much closer to zero.

In Figure 4, the mean of contamination distribution was changed to $\mu_{\text{out}} = 10$. The density power weight was investigated. In this scenario, the curves of latent bias shift left. This is because with increasing x^*, $f(x^*;\theta^*)$ decreases and then the necessary value of the tuning parameter decreases to give the same weight value for an outlier.

6.2. Mean squared error

The root of the mean squared error (RMSE) was investigated for the density power weight. The sample size was set to be $n = 40$. The RMSE was estimated from 500 replications.

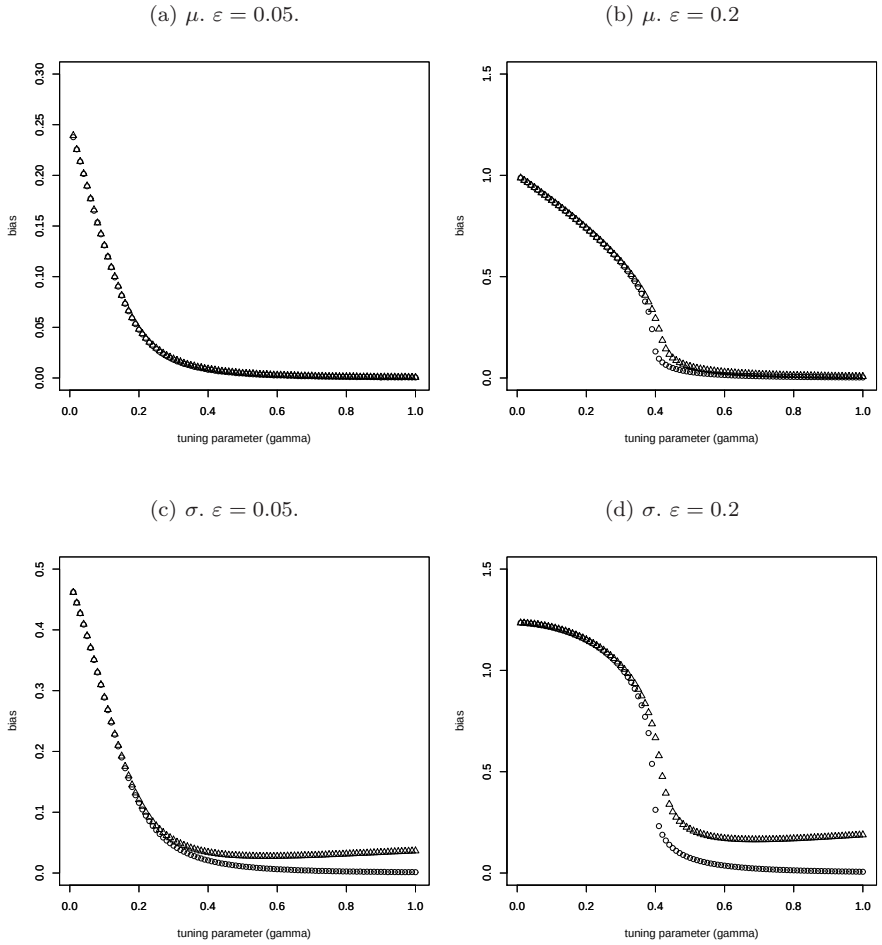


FIG 2. *Latent bias in the case of the normal distribution and the density power weight. The circles and triangles correspond to the latent biases for the normalized and non-normalized estimating equations, respectively.*

Figure 5 illustrates the RMSE for the normalized estimating equation. The trade-off between bias and variance was observed. The minimum of RMSE was attained at a certain value of tuning parameter. Let $\hat{\gamma}_\mu$ and $\hat{\gamma}_\sigma$ be the optimal tuning parameters that minimized the RMSE for the parameters μ and σ, respectively. It should be noted that the latent bias was small enough when the tuning parameter was $\hat{\gamma}_\mu$ or $\hat{\gamma}_\sigma$, as seen in Figure 2. The optimal tuning parameter would correspond to the case where the latent bias is small enough and the variance of the estimator is as small as possible. In addition, the value $\hat{\gamma}_\sigma$ was slightly larger from $\hat{\gamma}_\mu$. This might be because the latent bias for the standard deviation parameter was larger than that for the mean parameter.

Figure 6 depicts the RMSE for the non-normalized estimating equation. The minimum of RMSE for the non-normalized estimating equation is larger than

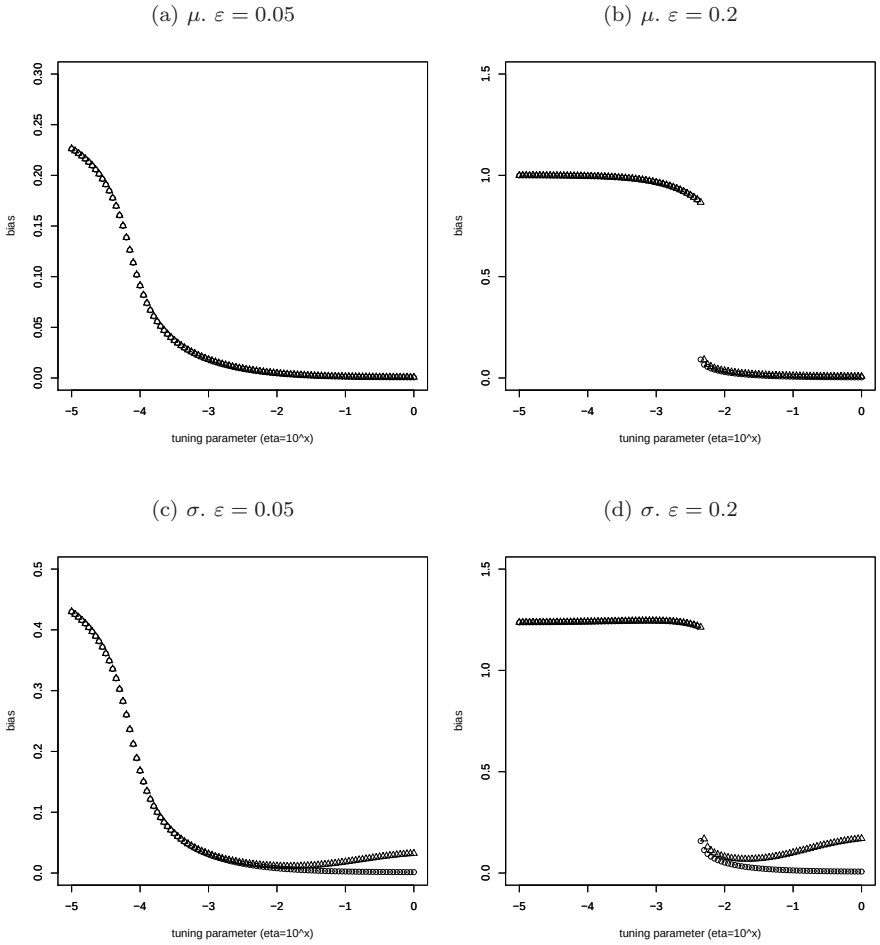


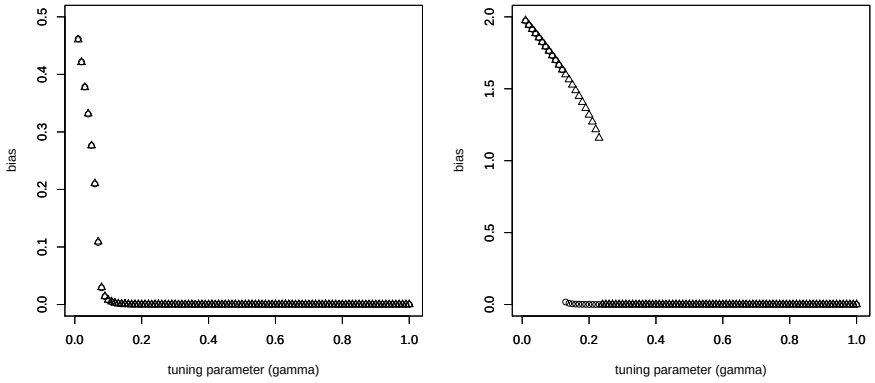
FIG 3. *Latent bias in the case of the normal distribution and the logistic weight. The circles and triangles correspond to the latent biases for the normalized and non-normalized estimating equations, respectively.*

that for the normalized estimating equation. The RMSE for the non-normalized estimating equation slowly increased with increasing γ after the minimum was attained.

7. Discussion

In this paper, a normalized estimating equation using a weighted score function was presented and compared with a non-normalized estimating equation. It was shown that the latent bias could be close to zero even if the proportion of outliers was not small. The latent bias and mean squared error were illustrated by some examples. In this section, the weight selection is further discussed.

(a) μ. $\varepsilon = 0.05$. (b) μ. $\varepsilon = 0.2$



(c) σ. $\varepsilon = 0.05$. (d) σ. $\varepsilon = 0.2$

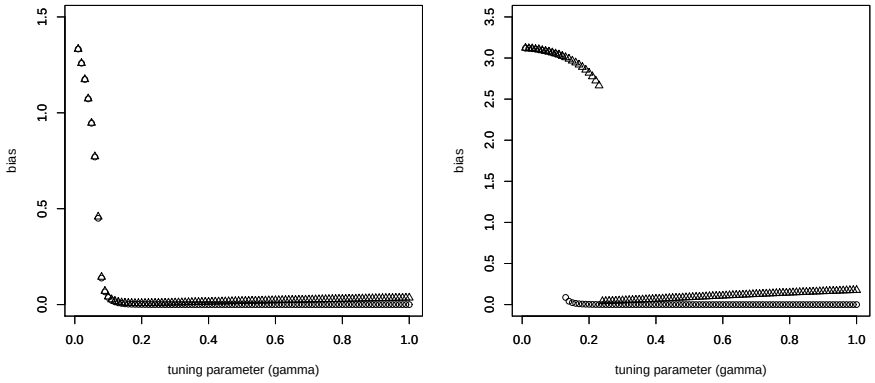


FIG 4. *Latent bias in the case of the normal distribution and the density power weight when the mean of contamination distribution is* $\mu_{\text{out}} = 10$. *The circles and triangles correspond to the latent biases for the normalized and non-normalized estimating equations, respectively.*

To obtain the robust estimate, we must set the tuning parameter. Remember that the tuning parameter controls the trade-off between bias and variance, as described in Section 6.2. As seen in Section 6.1, when the tuning parameter is larger than a certain value, the latent bias is close to zero, in other words, the estimate is close to the true value. Consider the set of tuning parameters that is larger than a certain value and show a similar estimate. In this set, a smaller value of tuning parameter would be favorable because the latent bias is close to zero and the variance is smaller. We can also use robust model selection criterion to select a good tuning parameter.

There are many candidates for the weight function. We might think what type of weight function is better. For example, among the weight functions satisfying

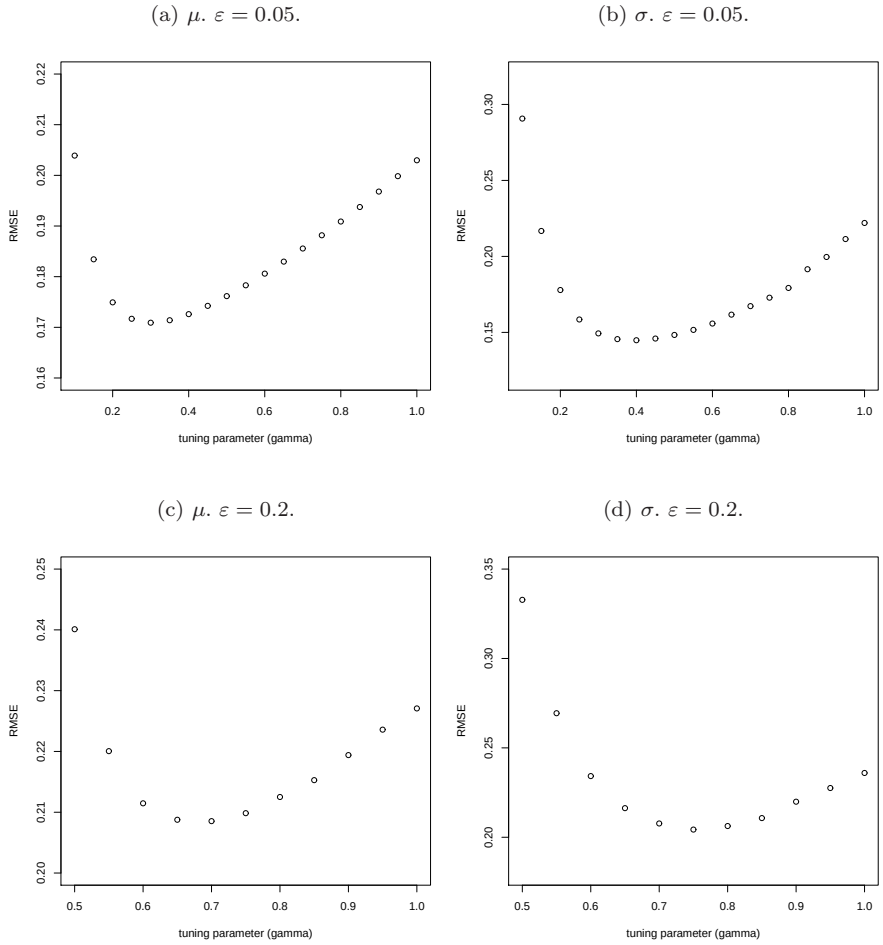


FIG 5. *RMSE for normalized estimating equation with density power weight.*

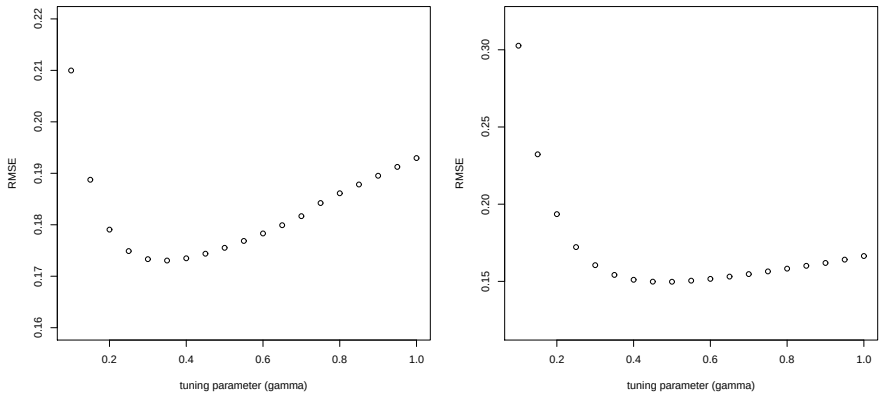
the assumption in Theorem 3.1, we might consider that a weight function with a smaller maximal bias is better. Note that this type of condition has not been assumed so far. This additional condition might imply a new problem about the optimality of maximal bias. This will be a future issue.

Appendix A: Examples satisfying the condition (3.1)

In this section, we consider the case where δ is the dirac function at x^* and illustrate that $\mathrm{E}_\delta[\xi(l(x;\theta^*))s(x;\theta^*)] = \xi(l(x^*;\theta^*))s(x^*;\theta^*)$ approaches zero as $|x^*|$ goes to infinity. This is enough to show the condition (3.1).

The density of the exponential distribution is $f(x;\theta) = \exp(-x/\theta)/\theta$. The log-likelihood function is $l(x;\theta) = -x/\theta - \log\theta$. The score function is $s(x;\theta) =$

(a) μ. $\varepsilon = 0.05$. (b) σ. $\varepsilon = 0.05$.



(c) μ. $\varepsilon = 0.2$. (d) σ. $\varepsilon = 0.2$.

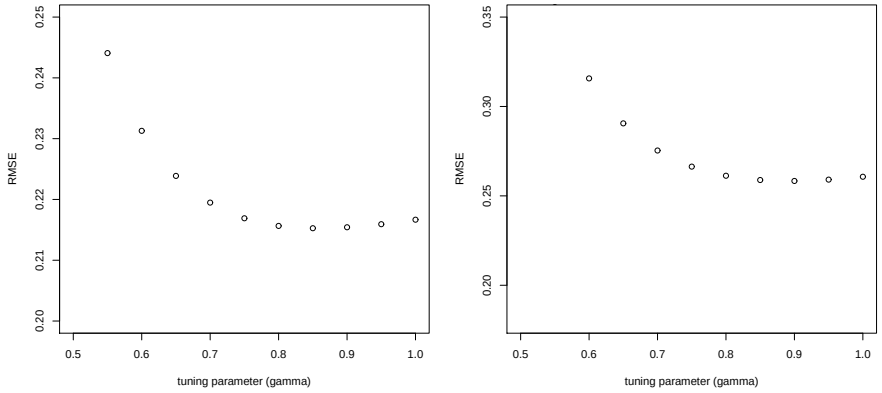


FIG 6. *RMSE for non-normalized estimating equation with density power weight.*

$-x/\theta^2 + 1/\theta$. In the case of the density power weight, we see that

$$f(x^*;\theta^*)^\gamma s(x^*;\theta^*) = \frac{1}{\theta^{*\gamma}} \exp\left(-\gamma \frac{x^*}{\theta^*}\right)\left(-\frac{x^*}{\theta^{*2}} + \frac{1}{\theta^*}\right),$$

which approaches zero as x^* goes to infinity.

The density of the normal distribution is $f(x;\theta) = \exp\{-(x-\mu)^2/2\sigma^2\}/\sqrt{2\pi\sigma^2}$ for $\theta = (\mu, \sigma)'$. The log-likelihood function is $l(x;\theta) = -(1/2)\log(2\pi) - (1/2)\log\sigma^2 - (x-\mu)^2/2\sigma^2$. The score function is $s(x;\theta) = \partial l/\partial\theta = (\partial l/\partial\mu, \partial l/\partial\sigma)'$, where $\partial l/\partial\mu = (x-\mu)/\sigma^2$ and $\partial l/\partial\sigma = -1/\sigma + (x-\mu)^2/\sigma^3$. In the case of the density power weight, we see that

$$f(x^*;\theta^*)^\gamma s(x^*;\theta^*) = \frac{1}{(2\pi\sigma^2)^{\gamma/2}} \exp\left\{-\frac{\gamma}{2\sigma^2}(x^*-\mu)^2\right\} s(x^*;\theta^*),$$

which approaches zero as x^* goes to infinity.

Similarly to the exponential and normal distributions, in particular, for an exponential family, it is often easy to prove that $f(x^*;\theta^*)^\gamma s(x^*;\theta^*)$ approaches zero as $|x^*|$ goes to infinity, because the density $f(x^*;\theta^*)$ approaches zero with exponential order but the score goes to infinity with polynomial order. When the tail order of the weight $\xi(l(x^*;\theta^*))$ is similar to that of the density power weight, we can also show that $\xi(l(x^*;\theta^*))s(x^*;\theta^*)$ approaches zero as $|x^*|$ goes to infinity. There are many examples of weight, including the logistic weight, $\log(1+f(x;\theta)^\gamma)$, $(\eta+f(x;\theta))^\gamma-\eta^\gamma$, and so on.

Appendix B: Proof of Theorem 3.1

From the implicit function theorem, there exists a sufficiently small $\eta>0$ and a closed set R_η such that $\lambda_{f_{\theta^*}}(\theta)$ is a homeomorphism from R_η to $B_\eta(\tau^*)$, where $\theta^*\in R_\eta$ and $\tau^*=\lambda_{f_{\theta^*}}(\theta^*)=0$. Prepare the function

$$h(\tau) \quad = \quad \tau-\frac{1}{1-\varepsilon}\lambda_g\left(\lambda^{-1}_{f_{\theta^*}}(\tau)\right)=\frac{\varepsilon}{1-\varepsilon}\lambda_\delta\left(\lambda^{-1}_{f_{\theta^*}}(\tau)\right).$$

The assumption implies that $\lambda_\delta(\lambda^{-1}_{f_{\theta^*}}(\tau))\in B_{\eta(1-\varepsilon)/\varepsilon}(0)$ for $\tau\in B_\eta(0)$, so that the function $h(\tau)$ maps from $B_\eta(0)$ into $B_\eta(0)$. By Brouwer's fixed point theorem, the function $h(\tau)$ has a fixed point τ_0 in $B_\eta(0)$, which implies that $\theta^*_N=\lambda^{-1}_{f_{\theta^*}}(\tau_0)$ is a root of $\lambda_g(\theta)=0$ from $h(\tau_0)=\tau_0$ and the formula of $h(\tau)$. For any small $\nu>0$, we can take η such that $R_\eta\subset B_\nu(\theta^*)$. We see that $\theta^*_N=\lambda^{-1}_{f_{\theta^*}}(\tau_0)\in\lambda^{-1}_{f_{\theta^*}}(B_\eta(0))=R_\eta\subset B_\nu(\theta^*)$. It is clear that $\theta^*_N(f_{\theta^*})=\theta^*$ since $\lambda_{f_{\theta^*}}(\theta)$ is a homeomorphism from R_η to $B_\eta(0)$.

Appendix C: Verification of assumption (d)

It holds that

$$\mathrm{E}_g\left[\frac{\partial\psi}{\partial\theta'}(\theta)\right]=(1-\varepsilon)\mathrm{E}_f\left[\frac{\partial\psi}{\partial\theta'}(\theta)\right]+\varepsilon\mathrm{E}_\delta\left[\frac{\partial\psi}{\partial\theta'}(\theta)\right].$$

Assume that $\mathrm{E}_f\left[\partial\psi/\partial\theta'(\theta^*)\right]$ is nonsingular, as seen in Theorem 3.1. It is easy to see that $\mathrm{E}_f\left[\partial\psi/\partial\theta'(\theta_N)\right]$ is nonsingular since $\theta_N\approx\theta^*$ from Theorem 3.1. Assume that $\mathrm{E}_\delta\left[\partial\psi/\partial\theta'(\theta^*)\right]\approx O$, which implies that $\mathrm{E}_\delta\left[\partial\psi/\partial\theta'(\theta_N)\right]\approx O$ from $\theta_N\approx\theta^*$. Therefore, it follows that $\mathrm{E}_g\left[\partial\psi/\partial\theta'(\theta_N)\right]$ is nonsingular.

Appendix D: Proof of Theorem 4.2

We see that

$$\begin{aligned}
J_g(\theta_N) &= \mathrm{E}_g\left[\partial\psi/\partial\theta'(\theta_N)\right]=(1-\varepsilon)J_f(\theta_N)+\varepsilon J_\delta(\theta_N)\\
&\approx (1-\varepsilon)J_f(\theta^*)+\varepsilon J_\delta(\theta^*)\approx(1-\varepsilon)J_f(\theta^*),\\
K_g(\theta_N) &= \mathrm{E}_g\left[\psi(\theta_N)\psi(\theta_N)'\right]=(1-\varepsilon)K_f(\theta_N)+\varepsilon K_\delta(\theta_N)\\
&\approx (1-\varepsilon)K_f(\theta^*)+\varepsilon K_\delta(\theta^*)\approx(1-\varepsilon)K_f(\theta^*).
\end{aligned}$$

The proof is complete.

Appendix E: Iterative algorithm

Suppose that the parametric density belongs to an exponential family, more precisely,

$$f(x;\theta) = \exp\{\theta' t(x) - \psi(\theta) + b(x)\}.$$

The score function is given by

$$s(x;\theta) = t(x) - \eta(\theta),$$

where $\eta(\theta) = \partial\psi(\theta)/\partial\theta$. The limit of the normalized estimating equation (3.2) becomes

$$\frac{\mathrm{E}_g\left[\xi(l(x;\theta))t(x)\right]}{\mathrm{E}_g\left[\xi(l(x;\theta))\right]} = \frac{\mathrm{E}_{f_\theta}\left[\xi(l(x;\theta))t(x)\right]}{\mathrm{E}_{f_\theta}\left[\xi(l(x;\theta))\right]}. \tag{E.1}$$

Note that the parameter η vanishes. To obtain the root, we propose the iterative algorithm given by

$$\frac{\mathrm{E}_g\left[\xi(l(x;\theta^{(a)}))t(x)\right]}{\mathrm{E}_g\left[\xi(l(x;\theta^{(a)}))\right]} = \frac{\mathrm{E}_{f_{\theta^{(a+1)}}}\left[\xi(l(x;\theta^{(a+1)}))t(x)\right]}{\mathrm{E}_{f_{\theta^{(a+1)}}}\left[\xi(l(x;\theta^{(a+1)}))\right]}. \tag{E.2}$$

Note that this type of algorithm was proposed for the density power weight by Fujisawa and Eguchi (2008) and the corresponding iterative algorithm monotonously increases the γ-cross entropy at each step.

Here we suppose that the parametric density is normal with mean μ and variance σ^2. The sufficient statistic is $t(x) = (x, x^2)$. In the case of the density power weight, the iterative algorithm was obtained by Fujisawa and Eguchi (2008), given by

$$\begin{aligned}
\mu^{(a+1)} &= \mathrm{E}_g\left[x\, f(x;\theta^{(a)})^\gamma\right] \Big/ \mathrm{E}_g\left[f(x;\theta^{(a)})^\gamma\right], \\
(\sigma^2)^{(a+1)} &= \left\{\mathrm{E}_g\left[x^2\, f(x;\theta^{(a)})^\gamma\right] \Big/ \mathrm{E}_g\left[f(x;\theta^{(a)})^\gamma\right] - (\mu^{(a+1)})^2\right\}(1+\gamma).
\end{aligned}$$

The empirical type can be obtained by replacing E_g by the sample mean. A typical initial value is the median and the median absolute value. Next, we consider the case of the logistic weight. Let $\phi(x)$ be the standard normal density. Let

$$B_j(\sigma) = \int \frac{\phi(y)}{\phi(y) + \eta\sigma} y^j \phi(y)\, dy \qquad \text{for } j = 0, 1, 2.$$

Note that $B_1(\sigma) = 0$ because $\phi(y)$ is an even function of y. The first component of the numerator of the right-hand side of (E.1) is

$$\begin{aligned}
\mathrm{E}_{f_\theta}\left[\xi(l(x;\theta))x\right] &= \int \frac{\phi((x-\mu)/\sigma)/\sigma}{\phi((x-\mu)/\sigma)/\sigma + \eta} x \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) dx \\
&= \int \frac{\phi(y)}{\phi(y) + \eta\sigma}(\mu + \sigma y)\phi(y)\, dy \\
&= \mu B_0(\sigma).
\end{aligned}$$

The denominator of the right-hand side of (E.1) is $B_0(\sigma)$. Hence, from (E.2), the iterative algorithm for the mean parameter μ is given by

$$\mu^{(a+1)} = \mathrm{E}_g\left[\xi(l(x;\theta^{(a)}))\, x\right] \Big/ \mathrm{E}_g\left[\xi(l(x;\theta^{(a)}))\right].$$

The second component of the numerator of the right-hand side of (E.1) is

$$\begin{aligned}\mathrm{E}_{f_\theta}\left[\xi(l(x;\theta))x^2\right] &= \int \frac{\phi((x-\mu)/\sigma)/\sigma}{\phi((x-\mu)/\sigma)/\sigma+\eta} x^2 \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) dx \\ &= \int \frac{\phi(y)}{\phi(y)+\eta\sigma}(\mu+\sigma y)^2 \phi(y)\, dy \\ &= \mu^2 B_0(\sigma) + \sigma^2 B_2(\sigma).\end{aligned}$$

Thus, from (E.2), $\sigma^{(a+1)}$ is a unique root of

$$\left(\mu^{(a+1)}\right)^2 + \sigma^2 B_2(\sigma)/B_0(\sigma) = \mathrm{E}_g\left[\xi(l(x;\theta^{(a)}))\, x^2\right] \Big/ \mathrm{E}_g\left[\xi(l(x;\theta^{(a)}))\right].$$

The uniqueness is easily shown since $\sigma^2 B_2(\sigma)/B_0(\sigma)$ is monotone increasing.

Acknowledgement

The author really thanks the Editor, Associate Editor and two reviewers for their constructive comments. In particular, Section 3 was much improved by their comments.

References

Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85** 549–559. MR1665873

Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *J. Amer. Statist. Assoc.* **96** 1022–1030. MR1947250

Cichocki, A. and Amari, S. (2010). Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy* **12** 1532–1568. MR2659408

Copt, S. and Heritier, S. (2007). Robust alternatives to the F-test in mixed linear models based on MM-estimates. *Biometrics* **63** 1045–1052. MR2414581

Croux, C., Gijbels, I. and Prosdocimi, I. (2012). Robust estimation of mean and dispersion functions in extended generalized additive models. *Biometrics* **68** 31–44. MR2909851

Croux, C. and Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Comput. Statist. Data Anal.* **44** 273–295. MR2020151

Eguchi, S. and Kano, Y. (2001). Robusting maximum likelihood estimation by psi-divergence. *Research Memorandam of the Institute of Statistical Mathematics* No.802.

Eguchi, S. and Kato, S. (2010). Entropy and divergence associated with power function and the statistical application. *Entropy* **12** 262–274. MR2608236

Eguchi, S., Komori, O. and Kato, S. (2011). Projective Power Entropy and Maximum Tsallis Entropy Distributions. *Entropy* **13** 1746–1764. MR2851127

Ferrari, D. and La Vecchia, D. (2012). On robust estimation via pseudo-additive information. *Biometrika* **99** 238–244. MR2899677

Ferrari, D. and Yang, Y. (2010). Maximum Lq-likelihood estimation. *Ann. Statist.* **38** 753–783. MR2604695

Field, C. and Smith, B. (1994). Robust estimation: A weighted maximum likelihood approach. *Internat. Statist. Rev.* **62** 405–424.

Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *J. Multivariate Anal.* **99** 2053–2081. MR2466551

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics*. John Wiley & Sons Inc., New York. MR0829458

Heritier, S. and Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models. *J. Amer. Statist. Assoc.* **89** 897–904. MR1294733

Huang, S.-Y., Yeh, Y.-R. and Eguchi, S. (2009). Robust kernel principal component analysis. *Neural Comput.* **21** 3179–3213. MR2604317

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*, Second ed. John Wiley & Sons Inc., Hoboken, NJ. MR2488795

Jones, M. C., Hjort, N. L., Harris, I. R. and Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika* **88** 865–873. MR1859416

Lee, A. J. (1990). *U-Statistics*. Marcel Dekker Inc., New York. MR1075417

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. John Wiley & Sons Inc., New York. MR2238141

Mattheou, K., Lee, S. and Karagrigoriou, A. (2009). A model selection criterion based on the BHHJ measure of divergence. *J. Statist. Plann. Inference* **139** 228–235. MR2474000

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. Chapman & Hall, London. MR0727836

Minami, M. and Eguchi, S. (2002). Robust blind source separation by beta-divergence. *Neural Comput.* **14** 1859–1886.

Miyamura, M. and Kano, Y. (2006). Robust Gaussian graphical modeling. *J. Multivariate Anal.* **97** 1525–1550. MR2275418

Mollah, M. N. H., Minami, M. and Eguchi, S. (2006). Exploring latent structure of mixture ICA models by the minimum β-divergence method. *Neural Comput.* **18** 166–190.

Murata, N., Takenouchi, T., Kanamori, T. and Eguchi, S. (2004). Information geometry of U-Boost and Bregman divergence. *Neural Comput.* **16** 1437–1481.

Scott, D. W. (2001). Parametric statistical modeling by minimum integrated square error. *Technometrics* **43** 274–285. MR1943184

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons Inc., New York. MR0595165

Takenouchi, T. and Eguchi, S. (2004). Robustifying AdaBoost by adding the naive error rate. *Neural Comput.* **16** 767–787.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *J. Statist. Physics* **52** 479–487. MR0968597

Tsallis, C. (2009). *Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World*. Springer, New York. MR2724662

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge. MR1652247

Villmann, T. and Haase, S. (2011). Divergence-based vector quantization. *Neural Comput.* **23** 1–50. MR2814848

Windham, M. P. (1995). Robustifying model fitting. *J. Roy. Statist. Soc. Ser. B* **57** 599–609. MR1341326