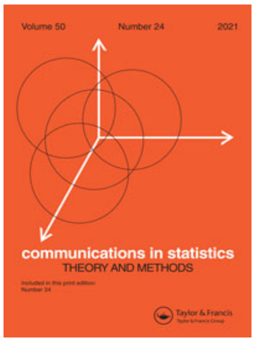

Communications in Statistics - Theory and Methods

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

# A multi-step kernel–based regression estimator that adapts to error distributions of unknown form

Jan G. De Gooijer & Hugo Reichardt

**To cite this article:** Jan G. De Gooijer & Hugo Reichardt (2021) A multi-step kernel–based regression estimator that adapts to error distributions of unknown form, Communications in Statistics - Theory and Methods, 50:24, 6211-6230, DOI: 10.1080/03610926.2020.1741625

**To link to this article:** https://doi.org/10.1080/03610926.2020.1741625

Published online: 19 Mar 2020.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2021, VOL. 50, NO. 24, 6211–6230
https://doi.org/10.1080/03610926.2020.1741625

Taylor & Francis
Taylor & Francis Group

# A multi-step kernel–based regression estimator that adapts to error distributions of unknown form

Jan G. De Gooijer[a] and Hugo Reichardt[b]

[a]Amsterdam School of Economics, University of Amsterdam, Amsterdam, The Netherlands; [b]London School of Economics, London, UK

**ABSTRACT**
For linear regression models, we propose and study a multi-step kernel density-based estimator that is adaptive to unknown error distributions. We establish asymptotic normality and almost sure convergence. An efficient EM algorithm is provided to implement the proposed estimator. We also compare its finite sample performance with five other adaptive estimators in an extensive Monte Carlo study of eight error distributions. Our method generally attains high mean-square-error efficiency. An empirical example illustrates the gain in efficiency of the new adaptive method when making statistical inference about the slope parameters in three linear regressions.

**ARTICLE HISTORY**
Received 16 August 2018
Accepted 6 March 2020

**KEYWORDS**
Adaptive estimation; EM algorithm; kernel density estimate; least squares estimate; linear regression

**MATHEMATICS SUBJECT CLASSIFICATION**
62F35; 62J05

## 1. Introduction

In parametric linear regression analysis one often imposes the model assumptions that the errors are independent and normally distributed. The normality assumption is convenient, as it is well known that the maximum likelihood estimator (MLE) of the unknown parameter vector simplifies to the least squares estimator (LSE). Naturally, an invalid assumption on the error distribution $F$ comes at a cost; the MLE is in general neither consistent nor asymptotically efficient under model misspecification. Moreover, in practice, it can lead to inaccurate or invalid statistical inference; see Sec. 5. This has motivated the search for alternative, (semi)parametric, estimators that retain asymptotic efficiency when $F$ is unknown.

One approach is adaptive estimation, which "adapts" to an unknown, or incorrectly specified, distribution $F$ by maximizing an estimated likelihood function based on an initial estimate of the error distribution; see Bickel (1982), Linton and Xiao (2007), Yuan and De Gooijer (2007), and the references therein. The adaptive idea has been studied for (non)linear regression models using non- and semiparametric methods to estimate $F$ or its probability density function (pdf) $f$.

There are various alternative adaptive estimation methods for non- and semiparametric regression problems with errors of unknown distributional form. For instance, the

**CONTACT** Jan G. De Gooijer j.g.degooijer@uva.nl Amsterdam School of Economics, University of Amsterdam, Amsterdam, The Netherlands.

empirical likelihood method of Owen (2001) has been used to obtain adaptive confidence limits and likelihood ratio test statistics for regression parameters; see also Owen (1988, 1990, 1991), Qin and Lawless (1994), Kitamura (1997), Kitamura (2007), among many others. Another example is the multivariate adaptive regression splines (MARS) (Friedman 1991) which is a global adaptive nonparametric method for fitting nonlinear regression models. PolyMARS (Kooperberg, Bose, and Stone 1997) is an extension of MARS that allows for multiple polychotomous regression. Time series MARS, or TSMARS, can be used for nonlinear time series analysis and forecasting; see, e.g., De Gooijer (2017). More recently, Wang and Yao (2012) proposed a minimum average variance estimation method, a dimension reduction technique, which can be adaptive to different error distributions. Also, Chen, Wang, and Yao (2015) developed an adaptive estimation method for varying coefficient models.

Recently, Yao and Zhao (2013) proposed an adaptive kernel density-based estimator for classical linear regression models, called KDRE. In particular, with an estimate of the "true" parameter vector, $f$ is modeled by a kernel density-based estimator of the regression residuals. In the second step, parameter estimates are obtained by maximizing a local, kernel-based, log-likelihood function using the first-step estimated density function as the true one. Through a simulation study, Yao and Zhao (2013) show that the resulting KDRE is asymptotically equivalent to the oracle estimator in which the true error pdf is known.

Now, it is well known that each LSE-based residual is the sum of two components: one is the true error, the other is a linear function of the entire vector of errors. Since, in finite samples, the second term will tend to be normally distributed (as long as the errors have finite variance), the residuals for small samples will appear more normal than would the unobserved values of the errors. This tendency is called supernormality; see Bassett and Koenker (1982), Bloomfield (1974), and White and MacDonald (1980). Hence, for the two-step KDRE method, it is likely that the finite-sample properties of the proposed estimator strongly depend on the empirical distribution of the residuals, more closely resembling normality than would be the case by using the pdf of the errors itself, if this were in fact possible. This makes the KDRE method nonoptimal.

In this paper, we remedy this deficiency of the KDRE method by further iteration. That is, we maximize over different kernel-based likelihood functions. These different likelihood functions follow iteratively from parameter estimates that result from maximizing the (previous) likelihood function. The algorithm iterates until the parameter estimates (and, hence, the estimated likelihood function) reaches a fixed point. The new estimator is called multi-step KDRE (M-KDRE). In finite samples, one may expect that M-KDRE yields better estimation results; see Robinson (1988). Actually, from an extensive Monte Carlo study where we compare the finite-sample performance of M-KDRE with estimates based on five other, parametric and (semi)nonparametric adaptive estimators (AEs), we find that iterating over the likelihood functions can indeed strongly increase finite-sample performance. In fact, we find that M-KDRE outperforms all other considered AEs. In an empirical example based on Andrabi, Das, and Khwaja (2017), we show that LSE estimates may be misleading. M-KDRE outperforms the other considered estimators in terms of out-of-sample prediction performance. Furthermore, M-KDRE provides strong evidence that the treatment effect as described in Andrabi, Das, and Khwaja (2017) does not exist.

Theoretically, we establish strong (almost sure) convergence to the true parameter vector, under relative weak conditions. We also show that the M-KDRE method is adaptive, i.e., asymptotically normal and efficient. Furthermore, its computation is made convenient by proposing an EM type algorithm.

The rest of the paper is organized as follows. Section 2 introduces the new adaptive M-KDRE method and contains our theoretical results. An efficient EM type algorithm to implement the proposed estimator is also given in this section. In Section 3, we describe and explain five alternative adaptive estimation methods. Section 4 contains results of a simulation-based study of the finite sample properties of the M-KDRE method and compares it with those of the adaptive estimation methods discussed in Section 3. In Section 5, we present the empirical application of our method to the educational data set as used in Andrabi, Das, and Khwaja (2017). Section 6 gives a summary and some concluding remarks. Proofs are presented in Appendix A.

## 2. Multi-Step kernel Density-Based regression estimation

### 2.1. Model and method

Consider the general linear regression problem with observations

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}_0 + \varepsilon_i, \quad (i = 1, \dots, n) \tag{1}$$

where $y_i$ is a univariate response variable, $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})'$ is a $p$-dimensional ($p < n$) vector of covariates, and $\boldsymbol{\beta}_0 \in \boldsymbol{\mathcal{B}} \subseteq \mathbb{R}^p$ is an unknown parameter vector including an intercept. Here $(y_i, \mathbf{x}_i, \varepsilon_i)$ are independent and identically distributed (i.i.d.) realizations from a common random source $(y, \mathbf{x}, \boldsymbol{\varepsilon})$. Moreover, the $\varepsilon_i$'s are assumed to have some common unknown pdf $f(\varepsilon)$, and $E[\varepsilon_i|\mathbf{x}_i, \boldsymbol{\beta}_0] = 0$ and $E[|\varepsilon||\mathbf{x}_i, \boldsymbol{\beta}_0] < \infty$ $(i = 1, \dots, n)$. Model (1) is semiparametric with $\boldsymbol{\beta}$ and $f(\cdot)$ its parametric and non-parametric part, respectively.

Let $\hat{\boldsymbol{\beta}}_{\text{LSE}}$ be the LSE of $\boldsymbol{\beta}_0$ in (1), which is a natural estimator to start the M-KDRE method. Also, let $\hat{\boldsymbol{\beta}}^{(u)}$ denote an estimator of $\boldsymbol{\beta}_0$ at iteration step $u = 0, 1, \dots$. Then, with the conditions introduced above, the proposed M-KDRE can be obtained as follows.

(i) Initial step: At $u=0$, start with $\hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}_{\text{LSE}}$. Compute the residuals $\hat{\varepsilon}_i^{(0)} = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}^{(0)}$ $(i = 1, \dots, n)$.

(ii) Compute the Rosenblatt-Parzen kernel-based estimator $\hat{f}_n^{(u)}(\cdot)$ of $f(\cdot)$. That is

$$\hat{f}_n^{(u)}(x) = (nh_n)^{-1}\sum_{i=1}^{n} K\left(\frac{x - \hat{\varepsilon}_i^{(u-1)}}{h_n}\right) \tag{2}$$

where $h_n > 0$ is the bandwidth.

(iii) Let $c(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})$. Then, using (2), compute

$$\hat{\boldsymbol{\beta}}^{(u)} = \text{argmax}_{\boldsymbol{\beta}\in\boldsymbol{\mathcal{B}}}\hat{Q}_u(\boldsymbol{\beta}) \quad \text{s.t.} \quad \mathrm{c}(\boldsymbol{\beta}^{(\mathrm{u})}) = 0 \tag{3}$$

where $\hat{Q}_u(\boldsymbol{\beta})$ is the local log-likelihood function

$$\hat{Q}_u(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln \hat{f}_n^{(u)}(y_i - \mathbf{x}_i'\boldsymbol{\beta}) \tag{4}$$

(iv) Repeat steps (ii)–(iii) until convergence at iteration step $(u+1)$ $(u = 1, 2, \ldots)$.

It is worth mentioning that Yao and Zhao (2013) only iterate over the empirical likelihood function of the first initial, LSE, estimate while in the above case we allow for stochastic fluctuations in $\hat{f}_n^{(u)}(\cdot)$. As we show, this generalization will increase mean-square error parameter efficiency.

### 2.2. Asymptotic properties

In this section we give the asymptotic properties of the $u$th M-KDRE, denoted hereafter by $\hat{\boldsymbol{\beta}}^{(u)}$. For convenience, when we emphasize the dependence on $\boldsymbol{\beta}$, we use $f(y_i|\boldsymbol{\beta})$ to denote $f(y_i - \mathbf{x}_i'\boldsymbol{\beta})$. Technical details, lemmas, and proofs are given in Appendix A.

**Theorem 2.1.** *(Almost sure (a.s.) convergence) Under the assumptions of Lemma A.1 and*

(i) $f(y_i|\hat{\boldsymbol{\beta}})$ *is non-parametrically identifiable,*
(ii) $\boldsymbol{\beta} \in \boldsymbol{\mathcal{B}}$ *where* $\boldsymbol{\mathcal{B}} \subseteq \mathbb{R}^p$ *is compact,*
(iii) $f(y_i|\boldsymbol{\beta})$ *is continuous for each* $\boldsymbol{\beta} \in \boldsymbol{\mathcal{B}}$,
(iv) $E[\sup_{\boldsymbol{\beta}\in\boldsymbol{\mathcal{B}}} |\ln f(y_i|\boldsymbol{\beta})|] < \infty$,
*then*

$$\hat{\boldsymbol{\beta}}^{(u)} \overset{a.s.}{\longrightarrow} \boldsymbol{\beta}_0 \tag{5}$$

The following notation is used throughout the next part of the paper. Let $L_n(y_i|\boldsymbol{\beta}) = \ln f(y_i|\boldsymbol{\beta}) = \ln f(y_i - \mathbf{x}_i'\boldsymbol{\beta})$. Then $\mathbf{d}_\beta(\boldsymbol{\beta}) = \partial L_n(y_i|\boldsymbol{\beta})/\partial\boldsymbol{\beta} = -f'(y_i|\boldsymbol{\beta})f^{-1}(y_i|\boldsymbol{\beta})\mathbf{x}_i, \mathbf{d}_{\beta\beta}(\boldsymbol{\beta}) = \partial L_n^2(y_i|\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}' = (f(y_i|\boldsymbol{\beta})f''(y_i|\boldsymbol{\beta}) - f'2(y_i|\boldsymbol{\beta})/f^2(y_i|\boldsymbol{\beta}))\mathbf{x}_i\mathbf{x}_i'$, where $f'(u) = \partial f(u)/\partial u$, $f''(u) = \partial f'(u)/\partial u$, and $f'2(u) = f'(u)f'(u)$. Given these notations, the Fisher information matrix (FIM) for the *unconstrained* linear regression problem, evaluated at $\boldsymbol{\beta}_0$, can be defined as

$$\mathcal{I}_{\beta\beta} = E(\mathbf{d}_\beta(\boldsymbol{\beta}_0)\mathbf{d}_\beta(\boldsymbol{\beta}_0)') = -E(\mathbf{d}_{\beta\beta}(\boldsymbol{\beta}_0)) \tag{6}$$

where the second equality holds under mild regularity conditions. If $\mathcal{I}_{\beta\beta}$ is nonsingular, then $\mathcal{I}_{\beta\beta}^{-1}$ is the unconstrained Cramér-Rao bound (CRB) for the mean-square error (MSE) covariance matrix of any unbiased estimator of $\boldsymbol{\beta}_0$.

Observe that incorporation of the one-dimensional linear constraint $c(\boldsymbol{\beta}^{(u)}) = 0$ in step (iii) of the M-KDRE method leads to a $p$-dimensional parameter vector that has only $p-1$ independent components. As a consequence, the FIM is singular and the CRB may not be an informative lower bound on the MSE matrix of the resulting estimator. So the asymptotic distribution of $\hat{\boldsymbol{\beta}}^{(u)}$ degenerates. For deterministic linear parameter constraints, Stoica and Ng (1998) formulated a constrained CRB (CCRB) that explicitly incorporates the active constraint information with the original FIM, singular or nonsingular. Their general setting is for $q$ $(q < p)$ continuously differentiable constraints $g(\boldsymbol{\beta}) = 0$. Assuming $\boldsymbol{\beta}$ is regular in the active set of linear constraints, the $q \times p$ gradient matrix $\mathbf{G} = \partial g(\boldsymbol{\beta})/\partial \boldsymbol{\beta}'$ has full row rank $q$, with $\mathbf{G}$ independent of $\boldsymbol{\beta}$. Hence, there exists a matrix $\mathbf{U} \in \mathbb{R}^{p\times(p-q)}$ whose $p$-dimensional columns form an orthonormal null space of the range space of the row vectors in $\mathbf{G}$, i.e., such that

$$\mathbf{G}\mathbf{U} = 0 \quad \text{and} \quad \mathbf{U}'\mathbf{U} = \mathbf{I}_{p-q} \tag{7}$$

where $\mathbf{I}_{p-q}$ denotes the identity matrix of size $p$ – $q$. For nonlinear deterministic constraints, $\mathbf{G}$ and $\mathbf{U}$ are functions of $\boldsymbol{\beta}$; see, e.g., Moore, Sadler, and Kozick (2008).

Recently, Ren et al. (2015) extended the deterministic CCRB of Stoica and Ng (1998) to a hybrid parameter vector with both nonrandom and random parameter constraints. In the case of the M-KDRE method the constraint is not deterministic, depending on the random variables $\mathbf{x}_i$. Then the matrix $\mathbf{U}$ in (7) depends on the estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$, say $\mathbf{U}(\hat{\boldsymbol{\beta}})$. The resulting CCRB, as a special case of the hybrid CCRB of Ren et al. (2015), can be stated as follows.

**Theorem 2.2.** *Let $\hat{\boldsymbol{\beta}}$ be an unbiased estimate of $\boldsymbol{\beta}_0$ satisfying the active functional constraints $g(\boldsymbol{\beta}) = \mathbf{0}$, and let $\mathbf{U} = \mathbf{U}(\hat{\boldsymbol{\beta}})$ be defined in (7). Then, under certain regularity conditions, if $\mathbf{U}'\mathcal{I}_{\beta\beta}\mathbf{U}$ is nonsingular,*

$$E((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)') \geq \mathbf{U}(\mathbf{U}'\mathcal{I}_{\beta\beta}\mathbf{U})^{-1}\mathbf{U}'$$

*where the equality is achieved if and only if $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \mathbf{U}(\mathbf{U}'\mathcal{I}_{\beta\beta}\mathbf{U})^{-1}\mathbf{U}'\mathbf{d}_\beta(\boldsymbol{\beta})$, in the mean- square sense.*

**Remark 1.** Note that rather than requiring a nonsingular FIM $\mathcal{I}_{\beta\beta}$, the alternative condition is that $\mathbf{U}'\mathcal{I}_{\beta\beta}\mathbf{U}$ is nonsingular. Thus, the unconstrained FIM may be singular, or, equivalently, the unconstrained model unidentifiable, but the constrained model must be identifiable, at least locally. Ren et al. (2015) show that the difference between the standard CRB-based covariance matrix and the CCRB-based covariance matrix is a positive semi-definite matrix. This result is expected since the presence of parameter constraints can be considered as additional information to improve the performance of the estimator under study.

**Theorem 2.3.** (Normality and efficiency) *If model (1) holds, $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. with unknown density f(x) where f is a continuous function symmetric around zero with bounded continuous derivatives that satisfy*

(i) $\int xf(x)\mathrm{d}x = 0,$

(ii) $E\left[\left(\frac{\partial \ln f(x)}{\partial x}\right)^2 + \left|\frac{\partial^2 \ln f(x)}{\partial x^2}\right| + \left|\frac{\partial^3 \ln f(x)}{\partial x^3}\right|\right] < \infty,$

$\{\mathbf{x}_i\}_{i=1}^{\infty}$ *satisfies,*

(iii) $\exists \;\; 0 < M < \infty$ *such that* $\|\mathbf{x}\| < M,$

$K(\cdot)$ *is a symmetric and four times continuously differentiable function such that*

(iv) $\exists \; 0 < \rho < \infty$ *such that* $\mathrm{K(x)} = 0 \; \forall x : \|\mathbf{x}\| \geq \rho$

*holds, and*

(v) when $n \to \infty, nh_n^4 \to \infty$ and $nh_n^8 \to 0,$

*then* $\hat{\boldsymbol{\beta}}^{(u)}$ $(u = 1, 2, \ldots)$ *is asymptotically normal and efficient. That is, as* $n \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{(u)} - \boldsymbol{\beta}_0) \xrightarrow{d} N(0, \mathbf{U}(\mathbf{U}'\mathcal{I}_{\beta\beta}\mathbf{U})^{-1}\mathbf{U}') \tag{8}$$

**Remark 2.** All conditions are practical and easy to satisfy. Condition (ii) is used to guarantee the adaptiveness of $\hat{\boldsymbol{\beta}}^{(u)}$.

### 2.3. EM algorithm

In this section, we propose an EM type algorithm by noticing that (4) has a mixture log-likelihood form with an imposed constraint. Specifically, given an initial parameter estimate $\hat{\boldsymbol{\beta}}^{(0)}$ and the set of initial estimates $\{\hat{\varepsilon}_i^{(0)}\}_{i=1}^n$, the $(k+1)$ th iteration of the EM algorithm to maximize (4) (the $u$th likelihood function) is as follows.

E-step: Calculate the classification probabilities,

$$p_{ij,(k+1)}^{(u)} = \frac{\exp\left\{-\frac{1}{2h_n^2}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(k)}^{(u)} - \hat{\varepsilon}_j^{(u-1)})^2\right\}}{\sum_{\ell=1}^n \exp\left\{-\frac{1}{2h_n^2}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(k)}^{(u)} - \hat{\varepsilon}_\ell^{(u-1)})^2\right\}} \tag{9}$$

M-step: Update $\hat{\boldsymbol{\beta}}_{(k)}^{(u)}$ with

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{(k+1)}^{(u)} = \hat{\boldsymbol{\beta}}_{\mathrm{LSE}} &- \left(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i=1}^n \left(\mathbf{x}_i \sum_{j=1}^n p_{ij,(k+1)}^{(u)} \hat{\varepsilon}_j^{(u-1)}\right) \\ &+ \frac{1}{n}\left(\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \sum_{i=1}^n \mathbf{x}_i \left(\sum_{i=1}^n\sum_{j=1}^n p_{ij,(k+1)}^{(u)} \hat{\varepsilon}_j^{(u-1)}\right)\end{aligned} \tag{10}$$

where (10) follows from using a Gaussian kernel for density estimation. The choice of the kernel is not critical. Any symmetric kernel can be used for our method. However, the Gaussian second order kernel provides an explicit solution of the EM algorithm.

**Theorem 2.4.** *Under the linearity constraint* $c(\boldsymbol{\beta}) = 0$, *each iteration of the above E- and M-steps will monotonically increase the local log-likelihood* (4), *i.e.,* $\hat{Q}_n(\hat{\boldsymbol{\beta}}_{(k+1))}^{(u)}) \geq \hat{Q}_n(\hat{\boldsymbol{\beta}}_{(k)}^{(u)})$, *for all k.*

**Remark 3.** For the EM type algorithm, we use a full-kernel method rather than a leave-one-out method as in Yao and Zhao (2013). The approach has the following advantage. If a certain residual $\hat{\varepsilon}_j^{(u-1)}$ is extremely large, $p_{ij,(k+1)}^{(u)}$ will be close to zero for all $i \neq j$ and close to one for $i = j$. This implies that the effect of the residual is limited to the observation for which the following iteration of $\boldsymbol{\beta}$ is likely to lead to a residual that is similar in magnitude. Hence, the effect of a large residual on the maximization of (4) is small. In the leave-one-out method, the effect of the residual may be considerably larger as $p_{ij,(k+1)}^{(u)}$ is likely to have a substantial value for several observations.

**Remark 4.** The EM type algorithm is considered converged when $\max|\hat{\boldsymbol{\beta}}^{(u)}_{(k)} - \hat{\boldsymbol{\beta}}^{(u)}_{(k+1)}|$ is smaller than a threshold value, where $\max|\mathbf{A}|$ denotes the largest (absolute) element in $\mathbf{A}$. In the $u$th step of the M-KDRE method, the EM algorithm is initialized by the estimate at the $(u-1)$ th step. That is, $\hat{\boldsymbol{\beta}}^{(u)}_{(0)} = \hat{\boldsymbol{\beta}}^{(u-1)}$.

## 3. Some alternative adaptive estimation methods

### 3.1. SBS method

Stone (1975), Bickel (1982), and Schick (1993) (henceforth SBS) introduce a two-step AE. Let $\tilde{\boldsymbol{\beta}}$ be a certain $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}_0$. Then, an infeasible two-step estimator can be defined as

$$\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}} + n^{-1}\mathcal{I}^{-1}_{\beta\beta}(\tilde{\boldsymbol{\beta}})\mathbf{d}_{\beta}(\tilde{\boldsymbol{\beta}})$$

where $\mathcal{I}_{\beta\beta}(\tilde{\boldsymbol{\beta}})$ is Fisher's information matrix evaluated at $\tilde{\boldsymbol{\beta}}$ and $\mathbf{d}_{\beta}(\tilde{\boldsymbol{\beta}})$ is a corresponding $p \times 1$ score vector. The infeasibility of $\hat{\boldsymbol{\beta}}$ follows from the fact that $f$ is unknown, and hence $\mathcal{I}_{\beta\beta}$ and $\mathbf{d}_{\beta}$ are unknown. The approach of SBS is to replace $\mathbf{d}_{\beta}(\tilde{\boldsymbol{\beta}})$ by $\hat{\mathbf{d}}_{\beta}(\tilde{\boldsymbol{\beta}}) = -\sum_{i=1}^{n}\hat{f}'_n(y_i|\tilde{\boldsymbol{\beta}})\hat{f}_n^{-1}(y_i|\tilde{\boldsymbol{\beta}})\mathbf{x}_i$ where $\hat{f}_n(x)$ is defined in a similar way as in (2) and $\hat{f}'_n(x)$ is its derivative with respect to $x$. Similarly, $\mathcal{I}^{-1}_{\beta\beta}(\tilde{\boldsymbol{\beta}})$ is replaced with $n^2[\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}'_i\sum(\hat{f}'_n(y_i|\tilde{\boldsymbol{\beta}})\hat{f}_n^{-1}(y_i|\tilde{\boldsymbol{\beta}}))^2]^{-1}$.

The conditions under which the two-step AE approach can be shown to be asymptotically efficient have been researched extensively. Most importantly, the kernel estimator of the score function must be (i) i.i.d., and (ii) independent of $\mathbf{x}_i$. These conditions are restrictive and not easy to verify in practice; see, e.g., Yuan and De Gooijer (2007). Bickel (1982) solved the i.i.d. problem by splitting the sample in two; one sub-sample to estimate the score and another to solve for $\boldsymbol{\beta}$. However, Manski (1984) finds that the estimator works much better when the sample is not split, i.e., if the estimated score and $\tilde{\boldsymbol{\beta}}$ are both computed using the entire sample. If (i) and (ii) are satisfied, a sufficient condition for adaptiveness is that (iii) $E[(f'(y_i|\boldsymbol{\beta})f^{-1}(y_i|\boldsymbol{\beta})\hat{f}'_n(y_i|\tilde{\boldsymbol{\beta}})\hat{f}_n^{-1}(y_i|\tilde{\boldsymbol{\beta}}))^2] \to 0$, as $n \to \infty$.

Since $\hat{f}_n$ is present in the denominator of $\hat{\mathbf{d}}_{\beta}$, unstable estimates may follow for near-zero values of $\hat{f}_n(\cdot)$. Hence, Bickel (1982) suggests to trim the estimator of the kernel score as follows

$$\frac{\hat{f}'_n(y_i|\tilde{\boldsymbol{\beta}})}{\hat{f}_n(y_i|\tilde{\boldsymbol{\beta}})} = \begin{cases} \dfrac{\hat{f}'_n(y_i|\tilde{\boldsymbol{\beta}})}{\hat{f}_n(y_i|\tilde{\boldsymbol{\beta}})}, & \text{if } |y_i - \mathbf{x}'_i\boldsymbol{\beta}| \le t_1,\ \hat{f}_n(y_i|\tilde{\boldsymbol{\beta}}) > t_2, \text{ and } \dfrac{\hat{f}'_n(y_i|\tilde{\boldsymbol{\beta}})}{\hat{f}_n(y_i|\tilde{\boldsymbol{\beta}})} < t_3, \\ 0, & \text{otherwise.} \end{cases}$$

This trimming mechanism ensures that near-zero values do not have unreasonably large influence on the estimate. If $t_1 \to \infty$, $t_2 \to 0, t_3 \to \infty, h_n \to 0, t_1/nh_n^3 \to 0$, and $h_n t_1 \to 0$ as $n \to \infty$ then condition (iii) is satisfied. Hence, adaptiveness is established under the proper trimming parameters and conditions (i) and (ii).

Naturally, the growth rates of the trimming parameters are of little use to a practitioner, and as such the choice for the trimming parameter is a practical disadvantage. Hsieh and Manski (1987) reduce the problem to selecting a one-dimensional parameter $t$ by suggesting the following relation between the trimming parameters: $t_1 = t, t_2 = \exp(-t^2/2),\ t_3 = t$. These authors vary $t$ between 3, 4, and 8. For a sample size of 50, $t = 8$ works best in almost all cases under study.

### 3.2. LGMM and LGMMS methods

Newey (1988) describes a two-step AE that avoids kernel estimation. His approach is based on moment conditions that can be derived from certain assumptions on the error distribution. Two situations are analyzed. First, the case where the error terms are i.i.d. and independent of $\mathbf{x}_i$. This model implies the moment condition that any function of the errors are uncorrelated with any function of the regressors. Second, the case where the distribution of $\varepsilon_i$ is symmetric (S) around zero conditional on $\mathbf{x}_i$. The assumption that the errors are symmetrically distributed around zero yields the moment conditions that any odd function of the errors are uncorrelated with any function of the regressors. Hence, in both situations we can exploit moment restrictions. In particular, in the first case, we refer to the linearized general method of moment estimator as LGMM. In the second case, we use the short-hand notation LGMMS. For LGMM, natural moment conditions arise from the fact that $E[\mathbf{x}_i(\varepsilon_i^j - E(\varepsilon_i^j))] = \mathbf{0}$ for $j = 1, 2, \ldots$. However, Newey (1988) finds that these high-order "raw" moments, $m_j(\varepsilon_i) = \varepsilon_i^j$, are sensitive to a fat-tailed error distribution.

Estimates that are more robust against fat tails can be obtained using the "transformed" powers with $m_j(\varepsilon_i) = (\varepsilon_i/(1 + |\varepsilon_i|))^j$ or the "weighted" powers with $m_j(\varepsilon_i) = \exp(-\varepsilon_i/2)\varepsilon_i^j$. Similarly, for LGMMS we may use $E[\mathbf{x}_i\varepsilon_i^{2j-1}] = \mathbf{0}$ $(j = 1, 2, \ldots)$. As for LGMM, performance may be improved if we use the odd powers of the transformed method instead. However, for technical reasons the weighted powers can not be used for LGMMS; Newey (1988). In general, both for LGMM and LGMMS, we use the moment condition $E[\mathbf{x}_i(m_j(\varepsilon_i) - \mu_j)] = \mathbf{0}$ $(j = 1, 2, \ldots)$ where $\mu_j = E[m_j(\varepsilon_i)]$.

To define the LGMM(S) estimator, we introduce the following notation for some fixed value $J = J(n)$ of $j$,

$$\boldsymbol{\zeta}_i = (m_1(\varepsilon_i) - \mu_1, \ldots, m_J(\varepsilon_i) - \mu_J)', \quad \mathbf{w} = E\big[(m_{1,\varepsilon}(\varepsilon_i), \ldots, m_{J,\varepsilon}(\varepsilon_i))'\big] \quad \mathbf{V}_{\zeta\zeta} = \mathrm{Cov}(\boldsymbol{\zeta}_i) \tag{11}$$

where $m_{j,\varepsilon}(\cdot) = \partial m_j(\cdot)/\partial\varepsilon$. Let $\{\hat{\varepsilon}_i\}_{i=1}^n$ denote the residuals corresponding to the initial estimate $\hat{\boldsymbol{\beta}}$, then the quantities in (11) can be consistently estimated by their corresponding sample statistics, i.e.,

$$\hat{\boldsymbol{\zeta}}_i = (m_1(\hat{\varepsilon}_i) - \hat{\mu}_1, \ldots, m_J(\hat{\varepsilon}_i) - \hat{\mu}_J)', \quad \hat{\mathbf{w}} = \left(n^{-1}\sum_i m_{1,\varepsilon}(\hat{\varepsilon}_i), \ldots, n^{-1}\sum_i m_{J,\varepsilon}(\hat{\varepsilon}_i)\right)'$$

and

$$\hat{\mathbf{V}}_{\zeta\zeta} = n^{-1}\sum_i \hat{\zeta}_i\hat{\zeta}'_i$$

The LGMM(S) estimator is given by

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\mathrm{LGMM(S)}} = \hat{\boldsymbol{\beta}} + [(\hat{\mathbf{w}}' \otimes \mathbf{X}'\mathbf{X})(\hat{\mathbf{V}}^{-1}_{\zeta\zeta} \otimes [\mathbf{X}'\mathbf{X}]^{-1})(\hat{\mathbf{w}}\otimes\mathbf{X}'\mathbf{X})]^{-1} \\ \times (\hat{\mathbf{w}}' \otimes \mathbf{X}'\mathbf{X})(\hat{\mathbf{V}}^{-1}_{\zeta\zeta} \otimes [\mathbf{X}'\mathbf{X}]^{-1})(\mathbf{I}_J \otimes \mathbf{X}')\mathrm{vec}(\hat{\boldsymbol{\zeta}}),\end{aligned} \tag{12}$$

where $\hat{\boldsymbol{\zeta}}$ is the $n \times J$ matrix $(\hat{\zeta}'_1, ..., \hat{\zeta}'_n)'$, $\mathbf{I}_J$ is an $J \times J$ identity matrix, and $\mathbf{X}$ an $n \times p$ matrix with its first column an $n \times 1$ vector of ones. Under certain assumptions, Newey (1988) proves asymptotic normality of the LGMM and LGMMS estimators. In particular, it should hold that $J \to \infty$ and $J \ln J / \ln n \to 0$, as $n \to \infty$. Only for LGMMS, asymptotic efficiency is obtained, but not for LGMM. By means of simulation, Newey (1988) finds for LGMM that $J = 3$ performs best for sample sizes between $n = 50$ and $n = 200$. However, the MSE efficiency of the estimator as a function of $J$ flattens out as $n$ increases. Also, the transformed method is in general preferred over the weighted method.

### 3.3. KDRE method

The KDRE method of Yao and Zhao (2013) can be viewed as the unconstrained two-step version of M-KDRE. That is, it follows from unconstrained maximization of the kernel likelihood function that is estimated on the basis of the residuals corresponding to an initial estimate. Under conditions (i)–(v) of Theorem 2.3 these authors prove that the KDRE method is adaptive. For technical reasons, this property is proven for a trimmed version. The untrimmed maximizer of the kernel-based likelihood is the solution to $n^{-1}\sum_{i=1}^n \left\{\hat{f}'_n(y_i|\boldsymbol{\beta})/\hat{f}_n(y_i|\boldsymbol{\beta})\right\}\mathbf{x}'_i = 0$. The trimmed version is then defined as the solution to

$$\frac{1}{n}\sum_{i=1}^n \frac{\hat{f}'_n(y_i|\boldsymbol{\beta})}{\hat{f}_n(y_i|\boldsymbol{\beta})}\mathbf{x}'_i G_b\left(\hat{f}_n(y_i|\boldsymbol{\beta})\right) = 0.$$

Here,

$$G_b(x) = \begin{cases} 0, & \text{if } x < b, \\ \int_b^x g_b(z)\mathrm{d}z & \text{if } b \le x \le 2b, \\ 1, & \text{if } x > 2b, \end{cases} \tag{13}$$

where $g_b(\cdot)$ is a four times continuously differentiable function with support on $[b,\ 2b]$, and $b \to 0$ if $n \to \infty$. This trimming function is introduced by Linton and Xiao (2007) and they suggest the use of the beta function. For the purpose of KDRE, the trimming parameter is only used to simplify the proof and is not a part of the actual implementation of the method.

### 3.4. YDG method

Yuan and De Gooijer (2007) (hereafter YDG) propose another estimator of $\boldsymbol{\beta}_0$ based on estimating the error density by means of a kernel. The method is a one-step approach and as such does not require an initial estimate. The proposed estimator is given by

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{i=1}^{n} \ln \frac{1}{(n-1)h_n} \sum_{j\neq i}^{n} K\left(\frac{r(y_i - \mathbf{x}_i'\boldsymbol{\beta}) - r(y_j - \mathbf{x}_j'\boldsymbol{\beta})}{h_n}\right) \tag{14}$$

where $r(z) = 10 \times e^z/(1+e^z)$. The nonlinear function $r(\cdot)$ is introduced to avoid cancelation of the intercept coefficient in $\mathbf{x}_j'\boldsymbol{\beta} - \mathbf{x}_i'\boldsymbol{\beta}$. However, as Yao and Zhao (2013) note, this comes with an efficiency loss; $r(z) = z$ is efficient in the sense that even though the intercept is canceled out, the slope coefficients are adaptively estimated. They suggest the use of the following estimator

$$\hat{\boldsymbol{\beta}}^*_{\text{YDG}} = \arg\max_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{i=1}^{n} \ln \frac{1}{(n-1)h_n} \sum_{j\neq i}^{n} K\left(\frac{(y_i - \mathbf{x}_i^{*\prime}\boldsymbol{\beta}^*) - (y_j - \mathbf{x}_j^{*\prime}\boldsymbol{\beta}^*)}{h_n}\right) \tag{15}$$

where $\mathbf{x}_i^* = (1,\ \mathbf{x}_i')'$, and $\hat{\boldsymbol{\beta}}_{\text{YDG}} = (\hat{\alpha}_{\text{YDG}},\ \hat{\boldsymbol{\beta}}^{*\prime}_{\text{YDG}})'$ and $\hat{\alpha}_{\text{YDG}} = n^{-1}\sum_i (y_i - \mathbf{x}_i^{*\prime}\hat{\boldsymbol{\beta}}^*_{\text{YDG}})$. The intercept estimate $\hat{\alpha}_{\text{YDG}}$, however, is not in general asymptotically efficient.

## 4. Simulation study

### 4.1. Setup

In order to assess the finite sample practical performance of all reviewed AEs, we conduct a Monte Carlo study. We generate i.i.d. data $\{(x_i, y_i)\}_{i=1}^n$ from the regression model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad \boldsymbol{\beta} = (1,\ -1,\ 2,\ -0.5,\ 3,\ 1,\ -1,\ 2,\ -0.5,\ 3)' \tag{16}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector containing an intercept and the parameters corresponding to $p-1$ explanatory variables. Here $p=10$, but we also consider the case $p=2$ and $p=5$ consisting of the first two and first five coefficients of $\boldsymbol{\beta}$, respectively. The sample size is set at $n = 50,\ 100,\ 500,$ and $1{,}000$. All simulation results are based on 500 replications. The explanatory variables in $\mathbf{x}_i$ are independent realizations of an $N(0,\ 1)$ distribution. The errors $\varepsilon_i$ are i.i.d., and we consider the following eight error distributions:

(a) standard normal;
(b) variance-contaminated normal, the mixture $0.9N(0, 1/9) + 0.1N(0, 9)$;
(c) $t$-distribution with two degrees of freedom;
(d) bimodal symmetric mixture of two normals, $0.5N(-3, 1) + 0.5N(3, 1)$;
(e) Unif$[-\sqrt{3}; \sqrt{3}]$;
(f) Gamma(2,2);
(g) skewed mixture of normals, $0.3N(-1.4, 1) + 0.7N(0.6, 0.16)$; and
(h) log-normal, being the distribution of $\exp(z)$ for $z \sim N(0, 1)$.

The distributions are centered and scaled to have mean zero and unit variance, when necessary and possible. The $t(2)$-distribution is left unscaled as its variance is infinite.

**Table 1.** Number of times the RMSE attains its lowest value for seven regression estimators, for each *n* and *p* and across all eight error distribution functions, for the slope coefficient and, in parentheses, for the intercept.

| *p* | 2 | | | | 5 | | | | 10 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *n* | 50 | 100 | 500 | 1,000 | 50 | 100 | 500 | 1,000 | 50 | 100 | 500 | 1,000 | Total |
| M-KDRE | $5_{(3)}$ | $6_{(3)}$ | $5_{(3)}$ | $6_{(2)}$ | $4_{(1)}$ | $5_{(2)}$ | $6_{(2)}$ | $6_{(4)}$ | $4_{(3)}$ | $4_{(3)}$ | $6_{(3)}$ | $7_{(3)}$ | $64_{(32)}$ |
| KDRE | $0_{(0)}$ | $0_{(0)}$ | $3_{(1)}$ | $3_{(0)}$ | $0_{(0)}$ | $0_{(0)}$ | $2_{(1)}$ | $3_{(1)}$ | $0_{(1)}$ | $0_{(0)}$ | $1_{(0)}$ | $4_{(1)}$ | $16_{(5)}$ |
| YDG | $1_{(2)}$ | $0_{(3)}$ | $2_{(5)}$ | $2_{(4)}$ | $3_{(1)}$ | $1_{(1)}$ | $3_{(4)}$ | $3_{(5)}$ | $1_{(0)}$ | $3_{(1)}$ | $2_{(5)}$ | $3_{(5)}$ | $24_{(36)}$ |
| SBS | $0_{(0)}$ | $0_{(0)}$ | $0_{(0)}$ | $1_{(0)}$ | $0_{(0)}$ | $0_{(0)}$ | $0_{(1)}$ | $0_{(1)}$ | $0_{(0)}$ | $0_{(0)}$ | $0_{(0)}$ | $0_{(1)}$ | $1_{(3)}$ |
| LGMM | $1_{(1)}$ | $2_{(4)}$ | $4_{(3)}$ | $3_{(3)}$ | $1_{(5)}$ | $0_{(4)}$ | $2_{(1)}$ | $4_{(4)}$ | $1_{(1)}$ | $1_{(2)}$ | $1_{(3)}$ | $4_{(3)}$ | $24_{(34)}$ |
| LGMMS | $0_{(4)}$ | $1_{(4)}$ | $3_{(4)}$ | $2_{(3)}$ | $1_{(3)}$ | $1_{(3)}$ | $0_{(4)}$ | $3_{(4)}$ | $1_{(3)}$ | $1_{(3)}$ | $1_{(3)}$ | $3_{(4)}$ | $17_{(42)}$ |
| LSE | $1_{(1)}$ | $1_{(3)}$ | $1_{(3)}$ | $1_{(3)}$ | $1_{(1)}$ | $1_{(1)}$ | $1_{(2)}$ | $1_{(3)}$ | $1_{(1)}$ | $1_{(1)}$ | $1_{(2)}$ | $1_{(3)}$ | $12_{(24)}$ |
| Total | 8<br>(11) | 10<br>(17) | 18<br>(19) | 18<br>(15) | 10<br>(11) | 8<br>(11) | 14<br>(15) | 20<br>(22) | 8<br>(9) | 10<br>(10) | 12<br>(16) | 22<br>(20) | 158<br>(176) |

**Note:** For the purpose of this table, the RMSE is measured up to three decimal figures such that more than one estimator may have the lowest RMSE.

If required, we use $\hat{\boldsymbol{\beta}}_{\mathrm{LSE}}$ as an initial parameter estimate. In addition, we adopt the standard normal kernel density $K(\cdot)$. For the SBS method, we set the trimming parameter at $t = 8$. Following Newey (1988), we compute the LGMM and LGMMS estimators for the transformed moments with $J = 3$. Implementing the kernel density-based estimators requires a method for choosing the bandwidth $h_n$. There is a vast literature on this topic, ranging from simple to involved methods. But none of the proposed methods has overall performance. In an extensive simulation study of model (1) with $n = 100$ and $p = 2$, Reichardt (2017) concludes that for M-KDRE $h_{n,1} = 1.06\hat{\sigma}_n n^{-1/5}$ is preferable for symmetric error distributions in terms of root mean squared error (RMSE) of the estimators. Here $\hat{\sigma}_n$ is the standard deviation of the data. For skewed distributions, he recommends $h_{n,2} = 0.9An^{-1/5}$ where $A = \min(\hat{\sigma}_n, R/1.34)$ with $R$ the inter-quartile range of the data. The KDRE and YDG estimators perform best under $h_{n,1}$. The SBS estimator generally shows the smallest RMSE for $h_{n,2}$. Hence, throughout the simulations, we use the above bandwidths.

### 4.2. Results

Averaged over all replications, Table 1 provides summary information on the computed RMSEs of the slope and intercept coefficients for all $n$ and $p$, and across all error distributions. Clearly, M-KDRE shows the best overall performance of all estimators for the slope coefficient, with 64 lowest RMSE values out of a total of 84, i.e., 7 estimation methods (M-KDRE, KDRE, YDG, SBS, LGMM, LGMMS, LSE), 3 values for $p$, and 4 values for $n$. On the other hand, the SBS estimator has only one lowest RMSE value for $n = 1{,}000$ and $p = 2$. The other estimators have low RMSE values lying in between the above two values, with the LSE results as a benchmark. Finally, from the last column of Table 1, we see that M-KDRE performs equally well with YDG, LGGM and LGMMS in estimating the intercept term, and M-KDRE markedly outperforms KDRE.

Reichardt (2017) reports RMSE values for each of the eight error distributions. For the sake of space we omit details. However, for the slope coefficient the simulation results can be summarized as follows.

**Table 2.** Number of times the bias of seven regression estimators attains it lowest value for each $n$ and $p$, and across all eight error distribution functions for the slope coefficient and, in parentheses, for the intercept.

| $p$ | 2 | | | | 5 | | | | 10 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 50 | 100 | 500 | 1,000 | 50 | 100 | 500 | 1,000 | 50 | 100 | 500 | 1,000 | Total |
| M-KDRE | $0_{(4)}$ | $3_{(2)}$ | $7_{(5)}$ | $8_{(6)}$ | $4_{(1)}$ | $6_{(6)}$ | $5_{(6)}$ | $8_{(3)}$ | $5_{(1)}$ | $7_{(3)}$ | $7_{(5)}$ | $6_{(5)}$ | $66_{(47)}$ |
| KDRE | $0_{(1)}$ | $1_{(2)}$ | $7_{(3)}$ | $7_{(3)}$ | $3_{(0)}$ | $4_{(3)}$ | $6_{(3)}$ | $8_{(3)}$ | $0_{(0)}$ | $4_{(1)}$ | $6_{(2)}$ | $7_{(2)}$ | $53_{(23)}$ |
| YDG | $3_{(2)}$ | $3_{(4)}$ | $4_{(3)}$ | $7_{(6)}$ | $0_{(2)}$ | $2_{(4)}$ | $5_{(5)}$ | $5_{(4)}$ | $2_{(5)}$ | $2_{(1)}$ | $6_{(8)}$ | $5_{(6)}$ | $44_{(50)}$ |
| SBS | $3_{(2)}$ | $1_{(1)}$ | $5_{(3)}$ | $4_{(3)}$ | $3_{(0)}$ | $1_{(2)}$ | $7_{(5)}$ | $7_{(3)}$ | $3_{(1)}$ | $2_{(0)}$ | $6_{(3)}$ | $4_{(1)}$ | $46_{(24)}$ |
| LGMM | $2_{(4)}$ | $3_{(2)}$ | $5_{(4)}$ | $5_{(6)}$ | $1_{(1)}$ | $4_{(2)}$ | $6_{(5)}$ | $8_{(3)}$ | $3_{(1)}$ | $3_{(2)}$ | $5_{(6)}$ | $4_{(4)}$ | $49_{(40)}$ |
| LGMMS | $3_{(3)}$ | $3_{(3)}$ | $5_{(4)}$ | $6_{(5)}$ | $2_{(4)}$ | $3_{(1)}$ | $5_{(4)}$ | $6_{(4)}$ | $1_{(2)}$ | $2_{(4)}$ | $6_{(1)}$ | $5_{(4)}$ | $47_{(39)}$ |
| LSE | $2_{(5)}$ | $1_{(2)}$ | $2_{(5)}$ | $4_{(6)}$ | $0_{(3)}$ | $0_{(5)}$ | $2_{(4)}$ | $2_{(3)}$ | $1_{(1)}$ | $1_{(3)}$ | $2_{(5)}$ | $1_{(2)}$ | $18_{(45)}$ |
| Total | 13<br>(21) | 15<br>(16) | 35<br>(27) | 41<br>(35) | 13<br>(11) | 20<br>(23) | 36<br>(32) | 44<br>(23) | 15<br>(11) | 21<br>(14) | 38<br>(30) | 32<br>(25) | 323<br>(268) |

- In terms of RMSE, the M-KDRE method performs very well for the log-normal error distribution (h). That is, the RMSE of the second most efficient estimators (KDRE and LGMM) is approximately 40% larger, even for $n = 1,000$. Under error distributions (b) and (c), M-KDRE is also most efficient, but here the efficiency is gained mostly for $n = 50$ and $n = 100$. Furthermore, M-KDRE has a superior performance in small samples of the $t(2)$ error distribution. It is also close to best for error distributions (d)–(g).
- The YDG method performs well for error distributions (d)–(g), but fails quite dramatically for error distributions with fat tails, i.e. (b), (c), and (h).
- Efficiency of the SBS estimator is in general low relative to alternatives, but performance is especially weak under error distributions (c) and (d).
- Overall, LGMM is a reasonable estimator, but its efficiency is lost under error distributions (e) and (f). This efficiency loss persists even for $n = 1,000$.
- LGMMS is by construction inefficient when the error distribution is skewed: (f)–(h). More interestingly, the LGMMS-estimate of the slope coefficient is no improvement over LGMM under symmetric error distributions. The inefficiency of LGMMS with respect to LGMM is likely to be due to the fact that LGMMS uses moment restrictions on odd powers of the disturbances only and, hence, for a particular value of $J$, uses higher order moments that may lead to less efficient estimation.

Table 2 shows summary results for the bias for both slope and intercept estimators for all $n$ and $p$, and across all error distributions. For the slope coefficient, M-KDRE has the best performance in terms of the lowest bias values. Again, from Reichardt (2017) we learn that the intercept bias of the different estimators is usually of similar magnitude in the symmetric cases. Under the asymmetric error distributions, the bias of the intercept is much larger for KDRE, SBS and LGMMS than for the other estimators.

## 5. Empirical application

Andrabi, Das, and Khwaja (2017) study the impact of providing information in the form of school report cards on educational outcomes such as school fees, test scores,

**Table 3.** Effect of report cards on school fees, test scores, and enrollment as given by parameter estimates of $\beta_j$ $(j = 1,\ 2,\ 3)$ using seven estimation methods. For columns 2–7, standard errors (in parentheses) are based on 500 bootstrap replicates.

| | LSE | M-KDRE | KDRE | YDG | SBS | LGMM | LGMMS |
|---|---|---|---|---|---|---|---|
| $\hat{\beta}_1$ | −187.0 | −4.418 | −85.55 | −154.4 | −47.74 | −172.3 | −182.9 |
| | (65.91) | (38.76) | (63.42) | (59.90) | (73.34) | (64.54) | (67.65) |
| $\hat{\beta}_2$ | 0.114 | 0.084 | 0.092 | 0.094 | 0.088 | 0.088 | 0.099 |
| | (0.046) | (0.050) | (0.048) | (0.060) | (0.054) | (0.045) | (0.050) |
| $\hat{\beta}_3$ | 0.032 | 0.028 | 0.030 | 0.018 | 0.030 | 0.029 | 0.027 |
| | (0.014) | (0.014) | (0.012) | (0.019) | (0.013) | (0.012) | (0.012) |

and enrollment in markets with multiple public and private providers. The report cards given to both households and schools in $n$ randomly sampled villages across three districts, in the Punjab province of Pakistan, include information on the performance of the child, the average score of different schools in the village, and the average village score in mathematics, English, and Urdu. The following three linear regression models are of interest:

$$Y_{i,j} = \alpha_j + \beta_j \mathrm{RC}_i + \gamma_j \mathrm{Y}^*_{i,j} + \delta_j \mathrm{X}_{i,j} + \varepsilon_i, \quad (i = 1, \ldots, n; j = 1,\ 2,\ 3) \tag{17}$$

where $Y_{i,1}, Y_{i,2}$, and $Y_{i,3}$ are average fees, test scores, and enrollment rate in the post-intervention year of village $i$, respectively. $Y^*_{i,j}$ denotes the baseline measurement of the same variables. $\mathrm{RC}_i$ is the treatment dummy assignment to village $i$, which makes $\beta_j$ the variable of interest, an estimate of the impact of the report card assignment. $X_{i,j}$ is a vector of village-level baseline controls. All models in the paper are estimated using LSE.

Table 3, column 1, shows the LSEs of $\beta_j$ $(j = 1,\ 2,\ 3)$ and their corresponding standard errors (in parentheses) as shown in, respectively, Table 3(1) panel C, Table 3(4) panel C, and Table 4(1) panel C of Andrabi, Das, and Khwaja (2017). The Shapiro-Wilk test for normal data indicates that the LSE residuals are far from normally distributed, with $p$-values 0.000 $(j = 1, n = 104)$, 0.002 $(j = 2, n = 112)$, and 0.000 $(j = 3, n = 112)$. Indeed, in all cases, diagnostic statistics show that the residuals have fatter tails than one would expect based on normality. Based on the LSE results, Andrabi, Das, and Khwaja (2017) report the following main findings. First, private schools decreased their annualized fees $(Y_{i,1})$ by an average of 187 rupees, about 17% of their baseline fees, in response to the report card intervention. Second, test scores $(Y_{i,2})$ increased by 0.11 standard deviation. Third, primary enrollment $(Y_{i,3})$ increased by 3.2 percent points or 4.5% in treatment villages.

Table 3, columns 2–7, shows estimates of the six AEs for models $j = 1,\ 2$, and 3. We see that these estimators pull the estimated treatment effect toward zero for all models. The results for model 1 are especially striking. The M-KDRE of $\beta$ is more than 40 times smaller than the LSE, in absolute value. Also, for model 1, the AEs differ substantially. In that respect, it is interesting to investigate the prediction performance of the respective methods.

Table 4 shows the ratio of the median absolute prediction error (MAPE) of an estimator relative to the LSE. The training set is a random sample of the data of size $\lceil 0.8n \rceil$. We see that M-RKDRE has the lowest MAPE for model 1. In addition, observe that the prediction performance is generally better for AEs with a low estimate of $\beta_1$

**Table 4.** Median absolute prediction error (MAPE) of six AEs relative to LSE.

| Model ($j$) | M-KDRE | KDRE | YDG | SBS | LGMM | LGMMS |
|---|---|---|---|---|---|---|
| 1 | 0.720 | 0.858 | 0.906 | 0.769 | 0.992 | 1.001 |
| 2 | 1.002 | 1.016 | 0.998 | 1.005 | 0.997 | 1.020 |
| 3 | 0.972 | 0.975 | 1.029 | 0.963 | 0.965 | 0.971 |

**Table 5.** Bootstrapped 95% confidence intervals of the effect of report cards for the M-KDRE method.

| Model ($j$) | Normal | Percentile |
|---|---|---|
| 1 | [−80.38, 71.55] | [−88.11, 59.25] |
| 2 | [−0.015, 0.183] | [−0.011, 0.187] |
| 3 | [0.001, 0.055] | [0.002, 0.055] |

such as KDRE and SBS. This suggests that the effect of the report cards on school fees, if it exists at all, is much lower than reported. For models 2 and 3, there is less difference between the estimates of the AEs. Also, the estimates are adjusted less strongly with respect to LSE.

Table 5 reports two bootstrapped 95% confidence intervals for $\beta_j$ $(j = 1,\ 2,\ 3)$ as estimated by M-KDRE. The confidence interval termed "Normal" is based on an asymptotic normality assumption, and the column called "Percentile" is based on the 95% inter quartile range obtained from the empirical distribution of 500 bootstrap replicates. Both intervals show that the estimated treatment effect for model 1 is not significantly different from zero.

In summary, the above results demonstrate the practical relevance of the AEs in general and that of the proposed M-KDRE method in particular. None of the other AEs adjusted the treatment effect on school fees as far toward zero as M-KDRE, while prediction performance suggests that this method should be preferred over other methods for this particular linear regression model and sample size. Thus, there is no support for the first finding of Andrabi, Das, and Khwaja (2017) at any reasonable significance level. Further, Table 3 shows that the AEs find that the effect of report cards on test scores is not significantly different from zero at the 5% nominal level. Lastly, the effect of report cards on the enrollment rate seems, even though marginally significant for M-KDRE, also questionable.

## 6. Summary and concluding remarks

In this paper, we proposed an adaptive multi-step kernel density-based regression estimator for linear regression models. We have established the theoretical properties of our estimation method, including asymptotic normality and almost sure convergence. In an extensive simulation study, we have shown that the finite sample performance of M-KDRE is second to none of five alternative AEs. For several error distributions, it is up to twice as efficient in terms of RMSE than the second best estimator. Further, for any other error distribution, it is either the most efficient or very close to the most efficient estimator. All other AEs show a loss of efficiency for certain specific error distributions. Our empirical application provides a good illustration of many of these issues. In particular, using the M-KDRE method and its corresponding bootstrap standard errors, we found fairly compelling evidence that the treatment effects that Andrabi, Das, and Khwaja (2017) find are not significantly different from zero.

The results raise several questions for further research. For instance, one may wish to estimate nonlinear regressions via the M-KDRE method. In that case the EM algorithm, at least in its present form, needs to be adjusted. Another issue concerns the fact that the multi-step method makes use of $\hat{\beta}_{\text{LSE}}$ in the initial step. This choice was primarily based on computational convenience. Perhaps, efficiency may be further enhanced by a more prudent choice of the initial estimator. It may also be of interest to assess the robustness of M-KDRE to a violation of the independence assumption. In particular, adaptive estimation is not in general possible when the vector of covariates and the error process are not mutually independent. We leave these questions for future research.

## Acknowledgments

The authors would like to thank two anonymous referees for their valuable comments and suggestions.

## References

Andrabi, T., J. Das, and A. I. Khwaja. 2017. Report cards: The impact of providing school and child test scores on educational markets. *American Economic Review* 107 (6):1535–63. doi:10.1257/aer.20140774.

Bassett, G. W., and R. W. Koenker. 1982. An empirical quantile function for linear models with iid errors. *Journal of the American Statistical Association* 77:407–15. doi:10.2307/2287261.

Bickel, P. J. 1982. On adaptive estimation. *The Annals of Statistics* 10 (3):647–71. doi:10.1214/aos/1176345863.

Bloomfield, P. 1974. *On the distribution of the residuals from a fitted linear model*. Technical report, Department of Statistics, Princeton University, Princeton, NJ, Series 2; 56.

Chai, G., Z. Li, and H. Tian. 1991. Consistent nonparametric estimation of error distributions in linear model. *Acta Mathematicae Applicatae Sinica* 7 (3):245–56. doi:10.1007/BF02005973.

Chen, Y., Q. Wang, and W. Yao. 2015. Adaptive estimation for varying coefficient models. *Journal of Multivariate Analysis* 137:17–31. doi:10.1016/j.jmva.2015.01.017.

Crowder, M. 1984. On constrained maximum likelihood estimation with non-i.i.d. observations. *Annals of the Institute of Statistical Mathematics* 36 (2):239–49. doi:10.1007/BF02481968.

De Gooijer, J. G. 2017. *Elements of nonlinear time series analysis and forecasting*. New York: Springer-Verlag.

Friedman, J. H. 1991. Multivariate adaptive regression splines. *The Annals of Statistics* 19 (1): 1–141. with discussion).

Hshieh, D. A., and C. F. Manski. 1987. Monte Carlo evidence on adaptive maximum likelihood estimation of a regression. *The Annals of Statistics* 15:541–51. doi:10.1214/aos/1176350359.

Kitamura, Y. 1997. Empirical likelihood methods with weakly dependent processes. *The Annals of Statistics* 25 (5):2084–102. doi:10.1214/aos/1069362388.

Kitamura, Y. 2007. Empirical likelihood methods in econometrics: Theory and practice. In *Advances in economics and econometrics: Theory and applications, Ninth World Congress*, Econometric Society Monographs, ed. R. Blundell, W. Newey, and T. Persson, pp. 174–237. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511607547.008.

Kooperberg, C., S. Bose, and C. J. Stone. 1997. Polychotomous regression. *Journal of the American Statistical Association* 92 (437):117–27. doi:10.1080/01621459.1997.10473608.

Linton, O., and Z. Xiao. 2007. A nonparametric regression estimator that adapts to error distribution of unknown form. *Econometric Theory* 23 (3):371–413. doi:10.1017/S026646660707017X.

Manski, C. F. 1984. Adaptive estimation of non-linear regression models. *Econometric Reviews* 3 (2):145–94. doi:10.1080/07474938408800060.

Moore, T. J., B. M. Sadler, and R. J. Kozick. 2008. Maximum-likelihood estimation, the Cramér-Rao bound, and the method of scoring with parameter constraints. *IEEE Transactions on Signal Processing* 56 (3):895–908. doi:10.1109/TSP.2007.907814.

Newey, W. K. 1988. Adaptive estimation of regression models via moment restrictions. *Journal of Econometrics* 38 (3):301–39. doi:10.1016/0304-4076(88)90048-6.

Newey, W. K., and D. McFadden. 1994. Large sample estimation and hypothesis testing. In *Handbook of econometrics*, ed. K. J. Arrow and D. Intriligator, Vol. 4, 2111–245. New York: Elsevier.

Osborne, M. R. 2000. Scoring with constraints. *The Anziam Journal* 42 (1):9–25. doi:10.1017/S1446181100011561.

Owen, A. B. 1988. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* 75 (2):237–49. doi:10.1093/biomet/75.2.237.

Owen, A. B. 1990. Empirical likelihood confidence regions. *The Annals of Statistics* 18 (1): 90–120. doi:10.1214/aos/1176347494.

Owen, A. B. 1991. Empirical likelihood for linear models. *The Annals of Statistics* 19 (4):1725–47. doi:10.1214/aos/1176348368.

Owen, A. B. 2001. *Empirical likelihood*. Boca Raton, FL: Chapman & Hall/CRC.

Qin, J., and J. Lawless. 1994. Empirical likelihood and general estimating equations. *The Annals of Statistics* 22 (1):300–25. doi:10.1214/aos/1176325370.

Reichardt, H. 2017. Adaptive estimation in linear regression using repeated kernel error density estimation. MSc. thesis. Econometrics, Erasmus University Rotterdam. https://thesis.eur.nl/pub/38903/.

Ren, C., J. Le Kernec, J. Galy, E. Chaumette, P. Larzabal, and A. Renaux. 2015. A constrained hybrid Cramér-Rao bound for parameter estimation. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (*ICASSP), Brisbane, Australia, 3472–6.

Robinson, P. M. 1988. Root-*n*-consistent semiparametric regression. *Econometrica* 56 (4):931–54. doi:10.2307/1912705.

Schick, A. 1993. On efficient estimation in regression models. *The Annals of Statistics* 21 (3): 1486–521. doi:10.1214/aos/1176349269.

Stoica, P., and B. C. Ng. 1998. On the Cramér-Rao bound under parametric constraints. *IEEE Signal Processing Letters* 5:177–9. doi:10.1109/97.700921.

Stone, C. J. 1975. Adaptive maximum likelihood estimators of a location parameter. *The Annals of Statistics* 3 (2):267–84. doi:10.1214/aos/1176343056.

Wade, W. 1974. The bounded convergence theorem. *The American Mathematical Monthly* 81 (4):387–9. doi:10.2307/2319009.

Wang, Q., and W. Yao. 2012. An adaptive estimation of MAVE. *Journal of Multivariate Analysis* 104 (1):88–100. doi:10.1016/j.jmva.2011.07.001.

White, H., and G. M. MacDonald. 1980. Some large-sample tests for nonnormality in the linear regression model. *Journal of the American Statistical Association* 75 (369):16–28. doi:10.1080/01621459.1980.10477415.

Yao, W., and Z. Zhao. 2013. Kernel density-based linear regression estimate. *Communications in Statistics - Theory and Methods* 42 (24):4499–512. doi:10.1080/03610926.2011.650269.

Yuan, A., and J. G. De Gooijer. 2007. Semiparametric regression with kernel error model. *Scandinavian Journal of Statistics* 34:841–69. doi:10.1111/j.1467-9469.2006.00531.x.

Zhang, W. Y. 1990. On the congruent kernel estimate of error distributions in linear model (in Chinese). *Journal of Sichuan University* (Natural Science Edition) 27:132–44.

## Appendix A: Proofs of results

**Lemma A.1.** (Zhang 1990, Theorem 5) *If model (1) holds, $\{\varepsilon_i\}_{i=1}^{n}$ are i.i.d. with unknown density f(x) where $f(\cdot)$ is a uniformly continuous function that satisfies (i) $\int xf(x)\mathrm{d}x = 0$, (ii) $0 < \int x^2 f(x)\mathrm{d}x < \infty$, the set of covariates $\{\mathbf{x}_i\}_{i=1}^{\infty}$ satisfies (iii) $\exists\ 0 < M < \infty$ such that $\|\mathbf{x}_i\| < M$ $\forall i = 1, ..., n$, (iv) $\mathbf{S}_n \to \mathbf{Q} = E(\mathbf{x}\mathbf{x}')$ where $\mathbf{S}_n = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'$, and the following assumptions on the kernel function $K(\cdot)$ hold: (v) K(x) is uniformly bounded and $\exists\ 0 < \rho < \infty$ such that $K(x) = \mathbf{0}\ \forall x : \|\mathbf{x}\| \geq \rho$ (vi) K(x) is Riemann integrable on $[-\rho,\ \rho]$ (vii) when $n \to \infty, 0 < h_n \to 0$ and $\sqrt{n}h_n/\ln n \to \infty$, then*

$$\sup_{x\in\mathbb{R}} \|\hat{f}_{n,\mathrm{LSE}}(x) - f(x)\| \xrightarrow{a.s.} 0 \tag{A.1}$$

*where $\hat{f}_{n,\mathrm{LSE}}$ is the kernel density-based estimator of the* LSE *residuals $\hat{\varepsilon}_{i,\mathrm{LSE}} = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\mathrm{LSE}}$ $(i = 1, ..., n)$.*

**Lemma A.2.** *Suppose that the assumptions of Lemma A.1 hold. Then, if any estimator $\boldsymbol{\beta}^*$ satisfies $Pr(\lim_{n\to\infty} \max|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0| \leq \max|\mathbf{S}_n^{-1}|\ \|\ln\max|\mathbf{S}_n^{-1}|\ \|) = 1$ where $\max|\mathbf{A}| = \max_{i,j}|a_{ij}|$ with $a_{ij}$ the elements of a matrix $\mathbf{A}$,*

$$\sup_{x\in\mathbb{R}} \|\hat{f}_n^*(x) - f(x)\| \xrightarrow{a.s.} 0 \tag{A.2}$$

*where $\hat{f}_n^*$ is the kernel density-based estimator of the residuals $\hat{\varepsilon}_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}^*$ $(i = 1, ..., n)$.*

*Proof.* This result follows immediately from Theorem 5 and Lemma 4 in Zhang (1990) in conjunction with Theorem 4 and (29) in Chai, Li, and Tian (1991). □

**Lemma A.3.** *If there is a function $Q_0(\boldsymbol{\beta})$ such that (i) $Q_0(\boldsymbol{\beta})$ is uniquely maximized at $\boldsymbol{\beta}_0$, (ii) $\mathcal{B}$ is compact, (iii) $Q_0(\boldsymbol{\beta})$ is continuous, (iv) $\sup_{\boldsymbol{\beta}\in\mathcal{B}} \|\hat{Q}_n(\boldsymbol{\beta}) - Q_0(\boldsymbol{\beta})\| \xrightarrow{a.s.} 0$, then for $u = 1, 2, ...,$ $\hat{\boldsymbol{\beta}}^{(u)} \xrightarrow{a.s.} \boldsymbol{\beta}_0$, where $\hat{\boldsymbol{\beta}}$ maximizes the objective function $\hat{Q}_n(\boldsymbol{\beta})$ subject to $\boldsymbol{\beta} \in \mathcal{B}$. The weak convergence result, i.e., $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ can be obtained by replacing condition (iv) by $\sup_{\boldsymbol{\beta}\in\mathcal{B}} \|\hat{Q}_n(\boldsymbol{\beta}) - Q_0(\boldsymbol{\beta})\| \xrightarrow{p} 0$.*

*Proof:* The proof is similar to the proof of Theorem 2.1 of Newey and McFadden (1994). □

**Lemma A.4.** *If $f_n : \mathcal{B} \to \mathbb{R}$ is a continuous function, $\mathcal{B}$ is compact, and $f_n \xrightarrow{a.s.} f$, then*

$$\lim_{n\to\infty} \int_{\mathcal{B}} f_n \mathrm{d}u = \int_{\mathcal{B}} f \mathrm{d}u \tag{A.3}$$

*Proof.* Since $\mathcal{B}$ is compact and $f_n$ is continuous, the image $f_n(\mathcal{B})$ is a compact subset of $\mathbb{R}$ and hence, closed and bounded. Then, the result follows from the bounded convergence theorem; see, e.g., Wade (1974). □

*Proof of Theorem 2.1.* Following (Newey and McFadden 1994, Thm. 2.5), we verify the conditions in Lemma A.3. Note that conditions (i)–(iii) are on the density $f(\cdot)$ of $\boldsymbol{\varepsilon}$, and on the parameter space $\mathcal{B}$. These conditions hold under the usual regularity conditions of MLE. Condition (iv) of Lemma A.3 implies that we have to prove that

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}\ln\hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})-E\big[\ln f(y_i|\boldsymbol{\beta})\big]\right\|\overset{a.s.}{\to}0.$$

Since $\hat{f}_n^{(1)}=\hat{f}_{n,\mathrm{LSE}}$, we have by Lemma A.1

$$\sup_{x\in\mathbb{R}}\|\hat{f}_n^{(1)}(x)-f(x)\|\overset{a.s.}{\to}0.$$

Now note that

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\|\hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})-f(y_i|\boldsymbol{\beta})\|\leq\sup_{\boldsymbol{\beta}\in\mathbb{R}^p}\|\hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})-f(y_i|\boldsymbol{\beta})\|\leq\sup_{x\in\mathbb{R}}\ \|\hat{f}_n^{(1)}(x)-f(x)\|$$

implying that

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\ \|\hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})-f(y_i|\boldsymbol{\beta})\|\overset{a.s.}{\to}0. \tag{A.4}$$

Condition (iii) implies that $\inf_{\boldsymbol{\beta}\in\mathcal{B}}f(y_i|\boldsymbol{\beta})>0$. Thus, $\exists\varepsilon>0$ such that $\inf_{\boldsymbol{\beta}\in\mathcal{B}}f(y_i|\boldsymbol{\beta})>\varepsilon$. Also, by (A.4) for any $\varepsilon>0$,

$$\Pr(\lim_{n\to\infty}\sup_{\boldsymbol{\beta}\in\mathcal{B}}\|\hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})-f(y_i|\boldsymbol{\beta})\|<\varepsilon)=1\Rightarrow\Pr(\lim_{n\to\infty}\inf_{\boldsymbol{\beta}\in\mathcal{B}}\hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})>0)=1.$$

This, together with condition (ii), ensures that for $n$ large enough both $\ln f(y_i|\boldsymbol{\beta})$ and $\ln\hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})$ are uniformly continuous with probability one such that by the uniform continuous mapping theorem

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\|\ln\hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})-\ln f(y_i|\boldsymbol{\beta})\|\overset{a.s.}{\to}0.$$

Note that by conditions (ii) and (iii), $\ln\hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})$ is bounded and we may invoke the uniform law of large numbers such that,

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}\ln\hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})-E\Big[\ln\hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})\Big]\right\|\overset{a.s.}{\to}0. \tag{A.5}$$

Also, by Lemma A.4,

$$\lim_{n\to\infty}E\Big[\ln\hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})\Big]=E\big[\ln f(y_i|\boldsymbol{\beta})\big] \tag{A.6}$$

Now define the following variables

$$A\equiv\left\|\frac{1}{n}\sum_{i=1}^{n}\ln\hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})-E\big[\ln f(y_i|\boldsymbol{\beta})\big]\right\|,\quad A_1\equiv\left\|\frac{1}{n}\sum_{i=1}^{n}\ln\hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})-E\Big[\ln\hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})\Big]\right\|,$$
$$A_2\equiv\left\|E\Big[\ln\hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})\Big]-E\big[\ln f(y_i|\boldsymbol{\beta})\big]\right\|.$$

Then, by the triangle inequality, we have $A\leq A_1+A_2$, and by (A.5) and (A.6), $\sup_{\boldsymbol{\beta}\in\mathcal{B}}A_1\overset{a.s.}{\to}0$ and $\lim_{n\to\infty}\sup_{\boldsymbol{\beta}\in\mathcal{B}}A_2=0$. From condition (iv) of Lemma A.3 it follows that

$$\sup_{\boldsymbol{\beta}\in\mathcal{B}}\left\|\frac{1}{n}\sum_{i=1}^{n}\ln\hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})-E\big[\ln f(y_i|\boldsymbol{\beta})\big]\right\|\overset{a.s.}{\to}0. \tag{A.7}$$

Thus, by Lemma A.3,

$$\hat{\boldsymbol{\beta}}^{(1)} \overset{a.s.}{\rightarrow} \boldsymbol{\beta}_0 \tag{A.8}$$

For the sake of completeness, we prove also that the constraint $c(\boldsymbol{\beta}) = 0$ does not affect this result. Let $\mathcal{C} \subseteq \mathcal{B}$ be the subset for which $c(\boldsymbol{\beta}) = 0$. That is, $\mathcal{C} = \left\{\boldsymbol{\beta} \in \mathcal{B} : c(\boldsymbol{\beta}) = 0\right\}$. First, note that $\mathcal{C}$ is the level set of the continuous function $c(\boldsymbol{\beta})$ such that $\mathcal{C}$ is closed. Also, $\mathcal{C}$ is bounded since $\mathcal{C} \subseteq \mathcal{B}$ and $\mathcal{B}$ is bounded. Hence, $\mathcal{C}$ is compact such that $\hat{\boldsymbol{\beta}}^{(1)} = \arg\sup_{\boldsymbol{\beta}\in\mathcal{C}} \sum_{i=1}^{n} \ln \hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})$. Denote $\tilde{\boldsymbol{\beta}} = \arg\sup_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{i=1}^{n} \ln \hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})$ as the global maximizer of the objective function over $\mathcal{B}$. (Newey and McFadden 1994, p. 2122) show that for (A.8) to hold, it suffices to prove that

$$\frac{1}{n}\sum_{i=1}^{n} \ln \hat{f}_n^{(1)}\left(y_i|\hat{\boldsymbol{\beta}}^{(1)}\right) - \frac{1}{n}\sum_{i=1}^{n} \ln \hat{f}_n^{(1)}(y_i|\tilde{\boldsymbol{\beta}}) \overset{a.s.}{\rightarrow} 0. \tag{A.9}$$

For that purpose, define

$$\begin{aligned}
B &\equiv \left\|\sup_{\boldsymbol{\beta}\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} \ln \hat{f}_n^{(1)}(y_i|\boldsymbol{\beta}) - \sup_{\boldsymbol{\beta}\in\mathcal{B}} \frac{1}{n}\sum_{i=1}^{n} \ln \hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})\right\|,\\
B_1 &\equiv \left\|\sup_{\boldsymbol{\beta}\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n} \ln \hat{f}_n^{(1)}(y_i|\boldsymbol{\beta}) - \sup_{\boldsymbol{\beta}\in\mathcal{C}} E\left[\ln f(y_i|\boldsymbol{\beta})\right]\right\|,\\
B_2 &\equiv \left\|\sup_{\boldsymbol{\beta}\in\mathcal{C}} E\left[\ln f(y_i|\boldsymbol{\beta})\right] - \sup_{\boldsymbol{\beta}\in\mathcal{B}} E\left[\ln f(y_i|\boldsymbol{\beta})\right]\right\|, B_3 \equiv \left\|\sup_{\boldsymbol{\beta}\in\mathcal{B}} E\left[\ln f(y_i|\boldsymbol{\beta})\right] - \sup_{\boldsymbol{\beta}\in\mathcal{B}} \frac{1}{n}\sum_{i=1}^{n} \ln \hat{f}_n^{(1)}(y_i|\boldsymbol{\beta})\right\|.
\end{aligned}$$

Again by the triangle inequality, $B \leq B_1 + B_2 + B_3$. From (A.7), it is easy to show that $B_1 \overset{a.s.}{\rightarrow} 0$ and $B_3 \overset{a.s.}{\rightarrow} 0$. To show that $B_2 \overset{a.s.}{\rightarrow} 0$, first observe that by conditions (i) and (ii) of Lemma A.1 and the strong law of large numbers,

$$\Pr\left(\lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta}_0)\right] = 0\right) = 1. \tag{A.10}$$

This implies

$$\begin{aligned}
\Pr(\lim_{n\to\infty} \boldsymbol{\beta}_0 \in \mathcal{C}) = 1 &\Rightarrow \Pr\left(\lim_{n\to\infty}\left[\arg\sup_{\boldsymbol{\beta}\in\mathcal{B}} E[\ln f(y_i|\boldsymbol{\beta})]\right] \in \mathcal{C}\right) = 1\\
&\Rightarrow \Pr(\lim_{n\to\infty} B_2 = 0) = 1,
\end{aligned}$$

and the last result implies, by definition of almost sure convergence, that $B_2 \overset{a.s.}{\rightarrow} 0$. Hence, $B \overset{a.s.}{\rightarrow} 0$ and the constraint does not affect the result.

Lastly, to show that the algorithm asymptotically converges to $\boldsymbol{\beta}_0$, remark that (A.8) implies by Lemma A.2 that $\sup_{x\in\mathbb{R}} \|\hat{f}_n^{(2)}(x) - f(x)\| \overset{a.s.}{\rightarrow} 0$ where $\hat{f}_n^{(2)}(x)$ is the kernel density-based estimator of the residuals corresponding to $\hat{\boldsymbol{\beta}}^{(1)}$. Thus, by identical reasoning, we obtain $\hat{\boldsymbol{\beta}}^{(u)} \overset{a.s.}{\rightarrow} \boldsymbol{\beta}_0$ for $u = 1, 2, \ldots$. □

**Remark 5.** Conditions (i)–(iv) of Theorem 2.1 are the regularity conditions that are necessary for the convergence of MLE under the true density. Thus, the only additional conditions imposed are those in Lemma A.1 of which condition (i) of zero mean goes without loss of generality in the context of linear regression as we can always adjust the intercept parameter in $\boldsymbol{\beta}$ if the center of $f(\cdot)$ is not zero. Condition (ii) of Lemma A.1 may be restrictive in some cases as it rules out, for instance, the $t(\nu)$-distribution with $1 < \nu \leq 2$. However, in Section 4.2 we observed that M-KDRE performs well for $t(2)$. In fact, its performance is best of all considered estimators under that error distribution. Hence, the practical use of M-KDRE does not seem to be restricted to distributions with finite variance. Conditions (iii) and (iv) of Lemma A.1 are easy to verify in practice, and conditions (v)–(vii) are technical requirements on the kernel and bandwidth. Note that (v) is not satisfied by the Gaussian kernel since that kernel does not have bounded support. In practice, however, the Gaussian kernel entails a significant computational advantage.

*Proof of Theorem 2.3* (sketch): For the case of $q$ $(q<p)$ with linear random or deterministic equality constraints, the proof of consistency and asymptotic distribution of $\hat{\boldsymbol{\beta}}$ can be based on results in Crowder (1984) and Osborne (2000). In particular, given these results it follows that $\hat{\boldsymbol{\beta}}^{(1)}$ is asymptotically normal and efficient. The only condition on the initial estimator $\hat{\boldsymbol{\beta}}^{(0)}$ is that $(\hat{\boldsymbol{\beta}}^{(0)}-\hat{\boldsymbol{\beta}}_0)=O_p(n^{-1/2})$. For $\hat{\boldsymbol{\beta}}^{(1)}$ this follows from the proof of (Yao and Zhao 2013, Thm. 2.1). Hence, all subsequent estimates $\hat{\boldsymbol{\beta}}^{(u)}$ $(u=2,3,\dots)$ also satisfy (8). □

*Proof of Theorem 2.4.* Under the Gaussian kernel, the linear constraint $c(\boldsymbol{\beta})=0$, and a full-kernel method, the M-step in (10) becomes

$$\hat{\boldsymbol{\beta}}^{(u)}_{(k+1)}=\arg\min_{\boldsymbol{\beta}}\sum_{i=1}^{n}\sum_{j=1}^{n}(p^{(u)}_{ij,(k+1)}(y_i-\mathbf{x}_i'\boldsymbol{\beta}-\hat{\varepsilon}_j^{(u-1)})^2)\ \text{s.t.}\ \sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y}_\mathrm{i}-\mathbf{x}_\mathrm{i}'\boldsymbol{\beta})=0 \tag{A.11}$$

This can be solved by Lagrangian optimization. Define the Lagrangian $\mathcal{L}$ as

$$\mathcal{L}(\boldsymbol{\beta},\lambda)=\sum_{i=1}^{n}\sum_{j=1}^{n}(p^{(u)}_{ij,(k+1)}(y_i-\mathbf{x}_i'\boldsymbol{\beta}-\hat{\varepsilon}_j^{(u-1)})^2)-\lambda\sum_{i=1}^{n}(y_i-\mathbf{x}_i'\boldsymbol{\beta}) \tag{A.12}$$

with first-order conditions

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\beta}}=-2\sum_{i=1}^{n}\sum_{j=1}^{n}(p^{(u)}_{ij,(k+1)}\mathbf{x}_i(y_i-\mathbf{x}_i'\boldsymbol{\beta}-\hat{\varepsilon}_j^{(u-1)}))-\lambda\sum_{i=1}^{n}\mathbf{x}_i=0 \tag{A.13}$$

$$\frac{\partial\mathcal{L}}{\partial\lambda}=\sum_{i=1}^{n}(y_i-\mathbf{x}_i'\boldsymbol{\beta})=0. \tag{A.14}$$

By setting $x_{i,1}\equiv 1$ $(i=1,\dots,n)$, the first element of the first-order condition in (A.13) implies

$$\begin{aligned}\lambda&=-\frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(p^{(u)}_{ij,(k+1)}(y_i-\mathbf{x}_i'\boldsymbol{\beta}-\hat{\varepsilon}_j^{(u-1)})\right)\\&=-\frac{2}{n}\sum_{i=1}^{n}(y_i-\mathbf{x}_i'\boldsymbol{\beta})\sum_{j=1}^{n}p^{(u)}_{ij,(k+1)}+\frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}p^{(m)}_{ij,(k+1)}\hat{\varepsilon}_j^{(u-1)}\\&=\frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}p^{(u)}_{ij,(k+1)}\hat{\varepsilon}_j^{(u-1)},\end{aligned} \tag{A.15}$$

where the last equality follows from (A.14). Then, by plugging $\lambda$ in (A.13), rearranging terms and using $\sum_{j=1}^{n}p^{(u)}_{ij,(k+1)}=1$, we obtain

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(u)}_{(k+1)}&=\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{n}\mathbf{x}_iy_i-\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{n}\left(\mathbf{x}_i\sum_{j=1}^{n}p^{(u)}_{ij,(k+1)}\hat{\varepsilon}_j^{(u-1)}\right)\\&\quad+\frac{1}{n}\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}\sum_{i=1}^{n}\mathbf{x}_i\left(\sum_{i=1}^{n}\sum_{j=1}^{n}p^{(u)}_{ij,(k+1)}\hat{\varepsilon}_j^{(u-1)}\right).\end{aligned} \tag{A.16}$$

Recognize that the first term is equal to $\hat{\boldsymbol{\beta}}_{\mathrm{LSE}}$. Then, the fact that (9) and (10) are the E- and M-step, respectively, of an EM type algorithm follows trivially from the proof of Theorem 2.2 in Yao and Zhao (2013). □