DALARNA UNIVERSITY

Working papers in transport, tourism, information technology and microdata analysis

A Finite Sample Improvement of the Fixed Effects Estimator – Applied to Technical Inefficiency

Author: Daniel Wikström
Editor: Hasan Fleyeh

Nr: 2013: 13

Working papers in transport, tourism, information technology and microdata analysis
ISSN: 1650-5581

Noname manuscript No.
(will be inserted by the editor)

A Finite Sample Improvement of the Fixed Effects Estimator – Applied to Technical Inefficiency

Daniel Wikström

Received: date / Accepted: date

Abstract The FE ('fixed effects') estimator of technical inefficiency performs poorly when N ('number of firms') is large and T ('number of time observations') is small. We propose estimators of both the firm effects and the inefficiencies, which have small sample gains compared to the traditional FE estimator. The estimators are based on nonparametric kernel regression of unordered variables, which includes the FE estimator as a special case. In terms of global conditional MSE ('mean square error') criterions, it is proved that there are kernel estimators which are efficient to the FE estimators of firm effects and inefficiencies, in finite samples. Monte Carlo simulations supports our theoretical findings and in an empirical example it is shown how the traditional FE estimator and the proposed kernel FE estimator lead to very different conclusions about inefficiency of Indonesian rice farmers.

Keywords Technical output inefficiency · Nonparametric kernel estimation · Panel data (*JEL Classification:* C13, C14, C23)

1 Introduction

Schmidt and Sickles (1984) proposed the FE estimator of technical inefficiency. The benefit of the FE estimator compared to frequently applied maximum likelihood estimators is two-fold: no distributional assumption of the inefficiencies and no random-effects assumption. The latter means that the selected input quantities of a specific firm may depend on the firm's lack of being technical efficiency, which to us appears very plausible. We think the random effects

Daniel Wikström
Swedish University of Agricultural Sciences, Department of Economics, Box 7013, 750 07 Uppsala, Sweden
Tel.: +46-18-671725
Fax: +46-18-673502
E-mail: daniel.wikstrom@slu.se

assumption is overly restrictive and regard it foremost as a 'statistical' assumption in contrary to 'economic'. That is, the random effects assumption is imposed on the economical model to make the likelihood function handleable, i.e. to simplify statistical estimation.

However, the FE estimator has its flaws too, of course. Consistency of the inefficiencies relies both on large number of firms, N and on a large number of time observations, T, where the latter condition usually is poorly satisfied in empirical applications (see Park and Simar 1994, on details of consistency).[1] The ML estimators, on the other hand are consistent as T grows given the random effects assumption and the assumed marginal distribution of the inefficiencies are correct.

Furthermore it is well-known that estimates of technical inefficiency based on the FE estimator can be seriously upward biased due to random error (Kim et al 2007; Wang and Schmidt 2009; Satchachai and Schmidt 2010), mainly in the context of comparison to the best firm measure. Much attention has been paid to reducing this bias by using bootstrap and jackknife estimators. Although jackknife estimators seem to effectively reduce the bias this is at the expense of larger variance and MSE. However, the bootstrap approach appears more promising since some results have shown reductions in both bias and MSE.

Additionally, the FE estimator is an unbiased estimator of the firm effects but works poorly in MSE terms if data is affected by random error.[2] In this case the variance of the estimator is large, which is transferred to the inefficiencies. Bias reduction of the FE estimator inherits this problem, since this only reduces the bias of the maximum firm effect, i.e. only shifts the distribution of the estimated inefficiencies.

Thus, the FE estimator of inefficiency works poorly in situations when the data is influenced by random error, in which case both the bias and the variance are relatively large. The bias can be handled quite well by bias-reduction methods but not the variance.

The purpose of this paper is to investigate the merits of kernel estimation of firm effects and of the inefficiencies. We prove that kernel estimation of firm effects and of the compared-to-the-best-firm-in-the-sample inefficiency measure are efficient compared to traditional FE estimation in terms of global conditional MSE-criterions.

In Monte Carlo simulations the theoretical results are supported. Kernel estimation outperforms the traditional FE estimator in most cases, in particular when the random error is influential.

We also show that kernel estimation has similar asymptotic properties as the FE estimator of the firm effects. The kernel estimator like the FE estimator is consistent as $T \to \infty$ and in fact asymptotically $\sqrt{T}-$equivalent to the FE

[1] If one consider the measure of inefficiency compared to the 'best' firm in the sample, consistency only relies on a large T. And this is the measure we primally will consider in this text.

[2] Throughout this paper, the individual effects of the fixed effects model are referred to as 'firm effects'.

estimator. This implies that the kernel estimator has the same asymptotic distribution as the FE estimator, i.e. normally distributed as T goes to infinity (and N is either bounded or unbounded).

The bias-reduction methods for the FE estimator are based on presumed convergence rates for the bias of the maximum firm effect estimator as $T \to \infty$, which in turn are based on the consistency and asymptotic normality of the estimator of the firm effects. Because the asymptotics of the kernel estimator and the FE estimator are so similar we conjecture the bias-reduction methods for the latter estimator also are appropriate for the former estimator. This also works very well in the simulations if there is not too much random error, or analogously, not too much of a 'finite sample situation'. If there is a lot of random error influencing the estimation, bias reduction does not work very well for any of the estimators and we recommend using kernel estimation without bias-reduction in this case, which is shown to be superior both in terms of bias and ASE ('average square error').

For estimating cell-probabilities of categorical data, it has been shown theoretically that gains can be made by kernel estimation, compared to maximum likelihood, in the sense of reducing cell-global MSE (see Hall 1981; Brown and Rundell 1985, and references therein). In the present study, we show that estimating firm effects by unordered kernel regression leads to similar gains in efficiency when estimating firm effects and inefficiency. For cross-section data, Ouyang, Li, and Racine (2009) presented asymptotical properties as well as promising Monte Carlo simulations that compare kernel regression using discrete regressors with the cell-average estimator.[3] Li, Racine, and Wooldridge (2009) partly use these findings to construct a nonparametric average treatment effect estimator.

A branch of the literature on panel data modeling also related, directly or indirectly, to stochastic frontier modeling is Park and Simar (1994) and Park, Sickles, and Simar (1998, 2003, 2007). In these papers the aim is to estimate the parameters of the parametric (frontier) function as efficiently as possible. Park and Simar (1994) recognized that the random and fixed effects models are actually semiparametric models and derive an efficient semiparametric estimator of the slope coefficients of the random effects model. In the following studies, Park et al (1998, 2003, 2007), results for several other panel data estimators are derived in the same fashion. Nevertheless this is a different focus compared with ours; we instead focus on estimating the firm effects efficiently. Usually the main interest in panel data modeling is the slope parameters, however, for estimation of technical efficiency the firm effect are at least as important. We especially focus on large N and small T cases, for which the estimation error in the slope coefficients is small (while it is large in the firm effects) and thus has little influence on the firm effects and the inefficiencies.

[3] Ouyang et al (2009) call it a 'cell-frequency' estimator, however, we find it more appropriate to call it a 'cell-average' estimator. The sample is split according to one or more grouping variables and the averages of a continuous dependent variable are computed for each group or 'cell'.

Henderson and Simar (2005) is the only study we know of where a similar categorical kernel, as we use, is employed to estimate technical inefficiencies with a fixed effects type of estimator. However, in this case Henderson and Simar (2005) model the production frontier and time-varying inefficiencies in a joint unspecified function. Thus, fully nonparametric estimation. Our estimator is instead an extension of the traditional fixed effects estimator, where the production function is parametrically specified. The problem with a fully unspecified production function is of course the efficiency loss, when there are several continuous input variables. The so called 'curse of dimensionality'.[4] Henderson and Simar (2005) also propose a slightly more efficient semiparametric estimator of time-constant inefficiency, with additivity between the production function and the firm effects. However, they do not apply kernel estimation of the firm effects, in this case, and consequently do not make the small sample gains analyzed in this study.

The structure of the paper is as follows: Section 2 presents the linear panel data model and estimation approaches. Section 3 compares the small-sample properties of the kernel estimator of the firm effects versus the FE estimator. Section 4 includes large-sample properties of the kernel estimator. In Section 5 the estimation of technical inefficiency is discussed. Section 6 contains the Monte Carlo simulations for the firm effects, while Section 7 is the counterpart for the inefficiencies. Section 8 includes an empirical example and Section 9 concludes the paper.

2 The stochastic frontier model and estimation

In this study, we consider a standard linear panel data model:

$$y_{it} = x'_{it}\beta + \alpha_i + \nu_{it}; \quad i = 1, \ldots, N, t = 1, \ldots, T, \tag{1}$$

where y_{it} is the dependent variable, x_{it} is a $K \times 1$ covariate vector, β is a $K \times 1$ coefficient vector and α_i is the firm effect of firm i. The error term, ν_{it}, is for simplicity assumed to be independent and identically distributed with finite variance, and $E(\nu_{it}|\alpha_i, X_i) = 0$, where $X_i = [x_{i1}\ x_{i2} \cdots x_{iT}]'$, $i = 1, 2, \ldots, N$. We also assume that the cross-section of firms is independently drawn from the population. Thus, the firm effects, $\alpha_i,\ i = 1, \ldots, N$, are independent.

The traditional fixed effects estimator of the firm effect of firm i is given as follows:

$$\hat{\alpha}_i = \frac{1}{T}\sum_{t}^{T}(y_{it} - x'_{it}\hat{\beta}) = \bar{y}_i - \bar{x}'_i\hat{\beta}, \tag{2}$$

on the other hand, the kernel estimator of the firm effect of firm i is written as:

$$\tilde{\alpha}_i = \frac{\sum_j^N \sum_t^T \left(y_{jt} - x'_{jt}\hat{\beta}\right) L_j(i, \lambda)}{T \sum_j^N L_j(i, \lambda)} \tag{3}$$

[4] The problem with the parametric approach is of course potential misspecification.

where $L_j(\cdot)$ is a kernel function defined as:

$$L_j(i,\lambda) = \begin{cases} 1, & j = i \\ \lambda \in [0,1], & \text{otherwise.} \end{cases} \tag{4}$$

Note that when the bandwidth $\lambda = 0$, the two estimators coincide, $\tilde{\alpha}_i = \hat{\alpha}_i$.[5] For the other extreme, $\lambda = 1$, the estimator collapses to $\tilde{\alpha} = \bar{\bar{y}} - \bar{\bar{x}}\hat{\beta}$, (where $\bar{\bar{z}} = \sum_{i=1}^{N} \bar{z}_i/N$), which is a pooled estimator of the intercept. Thus, the kernel estimator ranges from the sample splitting (the FE estimator) to pooled estimation of a common intercept. The within estimator of β is used for both estimators (e.g. Wooldridge 2010).

3 Small sample properties

This section describes the theoretical basis for the kernel estimator of α_i compared with the FE estimator. The small sample properties are treated in detail, since T usually is small in real-life applications. However, in Section 4 asymptotic properties are also considered.

The FE estimator of the firm effects is unbiased, with conditional variance:

$$V\left(\hat{\alpha}_i \middle| X\right) = \frac{\sigma_\nu^2}{T} + \bar{x}_i' V(\hat{\beta}|X)\bar{x}_i, \tag{5}$$

keeping the firm effects 'fixed', α_i, $i = 1, 2, \ldots, N$ and conditioning on $X = [X_1 X_2 \ldots X_N]'$.[6] The second term in the variance is due to the estimation error in $\hat{\beta}$, where $\bar{x}_i = \frac{1}{T}\sum_t x_{it}$ and $V(\hat{\beta}|X)$ is the conditional covariance matrix of $\hat{\beta}$.

Unlike the traditional FE estimator, the kernel estimator is biased:

$$Bias\left(\tilde{\alpha}_i \middle| X\right) = \frac{N\lambda(\bar{\alpha} - \alpha_i)}{1 + (N-1)\lambda}, \tag{6}$$

where $\bar{\alpha} = \sum_{i=1}^{N} \alpha_i/N$.[7]

The conditional variance of the kernel estimator takes the following form:

$$Var\left(\tilde{\alpha}_i \middle| X\right) = \frac{\frac{\sigma_\nu^2}{T} + \bar{x}_i' V(\hat{\beta}|X)\bar{x}_i}{[1 + (N-1)\lambda]^2} + \tag{7}$$
$$\frac{2(N-1)\bar{x}_i' V(\hat{\beta}|X)\bar{\bar{x}}_{-i}\lambda + (N-1)\frac{\sigma_\nu^2}{T}\lambda^2 + (N-1)^2\bar{\bar{x}}_{-i}' V(\hat{\beta}|X)\bar{\bar{x}}_{-i}\lambda^2}{[1 + (N-1)\lambda]^2},$$

[5] Thus, the kernel estimator is a generalization of the traditional FE estimator. There is actually a continuum of FE estimators, since $\lambda \in [0,1]$.

[6] The 'fixed effects' estimator is a misleading name which likely can be traced back in time to when the firm effects actually were considered as fixed (see the introduction to Ch. 10 Wooldridge 2010, for a discussion about this issue.). In this study we actually are conditioning on the firm effects, however, we follow the notational convention and do not write this out explicitly. However, we explicitly condition on X. Without this conditioning, assumptions about the relation between X and the firm effects will be necessary.

[7] The bias is the same conditional and unconditional on X

where $\bar{\bar{x}}_{-i} = \frac{1}{N-1}\sum_{j\neq i}\bar{x}_j$. Consequently, we can express the conditional MSE as:

$$MSE\left(\tilde{\alpha}_i\middle|X\right) = \left\{\frac{N\lambda\left(\bar{\alpha}-\alpha_i\right)}{1+(N-1)\lambda}\right\}^2 + \frac{\frac{\sigma_\nu^2}{T}+\bar{x}_i'V(\hat{\beta}|X)\bar{x}_i+2(N-1)\bar{x}_i'V(\hat{\beta}|X)\bar{\bar{x}}_{-i}\lambda}{\left[1+(N-1)\lambda\right]^2}+ \frac{(N-1)\frac{\sigma_\nu^2}{T}\lambda^2+(N-1)^2\bar{\bar{x}}_{-i}'V(\hat{\beta}|X)\bar{\bar{x}}_{-i}\lambda^2}{\left[1+(N-1)\lambda\right]^2}. \tag{8}$$

Note that when $\lambda = 0$ the conditional MSE collapses to the conditional MSE of the FE estimator, i.e. the conditional variance of the FE estimator. The bias is independent of T, but the variance decreases when T increases. Therefore, when T increases the bias gets more important for the MSE, and an optimal λ should decrease, which implies that the kernel estimator approaches the FE estimator. Thus, the kernel estimator has a bias-variance tradeoff.

It should be possible to show that the kernel estimator dominates the FE estimator in terms of the conditional MSE criterion in (8) for sufficiently large N and to minimize the criterion with respect to λ_i for all $i = 1, 2, \ldots, N$. However, it will be difficult to estimate these N bandwidths, especially in cases when T is small. Instead we follow the example of Hall (1981) and Brown and Rundell (1985) among others, in the case of cell-probability estimation, and consider the following global conditional MSE criterion:

$$\sum_i^N MSE\left(\tilde{\alpha}_i\middle|X\right) = \frac{N^2\lambda^2\sum_{i=1}^N\left(\bar{\alpha}-\alpha_i\right)^2}{\left[1+(N-1)\lambda\right]^2}+ \frac{\frac{N\sigma_\nu^2}{T}+\sum_{i=1}^N\bar{x}_i'V(\hat{\beta}|X)\bar{x}_i+2(N-1)\sum_{i=1}^N\bar{x}_i'V(\hat{\beta}|X)\bar{\bar{x}}_{-i}\lambda}{\left[1+(N-1)\lambda\right]^2}+ \frac{N(N-1)\frac{\sigma_\nu^2}{T}\lambda^2+(N-1)^2\sum_{i=1}^N\bar{\bar{x}}_{-i}'V(\hat{\beta}|X)\bar{\bar{x}}_{-i}\lambda^2}{\left[1+(N-1)\lambda\right]^2}. \tag{9}$$

This global conditional MSE criterion is straightforward to minimize with respect to λ.

The following theorem is akin to the MSE efficiency theorem for the kernel estimator of multinomial probabilities, given by Brown and Rundell (1985), as well as the centerpiece theorem for ridge regression in Hoerl and Kennard (1970).

Theorem 1 [8] *If N, T and $s_\alpha^2 = \sum_i^N (\bar{\alpha} - \alpha_i)^2 / (N-1)$ are finite and σ_ν^2 is finite and nonzero*[9]*, then there exists an interval $(0, \varepsilon]$, such that*

$$\begin{aligned} &\sum_i^N MSE\left(\tilde{\alpha}_i \middle| X; \lambda \in (0, \varepsilon]\right) < \\ &\sum_i^N MSE\left(\tilde{\alpha}_i \middle| X; \lambda = 0\right) = \sum_i^N MSE(\hat{\alpha}_i | X). \end{aligned} \tag{10}$$

Proof The first derivative of the squared bias with respect to λ is well-defined and nonnegative for all $\lambda \in [0,1]$, and only equals zero as $\lambda = 0$. The first derivative of the variance is also well-defined, but negative when $\lambda = 0$.[10] Therefore, we conclude that (37) holds for some arbitrary small $\varepsilon > 0$ Q.E.D.

By minimizing the conditional MSE criterion in (9) we get the following optimal bandwidth:[11]

$$\lambda^* = \frac{N\frac{\sigma_\nu^2}{T} - \sum_i^N \bar{x}_i' V(\hat{\beta}|X)(\bar{\bar{x}}_{-i} - \bar{x}_i)}{N\frac{\sigma_\nu^2}{T} + N^2 s_\alpha^2 + (N-1)\sum_i^N (\bar{\bar{x}}_{-i} - \bar{x}_i)' V(\hat{\beta}|X)\bar{\bar{x}}_{-i}}. \tag{11}$$

This bandwidth has a very interesting implication for the kernel estimator when there is no heterogeneity in the population, such that $\alpha_1 = \alpha_2 = \cdots = \alpha_N = \alpha$. This implies $s_\alpha^2 = 0$ and as given by Lemma 2 and Lemma 3 in Appendix I

$$-\sum_i^N \bar{x}_i' V(\hat{\beta}|X)(\bar{\bar{x}}_{-i} - \bar{x}_i) = (N-1)\sum_i^N (\bar{\bar{x}}_{-i} - \bar{x}_i)' V(\hat{\beta}|X)\bar{\bar{x}}_{-i} \geq 0, \tag{12}$$

[8] It is straightforward to adapt the theorem for global MSE efficiency unconditional of X as well by replacing

$$\bar{x}_i' V(\hat{\beta}|X)\bar{x}_i, \quad \bar{x}_i' V(\hat{\beta}|X)\bar{\bar{x}}_{-i} \quad \text{and } \bar{\bar{x}}_i' V(\hat{\beta}|X)\bar{\bar{x}}_{-i}$$

with

$$E\left[\bar{x}_i' V(\hat{\beta}|X)\bar{x}_i\right], \quad E\left[\bar{x}_i' V(\hat{\beta}|X)\bar{\bar{x}}_{-i}\right] \quad \text{and } E\left[\bar{\bar{x}}_i' V(\hat{\beta}|X)\bar{\bar{x}}_{-i}\right].$$

However, the expectations are conditional on α_i, $i = 1, \ldots, N$, which make it difficutl to find a feasible estimator of an optimal bandwidth without imposing stron assumptions on the relationship between the covariates and the firm effects.

[9] These assumptions are made so that the first order derivatives of the squared conditional bias and the conditional variance are well-defined and to assure strict MSE efficiency. Standard regularity assumptions are implicitly made for the FE covariance matrix, $V(\hat{\beta}|X)$ (e.g. Wooldridge 2010).

[10] The derivatives are provided in Appendix I. By 'well-defined' we mean that the derivatives exists for all $\lambda \in [0,1]$, which implies that the global conditional MSE function is continuous for all $\lambda \in [0,1]$.

[11] The derivations behind the bandwidth are presented in the Appendix I.

which gives $\lambda^* = 1$ and in turn that the firms are pooled. Thus, appropriately the kernel estimator produces a common intercept if there is no heterogeneity. However, to find an estimator of λ^* which tends to the upper bound as $N \to \infty$ and/or $T \to \infty$ seems like an impossible task, since this would require an estimator $\hat{s}_\alpha^2$ for which

$$NT(\hat{s}_\alpha^2 - s_\alpha^2) = O_p(1)$$

to avoid the estimator of λ^* from tending to zero.

Given (12) it is easy to see that $\lambda^* > 0$, if $\sigma_\nu^2 > 0$, in finite samples, i.e. $\lambda^* \in (0, 1]$.

4 Large sample properties

In this section large sample properties of the kernel estimator of the firm effects are derived. Consistency and asymptotic $\sqrt{T}-$equivalence of $\tilde{\alpha}_i$ to $\hat{\alpha}_i$ are shown, where the latter implies that the kernel-FE estimator has the same asymptotic distribution as the traditional FE estimator.[12]

We start out by stating results and assumptions needed for deriving consistency and asymptotic $\sqrt{T}-$equivalence. Because the random error is assumed to be i.i.d. the following is true, given $E(\nu_{it})$ and σ_ν^2 exist:

$$\bar{\bar{\nu}}_{-i} = O_p(N^{-1/2}T^{-1/2}) \tag{13}$$

$$\bar{\nu}_i = O_p(T^{-1/2}) \tag{14}$$

$$\tag{15}$$

Furthermore given standard least squares assumptions for the within estimator $\hat{\beta}$:

$$(\beta - \hat{\beta}) = O_p(N^{-1/2}T^{-1/2}) \tag{16}$$

and finally the followings condition of the bandwidth is sufficient:

$$\lambda = O_p(N^{-1}T^{-1}), \tag{17}$$

then

$$\tilde{\alpha}_i = \frac{\alpha_i + \bar{\nu}_i + \bar{x}_i'(\beta - \hat{\beta}) + (N-1)\lambda\left[\bar{\alpha}_{-i} + \bar{\bar{\nu}}_{-i} + \bar{\bar{x}}_{-i}'(\beta - \hat{\beta})\right]}{1 + (N-1)\lambda} = \tag{18}$$

$$\alpha_i + O_p(T^{-1/2}).$$

Thus, $\tilde{\alpha}_i$ is a consistent estimator of α_i as $T \to \infty$ while N is either bounded or unbounded. To show asymptotic equivalence to $\sqrt{T}\hat{\alpha}_i$, consider

[12] Under regularity conditions $\sqrt{T}(\hat{\alpha}_i - \alpha_i)$ is asymptotically normal if $T \to \infty$ (Hall et al 1995) as well as if $T \to \infty$ and $N \to \infty$ (Park et al 1998). In the latter case the variance is smaller since the influence of the estimation error given by $\sqrt{T}(\beta - \hat{\beta})$ vanishes as $N \to \infty$.

$$\sqrt{T}\tilde{\alpha}_i = \frac{\sqrt{T}\left\{\alpha_i + \bar{\nu}_i + \bar{x}_i'(\beta - \hat{\beta}) + (N-1)\lambda\left[\bar{\alpha}_{-i} + \bar{\bar{\nu}}_{-i} + \bar{\bar{x}}_{-i}'(\beta - \hat{\beta})\right]\right\}}{1 + (N-1)\lambda} =$$

$$\frac{\sqrt{T}\left\{\hat{\alpha}_i + (N-1)\lambda\left[\bar{\alpha}_{-i} + \bar{\bar{\nu}}_{-i} + \bar{\bar{x}}_{-i}'(\beta - \hat{\beta})\right]\right\}}{1 + (N-1)\lambda} =$$

$$\sqrt{T}(\hat{\alpha}_i) + O_p(T^{-1/2}), \tag{19}$$

Hence, given $\sqrt{T}(\hat{\alpha}_i - \alpha) \xrightarrow{d} N(0, V)$, the asymptotic equivalence implies $\sqrt{T}(\tilde{\alpha}_i - \alpha) \xrightarrow{d} N(0, V)$.

Bias-reduction methods are derived based on the asymptotic normality of the FE estimator of the maximum firm effect as well as the consistency as T tends to infinity. We will not derive these properties explicitly for the kernel estimator but since the asymptotic behavior is so similar between the estimators we conjecture the bias-reduction methods for the FE estimator also are applicable for the kernel estimator.

The asymptotic results put some requirements on the bandwidth. For consistency and asymptotic normality $\lambda = O_p(N^{-1}T^{-1})$ is sufficient. If s_α^2 is nonzero and bounded[13], it is straightforward to show that $\lambda^* = O_p(N^{-1}T^{-1})$.

An estimator of λ^* with the same rates of convergence requires estimators of all unknowns which are bounded in probability as $T \to \infty$ (and $N \to \infty$). There are such estimators (see Appendix III) and consequently there are feasible estimators with the same asymptotic properties as if $\tilde{\alpha}_i$ was based on λ^*.[14]

5 Estimation of technical inefficiency

To estimate the technical inefficiencies, we follow Schmidt and Sickles (1984) and use the following estimators:

$$\hat{\alpha} = \max_j \hat{\alpha}_j, \quad \hat{u}_i = \hat{\alpha} - \hat{\alpha}_i, \ i = 1, \ldots, N, \tag{20}$$

$$\tilde{\alpha} = \max_j \tilde{\alpha}_j, \quad \tilde{u}_i = \tilde{\alpha} - \tilde{\alpha}_i, \ i = 1, \ldots, N, \tag{21}$$

It is well-known that the FE estimator of u_i and of $u_i^* = \alpha_{(N)} - \alpha_i$ is associated with an upward bias, if not N is small and/or $\frac{\sigma_\nu^2}{T}$ is small compared to σ_u^2 (e.g. Wang and Schmidt 2009; Satchachai and Schmidt 2010).[15] It is

[13] If s_α^2 is unbounded the kernel estimator equals the FE estimator and if $s_\alpha^2 = 0$ the kernel estimator is pooling firms and is consistent and asymptotic normal as $N \to \infty$.

[14] A naive version is $\hat{\lambda} = \frac{1}{NT}$.

[15] Using conventional notation for order statistics, $\alpha_{(1)} \leq \alpha_{(2)} \leq \cdots \leq \alpha_{(N)}$.

quite informative to look on the variance of the FE estimator unconditional of α_i, which is equal to:

$$\frac{\sigma_\nu^2}{T} + \bar{x}_i' V(\hat{\beta}|X)\bar{x}_i + \sigma_u^2. \tag{22}$$

The population variance of α_i is $\sigma_u^2 = \sigma_\alpha^2$ and hence, the variance of the FE estimator unconditional of the firm effects is larger than the population variance if not $T \to \infty$. If N is large the second term is small, and the variance in (22) should be close to σ_u^2. However, if $\frac{\sigma_\nu^2}{T}$ is large compared to σ_u^2 this is not the case, and we can expect the FE estimator to overestimate the maximum firm effect. Furthermore, the larger N is the higher is the probability of $\hat{\alpha} > \hat{\alpha}_{(N)}$. Thus, if N is large and $\frac{\sigma_\nu^2}{T}$ is large compared to σ_u^2 the probability of $\hat{\alpha} > \hat{\alpha}_{(N)}$ is large.

Formally as shown by Satchachai and Schmidt (2010) the unbiasness of the firm effects implies an upwards bias in $\hat{\alpha}$:

$$E(\hat{\alpha}) \geq \alpha_{(N)},$$

since $\hat{\alpha} \geq \hat{\alpha}_{(N)}$.

For the kernel estimator, we cannot really predict whether the estimates will be upward biased, downward biased or devoid of bias. However, by investigating the bias of $\tilde{\alpha}$, given in (6), it is reasonable to expect that the bias of $\tilde{\alpha}_{(N)}$ is smaller than the bias of $\hat{\alpha}_{(N)}$. The kernel estimator is a 'shrinkage' estimator, i.e. it is biased towards the mean. The estimates of the large realizations of α_i ($> \bar{\alpha}$) are negatively biased and, on average, it is reasonable to expect that $E_{uc}(\tilde{\alpha}) < E_{uc}(\hat{\alpha})$.[16] Therefore, the kernel estimator of $u_i^* = \alpha_{(N)} - \alpha_i$ will be less upward biased compared to the FE estimator, which can be shown as follows:

$$E_{uc}(\hat{u}_i - u_i^*) = E_{uc}(\hat{\alpha} - \hat{\alpha}_i) - E(\alpha_{(N)}) + \mu_\alpha = E_{uc}(\hat{\alpha}) - E(\alpha_{(N)}) \tag{23}$$

$$E_{uc}(\tilde{u}_i - u_i^*) = E_{uc}(\tilde{\alpha}) - E(\alpha_{(N)}), \tag{24}$$

and if $E_{uc}(\tilde{\alpha}) < E_{uc}(\hat{\alpha}) \Rightarrow E_{uc}(\tilde{u}_i - u_i^*) < E_{uc}(\hat{u}_i - u_i^*)$.[17]

However, it might not always be a good property that the kernel estimator is less upward biased than the FE estimator of u_i^*. As discussed above and by Wang and Schmidt (2009), the FE estimator does not need to be very upward biased when N is small. In this case the kernel estimator could be downward biased.

We will now state a lemma which is both positive and a bit negative. It is positive since it enables a proof of conditional MSE-efficiency of the kernel estimator of u^* but it is a bit negative since it states that the ordering of the

[16] The index '*uc*' stands for 'unconditional'. The expectations are taken unconditionally of the firm effects (which are considered to be random variables).

[17] In (23) and (24) we are using that $E_{uc}(\hat{\alpha}_i) = E_{uc}(\tilde{\alpha}_i) = \mu_\alpha$ which can be derived from (2) and (3), where μ_α is the expected value of the unconditional population distribution of the firm effects.

estimated inefficiencies stays the same as the ordering of the FE estimator. Thus, kernel estimation will not improve the ordering of firms compared to the traditional FE estimator.

Lemma 1 *If* $\hat{\alpha}_i \geq \hat{\alpha}_j$ *and* $-\frac{1}{N-1} < \lambda \leq 1$, *then* $\tilde{\alpha}_i \geq \tilde{\alpha}_j$.

Proof

$$\tilde{\alpha}_i - \tilde{\alpha}_j = \frac{\hat{\alpha}_i - \hat{\alpha}_j + \lambda\left(\sum_{j\neq i}\hat{\alpha}_j - \sum_{i\neq j}\hat{\alpha}_i\right)}{1+(N-1)\lambda} = \frac{(\hat{\alpha}_i - \hat{\alpha}_j)(1-\lambda)}{1+(N-1)\lambda} \geq 0 \tag{25}$$

given $\hat{\alpha}_i \geq \hat{\alpha}_j$ and $-\frac{1}{N-1} < \lambda \leq 1$ Q.E.D.

We will now derive a MSE criterion similar to (9) for $\tilde{u}_i - u_i^* = (\tilde{\alpha} - \tilde{\alpha}_i) - (\alpha_{(N)} - \alpha_i)$. Consider the following difference

$$\tilde{u}_i - u_i^* = (\tilde{\alpha} - \tilde{\alpha}_i) - \left(\alpha_{(N)} - \alpha_i\right) = \frac{\left\{(\alpha_{[N]} - \alpha_i) + (\bar{\nu}_{[N]} - \bar{\nu}_i) + (\bar{x}_{[N]} - \bar{x}_i)'(\beta - \hat{\beta})\right\}(1-\lambda)}{1+(N-1)\lambda} - \left(\alpha_{(N)} - \alpha_i\right). \tag{26}$$

The bias conditioning on X is then:

$$E\left(\tilde{u}_i - u_i^* | X\right) = \frac{\left\{(\alpha_{[N]} - \alpha_i) + E\left(\bar{\nu}_{[N]}\right)\right\}(1-\lambda)}{1+(N-1)\lambda} - \left(\alpha_{[N]} - \alpha_i\right), \tag{27}$$

where the index $[N]$ is defined by $\tilde{\alpha}_{[1]} \leq \tilde{\alpha}_{[2]} \cdots \leq \tilde{\alpha}_{[N]} = \tilde{\alpha}$. The expected value $E\left(\bar{\nu}_{[N]}\right)$ is non-negative. Given Lemma 1 it can be shown as follows:

$$E(\hat{\alpha}) = \alpha_{[N]} + E\left(\bar{\nu}_{[N]}\right) + E\left[\bar{x}'_{[N]}(\beta - \hat{\beta})\right] = \alpha_{[N]} + E\left(\bar{\nu}_{[N]}\right), \tag{28}$$

and since $E(\hat{\alpha}) \geq \alpha_{(N)}$:

$$E\left(\bar{\nu}_{[N]}\right) \geq \alpha_{(N)} - \alpha_{[N]} \geq 0. \tag{29}$$

This result is quite intuitive, the estimated maximum $\hat{\alpha}_{[N]}$ is not always equal to $\hat{\alpha}_{(N)}$, since $\hat{\alpha}_{[N]}$ is influenced by a larger random error than $\hat{\alpha}_{(N)}$, on average.

The conditional variance of (26) conditioning on X is:

$$V\left(\tilde{u}_i - u_i^* | X\right) = \frac{\left\{V\left(\bar{\nu}_{[N]}\right) + \frac{\sigma_\nu^2}{T} + (\bar{x}_{[N]} - \bar{x}_i)'V(\hat{\beta}|X)(\bar{x}_{[N]} - \bar{x}_i)\right\}(1-\lambda)^2}{\left[1+(N-1)\lambda\right]^2}. \tag{30}$$

The squared conditional bias is:

$$\left[E\left(\tilde{u}_i - u_i^*|X\right)\right]^2 = \tag{31}$$
$$\frac{A_i^2(1-\lambda)^2}{\left[1+(N-1)\lambda\right]^2} - 2\frac{A_i\left(\alpha_{(N)}-\alpha_i\right)(1-\lambda)}{1+(N-1)\lambda} + \left(\alpha_{(N)}-\alpha_i\right)^2.$$

where $A_i = (\alpha_{[N]} - \alpha_i) + E\left(\bar{\nu}_{[N]}\right)$. The conditional MSE follows as:

$$MSE\left(\tilde{u}_i - u_i^*|X\right) = \tag{32}$$
$$\frac{A_i^2(1-\lambda)^2}{\left[1+(N-1)\lambda\right]^2} - 2\frac{A_i\left(\alpha_{(N)}-\alpha_i\right)(1-\lambda)}{1+(N-1)\lambda} + \left(\alpha_{(N)}-\alpha_i\right)^2 +$$
$$\frac{B_i(1-\lambda)^2}{\left[1+(N-1)\lambda\right]^2},$$

where $B_i = \left\{V\left(\bar{\nu}_{[N]}\right) + \frac{\sigma_\nu^2}{T} + (\bar{x}_{[N]} - \bar{x}_i)'V(\hat{\beta}|X)(\bar{x}_{[N]} - \bar{x}_i)\right\}$. Finally we consider the global conditional MSE criterion given as follows:

$$\sum_{i=1}^{N} MSE\left(\tilde{u}_i|X\right) = \tag{33}$$
$$\frac{\sum_{i=1}^N A_i^2(1-\lambda)^2}{\left[1+(N-1)\lambda\right]^2} - \frac{\sum_{i=1}^N A_i\left(\alpha_{(N)}-\alpha_i\right)(1-\lambda)}{1+(N-1)\lambda} + \sum_{i=1}^{N}\left(\alpha_{(N)}-\alpha_i\right)^2 +$$
$$\frac{\sum_{i=1}^N B_i(1-\lambda)^2}{\left[1+(N-1)\lambda\right]^2}.$$

For the firm effects we proved global conditional MSE-efficiency compared to the FE estimator by observing that the first-order derivative of the squared conditional bias was zero for $\lambda = 0$ while the derivative for the variance was negative. For (33) the squared bias is not zero for $\lambda = 0$, however, the first order derivative of the criterion results in one critical point and with similar assumptions as in Theorem 1 it is straightforward to show that this critical point is positive and satisfy the second order condition to qualify as global minimum on an interval including $\lambda = 0$, i.e. on an interval including the traditional FE estimator. The critical point is written as follows:

$$\lambda^{**} = \frac{\sum_{i=1}^N A_i^2 - \sum_{i=1}^N A_i\left(\alpha_{(N)}-\alpha_i\right) + \sum_{i=1}^N B_i}{\sum_{i=1}^N A_i^2 + (N-1)\sum_{i=1}^N A_i\left(\alpha_{(N)}-\alpha_i\right) + \sum_{i=1}^N B_i}. \tag{34}$$

This bandwidth has the support $\lambda^{**} = [0, 1]$. It is easily seen by the following results:

$$A_i^2 \geq A_i\left(\alpha_{(N)} - \alpha_i\right), \tag{35}$$

$$A_i \geq 0, \quad \text{for all } i = 1, 2, \ldots, N. \tag{36}$$

Both result are verified by noting $A_i = (\alpha_{[N]} - \alpha_i) + E\left(\bar{\nu}_{[N]}\right) = E(\hat{\alpha}) - \alpha_i$ (and $E(\hat{\alpha}) \geq \alpha_{(N)}$). And since obviously $B_i \geq 0$ for all $i = 1, 2, \ldots, N$, $\lambda^{**} \geq 0$. If $B_i > 0$ the inequality for the bandwidth is also strict $\lambda^{**} > 0$. Furthermore since $(N-1)\sum_{i=1}^{N} A_i\left(\alpha_{(N)} - \alpha_i\right) \geq -\sum_{i=1}^{N} A_i\left(\alpha_{(N)} - \alpha_i\right)$, $\lambda^{**} \leq 1$ with equality if and only if $\alpha_{(N)} - \alpha_i = 0$ for all $i = 1, 2, \ldots, N$. Thus, appropriately the bandwidth equals one when there in no inefficiency in the sample.

We now summarize these results into a theorem.

Theorem 2 [18] *If N, T are finite and σ_ν^2 is finite and nonzero*[19]*, then there exists a bandwidth $\lambda^{**} > 0$, such that*

$$\begin{aligned} &\sum_{i}^{N} MSE\left(\tilde{u}_i \middle| X; \lambda = \lambda^{**}\right) < \\ &\sum_{i}^{N} MSE\left(\tilde{u}_i \middle| X; \lambda = 0\right) = \sum_{i}^{N} MSE(\hat{u}_i|X). \end{aligned} \tag{37}$$

Proof A finite N assures that the denominator of λ^{**} does not make λ^{**} to tend to zero, while a finite T and non-zero σ_ν^2 assures that the numerator of λ^{**} is not zero and nor tends to this value. The first order derivative of (33) is well-defined for all $\lambda > -\frac{1}{N-1}$ and the second order derivative is positive for all $-\frac{1}{N-1} < \lambda < \frac{1}{2(N-1)} + \frac{3}{2}\lambda^{**}$. Thus, λ^{**} is a global minimum on an interval including $\lambda = 0$.[20] Q.E.D.

Hence, the traditional FE estimator of inefficiency comparing to the best firm in the sample. However, for u_i we conclude, without any formal proof, that a similar proof is not possible. On the contrary, the FE estimator may in some rare cases be efficient. This may occur when $E(\hat{\alpha}) \leq \alpha$. Thus, the FE estimator may underestimate α and in this case the kernel-FE estimator will underestimate even more, and when σ_ν^2/T is arbitrary small the FE estimator may be a better choice in terms of bias and MSE (conditional as unconditional), and optimally the estimators should coincide in this case.

[18] In Theorem 1 we condition on X since this enables feasible estimation of λ^*. However, this is not necessary to prove MSE-efficiency. And the same is true for this theorem. In this case the only change necessary to prove efficiency without condition on X is to change

$$(\bar{x}_{[N]} - \bar{x}_i)' V(\hat{\beta}|X)(\bar{x}_{[N]} - \bar{x}_i)$$

into

$$E\left[(\bar{x}_{[N]} - \bar{x}_i)'(\beta - \hat{\beta})(\beta - \hat{\beta})'(\bar{x}_{[N]} - \bar{x}_i)\right] = E\left[(\bar{x}_{[N]} - \bar{x}_i) V(\hat{\beta}|X)(\bar{x}_{[N]} - \bar{x}_i)\right].$$

[19] These assumptions are made so that the first order derivatives of the squared conditional bias and the conditional variance are well-defined and to assure strict MSE efficiency. Standard regularity assumptions are implicitly made for the FE covariance matrix, $V(\hat{\beta}|X)$ (e.g. Wooldridge 2010).

[20] See Appendix II for the results of the second order derivative.

Nevertheless, as an estimator of u_i^* the kernel-FE estimator has an advantage. However, unfortunately we cannot find a feasible estimator of λ^{**}. The estimator $\hat{\alpha}$ could be used for either $\alpha_{[N]}$ or $\alpha_{(N)}$ but not for both and there are no estimators for $E(\bar{\nu}_{[N]})$ and $V(\bar{\nu}_{[N]})$.

Therefore, we will use λ^* and two data-driven bandwidth selectors aiming to find an appropriate bandwidth for estimating α_i.

However, we will also derive a bandwidth based on estimating the differences between firm effects. We believe that this bandwidth gives more emphasis on the collective sample distribution of the firm effects than λ^*. If this bandwidth improves the estimation of the collective sample distribution this should give advantages when estimating the inefficiency, $u_i^* = \alpha_{(N)} - \alpha_i$, since it depends on an accurate spread of the sample distribution.

Consider the difference between the kernel estimator of two firms i and j:

$$\tilde{\delta}_{j,i} = \tilde{\alpha}_j - \tilde{\alpha}_i = \frac{\left[(\alpha_j - \alpha_i) + (\bar{\nu}_j - \bar{\nu}_i) + (\bar{x}_j - \bar{x}_i)'(\beta - \hat{\beta})\right](1-\lambda)}{1 + (N-1)\lambda} \tag{38}$$

and the square of difference between the this difference and the population counterpart is as follows:

$$\begin{aligned}\left(\tilde{\delta}_{j,i} - \delta_{j,i}\right)^2 = {} & \frac{\left[(\alpha_i - \alpha_j)N + (\bar{\nu}_i - \bar{\nu}_j) + (\bar{x}_i - \bar{x}_j)'(\beta - \hat{\beta})\right]^2 \lambda^2}{\left[1 + (N-1)\lambda\right]^2} + \\ & \frac{-2\left[(\alpha_i - \alpha_j)N + (\bar{\nu}_i - \bar{\nu}_j) + (\bar{x}_i - \bar{x}_j)'(\beta - \hat{\beta})\right]}{\left[1 + (N-1)\lambda\right]^2} \times \\ & \frac{\left[(\bar{\nu}_i - \bar{\nu}_j) + (\bar{x}_i - \bar{x}_j)'(\beta - \hat{\beta})\right]\lambda}{\left[1 + (N-1)\lambda\right]^2} + \frac{\left[(\bar{\nu}_i - \bar{\nu}_j) + (\bar{x}_i - \bar{x}_j)'(\beta - \hat{\beta})\right]^2}{\left[1 + (N-1)\lambda\right]^2}.\end{aligned} \tag{39}$$

Then take expected value conditioning on X:

$$E\left[\left(\tilde{\delta}_{j,i}-\delta_{j,i}\right)^2\Big|X\right]= \tag{40}$$
$$\frac{\left[(\alpha_i-\alpha_j)^2N^2+2\frac{\sigma_\nu^2}{T}+(\bar{x}_i-\bar{x}_j)'V(\hat{\beta}|X)(\bar{x}_i-\bar{x}_j)\right]\lambda^2}{\left[1+(N-1)\lambda\right]^2}+$$
$$\frac{-2\left[2\frac{\sigma_\nu^2}{T}+(\bar{x}_i-\bar{x}_j)'V(\hat{\beta}|X)(\bar{x}_i-\bar{x}_j)\right]\lambda}{\left[1+(N-1)\lambda\right]^2}+$$
$$\frac{\left[2\frac{\sigma_\nu^2}{T}+(\bar{x}_i-\bar{x}_j)'V(\hat{\beta}|X)(\bar{x}_i-\bar{x}_j)\right]}{\left[1+(N-1)\lambda\right]^2}.$$

We now simplify this expression a bit by assuming that N is large. The covariance matrix of $\hat{\beta}$ converges in probability to zero at a rate implied by:

$$V(\hat{\beta}|X)=O_p(N^{-1}),$$

and if $\lambda=O_p(N^{-1})$ then

$$E\left[\left(\tilde{\delta}_{j,i}-\delta_{j,i}\right)^2\Big|X\right]= \tag{41}$$
$$\frac{\left[(\alpha_i-\alpha_j)^2N^2\right]\lambda^2+2\frac{\sigma_\nu^2}{T}}{\left[1+(N-1)\lambda\right]^2}+O_p(N^{-1}).$$

This gives the following approximation:

$$E\left[\left(\tilde{\delta}_{j,i}-\delta_{j,i}\right)^2\Big|X\right]\sim\frac{\left[(\alpha_i-\alpha_j)^2N^2\right]\lambda^2+2\frac{\sigma_\nu^2}{T}}{\left[1+(N-1)\lambda\right]^2}. \tag{42}$$

A firm-average MSE-criterion based on this approximation is then

$$\sum_{i=N}^{N}E\left[\left(\tilde{\delta}_{j,i}-\delta_{j,i}\right)^2\Big|X\right]/N\sim\frac{\sum_{i=1}^{N}\left[(\alpha_i-\alpha_j)^2N^2\right]\lambda^2+2\frac{\sigma_\nu^2}{T}}{N\left[1+(N-1)\lambda\right]^2}, \tag{43}$$

and this criterion gives the following optimal bandwidth:

$$\lambda(\alpha_j)^*=\frac{2\frac{\sigma_\nu^2}{T}}{N\frac{\sum_{i=1}^{N}(\alpha_i-\alpha_j)^2}{N-1}}. \tag{44}$$

This bandwidth is undefined if there is no inefficiency, $\alpha_1 = \cdots = \alpha_N$. Because of this we will instead use

$$\lambda(\alpha_j)^{**} = \frac{2\frac{\sigma_\nu^2}{T}}{N\frac{\sum_{i=1}^{N}(\alpha_i-\alpha_j)^2}{N-1} + 2\frac{\sigma_\nu^2}{T}}. \tag{45}$$

For sufficiently large N, $\lambda(\alpha_j)^*$ and $\lambda(\alpha_j)^{**}$ are indistinguishable but if $\alpha_1 = \cdots = \alpha_N = \alpha$ the latter bandwidth will be defined and equal to one, i.e. will produce the appropriate pooled estimate of the common intercept, α.

It is tempting to set $\alpha_j = \alpha_{(N)}$ to use the bandwidth $\lambda(\alpha_{(N)})^{**}$. However, in this case the bandwidth actually should include $\alpha_{[N]}$, $E(\bar{\nu}_{[N]})$ and $V(\bar{\nu}_{[N]})$ as well, just as for λ^{**} in (34). We instead set $\alpha_j = \bar{\alpha}$ and estimate

$$\lambda(\bar{\alpha})^{**} = \frac{2\frac{\sigma_\nu^2}{T}}{Ns_\alpha^2 + 2\frac{\sigma_\nu^2}{T}}, \tag{46}$$

which is easy to estimate, satisfy convergence rates for the large sample properties presented in Section 4 and proves to work very well in simulations.[21]

In the following two sections Monte Carlo simulations are presented to further investigate the small sample performance of the kernel estimator compared to the FE estimator both when it comes to the firm effects and the technical inefficiencies.

6 Monte Carlo Simulations for the Firm effects

In Section 3 efficiency of the kernel estimator of α_i, is shown based on a global conditional MSE-criterion.

All the unknowns are straightforward to estimate consistently as N grows. Therefore, the bandwidth should be suitable in large N and small T settings. The estimators of the unknowns are also bounded in probability as $T \to \infty$ which assures that the asymptotic results in Section 4 also holds for the estimator of λ^* (and of $\lambda(\bar{\alpha})^{**}$). For the estimators of the unknowns in λ^* (and in $\lambda(\bar{\alpha})^{**}$) see Appendix III.

For testing the conditional MSE-efficiency of kernel estimation of the firm effects the MSE-criterion in (9) is used to determine the 'best' kernel estimator and then this kernel estimator is compared to the traditional FE estimator by the same criterion.

[21] We have also tried to use an estimate of $\lambda(\alpha_{(N)})^{**}$, however, it works poorly in most simulations. Nevertheless, in situations where there are a large probability for $\alpha_{(N)} = \alpha_{[N]}$: $E(\bar{\nu}_{[N]}) \approx 0$ and $V(\bar{\nu}_{[N]}) \approx \frac{\sigma_\nu^2}{T}$, and consequently $\lambda(\alpha_{(N)})^{**}$ should be a good choice of bandwidth for estimating u_i^*.

Except for the kernel estimator based on the optimal bandwidth, $\hat{\lambda}^*$, we also consider two kernel estimators based on data-driven bandwidths. The corrected Akaike information criterion (AIC), henceforth denoted $\hat{\lambda}_{aic}$, and the popular least squares cross-validation bandwidth, denoted $\hat{\lambda}_{cv}$ (see Li and Racine 2007, and references therein). Ouyang et al (2009) provide further details on the least squares cross-validation bandwidth used in the case of kernel regression with unordered regressors. Hurvich et al (1998) proposed the corrected AIC selector. They argue and show by simulations that the bandwidth selector can compete and often performs better than Plug-in selectors. Thus, if they are right we should expect that this bandwidth selector performs better than the Plug-in estimator we have derived. Both the least squares cross-validation and the AIC selectors are readily available in the *np* package in *R* for kernel regression with unordered kernels as e.g. the estimator given by (3) and (4) (Hayfield and Racine 2008).

Based on the conditional MSE-criterion in (9) we select the 'best' kernel estimator (bandwidth) based the following ratio-criterion:

$$C\left[\tilde{\alpha}_i(\hat{\lambda})\right] = \frac{1}{B}\sum_{b=1}^{B}\frac{\sum_i^N MSE\left(\tilde{\alpha}_i\middle|X;\lambda=\hat{\lambda}_b\right)}{\sum_i^N MSE\left(\tilde{\alpha}_i\middle|X;\lambda=\lambda_b^*\right)} \tag{47}$$

where B is the number of Monte Carlo replications (1000), λ_b^* is the bandwidth minimizing the global conditional MSE criterion at replication b, and $\hat{\lambda}_b$ is either one of the estimated bandwidths $\hat{\lambda}_b^*$, $\hat{\lambda}_{aic,b}$ and $\hat{\lambda}_{cv,b}$ at replication b. Thus, the closer $C\left[\tilde{\alpha}_i(\hat{\lambda})\right]$ is to one the better the performance is of the estimator measured by the global conditional MSE-criterion. When the 'best' kernel-estimator is selected it is compared to the traditional FE estimator by comparing 'average mean square errors' defined as

$$AMSE\left(\tilde{\alpha}_i\middle|X;\lambda=\hat{\lambda}\right) = \frac{1}{NB}\sum_{b=1}^{B}\sum_{i=1}^{N}MSE\left(\tilde{\alpha}_i\middle|X;\lambda=\hat{\lambda}_b\right)$$

and

$$AMSE\left(\tilde{\alpha}_i\middle|X;\lambda=0\right) = \frac{1}{NB}\sum_{b=1}^{B}\sum_{i=1}^{N}MSE\left(\tilde{\alpha}_i\middle|X;\lambda=0\right).$$

We consider two different DGPs ('data generating processes'). These DGPs are based on the linear panel data model in (1). The first model, denoted M_1, is defined as follows:

$$y_{M1,it} = \tag{48}$$
$$0.5x_{M1,it} + 0.05x_{M1,it}^2 + 2D_{M1,it} + \sqrt{T}(\bar{x}_{M1,i} + \bar{D}_{M1,i}) + U_{M1,i} + \nu_{M1,it},$$

where $x_{M1,it}$ and $U_{M1,i}$ are drawn from the uniform distribution $U(0,1)$, $D_{M1,it}$ is drawn from the Bernoulli distribution with probability 0.6. The random error $\nu_{M1,it}$ is drawn from the normal distribution $N(0,\sigma_\nu^2)$. The firm

effects are generated by $\sqrt{T}(\bar{x}_{M1,i} + \bar{D}_{M1,i}) + U_{M1,i}$. The variance of the firm effects is $\sigma_\alpha^2 = (1/12 + 0.6 * 0.4) + 1/12 \approx 0.4067$ and the variance of the random error is set in accordance to $\gamma = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\nu^2} = 0.25, 0.5$ and 0.75.

We especially expect the kernel estimators to have big advantages when the random error is influential, due to the bias-variance tradeoff, i.e. when γ is small and T is small. The time observations are varied as $T = 5, 15, 30$, while the number of firms are set as $N = 10, 150, 300$.

The second DGP, denoted M_2, is constructed as follows:

$$y_{M2,it} = 0.5x_{M2,1it} + 0.3x_{M2,2it} + 0.2x_{M2,3it} + \alpha_{M2,i} + \nu_{M2,it}, \tag{49}$$

where the regressors are constructed to incorporate auto-dependence over time given by $AR(1)$ processes:

$$x_{M2,kit} = \phi x_{M2,ki,t-1} + \epsilon_{M2,kit}, \tag{50}$$

where $\phi = 0.5$ and $\epsilon_{M2,kit}$ is a random draw from the standard normal distribution. The initial value x_{ki1} is drawn from the normal distribution with mean zero and variance $\frac{1}{1-\phi^2}$. The firm effects are generated as follows:[22]

$$\alpha_{M2,i} = \sqrt{T} \sum_{k=1}^{3} \bar{x}_{M2,kit}/5. \tag{51}$$

Again the random error is normally distributed with mean zero and variance set in accordance to $\gamma = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\nu^2} = 0.25, 0.5$ and 0.75, while N and T are set as for M_1.

6.1 Results

The comparison of the Kernel estimators are presented in table form in Appendix IV for both M_1 and M_2 (Table 9-12). All three kernel estimators gives very similar results and in a few cases when γ and T is large $C\left[\tilde{\alpha}_i(\hat{\lambda})\right]$ equals one with four decimal rounding, i.e. the kernel estimates are relatively close to optimal. Nevertheless, overall the least-squares cross-validation estimator works slightly better than the other two for both M_1 and M_2. The comparison with the traditional FE estimator is, therefore, conducted with the least-squares cross-validation kernel estimator.

In Table 1 the results for the comparison of the estimators for M_1 are provided and in Table 2 for M_2. For both M_1 and M_2 the kernel estimator

[22] The AR(1)-setup makes it relatively easy to show

$$\sigma_\alpha^2 = \frac{T}{5^2}\left(\frac{1}{1-\phi^2}\frac{1}{T} + \frac{2}{T^2}\frac{1}{1-\phi^2}\sum_{t=1}^{T-1}\sum_{s>t}\phi^{s-t}\right),$$

which is needed to compare the performance of the estimators to the optimality given by the global conditional MSE criterion.

outperforms the FE estimator in all cases. Thus, these results strongly supports the theoretical result given by Theorem 1.

As expected the difference in performance is largest when there is much influence of random error, i.e. T and γ are small. The unbiased FE estimator has no means to reduce the variance in this case, which kernel estimation enables through the bias-variance tradeoff.

Hence, kernel estimation has the upper hand when it come to estimating firm effects and the different kernel estimators give very similar results.

In the next section simulations are presented for comparing traditional FE estimation of inefficiency to kernel estimation.

7 Monte Carlo Simulations for the Inefficiencies

The DGP selected for the Monte Carlo simulations of the inefficiencies is based on the Cobb-Douglas stochastic frontier model with four inputs. The model is denoted M_3 and is constructed as:

$$y_{M3,it} = \prod_{k}^{4} x_{M3,kit}^{\beta_j} exp(0.5 - u_i + \nu_{M3,it}) \tag{52}$$

where the inefficiencies , $u_i,\ i = 1, \ldots, N$, are independently and identically half-normally distributed, $u_i \sim |N(0, \sigma^2)|$, and the random error $\nu_{M3,it}$ is drawn from the normal distribution, $\nu_{M3,it} \sim N(0, \sigma_\nu^2)$. The variance σ^2 is set such that $\sigma_u^2 = 0.1$ and σ_ν^2 is set in accordance to $\gamma = \frac{\sigma_u^2}{\sigma_u^2+\sigma_\nu^2} = 0.25, 0.5$ or 0.75. As described in previous sections and shown by e.g. Wang and Schmidt (2009) the performance of the FE estimator depends greatly on the ratio $\frac{\sigma_\nu^2}{T}$ contra σ_u^2, and on N. When N is large and the random error, σ_ν^2, is large compared to σ_u^2 and T is small, the upward bias in $\hat{\alpha}$ is large as well as the variance. In these cases kernel estimation may gain much due to the bias-variance tradeoff and the shrinkage toward the mean firm effect. As in the previous section we run the simulations both for a small, moderate sized and large N: 10, 150 and 300. We set T to 5 (small), 15 (relatively large) and 30 (large). To generate $x_{M3,kit}$ we use the standard uniform distribution

$$\tilde{x}_{kit} \sim \text{Uniform}(0, 1)$$

and

$$x_{M3,kit} = \tilde{x}_{kit} exp(u_i/\varepsilon), \tag{53}$$

where

$$\varepsilon = \beta_1 + \beta_2 + \beta_3 + \beta_4 = 0.5 + 0.25 + 0.15 + 0.1 = 1.$$[23]

[23] Thus, we consider constant returns to scale (CRS). However, we have also conducted simulations based on both decreasing and increasing returns to scale, but the results do not change any conclusions compared to the CRS case so the simulations are not included. The results are provided on request.

N	T	γ	FE	$\hat{\lambda}_{cv}$	Ratio
10	5	0.25	0.5887	0.5092	0.86
10	5	0.50	0.1962	0.1829	0.93
10	5	0.75	0.0654	0.0635	0.97
10	15	0.25	0.1689	0.1564	0.93
10	15	0.50	0.0563	0.0544	0.97
10	15	0.75	0.0188	0.0185	0.99
10	30	0.25	0.0816	0.0778	0.95
10	30	0.50	0.0272	0.0267	0.98
10	30	0.75	0.0091	0.0090	0.99
150	5	0.25	0.2648	0.1740	0.66
150	5	0.50	0.0883	0.0747	0.85
150	5	0.75	0.0294	0.0277	0.94
150	15	0.25	0.0870	0.0735	0.84
150	15	0.50	0.0290	0.0273	0.94
150	15	0.75	0.0097	0.0095	0.98
150	30	0.25	0.0434	0.0397	0.91
150	30	0.50	0.0145	0.0140	0.97
150	30	0.75	0.0048	0.0048	0.99
300	5	0.25	0.2543	0.1630	0.64
300	5	0.50	0.0848	0.0712	0.84
300	5	0.75	0.0283	0.0266	0.94
300	15	0.25	0.0842	0.0706	0.84
300	15	0.50	0.0281	0.0264	0.94
300	15	0.75	0.0094	0.0092	0.98
300	30	0.25	0.0420	0.0383	0.91
300	30	0.50	0.0140	0.0136	0.97
300	30	0.75	0.0047	0.0046	0.99

Table 1: $AMSE\left(\tilde{\alpha}_i \middle| X; \lambda = \hat{\lambda}\right)$ comparison for FE versus Kernel-FE ($\hat{\lambda}_{cv}$). DGP: M_1

N	T	γ	FE	$\hat{\lambda}_{cv}$	Ratio
10	5	0.25	0.6769	0.4502	0.67
10	5	0.50	0.2256	0.1864	0.83
10	5	0.75	0.0752	0.0692	0.92
10	15	0.25	0.2492	0.2084	0.84
10	15	0.50	0.0831	0.0772	0.93
10	15	0.75	0.0277	0.0270	0.97
10	30	0.25	0.1290	0.1168	0.91
10	30	0.50	0.0430	0.0414	0.96
10	30	0.75	0.0091	0.0090	0.99
150	5	0.25	0.4487	0.2810	0.63
150	5	0.50	0.1496	0.1243	0.83
150	5	0.75	0.0499	0.0467	0.94
150	15	0.25	0.2201	0.1833	0.83
150	15	0.50	0.0734	0.0687	0.94
150	15	0.75	0.0245	0.0239	0.98
150	30	0.25	0.1216	0.1105	0.91
150	30	0.50	0.0405	0.0392	0.97
150	30	0.75	0.0048	0.0048	0.99
300	5	0.25	0.4416	0.2760	0.63
300	5	0.50	0.1472	0.1224	0.83
300	5	0.75	0.0491	0.0459	0.94
300	15	0.25	0.2191	0.1825	0.83
300	15	0.50	0.0730	0.0684	0.94
300	15	0.75	0.0243	0.0238	0.98
300	30	0.25	0.1214	0.1103	0.91
300	30	0.50	0.0405	0.0392	0.97
300	30	0.75	0.0135	0.0133	0.99

Table 2: $AMSE\left(\tilde{\alpha}_i \middle| X; \lambda = \hat{\lambda}\right)$ comparison for FE versus Kernel-FE ($\hat{\lambda}_{cv}$). DGP: M_2

Thus, there is some dependence induced, which implies the following covariance:[24]

$$Cov\left[log(x_{M3,kit}), u_i\right] = \frac{1}{\varepsilon}\sigma_u^2.$$

Note that the variable $\tilde{x}_{kit}$ in this form has economical meaning, it is the efficient amount of input k, given output $y_{M3,it}$.

To compare estimators of inefficiency, u_i^*, we are not using the MSE-criterion in (33) since although the DGP is known we are not able to deduce the variance and the expected value of $\bar{\nu}_{[N]}$. Instead we use the following 'average square error' criterion:

$$ASE(\tilde{u}_i) = \frac{1}{N}\sum_{i=1}^{N}(\tilde{u}_i - u_i^*)^2 \tag{54}$$

to compare the small-sample performances of the different estimators. As in the previous section the kernel estimators are first compared by

$$C\left[\tilde{u}_i(\hat{\lambda})\right] = \frac{1}{B}\sum_{b=1}^{B}\frac{ASE\left(\tilde{u}_i(\hat{\lambda}_b)\right)}{ASE\left(\tilde{u}_i(\lambda_b^0)\right)}, \tag{55}$$

where $\hat{\lambda}_b$ is either one of the estimated bandwidths and λ_b^0 is the optimal bandwidth given the ASE-criterion at replication b. And in a second step the 'best' estimator is compared to the FE estimator.The kernel estimators compared are the three in the previous section and the estimator based on $\hat{\lambda}(\bar{\alpha})^{**}$ given by (46). The comparison of the 'best' kernel estimator and the FE estimator also includes the bootstrap bias-reduced versions of the two estimators. The bootstrap bias-reduction estimator was compared to two other bias-corrected methods in Satchachai and Schmidt (2010) and showed to have superior performance as an estimator of $\alpha_{(N)}$ in terms of MSE.[25].

The comparison of the kernel and the traditional FE estimators of course also includes bias performance, with the following average bias measure:

$$\frac{1}{B}\sum_{b=1}^{B}(\tilde{\alpha}_b - \alpha_{(N),b}) \tag{56}$$

where $\alpha_{(N),b}$ is the largest firm effect at replication b and $\tilde{\alpha}_b$ is the estimate for either one of the estimators included in the comparison.

[24] This is the main advantage of FE estimators compared to random effects and maximum likelihood estimators of technical (in-) efficiencies. It is hard to believe that the firm specific inefficiency should not be correlated (dependent) with the inputs selected by the same firm.

[25] See Kim et al (2007) for details of the bootstrap bias correction estimator.

7.1 Results

Among the kernel estimators the one based on the bandwidth $\hat{\lambda}(\bar{\alpha})^{**}$ proves to work best in almost all cases, except for when $N = 10$. The comparison is provided in Appendix IV, Table 13 and Table 14.

The results for the FE estimator are, therefore, compared to the kernel estimator with bandwidth $\hat{\lambda}(\bar{\alpha})^{**}$. With and without bias correction, first for the bias and then for the ASE-criterion, averaged over all b.

The results of the bias-comparison are contained in Table 3 and Table 4. 'Boot.' in front of the estimator's name stands for bias-corrected.

The bootstrap bias-reduction reduces the bias of the FE estimator in all cases. The estimator also produces smaller average bias among all four estimators when $N = 10$, otherwise not.

Bias reduction for the kernel estimator underestimates when the influence of random error is relatively large, i.e. when γ and T are small. In this case the kernel estimator without bias reduction produces relatively unbiased estimates and outperforms the other estimators. For example when $N = 150$, $T = 5$ and $\gamma = 0.25$ the average bias (absolute values) for the kernel estimator without bias-correction is only 5.7% of the average bias of the bias-corrected FE estimator, which is the second best estimator in this particular case.

However, when there is not too much influence of random error (and $N > 10$) bias reduction of the kernel estimator works very well. Overall this estimator works best in terms of average bias. And kernel estimation outperforms traditional FE estimation in all cases except for $N = 10$.

In practice it will be important to distinguish between cases when there is large influence of random error and when it is not. In the empirical example we show how this can be done.

The comparison based on average ASE is presented in Table 5 and Table 6. In terms of average ASE the kernel estimators outperforms the FE estimators in all cases. The bias-reduced kernel estimator when γ and T are not too small and otherwise the kernel estimator without bias-reduction. The gains compared to traditional FE estimation are largest in the latter case. This is again an example of the merits of the bias-variance tradeoff of kernel estimation. The large variance for FE estimation when γ and T are small is transferred to the inefficiencies.

The distributions in Figure 1 captures the patterns recognized in the tables.[26] When there is much random error, $\gamma = 0.25$, the kernel estimator produces a distribution centered around the average of the distribution generated by the DGP. The bias-reduced kernel underestimates while the FE estimators clearly overestimate. The variance of the distribution is too large for the FE estimators while it is a bit too small for the kernel counterparts. No

[26] Figure 1 includes three kernel density plots constructed from the ordered statistics of the GDP and the estimators. For the DGP the data points are constructed as: $\frac{\sum_{b=1}^{B} u^*_{(1),b}}{B} \leq \frac{\sum_{b=1}^{B} u^*_{(2),b}}{B} \leq \cdots \leq \frac{\sum_{b=1}^{B} u^*_{(N),b}}{B}$. The data points for the estimators are constructed in the same fashion. We call the distributions 'average order statistics distribution of inefficiency'.

N	T	p	FE	$\hat{\lambda}(\bar{\alpha})^{**}$	Boot. FE	Boot. $\hat{\lambda}(\bar{\alpha})^{**}$
10	5	0.25	0.2162	-0.0947	0.1136	-0.3202
10	5	0.50	0.0934	-0.0478	0.0442	-0.1300
10	5	0.75	0.0368	-0.0179	0.0146	-0.0461
10	15	0.25	0.0875	-0.0547	0.0310	-0.1401
10	15	0.50	0.0347	-0.0200	0.0088	-0.0509
10	15	0.75	0.0125	-0.0067	0.0017	-0.0184
10	30	0.25	0.0525	-0.0260	0.0180	-0.0695
10	30	0.50	0.0211	-0.0070	0.0066	-0.0228
10	30	0.75	0.0079	-0.0017	0.0020	-0.0078
150	5	0.25	0.4675	-0.0155	0.2723	-0.3856
150	5	0.50	0.2410	0.0520	0.1370	-0.0949
150	5	0.75	0.1217	0.0578	0.0672	-0.0054
150	15	0.25	0.2419	0.0522	0.1243	-0.1028
150	15	0.50	0.1220	0.0580	0.0602	-0.0111
150	15	0.75	0.0593	0.0389	0.0271	0.0052
150	30	0.25	0.1577	0.0608	0.0758	-0.0336
150	30	0.50	0.0775	0.0462	0.0342	0.0008
150	30	0.75	0.0370	0.0270	0.0149	0.0046
300	5	0.25	0.5354	0.0150	0.3266	-0.3841
300	5	0.50	0.2811	0.0808	0.1678	-0.0791
300	5	0.75	0.1450	0.0785	0.0842	0.0083
300	15	0.25	0.2807	0.0799	0.1492	-0.0914
300	15	0.50	0.1457	0.0790	0.0759	0.0014
300	15	0.75	0.0740	0.0529	0.0372	0.0148
300	30	0.25	0.1867	0.0853	0.0954	-0.0200
300	30	0.50	0.0961	0.0637	0.0479	0.0129
300	30	0.75	0.0485	0.0383	0.0236	0.0129

Table 3: Average bias of the estimators with respect to $\alpha_{(N)}$ (and u_i^*). DGP: M_3

N	T	p	FE	$\hat{\lambda}(\bar{\alpha})^{*}$	Boot. FE	Boot. $\hat{\lambda}(\bar{\alpha})^{*}$
10	5	0.25	0.4380	1.0000	0.8334	0.2957
10	5	0.50	0.4728	0.9238	1.0000	0.3395
10	5	0.75	0.3954	0.8114	1.0000	0.3157
10	15	0.25	0.3543	0.5670	1.0000	0.2213
10	15	0.50	0.2547	0.4415	1.0000	0.1737
10	15	0.75	0.1329	0.2472	1.0000	0.0905
10	30	0.25	0.3435	0.6942	1.0000	0.2593
10	30	0.50	0.3109	0.9413	1.0000	0.2884
10	30	0.75	0.2222	1.0000	0.8910	0.2249
150	5	0.25	0.0332	1.0000	0.0570	0.0402
150	5	0.50	0.2156	1.0000	0.3794	0.5479
150	5	0.75	0.0441	0.0929	0.0799	1.0000
150	15	0.25	0.2159	1.0000	0.4204	0.5083
150	15	0.50	0.0914	0.1920	0.1850	1.0000
150	15	0.75	0.0883	0.1347	0.1935	1.0000
150	30	0.25	0.2131	0.5526	0.4434	1.0000
150	30	0.50	0.0101	0.0170	0.0229	1.0000
150	30	0.75	0.1241	0.1701	0.3076	1.0000
300	5	0.25	0.0280	1.0000	0.0460	0.0391
300	5	0.50	0.2812	0.9780	0.4711	1.0000
300	5	0.75	0.0570	0.1053	0.0982	1.0000
300	15	0.25	0.2847	1.0000	0.5358	0.8745
300	15	0.50	0.0093	0.0172	0.0179	1.0000
300	15	0.75	0.1997	0.2792	0.3970	1.0000
300	30	0.25	0.1073	0.2349	0.2098	1.0000
300	30	0.50	0.1345	0.2028	0.2699	1.0000
300	30	0.75	0.2666	0.3377	0.5481	1.0000

Table 4: Ratios compared to the estimator with smallest average absolute value of bias. Value one indicate the estimator with smallest average absolute value of bias. DGP: M_3

N	T	p	FE	$\hat{\lambda}(\bar{\alpha})^{**}$	Boot. FE	Boot. $\hat{\lambda}(\bar{\alpha})^{**}$
10	5	0.25	0.1332	0.0823	0.1172	0.2185
10	5	0.50	0.0401	0.0353	0.0387	0.0604
10	5	0.75	0.0128	0.0125	0.0133	0.0163
10	15	0.25	0.0365	0.0317	0.0346	0.0557
10	15	0.50	0.0114	0.0109	0.0117	0.0146
10	15	0.75	0.0038	0.0038	0.0041	0.0044
10	30	0.25	0.0190	0.0173	0.0194	0.0246
10	30	0.50	0.0063	0.0059	0.0067	0.0072
10	30	0.75	0.0021	0.0021	0.0023	0.0023
150	5	0.25	0.2915	0.0491	0.1644	0.2098
150	5	0.50	0.0830	0.0246	0.0496	0.0358
150	5	0.75	0.0234	0.0116	0.0150	0.0101
150	15	0.25	0.0836	0.0244	0.0461	0.0362
150	15	0.50	0.0235	0.0116	0.0142	0.0101
150	15	0.75	0.0065	0.0045	0.0044	0.0036
150	30	0.25	0.0378	0.0157	0.0218	0.0157
150	30	0.50	0.0105	0.0065	0.0068	0.0054
150	30	0.75	0.0029	0.0023	0.0022	0.0019
300	5	0.25	0.3580	0.0467	0.1936	0.2021
300	5	0.50	0.1032	0.0275	0.0578	0.0313
300	5	0.75	0.0293	0.0140	0.0172	0.0096
300	15	0.25	0.1034	0.0277	0.0535	0.0344
300	15	0.50	0.0296	0.0142	0.0163	0.0099
300	15	0.75	0.0083	0.0056	0.0049	0.0037
300	30	0.25	0.0471	0.0187	0.0245	0.0144
300	30	0.50	0.0135	0.0082	0.0076	0.0053
300	30	0.75	0.0038	0.0029	0.0024	0.0020

Table 5: Average ASE for estimators of u_i^*. DGP: M_3

N	T	p	FE	$\hat{\lambda}(\bar{\alpha})^{**}$	Boot. FE	Boot. $\hat{\lambda}(\bar{\alpha})^{**}$
10	5	0.25	0.6178	1.0000	0.7020	0.3767
10	5	0.50	0.8803	1.0000	0.9115	0.5848
10	5	0.75	0.9769	1.0000	0.9445	0.7675
10	15	0.25	0.8688	1.0000	0.9172	0.5694
10	15	0.50	0.9567	1.0000	0.9339	0.7480
10	15	0.75	0.9831	1.0000	0.9246	0.8452
10	30	0.25	0.9116	1.0000	0.8910	0.7045
10	30	0.50	0.9493	1.0000	0.8935	0.8302
10	30	0.75	0.9743	1.0000	0.9116	0.8943
150	5	0.25	0.1685	1.0000	0.2988	0.2342
150	5	0.50	0.2963	1.0000	0.4965	0.6879
150	5	0.75	0.4310	0.8715	0.6710	1.0000
150	15	0.25	0.2922	1.0000	0.5304	0.6741
150	15	0.50	0.4278	0.8684	0.7071	1.0000
150	15	0.75	0.5535	0.8091	0.8215	1.0000
150	30	0.25	0.4150	1.0000	0.7203	0.9953
150	30	0.50	0.5162	0.8323	0.7998	1.0000
150	30	0.75	0.6600	0.8486	0.8955	1.0000
300	5	0.25	0.1306	1.0000	0.2414	0.2312
300	5	0.50	0.2663	1.0000	0.4758	0.8795
300	5	0.75	0.3280	0.6847	0.5589	1.0000
300	15	0.25	0.2678	1.0000	0.5174	0.8059
300	15	0.50	0.3350	0.6966	0.6089	1.0000
300	15	0.75	0.4442	0.6589	0.7501	1.0000
300	30	0.25	0.3044	0.7677	0.5863	1.0000
300	30	0.50	0.3936	0.6477	0.6966	1.0000
300	30	0.75	0.5193	0.6795	0.8275	1.0000

Table 6: Ratios of average ASE for estimators of u_i^* compared to the smallest – estimated – average ASE. Value one indicates smallest average ASE. DGP: M_3

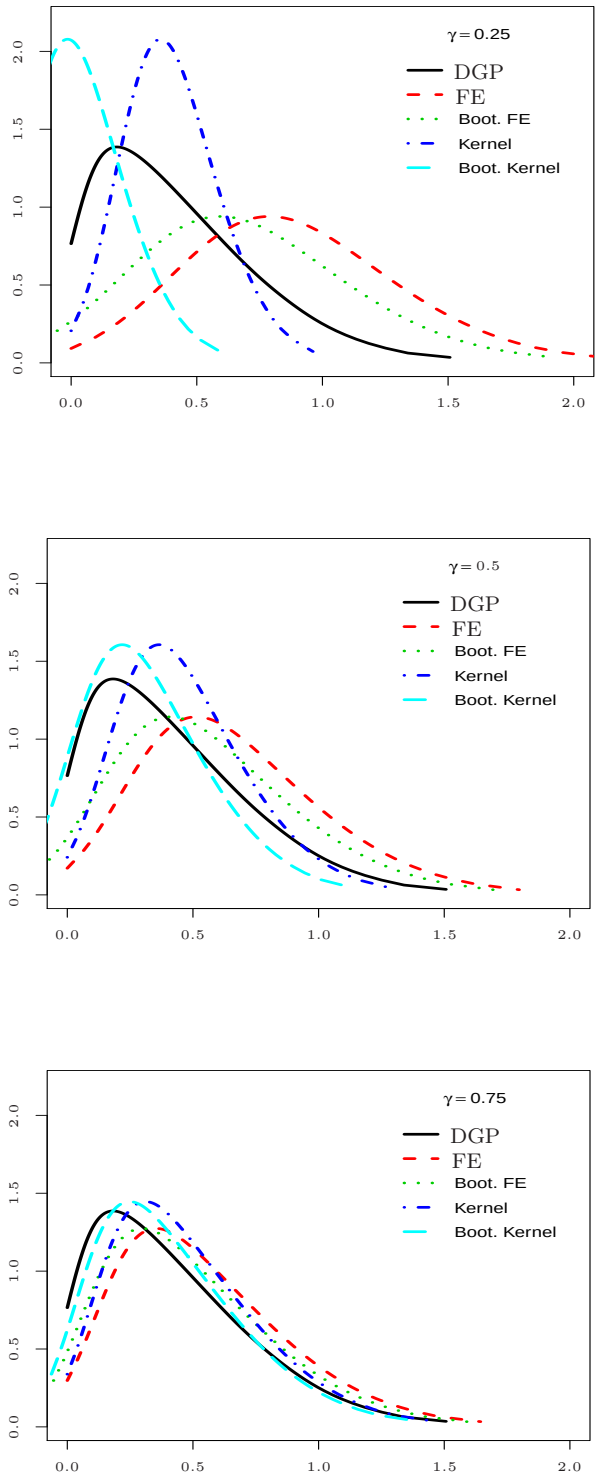

Fig. 1: Kernel densities of average order statistics distributions of inefficiency for $N = 150$, $T = 5$ and $\gamma = 0.25, 0.5, 0.75$.

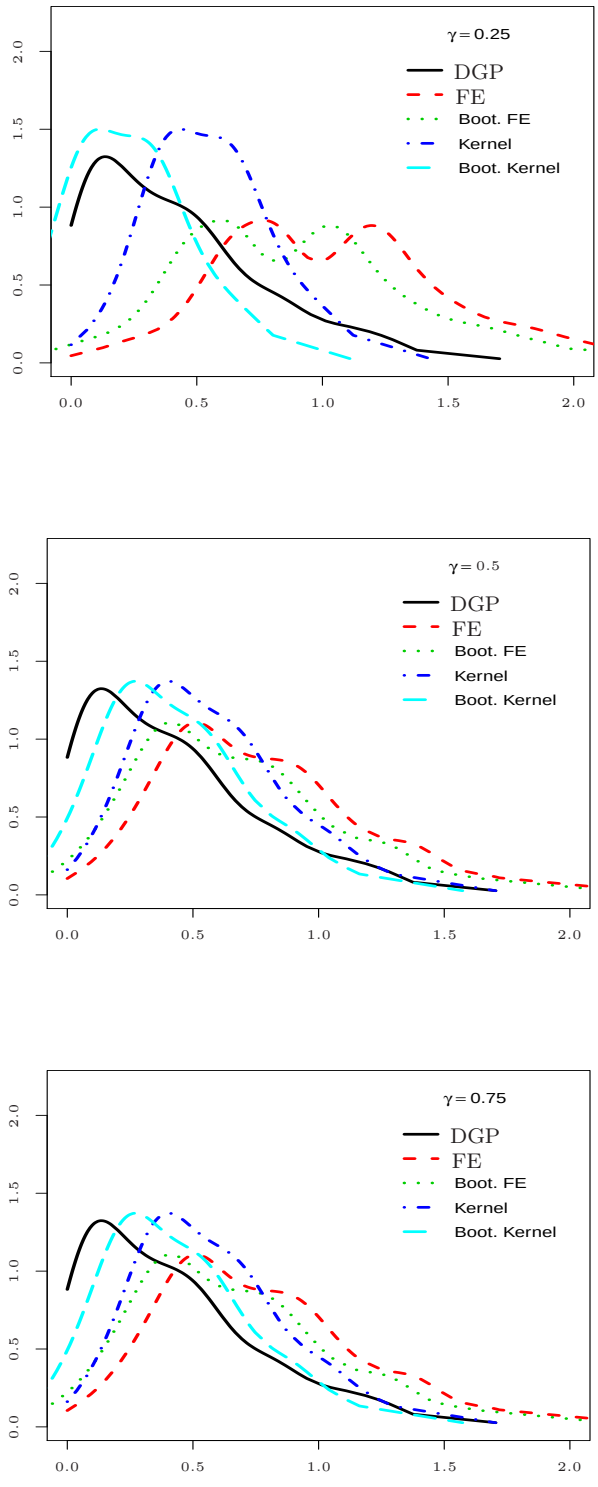

Fig. 2: Kernel densities of inefficiency distributions for a randomly drawn Monte Carlo replication ($b = 94$), $N = 150$, $T = 5$ and $\gamma = 0.25, 0.5, 0.75$.

estimator really captures the shape of the DGP-distribution. Nevertheless, the kernel estimator without bias-reduction give some information about average inefficiency. And although underestimated, the kernel estimators give better guidance about the variability.

On the other hand when γ is large the estimators are able to capture the shape quite well, where the bias-reduced kernel estimator comes closest.[27] A drawback with bias-reduction is though negative inefficiencies that occur since the maximum firm effect is corrected downwards.

Figure 2 contains distributions of inefficiencies from a randomly drawn Monte Carlo replication. Without the averaging the DGP distribution is a bit more 'shaky'. Nevertheless, the previous conclusions for Figure 1 about the accuracy and variance of the estimated distributions are valid for Figure 2 as well. When γ is small the kernel estimator without bias-reduction has the smallest bias among the estimators, while the bias-reduced kernel estimator encounters smallest bias when γ is large. And overall the kernel estimators capture the spread better than the traditional FE estimators, albeit underestimates the variation when γ is small.[28]

We summarize the result part by concluding that the expectations from the theoretical small sample properties of the kernel estimator are supported by the simulations. Although it is not doable to make the comparison based on the global conditional MSE-criterion of Theorem 2. Kernel estimation enables both reduction in bias and ASE ('average square error'), on average, compared to traditional FE estimation.

We conjecture that bias-reduction methods used for the FE estimator also can be applied on the kernel estimator due to the similar asymptotic properties as T grows. The simulations shows that if there is not too much random error, bootstrap bias-reduction of the kernel estimated inefficiencies performs very well.

In the next section a small empirical example is presented. We take the results from the simulations into account when selecting an appropriate estimator of inefficiency.

8 Empirical Example: Indonesian rice farmers

In this section Indonesian rice farmer data is analyzed. The Indonesian Ministry of Agriculture surveyed the data from six villages in West Java (Erwidodo 1990). This is a balanced panel of 171 rice farmers over six growing seasons (three wet and three dry). Output is measured in kilograms of rice produced, and inputs are seed (kg), urea (kg), trisodium phosphate (kg), labor (hours)

[27] The distributions are constructed for $T = 5$ and if it is increased the result will be similar to increasing γ.

[28] One Monte Carlo replication is not much to draw conclusions from, however, we have made similar conclusions from all replications we have looked at. Results for several single replications can be provided on request.

and land (hectares). We assume the commonly used Cobb-Douglas production function. Thus we are using log-transformed inputs and output.

We first estimate the production function coefficients and perform a Hausman test for random effects. We also add an estimate of $\gamma = \frac{\sigma_u^2}{\sigma_u^2+\sigma_\nu^2}$ with percentile bootstrapped 95 % confidence interval.[29] These results are supplied in Table 7. All coefficients are significant and the sum indicates constant returns to scale. The Hausman test rejects the null, i.e. rejects the hypothesis of random effects. Thus, there is statistical evidence that random effects estimators are inappropriate to use on this data.[30] The estimate of the measure of the influnce of random error is small: $\hat{\gamma} = 0.134$ and the upper limit of the confidence interval is 0.218.

Hence, there is strong evidence that random error is influential. Based on this we select the kernel estimator with bandwidth $\hat{\lambda}(\bar{\alpha})$. We do not employ bias reduction since this is a case for which bias reduction of the kernel estimator likely leads to underestimation of inefficiency.

Variable	Coefficients	P-values
Seed	0.12	0.0001
Urea	0.10	< 0.0001
Tri. phos.	0.10	< 0.0001
Labor	0.26	< 0.0001
Land	0.44	< 0.0001
Hausman test	χ_5^2=26.7	< 0.0001
$\gamma = \frac{\sigma_u^2}{\sigma_u^2+\sigma_\nu^2}$	$\hat{\gamma} = 0.134$	95 % CI $[0.067, 0.218]$

Table 7: The within estimation results of the production function

	FE	Boot. FE	Kernel-FE
Mean	0.60	0.53	0.19
Std. dev.	0.25	0.12	0.25
Min.	0.00	-0.07	0.00
1st Qu.	0.49	0.42	0.15
Median	0.61	0.55	0.19
3rd Qu.	0.70	0.64	0.22
Max.	1.03	0.97	0.32

Table 8: Summary Statistics for estimated inefficiencies u_i^*

[29] The estimator of γ is consistent and asymptotic normal as $N \to \infty$ this is shown by Wikström (2012) in Article III.

[30] We think this test should be regarded as strictly 'statistic' in contrast to 'economic'. Based on economic reasoning there is little support for random effects, on the contrary, it is difficult to imagine that random effects makes sense for any producers.

In Table 8 summary statistics of the kernel estimator of inefficiency u_i^* ('Kernel-FE') are provided along with summary statistics of the traditional FE estimator, ordinary ('FE') and bias-reduced ('Boot. FE'). Average inefficiency is considerably higher for the two traditional FE estimators. The bootstrapped bias-reduced a little bit less, 0.53 instead of 0.6, but still the difference compared to the kernel estimates, 0.19, is considerable. The spreads of the traditional FE estimators are also larger that of the kernel estimates. Thus, the results indicate exactly what to expect from traditional FE estimation when N is large, T and γ are small, i.e. upward bias and too large spread of the distribution of the inefficiencies.

In Figure 3 kernel densities are plotted for the estimated inefficiencies. As concluded from the Monte Carlo simulations, neither one of the estimator hardly captures the correct shape of the distribution given the conditions of the data, however, the kernel estimator likely give a better indication of the spread and of average inefficiency in this case.

As a tool for policymakers the kernel estimator gives a quite different picture of the inefficiency of the Indonesian rice farmers. The situation probably is not as bad as the traditional FE estimators indicate.

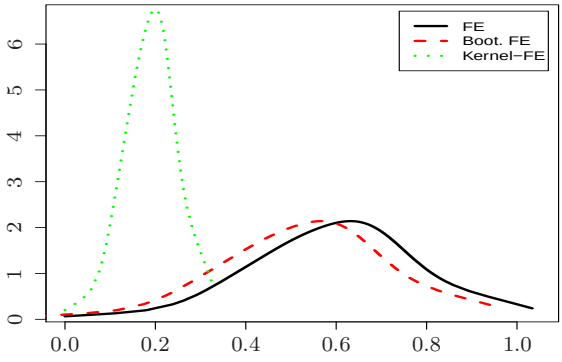

Fig. 3: Kernel densities of estimated inefficiency distributions

9 Conclusion

The FE estimator is theoretically attractive in the sense that we do not need to impose independence between the inputs and the technical inefficiencies. Furthermore, no explicit assumption is needed on the distribution of inefficiencies.

However, as Wang and Schmidt (2009) conclude, the FE estimator is seriously upward biased and only seems to be a realistic choice when N is small and/or T is large. In most microeconomic data sets (we would say the vast majority), this condition is not fulfilled. On the contrary, T is often small while

N can be large. Much effort has been devoted to reducing the bias of the intercept, $\hat{\alpha}$, by using bootstrap or jackknife estimators. However, the FE estimator of the inefficiencies also tends to have a large variance when the random error is influential. And the bias-reduction estimators inherit the same problem.

In this study, we do not primarily focus on reducing the bias in inefficiency as such. Instead, we aim to show that it is possible to estimate firm effects more efficiently, in global conditional MSE terms, without imposing restrictive assumptions.

We prove efficiency, in terms of global conditional MSE criterions, for the kernel estimators of both the firm effects and the inefficiencies compared to the traditional FE estimators. Unfortunately there is no feasible estimator of the optimal bandwidth for estimating inefficiency.

Nevertheless, bandwidths used for estimating the firm effects can also be used for the inefficiencies. We also derive an estimator which seems to work a bit better than the bandwidths designed for the firm effects, when estimating inefficiency.

Monte Carlo simulations support the theoretical results. Firm effects are estimated more efficiently and both ASE (average squared errors) and bias is reduced, on average, for the inefficiencies.

The bias-variance tradeoff of kernel estimation results in flexibility which especially give large small sample gains compared to the FE estimator in situations with much influence of random error.

Due to similarities of the asymptotic properties of the traditional FE estimator and the proposed kernel FE estimator, as T grows, we also conjecture bias-reduction methods, designed for the former estimator, can be applied on the latter estimator. Evidence for this is given by the Monte Carlo Simulations. When there is not too much random error the bootstrap bias-correction works very well for the kernel estimator. Otherwise the kernel estimator without bias-correction works well.

Only time-constant inefficiency is considered, however, the proposed estimation approach could easily be extended to time-varying inefficiencies. We believe kernel estimation is applicable to more or less all estimators which are based on averaging residuals for each separate firm, for which the traditional FE estimator is only one example. This type of sample splitting has been shown to be inefficient compared to kernel estimation, in terms of MSE-criterions, in this study as well as in several previous studies on cell-probability estimation.

The findings in this study runs against the pessimistic conclusion that FE type of estimators of technical (in-) efficiency are only competitive in cases when N is small and T relatively large.

We think the proposed estimation methodology could help researcher and policymaker to obtain better approximations of technical inefficiency without making the random effects assumption and strong distributional assumptions.

Acknowledgements We wish to acknowledge Jeffrey Racine and Yves Surry for their helpful comments. We would also like to thank Helena Hansson for helpful and inspiring discussions. Any errors are, of course, our own.

References

Brown P, Rundell P (1985) Kernel estimates for categorical data. Technometrics 27:293–299
Erwidodo (1990) Panel data analysis on farm-level efficiency, input demand and output supply of rice farming in west Java, Indonesia. PhD thesis, Michigan State University
Greene WH (2008) Econometric Analysis, sixth edn. Pearson Prentice Hall, Upper Saddle River, New Jersey
Hall P (1981) On Nonparametric Multivariate Binary Discrimination. Biometrika 1:287–294
Hall P, Hrdle W, Simar L (1995) Iterated Bootstrap with Applications to Frontier Models. Journal of Productivity Analysis 6:63–76
Hayfield T, Racine JS (2008) Nonparametric econometrics: The np package. Journal of Statistical Software 27(5), URL http://www.jstatsoft.org/v27/i05/
Henderson D, Simar L (2005) A fully nonparametric stochastic frontier model for panel data. Unpublished manuscript
Hoerl A, Kennard R (1970) Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12:55–67
Hurvich CM, Simonoff JS, Tsai CL (1998) Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. Journal of the Royal Statistical Society Series B (Statistical Methodology) 60(2):271–293
Kim M, Kim Y, Schmidt P (2007) On the accuracy of bootstrap confidence intervals for efficiency levels in stochastic frontier models with panel data. Journal of Productivity Analysis 28:165–181
Li Q, Racine JS (2007) Nonparametric Econometrics. Princeton University Press
Li Q, Racine JS, Wooldridge JM (2009) Efficient estimation of average treatment effects with mixed categorical and continuous data. Journal of Business & Economic Statistics 27(2):206–223
Ouyang D, Li Q, Racine J (2009) Nonparametric Estimation of Regression Functions with Discrete Regressors. Econometric Theory 25:1–42
Park B, Simar L (1994) Efficient Semiparametric Estimation in a Stochastic Frontier Model. Journal of the American Statistical Association 89:929–936
Park B, Sickles R, Simar L (1998) Stochastic panel frontiers: A semiparametric approach. Journal of Econometrics 84(2):273–301, DOI 10.1016/S0304-4076(97)00087-0
Park B, Sickles R, Simar L (2003) Semiparametric-efficient estimation of AR(1) panel data models. Journal of Econometrics 117(2):279–309, DOI 10.1016/S0304-4076(03)00149-0
Park B, Sickles R, Simar L (2007) Semiparametric efficient estimation of dynamic panel data models. Journal of Econometrics 136(1):281–301, DOI 10.1016/j.jeconom.2006.03.004
Satchachai P, Schmidt P (2010) Estimates of technical inefficiency in stochastic frontier models with panel data: generalized panel jackknife estimation. Journal of Productivity Analysis 34:83–97
Schmidt P, Sickles R (1984) Production frontiers and panel data. Journal of Business & Economic Statistics 2:367–374
Wang W, Schmidt P (2009) On the distribution of estimated technical efficiency in stochastic frontier models. Journal of Econometrics 148:36–45
Wikström D (2012) The Fixed Effects Estimator of Technical Efficiency. PhD thesis, Swedish University of Agricultural Sciencies
Wooldridge JM (2010) Econometric Analysis of Cross Section and Panel Data, 2nd edn. MIT Press

Appendix I

In this appendix the first order derivatives of the square conditional bias and the conditional variance of the global conditional MSE criterion in (11) are derived. The second order derivative of the criterion is also provided which is used to determine that the bandwidth given in (11) is the minimizer. But first we will prove two lemmas concerning the two sums:

$$\sum_i^N \bar{x}_i' V(\hat{\beta}|X)(\bar{\bar{x}}_{-i} - \bar{x}_i)$$
$$\sum_i^N (\bar{\bar{x}}_{-i} - \bar{x}_i)' V(\hat{\beta}|X)\bar{\bar{x}}_{-i}$$

which are useful both to show that the optimal bandwidth in fact is the minimum and also to conclude that the same bandwidth has the appropriate support, i.e. $\lambda^* \in (0, 1]$. To shorten the proof of the first lemma we first note, without any explicit proof, the following:

$$\sum_i^N \bar{\bar{x}}_{-i}' V(\hat{\beta}|X)\bar{\bar{x}}_{-i} = \frac{N^2}{N-1}\bar{\bar{x}}' V(\hat{\beta}|X)\bar{\bar{x}} - \frac{\sum_i^N \bar{\bar{x}}_{-i}' V(\hat{\beta}|X)\bar{x}_i}{N-1}, \tag{57}$$

and

$$\sum_i^N \bar{\bar{x}}_{-i}' V(\hat{\beta}|X)\bar{x}_i = \frac{N^2}{N-1}\bar{\bar{x}}' V(\hat{\beta}|X)\bar{\bar{x}} - \frac{\sum_i^N \bar{x}_i' V(\hat{\beta}|X)\bar{x}_i}{N-1}, \tag{58}$$

where $\bar{\bar{x}} = \frac{\sum_i^N \bar{x}_i}{N}$.

Lemma 2

$$(N-1)\sum_i^N (\bar{\bar{x}}_{-i} - \bar{x}_i)' V(\hat{\beta}|X)\bar{\bar{x}}_{-i} = -\sum_i^N \bar{x}_i' V(\hat{\beta}|X)(\bar{\bar{x}}_{-i} - \bar{x}_i) \tag{59}$$

Proof

$$(N-1)\sum_i^N (\bar{\bar{x}}_{-i} - \bar{x}_i)' V(\hat{\beta}|X)\bar{\bar{x}}_{-i} + \sum_i^N \bar{x}_i' V(\hat{\beta}|X)(\bar{\bar{x}}_{-i} - \bar{x}_i) =$$
$$(N-1)\left[\sum_i^N \bar{\bar{x}}_{-i}' V(\hat{\beta}|X)\bar{\bar{x}}_{-i} - \sum_i^N \bar{x}_i' V(\hat{\beta}|X)\bar{\bar{x}}_{-i}\right] +$$
$$\sum_i^N \bar{x}_i' V(\hat{\beta}|X)\bar{\bar{x}}_{-i} - \sum_i^N \bar{x}_i' V(\hat{\beta}|X)\bar{x}_i =$$

$$= (N-1)\sum_{i}^{N} \bar{\bar{x}}'_{-i} V(\hat{\beta}|X) \bar{\bar{x}}_{-i} -$$

$$(N-2)\sum_{i}^{N} \bar{x}'_i V(\hat{\beta}|X) \bar{\bar{x}}_{-i} - \sum_{i}^{N} \bar{x}'_i V(\hat{\beta}|X) \bar{x}_i = \qquad \text{[Use equation (57)]}$$

$$N^2 \bar{\bar{x}}' V(\hat{\beta}|X) \bar{\bar{x}} - (N-1)\sum_{i}^{N} \bar{x}'_i V(\hat{\beta}|X) \bar{\bar{x}}_{-i} -$$

$$\sum_{i}^{N} \bar{x}'_i V(\hat{\beta}|X) \bar{x}_i =$$

[Use equation (58)]

$$N^2 \bar{\bar{x}}' V(\hat{\beta}|X) \bar{\bar{x}} - N^2 \bar{\bar{x}}' V(\hat{\beta}|X) \bar{\bar{x}} +$$

$$\sum_{i}^{N} \bar{x}'_i V(\hat{\beta}|X) \bar{x}_i - \sum_{i}^{N} \bar{x}'_i V(\hat{\beta}|X) \bar{x}_i = 0 \qquad \text{Q.E.D.}$$

Lemma 3

$$-\sum_{i}^{N} \bar{x}'_i V(\hat{\beta}|X)(\bar{\bar{x}}_{-i} - \bar{x}_i) \geq 0 \tag{60}$$

Proof

$$\sum_{i}^{N} (\bar{\bar{x}}_{-i} - \bar{x}_i)' V(\hat{\beta}|X)(\bar{\bar{x}}_{-i} - \bar{x}_i) \geq 0 \quad \Big[\text{Since } V(\hat{\beta}|X) \text{ is a covariance matrix}\Big]$$

$$\Leftrightarrow$$

$$\sum_{i}^{N} (\bar{\bar{x}}_{-i} - \bar{x}_i)' V(\hat{\beta}|X) \bar{\bar{x}}_{-i} -$$

$$\sum_{i}^{N} (\bar{\bar{x}}_{-i} - \bar{x}_i)' V(\hat{\beta}|X) \bar{x}_i \geq 0$$

$$\Leftrightarrow$$

$$(N-1)\sum_{i}^{N} (\bar{\bar{x}}_{-i} - \bar{x}_i)' V(\hat{\beta}|X) \bar{\bar{x}}_{-i} -$$

$$(N-1)\sum_{i}^{N} (\bar{\bar{x}}_{-i} - \bar{x}_i)' V(\hat{\beta}|X) \bar{x}_i \geq 0$$

$$\Leftrightarrow$$

$$-N\sum_{i}^{N}(\bar{\bar{x}}_{-i}-\bar{x}_i)'V(\hat{\beta}|X)\bar{x}_i \geq 0 \qquad \text{[By Lemma 2]}$$

$$\Leftrightarrow$$

$$-\sum_{i}^{N}\bar{x}_i'V(\hat{\beta}|X)(\bar{\bar{x}}_{-i}-\bar{x}_i) \geq 0 \qquad \text{Q.E.D.}$$

Results for the proof of Theorem 1

The first derivative of the squared conditional bias is as follows:

$$\frac{\partial\sum_i^N Bias\left(\tilde{\alpha}_i\middle|X\right)^2}{\partial\lambda}=\frac{2N^2\lambda\sum_i^N\left(\bar{\alpha}-\alpha_i\right)^2}{\left[1+(N-1)\lambda\right]^3} \tag{61}$$

while the derivative for the conditional variance is:

$$\begin{aligned}&\frac{\partial\sum_i^N Var\left(\tilde{\alpha}_i\middle|X\right)}{\partial\lambda}= \\ &\frac{2(N-1)\left[\sum_i^N\bar{x}_i'V(\hat{\beta}|X)(\bar{\bar{x}}_{-i}-\bar{x}_i)-N\frac{\sigma_\nu^2}{T}\right]}{\left[1+(N-1)\lambda\right]^3}+ \\ &\frac{2(N-1)\left[N\frac{\sigma_\nu^2}{T}+(N-1)\sum_i^N(\bar{\bar{x}}_{-i}-\bar{x}_i)'V(\hat{\beta}|X)\bar{\bar{x}}_{-i}\right]\lambda}{\left[1+(N-1)\lambda\right]^3}.\end{aligned} \tag{62}$$

Since the first order derivative of the squared biased is zero when $\lambda=0$, the condition for the conditional MSE efficiency of Theorem 1 is given by the following derivation:

$$\begin{aligned}&\left.\frac{\partial\sum_i^N Var\left(\tilde{\alpha}_i\middle|X\right)}{\partial\lambda}\right|_{\lambda=0}= \\ &2(N-1)\left[\sum_i^N\bar{x}_i'V(\hat{\beta}|X)(\bar{\bar{x}}_{-i}-\bar{x}_i)-N\frac{\sigma_\nu^2}{T}\right]<0 \\ &\Leftrightarrow N\frac{\sigma_\nu^2}{T}-\sum_i^N\bar{x}_i'V(\hat{\beta}|X)(\bar{\bar{x}}_{-i}-\bar{x}_i)>0.\end{aligned} \tag{63}$$

This condition is always fulfilled when $N\frac{\sigma_\nu^2}{T}>0$ and $-\sum_i^N\bar{x}_i'V(\hat{\beta}|X)(\bar{\bar{x}}_{-i}-\bar{x}_i)\geq 0$. The first inequality is fulfilled by assumption and the second one by Lemma 3.

Determine the optimal bandwidth λ^*

The minimizer $\lambda^* = arg\min_\lambda \sum_i^N MSE\left(\tilde{\alpha}_i \middle| X\right)$ is obtained from the following first order condition:

$$\begin{aligned}
&\frac{\partial \sum_i^N MSE\left(\tilde{\alpha}_i \middle| X\right)}{\partial \lambda} = \\
&\frac{\partial \sum_i^N Bias\left(\tilde{\alpha}_i \middle| X\right)^2}{\partial \lambda} + \frac{\partial \sum_i^N Var\left(\tilde{\alpha}_i \middle| X\right)}{\partial \lambda} = 0 \\
&\Leftrightarrow \\
&\frac{\left[\sum_i^N \bar{x}_i' V(\hat{\beta}|X)(\bar{\bar{x}}_{-i} - \bar{x}_i) - N\frac{\sigma_\nu^2}{T}\right]}{\left[1 + (N-1)\lambda\right]^3} + \\
&\frac{\left[N\frac{\sigma_\nu^2}{T} + N^2\frac{\sum_i^N (\bar{\alpha} - \alpha_i)^2}{N-1} + (N-1)\sum_i^N (\bar{\bar{x}}_{-i} - \bar{x}_i)' V(\hat{\beta}|X)\bar{\bar{x}}_{-i}\right]\lambda}{\left[1 + (N-1)\lambda\right]^3} = 0 \\
&\Leftrightarrow \\
&\lambda = \frac{N\frac{\sigma_\nu^2}{T} - \sum_i^N \bar{x}_i' V(\hat{\beta}|X)(\bar{\bar{x}}_{-i} - \bar{x}_i)}{N\frac{\sigma_\nu^2}{T} + N^2\frac{\sum_i^N (\bar{\alpha} - \alpha_i)^2}{N-1} + (N-1)\sum_i^N (\bar{\bar{x}}_{-i} - \bar{x}_i)' V(\hat{\beta}|X)\bar{\bar{x}}_{-i}} = \lambda^*.
\end{aligned} \tag{64}$$

Thus, there is one critical point given by the first order derivative. And since the derivative is well-defined for all $\lambda \in [0,1]$ a sufficient second order condition is that the second order derivative is strictly positive at λ^* to make it the global minimizer on $\lambda \in [0,1]$. Thus,

$$\begin{aligned}
&\frac{\partial^2 \sum_i^N MSE\left(\tilde{\alpha}_i \middle| X\right)}{\partial \lambda^2} > 0 \\
&\Leftrightarrow \\
&2(N-1)\left\{\left[N\frac{\sigma_\nu^2}{T} + N^2 s_\alpha^2 + (N-1)\sum_i^N (\bar{\bar{x}}_{-i} - \bar{x}_i)' V(\hat{\beta}|X)\bar{\bar{x}}_{-i}\right] \times \right. \\
&\left. (1 - 2(N-1)\lambda) - 3(N-1)\sum_i^N \bar{x}_i' V(\hat{\beta}|X)(\bar{\bar{x}}_{-i} - \bar{x}_i) + 3(N-1)N\frac{\sigma_\nu^2}{T}\right\} \times \\
&\frac{1}{\left[1 + (N-1)\lambda\right]^4} > 0 \\
&\Leftrightarrow
\end{aligned} \tag{65}$$

$$
\begin{aligned}
&- 4(N-1)^2\left[N\frac{\sigma_\nu^2}{T}+N^2s_\alpha^2+(N-1)\sum_i^N(\bar{\bar{x}}_{-i}-\bar{x}_i)'V(\hat{\beta}|X)\bar{\bar{x}}_{-i}\right]\lambda > \\
&- 2(N-1)\left[N\frac{\sigma_\nu^2}{T}+N^2s_\alpha^2+(N-1)\sum_i^N(\bar{\bar{x}}_{-i}-\bar{x}_i)'V(\hat{\beta}|X)\bar{\bar{x}}_{-i}\right]+ \\
&6(N-1)^2\sum_i^N\bar{x}_i^{'}V(\hat{\beta}|X)(\bar{\bar{x}}_{-i}-\bar{x}_i)-6(N-1)^2N\frac{\sigma_\nu^2}{T} \\
&\Leftrightarrow \quad \text{[Given Lemma 3]} \\
&\lambda < \frac{N^2s_\alpha^2+N\frac{\sigma_\nu^2}{T}-\sum_i^N\bar{x}_i'V(\hat{\beta}|X)(\bar{\bar{x}}_{-i}-\bar{x}_i)}{2(N-1)\left[N^2s_\alpha^2+N\frac{\sigma_\nu^2}{T}-(N-1)\sum_i^N(\bar{\bar{x}}_{-i}-\bar{x}_i)'V(\hat{\beta}|X)\bar{\bar{x}}_{-i}\right]}- \\
&\frac{3(N-1)\left[\sum_i^N\bar{x}_i'V(\hat{\beta}|X)(\bar{\bar{x}}_{-i}-\bar{x}_i)-N\frac{\sigma_\nu^2}{T}\right]}{2(N-1)\left[N^2s_\alpha^2+N\frac{\sigma_\nu^2}{T}+(N-1)\sum_i^N(\bar{\bar{x}}_{-i}-\bar{x}_i)'V(\hat{\beta}|X)\bar{\bar{x}}_{-i}\right]} \\
&\Leftrightarrow \quad \text{[set } \lambda=\lambda^*\text{]} \\
&-\lambda^*\frac{N}{2(N-1)} < \\
&\frac{N^2s_\alpha^2}{2(N-1)\left[N^2s_\alpha^2+N\frac{\sigma_\nu^2}{T}+(N-1)\sum_i^N(\bar{\bar{x}}_{-i}-\bar{x}_i)'V(\hat{\beta}|X)\bar{\bar{x}}_{-i}\right]} \\
&\Leftrightarrow \\
&\lambda^* > -\frac{Ns_\alpha^2}{\left[N^2s_\alpha^2+N\frac{\sigma_\nu^2}{T}+(N-1)\sum_i^N(\bar{\bar{x}}_{-i}-\bar{x}_i)'V(\hat{\beta}|X)\bar{\bar{x}}_{-i}\right]}. \qquad (66)
\end{aligned}
$$

We conclude that λ^* is the global minimum on $\lambda \in [0,1]$, since it is strictly positive, given $N<\infty$, $T<\infty$ and $\sigma_\nu^2>0$, and therefore, satisfies the second order condition in (66).

Appendix II

In this appendix the second order condition for the proof of Theorem 2 is derived.

$$\begin{aligned}
&\frac{\partial^2 \sum_i^N MSE\left(\tilde{u}_i \middle| X\right)}{\partial \lambda^2} > 0 \qquad (67)\\
&\Leftrightarrow\\
&\frac{-2(N-1)A\lambda + A + 3(N-1)B}{\left[1+(N-1)\lambda\right]^4} > 0\\
&\Leftrightarrow\\
&\lambda < \frac{1}{2(N-1)} + \frac{3}{2}\frac{B}{A}
\end{aligned}$$

where B is the numerator of λ^{**} and the A is the denominator. The second order condition can, therefore, be written as:

$$\lambda < \frac{1}{2(N-1)} + \frac{3}{2}\lambda^{**} \qquad . \qquad (68)$$

This implies that the upper bound of this condition is something larger than λ^{**}. There is also a lower bound which falls out if one puts $\lambda = \lambda^{**}$. Thus, in the point λ^{**} the following should be true for the second order derivative to be larger than zero:

$$\begin{aligned}
&\lambda^{**} < \frac{1}{2(N-1)} + \frac{3}{2}\lambda^{**} \qquad (69)\\
&\Leftrightarrow\\
&\lambda^{**} > -\frac{1}{(N-1)}
\end{aligned}$$

Thus, λ^{**} is the global minimum on $-\frac{1}{(N-1)} < \lambda < \frac{1}{2(N-1)} + \frac{3}{2}\lambda^{**}$ and this is sufficient together with the assumptions of Theorem 2 to make the kernel-FE estimator MSE-efficient in respect to the criterion given by (33) compared to the traditional FE-estimator.

Appendix III

In this appendix the estimators used to compute the estimates of the two bandwidths proposed in this study are provided.

The bandwidths λ^* and $\lambda(\bar{\alpha})^{**}$ are estimated with help of estimators of s_α^2, σ_ν^2 and for the former bandwidth also $V(\hat{\beta}|X)$. These three unknowns are estimated with the following estimators:[31]

$$\hat{\sigma}_u^2 = (N-1)^{-1} \sum_i^N (\hat{\alpha}_i - \hat{\mu}_\alpha)^2 - \hat{\sigma}_\nu^2/T, \tag{70}$$

$$\hat{\sigma}_\nu^2 = \frac{\sum_i \sum_t \hat{\ddot{\nu}}_{it}^2}{N(T-1)}, \tag{71}$$

and

$$\widehat{V(\hat{\beta}|X)} = \hat{\sigma}_\nu^2 \left(\sum_{i=1}^N \sum_{t=1}^T \ddot{x}_{it} \ddot{x}_{it}' \right)^{-1} \tag{72}$$

where the dot-dot notation is defined as: $\ddot{z}_{it} = z_{it} - \bar{z}_i$, and $\hat{\ddot{\nu}}_{it} = \hat{\nu}_{it} - \hat{\bar{\nu}}_i = \ddot{y}_{it} - \ddot{x}_{it}'\hat{\beta}$. Given results in the third article of Wikström (2012) it is verified or straightforward to show the following:

$$\hat{\sigma}_u^2 - s_\alpha^2 = O_p(N^{-1/2}T^{-1/2}), \tag{73}$$

$$\hat{\sigma}_\nu^2 - \sigma_\nu^2 = O_p(N^{-1/2}T^{-1/2}), \tag{74}$$

And for the covariance matrix of $\hat{\beta}$ we have:[32]

$$\widehat{V(\hat{\beta}|X)} = O_p(N^{-1}T^{-1}). \tag{75}$$

The results presented above implies that the estimators of the unknowns in the proposed bandwidths are 'bounded in probability' both as N and T tends to infinity. This assures the following properties of the estimators of the proposed bandwidths:

$$\hat{\lambda}^* = O_p(N^{-1}T^{-1}), \tag{76}$$

$$\hat{\lambda}(\bar{\alpha})^{**} = O_p(N^{-1}T^{-1}), \tag{77}$$

as $N \to \infty$ and/or $T \to \infty$.

Thus, the asymptotic results presented in Section 4 holds for kernel estimation based on both $\hat{\lambda}^*$ and $\hat{\lambda}(\bar{\alpha})^{**}$.

[31] The estimator $\hat{\sigma}_u^2$ was provided by Jeffrey Wooldridge. The second estimator $\hat{\sigma}_\nu^2$ is included in Wooldridge (2010). An estimator of $\sigma_u^2 = \sigma_\alpha^2$ is also included in Wooldridge (2010) but the expression has a typo. In Article III of Wikström (2012) consistency as N grow is shown for both estimators.

[32] See e.g. p. 196 Greene (2008)

Appendix IV

N	T	γ	$\hat{\lambda}^*$	$\hat{\lambda}_{aic}$	$\hat{\lambda}_{cv}$
10	5	0.25	0.9557	0.9529	0.9588
10	5	0.50	0.9812	0.9764	0.9841
10	5	0.75	0.9958	0.9936	0.9963
10	15	0.25	0.9839	0.9826	0.9851
10	15	0.50	0.9974	0.9969	0.9976
10	15	0.75	0.9996	0.9995	0.9996
10	30	0.25	0.9955	0.9951	0.9957
10	30	0.50	0.9994	0.9993	0.9994
10	30	0.75	0.9999	0.9999	0.9999
150	5	0.25	0.9925	0.9802	0.9926
150	5	0.50	0.9983	0.9902	0.9983
150	5	0.75	0.9996	0.9962	0.9996
150	15	0.25	0.9991	0.9983	0.9991
150	15	0.50	0.9998	0.9995	0.9998
150	15	0.75	1.0000	0.9999	1.0000
150	30	0.25	0.9997	0.9996	0.9997
150	30	0.50	1.0000	0.9999	1.0000
150	30	0.75	1.0000	1.0000	1.0000
300	5	0.25	0.9963	0.9841	0.9964
300	5	0.50	0.9992	0.9911	0.9992
300	5	0.75	0.9998	0.9963	0.9998
300	15	0.25	0.9995	0.9987	0.9995
300	15	0.50	0.9999	0.9996	0.9999
300	15	0.75	1.0000	0.9999	1.0000
300	30	0.25	0.9999	0.9998	0.9999
300	30	0.50	1.0000	0.9999	1.0000
300	30	0.75	1.0000	1.0000	1.0000

Table 9: Average ratios of global conditional MSE for estimators of α_i compared to the theoretical minima: $C\left[\tilde{\alpha}_i(\hat{\lambda})\right]$. DGP: M_1

N	T	γ	$\hat{\lambda}^*$	$\hat{\lambda}_{aic}$	$\hat{\lambda}_{cv}$
10	5	0.25	0.9968	0.9938	1.0000
10	5	0.50	0.9971	0.9922	1.0000
10	5	0.75	0.9994	0.9972	1.0000
10	15	0.25	0.9988	0.9974	1.0000
10	15	0.50	0.9997	0.9993	1.0000
10	15	0.75	1.0000	0.9999	1.0000
10	30	0.25	0.9998	0.9994	1.0000
10	30	0.50	1.0000	0.9999	1.0000
10	30	0.75	1.0000	1.0000	1.0000
150	5	0.25	0.9998	0.9875	1.0000
150	5	0.50	1.0000	0.9918	1.0000
150	5	0.75	1.0000	0.9965	1.0000
150	15	0.25	1.0000	0.9992	1.0000
150	15	0.50	1.0000	0.9997	1.0000
150	15	0.75	1.0000	0.9999	1.0000
150	30	0.25	1.0000	0.9999	1.0000
150	30	0.50	1.0000	1.0000	1.0000
150	30	0.75	1.0000	1.0000	1.0000
300	5	0.25	1.0000	0.9877	1.0000
300	5	0.50	1.0000	0.9919	1.0000
300	5	0.75	1.0000	0.9965	1.0000
300	15	0.25	1.0000	0.9992	1.0000
300	15	0.50	1.0000	0.9997	1.0000
300	15	0.75	1.0000	0.9999	1.0000
300	30	0.25	1.0000	0.9999	1.0000
300	30	0.50	1.0000	1.0000	1.0000
300	30	0.75	1.0000	1.0000	1.0000

Table 10: Ratios of $C\left[\tilde{\alpha}_i(\hat{\lambda})\right]$. Value one indicates best – estimated – fit. DGP: M_1

N	T	γ	$\hat{\lambda}^*$	$\hat{\lambda}_{aic}$	$\hat{\lambda}_{cv}$
10	5	0.25	0.8743	0.8602	0.8379
10	5	0.50	0.9391	0.9390	0.9314
10	5	0.75	0.9839	0.9840	0.9785
10	15	0.25	0.9673	0.9687	0.9705
10	15	0.50	0.9941	0.9944	0.9944
10	15	0.75	0.9990	0.9990	0.9988
10	30	0.25	0.9910	0.9913	0.9918
10	30	0.50	0.9987	0.9987	0.9987
10	30	0.75	0.9999	0.9999	0.9999
150	5	0.25	0.9871	0.9782	0.9871
150	5	0.50	0.9972	0.9906	0.9971
150	5	0.75	0.9995	0.9967	0.9994
150	15	0.25	0.9985	0.9977	0.9985
150	15	0.50	0.9998	0.9995	0.9998
150	15	0.75	1.0000	0.9999	1.0000
150	30	0.25	0.9997	0.9996	0.9997
150	30	0.50	0.9999	0.9999	0.9999
150	30	0.75	1.0000	1.0000	1.0000
300	5	0.25	0.9934	0.9821	0.9935
300	5	0.50	0.9986	0.9910	0.9986
300	5	0.75	0.9997	0.9964	0.9997
300	15	0.25	0.9993	0.9985	0.9993
300	15	0.50	0.9999	0.9996	0.9999
300	15	0.75	1.0000	0.9999	1.0000
300	30	0.25	0.9998	0.9997	0.9998
300	30	0.50	1.0000	0.9999	1.0000
300	30	0.75	1.0000	1.0000	1.0000

Table 11: Average ratios of global conditional MSE for estimators of α_i compared to the theoretical minima: $C\left[\tilde{\alpha}_i(\hat{\lambda})\right]$. DGP: M_2

N	T	γ	$\hat{\lambda}^*$	$\hat{\lambda}_{aic}$	$\hat{\lambda}_{cv}$
10	5	0.25	1.0000	0.9839	0.9583
10	5	0.50	1.0000	0.9998	0.9917
10	5	0.75	0.9999	1.0000	0.9944
10	15	0.25	0.9967	0.9981	1.0000
10	15	0.50	0.9996	1.0000	1.0000
10	15	0.75	1.0000	1.0000	0.9998
10	30	0.25	0.9991	0.9995	1.0000
10	30	0.50	1.0000	1.0000	1.0000
10	30	0.75	1.0000	1.0000	1.0000
150	5	0.25	1.0000	0.9909	0.9999
150	5	0.50	1.0000	0.9934	0.9999
150	5	0.75	1.0000	0.9972	0.9999
150	15	0.25	1.0000	0.9992	1.0000
150	15	0.50	1.0000	0.9997	1.0000
150	15	0.75	1.0000	0.9999	1.0000
150	30	0.25	1.0000	0.9999	1.0000
150	30	0.50	1.0000	1.0000	1.0000
150	30	0.75	1.0000	1.0000	1.0000
300	5	0.25	0.9999	0.9885	1.0000
300	5	0.50	1.0000	0.9923	1.0000
300	5	0.75	1.0000	0.9967	1.0000
300	15	0.25	1.0000	0.9992	1.0000
300	15	0.50	1.0000	0.9997	1.0000
300	15	0.75	1.0000	0.9999	1.0000
300	30	0.25	1.0000	0.9999	1.0000
300	30	0.50	1.0000	1.0000	1.0000
300	30	0.75	1.0000	1.0000	1.0000

Table 12: Ratios of $C\left[\tilde{\alpha}_i(\hat{\lambda})\right]$. Value one indicates best – estimated – fit. DGP: M_2

N	T	p	$\hat{\lambda}^*$	$\hat{\lambda}_{aic}$	$\hat{\lambda}_{cv}$	$\hat{\lambda}(\bar{\alpha})^{**}$
10	5	0.25	0.5863	0.5853	0.5965	0.5611
10	5	0.50	0.6522	0.6488	0.6666	0.6017
10	5	0.75	0.7078	0.7052	0.7178	0.6600
10	15	0.25	0.6613	0.6608	0.6632	0.6149
10	15	0.50	0.7132	0.7129	0.7152	0.6729
10	15	0.75	0.7523	0.7523	0.7542	0.7244
10	30	0.25	0.6959	0.6956	0.6959	0.6598
10	30	0.50	0.7342	0.7341	0.7345	0.7126
10	30	0.75	0.7653	0.7654	0.7658	0.7515
150	5	0.25	0.7202	0.8127	0.7103	0.8571
150	5	0.50	0.6313	0.7131	0.6248	0.8733
150	5	0.75	0.6268	0.6755	0.6232	0.8050
150	15	0.25	0.6293	0.6537	0.6266	0.8823
150	15	0.50	0.6246	0.6385	0.6231	0.8081
150	15	0.75	0.6655	0.6730	0.6648	0.7698
150	30	0.25	0.6187	0.6268	0.6175	0.8358
150	30	0.50	0.6515	0.6559	0.6509	0.7824
150	30	0.75	0.7007	0.7030	0.7004	0.7708
300	5	0.25	0.6335	0.7657	0.6273	0.9142
300	5	0.50	0.5302	0.6197	0.5269	0.8460
300	5	0.75	0.5329	0.5830	0.5312	0.7342
300	15	0.25	0.5366	0.5617	0.5353	0.8500
300	15	0.50	0.5349	0.5488	0.5342	0.7354
300	15	0.75	0.5762	0.5841	0.5758	0.6923
300	30	0.25	0.5216	0.5297	0.5210	0.7678
300	30	0.50	0.5520	0.5566	0.5517	0.6955
300	30	0.75	0.6016	0.6043	0.6015	0.6835

Table 13: Average ratios of ASE for estimators of u_i^* compared to the theoretical minima: $C\left[\tilde{u}_i(\hat{\lambda})\right]$. DGP: M_3

N	T	p	$\hat{\lambda}^*$	$\hat{\lambda}_{aic}$	$\hat{\lambda}_{cv}$	$\hat{\lambda}(\bar{\alpha})^{**}$
10	5	0.25	0.9829	0.9813	1.0000	0.9408
10	5	0.50	0.9784	0.9734	1.0000	0.9026
10	5	0.75	0.9861	0.9824	1.0000	0.9194
10	15	0.25	0.9971	0.9963	1.0000	0.9271
10	15	0.50	0.9971	0.9968	1.0000	0.9409
10	15	0.75	0.9975	0.9974	1.0000	0.9604
10	30	0.25	1.0000	0.9995	1.0000	0.9480
10	30	0.50	0.9996	0.9995	1.0000	0.9702
10	30	0.75	0.9993	0.9995	1.0000	0.9814
150	5	0.25	0.8402	0.9482	0.8287	1.0000
150	5	0.50	0.7229	0.8165	0.7154	1.0000
150	5	0.75	0.7786	0.8391	0.7742	1.0000
150	15	0.25	0.7132	0.7409	0.7102	1.0000
150	15	0.50	0.7728	0.7901	0.7711	1.0000
150	15	0.75	0.8646	0.8743	0.8636	1.0000
150	30	0.25	0.7402	0.7499	0.7388	1.0000
150	30	0.50	0.8327	0.8384	0.8319	1.0000
150	30	0.75	0.9091	0.9121	0.9086	1.0000
300	5	0.25	0.6929	0.8375	0.6862	1.0000
300	5	0.50	0.6268	0.7326	0.6229	1.0000
300	5	0.75	0.7258	0.7941	0.7235	1.0000
300	15	0.25	0.6313	0.6608	0.6298	1.0000
300	15	0.50	0.7273	0.7462	0.7264	1.0000
300	15	0.75	0.8322	0.8438	0.8317	1.0000
300	30	0.25	0.6793	0.6899	0.6785	1.0000
300	30	0.50	0.7937	0.8003	0.7932	1.0000
300	30	0.75	0.8802	0.8841	0.8799	1.0000

Table 14: Ratios of $C\left[\tilde{u}_i(\hat{\lambda})\right]$. Value one indicates best – estimated – fit. DGP: M_3