Electronic Journal of Statistics
Vol. 15 (2021) 6544–6585
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1947

Regression using localised functional Bregman divergence[*]

Kanta Naito[†]

Department of Mathematics and Informatics, Chiba University
263-8522 Chiba, Japan
e-mail: naito@math.s.chiba-u.ac.jp

and

Spiridon Penev

Department of Statistics, UNSW Sydney and UNSW Data Science Hub
Sydney 2052 NSW, Australia
e-mail: s.penev@unsw.edu.au

Abstract: This paper is concerned with a unified approach to estimating regression methods based on a certain divergence and its localisation. Some past papers have demonstrated theoretically and numerically that infusing a little localisation in the likelihood-based methods for regression and for density estimation can actually improve the resulting estimators with respect to suitably defined global risk measures. Thus a variety of local likelihood methods have been suggested. We demonstrate that similar effect can also be observed in the general framework discussed in this paper and with respect to robust estimation procedures. Localised versions of robust regression estimation procedures perform better with respect to global risk measures based on minimisation of Bregman divergence measures. An intricate relationship between regression model's inadequacy and its robustness can be better analysed by using the local approach developed in this paper. We support our claims with a short simulation study.

MSC2020 subject classifications: Primary 62J02, 62J12; secondary 62E20.
Keywords and phrases: Bregman divergence, kernel, localisation, regression, risk, robustness.

Received March 2021.

1. Introduction

This paper discusses a divergence-based method for estimating a regression function and a local version of this method. The smoothing approaches using localisation have been discussed in literature: most of those are based on certain form of local likelihood, see [20], [7], [15], [8], [13], [9] and references therein. In this paper we aim to develop a general framework of localised regression which includes methods based on the local likelihood as special cases. The proposed

[*] The first author was supported in part by KAKENHI 19K11851. The second author was supported by Australian Research Council Discovery Project DP 160103489.

[†] The corresponding author.

framework naturally induces a robust version. The estimation scheme in this paper is composed via functional Bregman divergence composed by a strictly convex function ([4], [23]), while the parametric model for regression function is chosen simply by exploiting a composition of linear predictor and a smooth link function as in the generalized linear model ([16], [23] and [17]). The localisation is applied in the estimation scheme by slotting a kernel function, as in [20]. Hence the proposed localised regression inference in this paper is essentially composed by choosing the strictly convex function U in the functional Bregman divegence, the link function G in the model, and the kernel function K for the localisation. We claim that an appropriate choice of U and G leads to asymptotic risk improvement by the localised estimator over the global (non-localised) estimator. Furthermore we show that choosing a suitable U in the functional Bregman divergence naturally induces a robust version of the resulting localised regression, with a similar risk improvement by the localised estimator when an appropriate choice of G is made.

In a recent paper [18] we demonstrate theoretically and numerically that infusing a little localisation in a *robust* parametric density estimation procedure can bring benefits when the quality of the density estimator is measured by using a global risk measure based on a minimisation of a Bregman divergence measure. In this way, an extension in robustness context of a past results has been delivered. These past results ([15], [13], [8]) have demonstrated similar effects when localising *likelihood-based* methods. Specifically, they show that a little localisation helps to improve the non-localised estimators when the quality of the estimator is measured by using global risk measures.

In regression setting, the idea of localisation of the inference was pioneered in [20]. It was presented as an extension of the idea of scatterplot smoothing [6] to generalized linear regression models. To the best of our knowledge, there have not been efforts to apply the *localisation* approach to *robustify* inference in regression, neither for the simple linear regression setting, nor for the generalized linear regression models. Our paper represents and attempt to fill this gap. *However*, we stress on the fact that in the regression setting of the current paper, even the global estimator (that we then localise for better performance) suggested in this paper, seems to be new to the best of our knowledge.

It is well-known that likelihood-based inference is based on the idea of the minimisation of the Kullback-Leibler (KL) divergence between an ideal and empirical distribution. We illustrate below that the use of *Bregman divergence* (BD) as a generalization and replacement of the KL divergence can, when properly applied, bring about intrinsic robustification to standard likelihood-based inference when the quality of the regression fit is measured by using a suitable global risk measure. Further point of this paper is that when coupled with a suitable localisation, this robustification effect can be magnified, in a way similar to the one that has been demonstrated for density estimation in [18].

From the very beginning we stress that the general setting of our paper does allow for model misspecification. That is, we allow for the possibility that the regression function, defined as the conditional expected value of the output variable given the input vector, may not be equal to any of the parametric relationships used to model this function. Admittedly, this is the more realistic scenario in practice. Intuitively, if we are ready to admit (parametric) model misspecification then it is to be expected that a localisation would be helpful in improving the performance with respect to a global risk measure as a local estimator would be better adaptable to the unknown regression function. We give sufficient conditions for this effect to happen asymptotically by relating the choice of the function G to the choice of U.

The paper is structured as follows. We start with a detailed discussion of the setup of our paper in Section 2. The global estimation scheme is firstly introduced using the functional Bregman divergence [11] using the setup discussed in [23]. Then the localisation is naturally composed by exploiting the kernel function. Section 3 is devoted to asymptotic statements about the global and local estimators discussed in our framework. The risk difference between the global estimator and the local estimator is asymptotically evaluated in Section 3.3. The choice of U and G to yield the asymptotic risk improvement by the local estimator is addressed in Section 3.4. Section 4 discusses approaches to robustify our estimators by choosing a suitable divergence measure. It is seen in Section 4.1 that using a suitable U and applying it on the residuals helps us to robustify the method discussed in the previous sections. The advantage of localisation in the sense of the risk improvement is also observed in the robustness setting in Section 4.3. Numerical implementations and simulations are presented in Section 5. Some final discussions are presented in Section 6. Section 7 presents the proofs of our theoretical results.

2. Setup and estimators

2.1. Background

Bregman divergence was introduced in [4]. It can be defined for d-dimensional $(d \geq 1)$ vectors but also for matrices or for functions. For a given strictly convex function $U : A \to \mathbb{R}$, where $A \subset \mathbb{R}^d$ is a convex set, the Bregman divergence between two points $X \in A$ and $Y \in A$ is defined as

$$d_U(Y, X) = U(Y) - U(X) - \nabla U(X)(Y - X),$$

with ∇ denoting gradient-taking. This definition can be applied point wise for positive density functions f, g defined on a common domain. The point-wise application means that in this case $d = 1$, ∇ means a simple derivative U' and we interpret locally, for a fixed t

$$d_U(g(t), f(t)) = U(g(t)) - U(f(t)) - U'(f(t))\{g(t) - f(t)\}.$$

Using this localised divergence measure at the point t, we then define the global (or also called functional) Bregman divergence between the densities f and g:

$$D_U(g, f) := \int d_U(g(t), f(t))v(t)dt, \tag{2.1}$$

where v is some non-negative weight function.

Suppose an exponential family of distributions in canonical form

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\{\boldsymbol{\theta}^T \boldsymbol{x} - \psi(\boldsymbol{\theta})\}$$

is given, where $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{\theta} \in \mathbb{R}^d$ is an unknown parameter vector, with the convex function $\psi(\boldsymbol{\theta})$ being the cumulant generating function. There exists an intimate connection between such a family of distributions and the Bregman divergence. The latter is usually discussed via the notion of *dually flat Riemannian structure* [1] introduced in this family by using the function $\psi(\boldsymbol{\theta})$. This function uniquely characterizes the distribution $f(\boldsymbol{x}, \boldsymbol{\theta})$. In more details, the canonical parameter vector $\boldsymbol{\theta}$ of this exponential family is used as affine coordinate system. For two members of this exponential family, $f(\boldsymbol{x}, \boldsymbol{\theta})$ and $f(\boldsymbol{x}, \boldsymbol{\theta}')$, say, Kullback-Leibler (KL) divergence $D_{KL}(\boldsymbol{\theta}, \boldsymbol{\theta}')$ can be written as

$$\begin{aligned} D_{KL}(\boldsymbol{\theta}, \boldsymbol{\theta}') &= \int_{\mathbb{R}^d} f(\boldsymbol{x}, \boldsymbol{\theta}) \log\left(\frac{f(\boldsymbol{x}, \boldsymbol{\theta})}{f(\boldsymbol{x}, \boldsymbol{\theta}')}\right) d\boldsymbol{x} \\ &= \psi(\boldsymbol{\theta}') - \psi(\boldsymbol{\theta}) - \frac{\partial}{\partial \boldsymbol{\theta}} \psi(\boldsymbol{\theta})^T (\boldsymbol{\theta}' - \boldsymbol{\theta}) \\ &= D_\psi(\boldsymbol{\theta}', \boldsymbol{\theta}). \end{aligned}$$

That is to say the KL divergence in this exponential family in canonical form is tantamount the Bregman divergence defined via the convex function $\psi(\boldsymbol{\theta})$. In addition the Riemannian metric $(\partial^2/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T)\psi(\boldsymbol{\theta})$ derived from $\psi(\boldsymbol{\theta})$ in this exponential family is easily seen to equal the Fisher information matrix (compare also Theorem 2.1 in [1]). This intimate relation with KL divergence has been an important push to consider applications of Bregman divergence also outside the exponential family setting. There is still the belief that the data's distribution is "close to exponential family" but may "deviate slightly" as in the robustness paradigm. In such cases it is prudent to start from the very beginning with a proposal of a convex function U in (2.1) and proceed with it. The above divergence between parameters will be replaced by functional Bregman divergence (see [11]) when analysing the quality of the fit of such models to data.

2.2. Setup

Let $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n) \sim_{i.i.d.} f(y, \boldsymbol{x}) = p(y|\boldsymbol{x})q(\boldsymbol{x})$, where $(Y_i, \boldsymbol{X}_i) \in \mathbb{R} \times \mathbb{R}^d$, f is the joint density of $(Y, \boldsymbol{X})$, $p(\cdot|\boldsymbol{x})$ is the conditional density of Y given $\boldsymbol{X} = \boldsymbol{x}$, and q is the density of $\boldsymbol{X}$. The support of the density q is expressed as $\mathbb{D}$ which is assumed to be a compact set in $\mathbb{R}^d$.

Let $\boldsymbol{t} \in \mathbb{D} \subset \mathbb{R}^d$ be a target point at which we want to estimate the value of regression function $\mu(\boldsymbol{t}) = E[Y|\boldsymbol{X} = \boldsymbol{t}]$. Given that the precise distribution of the pairs $(Y_i, \boldsymbol{X}_i)$ is virtually never known in practice, there is a need to approximate this ultimate function $\mu(\boldsymbol{t}) = E[Y|\boldsymbol{X} = \boldsymbol{t}]$ and a simple well-known approach is to approximate certain (possibly nonlinear transformation of it) via a function that is linear in the input observations $\boldsymbol{x}$. As usually done in practice, one postulates a parametric model for μ in the form

$$m(\boldsymbol{x}, \boldsymbol{\theta}) = G^{-1}\left(\boldsymbol{\theta}^T \widetilde{\boldsymbol{x}}\right), \tag{2.2}$$

where $\widetilde{\boldsymbol{x}}^T = [1\ \boldsymbol{x}^T] \in \mathbb{R}^{d+1}$ is the vector of explanatory variables and $\boldsymbol{\theta} = [\theta_0\ \theta_1\ \cdots\ \theta_d]^T \in \Theta \subset \mathbb{R}^{d+1}$ is the parameter vector that also includes the component θ_0 as an intercept. The one-to-one transformation G (whose inverse is denoted as G^{-1} above) represents the so-called link function.

As said in the Introduction, we allow for model misspecification. That is, we allow for the possibility that for all $\boldsymbol{\theta} \in \Theta$, the relationship $m(\boldsymbol{x}, \boldsymbol{\theta}) \neq \mu(\boldsymbol{x})$ may hold. This general setting is sometimes called approgression [5]. For the particular case where model misspecification is excluded, some of our results in this paper coincide with results that have been obtained in [23].

2.3. Bregman divergence

Throughout this paper we denote by $\mathscr{U}$ the set of strictly convex functions on $\mathbb{R}$. Now we fix a $U \in \mathscr{U}$. Then the discrepancy between $\mu(\cdot)$ and its parametric model $m(\cdot, \boldsymbol{\theta}) = m_\theta(\cdot)$ can be measured by the *functional* Bregman divergence defined as

$$\begin{aligned} & D_U(\mu, m_\theta) \\ & = \int_{\mathbb{R}^d} \left[U(\mu(\boldsymbol{x})) - U(m(\boldsymbol{x}, \boldsymbol{\theta})) - u(m(\boldsymbol{x}, \boldsymbol{\theta}))\{\mu(\boldsymbol{x}) - m(\boldsymbol{x}, \boldsymbol{\theta})\}\right] q(\boldsymbol{x}) d\boldsymbol{x}, \end{aligned} \tag{2.3}$$

where $u = U'$: the derivative of U.The reason to add the term *functional* to the Bregman divergence above is that it can be interpreted as a weighted form of the point-wise Bregman divergence between $\mu(\boldsymbol{x})$ and $m_\theta(\boldsymbol{x})$, with a weight function given by the density of $\boldsymbol{X}$.

In what follows, for the purpose of easier tractability, we prefer to utilize

$$\begin{aligned} & D_{U^*}(u(m_\theta), u(\mu)) \\ & = \int_{\mathbb{R}^d} \left[U^*(u(m(\boldsymbol{x}, \boldsymbol{\theta}))) - U^*(u(\mu(\boldsymbol{x})))\right. \qquad (2.4) \\ & \qquad \left. - \mu(\boldsymbol{x})\{u(m(\boldsymbol{x}, \boldsymbol{\theta})) - u(\mu(\boldsymbol{x}))\}\right] q(\boldsymbol{x}) d\boldsymbol{x} \\ & = \int_{\mathbb{R} \times \mathbb{R}^d} \left[U^*(u(m(\boldsymbol{x}, \boldsymbol{\theta}))) - y \cdot u(m(\boldsymbol{x}, \boldsymbol{\theta}))\right] f(y, \boldsymbol{x}) dy d\boldsymbol{x} \qquad (2.5) \\ & \qquad + \int_{\mathbb{R}^d} \left[-U^*(u(\mu(\boldsymbol{x}))) + \mu(\boldsymbol{x}) u(\mu(\boldsymbol{x}))\right] q(\boldsymbol{x}) d\boldsymbol{x}, \end{aligned}$$

where U^* is the convex conjugate of U: $U^*(s) = \sup_{z \in \mathbb{R}}\{zs - U(z)\}$, and we have used a fact that $(U^*)' = u^{-1}$ in (2.4). The fact that

$$D_U(\mu, m_\theta) = D_{U^*}(u(m_\theta), u(\mu)) \tag{2.6}$$

is an easy consequence of the fundamental properties of the Legendre transformations. It has been derived explicitly, for example, in [1] (Equation 1.68 on page 17). The equality (2.6) implies that minimising either quantity with respect to $\boldsymbol{\theta}$ would deliver the same outcome hence we can be guided by the simplicity of the resulting optimization problem when making a choice. Up to terms not involving $\boldsymbol{\theta}$, it is clear that $D_{U^*}(u(m_\theta), u(\mu))$ is more tractable and we focus on this choice from now on.

The usual parametric regression can be carried out by using a certain estimator $\hat{\boldsymbol{\theta}}$ of the true value of $\boldsymbol{\theta}$. Necessary tools for this estimation scheme are

$$\begin{aligned} \rho(y, \boldsymbol{x}, \boldsymbol{\theta}) &= U^*(u(m(\boldsymbol{x}, \boldsymbol{\theta}))) - y \cdot u(m(\boldsymbol{x}, \boldsymbol{\theta})), && (2.7) \\ \psi(y, \boldsymbol{x}, \boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}} \rho(y, \boldsymbol{x}, \boldsymbol{\theta}) = -\{y - m(\boldsymbol{x}, \boldsymbol{\theta})\} \frac{u'(m(\boldsymbol{x}, \boldsymbol{\theta}))}{G'(m(\boldsymbol{x}, \boldsymbol{\theta}))} \widetilde{\boldsymbol{x}}, && (2.8) \end{aligned}$$

by which the estimator $\hat{\boldsymbol{\theta}}$ and the true value $\boldsymbol{\theta}_*$ of $\boldsymbol{\theta}$ can be defined as

$$\begin{aligned} \boldsymbol{\theta}_* &= \arg\min_{\theta \in \Theta} \int_{\mathbb{R} \times \mathbb{R}^d} \rho(y, \boldsymbol{x}, \boldsymbol{\theta}) dF(y, \boldsymbol{x}) \\ &= \text{solution}_{\theta \in \Theta} \left[\int_{\mathbb{R} \times \mathbb{R}^d} \psi(y, \boldsymbol{x}, \boldsymbol{\theta}) dF(y, \boldsymbol{x}) = \mathbf{0}_{d+1} \right], \qquad (2.9) \end{aligned}$$

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg\min_{\theta \in \Theta} \int_{\mathbb{R} \times \mathbb{R}^d} \rho(y, \boldsymbol{x}, \boldsymbol{\theta}) dF_n(y, \boldsymbol{x}) \\ &= \text{solution}_{\theta \in \Theta} \left[\int_{\mathbb{R} \times \mathbb{R}^d} \psi(y, \boldsymbol{x}, \boldsymbol{\theta}) dF_n(y, \boldsymbol{x}) = \mathbf{0}_{d+1} \right], \qquad (2.10) \end{aligned}$$

where F_n is the empirical distribution function based on $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$, F is the cumulative distribution function with its density f and $\mathbf{0}_{d+1}$ is the zero vector in $\mathbb{R}^{d+1}$.

Note that $\boldsymbol{\theta}_*$ in (2.9) is the minimiser of (2.5), and the minimiser of the empirical version of (2.5) is precisely $\hat{\boldsymbol{\theta}}$ in (2.10). The regression function estimator can be obtained by substituting $\hat{\boldsymbol{\theta}}$ into $\boldsymbol{\theta}$ in $m(\cdot, \boldsymbol{\theta})$:

$$\hat{\mu}_G(\boldsymbol{x}) = m(\boldsymbol{x}, \hat{\boldsymbol{\theta}}). \qquad (2.11)$$

2.4. Local Bregman divergence

Now we introduce a localisation of the Bregman divergence by slotting a kernel function K, where $K(\boldsymbol{z})$ is a smooth unimodal integrable function, symmetric around $\boldsymbol{z} = \mathbf{0}_d$ and satisfying $K(\mathbf{0}_d) = 1$. Our proposed local version of the Bregman divergence corresponding to (2.4) is defined as

$$D_{U^*}(u(m_\theta), u(\mu) | \boldsymbol{t})$$

$$= \int_{\mathbb{R}\times\mathbb{R}^d} K\left(\frac{\boldsymbol{x}-\boldsymbol{t}}{h}\right)\left[U^*(u(m(\boldsymbol{x},\boldsymbol{\theta}))) - y\cdot u(m(\boldsymbol{x},\boldsymbol{\theta}))\right] f(y,\boldsymbol{x})dyd\boldsymbol{x}$$
$$+ \int_{\mathbb{R}^d} K\left(\frac{\boldsymbol{x}-\boldsymbol{t}}{h}\right)\left[-U^*(u(\mu(\boldsymbol{x}))) + \mu(\boldsymbol{x})u(\mu(\boldsymbol{x}))\right] q(\boldsymbol{x})d\boldsymbol{x}.$$

Here $h > 0$ is the scalar bandwidth which controls the degree of localisation. This local divergence aims to evaluate the discrepancy between $u(m(\cdot,\boldsymbol{\theta}))$ and $u(\mu(\cdot))$ locally around the evaluation point $\boldsymbol{t}$.

For a better adaptation along the regression curve, we now allow the parameter $\boldsymbol{\theta}$ to vary with $\boldsymbol{t}$ and suggest a scheme to estimate $\boldsymbol{\theta}$ depending on $\boldsymbol{t}$. The necessary functions are listed as follows:

$$\rho(\boldsymbol{t},y,\boldsymbol{x},\boldsymbol{\theta}) = K\left(\frac{\boldsymbol{x}-\boldsymbol{t}}{h}\right)\rho(y,\boldsymbol{x},\boldsymbol{\theta}), \tag{2.12}$$
$$\psi(\boldsymbol{t},y,\boldsymbol{x},\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}\rho(\boldsymbol{t},y,\boldsymbol{x},\boldsymbol{\theta}). \tag{2.13}$$

We note that (2.12) and (2.13) are localised version of (2.7) and (2.8) respectively, with the use of the kernel K. Using these functions, we define the true parameter $\boldsymbol{\theta}_*(\boldsymbol{t})$ at $\boldsymbol{t}$ and its estimator $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$ as follows:

$$\begin{aligned}\boldsymbol{\theta}_*(\boldsymbol{t}) &= \arg\min_{\theta\in\Theta}\int_{\mathbb{R}\times\mathbb{R}^d}\rho(\boldsymbol{t},y,\boldsymbol{x},\boldsymbol{\theta})dF(y,\boldsymbol{x})\\ &= \text{solution}_{\theta\in\Theta}\left[\int_{\mathbb{R}\times\mathbb{R}^d}\psi(\boldsymbol{t},y,\boldsymbol{x},\boldsymbol{\theta})dF(y,\boldsymbol{x}) = \mathbf{0}_{d+1}\right], \qquad (2.14)\\ \hat{\boldsymbol{\theta}}(\boldsymbol{t}) &= \arg\min_{\theta\in\Theta}\int_{\mathbb{R}\times\mathbb{R}^d}\rho(\boldsymbol{t},y,\boldsymbol{x},\boldsymbol{\theta})dF_n(y,\boldsymbol{x})\\ &= \text{solution}_{\theta\in\Theta}\left[\int_{\mathbb{R}\times\mathbb{R}^d}\psi(\boldsymbol{t},y,\boldsymbol{x},\boldsymbol{\theta})dF_n(y,\boldsymbol{x}) = \mathbf{0}_{d+1}\right]. \qquad (2.15)\end{aligned}$$

This local estimator $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$ of $\boldsymbol{\theta}_*(\boldsymbol{t})$ also suggests us to make a regression estimator defined as

$$\hat{\mu}_L(\boldsymbol{x}) = m(\boldsymbol{x},\hat{\boldsymbol{\theta}}(\boldsymbol{x})), \tag{2.16}$$

which we call the *local estimator* of $\mu(\boldsymbol{x})$, because the involved estimator of parameter is determined locally. Since $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ can vary depending on $\boldsymbol{x}$, $\hat{\mu}_L$ would be expected to be more flexible than $\hat{\mu}_G$. On the other hand, we call the estimator $\hat{\mu}_G(\boldsymbol{x})$ in (2.11) the *global estimator* of $\mu(\boldsymbol{x})$.

Let us introduce the notation

$$\psi^{(\ell)}(\boldsymbol{t},y,\boldsymbol{x},\boldsymbol{\theta}) = ||\boldsymbol{x}-\boldsymbol{t}||^\ell\psi(y,\boldsymbol{x},\boldsymbol{\theta}),\ \ \ell = 2,4, \tag{2.17}$$

We also define the following $(d+1)\times(d+1)$ matrices

$$\Psi(\boldsymbol{\theta}) = \int_{\mathbb{R}\times\mathbb{R}^d}\frac{\partial}{\partial\boldsymbol{\theta}}\psi(y,\boldsymbol{x},\boldsymbol{\theta})^T dF(y,\boldsymbol{x}), \tag{2.18}$$

$$\widehat{\Psi}(\boldsymbol{\theta}) = \int_{\mathbb{R}\times\mathbb{R}^d} \frac{\partial}{\partial\boldsymbol{\theta}}\psi(y,\boldsymbol{x},\boldsymbol{\theta})^T dF_n(y,\boldsymbol{x}), \tag{2.19}$$

$$\Psi(\boldsymbol{t},\boldsymbol{\theta}) = \int_{\mathbb{R}\times\mathbb{R}^d} \frac{\partial}{\partial\boldsymbol{\theta}}\psi(\boldsymbol{t},y,\boldsymbol{x},\boldsymbol{\theta})^T dF(y,\boldsymbol{x}), \tag{2.20}$$

$$\widehat{\Psi}(\boldsymbol{t},\boldsymbol{\theta}) = \int_{\mathbb{R}\times\mathbb{R}^d} \frac{\partial}{\partial\boldsymbol{\theta}}\psi(\boldsymbol{t},y,\boldsymbol{x},\boldsymbol{\theta})^T dF_n(y,\boldsymbol{x}), \tag{2.21}$$

$$\Psi^{(\ell)}(\boldsymbol{t},\boldsymbol{\theta}) = \int_{\mathbb{R}\times\mathbb{R}^d} \frac{\partial}{\partial\boldsymbol{\theta}}\psi^{(\ell)}(\boldsymbol{t},y,\boldsymbol{x},\boldsymbol{\theta})^T dF(y,\boldsymbol{x}), \ell=2,4, \tag{2.22}$$

$$\widehat{\Psi}^{(\ell)}(\boldsymbol{t},\boldsymbol{\theta}) = \int_{\mathbb{R}\times\mathbb{R}^d} \frac{\partial}{\partial\boldsymbol{\theta}}\psi^{(\ell)}(\boldsymbol{t},y,\boldsymbol{x},\boldsymbol{\theta})^T dF_n(y,\boldsymbol{x}), \ell=2,4. \tag{2.23}$$

The matrix (2.18) plays a role similar to the Fisher information, and (2.20) is its local version. The matrix (2.22) is needed to develop our asymptotic statements. The hat symbol always designates corresponding empirical version based on F_n.

Remark 1. Of course, it matters which function U do we choose in the BD family. In [10] and in [18] the effect of this choice is discussed is more detail. In any case, when a robustification is sought, it makes sense to choose U which is parameterized by a parameter, r, say, whereby the parameterized function represents some sort of distortion of the logarithmic function. Very often this is achieved by replacing in the core definition a logarithm by the Box-Cox transformation. In such a way, the new (non-local) robust estimation procedure can be interpreted as one that tries (empirically) to minimise “the discrepancy between a distribution in an ideal parametric family and one that modifies the true distribution to diminish (or emphasize) the role of extreme observations” (see [10], p. 754). Their procedure is called maximum L_r-likelihood estimation procedure to reflect the distortion of the usual likelihood function (obtained as a limiting case when $r = 1$) via the distortion function $L_r(u) = \{u^{1-r} - 1\}/(1-r), r > 0$. These authors have not discussed robustness issues except to mention that these might be interesting to be discussed in a future work and they do not discuss any effects of localisation. Our estimation procedure is different to theirs and it can be implemented in such a way as to deliver robustification in regression setting that is consistent with the classical requirement for a robust estimator to have a bounded influence function.

A paper that deals explicitly with the aspects of Bregman divergence in regression is [23]. It is more closely related to our own paper although again it does not analyse any localisation effects and do not deal with model misspecification. Further, we note that in the global setting, the Bregman divergence they use is not the same as ours. They aim to measure the discrepancy between y and $m(\boldsymbol{x},\boldsymbol{\theta})$, rather than between $\mu(\boldsymbol{x})$ and $m(\boldsymbol{x},\boldsymbol{\theta})$ as we do. However it can be confirmed that, eventually, their estimation procedure (Equation (20) in p.126) for $\boldsymbol{\theta}$ coincides with ours in the no-misspecification case $\mu(\cdot) = m(\cdot,\boldsymbol{\theta}_*)$.

We should also mention that the paper [9] also deals explicitly with the model misspecification issue in regression. It points out that a localisation bandwidth

h controls the behaviour of the local estimator, with two distinctive effects. For nonparametric consistency, a bandwidth h tending to zero as sample size $n \to \infty$ is to be chosen. On the other hand, with h large, the estimator would share asymptotic efficiency with the parametric estimator if the parametric model is precisely correct but at the same time, it will suffer much less than the parametric estimator when a slight misspecification of the parametric model occurs. However the paper assumes that the conditional density $p(y|\boldsymbol{x})$ belongs to an exponential family. We do not need this assumption.

The authors of [23] investigate the question of how to tune the bandwidth h depending on the degree of model misspecification. In contrast, we focus on investigating the gain of using the large h asymptotic approach instead of purely parametric regression in cases where the model misspecification is small so that we happen to be in the large h regime. In addition, in [23] the conditional distribution of Y given $\boldsymbol{X}$ is again supposed to belong to an exponential family. We do not need this assumption in our methodological part.

We also investigate robustness properties of the estimators constructed by using the large h approach. Noticing the structure of the loss function Q in [10], it becomes apparent via simple calculations that the resulting influence function of their estimator (and hence also of our estimator), in terms of the notations introduced in our paper, is proportional to

$$\Psi(\boldsymbol{\theta})^{-1}\left\{y - m(\boldsymbol{x}, \boldsymbol{\theta})\right\} \frac{u'(m(\boldsymbol{x}, \boldsymbol{\theta}))}{G'(m(\boldsymbol{x}, \boldsymbol{\theta}))} \widetilde{\boldsymbol{x}}.$$

Even if we assume, as we do, that the input (design) variable $\boldsymbol{X}$ could be restricted to be in a compact set, this influence function is clearly unbounded in general (unless the Y_i observations also stay in a bounded set).

In contrast, our method, in its form presented in Section 4 can deliver robustification in a classical sense, with a bounded influence function.

3. Asymptotic theory

3.1. Assumptions

We will now state the assumptions for our asymptotic statements related to $\hat{\mu}_G$ and $\hat{\mu}_L$ to hold. In the sequel, for any $(d+1) \times (d+1)$ matrix A, we let A_{ij} denote the (i, j)-component of A.

(A0) The kernel K satisfies $K(\mathbf{0}_d) = 1$ and $K(\boldsymbol{z}) = 1 - \kappa_2 ||\boldsymbol{z}||^2 + \kappa_4 ||\boldsymbol{z}||^4 + o(||\boldsymbol{z}||^4)$, as $||\boldsymbol{z}|| \to 0$, where $\kappa_2, \kappa_4 > 0$.

(A1) As $n \to \infty$ and $h \to \infty$, $h^2 = O(\sqrt{n})$.

(A2) The parameter space Θ is a bounded open subset in $\mathbb{R}^{d+1}$, and $\rho(y, \boldsymbol{x}, \boldsymbol{\theta})$ is continuous on $\mathbb{R} \times \mathbb{R}^d \times \Theta_c$, where Θ_c is the closure of Θ. For almost all $(y, \boldsymbol{x})$, $\rho(y, \boldsymbol{x}, \boldsymbol{\theta})$ is sufficiently smooth on Θ.

(A3) The parameter $\boldsymbol{\theta}_* \in \Theta$ is the unique minimiser of

$$\int_{\mathbb{R} \times \mathbb{R}^d} \rho(y, \boldsymbol{x}, \boldsymbol{\theta}) dF(y, \boldsymbol{x}).$$

(A4) $E[Y^2] < \infty$, and $E[\,|Y|\,|\boldsymbol{X} = \boldsymbol{x}]$ is continuous on $\mathbb{D}$.

(A5) $\displaystyle\int_{\mathbb{R}^d} \sup_{\theta \in \Theta_c} |u(m(\boldsymbol{x}, \boldsymbol{\theta}))|^2 q(\boldsymbol{x}) d\boldsymbol{x} < \infty$, and for any $\varepsilon > 0$, there exists $L > 0$ such that

$$\sup_{\theta \in \Theta_c} \int_{||\boldsymbol{x}|| > L} |U^*(u(m(\boldsymbol{x}, \boldsymbol{\theta})))| q(\boldsymbol{x}) d\boldsymbol{x} < \varepsilon.$$

(A6) For any $\boldsymbol{t} \in \mathbb{D}$ and any vector $\tilde{\boldsymbol{\theta}}$ satisfying $||\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}|| < ||\hat{\boldsymbol{\theta}}(\boldsymbol{t}) - \hat{\boldsymbol{\theta}}||$, both

$$\max_{1 \le i,j \le d+1} \left| \Psi(\boldsymbol{t}, \tilde{\boldsymbol{\theta}})_{ij} - \Psi(\boldsymbol{t}, \boldsymbol{\theta}_*)_{ij} \right| = o_p\left(\frac{1}{h^2}\right)$$

and

$$\max_{1 \le i,j \le d+1} \left| \widehat{\Psi}(\boldsymbol{t}, \tilde{\boldsymbol{\theta}})_{ij} - \widehat{\Psi}(\boldsymbol{t}, \boldsymbol{\theta}_*)_{ij} \right| = o_p\left(\frac{1}{h^2}\right)$$

hold as $n \to \infty$ and $h \to \infty$.

(A7) $\Psi(\boldsymbol{\theta}_*)$ is positive definite, and for any $\boldsymbol{t} \in \mathbb{D}$ and any vector $\tilde{\boldsymbol{\theta}}$ satisfying $||\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}|| < ||\hat{\boldsymbol{\theta}}(\boldsymbol{t}) - \hat{\boldsymbol{\theta}}||$, $\widehat{\Psi}(\boldsymbol{t}, \tilde{\boldsymbol{\theta}})$ is nonsingular.

Remark 2. The following remarks should be made regarding the above assumptions for our theory:

(a) (A0) relates to the shape of kernel function used for the localisation. A typical choice is the Gaussian kernel $K(\boldsymbol{z}) = \exp(-||\boldsymbol{z}||^2/2)$.

(b) We aim to develop asymototics for $\hat{\mu}_L$ and hence $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$ under the scenario where the bandwidth h grows as n increases. The typical order is set to $h^2 = O(\sqrt{n})$ in (A1), however other orders might also be possible, see [8]. We do not pursue the optimal order of h in this paper.

(c) (A2), (A3), (A4) and (A5) are necessary to demonstrate consistency of our local estimator $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$, see [14], [21] and [22].

(d) (A6) claims that the matrix $\Psi(\boldsymbol{t}, \boldsymbol{\theta}_*)$ is approximated well by $\Psi(\boldsymbol{t}, \tilde{\boldsymbol{\theta}})$ for $\tilde{\boldsymbol{\theta}}$ close to $\hat{\boldsymbol{\theta}}$, and that this is also true for the estimated version $\widehat{\Psi}(t, \boldsymbol{\theta}_*)$.

(e) (A7) assures the non-singularity of $\Psi(\boldsymbol{\theta}_*)$ and $\widehat{\Psi}(\boldsymbol{t}, \tilde{\boldsymbol{\theta}})$ for $\tilde{\boldsymbol{\theta}}$ close to $\hat{\boldsymbol{\theta}}$.

We see from (A0), (2.17), (2.21), (2.22) and (2.23) that

$$\begin{aligned}
\psi(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta}) &= \psi(y, \boldsymbol{x}, \boldsymbol{\theta}) - \frac{\kappa_2}{h^2}\psi^{(2)}(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta}) + \frac{\kappa_4}{h^4}\psi^{(4)}(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta}) + o\left(\frac{1}{h^4}\right), \\
\Psi(\boldsymbol{t}, \boldsymbol{\theta}) &= \Psi(\boldsymbol{\theta}) - \frac{\kappa_2}{h^2}\Psi^{(2)}(\boldsymbol{t}, \boldsymbol{\theta}) + \frac{\kappa_4}{h^4}\Psi^{(4)}(\boldsymbol{t}, \boldsymbol{\theta}) + o\left(\frac{1}{h^4}\right), \\
\widehat{\Psi}(\boldsymbol{t}, \boldsymbol{\theta}) &= \widehat{\Psi}(\boldsymbol{\theta}) - \frac{\kappa_2}{h^2}\widehat{\Psi}^{(2)}(\boldsymbol{t}, \boldsymbol{\theta}) + \frac{\kappa_4}{h^4}\widehat{\Psi}^{(4)}(\boldsymbol{t}, \boldsymbol{\theta}) + o_p\left(\frac{1}{h^4}\right).
\end{aligned}$$

3.2. Asymptotic results for estimators

Theorem 1. *Assume that (A0)-(A5) hold. Then $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}_*$ as $n \to \infty$. Further, for any $\boldsymbol{t} \in \mathbb{D}$, $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$ converges in probability to $\boldsymbol{\theta}_*$ as $n, h \to \infty$.*

Theorem 2. *Under Assumptions (A0)-(A7), for any $\boldsymbol{t} \in \mathbb{D}$, it follows as $n, h \to \infty$ that*

$$\begin{aligned}
&\hat{\boldsymbol{\theta}}(\boldsymbol{t}) - \hat{\boldsymbol{\theta}} \\
&= \frac{\kappa_2}{h^2}\Psi(\boldsymbol{\theta}_*)^{-1}\left[\int_{\mathbb{R}\times\mathbb{R}^d} \psi^{(2)}(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta}_*)dF(y, \boldsymbol{x}) + \frac{1}{\sqrt{n}}\hat{v}^{(2)}(\boldsymbol{t}, \boldsymbol{\theta}_*)\right] \\
&\qquad + \frac{1}{h^4}\Psi(\boldsymbol{\theta}_*)^{-1}\left[\kappa_2 \widehat{V}(\boldsymbol{t}, \boldsymbol{\theta}_*)\Psi(\boldsymbol{\theta}_*)^{-1}\int_{\mathbb{R}\times\mathbb{R}^d} \psi^{(2)}(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta}_*)dF(y, \boldsymbol{x})\right. \\
&\qquad\qquad \left. - \kappa_4 \int_{\mathbb{R}\times\mathbb{R}^d} \psi^{(4)}(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta}_*)dF(y, \boldsymbol{x})\right] + o_p\left(\frac{1}{h^4}\right),
\end{aligned}$$

where $\hat{v}^{(2)}$ is given by (7.1) in Lemma 3.

Theorem 3. *Under Assumptions (A0)-(A7), for any $\boldsymbol{t} \in \mathbb{D}$, it follows as $n, h \to \infty$ that*

$$\begin{aligned}
&E\left[\hat{\boldsymbol{\theta}}(\boldsymbol{t}) - \hat{\boldsymbol{\theta}}\right] \\
&= \frac{\kappa_2}{h^2}\Psi(\boldsymbol{\theta}_*)^{-1}\int_{\mathbb{R}\times\mathbb{R}^d} \psi^{(2)}(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta}_*)dF(y, \boldsymbol{x}) \\
&\qquad + \frac{1}{h^4}\left[\kappa_2^2\Psi(\boldsymbol{\theta}_*)^{-1}\Psi^{(2)}(\boldsymbol{t}, \boldsymbol{\theta}_*)\Psi(\boldsymbol{\theta}_*)^{-1}\int_{\mathbb{R}\times\mathbb{R}^d} \psi^{(2)}(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta}_*)dF(y, \boldsymbol{x})\right. \\
&\qquad\qquad \left. - \kappa_4\Psi(\boldsymbol{\theta}_*)^{-1}\int_{\mathbb{R}\times\mathbb{R}^d} \psi^{(4)}(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta}_*)dF(y, \boldsymbol{x})\right] + o\left(\frac{1}{h^4}\right).
\end{aligned}$$

Asymptotic normality of $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$ and $\hat{\mu}_L(\boldsymbol{t})$ can be demonstrated as follows:

Theorem 4. *Under Assumptions (A0)-(A7), for any $\boldsymbol{t} \in \mathbb{D}$,*

$$\sqrt{n}\left\{\hat{\boldsymbol{\theta}}(\boldsymbol{t}) - \boldsymbol{\theta}_* - \frac{\kappa_2}{h^2}\Psi(\boldsymbol{\theta}_*)^{-1}\int_{\mathbb{R}\times\mathbb{R}^d} \psi^{(2)}(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta}_*)dF(y, \boldsymbol{x})\right\}$$

converges in distribution to a $(d+1)$-dimensional Gaussian distribution with mean vector zero and with covariance matrix

$$\Sigma_* = \Psi(\boldsymbol{\theta}_*)^{-1}Cov[\psi(Y, \boldsymbol{X}, \boldsymbol{\theta}_*)]\Psi(\boldsymbol{\theta}_*)^{-1}. \tag{3.1}$$

Furthermore, $\sqrt{n}\left\{\hat{\mu}_L(\boldsymbol{t}) - m(\boldsymbol{t}, \boldsymbol{\theta}_*)\right\}$ also converges in distribution to a univariate Gaussian distribution with mean $b_*(\boldsymbol{t})$ and variance $\sigma_*^2(\boldsymbol{t})$, where

$$\begin{aligned}
b_*(\boldsymbol{t}) &= \tau\kappa_2\frac{\partial}{\partial\boldsymbol{\theta}^T}m(\boldsymbol{t}, \boldsymbol{\theta}_*)\Psi(\boldsymbol{\theta}_*)^{-1}\int_{\mathbb{R}\times\mathbb{R}^d} \psi^{(2)}(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta}_*)dF(y, \boldsymbol{x}), \\
\sigma_*^2(\boldsymbol{t}) &= \frac{\partial}{\partial\boldsymbol{\theta}^T}m(\boldsymbol{t}, \boldsymbol{\theta}_*)\Sigma_*\frac{\partial}{\partial\boldsymbol{\theta}}m(\boldsymbol{t}, \boldsymbol{\theta}_*)
\end{aligned}$$

and $\tau > 0$ is the limit of $\sqrt{n}/h^2$.

3.3. Asymptotics for the risk

The effect of the localisation is revealed when the performance comparison of the local and the global estimator is completed by using global-performance risk measures. The risk of the global estimator $\hat{\mu}_G(\boldsymbol{x}) = m(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) = G^{-1}(\hat{\boldsymbol{\theta}}^T \widetilde{\boldsymbol{x}})$ is defined as the expected value of its divergence:

$$\mathcal{R}(\hat{\mu}_G) = E\left[D_{U^*}(u(\hat{\mu}_G)), u(\mu))\right],$$

where the expectation is based on the sample. The risk of the local estimator $\hat{\mu}_L(\boldsymbol{x}) = m(\boldsymbol{x}, \hat{\boldsymbol{\theta}}(\boldsymbol{x})) = G^{-1}(\hat{\boldsymbol{\theta}}(\boldsymbol{x})^T \widetilde{\boldsymbol{x}})$ is defined similarly as

$$\mathcal{R}(\hat{\mu}_L) = E\left[D_{U^*}(u(\hat{\mu}_L), u(\mu))\right].$$

The difference between the risks of the global and of the local estimators is thus defined as

$$\mathfrak{R}(\hat{\mu}_G, \hat{\mu}_L) = \mathcal{R}(\hat{\mu}_G) - \mathcal{R}(\hat{\mu}_L). \tag{3.2}$$

We have the following result for the risk difference:

Theorem 5. *Under Assumptions (A0)-(A7), it follows as $n, h \to \infty$ that*

$$\begin{aligned} &\mathfrak{R}(\hat{\mu}_G, \hat{\mu}_L) - E\left[D_{U^*}(u(\hat{\mu}_G), u(\hat{\mu}_L))\right] \\ &= \frac{2\kappa_2}{h^2} \sum_{j=1}^{d} \eta_j(\boldsymbol{\theta}_*)^T \Psi(\boldsymbol{\theta}_*)^{-1} \eta_j(\boldsymbol{\theta}_*) + o\left(\frac{1}{h^2}\right), \end{aligned} \tag{3.3}$$

where

$$\eta_j(\boldsymbol{\theta}_*) = \int_{\mathbb{R}\times\mathbb{R}^d} x_j \psi(y, \boldsymbol{x}, \boldsymbol{\theta}_*) dF(y, \boldsymbol{x}) \in \mathbb{R}^{d+1} \tag{3.4}$$

for $j = 1, ..., d$ and x_j is the j-th component of $\boldsymbol{x}$.

From Theorem 5 it is clear that (since under Assumption (A7) the matrix $\Psi(\boldsymbol{\theta}_*)$ is positive definite) the local estimator outperforms the global one asymptotically. It is not always the case, however, that Assumption (A7) would hold. In Theorem 6 below we investigate sufficient conditions for (A7) to hold. We also note that obviously when the parametric model holds then $\eta_j(\boldsymbol{\theta}_*)$ will be zero. In this particular case the claim is that the risk difference (3.2) is still approximated by a positive quantity $E\left[D_{U^*}(u(\hat{\mu}_G), u(\hat{\mu}_L))\right]$ up to a smaller order error of $o(1/h^2)$.

3.4. Efficient choice of the link function

Our methodology requires choices of the strictly convex function $U \in \mathscr{U}$ and of the link function G. The asymptotic result in Theorem 5 suggests us an efficient choice of the link function G for a fixed U.

Theorem 6. *Fix a strictly convex function U. The choice $G \equiv u + \alpha$ for a constant α makes $\Psi(\boldsymbol{\theta}_*)$ positive definite, for any μ. With this choice of G we have*

$$\Psi(\boldsymbol{\theta}_*) = \int_{\mathbb{R}^d} \frac{1}{U''(G^{-1}(\boldsymbol{\theta}_*^T \widetilde{\boldsymbol{x}}))} \widetilde{\boldsymbol{x}} \widetilde{\boldsymbol{x}}^T q(\boldsymbol{x}) d\boldsymbol{x}. \tag{3.5}$$

Remark 3. We are focusing now on analyzing Σ_* of (3.1) in Theorem 4. Consider the case $G(t) = u(t) + \alpha$ for some constant α. Then $\Psi(\boldsymbol{\theta}_*)$ is positive definite by Theorem 6. The same result as in Theorem 6 of [23] holds for Σ_*. That is, if U satisfies

$$U''(m(\boldsymbol{x}, \boldsymbol{\theta}_*)) = u'(m(\boldsymbol{x}, \boldsymbol{\theta}_*)) = \frac{c}{E[\{Y - m(\boldsymbol{x}, \boldsymbol{\theta}_*)\}^2]}$$

for some constant $c > 0$, then the asymptotic covariance matrix Σ_* of $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$ attains the lower bound

$$\left\{ E\left[t(\boldsymbol{X}, \boldsymbol{\theta}_*)^{-1} \tilde{\boldsymbol{X}} \tilde{\boldsymbol{X}}^T \right] \right\}^{-1},$$

where $t(\boldsymbol{X}, \boldsymbol{\theta}_*) = E_{Y|X}[\{Y - m(\boldsymbol{X}, \boldsymbol{\theta}_*)\}^2] U''(m(\boldsymbol{X}, \boldsymbol{\theta}_*))^2$. This lower bound coincides with that obtained in Theorem 6 of [23] provided that $\mu(\boldsymbol{x}) = m(\boldsymbol{x}, \boldsymbol{\theta}_*)$.

3.5. Remark on robustness

The influence function of $\hat{\boldsymbol{\theta}}$ can be obtained as

$$\begin{aligned} \mathrm{IF}(y, \boldsymbol{x} : F) &= -\Psi(\boldsymbol{\theta}_*)^{-1} \psi(y, \boldsymbol{x}, \boldsymbol{\theta}_*) \\ &= \Psi(\boldsymbol{\theta}_*)^{-1} \{y - m(\boldsymbol{x}, \boldsymbol{\theta}_*)\} \frac{u'(m(\boldsymbol{x}, \boldsymbol{\theta}_*))}{G'(m(\boldsymbol{x}, \boldsymbol{\theta}_*))} \widetilde{\boldsymbol{x}}, \end{aligned} \tag{3.6}$$

which is unbounded unless y stays in a bounded set. The influence function of $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$ at $\boldsymbol{t} \in \mathbb{D}$ can be obtained similarly. We do not pursue this here.

By examining (3.6) it becomes clear why the influence function can be unbounded. This is due to the appearance of the residual function $y - m(\boldsymbol{x}, \boldsymbol{\theta}_*)$ as a multiplier in (3.6). This observation also suggest ways to implement a robust estimator as a substitute to $\hat{\boldsymbol{\theta}}$. The way we adopt in this paper is presented in the next Section.

4. Choosing a divergence to yield robust estimators

4.1. Composition of the divergence

Define

$$\mathscr{U}_0 = \{U \in \mathscr{U} \mid u(0) = 0,\ |u| \text{ is bounded}\}. \tag{4.1}$$

A typical member of $\mathscr{U}_0$ is the so-called pseudo-Huber loss (see, e.g., [12])

$$L_\delta(t) = \delta^2 \left\{ \sqrt{1+\left(\frac{t}{\delta}\right)^2} - 1 \right\},\ \delta > 0. \tag{4.2}$$

The pseudo-Huber loss is similar to the Huber loss, but has continuous derivatives of all orders and is a strictly convex function of t for any fixed $\delta > 0$. It should be noted that L'_δ is bounded.

In what follows we shall denote the residual function as $r(\boldsymbol{\theta}) = y - m(\boldsymbol{x}, \boldsymbol{\theta})$.

Now fix a $U \in \mathscr{U}_0$. To yield a robust estimator of $\boldsymbol{\theta}$, we utilize an another feature of Bregman divegence defined as

$$\begin{aligned} D_U(\boldsymbol{\theta}) &= \int_{\mathbb{R}\times\mathbb{R}^d} \{U(r(\boldsymbol{\theta})) - U(0) - u(0)\{r(\boldsymbol{\theta}) - 0\}\}\, dF(y, \boldsymbol{x}) \\ &= \int_{\mathbb{R}\times\mathbb{R}^d} U(y - m(\boldsymbol{x}, \boldsymbol{\theta})) dF(y, \boldsymbol{x}) - U(0). \end{aligned} \tag{4.3}$$

In particular, if the pseudo-Huber loss (4.2) is used for U then also $U(0) = 0$ holds. By minimising $D_U(\boldsymbol{\theta})$ we aim to minimise the discrepancy between the residual $r(\boldsymbol{\theta})$ and 0 rather than between Y and $m(\cdot, \boldsymbol{\theta})$ as implemented in [23]. This is the key to robustification since now the influence function can be made bounded even when the Y-observations are not, as will be seen in the discussion that follows.

4.2. Estimators

In this setting, necessary functions corresponding to (2.7) and (2.8) are put into

$$\rho(y, \boldsymbol{x}, \boldsymbol{\theta}) = U(y - G^{-1}(\boldsymbol{\theta}^T \widetilde{\boldsymbol{x}})), \tag{4.4}$$

$$\psi(y, \boldsymbol{x}, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \rho(y, \boldsymbol{x}, \boldsymbol{\theta}) = -\frac{u(y - G^{-1}(\boldsymbol{\theta}^T \widetilde{\boldsymbol{x}}))}{G'(G^{-1}(\boldsymbol{\theta}^T \widetilde{\boldsymbol{x}}))} \tilde{\boldsymbol{x}}. \tag{4.5}$$

Associated functions corresponding to local version (2.12), (2.13), (2.17) and the matrix (2.18) can be defined in the same way, hence we will deal with these functions as before but by using (4.4) as a starting point.

The estimator $\hat{\boldsymbol{\theta}}$ as well as the true parameter value yielding the best approximation to Y can be defined in a same manner as in (2.10) and (2.9), respectively. Local estimator $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$ and the parameter $\boldsymbol{\theta}(\boldsymbol{t})$ at $\boldsymbol{t}$ are also defined in a similar way to (2.15) and (2.14), respectively.

Furthermore, the above parameter estimators suggest immediately the global estimator $\hat{\mu}_G$ and the local estimator $\hat{\mu}_L$ of the regression function as given in (2.11) and (2.16).

With the new $\rho(y, \boldsymbol{x}, \boldsymbol{\theta})$ and $\psi(y, \boldsymbol{x}, \boldsymbol{\theta})$ defined in (4.4) and (4.5), the influence function of $\hat{\boldsymbol{\theta}}$ can be derived as

$$\mathrm{IF}(y, \boldsymbol{x} : F) = -\Psi(\boldsymbol{\theta}_*)^{-1} \frac{u(y - G^{-1}(\boldsymbol{\theta}^T \widetilde{\boldsymbol{x}}))}{G'(G^{-1}(\boldsymbol{\theta}^T \widetilde{\boldsymbol{x}}))} \tilde{\boldsymbol{x}}, \tag{4.6}$$

where

$$\Psi(\boldsymbol{\theta}) = \int_{\mathbb{R}\times\mathbb{R}^d} \tilde{c}_2(y,\boldsymbol{x},\boldsymbol{\theta})\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T dF(y,\boldsymbol{x}) \tag{4.7}$$

and

$$\begin{aligned}&\tilde{c}_2(y,\boldsymbol{x},\boldsymbol{\theta})\\ &= \frac{1}{\{G'(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))\}^2}\Big\{U''(y-G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))\\ &\qquad\qquad + u(y-G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))\frac{G''(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))}{G'(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))}\Big\}.\end{aligned} \tag{4.8}$$

We now compare the estimator $\hat{\boldsymbol{\theta}}$ obtained via the estimating equation based on (4.5) with that based on another divergence.

The so-called density power divergence [2] is defined as a Bregman divergence with $\tilde{U}_\beta(t) = t^{1+\beta}$ for $t \geq 0$ and $\beta > 0$, see Section 9.2 of [3] as well. Robust inference using this divergence with $0 < \beta < 1$ was discussed in [2] and [3]. This divergence, however, cannot be utilized in its original form in regression setting since this $\tilde{U}_\beta(t)$ is defined on $t \geq 0$, though it is itself natural for the divergence of densities.

A suitable version of $\tilde{U}_\beta$ for the regression problem is $U_\beta(t) = |t|^{1+\beta}$ for $\beta > 0$. This U_β is strictly convex on $\mathbb{R}$ so that it is a member of $\mathscr{U}$. Hence it is possible to consider the divergence (2.3) and (4.3) based on U_β for inference.

However, it is easily confirmed that the inference using the divergence (4.3) with $U = U_\beta$ is not robust, since U'_β is not bounded. Using the form of U'_β and (4.6), we see that the influence function of estimator based on this power divergence U_β is *not* bounded. This is in stark contrast to the fact that the influence function based on the divergence associated with $U = L_\delta$ is bounded, due to the fact that $u = U' = L'_\delta$ is bounded.

Both U_β and L_δ are members of $\mathscr{U}$, and L_δ is also a member of $\mathscr{U}_0$ but U_β is not. This difference is essential when analysing robustness properties.

4.3. Risk improvement in the robust setting

Similarly to Theorem 5, the local regression estimator $\hat{\mu}_L$ can improve the risk of the global estimator $\hat{\mu}_G$ also in the robust setting when using the divergence (4.3) and the associated functions (4.4) and (4.5).

The risk of the global estimator can be defined as

$$\mathcal{R}(\hat{\mu}_G) = E\left[\int_{\mathbb{R}\times\mathbb{R}^d} U(y-m(\boldsymbol{x},\hat{\boldsymbol{\theta}}))dF(y,\boldsymbol{x})\right]$$

whereas the risk of the local estimator is

$$\mathcal{R}(\hat{\mu}_L) = E\left[\int_{\mathbb{R}\times\mathbb{R}^d} U(y-m(\boldsymbol{x},\hat{\boldsymbol{\theta}}(\boldsymbol{x})))dF(y,\boldsymbol{x})\right]$$

where the expectation is based on the sample. The risk difference between global and local estimators is therefore given as

$$\mathfrak{R}(\hat{\mu}_G, \hat{\mu}_L) = \mathcal{R}(\hat{\mu}_G) - \mathcal{R}(\hat{\mu}_L). \tag{4.9}$$

The following theorem sates that, under Assumptions (A0)-(A7), the local estimator is better than the global estimator when using the global risk (4.9) as a measure of performance:

Theorem 7. *Let $U \in \mathscr{U}_0$. Under Assumptions (A0)-(A7), it follows as $n, h \to \infty$ that*

$$\mathfrak{R}(\hat{\mu}_G, \hat{\mu}_L) \geq \frac{2\kappa_2}{h^2} \sum_{j=1}^{d} \zeta_j(\boldsymbol{\theta}_*)^T \Psi(\boldsymbol{\theta}_*)^{-1} \zeta_j(\boldsymbol{\theta}_*) + o\left(\frac{1}{h^2}\right),$$

where

$$\zeta_j(\boldsymbol{\theta}_*) = \int_{\mathbb{R}\times\mathbb{R}^d} x_j u(y - m(\boldsymbol{x}, \boldsymbol{\theta}_*)) \frac{\partial}{\partial \boldsymbol{\theta}} m(\boldsymbol{x}, \boldsymbol{\theta}_*) dF(y, \boldsymbol{x}) \in \mathbb{R}^{d+1}$$

for $j = 1, ..., d$.

We now discuss, in a similar way as in Theorem 6, examples of sufficient conditions for the positive definiteness of $\Psi(\boldsymbol{\theta}_*)$. We claim that a choice for $G = u$ associated with $U = L_\delta$, the pseudo-Huber loss in (4.2), leads to the desired positive definiteness. To see this, we introduce the constants

$$M(u^{-1}) = \max\{|u^{-1}(\boldsymbol{\theta}^T \tilde{\boldsymbol{x}})| \mid \boldsymbol{x} \in \mathbb{D}, \boldsymbol{\theta} \in \Theta_c\}$$

and

$$M = \max\{E[\,|Y| \,| \boldsymbol{X} = \boldsymbol{x}] \mid \boldsymbol{x} \in \mathbb{D}\},$$

both of which certainly exist by the smoothness of $G = u$ and the assumption (A4) as well as the assumed compactness of $\mathbb{D}$ and Θ_c.

We formulate the following theorem:

Theorem 8. *Let U be equal to the pseudo-Huber loss L_δ in (4.2), and let $G = u = U'$. Then $\Psi(\boldsymbol{\theta}_*)$ becomes positive definite for sufficiently large δ.*

Remark 4. Theorem 8 can be easily confirmed by looking at the structure of $\tilde{c}_2(y, \boldsymbol{x}, \boldsymbol{\theta})$ in (4.8). By the definition of the pseudo-Huber loss (4.2) and the setting $G = U'$, it follows that

$$\tilde{c}_2(y, \boldsymbol{x}, \boldsymbol{\theta}) = \tilde{c}_{2,1}(y, \boldsymbol{x}, \boldsymbol{\theta}) + \frac{1}{\delta^2} \tilde{c}_{2,2}(y, \boldsymbol{x}, \boldsymbol{\theta}), \tag{4.10}$$

where

$$\tilde{c}_{2,1}(y, \boldsymbol{x}, \boldsymbol{\theta}) = \frac{\{1 + (G^{-1}(\boldsymbol{\theta}^T \tilde{\boldsymbol{x}})/\delta)^2\}^3}{[1 + \{(y - G^{-1}(\boldsymbol{\theta}^T \tilde{\boldsymbol{x}}))/\delta\}^2]^{3/2}} \tag{4.11}$$

and

$$\tilde{c}_{2,2}(y, \boldsymbol{x}, \boldsymbol{\theta}) \tag{4.12}$$

$$= -3G^{-1}(\boldsymbol{\theta}^T\tilde{\boldsymbol{x}})\{y - G^{-1}(\boldsymbol{\theta}^T\tilde{\boldsymbol{x}})\}\frac{\{1 + (G^{-1}(\boldsymbol{\theta}^T\tilde{\boldsymbol{x}})/\delta)^2\}^2}{[1 + \{(y - G^{-1}(\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}))/\delta\}^2]^{1/2}}.$$

Note that $\tilde{c}_{2,1}(y, \boldsymbol{x}, \boldsymbol{\theta})$ in (4.11) is always positive for any $(y, \boldsymbol{x})$ as well as for any $\boldsymbol{\theta}$ and any positive δ. In addition, since $\tilde{c}_{2,2}(y, \boldsymbol{x}, \boldsymbol{\theta})$ is bounded in (4.12) for any positive δ we see that $\tilde{c}_2(y, \boldsymbol{x}, \boldsymbol{\theta})$ will be positive for large δ. Since the positive definiteness of $\Psi(\boldsymbol{\theta}_*)$ in (4.7) can be captured only through $\tilde{c}_2$ in (4.10), $\Psi(\boldsymbol{\theta}_*)$ will become positive definite for large δ.

Remark 5. The claim of Theorem 8 represents a new perspective on the cost of robustness guarantee. It is a part of the folklore that a strong robustness generally implies low efficiency of the method, and pursuing efficiency reduces the robustness of the method. Theorem 8 suggests that large value of δ guarantees the positive definiteness of $\Psi(\boldsymbol{\theta}_*)$, and hence the positiveness of the leading term of the risk difference in Theorem 7. Therefore, large δ guarantees the risk improvement by the proposed local estimator. But we note that by the definition of the pseudo-Huber loss in (4.2), this large δ reveals a weak robustness. Hence the dilemma between robustness and efficiency also appears in our current setting when analysing the risk improvement by the local method. We also note that our methodology, as implemented in the local estimator setting, allows us to analyse the separate effects of model misspecification and of robustness, as well as their interplay. Using a bandwidth $h^2 = O(\sqrt{n})$ as in Assumption (A1) represents a balance in the sense that using a smaller order bandwidth h will help if model misspecification of the conditional mean prevails whereas larger h would be helpful if there is a need for a more robustification. However, the robustification achieved by the local estimator has its limitations. Indeed, an attempt for a stronger robustification by choosing small δ may cause $\Psi(\boldsymbol{\theta}_*)$ to not be positive definite anymore. Hence, if for some reasons, a strong robustification is aimed at then the global estimator may do a better job.

From a practical point of view, one possible drawback of the choice $G = u$ in Theorem 8 is that the resulting $G^{-1} = u^{-1}$ requires $|\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}| < \delta$ since

$$G^{-1}(t) = \frac{\delta t}{\sqrt{\delta^2 - t^2}} = \frac{t}{\sqrt{1 - (t/\delta)^2}}.$$

This means that $m(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) = G^{-1}(\hat{\boldsymbol{\theta}}^T\widetilde{\boldsymbol{x}})$ would not be defined when $|\hat{\boldsymbol{\theta}}^T\widetilde{\boldsymbol{x}}| \geq \delta$. For large δs, this does not matter theoretically, but there is nonzero probability that $\hat{\boldsymbol{\theta}}^T\widetilde{\boldsymbol{x}}$ is bigger than δ for some $\boldsymbol{x}$ thus causing a numerical instability in practice. A less problematic choice of $G = u$ is the choice $G = u^{-1}$. We again achieve a risk improvement by the local estimator as shown in the following result:

Theorem 9. *Let U be equal to the pseudo-Huber loss L_δ in (4.2), and let $G = u^{-1} = (U')^{-1}$. Then $\Psi(\boldsymbol{\theta}_*)$ becomes positive definite for sufficiently large δ.*

5. Numerical implementations and simulations

The main purpose of this section is to discuss the numerical implementation of the estimations and to compare the performance of the global and local estimators. Furthermore, the risk reduction effect illustrated in our Theorems will be demonstrated in a short simulation study.

5.1. Numerical implementation of the estimation algorithms

First, we discuss the numerical implementation of the regression estimators $\hat{\mu}_G$ and $\hat{\mu}_L$. These are obtained via plug-in once the corresponding parameter estimators $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$ are obtained. Using the obtained regression estimators, we will be able to analyze and demonstrate the risk reduction achieved when using $\hat{\mu}_L$ instead of $\hat{\mu}_G$. Theoretical analysis of these improvement effects was been presented in Theorems 5 and 7 and in this section additional numerical support is demonstrated. Although the numerical implementations of the estimators are standard, we discuss them here for the purpose of making our discussion self-contained.

5.1.1. The Global Estimator

In a first step, we discuss the numerical implementation of the global estimator $\hat{\boldsymbol{\theta}}$ for $\hat{\mu}_G$. The Newton-Raphson iterative method is used for this purpose. Starting with an initial guess $\hat{\boldsymbol{\theta}}^{[0]}$, the updates are obtained as follows:

$$\hat{\boldsymbol{\theta}}^{[k+1]} = \hat{\boldsymbol{\theta}}^{[k]} - \left\{\widehat{\Psi}(\hat{\boldsymbol{\theta}}^{[k]})\right\}^{-1} \int_{\mathbb{R}\times\mathbb{R}^d} \psi(y, \boldsymbol{x}, \hat{\boldsymbol{\theta}}^{[k]}) dF_n(y, \boldsymbol{x}), \tag{5.1}$$

where $k = 0, 1, 2, ...$ denote the iteration steps.

For independent observations $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$ from $F(y, \boldsymbol{x})$, we introduce the $n \times (p+1)$ design matrix, and two n-dimensional vectors

$$\widetilde{\mathbf{X}} = \begin{bmatrix} \widetilde{\boldsymbol{X}}_1^T \\ \vdots \\ \widetilde{\boldsymbol{X}}_n^T \end{bmatrix}, \quad \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \boldsymbol{m}(\boldsymbol{\theta}) = \begin{bmatrix} m(\boldsymbol{X}_1, \boldsymbol{\theta}) \\ \vdots \\ m(\boldsymbol{X}_n, \boldsymbol{\theta}) \end{bmatrix}. \tag{5.2}$$

Here we note that $\widetilde{\boldsymbol{X}}_i = [1\ \boldsymbol{X}_i^T]^T$. Then we see from (2.8), (5.1) and (5.2) that

$$\int_{\mathbb{R}\times\mathbb{R}^d} \psi(y, \boldsymbol{x}, \boldsymbol{\theta}) dF_n(y, \boldsymbol{x}) = -\frac{1}{n} \widetilde{\mathbf{X}}^T \mathbf{W}(\boldsymbol{\theta})(\boldsymbol{Y} - \boldsymbol{m}(\boldsymbol{\theta})), \tag{5.3}$$

where

$$\mathbf{W} = \mathbf{W}(\boldsymbol{\theta}) = \operatorname{diag}\{w_1(\boldsymbol{\theta}), \ldots, w_n(\boldsymbol{\theta})\}$$

with entries

$$w_i(\boldsymbol{\theta}) = \frac{u'(m(\boldsymbol{X}_i, \boldsymbol{\theta}))}{G'(m(\boldsymbol{X}_i, \boldsymbol{\theta}))}, \ i = 1, ..., n. \tag{5.4}$$

On the other hand, by using (2.19) and (7.4) in the proof of Theorem 6, we get

$$\widehat{\Psi}(\boldsymbol{\theta}) = \frac{1}{n}\widetilde{\mathbf{X}}^T\boldsymbol{\Delta}(\boldsymbol{\theta})\widetilde{\mathbf{X}}, \tag{5.5}$$

where

$$\boldsymbol{\Delta} = \boldsymbol{\Delta}(\boldsymbol{\theta}) = \text{diag}\{c_2(Y_1, \boldsymbol{X}_1, \boldsymbol{\theta}), \dots, c_2(Y_n, \boldsymbol{X}_n, \boldsymbol{\theta})\}$$

and $c_2(y, \boldsymbol{x}, \boldsymbol{\theta})$ is that given in (7.5).

The relations (5.3) and (5.5) help us to arrive at a more convenient expression for the update (5.1):

$$\hat{\boldsymbol{\theta}}^{[k+1]} = \hat{\boldsymbol{\theta}}^{[k]} + \{\widetilde{\mathbf{X}}^T\boldsymbol{\Delta}(\hat{\boldsymbol{\theta}}^{[k]})\widetilde{\mathbf{X}}\}^{-1}\widetilde{\mathbf{X}}^T\mathbf{W}(\hat{\boldsymbol{\theta}}^{[k]})(\boldsymbol{Y} - \boldsymbol{m}(\hat{\boldsymbol{\theta}}^{[k]})). \tag{5.6}$$

Note that an efficient choice $G = u$ gives $\mathbf{W} = \mathbf{I}_n$, the identity matrix, as confirmed by (5.4). Furthermore the choice $G = u$ simplifies $\boldsymbol{\Delta}$ as

$$\boldsymbol{\Delta} = \boldsymbol{\Delta}(\boldsymbol{\theta}) = \text{diag}\left\{\frac{1}{G'(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{X}}_1))}, \cdots, \frac{1}{G'(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{X}}_n))}\right\}.$$

Therefore (5.6) is finally simplified to

$$\hat{\boldsymbol{\theta}}^{[k+1]} = \hat{\boldsymbol{\theta}}^{[k]} + \{\widetilde{\mathbf{X}}^T\boldsymbol{\Delta}(\hat{\boldsymbol{\theta}}^{[k]})\widetilde{\mathbf{X}}\}^{-1}\widetilde{\mathbf{X}}^T(\boldsymbol{Y} - \boldsymbol{m}(\hat{\boldsymbol{\theta}}^{[k]})).$$

As long as we utilize $G = u = U'$, we only need to care to update the matrix $\boldsymbol{\Delta}(\boldsymbol{\theta})$ and the vector $\boldsymbol{m}(\boldsymbol{\theta})$.

5.1.2. The Local Estimator

A similar algorithm can be implemented for the local estimator $\hat{\mu}_L(\boldsymbol{t})$ at $\boldsymbol{t} \in \mathbb{D}$. Starting with an initial guess $\hat{\boldsymbol{\theta}}(\boldsymbol{t})^{[0]}$, the k-th update of the Newton-Raphson iterative algorithm can be written as

$$\hat{\boldsymbol{\theta}}(\boldsymbol{t})^{[k+1]} = \hat{\boldsymbol{\theta}}(\boldsymbol{t})^{[k]} - \left\{\widehat{\Psi}(\boldsymbol{t}, \hat{\boldsymbol{\theta}}(\boldsymbol{t})^{[k]})\right\}^{-1}\int_{\mathbb{R}\times\mathbb{R}^d}\psi(\boldsymbol{t}, y, \boldsymbol{x}, \hat{\boldsymbol{\theta}}(\boldsymbol{t})^{[k]})dF_n(y, \boldsymbol{x}), \tag{5.7}$$

$k = 0, 1, 2, \dots$. The vector to which $\hat{\boldsymbol{\theta}}(\boldsymbol{t})^{[k]}$ converges is defined as $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$. Based on (5.7), a local version for (5.3) can be obtained by introducing the matrix

$$\mathbf{K} = \mathbf{K}(\boldsymbol{t}) = \text{diag}\left\{K\left(\frac{\boldsymbol{X}_1 - \boldsymbol{t}}{h}\right), \dots, K\left(\frac{\boldsymbol{X}_n - \boldsymbol{t}}{h}\right)\right\}, \boldsymbol{t} \in \mathbb{D}.$$

In fact, we have from (2.13), (2.20) and (5.3) that

$$\begin{aligned}&\int_{\mathbb{R}\times\mathbb{R}^d}\psi(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta})dF_n(y, \boldsymbol{x})\\ &= -\frac{1}{n}\sum_{i=1}^{n}K\left(\frac{\boldsymbol{X}_i - \boldsymbol{t}}{h}\right)\psi(Y_i, \boldsymbol{X}_i, \boldsymbol{\theta})\end{aligned}$$

$$= -\frac{1}{n}\widetilde{\mathbf{X}}^T\mathbf{K}(\boldsymbol{t})^{1/2}\mathbf{W}(\boldsymbol{\theta})\mathbf{K}(\boldsymbol{t})^{1/2}(\boldsymbol{Y} - \boldsymbol{m}(\boldsymbol{\theta})), \tag{5.8}$$

and

$$\begin{aligned}\widehat{\Psi}(\boldsymbol{t},\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{\boldsymbol{X}_i - \boldsymbol{t}}{h}\right) c_2(Y_i, \boldsymbol{X}_i, \boldsymbol{\theta})\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^T \\ &= \frac{1}{n}\widetilde{\mathbf{X}}^T\mathbf{K}(\boldsymbol{t})^{1/2}\boldsymbol{\Delta}(\boldsymbol{\theta})\mathbf{K}(\boldsymbol{t})^{1/2}\widetilde{\mathbf{X}}. \end{aligned}\tag{5.9}$$

The choice $G = u$, (5.8) and (5.9) lead to

$$\hat{\boldsymbol{\theta}}(\boldsymbol{t})^{[k+1]} = \hat{\boldsymbol{\theta}}(\boldsymbol{t})^{[k]} + \{\widetilde{\mathbf{X}}^T\mathbf{K}(\boldsymbol{t})^{1/2}\boldsymbol{\Delta}(\hat{\boldsymbol{\theta}}(\boldsymbol{t})^{[k]})\mathbf{K}(\boldsymbol{t})^{1/2}\widetilde{\mathbf{X}}\}^{-1}\widetilde{\mathbf{X}}^T\mathbf{K}(\boldsymbol{t})(\boldsymbol{Y} - \boldsymbol{m}(\hat{\boldsymbol{\theta}}(\boldsymbol{t})^{[k]})).$$

5.1.3. Estimators in Bernoulli Regression

Applications of above algorithm in the setting of Bernoulli regression with the sample size $n = 200$ and $d = 1$ are exhibited in Figures 1(a) and 1(b), where the true regression function $\mu(x)$ is

$$\mu(x) = 0.5 + 0.4 \cdot \cos x. \tag{5.10}$$

in Figure 1(a) and

$$\mu(x) = \frac{\exp(0.2 - x)}{1 + \exp(0.2 - x)} \tag{5.11}$$

in Figure 1(b). The bandwidth for the local estimator is $h = 0.25 \cdot (200)^{1/4}$. The utilized parametric model is

$$m(x, \boldsymbol{\theta}) = G^{-1}(\boldsymbol{\theta}^T\tilde{x}) = \frac{\exp(\boldsymbol{\theta}^T\tilde{x})}{1 + \exp(\boldsymbol{\theta}^T\tilde{x})}$$

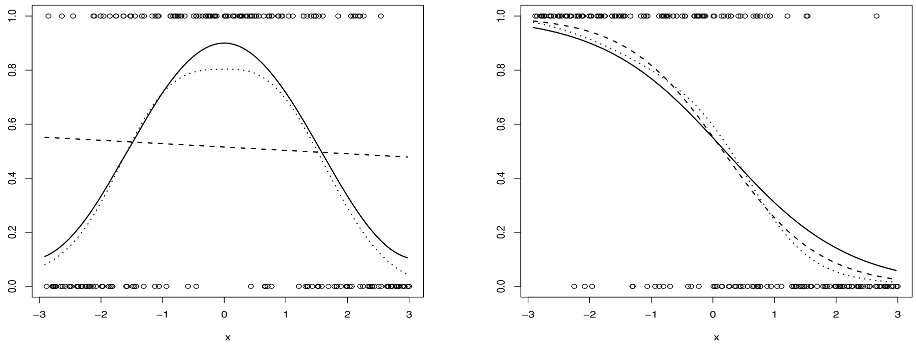
with $\tilde{x} = [1\ x]^T$ and $\boldsymbol{\theta} = [\theta_0\ \theta_1]^T$.

We see from Figure 1(a) that the local estimator $\hat{\mu}_L(x)$ (dotted) fits to the true $\mu(x)$, while the global estimator $\hat{\mu}_G(x)$ (dashed) cannot capture the structure of $\mu(x)$. In contrast, in Figure 1(b) the parametric model includes the true $\mu(x)$, hence both the global and local approaches give reasonable estimators.

5.2. Simulation related to Theorem 5

Referring to the proof of Theorem 5, we need to calculate an estimate of

$$\begin{aligned}&\Re(\hat{\mu}_G, \hat{\mu}_L) - E\left[D_{U^*}(u(\hat{\mu}_G), u(\hat{\mu}_L))\right] \\ &= E\left[\int_{\mathbb{R}^d} \{\hat{\mu}_L(\boldsymbol{x}) - \mu(\boldsymbol{x})\}\{u(\hat{\mu}_G(\boldsymbol{x})) - u(\hat{\mu}_L(\boldsymbol{x}))\} q(\boldsymbol{x})d\boldsymbol{x}\right],\end{aligned}\tag{5.12}$$



(a) (b)

FIG 1. *Simulated data(circle); $\mu(x)$ (solid); $\hat{\mu}_G(x)$ (dashed); $\hat{\mu}_L(x)$ (dotted); (a), $\mu(x)$ in (5.10); (b), $\mu(x)$ in (5.11).*

which corresponds to the left hand side of (3.3). To calculate

$$\sum_{j=1}^{d} \eta_j(\boldsymbol{\theta}_*)^T \Psi(\boldsymbol{\theta}_*)^{-1} \eta_j(\boldsymbol{\theta}_*),$$

the $O(h^2)$-term of the right hand side of (3.3), we need to obtain an approximation to $\Psi(\boldsymbol{\theta}_*)$ in (2.18) and $\eta_j(\boldsymbol{\theta}_*)$ in (3.4). The integrals involved in (5.12), (2.18) and (3.4) do not have a closed form in general and we replace them by Monte Carlo integral approximation.

To this end, we generate a random sample of size S drawn from $F(y, \boldsymbol{x})$: $(y_1^*, \boldsymbol{x}_1^*), \ldots, (y_S, \boldsymbol{x}_S^*) \sim F(y, \boldsymbol{x})$. Independently, we generate a data set

$$(y_1^{(t)}, \boldsymbol{x}_1^{(t)}), \ldots, (y_n^{(t)}, \boldsymbol{x}_n^{(t)})$$

from $F(y, \boldsymbol{x})$ for $t = 1, \ldots, T$, and calculate estimators $\hat{\boldsymbol{\theta}}(\cdot)^{(t)}$ and $\hat{\boldsymbol{\theta}}^{(t)}$ at each iteration $t = 1, \ldots, T$. Using these estimators at each iteration t, we have $\hat{\mu}_L(\cdot)^{(t)} = m(\cdot, \hat{\boldsymbol{\theta}}(\cdot)^{(t)})$ and $\hat{\mu}_G(\cdot)^{(t)} = m(\cdot, \hat{\boldsymbol{\theta}}^{(t)})$. We then have the estimates of (5.12) and $\sum_{j=1}^{d} \eta_j(\boldsymbol{\theta}_*)^T \Psi(\boldsymbol{\theta}_*)^{-1} \eta_j(\boldsymbol{\theta}_*)$ defined as

$$\hat{L} = \frac{1}{T \cdot S} \sum_{t=1}^{T} \sum_{s=1}^{S} \left[\left\{ \hat{\mu}_L(\boldsymbol{x}_s^*)^{(t)} - \mu(\boldsymbol{x}_s^*) \right\} \left\{ u(\hat{\mu}_G(\boldsymbol{x}_s^*)^{(t)}) - u(\hat{\mu}_L(\boldsymbol{x}_s^*)^{(t)}) \right\} \right] \tag{5.13}$$

and

$$\hat{R}(\boldsymbol{\theta}_\dagger) = \sum_{j=1}^{d} \hat{\eta}_j(\boldsymbol{\theta}_\dagger)^T \left[\frac{1}{S} \sum_{s=1}^{S} \frac{\partial}{\partial \boldsymbol{\theta}} \psi(y_s^*, \boldsymbol{x}_s^*, \boldsymbol{\theta}_\dagger) \right]^{-1} \hat{\eta}_j(\boldsymbol{\theta}_\dagger), \tag{5.14}$$

respectively, where

$$\hat{\eta}_j(\boldsymbol{\theta}) = \frac{1}{S} \sum_{s=1}^{S} x_{sj}^* \psi(y_s^*, \boldsymbol{x}_s^*, \boldsymbol{\theta}),$$

see (3.4). Here x^*_{sj} is the j-th component of $\boldsymbol{x}^*_s$. The vector $\boldsymbol{\theta}_\dagger$ can be chosen as $\boldsymbol{\theta}_\dagger = \boldsymbol{\theta}_*$ for the parametric case $\mu(\boldsymbol{x}) = m(\boldsymbol{x}, \boldsymbol{\theta}_*)$. For the more general situation where $\mu(\boldsymbol{x})$ differs from the parametric model (i.e., for the "approgression" case), it is not as easy to determine the "true" vector $\boldsymbol{\theta}_*$ as the solution of (2.9). In this case we would utilize $\boldsymbol{\theta}_\dagger$ defined as $\boldsymbol{\theta}_\dagger = T^{-1}\sum_{t=1}^T \hat{\boldsymbol{\theta}}^{(t)}$.

Under the choice of $G = u$, it is easily verified that (5.14) can be expressed as

$$\hat{R}(\boldsymbol{\theta}_\dagger) = \frac{1}{S}\mathrm{tr}\left[\widetilde{\mathbf{X}}_*(\widetilde{\mathbf{X}}_*^T\Delta_*\widetilde{\mathbf{X}}_*)^{-1}\widetilde{\mathbf{X}}_*^T\left[\{\boldsymbol{\mu}_* - \boldsymbol{m}_*(\boldsymbol{\theta}_\dagger)\}\{\boldsymbol{\mu}_* - \boldsymbol{m}_*(\boldsymbol{\theta}_\dagger)\}^T \odot \mathbf{X}_*\mathbf{X}_*^T\right]\right],$$

where $\odot$ is the Hadamard product, see Chapter 7 in [19], $\widetilde{\mathbf{X}}_* = [\mathbf{1}_S\ \mathbf{X}_*]$,

$$\mathbf{X}_* = \begin{bmatrix} \boldsymbol{x}_1^{*T} \\ \vdots \\ \boldsymbol{x}_S^{*T} \end{bmatrix}, \quad \boldsymbol{\mu}_* = \begin{bmatrix} \mu(\boldsymbol{x}_1^*) \\ \vdots \\ \mu(\boldsymbol{x}_S^*) \end{bmatrix}, \quad \boldsymbol{m}_*(\boldsymbol{\theta}) = \begin{bmatrix} m(\boldsymbol{x}_1^*, \boldsymbol{\theta}) \\ \vdots \\ m(\boldsymbol{x}_S^*, \boldsymbol{\theta}) \end{bmatrix},$$

$\Delta_* = \mathrm{diag}\{c_2(y_1^*, \boldsymbol{x}_1^*, \boldsymbol{\theta}_\dagger), \ldots, c_2(y_S^*, \boldsymbol{x}_S^*, \boldsymbol{\theta}_\dagger)\}$ and $\mathbf{1}_S$ stands for the S-dimensional vector of 1's.

We can check Theorem 5 by comparing $\hat{L}$ and $2\kappa_2\hat{R}(\boldsymbol{\theta}_\dagger)/h^2$ for several large values of n, where $h = c \cdot n^{1/4}$ for some constant $c > 0$. The Gaussian kernel $K(\boldsymbol{z}) = \exp(-||\boldsymbol{z}||^2/2)$ is adopted, hence $\kappa_2 = 1/2$.

For simplicity $d = 1$ is adopted in the following experiments. Also we set parameters to $S = 200$ and $T = 100$.

5.2.1. Normal regression example

Under the usual setting of linear regression with normal errors, we modelled the joint density $f(y, x)$ associated with the distribution function $F(y, x)$ as follows:

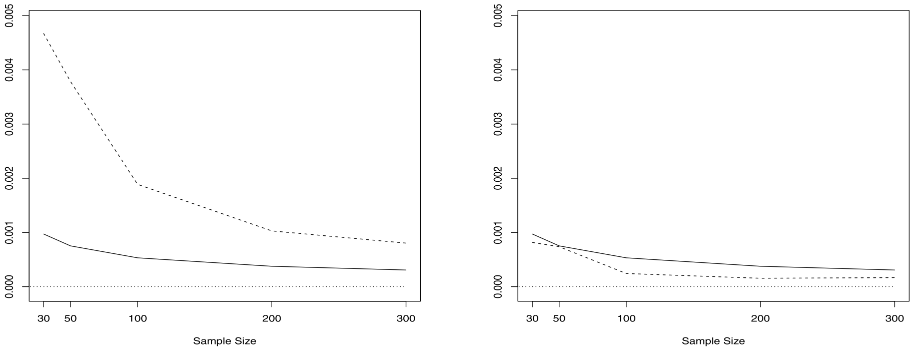
$$f(y, x) = p(y|x)q(x) \equiv (0.2)^{-1}\phi((y - \mu(x))/0.2) \cdot (1/6)\mathbf{1}(x \in [-3, 3]), \quad (5.15)$$

i.e., the conditional distribution of Y given $X = x$ is $N(\mu(x), (0.2)^2)$. We choose the true regression function as $\mu(x) = \sin x$ in this simulation design, and we utilize

$$m(x, \boldsymbol{\theta}) = G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}) = \theta_0 + \theta_1 x,$$

as a parametric model. This example has also been used in [9]. Further we note that $U(t) = t^2/2$ corresponds to the normal regression.

Results of simulations are exhibited in Figures 2(a) and (b). For both simulated cases $c = 1, 1.4$, $\hat{L}$ in (5.13) is positive (dashed), which means that the local estimator improves the risk. As h increases from $h = 1^2 \cdot \sqrt{n}$ to $h = (1.4)^2 \cdot \sqrt{n}$, $\hat{L}$ is getting uniformly smaller, which means that the difference between the local estimator $\hat{\mu}_L(x)$ and the global estimator $\hat{\mu}_G(x)$ is disappearing just as the theory in the previous section claims. The claim of Theorem 5 is clearly illustrated especially on Figure 2(b) for $c = 1.4$, $\hat{R}/(1.4^2\sqrt{n})$ (solid) is very close to $\hat{L}$ (dashed).



(a) (b)

Fig 2. *Case of Normal: (5.14) divided by* $c^2\sqrt{n}$ *(solid); (5.13)(dashed); (a)* $c = 1.0$*, (b)* $c = 1.4$.

5.2.2. Bernoulli regression example

Next we implement a Bernoulli regression example using the following experimental design:

$$f(y,x) = p(y|x)q(x) \equiv \mu(x)^y(1-\mu(x))^{1-y} \cdot (1/6)\mathbf{1}(x \in [-3,3]),\ y \in \{0,1\}.$$

That is, the conditional distribution of Y given $X = x$ is the Bernoulli distribution with the probability of success equal to $\mu(x)$. The true regression function $\mu(x)$ was chosen as in (5.10). The parametric model utilized in this case was

$$m(x,\boldsymbol{\theta}) = G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}) = \frac{\exp(\theta_0+\theta_1 x)}{1+\exp(\theta_0+\theta_1 x)}.$$

Notice that the associated convex function $U(t)$ for the Bernoulli (binomial) regression is given as

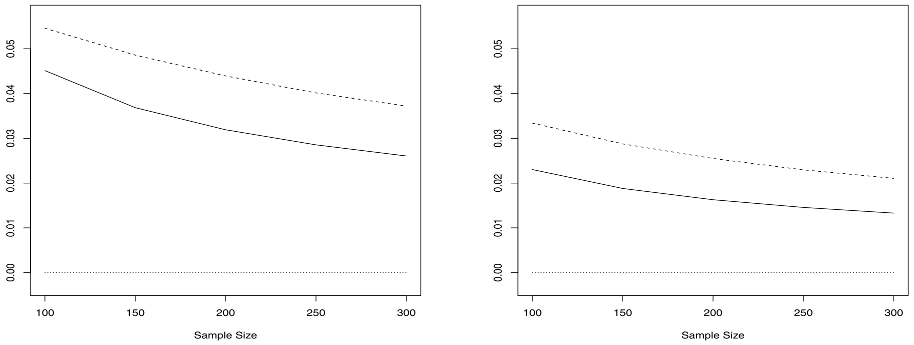
$$U(t) = t(\log t - 1) + (1-t)\{\log(1-t) - 1\},\ 0 < t < 1.$$

The results of simulations are exhibited in Figure 3(a)($c = 1$) and (b)($c = 1.4$).

Similar tendency to the one observed in Figures 2(a) and (b) can be observed now in Figures 3 (a) and (b): the risk difference is positive, which reveals a superiority of the local estimator to the global one. Although the curve of $\hat{R}(\boldsymbol{\theta}_\dagger)/h^2$ happens to be positioned uniformly below the risk difference curve in these two cases, these two curves are fairly close.

5.3. Robustness of Regression Estimators

In the rest of this section, we fix U as the pseudo-Huber loss in (4.2). We aim to check whether the global and the local estimators are robust against outliers. To do this, we generate the data set as follows.



(a) (b)

FIG 3. *Case of Bernoulli: (5.14) divided by* $c^2\sqrt{n}$ *(solid); (5.13)(dashed); (a)* $c = 1.0$*, (b)* $c = 1.4$.

Ordinal data (with no outliers) is generated according to the distribution model (5.15). The data *with* outliers is generated according to the model

$$(y_1, x_1), \ldots, (y_n, x_n) \sim_{i.i.d.} h(y, x) = (1 - p) \cdot f(y, x) + p \cdot g(y, x), \qquad (5.16)$$

where

$$g(y, x) = 1(-2 \leq y \leq -1) \times 1(2 \leq x \leq 3)$$

and $f(y, x)$ is that in (5.15). By (5.16), outliers will appear on the region

$$\{(y, x)| - 2 \leq y \leq -1, 2 \leq x \leq 3\} \qquad (5.17)$$

with a small probability p. For calculation of the estimate of the risk, we generate, in a similar manner, another independent data set

$$(y_1^*, x_1^*), \ldots, (y_S^*, x_S^*)$$

according to (5.15) and (5.16).

We investigate the robustness of estimators via the variation of the global risk. According to Section 4.3, the estimate of the risk for an estimator $\hat{\mu}$ can be defined as

$$\widehat{\mathcal{R}}(\hat{\mu}) = \frac{1}{T} \sum_{t=1}^{T} \ell(\hat{\mu}^{(t)}), \qquad (5.18)$$

where

$$\ell(\hat{\mu}^{(t)}) = \frac{1}{S} \sum_{s=1}^{S} U(y_s^* - \hat{\mu}^{(t)}(x_s^*)),$$

here $\hat{\mu}^{(t)}$ is the estimator obtained by using t-th simulated sample of size n drawn from (5.15) or (5.16). The risk comparison was carried out under the setting

$$\mu(x) = \sin x$$

TABLE 1
Comparisons by (5.18).

(5.18)×10^4			$p = 0$	$p = 0.05$	$p = 0.10$
$h = 1.5$	$\delta = 6$	$\hat{\mu}_{LSE}$	1131	2306	3249
		$\hat{\mu}_{G1}$	1122	2305	3260
		$\hat{\mu}_{G2}$	1139	2309	3258
		$\hat{\mu}_{L1}$	753	1844	2651
		$\hat{\mu}_{L2}$	772	1859	2662
$h = 3$	$\delta = 6$	$\hat{\mu}_{L1}$	1094	2245	3150
		$\hat{\mu}_{L2}$	1111	2253	3153
$h = 1.5$	$\delta = 12$	$\hat{\mu}_{LSE}$	1134	2336	3287
		$\hat{\mu}_{G1}$	1132	2336	3290
		$\hat{\mu}_{G2}$	1136	2337	3289
		$\hat{\mu}_{L1}$	763	1876	2682
		$\hat{\mu}_{L2}$	768	1880	2685
$h = 3$	$\delta = 12$	$\hat{\mu}_{L1}$	1103	2277	3179
		$\hat{\mu}_{L2}$	1108	2278	3179

$n = 100$, $p = 0, 0.05, 0.1$, $T = 100$, $S = 200$, $\delta = 6, 12$ and $h = 1.5, 3$. Table 1 includes 10^4 times the calculated estimates of the risk (5.18) for estimators:

$$\begin{aligned}\hat{\mu}_{G1}(x) &= G^{-1}(\hat{\boldsymbol{\theta}}_1^T\tilde{x}), \quad (G = u^{-1}),\\ \hat{\mu}_{G2}(x) &= G^{-1}(\hat{\boldsymbol{\theta}}_2^T\tilde{x}), \quad (G = u),\\ \hat{\mu}_{L1}(x) &= G^{-1}(\hat{\boldsymbol{\theta}}_1(x)^T\tilde{x}), \quad (G = u^{-1}),\\ \hat{\mu}_{L2}(x) &= G^{-1}(\hat{\boldsymbol{\theta}}_2(x)^T\tilde{x}), \quad (G = u),\end{aligned}$$

and $\hat{\mu}_{LSE}(x) = \hat{\boldsymbol{\theta}}_{LSE}^T\tilde{x}$, where $\hat{\boldsymbol{\theta}}_{LSE}$ is the usual least squared estimator, $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_1(x)$ are respectively obtained via the algorithms in Section 5.1.1 and 5.1.2 respectively, with $G = u^{-1}$, $\hat{\boldsymbol{\theta}}_2$ and $\hat{\boldsymbol{\theta}}_2(x)$ are those obtained with $G = u$, and $\tilde{x} = [1\ x]^T$.

We observe from Table 1 that the risk of $\hat{\mu}_{LSE}$ is mostly affected by outliers. For the case $\delta = 6$, the local estimators $\hat{\mu}_{L1}$ and $\hat{\mu}_{L2}$ perform better than the global estimators $\hat{\mu}_{G1}$ and $\hat{\mu}_{G2}$, and $\hat{\mu}_{LSE}$ irrespective of the value of p. The results for using smaller $h(= 1.5)$ are totally better than those using $h = 3$. The global estimators behave almost similarly to $\hat{\mu}_{LSE}$. The same tendency can be observed for the case $\delta = 12$. Hence, it can be claimed that the local estimators are more robust in the sense of small values of the risk (5.18).

We further investigated the behaviour of estimators by using yet another risk measure: the MISE. For an estimator $\hat{\mu}$ of μ, an estimate of integrated squared error of t-th estimator $\hat{\mu}^{(t)}$ is calculated as

$$\widehat{\text{ISE}}(\hat{\mu}^{(t)}) = \frac{1}{S}\sum_{s=1}^{S}\{\hat{\mu}^{(t)}(x_s^*) - \mu(x_s^*)\}^2,$$

Table 2
Comparisons by (5.19).

(5.19)$\times 10^4$			$p = 0$	$p = 0.05$	$p = 0.10$
		$\hat{\mu}_{LSE}$	1668	2088	2885
$h = 1.5$	$\delta = 6$	$\hat{\mu}_{G1}$	1648	2040	2813
		$\hat{\mu}_{G2}$	1687	2072	2830
		$\hat{\mu}_{L1}$	952	1479	2641
		$\hat{\mu}_{L2}$	994	1509	2644
$h = 3$	$\delta = 6$	$\hat{\mu}_{L1}$	1574	1912	2716
		$\hat{\mu}_{L2}$	1615	1946	2735
$h = 1.5$	$\delta = 12$	$\hat{\mu}_{G1}$	1663	2076	2866
		$\hat{\mu}_{G2}$	1673	2084	2870
		$\hat{\mu}_{L1}$	970	1522	2691
		$\hat{\mu}_{L2}$	980	1529	2692
$h = 3$	$\delta = 12$	$\hat{\mu}_{L1}$	1589	1949	2771
		$\hat{\mu}_{L2}$	1599	1957	2776

hence the estimate of MISE of $\hat{\mu}$ based on T-iterations can be obtained as

$$\widehat{\mathrm{MISE}}(\hat{\mu}) = \frac{1}{T}\sum_{t=1}^{T} \widehat{\mathrm{ISE}}(\hat{\mu}^{(t)}). \tag{5.19}$$

Table 2 includes 10^4 times the calculated estimates of MISE of simulated estimators.

It is seen from Table 2 that the local estimators $\hat{\mu}_{L1}$ and $\hat{\mu}_{L2}$ perform well even in this comparison using an usual risk accuracy measure such as the MISE. However there is a possibility that the local estimators might be defeated by the global estimators for a larger p.

Next, we calculated estimators using the two sets of data: with and without outliers, to investigate their robustness. A simple linear regression is considered under the same setting as above, where the true regression function is now

$$\mu(x) = 1.2 + 0.8x.$$

We calculated and compared the estimators $\hat{\mu}_{G1}(x)$, $\hat{\mu}_{L1}(x)$ and $\hat{\mu}_{LSE}(x)$. We report in Figure 4 the results of comparison for these three regression estimators under the setting $n = 100$, $p = 0, 0.1$, $\delta = 4$ and $h = 1.5$ for $\hat{\mu}_{L1}$.

Figure 4(a) illustrates the no-outliers case ($p = 0$). Obviously $\hat{\mu}_{LSE}$ (solid) gives a good fit, while $\hat{\mu}_{G1}$ (dashed) and $\hat{\mu}_{L1}$ (dotted) look slightly curved when using $G^{-1} = u$. Figure 4(b) exhibits the result where the data includes outliers generated according to (5.16) with $p = 0.1$. We observe from Figure 4(b) that all three estimators are affected by the outliers in the region (5.17), but clearly $\hat{\mu}_{LSE}$ is the most affected. The local estimator $\hat{\mu}_{L1}$ adjusts to some local features of the data, hence it becomes to be more close to the outliers than the global estimator $\hat{\mu}_{G1}$. However the fit of the local estimator around $-3 \leq x \leq 0.5$ is better than that of the global one. It becomes clear that $\hat{\mu}_{G1}$ and $\hat{\mu}_{L1}$ are more robust than $\hat{\mu}_{LSE}$. The $\hat{\mu}_{G1}$ seems more robust than the $\hat{\mu}_{L1}$ in this particular example which might be due to the fact that the outliers are clustered in one cluster. We further implemented a design where the outliers were spread in more clusters. The data *with* outliers is generated as

$$(y_1, x_1), \ldots, (y_n, x_n) \sim_{i.i.d.} h(y, x|p), \tag{5.20}$$

where

$$\begin{aligned} h(y, x|p) &= \frac{p}{2} g_L(y, x) + (1-p) f(y, x) + \frac{p}{2} g_U(y, x), \\ g_L(y, x) &= 1(-6 \le y \le -5) \times 1(-3 \le x \le -2), \\ g_U(y, x) &= 1(-2 \le y \le -1) \times 1(2 \le x \le 3) \end{aligned}$$

and $f(y, x)$ is in (5.15). The result for $p = 0.1$ is exhibited in Figure 4(c). The global estimator (dashed) behaves similar to the estimator based on LSE (solid), but the two clusters of outliers have more significant effect on the LSE estimator. The local estimator with $h = 1.5$ (dotted) fits the data in the main area $-2 \le x \le 1$, but also follows the outliers outside the main area. Hence the global estimator is more robust than the local one also in this case with two clusters of outliers.

Summarizing the results in this Section 5.3, we can say that the global and the local estimators perform well. The local estimator may sometimes be less competitive in comparison to the global with respect to resistance to outliers as it has not been constructed with this goal in mind. However, it is tailored well to minimise the combined effect of model misspecification and robustness. This is important property of the local estimator as typically in practice there is a need to deal with both these effects.

5.4. Simulation related to Theorem 7

We have designed a simulation to check Theorem 7 under the setting of Poisson regression. The ordinal data (with no outliers) is generated by

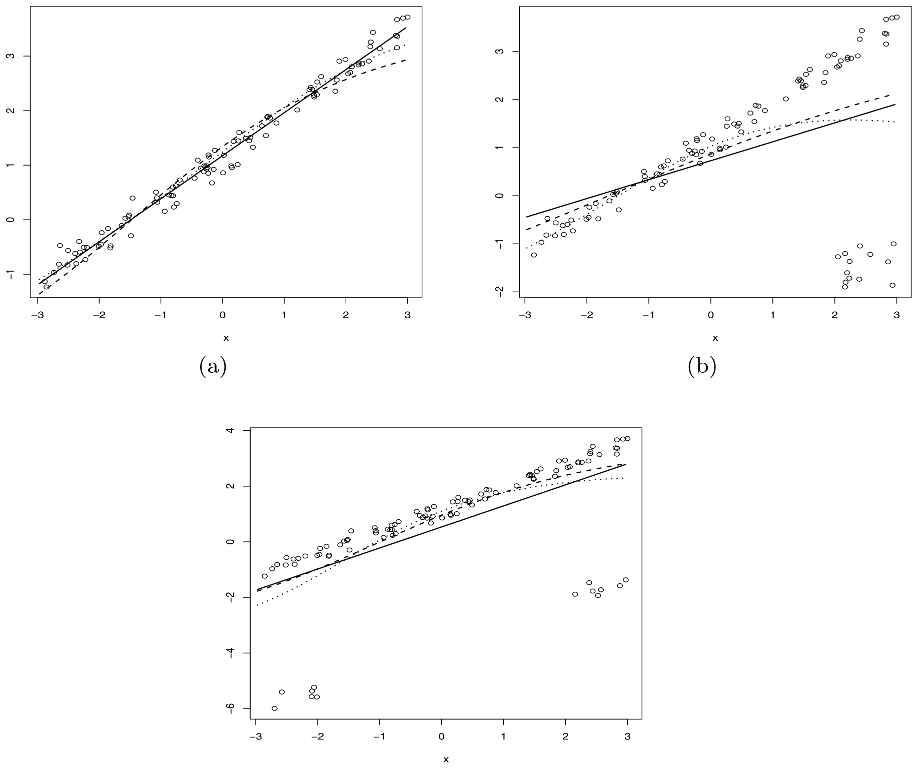
$$(y_1, x_1), \ldots, (y_n, x_n) \sim_{i.i.d.} f(y, x) = p(y|x) \cdot q(x), \tag{5.21}$$

where

$$\begin{aligned} p(y|x) &= \exp(-\mu(x)) \frac{\mu(x)^y}{y!}, \\ q(x) &= \frac{1}{6} 1(-3 \le x \le 3), \end{aligned}$$

which means that $Y|X = x \sim \text{Poisson}(\mu(x))$ and $X \sim U(-3, 3)$, a uniform distribution on the interval $[-3, 3]$. The data *with* outliers is made as

$$(y_1, x_1), \ldots, (y_n, x_n) \sim_{i.i.d.} h(y, x) = (1-p) f(y, x) + p g(y, x), \tag{5.22}$$



(a)

(b)

(c)

FIG 4. *Simulated data(circle); $\hat{\mu}_{LSE}(x)$ (solid); $\hat{\mu}_{G1}(x)$ (dashed); $\hat{\mu}_{L1}(x)$ (dotted): (a), $p = 0$ no outliers by (5.15); (b), $p = 0.1$ one cluster of outliers by (5.16); (c), $p = 0.1$ two clusters of outliers by (5.20).*

where

$$g(y, x) = 1(y = 30) \times \frac{1}{2} \cdot 1(-3 \leq x \leq -1).$$

This means that the outliers occur as $Y = 30$ on $-3 \leq x \leq -1$ with a small probability p. For calculation of the risk, we generate, in a similar manner, another independent data set $(y_1^*, x_1^*), \ldots, (y_S^*, x_S^*)$ according to (5.21) and (5.22).

To illustrate the effect of using Theorem 7, we focus on (4.9) and on the term

$$\sum_{j=1}^{d} \zeta_j(\boldsymbol{\theta}_*)^T \Psi(\boldsymbol{\theta}_*)^{-1} \zeta_j(\boldsymbol{\theta}_*)$$

in the right hand side of the Theorem. Exploiting similar Monte Carlo integrals to the ones we use in the previous Sections, we can compose, with refer to (5.18), an estimate of the risk difference between $\hat{\mu}_{G1}$ and $\hat{\mu}_{L1}$ as

$$\widehat{\mathcal{RD}}(\hat{\mu}_{G1}, \hat{\mu}_{L1})$$

$$= \widehat{\mathcal{R}}(\hat{\mu}_{G1}) - \widehat{\mathcal{R}}(\hat{\mu}_{L1})$$
$$= \frac{1}{T \cdot S} \sum_{t=1}^{T} \sum_{s=1}^{S} \left\{ U(y_s^* - \hat{\mu}_{G1}^{(t)}(x_s^*)) - U(y_s^* - \hat{\mu}_{L1}^{(t)}(x_s^*)) \right\}. \quad (5.23)$$

for an estimate of (4.9), and

$$\widehat{\mathcal{RHS}}(\boldsymbol{\theta}) = \sum_{j=1}^{d} \widetilde{\zeta}_j(\boldsymbol{\theta}_\dagger)^T \widetilde{\Psi}(\boldsymbol{\theta}_\dagger)^{-1} \widetilde{\zeta}_j(\boldsymbol{\theta}_\dagger) \quad (5.24)$$

for an estimate of $\sum_{j=1}^{d} \zeta_j(\boldsymbol{\theta}_*)^T \Psi(\boldsymbol{\theta}_*)^{-1} \zeta_j(\boldsymbol{\theta}_*)$, where

$$\widetilde{\zeta}_j(\boldsymbol{\theta}_\dagger) = \frac{1}{S} \sum_{s=1}^{S} \tilde{x}_{sj} u(y_s^* - m(x_s^*, \boldsymbol{\theta}_\dagger)) \frac{\partial}{\partial \boldsymbol{\theta}} m(x_s^*, \boldsymbol{\theta}_\dagger).$$

We have demonstrated the simulation to check Theorem 7 for $\hat{\mu}_{G1}$ and $\hat{\mu}_{L1}$ with the setting of

$$\mu(x) = 5 + 4 \sin x,$$

$T = 100$, $S = 200$ and the bandwidth $h = n^{1/4}$ was utilized for the local estimator $\hat{\mu}_{L1}$. Theorem 7 claims that the risk difference $\widehat{\mathcal{RD}}(\hat{\mu}_{G1}, \hat{\mu}_{L1})$ in (5.23) is asymptotically bounded from below by $\widehat{\mathcal{RHS}}(\boldsymbol{\theta})$ in (5.24) divided by $h^2 = n^{1/2}$:

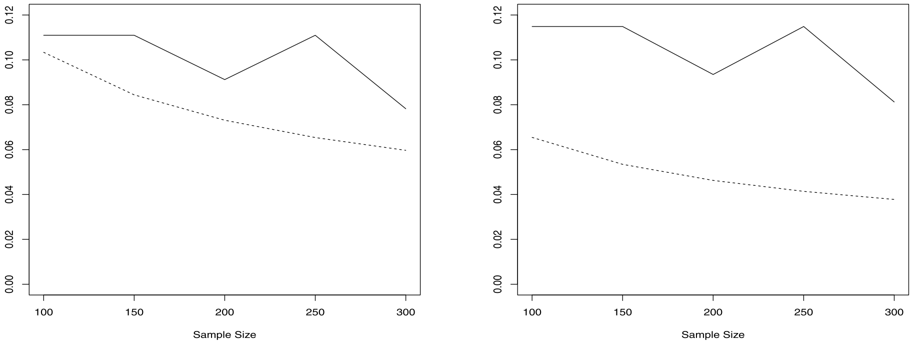
$$\widehat{\mathcal{RD}}(\hat{\mu}_{G1}, \hat{\mu}_{L1}) \geq \frac{\widehat{\mathcal{RHS}}(\boldsymbol{\theta})}{n^{1/2}}. \quad (5.25)$$

Here $\boldsymbol{\theta} = \boldsymbol{\theta}^\dagger$ is determined by the mean of 100 $\hat{\boldsymbol{\theta}}_1$'s obtained by 100 iterations for $n = 300$. Figures 5(a) and (b) illustrate the results for $n = 100, 150, 200, 250, 300$, with contamination rate $p = 0, 0.05$ and $\delta = 12$, where the solid line is $\widehat{\mathcal{RD}}(\hat{\mu}_{G1},$ $\hat{\mu}_{L1})$, and the dashed line designating $\widehat{\mathcal{RHS}}(\boldsymbol{\theta}^\dagger)/n^{1/2}$ in each figure.

We observe from Figures 5(a) and (b) that the risk difference is positive, which reveals a risk improvement by the proposed local method. The risk difference under the data including outliers (solid line in Figure 5(b)) is getting slightly bigger than that without outliers (solid line in Figure 5(a)). It can be recognized that the dashed line stays below the solid line in both Figures. The zigzag shape of the solid line could be attributed to small sample fluctuation. This demonstrates that in this example the inequality (5.25) holds not only in the case $p = 0$ but also in the case $p = 0.05$. It also holds for $p = 0.1$ although we did not include the graph for this case.

6. Discussion

This paper presents a unified way to compose a localised regression inference method by utilising the following triplet (U, G, K): a strictly convex function U for the estimation scheme with the functional Bregman divergence, the link



(a) (b)

FIG 5. *The left hand side of (5.25)(solid); The right hand side of (5.25)(dashed); (a)* $p = 0$, *(b)* $p = 0.05$.

function G in the parametric model utilised, and the kernel function K for localisation.

In most statistical analyses, the estimation scheme is considered separately from the model, and the model is often suggested irrespective of the estimation procedure: the choice of the estimation method and the construction of model are independent. A natural question arises: is there any *model* that is more suitable for the estimation scheme we would like to use? The results in Sections 3 and 4 provide an answer to this question: a specific relationship between the choice of U (for the estimation scheme in (2.3)) and G (for the model in (2.2)) yields an advantage of the use of the local estimator associated with the kernel K, see Theorems 5, 6, 7, 8 and 9. This suggests that it is beneficial to construct the model with a reference to the estimation scheme to be implemented. There is no need to care too much about the discrepancy between the model and the underlying true structure as the localisation helps to adjust to the latter.

As the choices on U and of G are interrelated (for example: $U' = u = G - \alpha$ in Theorem 6), it is not really important which one is chosen “first”. In practice, the choice can be dictated by our prior information and by the degree of confidence in this prior information. If, for example, we have a strong confidence in the form of the global regression function, then we can start with a choice of the link function G and then determining U from the relationship $U' = G - \alpha$. On the other hand, if we do not have a firm idea about the parametric form of the regression function or we wish to perform in a robust way, then we may wish to choose a suitable U first (for example, the pseudo-Huber loss) and then choose G accordingly as $U' + \alpha$. As in our paper we stressed on the possibility of model misspecification (i.e., on the case $m(\boldsymbol{x}, \boldsymbol{\theta}) \neq \mu(\boldsymbol{x})$) we have focused on the choice of the function U at a first step.

In conclusion, we can say that the approgression approach that we have adopted in this paper, seems to be the realistic practical setting. In that sense, the paper [23] can be considered a special case of our methodology. Sensible asymptotic statements about the behavior of the estimators of the true regression function (i.e., of the conditional expected value of the output given the input) in our setting can be attained when this regression function is close to some parametric model. This type of requirement is similar to the Fisher consistency in classical robustness theory. These are the types of statements that we have derived in this paper. We have revealed that the intricate relationships between the utilised regression model's inadequacy and its robustness can be analysed more conveniently by using the local approach developed in this paper. We supported our claims with a short simulation study. There are several extensions possible as a future research work. For example, additional sparseness-type penalties can be added in the main minimisation problem. Another problem to study is the data-driven choice of the constants c and δ in Section 5.

7. Proofs

7.1. Preliminary results

Lemma 1. *Under Assumptions (A0)-(A7), for any $\boldsymbol{t} \in \mathbb{D}$ and any vector $\tilde{\boldsymbol{\theta}}$ satisfying $||\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}|| < ||\hat{\boldsymbol{\theta}}(\boldsymbol{t}) - \hat{\boldsymbol{\theta}}||$, it follows that*

$$\widehat{\Psi}(\boldsymbol{t}, \tilde{\boldsymbol{\theta}}) = \Psi(\boldsymbol{\theta}_*) - \frac{1}{h^2}\widehat{V}(\boldsymbol{t}, \boldsymbol{\theta}_*) + o_p\left(\frac{1}{h^2}\right),$$

where

$$\widehat{V}(\boldsymbol{t}, \boldsymbol{\theta}) = \kappa_2 \Psi^{(2)}(\boldsymbol{t}, \boldsymbol{\theta}) - \left(\frac{h^2}{\sqrt{n}}\right)\sqrt{n}\left\{\widehat{\Psi}(\boldsymbol{\theta}) - \Psi(\boldsymbol{\theta})\right\}.$$

Lemma 2. *Under Assumptions (A0)-(A7), for any $\boldsymbol{t} \in \mathbb{D}$ and any vector $\tilde{\boldsymbol{\theta}}$ satisfying $||\tilde{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}|| < ||\hat{\boldsymbol{\theta}}(\boldsymbol{t}) - \hat{\boldsymbol{\theta}}||$, it follows that*

$$\widehat{\Psi}(\boldsymbol{t}, \tilde{\boldsymbol{\theta}})^{-1} = \Psi(\boldsymbol{\theta}_*)^{-1} + \frac{1}{h^2}\Psi(\boldsymbol{\theta}_*)^{-1}\widehat{V}(\boldsymbol{t}, \boldsymbol{\theta}_*)\Psi(\boldsymbol{\theta}_*)^{-1} + o_p\left(\frac{1}{h^2}\right).$$

Lemma 3. *Under Assumptions (A0)-(A7), it follows that*

$$\begin{aligned}&\int_{\mathbb{R}\times\mathbb{R}^d} \psi^{(\ell)}(\boldsymbol{t}, y, \boldsymbol{x}, \hat{\boldsymbol{\theta}}) dF_n(y, \boldsymbol{x}) \\ &= \int_{\mathbb{R}\times\mathbb{R}^d} \psi^{(\ell)}(\boldsymbol{t}, y, \boldsymbol{x}, \theta_*) dF(y, \boldsymbol{x}) + \frac{1}{\sqrt{n}}\hat{v}^{(\ell)}(\boldsymbol{t}, \boldsymbol{\theta}_*) + o_p\left(\frac{1}{\sqrt{n}}\right),\end{aligned}$$

for $\ell = 2, 4$, where

$$\begin{aligned}&\hat{v}^{(\ell)}(\boldsymbol{t}, \boldsymbol{\theta}) \\ &= \sqrt{n}\left\{\int_{\mathbb{R}\times\mathbb{R}^d} \psi^{(\ell)}(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta}) dF_n(y, \boldsymbol{x}) - \int_{\mathbb{R}\times\mathbb{R}^d} \psi^{(\ell)}(\boldsymbol{t}, y, \boldsymbol{x}, \boldsymbol{\theta}) dF(y, \boldsymbol{x})\right\} \\ &\quad + \Psi^{(\ell)}(\boldsymbol{t}, \boldsymbol{\theta})\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}). \qquad (7.1)\end{aligned}$$

7.2. Proof of Theorem 1

Proof Define

$$\begin{aligned}
\mathcal{F} &= \{\rho(\cdot,\cdot,\boldsymbol{\theta})|\ \boldsymbol{\theta}\in\Theta\},\\
\mathcal{F}_n(\boldsymbol{t}) &= \left\{\left(1-K\left(\frac{\cdot-\boldsymbol{t}}{h_n}\right)\right)\rho(\cdot,\cdot,\boldsymbol{\theta})|\ \boldsymbol{\theta}\in\Theta\right\}
\end{aligned}$$

Theorem 1 can be proven in almost the same way as the proof of Theorems 1 and 2 in [14], using the notions of Glivenko-Cantelli class of functions and bracketing numbers. First we note that

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|\int_{\mathbb{R}\times\mathbb{R}^d}\rho(y,\boldsymbol{x},\boldsymbol{\theta})dF_n(y,\boldsymbol{x})-\int_{\mathbb{R}\times\mathbb{R}^d}\rho(y,\boldsymbol{x},\boldsymbol{\theta})dF(y,\boldsymbol{x})\right|\\
&=\sup_{g\in\mathcal{F}}\left|\int_{\mathbb{R}\times\mathbb{R}^d}g(y,\boldsymbol{x})d(\mathbb{P}_n-P)(y,\boldsymbol{x})\right|. \qquad (7.2)
\end{aligned}$$

And, by assumptions (A2)-(A5), the bracketing number $N_{[]}(\varepsilon,\mathcal{F},L_1(P))$ is finite for any $\epsilon>0$, see Lemma 6.1 in [22]. This reveals that $\mathcal{F}$ is a Glivenko-Cantelli class and hence, the right hand side of (7.2) tends to zero as n grows. Further, with the use of (A3), we conclude that $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}_*$, see Theorem 2.4.1 in [21] and also Theorem 1 in [14].

Next our focus is on $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$. It is easily confirmed that

$$\begin{aligned}
&\sup_{\theta\in\Theta}\left|\int_{\mathbb{R}\times\mathbb{R}^d}K\left(\frac{\boldsymbol{x}-\boldsymbol{t}}{h_n}\right)\rho(y,\boldsymbol{x},\boldsymbol{\theta})dF_n(y,\boldsymbol{x})-\int_{\mathbb{R}\times\mathbb{R}^d}\rho(y,\boldsymbol{x},\boldsymbol{\theta})dF(y,\boldsymbol{x})\right|\\
&\le\sup_{\theta\in\Theta}\left|\int_{\mathbb{R}\times\mathbb{R}^d}\rho(y,\boldsymbol{x},\boldsymbol{\theta})d(\mathbb{P}_n-P)(y,\boldsymbol{x})\right|\\
&\quad+\sup_{\theta\in\Theta}\left|\int_{\mathbb{R}\times\mathbb{R}^d}\left(1-K\left(\frac{\boldsymbol{x}-\boldsymbol{t}}{h_n}\right)\right)\rho(y,\boldsymbol{x},\boldsymbol{\theta})d(\mathbb{P}_n-P)(y,\boldsymbol{x})\right|\\
&\quad+\sup_{\theta\in\Theta}\left|\int_{\mathbb{R}\times\mathbb{R}^d}\left(1-K\left(\frac{\boldsymbol{x}-\boldsymbol{t}}{h_n}\right)\right)\rho(y,\boldsymbol{x},\boldsymbol{\theta})dF(y,\boldsymbol{x})\right|\\
&=\sup_{g\in\mathcal{F}}\left|\int_{\mathbb{R}\times\mathbb{R}^d}g(y,\boldsymbol{x})d(\mathbb{P}_n-P)(y,\boldsymbol{x})\right|\\
&\quad+\sup_{g\in\mathcal{F}_n(t)}\left|\int_{\mathbb{R}\times\mathbb{R}^d}g(y,\boldsymbol{x})d(\mathbb{P}_n-P)(y,\boldsymbol{x})\right|\\
&\quad+\sup_{\theta\in\Theta}\left|\int_{\mathbb{R}\times\mathbb{R}^d}\left(1-K\left(\frac{\boldsymbol{x}-\boldsymbol{t}}{h_n}\right)\right)\rho(y,\boldsymbol{x},\boldsymbol{\theta})dF(y,\boldsymbol{x})\right|. \qquad (7.3)
\end{aligned}$$

The first term in the right hand side in (7.3) appears in (7.2), hence it tends to zero. It can be proven by using Lemmas 1 to 5 in [14] that the second and third terms also tend to zero. This together with (A3) leads the convergence in probability of $\hat{\boldsymbol{\theta}}(\boldsymbol{t})$ to $\boldsymbol{\theta}$. □

7.3. Proof of Theorem 2

Theorem 2 can be proven by almost the same way as Theorem 1 in [18].

7.4. Proof of Theorem 3

Theorem 3 can be obtained in a similar manner to Theorem 2 in [18].

7.5. Proof of Theorem 4

Proof It is easily confirmed from Theorem 1 that

$$\begin{aligned}
&\sqrt{n}\left\{\hat{\boldsymbol{\theta}}(\boldsymbol{t})-\boldsymbol{\theta}_*-\frac{\kappa_2}{h^2}\Psi(\boldsymbol{\theta}_*)^{-1}\int_{\mathbb{R}\times\mathbb{R}^d}\psi^{(2)}(\boldsymbol{t},y,\boldsymbol{x},\boldsymbol{\theta}_*)dF(y,\boldsymbol{x})\right\}\\
&=\sqrt{n}\left\{\hat{\boldsymbol{\theta}}(\boldsymbol{t})-\hat{\boldsymbol{\theta}}-\frac{\kappa_2}{h^2}\Psi(\boldsymbol{\theta}_*)^{-1}\int_{\mathbb{R}\times\mathbb{R}^d}\psi^{(2)}(\boldsymbol{t},y,\boldsymbol{x},\boldsymbol{\theta}_*)dF(y,\boldsymbol{x})\right\}\\
&\qquad\qquad+\sqrt{n}\{\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*\}\\
&=\sqrt{n}\{\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*\}+\frac{\kappa_2}{h^2}\Psi(\boldsymbol{\theta}_*)^{-1}\left\{\hat{v}^{(2)}(\boldsymbol{t},\boldsymbol{\theta}_*)+O_p\left(\frac{\sqrt{n}}{h^4}\right)\right\}\\
&=\sqrt{n}\{\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_*\}+O_p\left(\frac{1}{h^2}\right).
\end{aligned}$$

Hence the result follows by the asymptotic normality of $\hat{\boldsymbol{\theta}}$. A simple application of delta method immediately yields the result for $\hat{\mu}_L(\boldsymbol{t})$. □

7.6. Proof of Theorem 5

Proof By the definition of Bregman divergence, it is easily confirmed that

$$\begin{aligned}
&\mathfrak{R}(\hat{\mu}_G,\hat{\mu}_L)\\
&=E\left[D_{U^*}(u(\hat{\mu}_G),u(\hat{\mu}_L))\right]\\
&\qquad+E\left[\int_{\mathbb{R}^d}\{\hat{\mu}_L(\boldsymbol{x})-\mu(\boldsymbol{x})\}\{u(\hat{\mu}_G(\boldsymbol{x}))-u(\hat{\mu}_L(\boldsymbol{x}))\}q(\boldsymbol{x})d\boldsymbol{x}\right],
\end{aligned}$$

here we note that the first term of the right hand side is nonnegative. By referring to Theorem 2, we can put

$$\hat{\boldsymbol{\theta}}(\boldsymbol{x})=\hat{\boldsymbol{\theta}}+\frac{\kappa_2}{h^2}\Psi(\boldsymbol{\theta}_*)^{-1}\int_{\mathbb{R}\times\mathbb{R}^d}\psi^{(2)}(\boldsymbol{x},y,\boldsymbol{z},\boldsymbol{\theta}_*)dF(y,\boldsymbol{z})+O_P\left(\frac{1}{h^2\sqrt{n}}\right).$$

Using this, we can expand $u(\hat{\mu}_G)-u(\hat{\mu}_L)$ as

$$u(\hat{\mu}_G(\boldsymbol{x}))-u(\hat{\mu}_L(\boldsymbol{x}))$$

$$
\begin{aligned}
&= u(G^{-1}(\hat{\boldsymbol{\theta}}^T\tilde{\boldsymbol{x}})) - u(G^{-1}(\hat{\boldsymbol{\theta}}(\boldsymbol{x})^T\tilde{\boldsymbol{x}}))\\
&= -\frac{\kappa_2}{h^2}\cdot\frac{u'(G^{-1}(\hat{\boldsymbol{\theta}}^T\tilde{\boldsymbol{x}}))}{G'(G^{-1}(\hat{\boldsymbol{\theta}}^T\tilde{\boldsymbol{x}}))}\left[\int_{\mathbb{R}\times\mathbb{R}^d}\psi^{(2)}(\boldsymbol{x},y,\boldsymbol{z},\boldsymbol{\theta}_*)dF(y,\boldsymbol{z})\right]^T\Psi(\boldsymbol{\theta}_*)^{-1}\tilde{\boldsymbol{x}}\\
&\qquad\qquad + o_P\left(\frac{1}{h^2}\right).
\end{aligned}
$$

Further, since $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}_*$, Theorem 2 yields

$$\hat{\mu}_L(\boldsymbol{x}) - \mu(\boldsymbol{x}) = G^{-1}(\boldsymbol{\theta}_*^T\tilde{\boldsymbol{x}}) - \mu(\boldsymbol{x}) + o_P(1).$$

By also noting (2.8), (2.17) and (2.9), the latter calculations lead us to

$$
\begin{aligned}
&E\left[\int_{\mathbb{R}^d}\{\hat{\mu}_L(\boldsymbol{x})-\mu(\boldsymbol{x})\}\{u(\hat{\mu}_G(\boldsymbol{x}))-u(\hat{\mu}_L(\boldsymbol{x}))\}q(\boldsymbol{x})d\boldsymbol{x}\right]\\
&= -\frac{\kappa_2}{h^2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}||\boldsymbol{z}-\boldsymbol{x}||^2\psi(\mu(\boldsymbol{x}),\boldsymbol{x},\boldsymbol{\theta}_*)^T\Psi(\boldsymbol{\theta}_*)^{-1}\psi(\mu(\boldsymbol{z}),\boldsymbol{z},\boldsymbol{\theta}_*)q(\boldsymbol{x})q(\boldsymbol{z})d\boldsymbol{x}d\boldsymbol{z}\\
&\qquad\qquad + o\left(\frac{1}{h^2}\right)\\
&= \frac{2\kappa_2}{h^2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\boldsymbol{z}^T\boldsymbol{x}\psi(\mu(\boldsymbol{x}),\boldsymbol{x},\boldsymbol{\theta}_*)^T\Psi(\boldsymbol{\theta}_*)^{-1}\psi(\mu(\boldsymbol{z}),\boldsymbol{z},\boldsymbol{\theta}_*)q(\boldsymbol{x})q(\boldsymbol{z})d\boldsymbol{x}d\boldsymbol{z}\\
&\qquad\qquad + o\left(\frac{1}{h^2}\right)\\
&= \frac{2\kappa_2}{h^2}\sum_{j=1}^d\eta_j(\boldsymbol{\theta}_*)^T\Psi(\boldsymbol{\theta}_*)^{-1}\eta_j(\boldsymbol{\theta}_*) + o\left(\frac{1}{h^2}\right).
\end{aligned}
$$

This completes the proof. □

7.7. Proof of Theorem 6

Proof We start with the notation $\psi(y,\boldsymbol{x},\boldsymbol{\theta}) = c_1(y,\boldsymbol{x},\boldsymbol{\theta})\widetilde{\boldsymbol{x}}$, where

$$c_1(y,\boldsymbol{x},\boldsymbol{\theta}) = \left\{G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}) - y\right\}\frac{u'(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))}{G'(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))},$$

see (2.8). Hence, by (2.18), the partial derivative of $c_1(y,\boldsymbol{x},\boldsymbol{\theta})$ w.r.t. $\boldsymbol{\theta}$ is essential to obtain $\Psi(\boldsymbol{\theta}_*)$. A straightforward calculation yields

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\rho(y,\boldsymbol{x},\boldsymbol{\theta}) = c_2(y,\boldsymbol{x},\boldsymbol{\theta})\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T, \tag{7.4}$$

where

$$c_2(y,\boldsymbol{x},\boldsymbol{\theta})$$

$$= \frac{\left\{G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}) - y\right\}}{G'(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))} \left\{ \frac{u''(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))}{G'(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))} - \frac{u'(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))G''(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))}{\{G'(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))\}^2} \right\} + \frac{u'(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))}{G'(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))^2}. \tag{7.5}$$

This notation (7.4) enables us to express

$$\begin{aligned}\Psi(\boldsymbol{\theta}_*) &= \int_{\mathbb{R}\times\mathbb{R}^d} \frac{\partial}{\partial\boldsymbol{\theta}}\psi(y,\boldsymbol{x},\boldsymbol{\theta}_*)^T dF(y,\boldsymbol{x}) \\ &= \int_{\mathbb{R}\times\mathbb{R}^d} \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\rho(y,\boldsymbol{x},\boldsymbol{\theta}_*) f(y,\boldsymbol{x}) dy d\boldsymbol{x} \\ &= \int_{\mathbb{R}^d} c_2(\mu(\boldsymbol{x}),\boldsymbol{x},\boldsymbol{\theta}_*)\widetilde{\boldsymbol{x}}\widetilde{\boldsymbol{x}}^T q(\boldsymbol{x}) d\boldsymbol{x}.\end{aligned}$$

Therefore $\Psi(\boldsymbol{\theta}_*)$ is positive definite, provided that $c_2(\mu(\boldsymbol{x}),\boldsymbol{x},\boldsymbol{\theta}_*)$ is positive for any $\boldsymbol{x}$.

Now we see by a careful calculation and by referring to (7.5) that

$$\begin{aligned} c_2(\mu(\boldsymbol{x}),\boldsymbol{x},\boldsymbol{\theta}) &= \frac{\left\{G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}) - \mu(\boldsymbol{x})\right\}}{G'(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))} \cdot \frac{d}{dt}\left(\frac{u'(t)}{G'(t)}\right)\Big|_{t=G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}})} \\ &\qquad + \frac{u'(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))}{G'(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))^2},\end{aligned}$$

which becomes positive by the choice of $G \equiv u + \alpha$, without any dependence on μ. Under this choice we get (3.5). □

7.8. *Proof of Theorem 7*

Proof Since U is strictly convex, it follows that

$$\mathfrak{R}(\hat{\mu}_G,\hat{\mu}_L) \geq E\left[\int_{\mathbb{R}\times\mathbb{R}^d} u(y - m(\boldsymbol{x},\hat{\boldsymbol{\theta}}(\boldsymbol{x})))\{m(\boldsymbol{x},\hat{\boldsymbol{\theta}}(\boldsymbol{x})) - m(\boldsymbol{x},\hat{\boldsymbol{\theta}})\}dF(y,\boldsymbol{x})\right]. \tag{7.6}$$

By a repeated use of Theorems 2 and 3 and the fact (induced from (4.4) and (4.5)) that

$$\psi(y,\boldsymbol{x},\boldsymbol{\theta}) = u(y - m(\boldsymbol{x},\boldsymbol{\theta}))\left(-\frac{\partial}{\partial\boldsymbol{\theta}}m(\boldsymbol{x},\boldsymbol{\theta})\right),$$

the right hand side of (7.6) can be evaluated as

$$E\left[\int_{\mathbb{R}\times\mathbb{R}^d} u(y - m(\boldsymbol{x},\hat{\boldsymbol{\theta}}(\boldsymbol{x})))\{m(\boldsymbol{x},\hat{\boldsymbol{\theta}}(\boldsymbol{x})) - m(\boldsymbol{x},\hat{\boldsymbol{\theta}})\}dF(y,\boldsymbol{x})\right]$$

$$
\begin{aligned}
&= \int_{\mathbb{R}\times\mathbb{R}^d} u(y - m(\boldsymbol{x},\boldsymbol{\theta}_*))\frac{\partial}{\partial\boldsymbol{\theta}} m(\boldsymbol{x},\boldsymbol{\theta}_*)^T E[\hat{\boldsymbol{\theta}}(\boldsymbol{x}) - \hat{\boldsymbol{\theta}}]dF(y,\boldsymbol{x}) + o\left(\frac{1}{h^2}\right)\\
&= -\frac{\kappa_2}{h^2}\int_{\mathbb{R}\times\mathbb{R}^d}\left[\psi(y,\boldsymbol{x},\boldsymbol{\theta}_*)^T\Psi(\theta_*)^{-1}\int_{\mathbb{R}\times\mathbb{R}^d}\psi^{(2)}(\boldsymbol{x},w,\boldsymbol{z},\boldsymbol{\theta}_*)dF(w,\boldsymbol{z})\right]dF(y,\boldsymbol{x})\\
&\qquad + o\left(\frac{1}{h^2}\right)\\
&= -\frac{\kappa_2}{h^2}\int_{\mathbb{R}\times\mathbb{R}^d}\int_{\mathbb{R}\times\mathbb{R}^d}||\boldsymbol{z}-\boldsymbol{x}||^2\psi(y,\boldsymbol{x},\boldsymbol{\theta}_*)\Psi(\boldsymbol{\theta}_*)^{-1}\psi(w,\boldsymbol{z},\boldsymbol{\theta}_*)dF(w,\boldsymbol{z})dF(y,\boldsymbol{x})\\
&\qquad + o\left(\frac{1}{h^2}\right),
\end{aligned}
$$

which is the claim of Theorem 7. □

7.9. *Proof of Theorem 8*

Proof It is easily confirmed that

$$
\begin{aligned}
u(t) = U'(t) &= \frac{\delta t}{(\delta^2+t^2)^{1/2}} = \frac{t}{\sqrt{1+(t/\delta)^2}}, && (7.7)\\
U''(t) &= \frac{\delta^3}{(\delta^2+t^2)^{3/2}} = \frac{1}{(\sqrt{1+(t/\delta)^2})^3}, && (7.8)\\
U'''(t) &= \frac{-3\delta^3 t}{(\delta^2+t^2)^{5/2}} = \frac{-3t}{\delta^2(\sqrt{1+(t/\delta)^2})^5}. && (7.9)
\end{aligned}
$$

Now we focus on $\tilde{c}_2(y,\boldsymbol{x},\boldsymbol{\theta}_*)$ in (4.8). Clearly, from the form of $\Psi(\boldsymbol{\theta}_*)$ in (4.7), it is seen that it is positive definite provided that

$$E[\tilde{c}_2(Y,\boldsymbol{X},\boldsymbol{\theta}_*)|\boldsymbol{X}=\boldsymbol{x}] = \frac{1}{\{G'(G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}))\}^2}E[J(Y,\boldsymbol{X},\boldsymbol{\theta}_*)|\boldsymbol{X}=\boldsymbol{x}] > 0$$

for almost all $\boldsymbol{x}\in\mathbb{D}$, where

$$J(Y,\boldsymbol{X},\boldsymbol{\theta}_*) = U''(Y - G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{X}})) + U'(Y - G^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{X}}))\frac{U'''(u^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{X}}))}{U''(u^{-1}(\boldsymbol{\theta}^T\widetilde{\boldsymbol{X}}))}$$

Noting that $U''>0$, it suffices to show that for any $\boldsymbol{x}\in\mathbb{D}$

$$\tilde{J}(\boldsymbol{\theta}_*) = E[U''(Y-t)]U''(t) + E[U'(Y-t)]U'''(t) > 0$$

holds for sufficiently large $\delta>0$, where we have put $t = G^{-1}(\boldsymbol{\theta}_*^T\widetilde{\boldsymbol{x}}) = u^{-1}(\boldsymbol{\theta}_*^T\widetilde{\boldsymbol{x}})$.

Using (7.7), (7.8) and (7.9), we see that

$\tilde{J}(\boldsymbol{\theta}_*)$

$$\begin{aligned}
&= \frac{\delta^6}{(\delta^2+t^2)^{3/2}} E\left[\frac{1}{\{\delta^2+(Y-t)^2\}^{3/2}}\right] - \frac{3\delta^4 t}{(\delta^2+t^2)^{5/2}} E\left[\frac{Y-t}{\{\delta^2+(Y-t)^2\}^{1/2}}\right] \\
&\geq \frac{\delta^6}{(\delta^2+t^2)^{3/2}} E\left[\frac{1}{\{\delta^2+(Y-t)^2\}^{3/2}}\right] - \frac{3\delta^4 t}{(\delta^2+t^2)^{5/2}} E\left[\frac{Y}{\{\delta^2+(Y-t)^2\}^{1/2}}\right] \\
&\geq \frac{\delta^6}{(\delta^2+t^2)^{3/2}} E\left[\frac{1}{\{\delta^2+(Y-t)^2\}^{3/2}}\right] - \frac{3\delta^4 |t|}{(\delta^2+t^2)^{5/2}} \left|E\left[\frac{Y}{\{\delta^2+(Y-t)^2\}^{1/2}}\right]\right| \\
&\geq \frac{\delta^6}{(\delta^2+t^2)^{3/2}} E\left[\frac{1}{\{\delta^2+(Y-t)^2\}^{3/2}}\right] - \frac{3\delta^4 |t|}{(\delta^2+t^2)^{5/2}} E\left[\frac{|Y|}{\{\delta^2+(Y-t)^2\}^{1/2}}\right] \\
&\geq \frac{\delta^6}{(\delta^2+t^2)^{3/2}} E\left[\frac{1}{\{\delta^2+(Y-t)^2\}^{3/2}}\right] - \frac{3\delta^4 |t|}{(\delta^2+t^2)^{5/2}} E\left[\frac{|Y|}{\delta}\right] \\
&= \frac{\delta^3}{(\delta^2+t^2)^{3/2}} \left[E\left[\frac{\delta^3}{\{\delta^2+(Y-t)^2\}^{3/2}}\right] - \frac{3|t|}{(\delta^2+t^2)} E[|Y|]\right] \\
&= \frac{3\delta^3}{(\delta^2+t^2)^{5/2}} \left[\frac{(\delta^2+t^2)}{3} E\left[\frac{\delta^3}{\{\delta^2+(Y-t)^2\}^{3/2}}\right] - |t| E[|Y|]\right] \\
&\geq \frac{3\delta^3}{(\delta^2+t^2)^{5/2}} \left[\frac{\delta^5}{3} E\left[\frac{1}{\{\delta^2+(Y-t)^2\}^{3/2}}\right] - |t| E[|Y|]\right] \\
&= \frac{3\delta^3}{(\delta^2+t^2)^{5/2}} \left[\frac{\delta^2}{3} E\left[\frac{1}{\{1+\{(Y-t)/\delta\}^2\}^{3/2}}\right] - |t| E[|Y|]\right] \\
&\geq \frac{3}{(1+(t/\delta)^2)^{5/2}} \left[\frac{1}{3} E\left[\frac{1}{\{1+\{(Y-t)/\delta\}^2\}^{3/2}}\right] - \frac{|t| \cdot E[|Y|]}{\delta^2}\right].
\end{aligned}$$

Since

$$\left|\frac{1}{\{1+\{(y-t)/\delta\}^2\}^{3/2}}\right| \leq 1,$$

the dominated convergence theorem yields that

$$E\left[\frac{1}{\{1+\{(Y-t)/\delta\}^2\}^{3/2}}\right] = E\left[\frac{1}{\{1+\{(Y-u^{-1}(\boldsymbol{\theta}_*^T\widetilde{\boldsymbol{x}}))/\delta\}^2\}^{3/2}}\right] \to 1 \tag{7.10}$$

for any $\boldsymbol{x} \in \mathbb{D}$ as δ increases. Furthermore,

$$\frac{|t| \cdot E[|Y|]}{\delta^2} = \frac{|u^{-1}(\boldsymbol{\theta}_*^T\widetilde{\boldsymbol{x}})| \cdot E[\,|Y|\,|\boldsymbol{X}=\boldsymbol{x}]}{\delta^2} \leq \frac{M(u^{-1})M}{\delta^2} \to 0, \tag{7.11}$$

as δ increases. Hence, by (7.10) and (7.11), for arbitrary $0 < \epsilon < 1/3$, there exists $\delta_0 > 0$ such that

$$\frac{1}{3} E\left[\frac{1}{\{1+\{(Y-u^{-1}(\boldsymbol{\theta}_*^T\widetilde{\boldsymbol{x}}))/\delta\}^2\}^{3/2}}\right] - \frac{M(u^{-1})M}{\delta^2} > \left(\frac{1}{3}-\frac{\epsilon}{3}\right) - \frac{\epsilon}{3} > \frac{1}{9}$$

when $\delta > \delta_0$. This implies for $\delta > \delta_0$ that

$$\begin{aligned}
&\tilde{J}(\boldsymbol{\theta}_*) \\
&\geq \frac{3}{(1+(u^{-1}(\boldsymbol{\theta}_*^T\widetilde{\boldsymbol{x}})/\delta)^2)^{5/2}}
\end{aligned}$$

$$
\begin{aligned}
&\times \left[\frac{1}{3}E\left[\frac{1}{\{1+\{(Y-u^{-1}(\boldsymbol{\theta}_*^T\widetilde{\boldsymbol{x}}))/\delta\}^2\}^{3/2}}\right]-\frac{M(u^{-1})\cdot M}{\delta^2}\right] \\
&\geq \frac{3}{(1+(u^{-1}(\boldsymbol{\theta}_*^T\widetilde{\boldsymbol{x}})/\delta)^2)^{5/2}}\left(\frac{1}{9}\right) \\
&= \frac{1}{3(1+(u^{-1}(\boldsymbol{\theta}_*^T\widetilde{\boldsymbol{x}})/\delta)^2)^{5/2}} \\
&\geq \frac{1}{3(1+(M(u^{-1})/\delta)^2)^{5/2}}
\end{aligned}
$$

which is positive for any $\boldsymbol{x} \in \mathbb{D}$. □

7.10. Proof of Theorem 9

Proof For a given $U \in \mathscr{U}_0$, we choose the link function $G = u^{-1} = (U')^{-1}$. By observing the form (4.8), we need to derive the following equalities:

$$
\begin{aligned}
G'(t) &= (u^{-1})'(t) = \frac{1}{u'(u^{-1}(t))} \\
G''(t) &= (u^{-1})''(t) = \left(\frac{1}{u'(u^{-1}(t))}\right)' = -\frac{u''(u^{-1}(t))}{\{u'(u^{-1}(t))\}^3},
\end{aligned}
$$

from which it follows that

$$
G'(G^{-1}(\eta)) = (u^{-1})'(u(\eta)) = \frac{1}{u'(u^{-1}(u(\eta)))} = \frac{1}{u'(\eta)} = \frac{1}{U''(\eta)}, \tag{7.12}
$$

$$
G''(G^{-1}(\eta)) = G''(u(\eta)) = -\frac{u''(\eta)}{u'(\eta)^3} = -\frac{U'''(\eta)}{U''(\eta)^3}, \tag{7.13}
$$

$$
\frac{G''(G^{-1}(\eta))}{G'(G^{-1}(\eta))} = -\frac{U'''(\eta)}{U''(\eta)^2}, \tag{7.14}
$$

where $\eta = \boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}$. By substituting (7.12), (7.13) and (7.14) into (4.8), we have

$$
\begin{aligned}
\tilde{c}_2(y,\boldsymbol{x},\boldsymbol{\theta}) &= \left(\frac{1}{U''(\eta)}\right)^{-2}\left\{U''(y-u(\eta))+u(y-u(\eta))\left(-\frac{U'''(\eta)}{U''(\eta)^2}\right)\right\} \\
&= U''(\eta)^2\left\{U''(y-u(\eta))-U'(y-u(\eta))\left(\frac{U'''(\eta)}{U''(\eta)^2}\right)\right\} \\
&= U''(\eta)^2U''(y-u(\eta))-U'(y-u(\eta))U'''(\eta).
\end{aligned} \tag{7.15}
$$

Now we start to deal with the pseudo-Huber loss defined in (4.2). Substituting (7.7), (7.8) and (7.9) into (7.15) we get

$$
\tilde{c}_2(y,\boldsymbol{x},\boldsymbol{\theta})
$$

$$
\begin{aligned}
&= \frac{1}{(\sqrt{1+(\eta/\delta)^2})^6}\frac{1}{[\sqrt{1+\{(y-u(\eta))/\delta\}^2}]^3} \\
&\qquad -\frac{(y-u(\eta))}{\sqrt{1+\{(y-u(\eta))/\delta\}^2}}\frac{-3\eta}{\delta^2(\sqrt{1+(\eta/\delta)^2})^5} \\
&= \frac{1}{(\sqrt{1+(\eta/\delta)^2})^6}\frac{1}{[\sqrt{1+\{(y-u(\eta))/\delta\}^2}]^3} \\
&\qquad +\frac{3\eta(y-u(\eta))}{\delta^2\sqrt{1+\{(y-u(\eta))/\delta\}^2}}\frac{1}{(\sqrt{1+(\eta/\delta)^2})^5}.
\end{aligned}
$$

Hence, noting $\eta = \boldsymbol{\theta}^T\widetilde{\boldsymbol{x}}$ again, we have the evaluation

$$
\begin{aligned}
&E[\tilde{c}_2(Y,\boldsymbol{X},\boldsymbol{\theta})|\boldsymbol{X}=\boldsymbol{x}] \\
&= \frac{1}{(\sqrt{1+(\eta/\delta)^2})^6}E\left[\frac{1}{[\sqrt{1+\{(Y-u(\eta))/\delta\}^2}]^3}\right] \\
&\qquad +\frac{1}{\delta^2(\sqrt{1+(\eta/\delta)^2})^5}E\left[\frac{3\eta(Y-u(\eta))}{\sqrt{1+\{(Y-u(\eta))/\delta\}^2}}\right] \\
&\geq \frac{1}{(\sqrt{1+(\eta/\delta)^2})^6}\left[E\left[\frac{1}{[\sqrt{1+\{(Y-u(\eta))/\delta\}^2}]^3}\right]\right. \\
&\qquad \left. -\frac{\sqrt{1+(\eta/\delta)^2}}{\delta^2}E\left[\frac{3|\eta(Y-u(\eta))|}{\sqrt{1+\{(Y-u(\eta))/\delta\}^2}}\right]\right] \\
&\geq \frac{1}{(\sqrt{1+(\eta/\delta)^2})^6}\left[E\left[\frac{1}{[\sqrt{1+\{(Y-u(\eta))/\delta\}^2}]^3}\right]\right. \\
&\qquad \left. -\frac{3\sqrt{1+(\eta/\delta)^2}}{\delta^2}E\left[|\eta(Y-u(\eta))|\right]\right] \\
&\geq \frac{1}{(\sqrt{1+(\eta/\delta)^2})^6}\left[E\left[\frac{1}{[\sqrt{1+\{(Y-u(\eta))/\delta\}^2}]^3}\right]\right. \\
&\qquad \left. -\frac{3\sqrt{1+(\eta/\delta)^2}\cdot|\eta|\cdot\{E[|Y|]+|u(\eta)|\}}{\delta^2}\right].
\end{aligned}
\tag{7.16}
$$

Since

$$
\left|\frac{1}{\{1+\{(y-u(\eta))/\delta\}^2\}^{3/2}}\right| \leq 1,
$$

the dominated convergence theorem yields that

$$
E\left[\frac{1}{\{1+\{(Y-u(\eta))/\delta\}^2\}^{3/2}}\right] = E\left[\frac{1}{\{1+\{(Y-u(\boldsymbol{\theta}_*^T\widetilde{\boldsymbol{x}}))/\delta\}^2\}^{3/2}}\right] \to 1 \tag{7.17}
$$

for any $\boldsymbol{x} \in \mathbb{D}$ as δ increases. On the other hand, the compact support $\mathbb{D}$ of $\boldsymbol{X}$, the assumptions (A2) and (A4) guarantee that there exist constants L, $M(u)$

and M such that

$$|\eta| = |\boldsymbol{\theta}_*^T \widetilde{\boldsymbol{x}}| \leq L, \ |u(\eta)| = |u(\boldsymbol{\theta}_*^T \widetilde{\boldsymbol{x}})| \leq M(u), \ E[|Y||\boldsymbol{X} = \boldsymbol{x}] \leq M \tag{7.18}$$

uniformly on $\boldsymbol{x}$. By combining (7.17) and (7.18), for any ε satisfying $0 < \varepsilon < 3^{-1}$, there exists δ_0 such that

$$E\left[\frac{1}{[\sqrt{1+\{(Y-u(\eta))/\delta\}^2}]^3}\right] > 1-\varepsilon, \tag{7.19}$$

$$\begin{aligned} \frac{3\sqrt{1+(\eta/\delta)^2}\cdot|\eta|\cdot\{E[|Y|]+|u(\eta)|\}}{\delta^2} &\leq \frac{3\sqrt{1+L^2}\cdot L\cdot\{M+M(u)\}}{\delta^2} \\ &< \varepsilon \end{aligned} \tag{7.20}$$

for any $\delta > \delta_0$. Evaluations (7.16), (7.19) and (7.20) furnish to reach, for $\delta > \delta_0$,

$$\begin{aligned} E[\tilde{c}_2(Y, \boldsymbol{X}, \boldsymbol{\theta})|\boldsymbol{X} = \boldsymbol{x}] &\geq \frac{1}{(\sqrt{1+(\eta/\delta)^2})^6}\,[(1-\varepsilon)-\varepsilon] \\ &\geq \frac{1-2\varepsilon}{(\sqrt{1+L^2})^6} \\ &\geq \frac{1}{3(\sqrt{1+L^2})^6}, \end{aligned}$$

which is positive for any $\boldsymbol{x} \in \mathbb{D}$. This confirms that the matrix (4.7) is positive definite for sufficiently large δ. □

7.11. Proof of Lemmas

Lemmas 1, 2 and 3 can be established in a similar way as in Lemmas 1, 2 and 3 of [18].

Acknowledgments

The authors would like to thank the associate editor and a referee for their useful comments.

References

[1] Amari, S. I. (2016). *Information Geometry and Its Applications.* Springer. MR3495836

[2] 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

[3] Basu, A., Shioya, H. and Park, C. (2011). *Statistical Inference. The Minimum Distance Approach.* CRC press. MR2830561

[4] Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics* **7** 200–217. MR0215617

[5] Bunke, O. and Bunke, H., eds. (1986). *Statistical Inference in Linear Models.* Wiley, Chichester.

[6] Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association* **74** 829–836. MR0556476

[7] Copas, J. (1995). Local likelihood based on kernel censoring. *Journal of the Royal Statistical Society: Series B* **57** 221–235. MR1325387

[8] Eguchi, S. and Copas, J. (1998). A class of local likelihood methods and near-parametric asymptotics. *Journal of the Royal Statistical Society: Series B* **60** 709–724. MR1649531

[9] Eguchi, S., Kim, T. Y. and Park, B. U. (2003). Local likelihood method: a bridge over parametric and nonparametric regression. *Nonparametric Statistics* **15** 665–683. MR2030279

[10] Ferrari, D. and Yang, Y. (2010). Maximum L_q-likelihood estimation. *Annals of Statistics* **38** 753–783. MR2604695

[11] Frigyik, B. A., Srivastava, S. and Gupta, M. R. (2008). Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Transactions on Information Theory* **54** 5130–5139. MR2589887

[12] Hartley, R. and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision. 2nd Edition.* Cambridge University Press. MR1823669

[13] Hjort, N. and Jones, M. C. (1996). Locally parametric nonparametric density estimation. *Annals of Statistics* 1619–1647. MR1416653

[14] Kawamura, K. and Naito, K. (2019). Asymptotic theory for local estimators based on Bregman divergence. *Canadian Journal of Statistics* **47** 628–652. MR4035793

[15] Loader, C. (1996). Local likelihood density estimation. *Annals of Statistics* **24** 1602–1618. MR1416652

[16] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models. 2nd Edition*. Chapman and Hall. MR3223057

[17] McCulloch, C. E., Searle, S. R. and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models. 2nd Edition*. Wiley. MR2431553

[18] Penev, S. and Naito, K. (2018). Locally robust methods and near-parametric asymptotics. *Journal of Multivariate Analysis* **167** 395–417. MR3830654

[19] Schott, J. R. (1997). *Matrix Analysis for Statistics*. Wiley. MR1421574

[20] Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association* **82** 559–567. MR0898359

[21] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer. MR1385671

[22] Wellner, J. A. (2005). Empirical processes: theory and applications.

Notes for a course given at Delft University of Technology. MR1394050

[23] Zhang, C., Jiang, Y. and Shang, Z. (2009). New aspects of Bregman divergence in regression and classification with parametric and nonparametric estimation. *Canadian Journal of Statistics* **37** 119–139. MR2509465