December 2020

NONPARAMETRIC IDENTIFICATION AND ESTIMATION OF STOCHASTIC FRONTIER MODELS

Jun Cai
Syracuse University

Abstract

This dissertation studies nonparametric identification and estimation of stochastic frontier models. It is composed of three chapters. The first chapter investigates the identification and estimation of a cross sectional stochastic frontier model with Laplacian errors and unknown variance, which is built on a nonparametric density deconvolution strategy. Chapter two studies a zero-inefficiency stochastic frontier model utilizing a penalized sieve estimator, which allows flexible function forms and arbitrary distributions of inefficiency. The third chapter explores identification and estimation of a nonparametric panel stochastic frontier model based on Kotlarski's Lemma and moments derived from conditional characteristic functions.

NONPARAMETRIC IDENTIFICATION AND ESTIMATION OF STOCHASTIC FRONTIER MODELS

by

Jun Cai

M.S., Shanghai University of Finance and Economics, 2015

B.S., Shanghai Jiao Tong University, 2011

Dissertation

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in *Economics*.

Syracuse University

December 2020

Acknowledgements

First of all, I would like to thank my supervisor, Dr. William C. Horrace, for his invaluable guidance and dedicated support through each stage of my research. His immense knowledge, motivation and patience have inspired and empowered me to excel in research and what's more important, to be a better person.

I would like to acknowledge Dr. Alfonso Flores-Lagunes and Dr. Yoonseok Lee for their insightful suggestions during the development of my dissertation and my research partner, Dr. Christopher F. Parmeter, for being instrumental in defining the path of my research. I also want to express my gratitude to faculty members Dr. Hugo Jales, Dr. Jan Ondrich Dr. Yilin Hou and colleague Yoon Jung Choi in Center for Policy Research, who inspired me deeply with their rigorous academic pursuits and professionalism.

At last, I cannot forget to thank my family who supported me unconditionally and consistently in the past five years and without whom I would not have been able to complete this research, and without whom I would not have made it through my doctorate degree!

Contents

List of Figures

List of Tables

Chapter 1

Density Deconvolution with Laplace Errors and Unknown Variance

Jun Cai[1], William C. Horrace[2], Christopher F. Parmeter[3]

1.1 Introduction

Deconvolution uses kernel techniques to estimate the density (the *target density*) of a random variable (u) in the presence of an independent and additive noise term (v). Most deconvolution estimators are for a random cross-section of observations from a noisy random variable (i.e., $\varepsilon = u + v$), where the noise distribution (f_v) is known. If we know f_v and (hence) its characteristic function, then under regularity conditions we can calculate the empirical characteristic function of ε and use the Fourier inversion formula to consistently point estimate f_u. Fan (1991) shows that convergence rates for kernel deconvolution estimators depend on the smoothness of the noise distribution, where smoothness is characterized by the tail behavior of the associated characteristic function. Specifically, if v is from the super-smooth family (e.g., normal or Cauchy), the fastest convergence rate is logarithmic in the sample

[1] Department of Economics, Syracuse University, jcai106@syr.edu.

[2] Department of Economics, Syracuse University, whorrace@syr.edu.

[3] Department of Economics, University of Miami, cparmeter@bus.miami.edu.

size (n), and if noise is from the ordinary-smooth family (e.g. Laplace or gamma), the fastest rate is polynomial in n.[4]

However, in applications (like the stochastic frontier model) it may be more practical to assume that the noise distribution is known *up to its variance*. Hence, Meister (2006) develops a semi-uniformly consistent estimator of the target density and a "truncation" device estimator for the unknown noise variance, when the noise density is super-smooth (e.g., normal) and the target density is ordinary-smooth (e.g., gamma), which bounds the decay of the tails of its characteristic function.[5] Horrace and Parmeter (2011) adapt the estimator of Meister (2006) to the stochastic frontier model (Aigner et al., 1977), where the noisy random variable (ε) is appended to a linear regression model, v is normally distributed, and u is ordinary-smooth and non-negative.[6] That is, for a linear production function with normally distributed (super-smooth) noise (v), we may estimate the density of technical inefficiency (u), if it belongs to the ordinary-smooth family (e.g., exponential or gamma). Unfortunately, the convergence results of Fan (1991) still apply: both the Meister (2006) and Horrace and Parmeter (2011) estimators converge at logarithmic rates. Therefore, it natural to consider a version of Horrace and Parmeter (2011) where noise is Laplace (ordinary-smooth), so as to achieve polynomial convergence rates for estimators of the density of technical inefficiency. This is the goal of this paper.

Laplace noise is not unprecedented in the literature. Horrace and Parmeter (2018) develop a parametric stochastic frontier model with Laplace noise which possess useful features for ranking and selecting efficient firms.[7] Meister (2004) shows that in a deconvolution problem if the noise distribution is misspecified, it is always better to assume Laplace noise rather

[4] We give a precise definition of smoothness in the sequel. Deconvolution applications for v normal (super-smooth) abound. See Stefanski and Carroll, 1990; Neumann, 1997; Johannes, 2009; Wang and Ye, 2012.

[5] Others are Butucea and Matias (2004) and Butucea, Matias, and Pouet (2008). The Meister (2006) estimator is uniformly consistent relative to the target distributional family but individually relative the noise distributional family. That is, consistency of the estimator does not hold uniformly over all noise distributions.

[6] Horowitz and Markatou (1996) consider deconvolution in the linear regression model for panel data.

[7] Horrace and Parmeter do maximum likelihood estimation of the stochastic frontier model, not deconvolution.

than normal, because normal noise produces infinite risk while Laplace noise produces finite risk. A similar result arises in the simulations of Horrace and Parmeter (2018) who find that the mean squared error (MSE) of the parametric stochastic frontier model is smaller with Laplace noise than with normal noise under misspecification. Errors-in-variable models have recently considered Laplace errors. See Carroll et al. (2006), Koul and Song (2014), Song et al. (2016), Cao (2016) and references therein. Finally, maximum likelihood estimation with Laplace errors produces the least absolute deviations (LAD) estimator, and applications of this method are plentiful in statistics, finance, engineering, and other applied sciences (see Dodge, 1987, 1992, 1997 and Dodge and Falconer, 2002).

Our aim here is to provide a complete account of Laplace kernel deconvolution and to develop a regression-based deconvolution estimator that does not require the variance of the Laplace distribution to be known. We modify the "variance truncation device" of Meister (2006) to bound of the variance of the noise (v) with the variance of the noisy random variable (ε). Target density estimation is drastically improved (in terms of convergence) with Laplace noise and is robust to misspecification of the noise distribution (per Meister, 2004). Moreover, we offer practical guidance and an adaptive procedure for selecting the smoothness parameters which are key to implementation of the proposed techniques (and which will be discussed later). This adaptive procedure is new in the literature and offers sound footing for practical use of these methods. Lastly, we apply the Laplace deconvolution estimator to two restricted versions of the model: a stochastic (cost) frontier model (SFM), where u is restricted non-positive, and a pure deconvolution problem, where the linear regression parameters are restricted to equal zero.

The paper is organized as follows. In Section 1.2 we discuss the basic issues surrounding deconvolution in the regression model and introduce the modified variance truncation device under Laplace errors (noise). Section 1.3 derives large sample properties of the estimator under certain regularity conditions. Two extensions are considered in Section 1.4. Section 1.5 contains a variety of Monte Carlo results demonstrating the finite sample performance of the proposed estimator as well as issues pertaining to robustness of the choice of the Laplace noise. In Section 1.6 we provide two practical applications to illustrate the utility of the proposed methodology. Conclusions are in Section 1.7.

1.2 The Laplace Convolution Problem

Consider the error component model (ECM) in the cross sectional setting:

$$y_j = x_j'\beta + u_j + v_j = x_j'\beta + \varepsilon_j, \quad j = 1, \ldots, n. \tag{1.1}$$

Here j indexes individuals or firms, β is a parameter vector of dimension q to be estimated and exogenous covariates $x \in \mathcal{R}^q$. The ε is a composed error term, u is the target error component, and v is statistical noise. Depending on assumptions on u, the model in (1) can be a cross sectional stochastic frontier model (e.g., $u \sim Exp(\sigma_u^2)$), a linear regression with measurement error (e.g., $y_j = x_j^*\beta + v_j$, where $x_j^* = x_j + e_j$, $u_j = \beta * e_j$), or a pure measurement error model (e.g., $\beta = 0$). A large statistical literature investigates the $\beta = 0$ model with known or partially-known error distribution of v (see Meister, 2009).[8] In this setting, deconvolution is complicated by the fact that only cross sectional data are available. Following the literature (i.e., Fan, 1991; Meister, 2006; Horrace and Parmeter, 2011), we make the following assumptions on the random components of the model and the covariates when present.

Assumption 1. *The x_j, v_j and u_j are pairwise independent for all $j = 1, \ldots, n$.*

Let the probability densities of the error components be $f_v(z)$, $f_u(z)$ and $f_\varepsilon(z)$ with corresponding characteristic functions $h_v(\tau)$, $h_u(\tau)$ and $h_\varepsilon(\tau)$. Based on the independence

[8] Neumann (1997), Johannes (2009), and Wang and Ye (2012) study deconvolution with *fully unknown* error distribution but require either an additional sample of the error or repeated observations, y_{jt}.

between v_j and u_j in Assumption 1,

$$h_\varepsilon(\tau) = h_v(\tau)h_u(\tau). \tag{1.2}$$

We restrict v to the family of Laplace densities with the following assumption.

Assumption 2. *The distribution of v is a member of the Laplace family with zero mean and unknown variance, i.e. $\mathcal{L} = \{Laplace(0, b) : b^2 > 0\}$.*

Hence, the density of v is known up to its variance ($2b^2$), and the characteristic function of v is $h_v(\tau) = (1 + b^2\tau^2)^{-1}$, so that,

$$h_u(\tau) = \frac{h_\varepsilon(\tau)}{h_v(\tau)} = (1 + b^2\tau^2)h_\varepsilon(\tau). \tag{1.3}$$

We restrict u to be ordinary-smooth (Fan, 1991) with the following assumption.

Assumption 3. *Assume u is ordinary-smooth. Namely, u belongs to the family $\mathcal{F}_u = \big\{h_u : C_1|\tau|^{-\delta} \leq |h_u(\tau)| \leq C_2|\tau|^{-\delta}, for \quad |\tau| \geq T > 0\big\}$ where $0 < C_1 < C_2$ and $\delta > 1$, $\delta \neq 2$.*

Assumption 3 dictates tail behavior of the characteristic function of u (smoothness of the density of u), and positive constants C_1, C_2 and δ are *smoothness parameters.* The lower bound, C_1, and upper bound, C_2, ensure the rate of decay of the tails of the characteristic function does not approach zero too rapidly or too slowly and are needed for identification. Constants C_1 and C_2 become irrelevant when T gets large. Practically speaking, we only use the lower bound to define our variance truncation device, so only C_1 is relevant to our estimator. We assume C_1 and δ to be known for now but will relax this in the sequel.[9]

Constant δ is the *smoothness order*, ensuring polynomial tail behavior of the characteristic function, and includes a wide array of nonparametric and analytical families (Horrace and Parmeter, 2011). Common families and their polynomial smoothness orders are tabulated in Table 3.1. For example, the Symmetric Uniform family of distributions has polynomial order

[9] Knowing C_1 and δ does not imply knowing $V(u)$ nor does it uniquely determine the analytic family.

$\delta = 1$, and the Laplace family has $\delta = 2$. We restrict $\delta \neq 2$ so that the target density cannot be Laplace, allowing our estimator to appropriately assign the target and noise distributions. That is, if u and v are both Laplace, we cannot determine which distribution is the target and which is the noise.[10] In the parlance of frontier estimation, when $\delta = 2$ we cannot distinguish the signal from the noise. Letting $\delta = 2$ does not preclude deconvolution *per se*. For example, the deconvolution convolution estimator of Dattner et al. (2011) relies on very general classes of distributions for the target and noise densities that includes the Laplace-Laplace convolution as a special case, and consistent target density estimation is achieved as long as the error variance is known. The restriction in Assumption 3 that $\delta > 1$ does not preclude a nonparamteric family of densities in Table 3.1 that is arbitrarily close a family with $\delta = 1$ like the Uniform or the Exponential (i.e, a Gamma with $k = 1$ in the Table), which have both been employed in Stochastic Frontier Analysis.

Note that Meister (2006) assumes different distributional families for u and v (i.e., ordinary-smooth and super-smooth, respectively) and that simplifies derivation of the convex upper bound of the criterion function in that paper. The intuition is that as n goes to infinity the tail of h_v (normal noise) decays faster than that of h_u. Turning to Table 3.1, we see that the normal distribution has polynomial order $\delta \to \infty$, so the intuition is justified.[11] In the current paper similar intuition applies, but the key here is that the tails of characteristic function of u and v decay at *different* rates with the noise characteristic function decaying at a rate fixed at $\delta = 2$ by design.

Under Assumptions 2 and 3, the Fourier inversion formula returns the density of u,

$$f_u(z) = \frac{1}{2\pi} \int e^{-i\tau z}(1 + b^2\tau^2)h_\varepsilon(\tau)d\tau, \tag{1.4}$$

where $i = \sqrt{-1}$. If noise $v \sim \mathcal{G} = \{N(0, \sigma^2) : \sigma^2 > 0 \text{ unknown}\}$, Meister (2006) shows

[10] We are grateful to an anonymous reviewer who alerted us to this identification issue. It should be noted that the restriction eliminates a broad class of ordinary-smooth distributions, not just the Laplace.

[11] Indeed neither the Normal nor Cauchy families of distribution are ordinary-smooth; they are *super-smooth*. See Fan (1991).

that there is no uniformly consistent estimator of $f_u(z)$. His deconvolution estimator of $f_u(z)$ is semi-uniformly consistent in the sense that for a given density in $\mathcal{G}$ whose variance is bounded, a deconvolution estimator is uniformly consistent but not uniformly consistent over all densities within $\mathcal{G}$. This is the price one pays for not knowing the variance. Here we focus on the Laplace noise case with unknown variance. As we shall demonstrate, with Laplace noise one still pays a price for not knowing the variance, but the cost is not as high as in the case with normally distributed noise.

Since h_ε is unknown, we may rely on the empirical characteristic function to recover the density of u based on equation (1.4),

$$\hat{h}_\varepsilon(\tau) = \left| \frac{1}{n} \sum_{j=1}^{n} e^{i\tau\varepsilon_j} \right|. \tag{1.5}$$

As mentioned previously, ε_j is unobserved when $\beta \neq 0$. Therefore, we must estimate it by consistently estimating the unknown parameter β first. That is, for a consistent estimator β_n, define $\hat{\varepsilon}_j = y_j - x_j'\beta_n$. Again, we take advantage of the empirical characteristic function of the residuals, which is defined as

$$\hat{h}_{\hat{\varepsilon}}(\tau) = \left| \frac{1}{n} \sum_{j=1}^{n} e^{i\tau\hat{\varepsilon}_j} \right|. \tag{1.6}$$

Replacing h_ε with $\hat{h}_\varepsilon$ or $\hat{h}_{\hat{\varepsilon}}$ in equation (4) does not ensure that the integration exists, so we convolve the integrand with a smoothing kernel (Stefanski and Carroll, 1990). Define a random variable z with the usual Parzen (1962) kernel density $K(z)$ and corresponding (invertible) characteristic function $h_K(\tau)$. Finite support of the characteristic function $h_K(\tau)$ is required to ensure the integrand exists and the resulting estimate is a valid density function.

Using $K(z) = (\pi z)^{-1} sin(z), (h_K(\tau) = 1\{|\tau| \leq 1\})$, our estimator of the density of u is,

$$\hat{f}_u(z) = \frac{1}{2\pi} \int_{-w_n}^{w_n} e^{-i\tau z}(1 + \hat{b}_n^2\tau^2) \left| \frac{1}{n} \sum_{j=1}^{n} e^{i\tau\hat{\varepsilon}_j} \right| d\tau, \tag{1.7}$$

where the limits of the integration are a function of an increasing sequence of positive constants w_n, which represent the degree of smoothing. In the sequel, $\{w_n\}_{n\in\mathbb{N}}$, $\{k_n\}_{n\in\mathbb{N}}$ and $\{b_n^2\}_{n\in\mathbb{N}}$ denote sequences of positive numbers which will be determined later. k_n is an intermediate sequence that will be useful for the case where C_1 and δ are unknown. When C_1 and δ are known, set $w_n = k_n$.[12]

Due to the upper and lower bound conditions on the target density function in Assumption 3, we are propose an estimator of unknown error variance parameter, b^2, that is (semi-uniformly) consistent. Therefore, setting $\tilde{b}_n^2 = k_n^{-2}\left(\frac{C_1 k_n^{-\delta}}{\hat{h}_{\hat{\varepsilon}}(k_n)} - 1\right)$ with constants $\delta > 1$ and $C_1 > 0$, we propose an explicit truncation device for the unknown variance parameter:

$$\hat{b}_n^2 = \begin{cases} 0 & \text{if } \tilde{b}_n^2 < 0 \\ \tilde{b}_n^2 & \text{if } \tilde{b}_n^2 \in [0, b_n^2] \\ b_n^2 & \text{if } \tilde{b}_n^2 > b_n^2, \end{cases} \tag{1.8}$$

where the variance parameter bound is $b_n^2 = \frac{1}{2}V(\hat{\varepsilon})$, half the variance of the estimated sum of the error components. The intuition is that we choose an increasing sequence to cover the unknown variance parameter, $\tilde{b}_n^2$, but bound it by half the total variance.[13] This is a modified version of the variance truncation device of Meister (2006).

What distinguishes our truncation device from that in Meister (2006) is that the variance of the estimated compound error is incorporated as a natural upper bound of the unknown variance of random noise v. Compared to the variance truncation device of Meister (2006), ours is more informative and converges faster, while still covering the unknown error variance associated with Laplace errors. Meister (2006) uses the bound $b_n^2 = \frac{1}{4}\ln\ln n$ for deconvolution with normal errors, and his bound arises directly from the characteristic function of the normal distribution and implicitly requires a very large sample size n. The modified truncation device, $\hat{b}_n^2$, is an important contribution of this paper which can also be applied

[12] In Section 1.4, we propose setting $w_n = k_n/\ln k_n$ in the case C_1 and δ are not fully known.

[13] Recall that for a Laplace distribution as defined in Assumption 2, the variance is $V(v) = 2b^2$. Moreover, $V(v) < V(\varepsilon)$ under Assumption 1. Hence, a natural upper bound for b^2 is one-half the variance.

in the setting of Meister (2006). Its attractiveness and usefulness will be demonstrated in the simulation section. We now discuss semi-uniform consistency of the Laplace deconvolution estimator in equation 7 within current setting.

1.3 Asymptotic Theory

To demonstrate that the unknown variance deconvolution estimator retains its asymptotic properties when the composed error is estimated, we introduce two additional conditions that will be useful in the Lemmas and Theorem to follow.

Assumption 4. *The distribution of x has bounded support.*

Assumption 5. *The estimator β_n converges at a rate of square root n. That is, $\sqrt{n}(\beta_n - \beta) = O_p(1)$ as $n \to \infty$.*

Assumption 4 follows Horowitz and Markatou (1996) while Assumption 5 guarantees that the difference between the composed errors and estimated errors is asymptotically negligible. In the pure deconvolution problem, $\beta = 0$, Assumption 5 is trivially satisfied. Moreover, the conditional mean function $x_j'\beta$ may suffer from misspecification but can be estimated with a nonparametric n^a convergence rate and $a = \frac{2}{4+q}$. We will discuss this case in the extensions in Section 1.4.

To establish semi-uniform consistency of $\hat{f}_u$, we introduce the following lemmas.

Lemma 1. *For Assumptions 1, and 3-5 and $\mathcal{L}_n = \{Laplace(0,b) : b^2 \in (0, b_n^2]\}$, the mean integrated squared error (MISE) of (1.7) is*

$$\sup_{g \in \mathcal{L}_n} \sup_{f \in \mathcal{F}_u} E_{f,g} ||\hat{f}_u - f_u||_{L_2}^2 \leq B + V + E,$$

where $B \leq const_1 \times w_n^{1-2\delta}$,

$V \leq const_2 \times n^{-1} w_n (1 + b_n^2 w_n^2)^2 + const_3 \times n^{-1} w_n^3 (1 + b_n^2 w_n^2)^2$,

$E \leq const_4 \times \sup_{g \in \mathcal{L}_n} \sup_{f \in \mathcal{F}_u} \Big(w_n \int_{-1}^{1} |h_u(sw_n)|^2 \left(\frac{d_n}{b^2}\right)^2 ds + w_n \int_{-1}^{1} |h_u(w_n s)|^2 \frac{b_n^4}{b^4} \times P_{f,g}(|\hat{b}_n^2 - b^2| > d_n) ds \Big)$, with $d_n := \frac{1}{w_n}$; f and g are the probability density function in distribution family $\mathcal{F}_u$ and $\mathcal{L}_n$, respectively, and $const_j$ are positive constants for $j = 1, 2, 3, 4$.

The proof is in the appendix. Notice the distinction between $\mathcal{L}_n$ above and $\mathcal{L}$ in Assumption 2. The former is the family of Laplace distributions with an upper bound on the variance and is a subset of the latter.[14] Following Horrace and Parmeter (2011), the B term is a bias component which is bounded by the ordinary-smoothness of f_u under Assumption 3. The V terms are variance components. The E term is a hybrid bias-variance component in which the first integral behaves like squared bias and the second integral looks like a variance. This entire bound exhibits the usual bias-variance trade-off in nonparametric density estimation. Note that the second addend of V arises from the regression function, which does not appear in the pure deconvolution setting of Meister (2006).

Establishing the convergence rate of E is not straight-forward. We need the following Lemma to assist in determining it.

Lemma 2. *Let d_n, f and g be the same as in Lemma 1. Then $\sup_{g \in \mathcal{L}_n} \sup_{f \in \mathcal{F}_u} P_{f,g}\big(|\hat{b}_n^2 - b^2| > d_n\big) \leq const \times n^{-1} k_n^{2\delta}(1 + b_n^2 k_n^2)(1 + k_n^2)$.*

The proof is in the appendix. Compared to deconvolution with normal noise as in Horrace and Parmeter (2011), estimation of ε matters here. That is, the conditional mean function in Horrace and Parmeter (2011) is linear, so their estimated error converges at a rate $n^{1/2}$, which is much faster than the logarithmic rate of their target density estimator. Therefore, estimation of the error can effectively be ignored. Here, both β_n and $\hat{f}_u$ converge at polynomial rates, so there is an additional effect on the convergence rate of the estimator of the target density.[15] Given that we replace ε with a consistent estimator, we have an additional

[14] In Meister (2006), the bounding of the normal variance is what leads to semi-uniformly consistency (as opposed to uniform consistency). Here, for Laplace errors, we still impose this "strong" condition for ease of proof. However, it may not be a necessary condition.

[15] The compound effect of estimating the regression function will slow the target density rate compared to pure (non-regression) deconvolution, but the final rate is not a simple algebraic sum of the rates.

term k_n^2 in Lemma 2, as well as the characteristic function of the Laplace distribution, embodied in the term $(1+b_n^2k_n^2)$. The second addend of E in Lemma 1, together with the upper bound of B and the first term in E, ensures convexity of the entire bound with respect to the bandwidth parameter k_n. Therefore, the optimal bandwidth w_n, which is a function of k_n, and the entire convergence rate of the density estimator can be determined.

Notice that neither of the proofs of the above two lemmas leverage anything on the assumption that the smoothness parameters of the target density are known (or not). However, for joint minimization of the upper bounds of MISE of Lemma 1, this assumption plays a role. That is, if the smoothness parameters are fully known (i.e., C_1 and δ) tight bounds can be achieved by setting $w_n = k_n$; otherwise, the best general upper bound can be reached by setting $w_n = k_n/\ln k_n$. The latter case is considered in the next section. First, we introduce the following theorem when C_1 and δ are known.

Theorem 1. *Assume δ and C_1 are known. Under Assumption 1, 3-5, set $\{b_n^2\}_{n\in\mathbb{N}} = \frac{1}{2}V(\hat{\varepsilon})$ and $w_n = k_n$ with $\{k_n\}_{n\in\mathbb{N}} = \{(\frac{n}{b_n^2})^{\frac{1}{6+2\delta}}\}_{n\in\mathbb{N}}$, if $1 < \delta \leq 1.5$ or $\{k_n\}_{n\in\mathbb{N}} = \{(\frac{n}{b_n^8})^{\frac{1}{3+4\delta}}\}_{n\in\mathbb{N}}$, if $\delta > 1.5$. For any $g \in \mathcal{L}_n$, the proposed deconvolution kernel density estimator in equation (1.7) is bounded from above as follows:*

$$\sup_{f_u\in\mathcal{F}_u} E_{f,g}||\hat{f}_u - f_u||_{L_2}^2 \leq n^{-\frac{2\delta-1}{6+2\delta}} \quad if \quad 1 \leq \delta \leq 1.5,$$

and

$$\sup_{f_u\in\mathcal{F}_u} E_{f,g}||\hat{f}_u - f_u||_{L_2}^2 \leq n^{-\frac{2\delta-1}{3+4\delta}} \quad if \quad \delta > 1.5$$

where δ is defined in Assumption 3.

The proof is in the appendix. The proposed density estimator is semi-uniformly consistent. That is, $\hat{f}_u$ is uniformly consistent over a given class of Laplace distributions $\mathcal{L}_n$. The optimal convergence rate for an ordinary-smooth target density is achieved in a minimax sense. It is similar to the conclusions in Fan (1991), even though in this exercise the variance of the error distribution is unknown and the composed error needs to be estimated. The polynomial convergence rate plays a role in the following sense. After imposing the modified variance truncation device, which is the proposed best choice one can use for unknown variance, and after deriving the optimal sequences for convergence (i.e., the order of the positive sequence $\{k_n\}_{n\in\mathbb{N}}$), we still achieve a polynomial convergence rate which is consistent with the lower bound derived by Fan (1991).

At first glance the Theorem 1 is similar to Theorem 2 in Meister (2006), but there are three major differences: (i) the upper bound of the noise v is not a known constant but a consistently estimated (at $\sqrt{n}$ rate) quantity (i.e., $\frac{1}{4}\ln\ln n$ versus $\frac{1}{2}V(\hat{\varepsilon})$); (ii) the chosen sequences are functions of the target density smoothness order, δ, which is due to the characteristic function of the Laplace noise, leading to different convergence rates (or effective sample size) as shown in Table 3.2; and (iii) we consider estimation in the regression setting, which is more general than the pure deconvolution setting ($\beta = 0$), and yields different convergence rates with Laplace noise. In Horrace and Parmeter (2011) this last difference was easily handled, given the slow convergence of the density estimator due to the assumption of super-smooth noise. It is more nuanced in the context of Laplace noise, given the polynomial rate of convergence. This has important implications if one were to estimate the unknown conditional mean using nonparametric methods. We discuss this and other extensions of the Laplace deconvolution estimator in the next section.

1.4 Some Useful Extensions

We discuss two useful extensions to the Laplace deconvolution estimator which are likely to arise in applications: (i) C_1 and δ are unknown in Assumption 3 and (ii) deploying nonparametric regression to estimate the unknown conditional mean needed to subsequently recover $\hat{\varepsilon}$. It is rare in applications that researchers have information on the target density. This leads to uncertainty in C_1 and δ, two parameters which are important in the implementation of our estimator.[16] Also, if we wish to follow the work of Fan, Li and Weersink (1996) and

[16]... and the estimator of Meister (2006) as well.

estimate the unknown regression function nonparametrically, then we must think carefully about the relative polynomial convergence rates of the deconvolution estimator and the nonparametric regression estimator. This is not a consideration with normal noise due to the logarithmic convergence rates it produces.

1.4.1 Selection of Unknown C_1 and δ

In the usual case that δ and C_1 are unknown and, therefore, might be misspecified,[17] we could apply the following selection rule due to Meister (2006):

Selection rule 1. *If C_1 and δ are unknown, we specify one set of $\{C_1, \delta\}$ and choose $w_n = k_n / \ln k_n$.*

An alternative rule may be based on our procedure when δ and C_1 are known. First, we specify one set of parameters $\{C_1, \delta\}$ to pin down the variance truncation device defined in Section 1.2, and then by Lemmas 1 and 2 we determine the optimal choice for the sequence $\{k_n\}_{n\in\mathbb{N}}$. The trade-off is a slower convergence rate of the estimated target density compared with that in the fully-known case due to lack of information about the target density. This implicitly requires a larger n to achieve a reliable estimate of the target density. This can be seen from following theorem.

Theorem 2. *Assume δ and C_1 are unknown. Under Assumption 1, 3, 4, and 5 set $\{b_n^2\}_{n\in\mathbb{N}} = \frac{1}{2}V(\hat{\varepsilon})$ and $w_n = k_n/\ln k_n$ with $\{k_n\}_{n\in\mathbb{N}} = \{(\frac{n}{b_n^2})^{\frac{1}{6+2\delta}}\}_{n\in\mathbb{N}}$, if $1 < \delta \leq 1.5$, or $\{k_n\}_{n\in\mathbb{N}} = \{(\frac{n}{b_n^8})^{\frac{1}{3+4\delta}}\}_{n\in\mathbb{N}}$, if $\delta > 1.5$. For any $g \in \mathcal{L}_n$, the proposed deconvolution kernel density estimator in equation (1.7) is bounded from above as following:*

$$\sup_{f_u\in\mathcal{F}_u} E_{f,g}||\hat{f}_u - f_u||_{L_2}^2 \leq (n/\ln n)^{-\frac{2\delta-1}{6+2\delta}} \quad if \quad 1 < \delta \leq 1.5$$

[17] Actually, if one wants to assume the random noise is super-smooth with similarity index s, the smoothness parameter δ of target density can be estimated as well as the s by an adaptive procedure proposed by Butucea, Matias and Pouet (2008).

and

$$\sup_{f_u \in \mathcal{F}_u} E_{f,g}||\hat{f}_u - f_u||^2_{L_2} \leq (n/\ln n)^{-\frac{2\delta-1}{3+4\delta}} \quad if \quad \delta > 1.5$$

where δ is defined by Assumption 3.

The proof is similar to that of Theorem 1 in the appendix and is contained therein. The only difference between the bounds in Theorem 1 and in Theorem 2 is that the bounds are negative exponents of n in the former and of $n/\ln n$ in the latter, and this is the price one pays for not knowing the smoothness parameters of the target density. Based on the Theorem 2 and Table 3.1, we propose a rule-of-thumb adaptive procedure as follows:

Step 1: Set initial estimates for C_1 and δ. A useful rule-of-thumb is C_1 is commonly between 0 and 1; δ is between 1 and 10.

Step 2: Treating this C_1 and δ as "known," select $k_n = w_n$ and apply the proposed deconvolution techniques to construct the estimated target density, $\hat{f}_{known}(u)$, say.

Step 3: Now, with the same C_1 and δ assume they are unknown and select $w_n = k_n/\ln k_n$. Again, apply the proposed deconvolution estimator to construct the estimated target density as $\hat{f}_{unknown}(u)$, say.

Step 4: Compare the *vector* of values $\hat{f}_{known}(u)$ and $\hat{f}_{unknown}(u)$ over a discretized support with a Euclidean distance measure (e.g., $\Delta = ||\hat{f}_{known}(u) - \hat{f}_{unknown}(u)||^2$). Iterate Steps 1 to 3 until Δ is smaller than a pre-specified threshold, say 0.0001.

One caveat with this iterative approach is that Δ may be quite large initially. The essential point is that more information about the underlying distribution is revealed after several trials with combinations of the smoothness parameters. This is similar in spirit to the adaptive procedure proposed by Butucea, Matias and Pouet (2008), but their targets are a "self-similarity index" and a smoothness parameter with super-smooth errors, and not a density.

1.4.2 Nonparametric Estimation of the Conditional Mean

If one is unsure of the linear specification of the conditional mean, equation (1.1) can be generalized to the nonparametric case as follows:

$$y_j = g(x_j) + u_j + v_j \quad j = 1, 2, \ldots, n \tag{1.9}$$

where $g(.)$ is unknown and $x \in \mathcal{R}^q$. Under certain regularity conditions,[18] a straightforward nonparametric kernel estimator for the unknown function $g(x)$ is:

$$\hat{g}(x) = \frac{\sum_{j=1}^{n} Y_j K(\frac{X_j - x}{h})}{\sum_{j=1}^{n} K(\frac{X_j - x}{h})}$$

where $K(\cdot)$ is the standard Gaussian kernel with bandwidth h. Note that since the convergence rate of the nonparametric estimator is a polynomial function of the number of covariates, this may impact application of the Laplace deconvolution estimator.

By Theorem 2.6 (with Condition 2.1) of Li and Racine (2007), the convergence rate of the estimated function is:

$$\sup_{x \in S} |\hat{g}(x) - g(x)| = O\left(\frac{(\ln n)^{0.5}}{(nh_1 \cdots h_q)^{0.5}} + \sum_{s=1}^{q} h_s^2 \right) \quad a.s.$$

Assuming each bandwidth (h_s) has the same order of magnitude, the optimal choice of h_s that minimizes $MSE[\hat{g}(x)]$ is $h_s \sim n^{-\frac{1}{4+q}}$, and the resulting MSE is therefore of order $O(n^{-\frac{4}{4+q}})$. Consequently, the estimated error, $\hat{\varepsilon}$, is n^a consistent where $a = \frac{2}{4+q}$. That is, $n^a(\hat{\varepsilon} - \varepsilon) = O_p(1)$ as $n \to \infty$.

Similarly, we can establish the convergence rate as follows:

Theorem 3. *Under Assumptions 3-5, and Condition 2.1 in Li and Racine (2007) consider equation* (1.7) *and take* $\{b_n^2\}_{n \in \mathbb{N}} = \frac{1}{2}V(\hat{\varepsilon})$ *and* $w_n = k_n$ *with* $\{k_n\}_{n \in \mathbb{N}} = \{(\frac{n}{b_n^2})^{\frac{2a}{6+2\delta}}\}_{n \in \mathbb{N}}$*, if* $1 < \delta \leq 1.5$*, and* $\{k_n\}_{n \in \mathbb{N}} = \{(\frac{n}{b_n^8})^{\frac{2a}{3+4\delta}}\}_{n \in \mathbb{N}}$*, if* $\delta > 1.5$*. For any* $g \in \mathcal{L}_n$*, the proposed*

[18] Details see Condition 2.1 in Li and Racine (2007).

deconvolution kernel density estimator in equation (1.7) *is bounded from above as follows:*

$$\sup_{f_u \in \mathcal{F}_u} E_{f,g}||\hat{f}_u - f_u||_{L_2}^2 \leq n^{-\frac{2a(2\delta-1)}{6+2\delta}} \quad if \quad 1 < \delta \leq 1.5$$

and

$$\sup_{f_u \in \mathcal{F}_u} E_{f,g}||\hat{f}_u - f_u||_{L_2}^2 \leq n^{-\frac{2a(2\delta-1)}{3+4\delta}} \quad if \quad \delta > 1.5$$

where $a = \frac{2}{4+q}$ *and* δ *is defined by Assumption 3.*

The proof is very similar to the proof of Theorem 1 in the appendix, and a sketch of the proof is contained therein.

1.5 Monte Carlo Simulations

We present a Monte Carlo study of the finite sample properties of the Laplace deconvolution estimator. For ease of comparison, we follow the sample design from Meister (2006) and Horrace and Parmeter (2011) except that we consider performance of the Laplace deconvolution with both Laplace errors (correctly specified) and normal errors (misspecified). We focus on sample sizes of $n = 500$, 1,000, and 3,000 with the linear model:

$$y_j = 4 + 3x_j + v_j + u_j, \quad j = 1, \ldots, n. \tag{1.10}$$

The x_js are generated from a standard normal distribution. Random noise v_j is generated from either a standard Laplace (correctly specified) or normal (misspecified) distribution for a range of values of the variance to produce several signal-to-noise settings. The u_js are generated from the twice convolved, zero-mean Laplace density for which the probability density function is $\tilde{L}(x) = \frac{1}{4}e^{-|x|}(|x| + 1)$.[19] We fix the variance of u to 2. In this setting it is known that $C_1 = 1/4$, $\delta = 4$ and $T = 1$.[20]

[19] This follows from the setting in Meister (2006).

[20] We are not concerned with C_2, since it has no bearing on any calculations for the estimator.

Following Theorem 1, we choose $b_n^2 = \frac{1}{2}V(\hat{\varepsilon})$ where $\hat{\varepsilon}$ is the residual from the first-step ordinary least squares (OLS) estimation, and $k_n = n^{\frac{1}{4\delta+3}}(b_n^2)^{\frac{-4}{4\delta+3}} = n^{\frac{1}{19}}(b_n^2)^{-\frac{4}{19}}$, correspondingly as $\delta = 4 > 1.5$. To explore the impact of the relative ratio of the component variances, we consider different scenarios of the signal-to-noise ratio which is defined as the ratio of $V(u)$ and $V(v)$: $\gamma := \sigma_u^2/\sigma_v^2 \in \{1/2, 1, 2\}$. We also apply our Laplace deconvolution estimator in the misspecified case where the errors are normally distributed. We compare the performance of our estimator under misspecification to the normal deconvolution estimator of Meister (2006) which is correctly specified. Even in this case, our estimator performs fairly well. We also explore the finite sample performance of our proposed rule-of-thumb adaptive procedure when the smoothness parameters of the target density are unknown.

The performance of our estimator is assessed through the root mean integrated square error (RMISE):

$$RMISE(\hat{f}_u) = \sqrt{\frac{1}{R}\sum_{l=1}^{R}\frac{1}{M}\sum_{i=1}^{M}(\hat{f}_l(u_i) - f(u_i))^2} \tag{1.11}$$

where R is the number of replications and $M = 256$ is the number of evaluation points over $u \in (-5, 5)$, which is fixed across the R replications.

1.5.1 Laplace Deconvolution with Laplace Errors

First, we consider the case that the random noise v_j is correctly specified (i.e., v_j is drawn from a Laplace distribution with variance 1). Figures 1-3 show the results for a single random draw ($R = 1$) across various sample sizes {500, 1,000, 3,000} and compare the proposed estimator (CHP) to the true unknown density ($True$). The graphical fit of the proposed estimator is quite good with only 500 observations (Figure 1). Most of the bias comes from estimation around the mode.[21] As the sample size increases, the RMISE of the proposed estimator (CHP) decreases from 0.0148 (Figure 1) to 0.0142 (Figure 2) and to 0.0125 (Figure 3).

[21] Estimation of a density around the mode is difficult due to the derivative at the mode being zero (Henderson and Parmeter, 2015).

Figures 4-6 show the results for a single draw ($R = 1$) and fixed sample size n =1,000 but varying the signal-to-noise ratio σ_u^2/σ_v^2=2/1, 2/2, 2/4. The proposed estimator (CHP) works very well when σ_u^2/σ_v^2=2/1 with 1,000 observations. As the signal-to-noise ratio decreases, the RIMSE of proposed estimator (CHP) increases from 0.0136 (Figure 4) to 0.0142 (Figure 5) and to 0.0180 (Figure 6). Even for the noisiest case (Figure 6) with $\sigma_u^2/\sigma_v^2 = 2/4$, the fit is very good except in an interval around the mode.

Table 3.3 contains detailed results from $R = 500$ simulations with varying sample sizes {500, 1,000, 3,000} and signal-to-noise ratios {1/2, 1, 2}. For each signal-to-noise setting (each column), the RMISE decreases monotonically as the sample size increases from 500 to 3,000 (down the rows), demonstrating the consistency of the proposed estimator (CHP). Unexpectedly, the RMISE is not increasing as the signal-to-noise ratio increases across the columns. This is an atypical finding that is due to the variance truncation device: when the variance of the random noise is relatively small, the estimated variance $\hat{b}_n^2$ is more likely to be closer to zero which dilutes the ability of the deconvolution estimator to recover the target density. Alternatively, when the variance of the random noise is relatively large, the estimated variance is no longer near zero, but the performance of the deconvolution estimator deteriorates as there is little information in the target density taken from the compound errors. This is a limitation of the variance truncation device.

1.5.2 Laplace Deconvolution with Misspecified Errors

To understand the impact of misspecification of the error distribution, we consider the performance of the proposed estimator when the true error is distributed normal. We compare the performance of our proposed estimator (CHP) with that of Meister (2006).

As a first pass on the empirical performance, Figures 7-9 show the results for the case with fixed $\sigma_u^2/\sigma_v^2 = 2/2$ for a single draw ($R = 1$) across various sample sizes. The proposed estimator (CHP) shows decent performance even with sample size of $n = 500$ (Figure 7). The figure contains plots of the proposed estimator (CHP), the estimator of Meister (2006)

($Meister06$), and the true normal density ($True$). As the sample size increases, the RMISE of the proposed estimator (CHP) changes from 0.0151 (Figure 7) to 0.0156 (Figure 8) to 0.0137 (Figure 9). Our estimator (CHP) performs as well as Meister's when the sample size is large ($n = 3,000$). An intuitive explanation is that the proposed estimator converges faster than Meister's estimator (even under misspecification).

Figures 10-12 show the results for $R = 1$ and fixed sample size $n = 1,000$ across the various signal-to-noise ratios. The proposed estimator (CHP) performs quite well in the least noisy case even though the error distribution is misspecified. As the signal-to-noise ratio decreases, the RIMSE of the proposed estimator increases from 0.0155 (Figure 10) to 0.0156 (Figure 11) and to 0.0191 (Figure 12) whereas the RMISE of Meister's estimator increases from 0.0120 (Figure 10) to 0.0172 (Figure 11) to 0.0260 (Figure 12). When the signal-to-noise ratio decreases from 1 to 0.5 (Figures 11 and 12, respectively) the misspecified estimator even outperforms Meister's estimator.

Table 3.4 presents the results of $R = 500$ replications across various sample sizes and signal-to-noise ratios under misspecification. Though misspecified, the RMISE of the proposed estimator decreases monotonically as the sample size increases (down each column) for each signal-to-noise ratio setting, and it is comparable to that of Meister's correctly specified estimator. In the most noisy setting, $\sigma_u^2/\sigma_v^2 = 2/4$, the proposed estimator outperforms Meister's estimator across all sample sizes. This may be due to the faster convergence rate of the proposed estimator coupled with the fact that the characteristic functions of the normal and the Laplace are quite similar.[22] Fixing the sample size (within each row), both RMISEs increase when the signal-to-noise ratio decreases as the information that can be recovered is reduced. Overall, the proposed estimator is robust to misspecification of the error distribution and its convergence rate is faster than that of Meister's estimator.

[22] Actually the characteristic function of the Laplace distribution is the second order Taylor expansion of that of a normal random variable with same variance (Hesse, 1999).

1.5.3 Deconvolution With Unknown Smoothness Parameters

To verify the feasibility and performance of the proposed rule-of-thumb adaptive procedure for unknown smoothness parameters of section 4.1, a set of simulations are performed. We employ the same simulation design. Specifically, the true target density is still a twice-convolved Laplace with true smoothness parameters of $C_1 = 1/4$ and $\delta = 4$. We search on a two-dimension grid of $C_1 \in \{0.1, 0.25, 0.40, 0.55, 0.70, 0.85\}$ and $\delta \in \{2, 4, 6, 8\}$ to minimize the Euclidean distance of the two estimated densities: the estimated density assuming the chosen C_1 and δ are known and the estimated density assuming these parameters are unknown. We restrict the range of u to be $(-5, 5)$ and evaluate over 128 evenly spaced points within this range.

Figure 13 shows the estimated densities (labeled CHP for the estimate with known smoothness parameters and CHP_{UN} for the estimate with unknown parameter) and the true density (labeled $True$) for one simulation ($R = 1$) with sample size $n = 1,000$ and signal-to-noise ratio equal to 1. The chosen smoothness parameters are: $C_1 = 0.1$ and $\delta = 2$. Even though the chosen smoothness parameters are misspecified (not exactly equal to the their true values $C_1 = 1/4$ and $\delta = 4$), the overall fit of the density with estimated parameters is quite good (CHP_{UN}) and appears to be better than the fit assuming the true values of the parameters, particularly around the mode.[23]

A more comprehensive analysis is conducted in Figures 14-16. Figure 14 shows the Euclidean distance of the estimated densities: $\Delta = ||\hat{f}_{known} - \hat{f}_{unknown}||^2$, as a function of the smoothness parameters for a single draw ($R = 1$). Figures 15 and 16 show the Euclidean distance between the true density and the estimated density taking the chosen C_1 and δ as known, $||\hat{f}_{known} - f_{true}||^2$, and unknown, $||\hat{f}_{unknown} - f_{true}||^2$, respectively. A straight comparison of the three figures indicates that the convergence pattern is almost identical

[23] The reader is reminded that the fit of the estimated densities, whether with or without known smoothness parameters, is a function of the Euclidean distance evaluated over the 128 points in their support. Therefore, the relative fit of the densities with known and unknown parameters will vary over this support. That is, we should not expect the density with known parameters to always have better fit than the estimated density with unknown parameters. This is reflected in Figure 13

which means that minimizing the Euclidean distance of the estimated densities (Figure 14) is almost equivalent to minimizing the Euclidean distance of the estimated density and the true underlying density (Figures 15 and 16). Obviously, the Euclidean distance is smaller for values around the true smoothness parameters ($C_1 = 1/4$ and $\delta = 4$) in this context.

Though it is a useful tool, two caveats are worth mentioning. First, our Laplace deconvolution estimator assumes that the error distribution is Laplace. If this assumption is violated, the adaptive procedure may not perform as well as we see here. Second, the Euclidean distance between the true density and the estimated density achieves small values in a range of smooth parameters rather than at one specific point in Figure 14. It indicates that the proposed rule-of-thumb adaptive procedure is informative for providing a small range of the smoothness parameters rather than one optimal point.

To calculate the RMISE when the smoothness parameters are unknown, we replicate the above simulations for $R = 100$ with various sample sizes and signal-to-noise ratios.[24] The results are presented in Table 2.5.[25] Similar to Table 3.3, the convergence pattern still holds when the sample size increases with fixed signal-to-noise ratios. That is, reading down the columns, RMISE is decreasing in the sample size. As we read across RMISE columns within a row, the RMISE is decreasing slightly and then increasing. We also report the chosen smoothness parameters, δ and C_1, based on minimizing the Euclidean distance in Table 2.5. They vary slightly around 2 and 0.1, respectively. They are not always accurate (compared to the true values) but still render reasonably good estimates of the target density.

1.6 Application

IIn this section two applications demonstrate the utility of the proposed method. We consider the parametric Laplace stochastic frontier model (Horrace and Parmeter, 2018), a regression-based application of the method, and a second application where the outcome of interest,

[24] We reduce the replication size from 500 to save computation time.

[25] We report the RMISE of $\hat{f}_{known}$ here.

daily saturated fat intake, is contaminated with measurement error (which we assume to be Laplace) and $\beta = 0$ in equation (1.1). In the first application we assume the smoothness parameters are known; in the second we use our adaptive rule-of-thumb to select them.

1.6.1 Stochastic Frontier Analysis

A typical parametric stochastic frontier model is equation (1), but restricting $u < 0$ (for a production frontier) or $u > 0$ (for a cost frontier). Given distributional assumptions on inefficiency, u (e.g., exponential or half-normal) and noise, v (e.g., normal or Laplace), β may be consistently estimated and used to calculate the conditional distribution of firm-level inefficiency, which is typically characterized by the empirical distribution of u conditional on ε (e.g., Jondrow et al. 1982). Much of the existing literature assumes normality of v (i.e., super-smooth v) and then applies maximum likelihood estimation (MLE). Relaxing parametric assumptions on the inefficiency distribution in these models is important, as articulated by Kneip, Simar, and Van Keilegom (2015, p.380) who note that "...there does usually not exist any information justifying particular distributional assumptions on (inefficiency)." Additionally, Tsionas (2017, p.1169) suggests that a model constructed to provide microfoundations for the presence of inefficiency "...does not make a prediction about the distribution." These statements underlie the importance of seeking alternative estimation approaches to recover important features of the stochastic frontier model; those approaches which eschew restrictive parametric assumptions are likely to curry favor among practitioners and regulators alike.

There is also no reason to favor normally distributed errors in the stochastic frontier model (Horrace and Parmeter, 2018). As such we apply our Laplace deconvolution estimator to estimate the distribution of inefficiency from a cost frontier for US banks. The data come from Feng and Serletis (2009) and are obtained from the Reports of Income and Condition (Call Reports).[26]

[26] The data are publicly available on the Journal of Applied Econometrics data archive website

The data are a sample of US banks covering the period from 1998 to 2005 (inclusive). After deleting banks with negative or zero input prices, we are left with a balanced panel of 6,010 banks observed annually over the 8-year period. A more detailed description of the data may be found in Feng and Serletis (2009). For our purposes we ignore the panel structure of the data and choose the most recent year data, 2005, for our example. The goal of this exercise is to estimate the marginal distribution of u and compare it with the typical half-normal distribution which informs practical choice of parametric assumption on u, which , in turn, informs estimation of $E(u|\varepsilon)$.[27]

The data contain information on three output quantities and three input prices. The three outputs are consumer loans, Y_1; non-consumer loans, Y_2, which consists of industrial and commercial loans and real estate loans; and securities, Y_3, including all non-loan financial and physical assets minus the sum of consumer loans, non-consumer loans, securities and equity. All outputs are deflated by the Consumer Price Index (CPI) to the base year of 1988. The three input prices are: the wage rate for the labor, P_1; the interest rate for borrowed funds, P_2 and the prices of physical capital, P_3. The total cost, C, is the sum of three corresponding input costs: total salaries and benefits, expenses on premises and equipment, and total interest expenses. Our specification of output and input prices is the same as (or very similar to) what is typical in the literature (see, for example, Feng and Serletis, 2009; Kumbhakar and Tsionas, 2005.) The cost frontier model is

$$c_j = \alpha + x_j'\beta + u_j + v_j \quad j = 1, \ldots, n, \tag{1.12}$$

where $c_j = \ln C_j$; $x_j = \ln X_j$ with X_j including the three output quantities and three input prices: Y_1, Y_2, Y_3, P_1, P_2, P_3; and $u_j > 0$ is firm-specific inefficiency.

We estimate the distribution of cost inefficiency in three ways. First, we estimate a fully parametric model, assuming v is distributed $N(0, \sigma_v^2)$ and u is distributed $|N(0, \sigma_u^2)|$.

http://qed.econ.queensu.ca/jae/2009-v24.1/feng-serletis/.

[27] Once $\hat{f}_u$ is obtained, one can estimate the efficiency score using numerical integration on a grid of $\hat{\varepsilon}$. To avoid an overloading of present paper, we stick to the estimation of marginal density of u.

Our maximum likelihood estimates of the distributional parameters are $\hat{\sigma}_u = 1.294$ and $\hat{\sigma}_v = 0.989$, implying $E(u) = \hat{\sigma}_u\sqrt{2/\pi} = 1.033$. Then, our estimate of the density of u is $|N(0, 1.294^2)|$, which is shown as the dotted line (SFA) in Figure 18. Second, we estimate equation (3.8) by OLS. Figure 3.1 shows a histogram of the OLS residuals, $\hat{\varepsilon}_j$. The asymmetry of the distribution (skew equals 1.550) suggests non-zero cost inefficiency.[28] Selecting $\delta = 3$ and $C_1 = 1$ and using Theorem 1, the deconvolution estimator yields an estimate of σ_v^2 equal to 0.0403.[29] A plot of the density estimate, $\hat{f}_u(u)$, is shown as the dashed line (CHP) in Figure 3.2. Third, using the procedure of Hall and Simar (2002) with a bandwidth of 0.3052, we detect a jump discontinuity point in $\hat{f}_u(u)$ at $u = -0.355$ which implies an estimate of $\hat{E}(u) = 0.355$. Then using the boundary kernel proposed by Zhang and Karunamuni (2000), with an estimated error variance of 0.0403 (as before), the boundary bias corrected density estimate is shown as the solid line (CHP_$E(u)$_bc) in Figure 3.2.[30]

Figure 3.2 shows all three density estimators for US bank inefficiency in 2005. Notice that even without a boundary correction, the deconvolution estimator (CHP) has a thinner right tail than the estimated half normal density (SFA). With boundary correction in place, the deconvolution estimator (CHP_$E(u)$_bc) implies that US banks in 2005 have a much smaller average inefficiency than parametric SFA would have predicted. This corresponds to the fact that in 1998 there are 10,139 banks in the US and this number declined to 8,390 in 2005 due to industry consolidation (Feng and Serletis, 2009).

Finally, there are at least two reasons to employ the proposed estimator: 1) the proposed method provides a robustness check for the distributional assumptions made in a parametric stochastic frontier model and 2) the skewness of the OLS residuals is greater than one, which invalidates the choice of the half-normal assumption for the distribution of u (which

[28] It is interesting to note that with a skew of 1.55, this provides evidence against use of the half-normal distribution.

[29] For the Laplace distribution, $\delta = 2$; for convolved Laplace, $\delta = 4$. The choice $\delta = 3$ is between Laplace and convolved Laplace.

[30] For Laplace deconvolution, we can apply directly Example 1 in Zhang and Karunamuni (2000).

has maximal skewness of 1 by definition).

1.6.2 Daily Saturated Fat Intake With Measurement Errors

The data come from Wave III (1988-1994) of the National Health and Nutrition Examination Survey, abbreviated NHANES III. Our interest is the survey response to daily saturated fat intake of 3,551 women between the ages of 25 and 50. This data set is ideally suited to our Laplace deconvolution estimator as it is well established that saturated fat consumption is recorded with measurement errors. In fact, previous analysis of the NHANES Wave I (1971-75) and Wave II (1976-1980) data suggest that more than 50% of the variability in the observed data may be due to measurement errors. See Stefanski and Carroll (1990), Carroll, Ruppert and Stefanski (2006) and Delaigle and Gijbels (2004).

The data were originally recorded to explore the relationship between breast cancer and dietary fat intake, see Jones et al. (1987). Stefanski and Carroll (1990) were the first to consider nonparametric deconvolution techniques to estimate the underlying true density of saturated fat intake, using NHANES I. Subsequently, Carroll, Ruppert and Stefanski (2006), Delaigle and Gijbels (2004) and others applied deconvolution estimators to NHANES II. In each of these applications a normal error distribution was assumed. To the best of our knowledge we are the first to apply deconvolution techniques to NHANES III (and certainly the first to apply Laplace deconvolution to any of these data). Here, saturated fat (fat) is measured in milligrams per day, and we apply the same data transformation as Delaigle and Gijbels (2004): $log(fat + 5)$.

To these data we implement a) the proposed estimator with Laplace errors (CHP), b) the estimator with normal errors due to Meister (2006) ($Meister$), and c) an error free estimator ($ErrorFree$), based on pure kernel density estimation of the observed data assuming there is no measurement error.[31]

First, we apply the proposed rule-of-thumb adaptive procedure to get a preliminary

[31] We use the package "ksdensity" in Matlab for the $ErrorFree$ case.

estimate of the smoothness parameters since they are unknown. Specifically, we search for the minimum of the Euclidean distance between the density estimator with unknown smoothness parameters and density estimator with known smoothness parameter, Δ, over a grid of $\delta \in \{1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3\}$ and $C_1 \in \{0.1, 0.25, 0.40, 0.55, 0.70, 0.85, 1\}$.[32] Figure 19 shows the surface of the Euclidean distance as a function of the smoothness parameters over the grid. The Δ increases as C_1 rises from 0 to 1 except when δ is around 2. It seems that $\delta = 1.5$ and $\delta = 3$ yield the minimum distance. It turns out that when $\delta = 3$, the estimated density decreases very quickly and goes below zero and becomes volatile when $log(fat + 5) < 2$ or $log(fat + 5) > 4.5$. Therefore, we consider the $\delta = 1.5$ case to be optimal. Specifically, we choose $C_1 = 1$ and $\delta = 1.5$ as our baseline model. We then consider alternative specifications of the smoothness parameters as a robustness check.

Figure 1.17 presents the final results of the analysis. The estimated error variance is 0.065 based on the CHP estimator and 0.525 based on the $Meister$ estimator in the baseline model. The $Meister$ error variance estimate is exceedingly large compared to the variance of the observed (convoluted) data, 0.236.[33] The CHP error variance estimate is more reasonable in the sense of being less than the total observed variance, and its corresponding signal-to-noise ratio is 0.275. This is consistent with the finding in the existing literature that about 30-50% of the variability of observed data is due to measurement error. The tail behaviors in Figure 1.17 shows that the $Meister$ estimator assigns more variance to the error variance than expected and it decreases to zero very quickly. The CHP estimator extracts the target density information based on the smoothness assumptions, which gives a reasonable variance estimate and tends to have longer tails.[34]

The CHP density estimator based on the NHANES III data is quite similar to that of Delaigle and Gijbels (2004), despite the fact that they used the NHANES II data, assumed

[32] We also tried larger range of δ and narrow down to this specific range by searching the minimum of Δ.

[33] It seems to violate the independence assumption between the target variable and the measurement error.

[34] Under Assumption 1, i.e., u and v are independent, the variance of Y should be the sum of the variances of u and v. Empirically, this may not be the case for real data.

the error to be normal, along with differing identification assumptions. They experiment with different "known" values of the signal-to-noise ratio, while we have to select the smoothness parameters. The minor difference is that our estimated tails are slightly thicker than theirs, however the means of the estimated densities are nearly identical.

As a robustness check, different combinations for the values of δ and C_1 are considered for the CHP estimator: $C_1 = 1$ and $\delta = 1.5$; $C_1 = 1$ and $\delta = 2$; $C_1 = 0.6$ and $\delta = 1.5$; $C_1 = 0.6$ and $\delta = 2$ in Figure 1.18. The baseline ($C_1 = 1$, $\delta = 1.5$) is in the upper-left panel of the figure. As we move to different panels in the figures we change the values of the smoothness parameters, so the CHP estimator is changing across panels, while the $ErrorFree$ estimator is fixed. For $C_1 = 0.6$, $\delta = 1.5$ (lower-left panel), the estimated error variance of CHP is 0.019 which is less than the baseline model, and it has less fat tails. For $C_1 = 1$, $\delta = 2$ (upper-right panel), the estimated error variance of CHP is 0 which makes it nearly coincide with the $ErrorFree$ case.[35] This means that it is more difficult for information on the measurement error to be be disentangled under these smoothness assumptions. We can also vary C_1 to recover certain information concerning the noise or the error term. For instance, $C_1 = 0.6$, $\delta = 2$ (lower-right panel), the estimated error variance of CHP is still 0 which renders an identical deconvolution density estimate. It seems that the variability of δ dominates that of C_1. This is intuitive as $\tau \to \infty$, the effect of C_1 is ignorable in Assumption 3.

1.7 Conclusion

This paper proposes a semiparametric estimator for a cross-sectional error component model. Instead of focusing on the estimation of the model parameters with the typical assumption of normality, we are interested in the density of the target error component. To estimate the target density without fully known random noise, we modify the variance truncation

[35] One point worth mentioning is that these minimax deconvolution techniques can produce error variance estimates equal to zero as we vary the choice of C_1 and δ. Recall that $\hat{b}_n^2$ is bound between 0 and $0.5V(\hat{\varepsilon})$. When it happens, the deconvolution estimators will be very similar to the $ErrorFree$ estimator.

device proposed by Meister (2006) and extend the methodology to the framework of an error component model with a Laplace error term with unknown variance.

The density deconvolution estimator with Laplace error has at least two attractive characteristics for applied researchers: 1) it possesses a faster convergence rate than that of normal distributed errors (i.e., $O(n^c)$ versus $O((\ln n)^c)$) and 2) it is robust to misspecification of the true underlying error distribution. A third (potential) feature that practitioners may find appealing is the Laplace errors generate different insights than normal errors: for example, the LAD estimator rather than OLS, the Laplace stochastic frontier model (Horrace and Parmeter, 2018) and the L-SIMEX estimator (Koul and Song, 2014).

For future research, it may be useful to extend the model to panel data and use it to estimate both the interest component's and noise's distributions nonparametrically. For example, with a nonparametric production or cost function this would imply a fully nonparametric stochastic frontier model. Jirak, Meister and Reiss (2014) studied the adaptive function estimation in nonparametric regression with one-sided errors. Another interesting strand in this area is to investigate the distribution of the unobserved heterogeneity with proposed deconvolution techniques. Recently, Evdokimov (2010) takes an initial step to explore that in a panel data model and Ju, Gan and Li (2019) applies it with a real data set.

Table 1.1: Effective Sample Size Compared with OLS

		Sample Size n				
Method	Convergence Rate	100	1000	3000	5000	10000
Parametric (OLS)	$n^{-1/2}$	100	1000	3000	5000	10000
Laplace Deconv.	$n^{-2/9}(\delta = 3/2)$	8	22	35	44	60
	$n^{-1/3}(\delta = 3)$	22	100	208	292	464
	$n^{-7/19}(\delta = 4)$	30	162	365	531	886
Normal Deconv.	$\frac{\ln(\ln n)}{\ln n}(\delta = 3/2)$	9	13	15	16	17
	$(\frac{\ln(\ln n)}{\ln n})^2(\delta = 3, 4)$	83	163	219	250	296

Notes: Assume δ is known for both deconvolution cases. For a $n^{-\alpha}$ convergence rate, the effect sample size is calculated by $n^{2\alpha}$. Similarly, for a $(\ln(\ln n)/\ln n)^2$ convergence rate, it could be calculated as $(\ln n/\ln(\ln n))^{2*2}$.

Table 1.2: Smoothing Parameters of Some Popular Continuous Distributions

				Smoothness Parameters			
Name	Parameter	Density	Chara. Function	C_1	C_2	δ	T
Symm. Uniform	$a > 0$	$\frac{1}{2a}1_{[-a,a]}(x)$	$\frac{\sin(at)}{at}$	0^+	1	1	
Laplace	$b > 0$	$\frac{1}{2b}e^{-\frac{\lvert x\rvert}{b}}$	$\frac{1}{1+b^2t^2}$	$\frac{1}{b^2}$	$\frac{1}{b^2}$	2	1
Uniform	$a, b(b > a)$	$\frac{1}{2(b-a)}1_{[a,b]}(x)$	$\frac{e^{itb}-e^{ita}}{it(b-a)}$	$\frac{\lvert\cos(b)-\cos(a)\rvert}{b-a}$	$\frac{2}{b-a}$	1	
χ^2_k	$k > 0$	$\frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-\frac{x}{2}}$	$\frac{1}{(1-2it)^{k/2}}$	$\frac{1}{(2^{k/2})^+}$	1	$k/2$	1
Gamma	$k > 0, \theta > 0$	$\frac{1}{\Gamma(k)\theta^k}x^{k-1}e^{-\frac{x}{\theta}}$	$\frac{1}{(1-i\theta t)^k}$	$\frac{1}{(\theta^k)^+}$	$1(\theta > 1)$	k	$\frac{1}{\theta}$
Twice-convolved Laplace	$b > 0$	$\frac{1}{4b}e^{-\frac{\lvert x\rvert}{b}}(\lvert x\rvert + b)$	$\frac{1}{(1+b^2t^2)^2}$	$\frac{1}{4}$	1	4	$\frac{1}{b^2}$
Cauchy	$\mu = 0, \theta > 0$	$\frac{\theta}{\pi(\theta^2+x^2)}$	$e^{-\theta\lvert t\rvert}$	NA	NA	∞	
Normal	$\mu = 0, \sigma^2 > 0$	$\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$	$e^{-\frac{1}{2}\sigma^2t^2}$	NA	NA	∞	

Notes: The ordinary-smoothness parameter are defined by the Fan(1991): $C_1|\tau|^{-\delta} \leq |h_x(\tau)| \leq C_2|\tau|^{-\delta}$ for $|\tau| \geq T > 0$ where $0 < C_1 < C_2$, $\delta > 1$ and $h_x(\tau)$ is the characteristic function of the corresponding distribution. $\Gamma(s) = \int_0^\infty t^{s-1}e^{-t}dt$. The last two rows are from the super-smooth family.

Table 1.3: RMISE for Laplacian Noise Deconvolution

n	$\sigma_u^2/\sigma_v^2 = 2/1$	$\sigma_u^2/\sigma_v^2 = 2/2$	$\sigma_u^2/\sigma_v^2 = 2/4$
500	0.0162	0.0155	0.0204
1000	0.0150	0.0143	0.0197
3000	0.0138	0.0126	0.0190

Notes: Replication 500 times. $\frac{\sigma_u^2}{\sigma_v^2}$ stands for the signal-to-noise ratio

Table 1.4: RMISE under Misspecification: Normal Noise Deconvolution

n	$\sigma_u^2/\sigma_v^2 = 2/1$		$\sigma_u^2/\sigma_v^2 = 2/2$		$\sigma_u^2/\sigma_v^2 = 2/4$	
	CHP	Meister06	CHP	Meister06	CHP	Meister06
500	0.0155	0.0128	0.0186	0.0170	0.0242	0.0340
1000	0.0143	0.0116	0.0168	0.0156	0.0234	0.0337
3000	0.0129	0.0108	0.0152	0.0146	0.0230	0.0330

Notes: Replication 500 times. $\frac{\sigma_u^2}{\sigma_v^2}$ stands for the signal-to-noise ratio.

Table 1.5: Simulation by Rule-of-Thumb Adaptive Procedure with Laplace Noise

N	$\sigma_u^2/\sigma_v^2 = 2/1$			$\sigma_u^2/\sigma_v^2 = 2/2$			$\sigma_u^2/\sigma_v^2 = 2/4$		
	RMISE	Ave. δ	Ave. C_1	RMISE	Ave. δ	Ave. C_1	RMISE	Ave. δ	Ave. C_1
500	0.0139	2.02	0.10	0.0133	2.02	0.10	0.0244	2.04	0.10
1000	0.0125	2.00	0.10	0.0112	2.00	0.10	0.0231	2.08	0.10
3000	0.0110	2.00	0.10	0.0094	2.00	0.10	0.0221	2.00	0.10

Notes: Replication 100 times. $\frac{\sigma_u^2}{\sigma_v^2}$ stands for the signal-to-noise ratio.

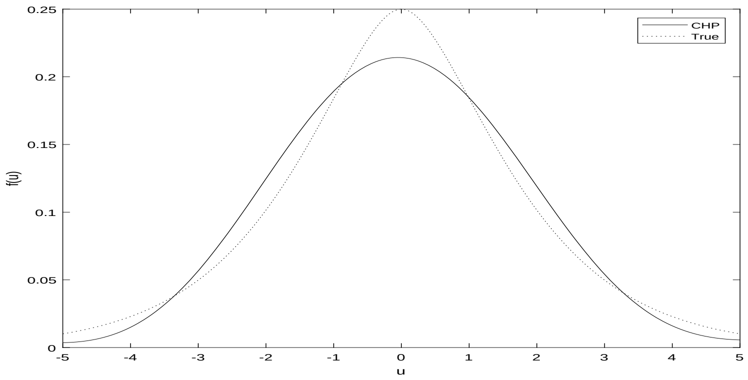


Figure 1.1: Laplace Deconvolution (CHP): $n = 500, \sigma_u^2/\sigma_v^2 = 2/2$

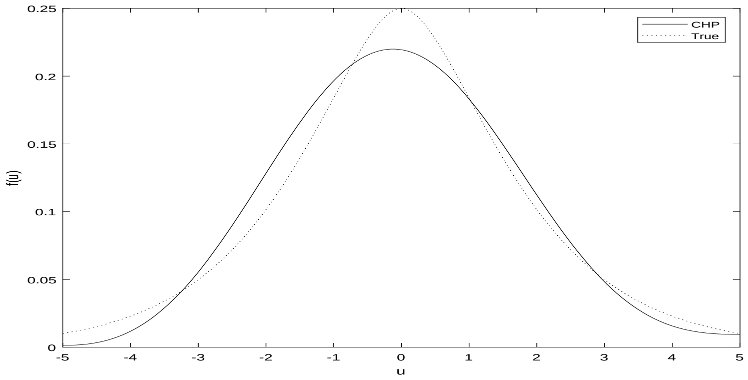


Figure 1.2: Laplace Deconvolution (CHP): $n = 1000, \sigma_u^2/\sigma_v^2 = 2/2$

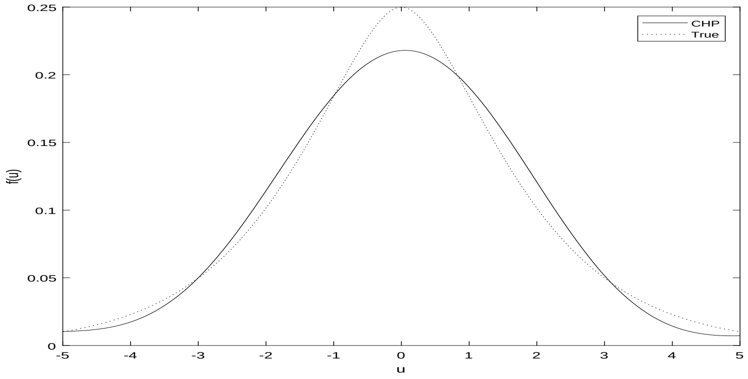


Figure 1.3: Laplace Deconvolution (CHP): $n = 3000, \sigma_u^2/\sigma_v^2 = 2/2$

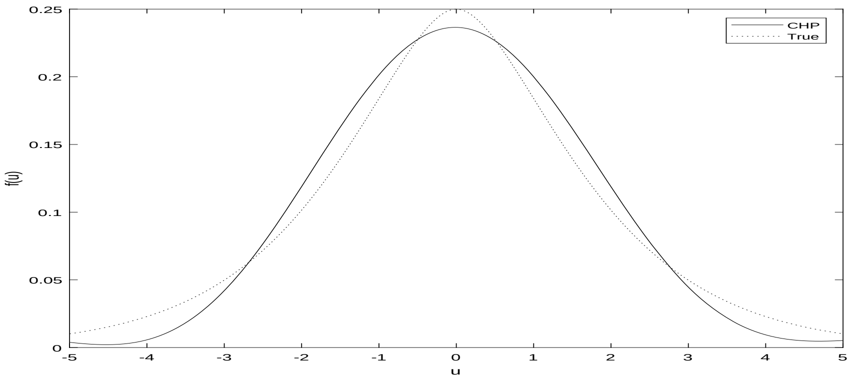


Figure 1.4: Laplace Deconvolution (CHP): $n = 1000, \sigma_u^2/\sigma_v^2 = 2/1$

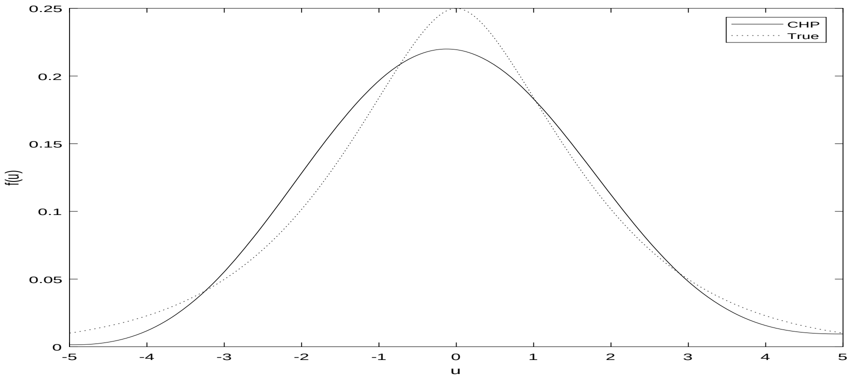


Figure 1.5: Laplace Deconvolution (CHP): $n = 1000, \sigma_u^2/\sigma_v^2 = 2/2$

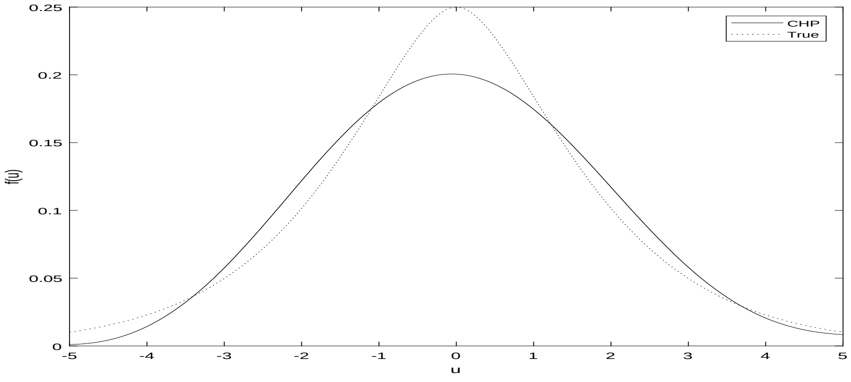


Figure 1.6: Laplace Deconvolution (CHP): $n = 1000, \sigma_u^2/\sigma_v^2 = 2/4$

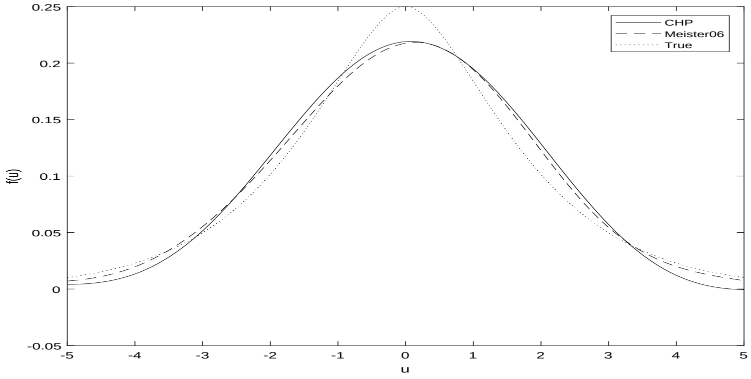


Figure 1.7: Misspecified Laplace (CHP) Deconvolution: $n = 500, \sigma_u^2/\sigma_v^2 = 2/2$

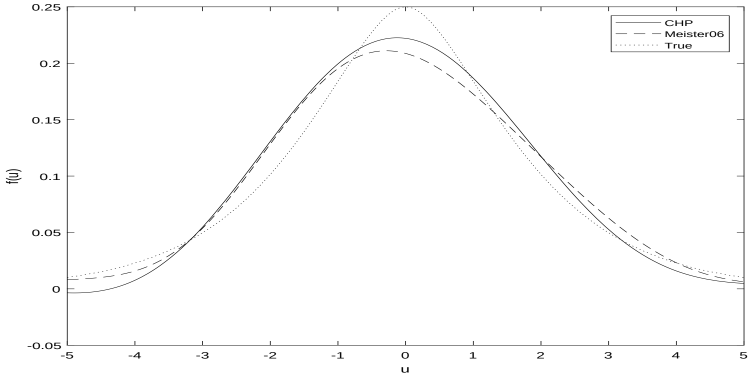


Figure 1.8: Misspecified Laplace (CHP) Deconvolution: $n = 1000, \sigma_u^2/\sigma_v^2 = 2/2$

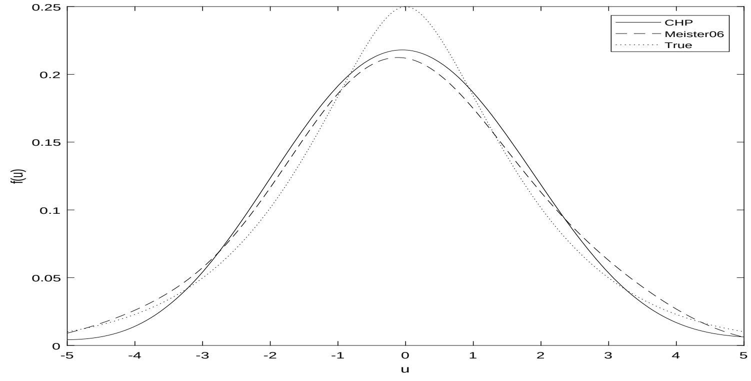


Figure 1.9: Misspecified Laplace (CHP) Deconvolution: $n = 3000, \sigma_u^2/\sigma_v^2 = 2/2$

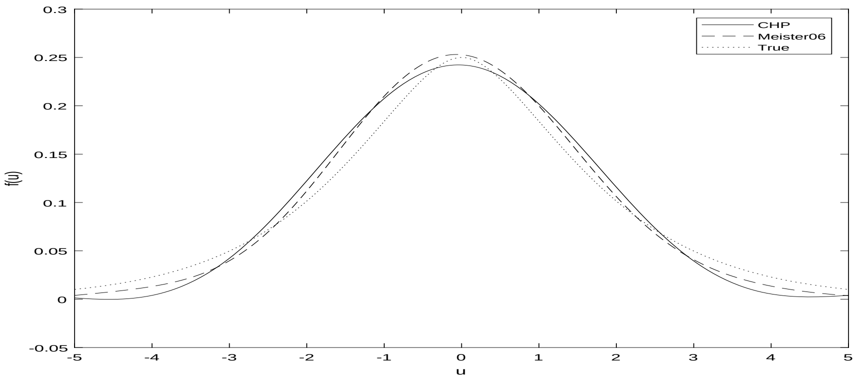


Figure 1.10: Misspecified Laplace (CHP) Deconvolution: $n = 1000, \sigma_u^2/\sigma_v^2 = 2/1$

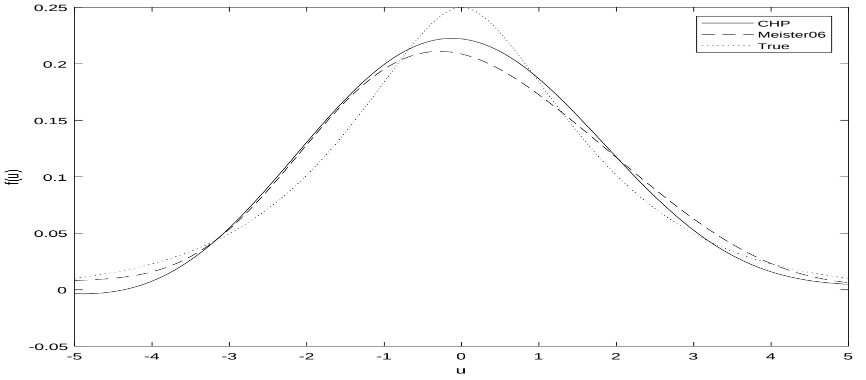


Figure 1.11: Misspecified Laplace (CHP) Deconvolution: $n = 1000, \sigma_u^2/\sigma_v^2 = 2/2$

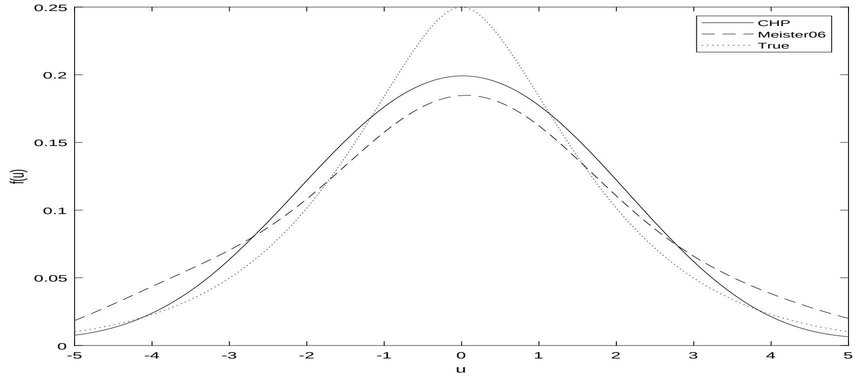


Figure 1.12: Misspecified Laplace (CHP) Deconvolution: $n = 1000, \sigma_u^2/\sigma_v^2 = 2/4$

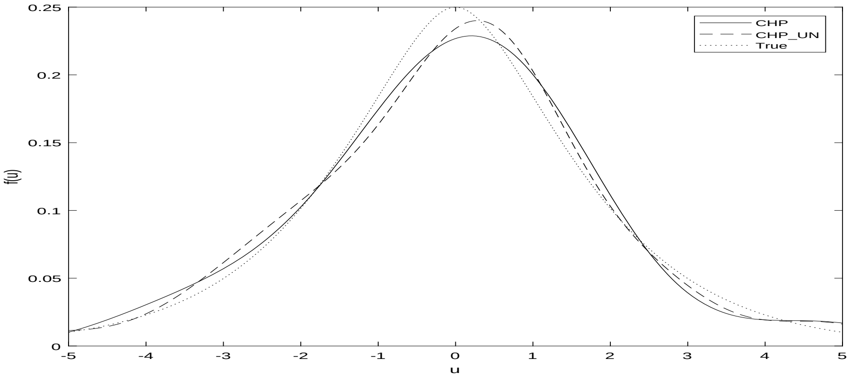


Figure 1.13: Deconvolution with Unknown Smooth Parameters, $n = 1000$, $\sigma_u^2/\sigma_v^2 = 2/2$

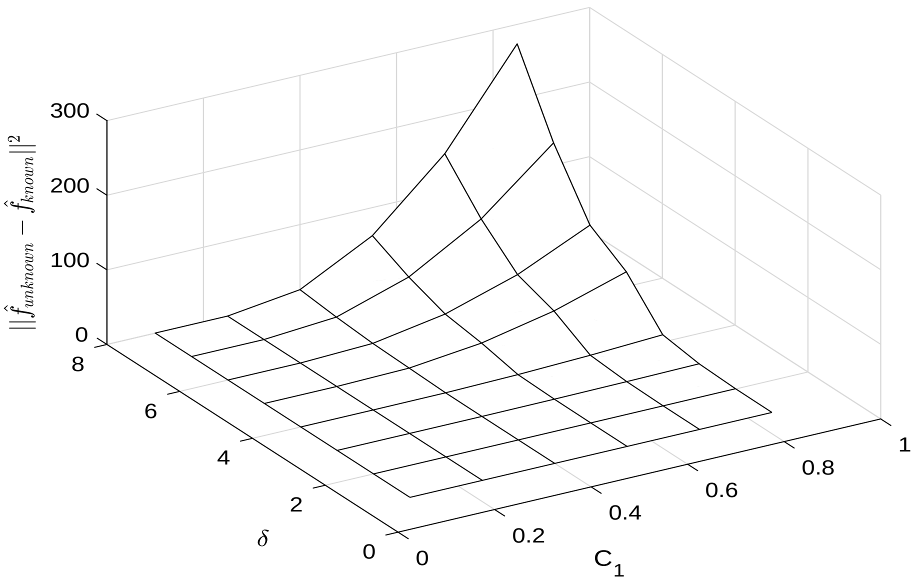


Figure 1.14: Euclidean Distance Between $\hat{f}_{unknown}$ and $\hat{f}_{known}$

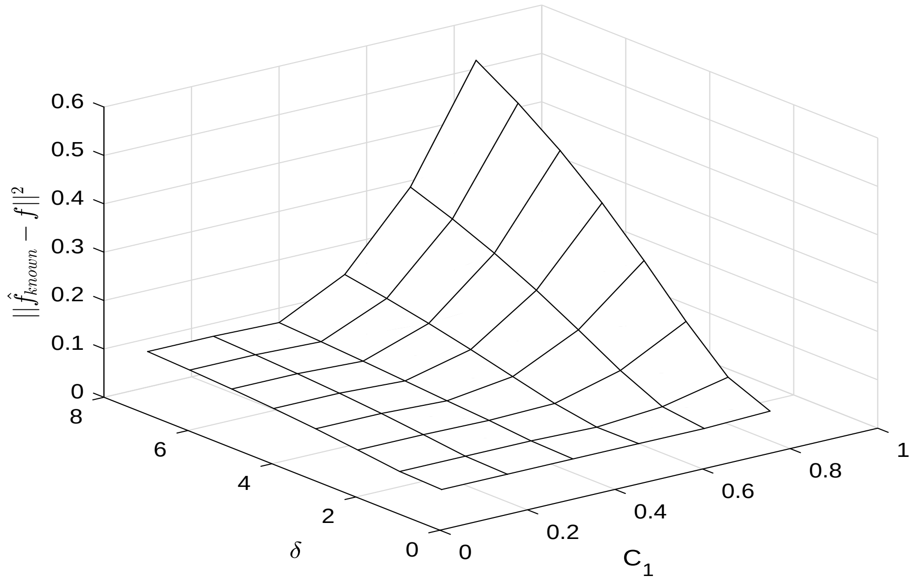


Figure 1.15: Euclidean Distance Between $\hat{f}_{known}$ and True Density

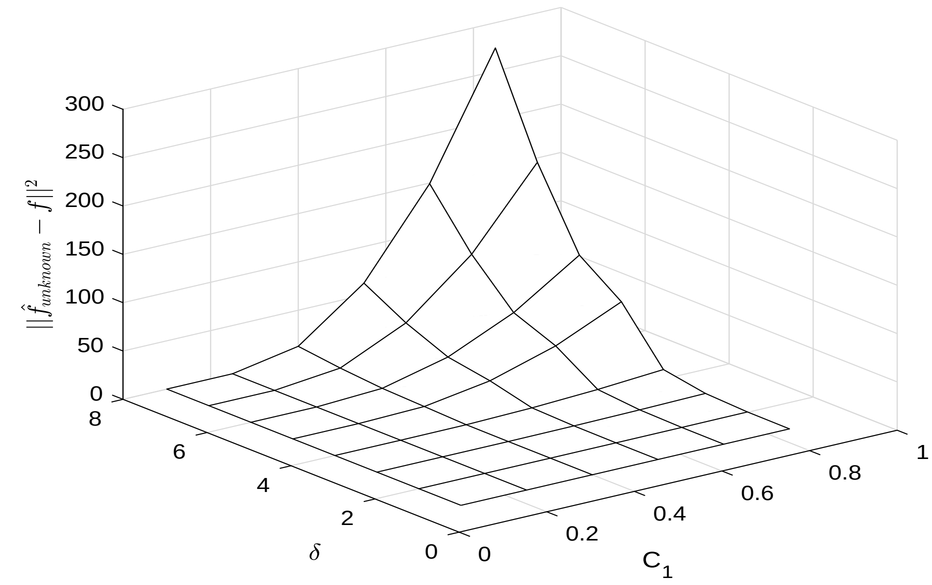


Figure 1.16: Euclidean Distance Between $\hat{f}_{unknown}$ and True Density

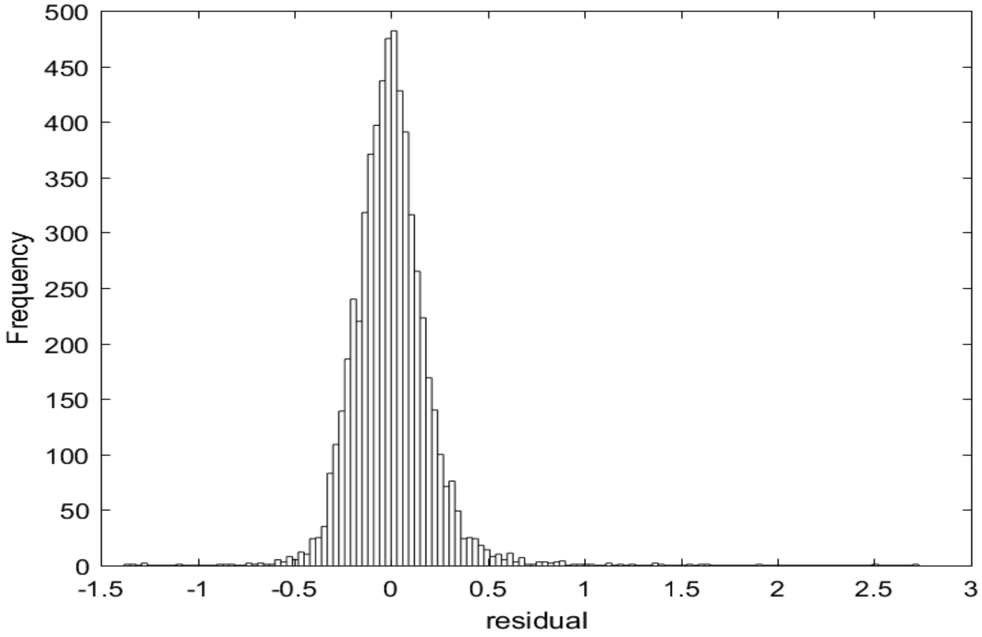


Figure 1.17: Histogram of the Residuals

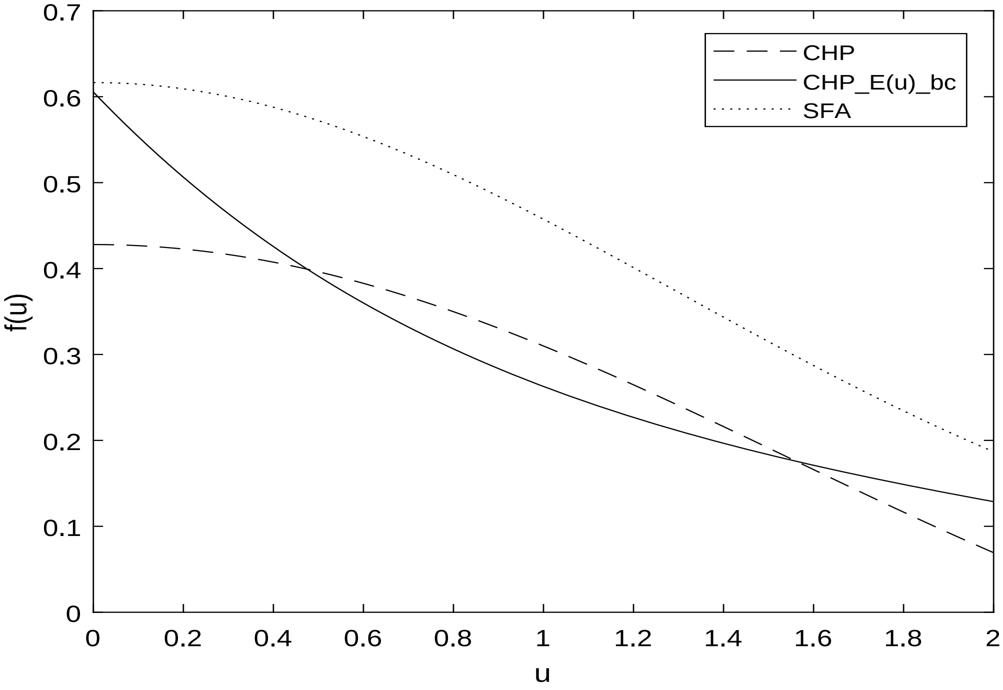


Figure 1.18: Estimated density of inefficiency

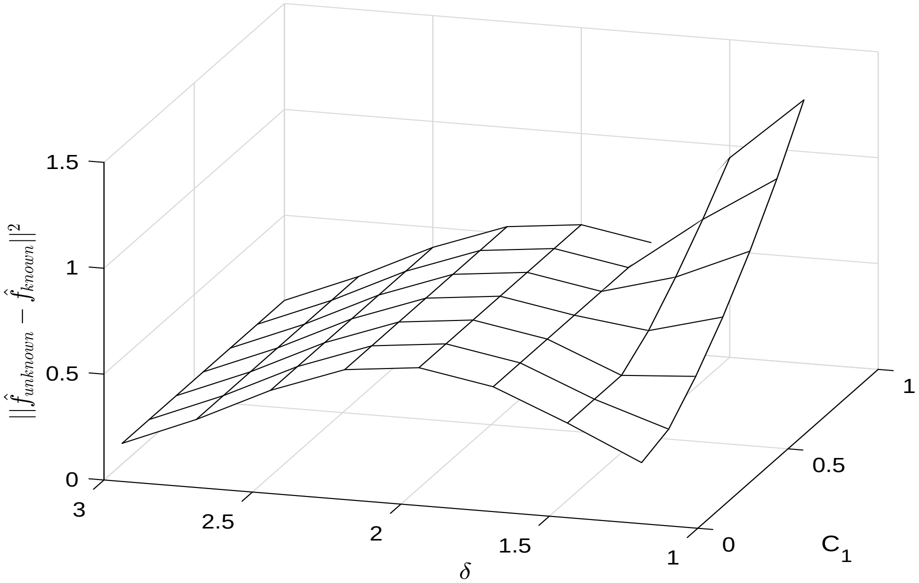


Figure 1.19: Euclidean Distance Between $\hat{f}_{unknown}$ and $\hat{f}_{known}$

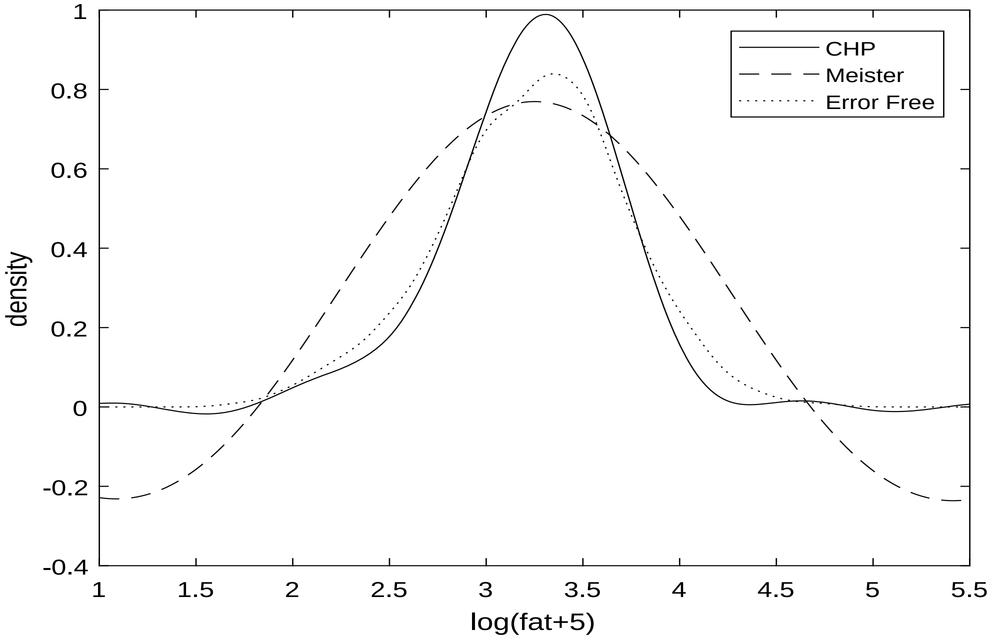


Figure 1.20: Density of the Logarithm of Daily Saturated Fat Intake, $C_1 = 1$, $\delta = 1.5$

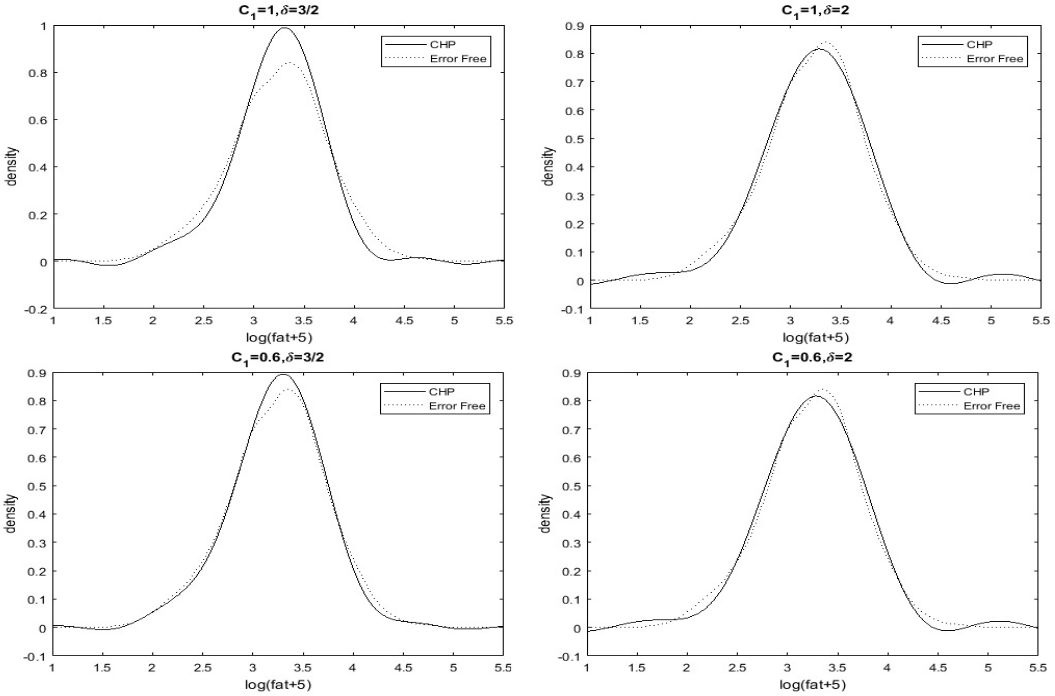


Figure 1.21: Saturated Fat Intake with Various Values of C_1 and δ

Appendices

1.A General Appendix

Definition: ε is ordinary-smooth of order δ (Fan 1991): characteristic function $\phi_\varepsilon(t)$ satisfies $d_0|t|^{-\delta} \leq |\phi_\varepsilon(t)| \leq d_1|t|^{-\delta}$ as $t \to \infty$.This is literally the same with the Assumption 3, just replacing $\phi_\varepsilon(t)$ with $h_\varepsilon(\tau)$.

A generalized result of Parseval's identity (or the Plancherel theorem) asserts that the integral of the square of the Fourier transform of a function is equal to the integral of the square of the function itself.

In one-dimension, for $f \in L_2(R)$,

$$\int_{-\infty}^{\infty} |\hat{f}(z)|^2 dz = \int_{-\infty}^{\infty} |f(\tau)|^2 d\tau$$

where $\hat{f}(z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\tau z} f(\tau) d\tau$ is the Fourier transform of the function $f(\tau)$. Specifically, we have

$$\int_{-\infty}^{\infty} |\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\tau z} f(\tau) d\tau|^2 dz = \int_{-\infty}^{\infty} |f(\tau)|^2 d\tau \quad (\star).$$

1.B Proof of Lemma 1

There is a N so that $w_n > T$ holds for all $n \geq N$. Hence the upper and lower bound of the Fourier Transform can be used. Similar to Lemma 1 in Meister(2006), using Parseval's identity and Fubini's theorem, we have:

$$\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} E_{f,g}||\hat{f}_u - f_u||_{L_2}^2 = \sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} E_{f,g}||\frac{1}{2\pi}\int_{-w_n}^{w_n} e^{-i\tau z}(1+\hat{b}_n^2\tau^2)\hat{h}_{\hat{\varepsilon}}d\tau - \frac{1}{2\pi}\int e^{-i\tau z}h_u(\tau)d\tau||^2$$

$$= \sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} E_{f,g}||\frac{1}{2\pi}\int_{-w_n}^{w_n} e^{-i\tau z}(1+\hat{b}_n^2\tau^2)\hat{h}_{\hat{\varepsilon}}d\tau - \frac{1}{2\pi}\int_{-w_n}^{w_n} e^{-i\tau z}h_u(\tau)d\tau - \frac{1}{2\pi}\int_{|\tau|>w_n} e^{-i\tau z}h_u(\tau)d\tau||^2$$

$$= \sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} E_{f,g}\int\Big(\big|\frac{1}{2\pi}\int_{-w_n}^{w_n} e^{-i\tau z}\big((1+\hat{b}_n^2\tau^2)\hat{h}_{\hat{\varepsilon}} - h_u(\tau)\big)d\tau\big|^2 + \big|\frac{1}{2\pi}\int_{|\tau|>w_n} e^{-i\tau z}h_u(\tau)d\tau\big|^2\Big)dz$$

$$= \sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} E_{f,g}\int\big|\frac{1}{2\pi}\int_{-w_n}^{w_n} e^{-i\tau z}\big(\hat{h}_{\hat{\varepsilon}}(1+\hat{b}_n^2\tau^2) - h_u(\tau)\big)d\tau\big|^2 dz + E_{f,g}\int\big|\frac{1}{2\pi}\int_{|\tau|>w_n} e^{-i\tau z}h_u(\tau)d\tau\big|^2 dz$$

By Fubini's Theorem, we could switch the order of integrals. That is

$$= \sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} E_{f,g}\int_{-w_n}^{w_n}\big|\frac{1}{2\pi}\int e^{-i\tau z}\big(\hat{h}_{\hat{\varepsilon}}(1+\hat{b}_n^2\tau^2) - h_u(\tau)\big)d\tau\big|^2 dz + E_{f,g}\int_{|\tau|>w_n}\big|\int\frac{1}{2\pi}e^{-i\tau z}h_u(\tau)d\tau\big|^2 dz$$

By Parseval's identity (let $f(\tau) = \hat{h}_{\hat{\varepsilon}}(1+\hat{b}_n^2\tau^2) - h_u(\tau)$ or $h_u(\tau)$ in equation $(\star)$), we have

$$\overset{Parseval}{=} \sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u}\big(\int_{-w_n}^{w_n} E_{f,g}|\hat{h}_{\hat{\varepsilon}}(1+\hat{b}_n^2\tau^2) - h_u(\tau)|^2 d\tau + \int_{|\tau|>w_n}|h_u(\tau)|^2 d\tau\big)$$

The expectation of the second integral over distribution family g, f is the integral itself as $h_u(\tau)$ is the characteristic function of the true distribution of u. Using $|A-B|^2 \leq 2A^2 + 2B^2$ and $\int_{|\tau|>w_n}\big|h_u(\tau)\big|^2 d\tau = 2\int_{w_n}^{\infty}\big|h_u(\tau)\big|^2 d\tau$, we have following inequality

$$\leq \sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} 2\int_{w_n}^{\infty}|h_u(\tau)|^2 d\tau + \sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} 2\int_{-w_n}^{w_n} E_{f,g}|(1+\hat{b}_n^2\tau^2)(\hat{h}_{\hat{\varepsilon}}(\tau) - h_{\varepsilon}(\tau))|^2 d\tau +$$

$$\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} 2\int_{-w_n}^{w_n} E_{f,g}|h_{\varepsilon}(\tau)/\frac{1}{(1+\hat{b}_n^2\tau^2)} - h_u(\tau)|^2 d\tau$$

The first term, which we call B, represents the bias which does not depend on the fact that the convoluted errors are estimated and can be bounded as in Lemma 1 of Meister (2006). The second term can be split into two pieces, V_1 and V_2, where V_1 is similar to V in Lemma 1 of Meister (2006) while V_2 is an additional component of the variance due to estimating the composed errors. Our third term, which we call E, can be found almost as that in Lemma 1 of Meister (2006) but the form of the bound is more complicated due to the fact that the empirical characteristic function used to construct the variance of the Laplace noise is constructed with $\hat{\varepsilon}$ instead of ε. The nonparametric regression in the first step impacts the convergence rate through the estimation of $\hat{\varepsilon}$.

The following proof is similar to Meister (2006) and Horrace and Parmeter (2011) except now we deal with Laplace noise and a nonparametric first-step regression estimator rather than just normal noise for the linear (stochastic frontier) model. There are three steps to the proof.

(1) $B \leq const \times w_n^{1-2\delta}$ by Assumption(3) $C_1|\tau|^{-\delta} \leq |h_u(\tau)| \leq C_2|\tau|^{-\delta}$ where $0 < C_1 < C_2$ and $\delta > 1$.

(2) By assumption 5,

$$\hat{h}_{\hat{\varepsilon}}(\tau) = \Big|\frac{1}{n}\sum_{j=1}^{n} e^{i\tau\hat{\varepsilon}_j}\Big| = \Big|\frac{1}{n}\sum_{j=1}^{n} e^{i\tau\varepsilon_j}(1+O_p(\tau n^{-a}))\Big| = (1+O_p(\tau n^{-a}))\hat{h}_{\varepsilon}(\tau)$$

where $a = \frac{2}{4+q}$ for the nonparametric first-step regression and $a = 0.5$ for parametric first-step regression, e.g, translog in the stochastic frontier model. We focus on the parametric setting hereafter for the main formulas and lay out the details of the differences when first-step nonparametric regression is implemented.[36] So

$$\hat{h}_{\hat{\varepsilon}}(\tau) = |\frac{1}{n}\sum_{j=1}^{n} e^{i\tau\hat{\varepsilon}_j}| = |\frac{1}{n}\sum_{j=1}^{n} e^{i\tau\varepsilon_j}(1+O_p(\tau n^{-1/2}))| = (1+O_p(\tau n^{-1/2}))\hat{h}_{\varepsilon}(\tau)$$

Let $A(\hat{h}_{\varepsilon}) = \int_{-w_n}^{w_n} E_{f,g}|\hat{h}_{\varepsilon}(\tau) - h_{\varepsilon}(\tau)|^2 d\tau = \int_{-w_n}^{w_n} E_{f,g}|\frac{1}{n}\sum_{j=1}^{n} e^{i\tau\varepsilon_j} - E(e^{i\tau\varepsilon})|^2 d\tau = O_p(n^{-1}w_n)$,

[36] Basically, there is $2a$ instead of 1 in the power of n.

$$\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} 2\int_{-w_n}^{w_n} E_{f,g}(1+\hat{b}_n^2\tau^2)^2|\hat{h}_{\hat{\varepsilon}}(\tau)-h_\varepsilon(\tau)|^2 d\tau \le 4(1+\hat{b}_n^2 w_n^2)^2 \sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u}\int_{-w_n}^{w_n}[E_{f,g}|\hat{h}_{\hat{\varepsilon}}(\tau)-\hat{h}_\varepsilon(\tau)|^2+$$

$$E_{f,g}|\hat{h}_\varepsilon(\tau)-h_\varepsilon(\tau)|^2]d\tau$$

$$=4(1+\hat{b}_n^2 w_n^2)^2\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u}\int_{-w_n}^{w_n} E_{f,g}|\hat{h}_{\hat{\varepsilon}}(\tau)-\hat{h}_\varepsilon(\tau)|^2 d\tau + 4(1+\hat{b}_n^2 w_n^2)^2 A(\hat{h}_\varepsilon)$$

$$\le 4(1+b_n^2 w_n^2)^2\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u}\int_{-w_n}^{w_n}\tau^2 E_{f,g}(n^{-1}\sum_{j=1}^n|\hat{\varepsilon}_j-\varepsilon_j|)^2 d\tau + 4(1+b_n^2 w_n^2)^2 A(\hat{h}_\varepsilon) = V_1+V_2$$

where $V_1 \le const \times (n^{-1}w_n^3)(1+b_n^2w_n^2)^2$ and $V_2 \le const \times (n^{-1}w_n)(1+b_n^2w_n^2)^2$.

(3) Similar to Lemma 1 in Meister (2006), for the last term we can derive:

$$E=\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} 2\int_{-w_n}^{w_n} E_{f,g}|\frac{h_u(\tau)(1+\hat{b}_n^2\tau^2)}{1+b^2\tau^2}-h_u(\tau)|^2 d\tau$$

$$\overset{\tau=sw_n}{=}\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} 2\int_{-1}^{1} E_{f,g}|\frac{s^2w_n^2(\hat{b}_n^2-b^2)}{1+b^2s^2w_n^2}|^2|h_u(sw_n)|^2 w_n ds,$$

where

$$\begin{aligned}E_{f,g}|\frac{s^2w_n^2(\hat{b}_n^2-b^2)}{1+b^2s^2w_n^2}|^2 &= E_{f,g}|\frac{s^2w_n^2(\hat{b}_n^2-b^2)}{1+b^2s^2w_n^2}|^2\chi(|\hat{b}_n^2-b^2|\le d_n)+E_{f,g}|\frac{s^2w_n^2(\hat{b}_n^2-b^2)}{1+b^2s^2w_n^2}|^2\chi(|\hat{b}_n^2-b^2|>d_n)\\ &\le |\frac{s^2w_n^2d_n}{1+s^2w_n^2b^2}|^2+|\frac{s^2w_n^2b_n^2}{1+b^2w_n^2s^2}|^2 Pr(|\hat{b}_n^2-b^2|>d_n)\\ &\le (\frac{s^2w_n^2d_n}{s^2w_n^2b^2})^2+(\frac{s^2w_n^2b_n^2}{s^2b^2w_n^2})^2 Pr(|\hat{b}_n^2-b^2|>d_n)\\ &\le (\frac{d_n}{b^2})^2+(\frac{b_n^2}{b^2})^2 Pr(|\hat{b}_n^2-b^2|>d_n)\\ &= const\times w_n^{-2}+const\times(b_n^2)^2 Pr(|\hat{b}_n^2-b^2|>d_n)\end{aligned}$$

Where the last inequality for the first term comes from the fact that $d_n = O(w_n^{-1})$.

1.C Proof of Lemma 2

Let d_n and f, g be the same as in Lemma 1, the term $\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} P_{f,g}(|\hat{b}_n^2 - b^2| > d_n)$ is bounded by two addends. We derive an upper bound for each of them. First,

$$\begin{aligned}
\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} P_{f,g}(\hat{b}_n^2 - b^2 > d_n) &= \sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} P_{f,g}(k_n^{-2}(\frac{C_1 k_n^{-\delta}}{\hat{h}_{\hat{\varepsilon}}(k_n)} - 1) > d_n + b^2) \\
&= \sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} P_{f,g}((\frac{C_1 k_n^{-\delta}}{\hat{h}_{\hat{\varepsilon}}(k_n)} - 1) > d_n k_n^2 + b^2 k_n^2) \\
&= P_{f,g}(|\hat{h}_{\hat{\varepsilon}}(k_n)| < \frac{C_1 k_n^{-\delta}}{1 + d_n k_n^2 + b^2 k_n^2}) \\
&\leq \sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} P_{f,g}(|\hat{h}_{\hat{\varepsilon}}(k_n)| < \alpha_n \frac{C_1 k_n^{-\delta}}{1 + b^2 k_n^2}) \\
&= \sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} P_{f,g}(\hat{h}_{\hat{\varepsilon}}(k_n) < \alpha_n |h_\varepsilon(k_n)|)
\end{aligned}$$

where $\alpha_n = \frac{1+b^2k_n^2}{1+d_nk_n^2+b^2k_n^2}$, hence, $\alpha_n \to 0$ as $d_n = w_n^{-1} = O(k_n^{-1})$, $d_n k_n^2 = O(k_n)$ for known δ and C_1 case and $d_n = w_n^{-1} = O(lnk_n/k_n), d_n k_n^2 = O(ln(k_n)k_n)$ for other cases.[37]

There exists a constant $c \in (0, 1)$ that guarantees that the above formula is bounded above by $\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} P_{f,g}(\hat{h}_{\hat{\varepsilon}}(k_n) < \alpha_n|h_\varepsilon(k_n)|) \leq \sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} P_{f,g}(\hat{h}_{\hat{\varepsilon}}(k_n) < c|h_\varepsilon(k_n)|)$ which by Chebyshev's inequality yields

$$\begin{aligned}
\leq&(1-c)^{-2}\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} |h_\varepsilon(k_n)|^{-2} E_\varepsilon|\hat{h}_{\hat{\varepsilon}}(k_n) - h_\varepsilon(k_n)|^2 \\
\leq&2(1-c)^{-2}\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} |h_\varepsilon(k_n)|^{-2}[E_\varepsilon|\hat{h}_{\hat{\varepsilon}}(k_n) - \hat{h}_\varepsilon(k_n)|^2 + E_\varepsilon|\hat{h}_\varepsilon(k_n) - h_\varepsilon(k_n)|^2] \\
\leq&2(1-c)^{-2}\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} |h_\varepsilon(k_n)|^{-2}[O_p(k_n n^{-1})E_\varepsilon|\frac{1}{n}\sum_j exp(ik_n\varepsilon_j)|^2 + E_\varepsilon|\frac{1}{n}\sum_j exp(ik_n\varepsilon_j) - h_\varepsilon(k_n)|^2] \\
=&const \times (E_1 + E_2),
\end{aligned}$$

[37] This is discussed in Section 1.4.

where the first term is bounded by $|h_\varepsilon(k_n)|^{-2} \leq k_n^{2\delta+2}(1+b_n^2k_n^2)$ as that for V_1; $E_1 \leq const \times k_n^{2\delta+2}(1+b_n^2k_n^2)n^{-1}$ and $E_2 \leq const \times k_n^{2\delta}(1+b_n^2k_n^2)n^{-1}$ are similar to that in Lemma 2 of Meister (2006).

The second addend can be bounded in a similar way:

$$\begin{aligned}\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} P_{f,g}(\hat{b}_n^2 - b^2 < -d_n) &= \sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} P_{f,g}(k_n^{-2}(\frac{C_1k_n^{-\delta}}{\hat{h}_{\hat{\varepsilon}}(k_n)} - 1) < b^2 - d_n)\\ &\leq \sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} P_{f,g}(\hat{h}_{\hat{\varepsilon}}(k_n) > \gamma_n|h_\varepsilon(k_n)|)\end{aligned}$$

where $\gamma_n = \frac{1+b^2k_n^2}{1+b^2k_n^2-d_nk_n^2}$, hence, $\gamma_n \to 1^+$ as $d_n = w_n^{-1} = O(k_n^{-1})$, $d_nk_n^2 = O(k_n)$ for known δ and C_1 case and $d_n = w_n^{-1} = O(lnk_n/k_n), d_nk_n^2 = O(ln(k_n)k_n)$ for other cases.

Again there exists a constant $C \in (0,1)$ that guarantees the above formula is bounded above by $\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} P_{f,g}(\hat{h}_{\hat{\varepsilon}}(k_n) > \gamma_n|h_\varepsilon(k_n)|) \leq \sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} P_{f,g}(\hat{h}_{\hat{\varepsilon}}(k_n) > C|h_\varepsilon(k_n)|)$ which by Chebyshev's inequality yields

$$\begin{aligned}&\leq (C-1)^{-2}\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u}|h_\varepsilon(k_n)|^{-2}E_\varepsilon|\hat{h}_{\hat{\varepsilon}}(k_n) - h_\varepsilon(k_n)|^2\\ &\leq 2(C-1)^{-2}\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u}|h_\varepsilon(k_n)|^{-2}[E_\varepsilon|\hat{h}_{\hat{\varepsilon}}(k_n) - \hat{h}_\varepsilon(k_n)|^2 + E_\varepsilon|\hat{h}_\varepsilon(k_n) - h_\varepsilon(k_n)|^2],\end{aligned}$$

which leads to the same upper bound as derived for the first addend.

1.D Proof of Theorem 1

Combining the results from Lemma 1 and 2, we obtain the upper bound of MISE of the density $\hat{f}_u$ as

$$\begin{aligned}\max(B,V,E) = \max\Big[& const_1 \times w_n^{1-2\delta}, const_2 \times n^{-1}w_n(1+b^2w_n^2)^2 + const_3 \times n^{-1}(1+b_n^2w_n^2)^2w_n^3,\\ & const_4 \times w_n^{1-2\delta}w_n^{-2} + const_5 \times k_n^{2\delta}b_n^4(1+b_n^2k_n^2)(1+k_n^2)n^{-1}\Big]\end{aligned}$$

Under Assumption 3, if C_1 and δ are known, then $w_n = k_n$, $b_n^2 = 0.5V(\hat{\varepsilon}\,)$, and collecting the leading maximum terms leads to

$$\max(B, V, E) = \max\Bigg[const_1 \times w_n^{1-2\delta}, const_2 \times \frac{b^4 w_n^5}{n} + const_3 \times \frac{b_n^4 w_n^7}{n}, \\ const_4 \times w_n^{-1-2\delta} + const_5 \times \frac{w_n^{2\delta+4} b_n^6}{n}\Bigg]$$

Observe that $w_n^{-1-2\delta} < w_n^{1-2\delta}$ and $w_n^7 > w_n^5$. Comparing the order of $const_3 \times \frac{b_n^4 w_n^7}{n}$ and $const_5 \times \frac{w_n^{2\delta+4} b_n^6}{n}$ leads to the cutoff value $\delta = 3/2$. Note that $w_n \to \infty$ as $n \to \infty$, and $w_n^{-a} + \frac{w_n^b}{n} \geq 2\sqrt{\frac{w_n^{b-a}}{n}}$ with equality holding when $w_n = n^{\frac{1}{a+b}}$, minimizing the above piecewise maximum leads to following two cases

(i) If $1 < \delta \leq 1.5$, $k_n = n^{\frac{1}{2\delta+6}} (b_n^2)^{\frac{-1}{2\delta+6}}$. Consequently, a $n^{-\frac{(2\delta-1)}{2\delta+6}} (b_n^2)^{\frac{(2\delta-1)}{2\delta+6}} \to D_1 n^{-\frac{(2\delta-1)}{2\delta+6}}$ convergence rate is determined by the equality of the first term and the second addend of the second term where $D_1 = V(\varepsilon)^{\frac{(2\delta-1)}{2\delta+6}}$.

(ii) If $\delta > 1.5$, $k_n = n^{\frac{1}{4\delta+3}} (b_n^2)^{\frac{-4}{4\delta+3}}$. Consequently, a $n^{-\frac{2\delta-1}{4\delta+3}} (b_n^2)^{\frac{(2\delta-1)}{4\delta+3}} \to D_2 n^{-\frac{2\delta-1}{4\delta+3}}$ convergence rate is determined by the equality of the first term and the second addend of the third term where $D_2 = V(\varepsilon)^{\frac{(2\delta-1)}{4\delta+3}}$. We exclude the case with $\delta = 2$ (Laplace-Laplace convolution) here as $\hat{b}_n^2 < b_n^2 = 0.5V(\hat{\varepsilon})$ converges to $\min\{V(u), V(v)\}$ which cannot distinguish the target from the noise[38].

1.E Proof of Theorem 2

For the case where C_1 and δ are unknown, similar argument applies, k_n stays the same with guess C_1 and δ since $w_n = k_n / \ln k_n = O(k_n)$ and the convergence rates are $n^{-\frac{(2\delta-1)}{2\delta+6}} (\ln n)^{\frac{(2\delta-1)}{2\delta+6}}$ if $1 < \delta \leq 1.5$ and $n^{-\frac{2\delta-1}{4\delta+3}} (\ln n)^{\frac{2\delta-1}{4\delta+3}}$ if $\delta > 1.5$.[39]

[38] This is a rare case related to identification given that $\delta = 2$ is negligible in the range of $\delta > 1$ but it does not impact the estimation

[39] See the rule-of-thumb adaptive procedure in section 4.1.

1.F Proof of Theorem 3

When nonparametric kernel estimation is implemented for the first-step regression, we can easily derive similar Lemmas (as those for the parametric case) as follows:

Lemma $1'$**.** *For Assumptions 3-5, Condition 2.1 in Li and Racine (2007) and* $\mathcal{L}_n = \{Laplace(0,b) : b^2 \in (0, b_n^2]\}$*, the MISE of (1.7) is*

$$\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} E_{f,g}||\hat{f}_u - f_u||^2_{L_2} \le B+V+E,$$

where $B \le const_1 \times w_n^{1-2\delta}$*,*
$V \le const_2 \times n^{-2a}w_n(1+b_n^2w_n^2)^2 + const_3 \times n^{-2a}w_n^3(1+b_n^2w_n^2)^2$ *,*
$E \le const_4 \times \sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u}\Big(w_n\int_{-1}^{1}|h_u(sw_n)|^2\left(\frac{d_n}{b^2}\right)^2 ds + w_n\int_{-1}^{1}|h_u(w_ns)|^2\frac{b_n^4}{b^4}\times P_{f,g}(|\hat{b}_n^2 - b^2| > d_n)ds\Big)$*, with* $d_n := \frac{1}{w_n}$*;* f *and* g *are the probability density function in distribution family* $\mathcal{F}_u$ *and* $\mathcal{L}_n$ *respectively.* $const_j$ *are positive constants for* $j = 1, 2, 3, 4$*.*

Lemma $2'$**.** *Let* d_n *and* f, g *be the same as in Lemma 3, then* $\sup_{g\in\mathcal{L}_n}\sup_{f\in\mathcal{F}_u} P_{f,g}(|\hat{b}_n^2 - b^2| > d_n) \le const \times k_n^{2\delta}(1+b_n^2k_n^2)(1+k_n^2)n^{-2a}$*.*

Then by a parallel argument, combining the results from Lemma 1 and 2, we can obtain the upper bound of MISE of the density $\hat{f}_u$ as

$$\max(B,V,E) = \max\Big[const_1 \times w_n^{1-2\delta}, const_2 \times n^{-2a}w_n(1+b^2w_n^2)^2 + const_3 \times n^{-2a}(1+b_n^2w_n^2)^2w_n^3,$$
$$const_4 \times w_n^{1-2\delta}w_n^{-2} + const_5 \times k_n^{2\delta}b_n^4(1+b_n^2k_n^2)(1+k_n^2)n^{-2a}\Big]$$

Under Assumption 3, if C_1 and δ are known, then $w_n = k_n$, $b_n^2 = 0.5V(\hat{\varepsilon}\,)$, and minimizing the above maximum leads to

(i) If $1 < \delta \le 1.5$, $k_n = n^{\frac{2a}{2\delta+6}} \times (b_n^2)^{\frac{-1}{2\delta+6}}$. Consequently, an $n^{-\frac{2a(2\delta-1)}{2\delta+6}} \times (b_n^2)^{\frac{2a(2\delta-1)}{2\delta+6}} \to D_1 \times n^{-\frac{2a(2\delta-1)}{2\delta+6}}$ convergence rate is determined by the equality of the first term and the first addend of the third term where $D_1 = V(\varepsilon)^{\frac{2a(2\delta-1)}{2\delta+6}}$.

(ii) If $\delta > 1.5$, $k_n = n^{\frac{2a}{4\delta+3}} \times (b_n^2)^{\frac{-4}{4\delta+3}}$. Consequently, an $n^{-\frac{2a(2\delta-1)}{4\delta+3}} \times (b_n^2)^{\frac{2a(2\delta-1)}{4\delta+3}} \to D_2 \times n^{-\frac{2a(2\delta-1)}{4\delta+3}}$ convergence rate is determined by the equality of the first term and the second addend of the third term where $D_2 = V(\varepsilon)^{\frac{2a(2\delta-1)}{4\delta+3}}$. We exclude the case with $\delta = 2$ (Laplace-Laplace convolution) here as similar reasoning in the proof of Theorem 1 applies.

Bibliography

[1] Butucea, C. and C. Matias, Minimax Estimation of the Noise Level and of the Deconvolution Density in a Semiparametric Convolution Model. Bernoulli. 2005, pp. 309-340.

[2] Butucea, C., C. Matias, and C. Pouet, Adaptivity in Convolution Models with Partially Known Noise Distribution. Electronic Journal of Statistics. 2008, pp. 897-915.

[3] Cao, C. D. Linear Regression with Laplace Measurement Errors. Master Thesis. Department of Statistic, Kansas State University.

[4] Carroll, R.J., D. Ruppert, L. A. Stefanski, and C. M., Crainiceanu. Measurement Error in Nonlinear Models: A Modern Perspective, volume 2nd Ed. CRC Press, 2006.

[5] Delaigle, A. and I. Gijbels. Practical bandwidth selection in deconvolution kernel density estimation. Computational Statistics & Data Analysis. 2004. 45 (2) pp. 249-267.

[6] Evdokimov, K., Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity. 2010, Working paper version.

[7] Fan, J., On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems. Annals of Statistics. 1991, pp. 1257-1272.

[8] Feng, G. H., and A. Serletis, Efficiency and Productivity of the US Banking Industry,1998-2005: Evidence from the Fourier Cost Function Satisfying Global Regularity Conditions. Journal of Applied Econometrics. 2009, 24, pp.105-138.

[9] Greene, W. H., A gamma-distributed stochastic frontier model, Journal of Econometrics. 1990, 46(1-2), 141–164.

[10] Greene, W. H., Simulated likelihood estimation of the normal-gamma stochastic frontier function, Journal of Productivity Analysis. 2003, 19(2), 179–190.

[11] Henderson, D. J. and C. F. Parmeter. Applied Nonparametric Econometrics. 2015, Cambridge University Press.

[12] Hesse, C. H., Data-driven Deconvolution. Nonparametric Statistics. 1999, 10, pp. 343-373.

[13] Horowitz, J. L., and M. Markatou, Semiparametric Estimation of Regression Models for Panel Data. Review of Economics Studies. 1996, 63, pp. 145-168.

[14] Horrace, W. C., Moments of the Truncated Normal Distribution. Journal of Productivity Analysis. 2015, 43, pp. 133-138.

[15] Horrace, W. C., and C.F. Parmeter, Semiparametric Deconvolution with Unknown Error Variance. Journal of Productivity Analysis. 2011, 35, pp. 129-141.

[16] Horrace, W. C., and C.F. Parmeter, A Laplace Stochastic Frontier Model. Econometric Reviews. 2018, 37, pp. 260-280.

[17] Jirak, M., A. Meister, and M., Reiss. Adaptive function estimation in nonparametric regression with one-sided errors. Annals of Statistics. 2014, 42(5), pp. 1970-2002.

[18] Jondrow, J., C. K. Lovell, I. S. Materov, and P. Schmidt, On the Estimation of Technical Inefficiency in the Stochastic Frontier Model. Journal of Econometrics. 1982, 19, pp. 233-238.

[19] Jones, D.Y., et al. Dietary fat and breast cancer in the National Health and Nutrition Examination Survey I Epidemiologic Follow-up Study. J Natl Cancer Institute. 1987. 79(3), pp. 465-71.

[20] Johannes, J., Deconvolution with Unknown Error Distribution. Annals of Statistics. 2009, 37, pp. 2301-2323.

[21] Ju, G., L. Gan, and Q. Li. Nonparametric panel estimation of labor supply. Journal of Business & Economic Statistics. 2019, 37(2), pp 260-274

[22] Koul, H.L., and W.X. Song. Simulation Extrapolation Estimation in Parametric Models with Laplace Measurement Error. Electronic Journal of Statistics. 2014, 8, pp. 1973-1995.

[23] Kneip, A., L. Simar, and I. Van Keilegom. Frontier estimation in the presence of measurement error with unknown variance. Journal of Econometrics. 2015, 184(2), 379-393.

[24] Kumbhakar, S. C. and E. G. Tsionas, Measuring Technical and Allocative Inefficiency in the Translog Cost System: A Bayesian Approach. Journal of Econometrics. 2005, 126, pp. 355-384.

[25] Kumbhakar S. C., B.U. Park, L. Simar, E.G. Tsionas. Nonparametric stochastic frontiers: a local likelihood approach. Journal of Econometrics. 2007, 137(1), pp. 1–27.

[26] Kumbhakar, S.C., C.F. Parameter, and V. Zelenyuk. Stochastic Frontier Analysis: Foundations and Advances. CEPA Working Papers Series WP022018, School of Economics, University of Queensland, Australia.

[27] Li, Q. and J.S. Racine. Nonparametric Econometrics: Theory and Practice. 2007, Princeton University Press.

[28] Meeusen, W. and J. van den Broeck. Efficiency estimation from Cobb-Douglas production functions with composed error, International Economic Review. 1977, 18(2), pp. 435–444.

[29] Meister, A., On the Effect of Misspecifying the Error Density in a Nonparametric Deconvolution Problem. Canadian Journal of Statistics. 2004, 32, pp. 439-449.

[30] Meister, A., Density Estimation with Normal Measurement Error with Unknown Variance. Statistica Sinica. 2006, 16, pp. 195-211.

[31] Meister, A., Deconvolution Problems in Nonparametric Statistics. 2009. Lecture Notes in Statistics 193. Springer, Berlin

[32] Neumann, M. H., On the effect of Estimating the Error Density in Nonparametric Deconvolution. Nonparametric Statistics. 1997, 7, pp. 307-330.

[33] Nguyen, N. B., Estimation of technical efficiency in stochastic frontier analysis. 2010, PhD thesis, Bowling Green State University.

[34] Parmeter, C. F., H. J. Wang, and S. C. Kumbhakar. Nonparametric estimation of the determinants of inefficiency. Journal of Productivity Analysis. 2017, 47, pp. 205-221.

[35] Parzen, E., On Estimation of A Probability Density Function and Mode. Annals of Math Statistics. 1962, 35, pp. 1065-1076.

[36] Ritter, C. and L. Simar. Pitfalls of normal-gamma stochastic frontier models. Journal of Productivity Analysis. 1997, 8(2), pp. 167–182.

[37] Simar, L., I.Van Keilegom, and V. Zelenyuk. Nonparametric least squares methods for stochastic frontier models. Journal of Productivity Analysis. 2017, 47, pp. 189-204.

[38] Song, W.X., J.H. Shi, and C.X. Zhang. Linear errors-in-variables regression and tweedie's formula. 2016. Working paper

[39] Stevenson, R. Likelihood functions for generalized stochastic frontier estimation. Journal of Econometrics. 1980, 13(1), pp. 58–66.

[40] Tran, K. C. and E. G. Tsionas. Local GMM estimation of semiparametric panel data with smooth coefficient models. Econometric Reviews. 2010, 29, pp. 39–61.

[41] Wang, W. S., C., Amsler, and P., Schmidt, Goodness of fit tests in stochastic frontier models. Journal of Productivity Analysis. 2011, 35, pp. 95-118.

[42] Wang, W.,S., and P., Schmidt, On the Distribution of Estimated Technical Efficiency in Stochastic Frontier Models. Journal of Econometrics. 2009, 148, pp. 36-45.

[43] Wang, X. F., and D., Ye, The Effects of Error Magnitude and Bandwidth Selection for Deconvolution with Unknown Error Distribution. Journal of Nonparametric Statistics. 2012, 24, pp. 153–167.

[44] Zhang, S., and R. J., Karunramuni, Boundary Bias correction for Nonparametric Deconvolution. Annals of Institute Statistic Mathematics. 2000, 52, pp. 612-629.

Chapter 2

Penalized Sieve Density Estimation with An application to Zero-Inefficiency Stochastic Frontiers

Jun Cai[1], William C. Horrace[2], Christopher F. Parmeter[3]

2.1 Introduction

The notion of the frontier arises naturally in economics in the context of productivity analysis usually through the estimation of a production frontier, though this idea extends quite generally to cost, revenue, profit and distance frontiers. The econometric tools that have been developed for the estimation of, and inference for, the frontier have been applied across many different economic milieus, including energy, healthcare, transportation, schools, banks, and public service (see Kumbhakar, et al. 2017 for examples). Within the arena of frontier estimation, stochastic frontier analysis (SFA) is one of the most important tools wielded by researchers to both recover the frontier and to assess firm level inefficiency. One of the strongest appeals of SFA is the allowance, or recognition of, stochastic shocks that lead to deviations from the frontier.

[1] Department of Economics, Syracuse University, jcai106@syr.edu. Special thanks to Hugo Jales, Yulong Wang, and participants at 29th Annual Meeting of the Midwest Econometrics Group (MEG 2019).

[2] Department of Economics, Syracuse University, whorrace@syr.edu.

[3] Department of Economics, University of Miami, cparmeter@bus.miami.edu.

More explicitly, the stochastic frontier model assumes that output Y_i is produced from an underlying technology $\tau(\cdot)$ through inputs W_i, contaminated with firm level inefficiency U_i and random shocks (or random noise) V_i:

$$Y_i = \tau(W_i) \cdot \exp(-U_i) \exp(V_i), \tag{2.1}$$

where $U_i > 0$. It is commonly assumed that $V \sim N(0, \sigma_v^2)$ with σ_v^2 unknown. In most applications of SFA, a fully parametric model is assumed. For instance, in the pioneering work of Aigner et al. (1977) and Meeusen and van den Broek (1977), $\tau(W_i)$ is a parametric Cobb-Douglas production frontier, with inefficiency U_i assumed to be either half normally or exponentially distributed, respectively. Based on the conditional independence of U_i and V_i, the model in (3.1) can be estimated using standard maximum likelihood techniques or corrected least squares methods. See Greene (2008) and more recently Kumbhakar et al. (2017) for surveys.

However, the model as constituted in (3.1) *cannot* readily accommodate the presence of a mass of fully-efficient firms, as it presumes U follows a one-sided continuous distribution and probability of the event $U_i = 0$ is zero. Specifying a model with a mass of efficient firms makes sense in highly competitive or well established industries where firms may have an incentive to move toward the frontier. In these instances ignoring a mass at zero in the distribution of U_i would yield biased estimates for both mean inefficiency and firm specific inefficiency scores. Kumbhakar, Parmeter and Tsionas (2013) were the first to tackle this problem proposing a zero-inefficiency stochastic frontier (ZISF) model, which specifically assigns a nontrivial probability p to the event $U_i = 0$ and which is estimated by maximum likelihood. More recently, Rho and Schmidt (2015) discussed extreme cases (all firms are efficient and all firms are inefficient) in the ZISF model and related tests. Tran and Tsionas (2016) extended the ZISF model by allowing for the point mass probability of fully-efficient firms to depend on a set of covariates via an unknown, smooth function. However, all of the previous work in this area has remained tethered to parametric assumptions on both the frontier and the shape of the inefficiency distribution. The purpose of this paper is to relax many of the parametric assumptions used in the extant literature but still provide consistent estimators of both the frontier and the distribution of inefficiency.

When there is no mass of efficient firms, there are several semiparametric or nonparametric options for practitioners. Fan, Li and Weersink (1996), Hall and Simar (2002), Kumbhakar et al. (2007), Martins-Filho and Yao (2015) and Simar, Van Keilegom and Zelenyuk (2017) relax parametric assumptions on the structure of $\tau(W_i)$. Horrace and Parmeter (2011), Parmeter, Wang and Kumbhakar (2017) and Cai, Horrace and Parmeter (2019) relax the distributional assumption on inefficiency. See Parmeter and Zelenyuk (2019) for a thorough review of the state of the art. More recently, Kneip, Simar and Van Keilegom (2015) (KSVK, hereafter) consider an estimator for the stochastic frontier with unknown structure on the frontier $\tau(W)$ and unspecified distribution of inefficiency U_i but with a normally distributed error V. Florens, Simar and Van Keilegom (2019) generalize KSVK by only requiring a symmetric error. The price of this generality is that they can no longer nonparametrically identify the distribution of U_i. None of the previous semiparametric and nonparametric stochastic frontier literature considers a ZISF specification.

This paper goes one step further in the sense that we investigate a ZISF model with an unknown structure on the frontier $\tau(W_i)$ and an unspecified distribution of inefficiency U_i. Different from previous nonparametric methods, we identify the frontier τ based on the mode of the mixed density of $\tau - U$ and the latter can be estimated through a penalized minimum distance sieve method under a unimodality assumption of U. A discretized version of the mixed density of the inefficiency U is also identified, which was previously unavailable in the literature.[4] We propose a two-step procedure to estimate the frontier τ and the mixed distribution of inefficiency U. Similar to KSVK, this technique can be readily generalized to define the estimator for the stochastic frontier model in Equation (3.1) based on a local

[4] Much effort has been made in the SFA literature to relax or generalize the distributional assumption on inefficiency.

constant or local linear approximation.

Our approach, penalized minimum distance estimation based on characteristic functions, follows in spirit the approach of Lee et al. (2013) and Lee et al. (2015),[5] while our large sample theory is predicated on existing results in Chen and Pouzo (2012, 2015) for penalized sieve minimum distance (PSMD) estimators. Lee et al. (2015) proposed a penalized least squares sieve (hereafter, LS sieve) approach to estimate a mixture distribution with boundary effects, which they find has some advantage over the penalized maximum likelihood sieve (ML sieve) estimation advocated by Lee et al. (2013). Lee et al. (2015) considered estimation of both the probability mass and the continuous density for the class of mixture distributions with finite but *unknown* point mass locations contaminated with *known* random noise. We extend their penalized LS sieve method to the case with *known* point mass location (i.e., $U = 0$, zero point mass) and random noise with *unknown* variance σ_v^2. Moreover, we derive the large sample properties of the proposed penalized LS sieve estimator and propose a quasi-likelihood ratio (QLR) test on the zero inefficiency hypothesis.

The remainder of the paper is organized as follows. Section 2 describes the models we study along with our identification strategy and estimation procedures. Section 3 presents the asymptotic properties of the proposed estimators and inference on the estimated zero inefficiency probability with a QLR statistic. The selection of tuning parameters is discussed in Section 4. Section 5 presents numerical illustrations with various designs to demonstrate the performance of the method. We implement the proposed procedures to estimate a zero-inefficiency cost frontier for a cross section of US banks in Section 6. Section 7 concludes the paper and lays out directions for future research. All statistical proofs and details on the core computational algorithms can be found in the Appendix.

[5] More recently, Madrid-Padilla, Polson, Scott (2018) also consider a class of estimators for deconvolution in mixture models based on a simple two-step "bin-and-smooth" procedure applied to histogram counts.

2.2 Model

Consider a simplified production frontier model in Equation (3.1), which we rewrite as follows:

$$logY = \underbrace{log(\tau) - \overbrace{U}}_{KSVK}\overbrace{+V}^{LS-Sieve} \tag{2.2}$$

where τ is the production frontier, $U > 0$ denotes production inefficiency and $V \sim N(0, \sigma_v^2)$ is measurement error or statistical noise. For the cost frontier model, one just needs to replace $-U$ with $+U$ where $U > 0$ represents cost inefficiency.

In KSVK, they treat $log(\tau) - U$ as a single object and denote it as $logX$. In this setting the model in Equation (3.2) can be written as a traditional deconvolution problem

$$logY = logX + V$$

where Y is observed and $V \sim N(0, \sigma_v^2)$. The $X = \tau \cdot \exp(U)$ is a latent unobserved true signal having a density f on the support $[0, \tau]$, with $f(\tau) > 0$ for some unknown boundary point τ. The target is to consistently estimate τ and the unknown variance of the noise σ_v^2.

In the proposed LS sieve method, we view the model in (3.2) as

$$logY = \alpha - U + V \tag{2.3}$$

where Y is observed, $\alpha = log(\tau)$ is the intercept, U stands for production inefficiency which follows an unknown truncated distribution and $V \sim N(0, \sigma_v^2)$ as before.[6] This is a deconvolution problem with unknown intercept and unknown noise variance.

[6] We treat τ as a constant at the moment. This will be relaxed in section 2.2.

2.2.1 Constant Frontier

When τ is constant, Equation (3.1) becomes a constant frontier model. We discuss both KSVK's method and the proposed LS sieve method here to highlight the differences.

KSVK's method

In KSVK, Equation (3.1) is rewritten as $Y = X \cdot Z$ where $Z = \exp(V)$ is log-normally distributed. Let $\phi(z)$ denote the standard normal density and recall that the density ρ_{σ_v} of a log-normal random variable with parameters $\mu = 0$ and $\sigma_v^2 > 0$ is given by $\rho_\sigma(z) = \frac{1}{z\sigma_v}\phi(\frac{logz}{\sigma_v})$ for $z > 0$. Then for all $y > 0$ the density of y is represented as follows:

$$g(y) = \int_0^\tau f(x)\frac{1}{x}\rho_{\sigma_v}\big(\frac{y}{x}\big)dx = \int_0^1 h(t)\frac{1}{t\tau}\rho_{\sigma_v}\big(\frac{y}{t\tau}\big)dt = \frac{1}{\sigma_v y}\int_0^1 h(t)\phi\big(\frac{1}{\sigma_v}log\frac{y}{t\tau}\big)dt \tag{2.4}$$

where $h(t) = \tau f(t\tau)$ $(0 \leq t \leq 1)$ is a density function.

Discretize the density h as follows

$$h_\gamma(t) = \gamma_1 I(t=0) + \sum_{k=1}^M \gamma_k I(q_{k-1} < t \leq g_k)$$

for $0 \leq t \leq 1$, where $q_k = k/M(k = 0, 1, \ldots, M)$ and $\gamma_k > 0$, $\sum_{k=1}^M \gamma_k = M$. Then they have

$$g_{\tau,\sigma_v,h_\gamma}(y) = \frac{1}{\sigma_v y}\int_0^1 h(t)\phi\big(\frac{1}{\sigma_v}log\frac{y}{t\tau}\big)dt = \frac{1}{\sigma_v y}\sum_{k=1}^M \gamma_k \int_{q_{k-1}}^{q_k} \phi\big(\frac{1}{\sigma_v}log\frac{y}{t\tau}\big)dt.$$

KSVK then estimate the unknown frontier τ, unknown variance σ_v and the nuisance density h_γ by maximizing the following penalized likelihood:

$$(\hat\tau, \hat\sigma_v, \hat h_\gamma) = \underset{\tau>0,\sigma_v>0,h_\gamma}{\arg\max}\ \big\{n^{-1}\sum_{i=1}^n log\big(g_{\tau,\sigma_v,h_\gamma}(Y_i)\big) - \lambda_n pen(g_{\tau,\sigma_v,h_\gamma,})\big\},$$

where $\lambda_n \geq 0$ is a tuning (penalty) parameter, and

$$pen(g_{\tau,\sigma_v,h_\gamma}) = max_{3\leq j\leq M}|\gamma_j - 2\gamma_{j-1} + \gamma_{j-2}|.$$

The Proposed LS Sieve Method

We propose a least squares sieve method (LS sieve) to estimate the mixture distribution of inefficiency $f(u)$, and the production frontier τ with unknown noise variance. The mixture distribution of inefficiency means allows a point mass at $u = 0$ which admits the existence of a nontrivial amount of fully efficient firms and a continuous distribution on u. The proposed method is similar to the least squares sieve estimation proposed in Lee at al. (2015) except that we consider the case with unknown noise variance and unknown intercept.

Case 1: $\alpha = 0$**.** Considering the most basic setting, $\alpha = 0$, namely, the frontier $\tau = 1$, then Equation (3.2) collapses to a mixture distribution deconvolution problem as follows

$$logY = -U + V$$

where Y is observed, U comes from a mixture distribution with discrete atoms at $u = 0$ and a continuous distribution for the remaining support, and random noise (or measurement error) $V \sim N(0, \sigma_v^2)$ with unknown variance.

Let δ_a be the Dirac delta function at a. The point mass (if exists) located at $u = 0$ can be represented as δ_0. Then the generalized density of U has the following form

$$f_U(u) = \pi_1\delta_0(u) + \pi_2 f_c(u), \tag{2.5}$$

where π_1 and π_2 are nonnegative weights with $\pi_1 + \pi_2 = 1$, and f_c is an arbitrary probability density function of a continuous nonnegative random variable.

Theorem ID 1. *There exists a unique* $\pi_1 \in (0, 1)$ *(as well as* π_2*), a unique* $\sigma_v > 0$ *and a unique sequence* $\{\gamma_k\}_{k=1}^{M(n)}$ *which determines a unique density* $f_c(u)$ *(which can be uniquely* *discretized as Equation* (3.12) *in the sequel), such that the model in Equations* (3.3) *and* (3.11) *holds true with* $\alpha = 0$*, i.e., such that the model is identified.*

The proof is straightforward and it follows from Theorem 2.1 in Schwarz and Van Bellegem (2010), in which they prove the identifiability of any pair (σ_v, f_U) (where f_U is the probability distribution of U) belonging to

$$(0, \infty) \times \{P \in \mathcal{P} | \exists A \in \mathcal{B}(\mathbb{R}) : |A| > 0 \quad \textit{and} \quad P(A) = 0\},$$

where $\mathcal{B}(\mathbb{R})$ is the set of Borel sets in $\mathbb{R}$, $\mathcal{P}$ is the set of all probability distributions on $\mathbb{R}$, and $|A|$ is the Lebesgue measure of A. For the $\mathcal{P}$ distributions that have density like what we consider here, the only requirement for identification is that the density has to vanish on a set of positive Lebesgue measures.[7] The theorem result follows here by choosing A equal to the complement in $\mathbb{R}$ of the interval $[0, \infty)$, and by noting that the true mixed density of U in the proposed model is completely characterized by the true values of $\{\gamma_k\}_{k=1}^{M(n)}$, π_1 and σ_v.

This basic model is similar to the setting in Lee at al. (2015) except now we consider a case with *known* point mass location (i.e., $U = 0$, zero point mass) and random noise with *unknown* variance σ_v^2. We outline a similar "bin-and-smooth" procedure as follows. To estimate the unknown parameters, the first step is to discretize the continuous part of U, which is denoted by U_c. We approximate U_c by a discrete random variable $\tilde{U}_c$ taking values on an equally spaced grid, with grid spacing h. The discrete random variable $\tilde{U}_c$ takes on values u_j: $u_{j+1} - u_j = h$, $j = 1, \ldots, M(n)$, which covers the support of f_c. $M(n)$ is the number of bins and it is a function of sample size n. In practice, we choose $\tilde{U}_c$ satisfying

$$\tilde{U}_c = u_j \quad \text{if and only if} \quad \tilde{U}_c \in [u_j - 0.5h, u_j + 0.5h]$$

[7] Most of the typical truncated distributions belong to this category. For details, please refer to Schwarz and Van Bellegem (2010).

The parameter h is similar to the bin width in a histogram and the bandwidth in the kernel density estimator. Let $\theta = (\theta_1, \ldots, \theta_r)^T$ be the probability distribution of $\tilde{U}_c$, i.e.

$$\theta_j = P(\tilde{U}_c = u_j) \quad \text{for each} \quad j = 1, \ldots, M(n)$$

where $\theta_j \geq 0$ and $\sum_{j=1}^{M(n)} \theta_j = 1$. Then each θ_j represents the probability that $\tilde{U}_c = u_j$, which is an approximation of the probability that U_c lies in the interval $[u_j - 0.5h, u_j + 0.5h]$.

Replacing U_c by $\tilde{U}_c$, the distribution of U is purely discrete and the generalized density f_U in equation (3.11) can be written as

$$\tilde{f}_U(u|\pi, \theta) = \pi_1 \delta_0(u) + \pi_2 \sum_{j=1}^{M(n)} \theta_j \delta_{u_j}(u). \tag{2.6}$$

Based on this approximation, the problem turns into the estimation of π, θ as well as the unknown noise variance σ_v^2 (which plays a role in the empirical characteristic function as we will see in the sequel).

A natural idea is to minimize the distance between the empirical characteristic function and the approximated characteristic function as Schwarz and Van Bellegem (2010) did.[8] Recall the fact that convergence of characteristic functions implies convergence of corresponding distributions (Theorem 6.3.3 in Chung, 2001) and the empirical characteristic function converges to the true counterpart as the sample size goes to infinity. We expect the distance between the empirical characteristic function and the characteristic function corresponding to Equation (3.11) might be small if the parameters are close to the truth. Based on that,

[8] Schwarz and Van Bellegem (2010) applied a minimum distance method based on characteristic functions to estimate the distribution of a latent variable in the context of normal contamination errors with unknown variance.

we can estimate π, θ and σ_v by minimizing

$$\begin{aligned}(\hat{\pi}, \hat{\theta}, \hat{\sigma}_v) &= \underset{\pi\in\Pi,\theta\in\Theta,\sigma\in\Sigma}{\arg\min}\ S_{chf}(\pi, \theta, \sigma_v) \\ &= \underset{\pi\in\Pi,\theta\in\Theta,\sigma\in\Sigma}{\arg\min} \int |\hat{\varphi}_n(t) - \tilde{\varphi}_Y(t|\pi, \theta, \sigma_v)|^2 w(t)dt\end{aligned} \tag{2.7}$$

where $\Pi = \{(\pi_1, \pi_2) : \pi_1 \geq 0, \pi_2 \geq 0, \pi_1 + \pi_2 = 1\}$, $\Theta = \{(\theta_1, \ldots, \theta_M) : \theta_j \geq 0, j = 1, \ldots, M(n), \sum_{j=1}^{M(n)} \theta_j = 1\}$, $\Sigma = \{\sigma_v : 0 < \sigma_v^2 \leq Var(logY)\}$.[9] $\hat{\varphi}_n(t) = \frac{1}{n}\sum_{k=1}^{n} exp(it \cdot logY_k)$ is the empirical characteristic function of $logY$ following equation (3.2) and

$$\tilde{\varphi}_Y(t|\pi, \theta, \sigma_v) = \pi_1 e^{it\cdot 0}\varphi_V(t) + \pi_2 \sum_{j=1}^{M(n)} \theta_j e^{itu_j}\varphi_V(t) \tag{2.8}$$

where $\varphi_V(t) := E(e^{itV}) = \int e^{itv} f_v(v)dv$, $f_v(v) = \frac{1}{\sqrt{2\pi}\sigma_v} exp(\frac{v^2}{2\sigma_v^2})$ as $V \sim N(0, \sigma_v^2)$. Actually, $\varphi_V(t) = exp(-0.5\sigma_v^2 t^2)$. This is where the unknown noise variance σ_v plays a role in the minimization problem.

Note that S_{chf} is a quadratic function of both π, θ and σ_v^2 and all of these parameters are defined on compact subsets of Euclidean space. Hence there exists a unique minimizer of Equation (2.7). One unfortunate problem is that we cannot obtain closed form solutions due to the presence of the constraints on the parameters. Due to this an iterative EM algorithm is proposed to solve this constrained minimization problem, the details of which appear in Appendix A.

Here we propose a LS sieve method based on characteristic functions (LS-ChF) to estimate π, θ and σ_v. An alternative approach which could be similarly employed is a LS sieve method based on cumulative probability functions (LS-CDF). The idea is to minimize the distance between the empirical cumulative probability function and the approximated cumulative probability function. We omit further discussion of this alternative estimation approach.

[9] The upper bound of the unknown noise variance comes from the independence assumption between U and V.

Case 2: $\alpha \neq 0$**.** When $\alpha \neq 0$, which is the typical case in stochastic frontier analysis, equation (3.2) becomes a mixture distribution deconvolution problem with both unknown variance and unknown intercept:

$$logY = \alpha - U + V.$$

This is a very difficult problem. To identify this model, an additional assumption is needed:

Assumption. *(Unimodality) U is unimodal and its mode is zero.*

This unimodality assumption is weak and it admits most typical distributions in stochastic frontier analysis, such as half-normal and exponential. Hall and Simar (2002) introduced the same unimodality assumption in a similar setting to identify the boundary of a latent variable. For a mixture distribution with point masses at zero, which is the case of interest here, this assumption is even weaker as it allows the non-zero truncated distribution for the continuous part U_c with a large point mass probability π_1. With $\pi_1 > max(f_c)$, the intercept is still identified as the mode stays at zero by definition.[10]

Theorem ID 2. *There exists a unique $\pi_1 \in (0,1)$ (as well as π_2), a unique $\sigma_v > 0$ and a unique sequence $\{\gamma_k\}_{k=1}^{M(n)}$ which determines a unique density $f_c(u)$ which is discretized in equation* (3.12) *and satisfies unimodality assumption, such that the model* (3.3) *and* (3.11) *hold true with $\alpha = const$, i.e., such that the model is identified.*

The proof is straightforward based on the results of Theorem ID 1 and the unimodality assumption. Under the unimodality assumption, we can identify the constant intercept α in the following sense:

$$mode(\alpha - U) = \alpha - mode(U) = \alpha \approx u_{j^*}, \quad where \quad j^* = index(\max\{\theta_j\})$$

[10] If the true intercept is zero, we can identify the mixture distribution parameters with non-zero truncated distributed inefficiency. This comes back to the previous case with $\tau = 0$.

where the second equality holds due to the unimodality assumption, and specifically zero modality of U. j^* is the index of the maximum evaluated θ. After obtaining the intercept α, we can shift $\log Y$ and transform it as $\tilde{Y} = logY - \hat{\alpha} = -U + V$, which collapses to the $\alpha = 0$ case and is identified in Theorem ID 1.

Hence, a two-step procedure is proposed to estimate the mixture distribution with both unknown noise variance and intercept:

1. Estimate the mode of $U + \alpha$ using the procedures in case 1 as we can always rewrite the model as

   $$logY = \underbrace{\alpha - U} + V = U' + V$$

   where $U' = \alpha + U$ and mode(U')=α. Once we obtain the estimator for θ: $\hat{\theta}_1$, the mode can be obtained by searching the maximum index of $\hat{\theta}_1$ and locating the corresponding u_j, i.e., $\hat{\alpha} = u_{j^*}$ where $j^* = index(\max_j \hat{\theta}_{1j})$.

2. After obtaining the estimator for the intercept $\hat{\alpha}$, we can plug it back into Equation (3.3) and transform the model to

   $$\tilde{Y} = U + V$$

   where $\tilde{Y} = logY - \hat{\alpha}$. Then the model reduces to Case 1 with $\alpha = 0$. We can implement the procedure in the zero intercept case to estimate the following unknown parameters: π, θ and σ_v.

It is well documented in both the statistics and econometrics literature that the inverse problem of estimating the distribution of a latent variable U from an observed sample of $\tilde{Y}$, a contaminated measurement of U, is ill-posed (Tikhonov, 1963; Fan, 1991; Chen and Reiss, 2011, etc). Therefore a penalty function is added to smooth the estimated density $\hat{f}_c$:

$$(\hat{\pi}, \hat{\theta}, \hat{\sigma}_v) = \underset{\pi\in\Pi,\theta\in\Theta,\sigma\in\Sigma}{\arg\min} S_{chf}(\pi, \theta, \sigma_v) + \lambda_n P(\theta) \tag{2.9}$$

where $P(\theta)$ is the roughness penalty. It can be an arbitrary nonnegative function which has a smaller value when θ is smoother. Following Lee et al. (2015), we choose the sum of squared first order differences, i.e. $P(\theta) = \sum_{j=2}^{M(n)} (\theta_j - \theta_{j-1})^2$ as the penalty function. The $\lambda_n > 0$ is the penalty parameter which is crucial in practice. We suggest selecting λ_n based on root mean squared error (RMSE) of the corresponding estimators with a bootstrap procedure the details of which are provided in Section 3.4.

Based on estimated $\hat{\pi}, \hat{\theta}, \hat{\sigma_v}$, we could construct the corresponding density estimator as follows

$$\tilde{f}_U(u|\hat{\pi}, \hat{\theta}) = \hat{\pi}_1 \delta_0(u) + \hat{\pi}_2 \sum_{j=1}^{M(n)} \hat{\theta}_j \delta_{u_j}(u). \tag{2.10}$$

This estimator can be improved by linear interpolation

$$\tilde{f}_U(u|\hat{\pi}, \hat{\theta}) = \hat{\pi}_1 \delta_0(u) + \hat{\pi}_2 \sum_{j=1}^{M(n)} \hat{\theta}_j \hat{f}_c(u|\hat{\theta}),$$

where

$$\hat{f}_c(u|\hat{\theta}) = \begin{cases} \frac{\hat{\theta}_{j-1}}{h} + \frac{\hat{\theta}_j - \hat{\theta}_{j-1}}{h(u_j - u_{j-1})}(u - u_{j-1}) & \text{if } u \in [u_j - 0.5h, u_j + 0.5h], \quad j = 2, \ldots, M(n) \\ 0 & \text{otherwise} \end{cases}$$

2.2.2 Heteroskedastic Frontiers

Heteroskedastic frontiers allow for the presence of covariates in the frontier, namely, $\tau = \tau(W)$ where $W \in R^d$ are the covariates. Redefine $U \geq 0$ to allow zero inefficiency. Then Equation (3.1) becomes

$$Y = \tau(W) \cdot \exp(-U) \exp(V) \tag{2.11}$$

where $\tau(W)$ is a heteroskedastic frontier incorporating covariates, $U \geq 0$ denotes inefficiency, and $V \sim N(0, \sigma_v^2)$ is measurement error or statistical noise as before. For instance, considering a production stochastic frontier model where W denotes inputs, $\tau(W)$ is the frontier and $\tau(W) \cdot \exp(-U)$ is the “true” output. This is the more typical case for the practitioner except that we now allow the presence of some fully efficient firms.

Taking logarithms on Equation (2.11), we have

$$logY = \alpha(W) - U + V \tag{2.12}$$

where $\alpha(W) = log(\tau(W))$ is the logarithmic transformation of the heteroskedastic frontier. The problem now turns into estimating the intercept $\alpha(w_0)$, the mixture distribution of U and the noise variance σ_v^2 conditional on $W = w_0$.

Suppose the following conditions hold: (i) $\{W_i, Y_i\}_{i=1}^n$ are i.i.d observations; (ii) the conditional distribution of V given $W = w_0$ is $N(0, \sigma_v^2(w_0))$ with a true variance $\sigma_v^2(w_0)$ (possibly) depending on w_0; (iii) U and V are conditionally independent given $W = w_0$ and U is nonpositive for production frontier model and nonnegative for cost frontier model; (vi) U is conditionally unimodal and its mode is zero.

Under above conditions, estimation of model (2.12) is a straightforward generalization of the estimation of the baseline model in equation (3.3) with $\alpha \neq 0$. Since given $W = w_0$, $\alpha = \log(\tau(w))$ is a constant and model (2.12) can be view as a constant frontier model in the neighborhood of $W = w_0$. This is in the same spirit of local boundary estimation as in Hall and Simar (2002) and KSVK (2015).

By condition (ii), $E(V|W = w) = 0$. Assume that $E(U|W = w)$ is a constant in a small neighborhood of the evaluation point w_0 and $\tau(\cdot)$ is sufficiently smooth and can be well represented by a Taylor expansion, i.e., $\log(\tau(w)) \approx \log(\tau(w_0)) + \beta^T(w_0)(w - w_0)$ with $\beta(w_0) = \frac{\partial}{\partial w}(\log \tau)(w)|_{w=w_0}$. For a small $b > 0$, we then obtain

$$\log Y_i \approx c(w_0) + \beta^T(w_0)(W_i - w_0) - U_i^0 + V_i, \quad \text{if} \quad ||W - w_0|| \leq b$$

where $c(w_0) = \log \tau(w_0) - E(U|W = w_0)$ and $U_i^0 = U_i - E(U|W = w_0)$. The term $-U_i^0 + V_i$ is the compound error term with zero mean and hence $c(w_0)$ and $\beta(w_0)$ can be estimated with ordinary least square. The $\beta(w_0)$ contains the local variation of $\alpha(w)$ (i.e., $\log \tau(w)$), which can be used to calculate a suitable correction of the mixture distribution of the inefficiency. In order to apply the estimation strategy in the baseline model, we can transform the dependent variable $\log Y_i$ as follows:

$$\begin{aligned} \log \tilde{Y_i} :&= \log Y_i - \beta^T(w_0)(W_i - w_0) \\ &\approx c(w_0) - U_i^0 + V_i \\ &= \log(\tau(w_0)) - U_i + V_i \\ &= \alpha(w_0) - U_i + V_i \end{aligned} \tag{2.13}$$

where $\alpha(w_0) := \log(\tau(w_0))$; or alternatively, we can apply following transformation

$$\begin{aligned} \log \tilde{Y_i} :&= \log Y_i - c(w_0) - \beta^T(w_0)(W_i - w_0) \\ &\approx -U_i^0 + V_i \\ &= E(U|W = w_0) - U_i + V_i \\ &= \alpha(w_0) - U_i + V_i \end{aligned} \tag{2.14}$$

where $\alpha(w_0) := E(U|W = w_0)$. Then apply the estimation procedure in equation (3.3) with $\alpha \neq 0$. Naturally, we can obtain the estimator of α, π and θ, consequently the continuous density f_c, conditional on $W = w_0$.

Based on the estimation strategy, we develop following estimation procedure for the heteroskedastic frontier model:

1. Fix a bandwidth $b > 0$ and estimate $c(w_0)$ and $\beta(w_0)$ by minimizing the following local least squares problem

$$\sum_{||W_i - w_0|| \leq b} (\log Y_i - c - \beta^T(W_i - w_0))^2$$

   Denote the estimates as $\hat{c}(w_0)$ and $\hat{\beta}(w_0)$.

2. Transform Y_i by either $\log \tilde{Y}_{1i} := \log Y_i - \beta^T(w_0)(W_i - w_0)$ or $\log \tilde{Y}_{2i} := \log Y_i - c(w_0) - \beta^T(w_0)(W_i - w_0)$.

3. Following the two-step procedure in the constant frontier estimation with $\alpha \neq 0$ to estimate $\alpha(w_0)$, $\pi(w_0)$ and $\theta(w_0)$, consequently the continuous density $f_c(u|w_0)$.

4. For the first transformation of Y_i, $\hat{\alpha}(w_0) = \log \hat{\tau}(w_0)$, we can get the conditional expected inefficiency level $\hat{E}(U|W = w_0) = \hat{\alpha}(w_0) - \hat{c}(w_0)$; for the second transformation of Y_i, $\hat{\alpha}(w_0) = \hat{E}(U|W = w_0)$, we can get the heteroskedastic frontier $\log \hat{\tau}(w_0) = \hat{c}(w_o) + \hat{\alpha}(w_0)$.

Unlike KSVK where the boundary $\log \tau(w)$ is identified without point mass in the distributions of U, we now estimate the heterogeneous intercept $\alpha(w)$ through the unimodality property of $\alpha - U$. Similar to KSVK, our method can also identify the conditional expected inefficiency level $E(U|W = w_0)$ with the proposed procedures. Moreover, our estimation procedure recovers the entire mixture distribution of inefficiency at a neighborhood around w_0, especially the probability of zero-inefficiency among the firms which are of primary interest for policy makers or regulators. The (conditional) unimodality of U yields the identification of the (heterogeneous) intercepts which lays the foundation for the point mass probability estimation as well as the continuous distribution of the heteroskedastic inefficiency at each of these neighborhoods.

2.3 Asymptotic Properties

In this section, we investigate the asymptotic properties of the proposed estimators and inference on zero inefficiency probability with a QLR statistic based on the results derived in Chen and Pouzo (2012, 2015).[11] We deal with the unknown noise variance extension first,

[11] Schwarz and Van Bellegem (2010) and Lee et al. (2015) also provided consistency of similar minimum distance estimation routines.

namely, the $\alpha = 0$ case, then generalize the theorems to the unknown noise variance with nonzero intercepts (i.e., $\alpha \neq 0$).

The proposed penalized LS sieve estimator can be regarded as a special case of the general PSMD estimator studied by Chen (2007) and Chen and Pouzo (2012, 2015). Define $\rho(Y,h) = |\phi_n(t_i) - \phi_Y(t_i|\pi,\theta)|$, then $E(\rho(Y,h)|t) = 0$ as $\phi_n(t) = \phi_Y(t|\pi,\theta)$, $\forall\ t$. Define $m(\cdot,h) = E(\rho(Y,h)|t = \cdot)$. We can rewrite the objective function in equation (3.15) with a special weight function $w(t) = 1(M_1 \leq t \leq M_2)$ ($M_1 = \min(\log Y_i)$, $M_2 = \max(\log Y_i)$) as

$$\hat{Q}_n(h) = \frac{1}{M_1 - M_2} \sum_{t_i = M_1}^{M_2} (\hat{m}(t_i,h)^T \hat{m}(t_i,h)) + \lambda_n \hat{P}_n(h) \tag{2.15}$$

where $\hat{m}(t_i,h) = E(\hat{\rho}(Y,h)|t_i) = |\hat{\phi}_n(t_i) - \tilde{\phi}_Y(t_i|\pi,\theta)|$, and

$$\mathcal{H}_n = \{h \in H : h(t) = [\pi_1 + \pi_2 \sum_{j=1}^{k(n)} \theta_j (i \sin(tx_j) + cos(tx_j))]\phi_v(t,\sigma_v)\}$$

where $k(n) \to \infty$ slowly as $n \to \infty$. Specifically, $k(n)/n \to 0$ as $n \to \infty$. $P(\cdot)$ is a penalty function and λ_n is the tuning or penalty parameter. Essentially the proposed method is a PSMD approach using a slowly growing infinite-dimensional linear (Fourier series) sieve studied in Chen and Pouzo (2012). As π, σ_v and the infinite-dimension θ which determines the continuous part of the inefficiency U are all parameters of interest in our context, the proposed model is semi-nonparametric based on the definition proposed by Chen (2007).

In order to demonstrate consistency, we need to verify the following assumptions in Chen and Pouzo (2012) (CP(2012), hereafter) first.

Notation: Denote $L^p(\Omega, d\mu)$ as the space of measurable functions with $||f||_{L^p(\Omega,d\mu)} \equiv \{\int_\Omega |f(t)|^p d\mu(t)\}^{1/p} < \infty$, where Ω is the support of a sigma-finite positive measure $d\mu$ (sometimes $L^p(d\mu)$ and $||f||_{L^p(d\mu)}$ are used for simplicity). For any positive sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, $a_n \asymp b_n$ means that there exists two constants $0 < c_1 \leq c_2 < \infty$ such that $c_1 a_n \leq b_n \leq c_2 a_n$; $a_n = O_p(b_n)$ means that $\lim_{c\to\infty} \limsup_n Pr(a_n/b_n > c) = 0$ and $a_n =$

$o_p(b_n)$ means that for all $\varepsilon > 0$, $\lim_{n\to\infty} Pr(a_n/b_n > \varepsilon) = 0$. For any vector-valued A, we let A^T denote the transpose and $||A||_W \equiv \sqrt{A^TWA}$ for its weight norm, although sometimes we also use $|A| = ||A||_I \equiv= \sqrt{A^TA}$ without too much confusion. We use $\mathcal{H}_n \equiv \mathcal{H}_{k(n)}$ to denote the sieve spaces.

Suppose $V \sim N(0, \sigma_v^2)$ with unknown variance and we observe $\{Y\}_{i=1}^n$ (or $\{\log Y\}_{i=1}^n$) which is a contaminated version of the "true" signal that follows a mixture distribution of a point mass at zero and a truncated unknown distribution (for the constant frontier case). The unknown distribution of Y satisfies $E(\rho(Y,h)|t) = 0$ for any t, where $\rho : \mathcal{Y} \times \mathcal{H} \to \mathbb{R}^{d_\rho}$ is a measurable mapping known up to a vector of unknown functions, $h_0 \in \mathcal{H} \equiv \mathcal{H}^1 \times \ldots \times \mathcal{H}^q$, a separable Banach space with norm $||h||_s \equiv \sum_{l=1}^q ||h_l||_{s,l}$.

Assumption 1. *Identification, Sieves: (i) W(t) is a positive definite matrix for all t; (ii)* $E(\rho(Y,h_0)|t) = 0$ *and* $||h_0 - h||_s = 0$ *for any* $h \in (\mathcal{H}, ||\cdot||_s)$ *with* $E(\rho(Y,h_0)|t) = 0$*; (iii)* $\{\mathcal{H}_k : k \geq 1\}$ *is a sequence of nonempty closed subsets satisfying* $\mathcal{H}_k \subseteq \mathcal{H}_{k+1} \subseteq \mathcal{H}$*, and there is* $\Pi_n h_0 \in \mathcal{H}_{k(n)}$ *such that* $||\Pi_n h_0 - h_0||_s (\equiv \sqrt{E[(\Pi_n h_0 - h_0)^2]}) = o(1)$*; (iv)* $E[||m(t, \Pi_n h_0)||_W^2] (\equiv E[m(t,\Pi_n h_0)^T m(t, \Pi_n h_0)]) = o(1)$*.*

Assumption 2. *Penalty: one of the following holds: (a)* $\lambda_n = 0$*, (b)* $\lambda_n > 0$*,* $\lambda_n = o(1)$*,* $\sup_{h\in\mathcal{H}_{k(n)}} |\hat{P}_n(h) - P(h)| = O_p(1)$*, and* $|P(\Pi_n h_0) - P(h_0)| = O(1)$ *with* $P : \mathcal{H} \to [0, \infty)$*,* $P(h_0) < \infty$*, or (c)* $\lambda_n > 0$*,* $\lambda_n = o(1)$*,* $\sup_{h\in\mathcal{H}_{k(n)}} |\hat{P}_n(h) - P(h)| = o_p(1)$*, and* $|P(\Pi_n h_0) - P(h_0)| = o(1)$ *with* $P : \mathcal{H} \to [0, \infty)$*,* $P(h_0) < \infty$*.*

Assumption 3. *Sample Criterion: (i)* $\frac{1}{n}\sum_{i=1}^n ||\hat{m}(t_i, \Pi_n h_0)||_{\hat{W}}^2 \leq c_0 \times E[||\hat{m}(t_i, \Pi_n h_0)||_W^2] + O_p(\eta_{0,n})$ *for some* $\eta_{0,n} = o(1)$ *and a finite constant* $c_0 > 0$*; (ii)* $\frac{1}{n}\sum_{t_i=M_1}^{M_2} ||\hat{m}(t_i, h_0)||_{\hat{W}}^2 \geq c \times E[||\hat{m}(t_i, h_0)||_W^2] - O_p(\bar{\delta}_{m,n}^2)$ *uniformly over* $\mathcal{H}_{k(n)}^{M_0}$ *for some* $\bar{\delta}_{m,n}^2 = o(1)$ *and a finite constant* $c > 0$*.*

Assumption 1 (i) and (ii) are trivially satisfied since $W(t) = 1(M_1 \leq t \leq M_2)$ and $E(\rho(Y,h)|t) = 0$ as $\phi_n(t) = \phi_Y(t|\pi, \theta)$, $\forall\ t$. Assumption 1 (ii) is for global identification. Assumption 1 (iii) defines the sieves. Under Assumption 1 (ii) and (iii), Assumption 1 (iv) is satisfied if $E[||\hat{m}(t_i,h)||_W^2]$ is continuous at h_0 under $||\cdot||_s$. And this is the case here since $E[||\hat{m}(t_i,h)||_W^2] \equiv E[m(t,h)^T m(t,h)]$ where $m(t,h) = E(\rho(Y,h)|t) = E(|\phi_n(t) - \phi_Y(t|\pi,\theta)|)$ which are continuous in h_0.

For Assumption 2, (a) means no penalty; (b) and (c) are trivially satisfied when $\mathcal{H}_{k(n)} = \mathcal{H}$ and $\hat{P}_n = P$. Assumption 2 (c) is a stronger version of Assumption 2 (b). Under Assumption 1 (iii) and $P(h_0) < \infty$, a sufficient condition for $|P(\Pi_n h_0) - P(h_0)| = o(1)$ is that $P(\cdot)$ is continuous at h_0. As we choose the penalty function $P(\cdot)$ as $\sum(\theta_j - \theta_{j+1})^2$, the above conditions are trivially satisfied.

Assumption 3 is satisfied as $\hat{m}(Y,h)$ can be written as a series least squares (series LS) estimator:

$$\hat{m}(t,h) = p^{k(n)}(t)^T \cdot (P^T P)^- \sum p^{k(n)}(t_i)\rho(Y_i,h) \quad (*)$$

where $p^{k(n)}(X) = (p_1(X), \dots, p_{k(n)}(X))^T$, $P = (p^{k(n)}(X_1), \dots, p^{k(n)}(X_n))^T$, $(P^T P)^-$ is the Moore-Penrose generalized inverse and $\{p_j(\cdot)\}_{j=1}^{\infty}$ is a sequence of (generalized) Fourier series from $\mathcal{H}_n$. The $k(n)$ is the number of approximating terms as before in the definition of h. The Lemma C.2 in Appendix C in Chen and Pouzo (2012) shows that the series LS estimator defined in (*) satisfies Assumption 3. For the series LS estimator $\hat{m}(t,h)$ defined above, we can choose $\delta_{m,n}^2 = \eta_{0,n} = const \times \frac{k(n)}{n} = o(1)$.

This leads to our first theorem on consistency:

Theorem 1. *Let $\hat{h}_n$ be the PSMD estimator with $\lambda_n \geq 0$ and $\eta_n = O(\eta_{0,n})$, and suppose Assumption 1, 2 and 3 hold. If $\max\{\eta_{0,n}, E[||m(t,\Pi_n h_0)||_W^2], \bar{\delta}_{m,n}^2, \lambda_n\} = o(g(k(n),\epsilon))$ for all $\epsilon > 0$ where $g(k(n),\epsilon) \equiv \inf_{h \in \mathcal{H}_{k(n)}^{M_0} : ||h-h_0||_s \geq \epsilon} E[||m(t,h)||_W^2]$, then*

$$||\hat{h}_0 - h_0||_s = o_p(1) \quad and \quad P(\hat{h}_n) = O_p(1) \quad if \quad \lambda_n > 0.$$

The proof is straightforward following Theorem 3.1 in CP (2012) requiring verification of the conditions required there for the present context. Full details are provided in Appendix B.

Next, we proceed to propose a unified theorem which demonstrates the convergence rate of the proposed estimator and incorporates KSVK's estimator as a special case. Given the consistency result above, we now restrict our attention to a shrinking $|| \cdot ||_s$ neighborhood around h_0 to derive the convergence rate. Following CP (2012), define

$$\mathcal{H}_{os} \equiv \{h \in \mathcal{H} : ||h - h_0||_s \leq \epsilon, ||h||_s \leq M_1, \lambda_n P(h) \leq \lambda_n M_0\}$$

$$\mathcal{H}_{osn} \equiv \mathcal{H}_{os} \cap \mathcal{H}_n$$

for some positive finite constants M_1 and M_0, a sufficiently small positive ϵ such that $Pr(\hat{h}_n \notin \mathcal{H}_{os}) < \epsilon$. We can treat $\mathcal{H}_{os}$ as the new parameter space and $\mathcal{H}_{osn}$ as the new sieve space.

Assume that $\mathcal{H}_{os}$ is an infinite-dimensional subset of a real-valued separable Hilbert space $\mathbf{H}$ with an inner product $< \cdot, \cdot >_s$ and the inner product induced norm $|| \cdot ||_s$. Let $\{q_j\}_{j=1}^{\infty}$ be a Riesz basis associated with the Hilbert space $(\mathcal{H}, || \cdot ||_s)$; that is , any $h \in \mathcal{H}$ can be expressed as $h = \sum_j < h, q_j >_s q_j$, and there are two finite constants $c_1, c_2 > 0$ such that $c_1 ||h||_s^2 \leq \sum_j | < h, q_j >_s |^2 \leq c_2 ||h||_s$ for all $h \in \mathcal{H}$.

Assumption 4. *Sieve Approximation Error:* $||h_0 - \sum_{j=1}^{k(n)} < h, q_j >_s q_j||_s = O({v_{k(n)}}^{-\kappa})$ *for a finite* $\kappa > 0$ *and a positive sequence* $\{v_j\}_{j=1}^{\infty}$ *that strictly increase to* ∞ *as* $j \to \infty$.

Assumption 5. *Sieve Link Condition: There are finite constants* c, $c > 0$ *and a continuous increasing function* $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ *such that (i)* $||h||^2 \geq c \sum_{j=1}^{\infty} \varphi(j^{-2}) | < h, q_j >_s |^2$ *for all* $h \in \mathcal{H}_{osn}$ *and (ii)* $||\Pi_n h_0 - h_0||^2 \leq c \sum_{j=1}^{\infty} \varphi(j^{-2}) \times | < \Pi_n h_0 - h_0, q_j >_s |^2$.

Fourier series bases satisfy Assumption 4 with $v_{k(n)} = k(n)^{1/d}$ where d is the dimension of target parameter U here. So $d = 1$ and $v_{k(n)} = k(n)$. Assumption 5 (i) relates the weak pseudometric $||h||$ to the strong norm in a shrinking sieve neighborhood $\mathcal{H}_{osn}$ of h_0. Assumption 5 (ii) is the so called "stability condition" in CP (2012) with $v_{k(n)} = k(n)$ here. It is required to hold only in terms of the sieve approximation error $\Pi_n h_0 - h_0$.

Under the above assumptions, we can apply corollary 5.1 in CP (2012) to establish convergence of the smoothing parameter in present context:

Lemma 3. *Let $\hat{h}_n$ be the PSMD estimator with $\lambda_n \geq 0$ and $\lambda_n = o(1)$. Let Assumptions 1-5 hold and max$\{\delta_{m,n}, \lambda_n\} = \delta^2_{m,n} = const \times \frac{k(n)}{n} = o(1)$. Then*

$$||\hat{h}_n - h_0||_s = O_p\big(k(n)^{-\kappa} + \sqrt{\frac{k(n)}{n \times \varphi(k(n)^{-2})}}\big).$$

Thus, $||\hat{h}_n - h_0||_s = O_p(n^{-\kappa/2(\kappa+\varsigma)+1})$ if $\varphi(\tau) = \tau^{\varsigma}$ for some $\varsigma \geq 0$ and $k(n) \asymp n^{1/2(\kappa+\varsigma)+1}$, and $||\hat{h}_n - h_0||_s = O_p([\ln n]^{-\kappa/\varsigma})$ if $\varphi(\tau) = \exp(-\tau^{-\varsigma/2})$ for some $\varsigma > 0$ and $k(n) = c[\ln n]^{1/\varsigma}$ for some $c \in (0,1)$.

Proof of Lemma 3 is a direct application of Corollary 5.1 in CP (2012) with Fourier series LS estimation and $d = 1$. Basically we choose $k(n)$ to balance the sieve approximation error rate ($\{\upsilon_{k(n)}\}^{-\kappa}$) and the model complexity ($\sqrt{\frac{k(n)}{n\times\varphi(k(n)^{-2})}}$). Different convergence rates are derived based on the degree of ill-posedness which is represented by $\varphi(\cdot)$. Note that this also lays a foundation for considering different distributions of the random shocks V rather than normal concerning that the characteristic functions of V (e.g., $\phi_v(t, \sigma_v)$) determines the function form of $\varphi(\cdot)$. For instance, Cauchy distribution with location parameter $\mu = 0$ or, more generally, for stable distributions with fixed exponent $\alpha \in (0, 2]$, skewness parameter $\beta = 0$, and location $\mu = 0$ are admitted here.

Consider the Hellinger distance $H^2(f_1, f_2) = \frac{1}{2}\int(\sqrt{f_1(u)} - \sqrt{f_2(u)})^2 du$ for any arbitrary density functions of u: $f_1(u)$ and $f_2(u)$. Then we have following theorem:

Theorem 2. *Let $\tilde{\varphi}_Y(t|\pi, \theta, \sigma_v)$ be the PSMD estimator defined in equation (2.8) which equals to $\hat{h}_n$ with $k(n) = M(n)$ where $M(n)$ as the number of bins, $\lambda_n \geq 0$ and $\lambda_n = o(1)$, $\varphi_0(t)$ is the true characteristic function of Y_i, $i = 1, \ldots, n$. Let Assumptions 1-5 and Unimodality hold and max$\{\delta_{m,n}, \lambda_n\} = \delta^2_{m,n} = const \times \frac{M(n)}{n} = o(1)$. Then*

$$||\tilde{\varphi}(t) - \varphi(t)||_s = O_p\big(n^{-1/2} + M(n)^{-\kappa} + \sqrt{\frac{M(n)}{n \times \varphi(M(n)^{-2})}}\big) = O_p([\ln n]^{-\kappa/\varsigma})$$

where $\varphi(\tau) = \exp(-\tau^{-\varsigma/2})$ for some $\varsigma > 0$ and $M(n) = c[\ln n]^{1/\varsigma}$ for some $c \in (0,1)$.

Consequently,

$$H^2(\tilde{f}_U, f_U) = O_p\big(\max([\ln n]^{-1/\varsigma}, \quad [\ln n]^{-\kappa/\varsigma})\big)$$

$$||\hat{\alpha} - \alpha|| = O_p\big(\max([\ln n]^{-1/\varsigma}, \quad [\ln n]^{-\kappa/\varsigma})\big),$$

where $\tilde{f}_U$ *is defined in equation* (3.16) *and* $\alpha \neq 0$ *is defined in equation* (3.3).

The proof is based on Lemma 3 and details are in Appendix 2.C. We derive $\ln(n)$ convergence rate due to the severe ill-posedness caused by the normally distributed random shock V. In practice, one needs to choose the number of bins which is also the number of evaluation points for the continuous density $f_c(u)$. A rule of thumb is to try ς =1, 2, 3, 4 to choose $M(n)$ with normally distributed random errors. We discuss this in detail in Section 4.

Though proved in a different manner, Theorem 5 incorporates Theorem 3.2 in KSVK as a special case: $\varphi(\tau) = exp(-\tau^{-\varsigma/2})$ with $\kappa = \varsigma = 2$. Essentially, the sieve they used is $\phi(\frac{1}{\sigma}\log\frac{y}{t\alpha})$ (when they discretized the target density) which is exactly $\varsigma = 2$ for $\varphi(k(n)^{-2}) = \exp(-k(n)^{\varsigma})$ here (recall $\phi(y) = O(\exp\{-0.5y^2\})$). Therefore, a direct application yields $||\tilde{\varphi}(t) - \varphi(t)||_s = ||\hat{h}_n - h_0||_s = O_p([\ln n]^{-1})$. They derived a slightly different convergence rate for the unknown frontier and variance due to choosing a special order of the tuning parameter in the penalty function.[12]

Next we propose a sieve quasi-likelihood ratio (QLR) test on the zero inefficiency hypothesis which is very useful and informative in the ZISF model.

Theorem 3. *Let* $\hat{m}$ *be the series LS estimator in (*) for* $E(\hat{\rho}(Y,h)|t_i) = 0$ *a.s. any* t_i*. Let Assumption 1-5 hold and define* $\phi(h) = \int \hat{m}(t_i,h)'\hat{m}(t_i,h)dt_i$*, then*

$$\sqrt{n}\frac{\phi(\hat{h}_n) - \phi(h_0)}{||v_n^*||_{sd}} \rightarrow \mathcal{N}(0,1),$$

where $||v_n^*||_{sd} = 2\int \hat{m}(t_i,h_0)p^{k(n)}(t_i)dt_i D_n^- \Phi_n D_n^- 2\int \hat{m}(t_i,h_0)p^{k(n)}(t_i)dt_i,$

[12] This difference also leads to a slightly different choice of order of $k(n)$ which is M in KSVK.

$D_n^- = E\big(E[p^{k(n)}(Y)|t_i]E[p^{k(n)}(Y)|t_i]'\big)$, $\Phi_n = E\big(E[p^{k(n)}(Y)|t_i]\rho^2(t_i, h_0)E[p^{k(n)}(Y)|t_i]'\big)$ *and* $\rho(t_i, h_0) = |\phi_n(t_i) - \phi_Y(t_i|\pi, \theta)|$.

The proof of Theorem 6 directly follows from Theorem 4.1 (or Remark 6.1) in Chen and Pouzo (2015). Since $\phi(h) = \int \hat{m}(t_i, h)'\hat{m}(t_i, h)dt_i$ is a regular quadratic functional of h which satisfies the regularity conditions and note that $\frac{d\phi(h)}{dh}[p^{k(n)(.)}] = 2\int \hat{m}(t_i, h)p^{k(n)}(t_i)dt_i$, direct adoption of Theorem 4.1 in Chen and Pouzo (2015) yields the conclusion in Theorem 3. Details are in Appendix 2.D.

Remark 3. 1. *Theorem 6 shows the asymptotic normality of a plug-in PSMD estimator of a quadratic functional regardless of whether the latter is root-n estimable or not. The close form expression of* $||v_n^*||_{sd}^2$ *immediately leads to simple consistent plug-in sieve variance estimator as follows:*

$$||\hat{v}_n^*||_{n,sd} = \frac{2}{M_2 - M_1}\sum_{t_i=M_1}^{M_2} \hat{m}(t_i, h_0)p^{k(n)}(t_i)\hat{D}_n^- \hat{\Phi}_n \hat{D}_n^- \frac{2}{M_2 - M_1}\sum_{t_i=M_1}^{M_2} \hat{m}(t_i, h_0)p^{k(n)}(t_i)$$

where $\hat{D}_n = \frac{1}{M_2 - M_1}\sum_{t_i=M_1}^{M_2}(\frac{1}{N}\sum_{i=1}^{N} p^{k(n)}(y_i, t_j))^2$,

$$\hat{\Phi}_n = \frac{1}{M_2 - M_1}\sum_{t_i=M_1}^{M_2}\big(\frac{1}{N}\sum_{i=1}^{N} p^{k(n)}(y_i, t_j)\hat{\rho}^2(t_i, h)\frac{1}{N}\sum_{i=1}^{N} p^{k(n)}(y_i, t_j)\big)$$

and $\hat{\rho}(t_i, h) = |\hat{\phi}_n(t_i) - \tilde{\phi}_Y(t_i|\pi, \theta, \sigma_v)|$. *This simple sieve variance estimator is consistent according to Chen and Pouzo (2015).*

Remark 3. 2. *Theorem 6 is essential to test the zero inefficiency hypothesis using a quasi-likelihood ratio (QLR) test which is subsequently derived based on the objective function* $\hat{Q}_n(\hat{h})$. *For example, we could estimate a restricted zero-inefficiency stochastic frontier model and an unrestricted model as proposed in Section 3.2 which yield the estimates* $\hat{\theta^r}$, $\hat{\sigma^r}_v$ *and* $\hat{\pi}$, $\hat{\theta}$, $\hat{\sigma}_v$ *respectively. Then the corresponding* $\hat{h}_0(t) = [1 * \sum_{j=1}^{r_n} \hat{\theta^r}_j(\mathbf{i}sin(tY_j) + cos(tY_j))]\phi_z(t, \hat{\sigma^r}_v)$ *and* $\hat{h}_1(t) = [\hat{\pi}_1 + (1 - \hat{\pi}_1) * \sum_{j=1}^{r_n} \hat{\theta}_j(\mathbf{i}sin(tY_j) + cos(tY_j))]\phi_z(t, \hat{\sigma}_v)$ *and*

$\hat{Q}_n(\hat{h}_k) = \frac{1}{M_1 - M_2}\sum_{t_i=M_1}^{M_2} \hat{m}(t_i, \hat{h}_k)'\hat{m}(t_i, \hat{h}_k)$, $k = 0, 1$. *Hence we could test the following hypothesis:*

$$H_0 : \pi_1 = 0, \quad H_1 : \pi_1 > 0$$

with a QLR statistic test as

$$\frac{n(\hat{Q}_n(\hat{h}_0) - \hat{Q}_n(\hat{h}_1))}{||\hat{v}_n^*||_{n,sd}^2} \rightarrow_d 0.5 * \chi_1^2 + 0.5 * \chi_0^2.$$

The sieve variance estimator can be obtained as illustrated in Remark 3.1. The denominator $||\hat{v}_n^*||_{n,sd}^2$ *is to normalize the test statistic as the weight applied here* $w(t) = 1(M_1 \leq t \leq M_2)$ *is not the optimal weight. Due to that fact that null hypothesis lies on the boundary of the parameter space (*$0 \leq \pi_1 \leq 1$*), the asymptotic distribution of the QLR statistic is a* $50 : 50$ *mixture of a* χ_0^2 *distribution and a* χ_1^2 *distribution rather than the typical* χ_1^2 *distribution (Andrews, 2001). A straightforward bootstrap test statistics is feasible based on this argument. We implement the above test with a bootstrap procedure when dealing with a US bank data in the application.*

2.4 Tuning Parameter Selection

2.4.1 Selection of the Number of Bins $M(n)$

Based on Theorem 2 in Section 3.3, there is a trade-off between the number of discretized points $M(n)$ and the order of $||\hat{h}_n - h_0||_s$, while they jointly determine the convergence rate of $\hat{f}_U$. Lee et al.(2013) proposes using Akaike Information Criterion (AIC, Akaike, 1974) to choose the number of bins $M(n)$ which is computationally intensive. Alternatively here we propose $M(n) = \max(3, O(\ln(n)))$.

KSVK applied a similar rule of thumb on $M(n)$ but with a polynomial order. A useful criterion is to choose the coefficient before $\ln(n)$ as $c * \ln(n) = n^{0.2}$. The intuition here is that the number of bins $M(n)$ here plays a role similar to (the inverse of) the bandwidth in kernel estimation and the optimal bandwidth for a univariate kernel estimation is $h^* = O(n^{-0.2})$. The max operator here is to avoid over-small number of bins in small sample sizes. Our simulation results show that a small number of sieve terms can deliver adequate approximations across different distribution designs.

2.4.2 Selection of the Penalty Parameter λ_n

To derive the asymptotics in Section 3.3, the only requirement on the penalty parameter is $\lambda_n = o(1)$. In practice, the penalty parameter λ_n plays the role of a smoothing parameter in the proposed penalized sieve estimation and it is crucial for the estimated density parameters $\{\theta_j\}$. Here we propose a bootstrap procedure following KSVK and Florens et al. (2019) to choose λ_n:

1. For a given sample n, draw bootstrap random samples of size n, $(Y_1^{*,m}, \ldots, Y_n^{*,m})$ for $m = 1, \ldots, B$, by sampling with replacement from the n values Y_j in the original sample. For the heteroskedastic frontier estimation, at each value of w, draw bootstrap random samples of size n_b, $(Y_1^{*,m}, \ldots, Y_{n_b}^{*,m})$ for $m = 1, \ldots, B$, by sampling with replacement from the n_b values Y_j in the original sample such that $||W_j - w|| \leq b$.

2. For a given λ_n, compute the original estimators $\hat{\pi}_1$, $\hat{\theta}_\lambda(w)$ and its bootstrap analogue $\hat{\theta}^*_\lambda(w)$, for $m = 1, \ldots, B$. For the first step estimation in Case 2 and heteroskedastic frontier estimation, $\hat{\alpha}$ is computed instead.

3. Select the optimal λ_n over a grid of points (e.g., $\log \lambda_n = -1.5, -1, -0.5, 0., 0.5, 1, 1.5$) using the sum of two Root Mean Squared Error (RMSE): RMSE($\hat{\pi}_1$)+RMSE ($\hat{\theta}$).[13] For the first step estimation in Case 2 and heteroskedastic frontier estimation, we can use RMSE($\hat{\alpha}$).

Some other methods like the simulation-based approach by Lee et al. (2013) is also applicable here. One point worth mentioning is that the optimal choice of penalty parameter

[13] Here, RMSE($\hat{\theta}$):=$\sqrt{\sum_{j=1}^{M(n)} MSE(\hat{\theta}_j)}$.

λ_n largely depends on the target of each step. We may have different penalty parameters for the first step LS sieve estimation and the second step LS sieve estimation in the two-step procedures.

2.5 Simulation

In this section, we provide three simulation studies to investigate the finite sample properties of the proposed estimators which correspond to Case 1 (mixture distribution), Case 2 (mixture distribution with an intercept) and the heteroskedastic frontier (mixture distribution with heteroskedastic frontiers) illustrated in Section 2. The last case is motivated by the econometric literature on stochastic frontier models, especially the ZISF models as described in the introduction.

For the data generating process, we consider a simple model

$$logY = \alpha + U + V \tag{2.16}$$

with three cases: $\alpha = 0$, $\alpha = constant$ and $\alpha = \alpha(W)$ where W is an exogenous covariate. The inefficiency term U, which follows a mixture distribution, is generated as follows:

$$U \sim \begin{cases} 0 & \text{with probability } \pi_1 \\ U_c & \text{with probability } \pi_2 \end{cases}$$

where $0 \leq \pi_1 \leq 1$, $\pi_1 + \pi_2 = 1$ and U_c is a positive continuous random variable which follows a one-sided distribution. If $\pi_1 = 0$, Equation (2.16) reduces to a typical stochastic frontier model with a continuously distributed inefficiency U; if $\pi_1 = 1$, it reduces to a traditional ordinary least square (OLS) model. In the following simulations, we consider two cases with $\pi_1 = 0.4$ and $\pi_1 = 0.7$ respectively.

To satisfy the unimodality assumption mentioned in Case 2 and heteroskedastic frontier

Case in Section 2, we consider exponentially or half normally distributed inefficiency U_c as in the typical stochastic frontier model (Aigner et al.,1977; Meeusen and van den Broek ,1977) and the newly proposed ZISF model (Kumbhakar et al., 2013; Rho and Schmidt, 2015).[14] In the exponential case the density of U_c is

$$U_c \sim Exp(b) \Longleftrightarrow f_{U_c}(u) = \frac{1}{b}\exp(-\frac{u}{b}) \quad u > 0$$

For the half normal case, the density of U_c is

$$U_c \sim |N(0,\sigma^2)| \Longleftrightarrow f_{U_c}(u) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}}\exp(-\frac{u^2}{2\sigma^2}) \quad u > 0$$

Experiments with various noise-to-signal ratios under the above distribution and zero probability settings are considered. Details are provided in Table 1. To facilitate the comparison, we increase the noise-to-signal ratio (i.e., $\rho_{nts} := \sigma_v/\sigma_u$) by doubling the variance of the random noise V and fixing the (mixture) distribution of inefficiency U.

For all the simulations, the following three measurements are reported for the point estimates $\hat{\pi}_1$:

$$\begin{aligned} Bias^2(\hat{\pi}_1) &= (\frac{1}{R}\sum_{r=1}^{R}\hat{\pi}_{1r} - \pi_1)^2 \\ Var(\hat{\pi}_1) &= \frac{1}{R}\sum_{r=1}^{R}(\hat{\pi}_{1r} - \frac{1}{R}\sum_{r=1}^{R}\hat{\pi}_{1r})^2 \\ MSE(\hat{\pi}_1) &= \frac{1}{R}\sum_{r=1}^{R}(\hat{\pi}_{1r} - \pi_1)^2 \end{aligned}$$

where $R = 100$ is the replication times. Similar measurements for frontier estimates $\hat{\alpha}$. For

[14] For the simulation design for Case 1, we do not need unimodality assumption.

the continuous density estimator $\hat{f}_c(u)$(e.g., the estimator of $f_{U_c}(u)$), we report

$$\int Bias^2(\hat{f}_c(u))du = h\sum_{j=1}^{M(n)}\big(\frac{1}{R}\sum_{r=1}^{R}\hat{f}_{cr}(u_j) - f_c(u_j)\big)^2$$

$$\int Var(\hat{f}_c(u))du = h\sum_{j=1}^{M(n)}\frac{1}{R}\sum_{r=1}^{R}\big(\hat{f}_{cr}(u_j) - \frac{1}{R}\sum_{j=1}^{M(n)}\hat{f}_{cr}(u_j)\big)^2$$

$$\int MSE(\hat{f}_c(u))du = h\sum_{j=1}^{M(n)}\frac{1}{R}\sum_{r=1}^{R}\big(\hat{f}_{cr}(u_j) - f_c(u_j)\big)^2$$

where $h := u_{j+1} - u_j$ and $M(n)$ is the number of evaluating points defined in Section 2. Here we choose equidistant points so $h = \frac{M_2 - M_1}{M(n)}$, where $M_1 = \min(\log Y)$ and $M_2 = \max(\log Y)$. We choose $M(n) = \max(3, c * \ln(n))$ with $c = 1$ as we proposed in Section 4.

2.5.1 Simulation Design I: $logY = U + V$

In this simulation experiment, we consider the basic setting with Case 1, namely, $\alpha = 0$. As mentioned before, the proposed method extends the LS sieve estimation in Lee et al. (2015) to the case with unknown variance but with known point mass. Simulations with three sample sizes are conducted: $N = 500$ (small), $N = 1000$ (medium) and $N = 3000$ (large), for the two sets of noise-to-signal ratio scenarios (i.e., $V \sim N(0, 0.2^2)$ and $V \sim N(0, 0.4^2)$) in Table 1.

Table 2 reports the results for simulation design I under the first noise-to-signal ratio set, e.g., $V \sim N(0, 0.2^2)$, with different sample sizes. The first panel of the table represents the case with $U_c \sim Exp(1)$ where U_c is the continuous portion of the inefficiency, and the second panel is for the case with $U_c \sim |N(0, 1)|$. In the last row of each panel, optimal penalty parameter λ_n are reported. These are chosen by minimizing the sum of RMSE(π_1) and RMSE($\hat{\theta}$). Vertically, the three pair columns correspond to the three scenarios with aforementioned sample sizes. Each column in a pair represents a specific case for the zero point mass probability, specifically, $\pi_1 = 0.4$ and $\pi_1 = 0.7$. Looking through the table, we can observe that the proposed estimators behave as expected. Horizontally, when the sample size increases, the estimation improves for both π_1 and $f_c(u)$. Vertically, the proposed estimators ($\hat{\pi}_1$ and $\hat{f}_c(u)$) perform better with half-normally distributed U_c than with exponentially distributed U_c in general. When zero point mass probability increases, exponentially distributed U_c yields a better estimator of π_1 but a worse estimator of $f_c(u)$ compared with those with $U_c \sim |N(0,1)|$, see, e.g., for the case of $\pi_1 = 0.7$ with $N = 1000$. For the penalty parameter, optimal value for exponentially distributed U_c is more varied than those with half-normally distributed U_c. This may partially explain the variability of the estimates with exponentially distributed U_c.

Similarly, Table 3 reports the results for simulation design I under the large noise-to-signal ratio setting, e.g., $V \sim N(0, 0.4^2)$, with different sample sizes. Layout in Table 3 is the same as that in Table 2 for ease of comparison. All the patterns in Table 2 still exist in Table 3. Compared with Table 2, the absolute values for all three measurements are relatively larger due to the increased (doubled) noise-to-signal ratios. However, we can obtain more precise and stable estimates with half-normally distributed U_c than those with exponentially distributed U_c.

2.5.2 Simulation Design II: $\log Y = \alpha + U + V$

In simulation design II, we consider Case 2 in Section 2.1, namely, α is a nonzero constant. We choose $\alpha = 1$ here. The estimation is accomplished deploying the two-step procedure proposed in Section 2 under the unimodality assumption. The first step is to estimate the mode of $\alpha - U$ without considering the zero point mass, then subtract the estimated mode (which is also the unknown intercept α under unimodality) from $logY$, and apply the same procedure as in Case 1 to estimate the zero point mass probability π_1 and the continuous density $f_c(u)$. To improve performance of the density estimators, roughness penalties are considered for both steps.

We consider a small sample size N=1000 and a large sample size N=3000 under two sets of noise-to-signal ratio scenarios. Table 3.4 reports the results under the small noise-to-signal setting, namely, $V \sim N(0, 0.2^2)$. There are two row panels in the table: each represents one estimation step. Four pair columns in the table correspond to the four combinations of two sample sizes and two distributions for U_c. Each column in the pair represents a specific case with certain zero point mass probability, specifically, $\pi_1 = 0.4$ or $\pi_1 = 0.7$ as before. Horizontally, the proposed estimators for α, π_1 and $f_c(u)$ converge in MSE as sample size increases with fixed ρ_{nts}, e.g., pair column one VS pair column three. Scenario with $\pi_1 = 0.4$ and $U_c \sim |N(0, 1)|$ seems to be the hardest case for estimation.[15] The MSE of estimated intercept increases as sample size increases from 1000 to 3000 (column 3 VS column 7). The reason is that in the first step, the mode (also the intercept) is not precisely estimated due to the fact that the evaluating points of the continuous density are a little far from the true mode in the large sample size N=3000 (column 7). This consequently results in the under-performance of the estimates for π_1 and $f_c(u)$ in the second step. Vertically, the mode estimates in the first step are more precise than the second step estimates, as expected. Large zero point mass probability (i.e., π_1) tends to result in a better estimation of the mode as well as better fit of the target continuous density $f_c(u)$.

Table 2.5 reports the simulation results in the same layout with a large noise-to-signal setting, namely, $V \sim N(0, 0.4^2)$. All patterns in Table 4 still hold in Table 5. The hardest cases (the case with $\pi_1 = 0.4$ and $U_c \sim |N(0, 1)|$, column 3 and column 7) perform worse than those with smaller noise-to-signal ratios in Table 4. Table 5 and Table 4 are comparable with each other in the remaining cases and do not show significant differences for the change of noise variance. This is due to the optimal choices of penalty parameters in both steps, which absorb most of the differences caused by the increasing noise variance.[16]

[15] One reason may be that it is very hard to estimate the mode with a mixture distribution containing a small probability of point mass mixed with a half normal distribution as what is observed in Table 2 and Table 3.

[16] Note that the optimal choice of penalty parameter λ_n is very different between Table 4 and Table 5. It may absorb much of the differences which should exist without penalty optimization.

2.5.3 Simulation Design III: $\log Y = \alpha(W) + U + V$

In this simulation experiment, we consider the heteroskedastic frontier case with $\alpha = \alpha(W)$ in Section 2, which is motivated by the zero-inefficiency stochastic frontier model. Specifically, we consider a cost frontier model:

$$logY = \alpha(W) + U + V$$

where $\alpha(W) = 2W^2$ reflects the convexity of a cost function, the continuous part of cost inefficiency $U_c \sim Exp(1/3)$ with $\pi_1 = 0.4$ and 0.7, $V \sim 0.2 * N(0, 1)$ and $W \sim Uniform[0, 1]$ representing the output. This is similar to the setting in KSVK, Hall and Simar (2002). However, there are three major differences in the present setting: firstly, this is a cost frontier design with convex $\alpha(W)$ and non-negative inefficiency rather than a production frontier with concave $\alpha(W)$ and non-positive inefficiency in KSVK[17] ; secondly, we consider a zero-inefficiency stochastic frontier in present paper, which is different and exclusive from above references; lastly, the present setting is more noisy compared with previous literature, that is, the noise-to-signal ratio $\rho_{nts} = 0.60$ here, comparing the noise V and continuous inefficiency U_c.

Based on the procedures in Section 2.2, we transform $logY_i$ by a local least squares regression and then implement the two-step procedure for Case 2 (i.e., $\alpha \neq 0$) to estimate $\alpha(w)$, $\pi_1(w)$ and $\theta(w)$. For simplicity, we implement the first transformation proposed in Section 2.2, i.e., $\log \tilde{Y_i} := \log Y_i - \beta^T(w_0)(W_i - w_0)$, and focus on the heteroskedastic frontiers $\alpha(w)$ here.[18] We consider two sample sizes $N = 500$ and $N = 1000$. For $N = 500$, the frontiers are evaluated at 5 quintiles of W from 0.1 to 0.9, and 10 quintiles for $N = 1000$. The penalty parameters are chosen by minimizing the corresponding RMSE with 50 replications. Details about the selected penalty parameters are reported in Table 6.

[17] The methodology proposed in KSVK is not readily applied to a cost frontier. But a modified version is applicable. See Cai (2020).

[18] One could also look at the conditional expected inefficiency $\hat{E}(U|W = w)$ which may contain some useful information on the inefficiency.

Figure 3.1-3.4 shows the four combinations for small/large sample sizes and low/high probability of zero point mass. The blue lines are the estimated frontiers and the smooth red lines are the true frontier. A quick look at the figures confirms that the estimated heteroskedastic frontiers are going in the expected directions: for a fixed level of zero probability level π_1, as the sample size N increase, the fit of frontier gets better. One reason is that we could evaluate the heteroskedastic frontiers at more points as N increases, and another reason is that each evaluated frontier point is estimated more precisely as there are more observations in each quintile with N increasing. Second, comparing Figure 3.1 (Figure 3.2) with Figure 3.3 (Figure 3.4), the estimated frontiers perform better with a high level of zero point mass probability, namely, large π_1. This follows from the two-step procedures in Case 2. We also find large zero point mass probability leads to a better estimation of the mode, i.e., $\alpha(w)$, in Table 4 and 5. Finally we observe that the estimated frontiers perform very well with confounding zero point mass even in a decent sample size such as $N = 1000$, concerning that multiple steps are implemented.[19]

2.6 Application

In this section, we apply the proposed method to data from 6,010 US banks observed in 2005. For each bank, we have data on the total cost Y_i which includes labor salary, interest of borrowed funds and depreciation of physical capital along with total output W_i consisting of three output quantities: consumer loans, commercial loans and securities. All outputs are deflated by the Consumer Price Index (CPI) to the base year of 1988. This data set comes from the *call report* in the Chicago Fed website and has been used in a different context (nonparametric panel heteroskedastic frontiers) in Cai et al. (2019).[20]

We are interested in the cost frontier, conditional on the value $W = w$, from the sample

[19] Here we just connect the evaluated points on the frontier by straight lines in the figures. One can implement spline interpolation to get a smoother and more precise frontier as Florens, Simar and Van Keilegom (2019) did.

[20] For details of the data, please refer to Cai et al. (2019) or Feng and Serletis (2009).

of $N = 6010$ observations (Y_i, W_i), where $i = 1, \dots, N$ and Y_i is the noisy version of the true total cost. We consider the following stochastic frontier model:

$$\log Y_i = \alpha(\log W_i) + U_i + V_i, \quad i = 1, \dots, N \tag{2.17}$$

where $V \sim N(0, \sigma_v^2)$ is the random noise or shock with unknown variance,

$$U \sim \begin{cases} 0 & \text{with probability } \pi_1(W) \\ U_c & \text{with probability } \pi_2(W) \end{cases}$$

where $0 \leq \pi_1(W) \leq 1$, $\pi_1 + \pi_2 = 1$ and U_c is a continuous positive random variable which follows a one-sided distribution. Our primary interests focus on the heteroskedastic frontiers $\alpha(logW)$ and the zero inefficiency probability $\pi_1(W)$.

Estimation consists of two steps. The localization of first-step least square estimation is done by choosing 10 equidistant groups of the log total output $logW$ (with a bandwidth $b = 1.254$). We also try 6, 8, and 12 groups which yields similar results. As explained earlier, we use a local linear approximation suggested by KSVK and Hall and Simar (2002), to estimate the cost frontier around $W = w$. Specifically, we use the local linear transformation of $\log Y_i$: $\log \tilde{Y_i} := \log Y_i - \beta^T(w)(W_i - w)$ obtained from application of local least squares, to estimate the frontiers evaluated at each of $W = w$ points in the second step (see details in Section 2.2).

The estimates are computed for a fixed grid of 10 values of $logw$. Within each equidistant group, the number of evaluation points is chosen by $M(n) = \max(3, \ln(n_b))$ as in the simulations. We apply the proposed procedures in Section 2.2 to estimate the heteroskedastic frontiers and zero inefficiency probabilities. For the penalty parameters in both steps, we search on a grid of $log\lambda = -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2$. The optimal penalty parameters $log\lambda_{1n}$ and $log\lambda_{2n}$ are obtained by minimizing the bootstrap estimate of RMSE($\hat{\alpha}$) and RMSE($\hat{\pi}_1$)+RMSE ($\hat{\theta}$) respectively as discussed earlier (100 bootstrap replications at each evaluation point w).

The results are displayed in Figure 3.5 and the estimated 10 values of the frontiers are shown in the blue circles. We connect the 10 blue circles by dash blue line to form the final frontier. We can see how the estimated cost frontier "envelope" the cloud of bank points but allowing some observations stay below the frontiers at certain intervals of the total output. This is very different from the typical stochastic frontier model which does not consider the (potential) zero inefficiency, i.e., the existence of (most) efficient firms. Based on the estimates of zero inefficiency probability $\hat{\pi}_1$, 28.5% of the banks are estimated to be fully cost efficient. About 8.49% of the banks around the fourth equidistant point of $logw$, namely, the banks with $logw \in [11.58, 12.78]$, are (most) efficient. However, 60% of banks around the seventh evaluating point, namely the banks with $logw \in [14.90, 16.10]$, are (most) efficient. Details on the ten subgroup analysis along with the proposed QLR test statistics are reported in Table 2.7.[21] The mean cost inefficiency (i.e., $E(U|W)$) decreases as the output increases. The QLR test shows that we can reject the full inefficiency null hypothesis, i.e., $H_0 : \pi_1 = 0$ (no zero inefficiency banks) at 5% significance level in seven out of the ten groups: Group 2-3 and Group 6-10. Statistic evidence suggests that there are (cost) efficient banks in the groups with bigger output and medium size banks tend to be cost inefficient as a whole.

We do find some evidence suggesting the presence of scale economies in the US bank industry in 2005. Big banks are more likely to be fully efficient, though some medium sized banks also obtain low levels of estimated cost inefficiency. This is intuitive when considering the consolidation process of US banks over the 1998-2005 period, a fact mentioned in Feng and Serletis (2009) who used the same data for their analysis.

2.7 Conclusion

This paper presents a novel nonparametric analysis of the model $Y = \alpha + U + V$ where U follows a mixture distribution with zero point mass and a continuous one-sided distribution,

[21] The variance of test statistics are obtained from a bootstrap procedure with 100 replications.

and $V \sim N(0, \sigma_v^2)$ with unknown variance. We also consider extension to the newly proposed zero-inefficiency stochastic frontier model (Kumbhakar et al., 2013). A penalized minimum distance estimation procedure based on the least square sieve methods studied in Lee et al. (2015) is proposed and the asymptotic properties of the proposed estimators are derived following the theorems of Chen and Pouzo (2012, 2015). Moreover, we propose an useful QLR test on the zero inefficiency hypothesis based on the penalized sieve least square estimator. Several practical procedures are proposed concerning different cases and the power of our estimators is shown via simulations and an application to a US bank data.

In this paper, we assume the random noise $V \sim N(0, \sigma_v^2)$ for convenience which may not be true though we leave its variance unspecified. We can easily extend the distribution assumption of V to be any one parameter distribution with unknown variance and symmetric around zero, for instance, $V \sim Laplace(b)$, as Horrace and Parmeter (2018) proposed in the context of stochastic frontier model. In addition, even though we estimate the zero-inefficiency frontiers i.e., the minimum (maximum) value of the cost (production) with a consideration of the existence of most efficient firms, we do not investigate its inference or related test in this paper. These are left for future research.

Table 2.1: Noise to signal ratios for different scenarios

	$V \sim N(0, 0.2^2)$				$V \sim N(0, 0.4^2)$			
	$U_c \sim Exp(1)$		$U_c \sim \lvert N(0,1)\rvert$		$U_c \sim Exp(1)$		$U_c \sim \lvert N(0,1)\rvert$	
	$\pi_1 = 0.4$	$\pi_1 = 0.7$	$\pi_1 = 0.4$	$\pi_1 = 0.7$	$\pi_1 = 0.4$	$\pi_1 = 0.7$	$\pi_1 = 0.4$	$\pi_1 = 0.7$
$Var(u)$	0.840	0.510	0.371	0.243	0.840	0.510	0.371	0.243
ρ_{nts}	0.22	0.28	0.33	0.41	0.44	0.56	0.66	0.82

Notes: The noise-to-signal ratio is defined as $\rho_{nts} := \sigma_v/\sigma_u$. U_c is the continuous part of inefficiency U.

Table 2.2: DESIGN I, $\log Y = U + V$, $V \sim N(0, 0.2^2)$

		$N = 500$		$N = 1000$		$N = 3000$	
		$\pi_1 = 0.4$	$\pi_1 = 0.7$	$\pi_1 = 0.4$	$\pi_1 = 0.7$	$\pi_1 = 0.4$	$\pi_1 = 0.7$
	$Bias^2(\hat{\pi}_1)$	0.0785	0.0242	0.0734	0.0223	0.0701	0.0190
	$Var(\hat{\pi}_1)$	0.0023	0.0007	0.0015	0.0006	0.0006	0.0003
	$MSE(\hat{\pi}_1)$	0.0808	0.0249	0.0749	0.0229	0.0708	0.0193
$U_c \sim Exp(1)$	$\int Bias^2(\hat{f}_c(u))du$	0.0435	0.0217	0.0292	0.0215	0.0213	0.0143
	$\int Var(\hat{f}_c(u))du$	0.0031	0.0027	0.0010	0.0016	0.0002	0.0005
	$\int MSE(\hat{f}_c(u))du$	0.0466	0.0245	0.0302	0.0230	0.0215	0.0148
	$\log(\lambda_n)$	0.5	-0.5	0	-0.5	0	-0.5
	$Bias^2(\hat{\pi}_1)$	0.0318	0.0345	0.0306	0.0324	0.0212	0.0311
	$Var(\hat{\pi}_1)$	0.0037	0.0010	0.0025	0.0003	0.0018	0.0003
	$MSE(\hat{\pi}_1)$	0.0354	0.0354	0.0330	0.0326	0.0230	0.0314
$U_c \sim \lvert N(0,1) \rvert$	$\int Bias^2(\hat{f}_c(u))du$	0.0189	0.0089	0.0138	0.0047	0.0098	0.0027
	$\int Var(\hat{f}_c(u))du$	0.0005	0.0022	0.0002	0.0009	0.0001	0.0004
	$\int MSE(\hat{f}_c(u))du$	0.0194	0.0110	0.0141	0.0057	0.0099	0.0031
	$\log(\lambda_n)$	0	0	0	0	0	0

Notes: The number of evaluation points $M(n) = \max(3, \ln(n))$. U comes from a mixed distribution with π_1 probability being a point mass at zero and $1 - \pi_1$ probability being a continuous random variable U_c with a pdf f_c. For the tuning parameter λ_n, we search with a grid $\log(\lambda_n) = -1, -0.5, 0, 0.5, 1$.

Table 2.3: DESIGN I, $\log Y = U + V$, $V \sim N(0, 0.4^2)$

		$N = 500$		$N = 1000$		$N = 3000$	
		$\pi_1 = 0.4$	$\pi_1 = 0.7$	$\pi_1 = 0.4$	$\pi_1 = 0.7$	$\pi_1 = 0.4$	$\pi_1 = 0.7$
	$Bias^2(\hat{\pi}_1)$	0.1080	0.0298	0.0909	0.0307	0.0906	0.0314
	$Var(\hat{\pi}_1)$	0.0024	0.0005	0.0007	0.0002	0.0003	0.0001
	$MSE(\hat{\pi}_1)$	0.1104	0.0303	0.0916	0.0308	0.0908	0.0315
$U_c \sim Exp(1)$	$\int Bias^2(\hat{f}_c(u))du$	0.0706	0.0582	0.0449	0.0307	0.0248	0.0268
	$\int Var(\hat{f}_c(u))du$	0.0040	0.0044	0.0007	0.0026	0.0006	0.0019
	$\int MSE(\hat{f}_c(u))du$	0.0746	0.0626	0.0457	0.0333	0.0254	0.0286
	$\log(\lambda_n)$	0.5	-0.5	0.5	-0.5	0	-0.5
	$Bias^2(\hat{\pi}_1)$	0.0434	0.0431	0.0219	0.0386	0.0083	0.0376
	$Var(\hat{\pi}_1)$	0.0036	0.0009	0.0032	0.0005	0.0029	0.0003
	$MSE(\hat{\pi}_1)$	0.0470	0.0440	0.0250	0.0391	0.0112	0.0379
$U_c \sim \lvert N(0,1) \rvert$	$\int Bias^2(\hat{f}_c(u))du$	0.0328	0.0156	0.0152	0.0087	0.0106	0.0045
	$\int Var(\hat{f}_c(u))du$	0.0008	0.0050	0.0004	0.0019	0.0002	0.0004
	$\int MSE(\hat{f}_c(u))du$	0.0336	0.0205	0.0156	0.0106	0.0108	0.0049
	$\log(\lambda_n)$	0	-0.5	0	0	0	0

Notes: The number of evaluation points $M(n) = \max(3, \ln(n))$. U comes from a mixed distribution with π_1 probability being a point mass at zero and $1 - \pi_1$ probability being a continuous random variable U_c with a pdf f_c. For the tuning parameter λ_n, we search with a grid $\log(\lambda_n) = -1, -0.5, 0, 0.5, 1$.

Table 2.4: DESIGN II, $\log Y = \alpha + U + V$, $V \sim N(0, 0.2^2)$, $\alpha = 1$

	$N = 1000$				$N = 3000$			
	$U_c \sim Exp(1)$		$U_c \sim \|N(0,1)\|$		$U_c \sim Exp(1)$		$U_c \sim \|N(0,1)\|$	
	$\pi_1 = 0.4$	$\pi_1 = 0.7$	$\pi_1 = 0.4$	$\pi_1 = 0.7$	$\pi_1 = 0.4$	$\pi_1 = 0.7$	$\pi_1 = 0.4$	$\pi_1 = 0.7$
$Bias^2(\hat{\alpha})$	0.0002	0.0048	0.0170	0.0027	0.0027	0.0105	0.0275	0.0001
$Var(\hat{\alpha})$	0.0350	0.0313	0.0147	0.0090	0.0231	0.0296	0.0310	0.0217
$MSE(\hat{\alpha})$	0.0352	0.0361	0.0318	0.0118	0.0258	0.0401	0.0585	0.0218
$\log(\lambda_{1n})$	-1	-1.5	0	-1.5	-0.5	-0.5	0	-1.5
$Bias^2(\hat{\pi}_1)$	0.0722	0.0031	0.1192	0.0428	0.0542	0.0050	0.1349	0.0076
$Var(\hat{\pi}_1)$	0.0153	0.0198	0.0173	0.0056	0.0121	0.0162	0.0253	0.0163
$MSE(\hat{\pi}_1)$	0.0875	0.0228	0.1365	0.0483	0.0663	0.0212	0.1602	0.0239
$\int Bias^2(\hat{f}_c(u))du$	0.0436	0.0177	0.0182	0.0358	0.0351	0.0130	0.0117	0.0063
$\int Var(\hat{f}_c(u))du$	0.0225	0.0460	0.0115	0.0038	0.0168	0.0338	0.0184	0.0114
$\int MSE(\hat{f}_c(u))du$	0.0661	0.0637	0.0297	0.0395	0.0519	0.0468	0.0301	0.0177
$\log(\lambda_{2n})$	0	0	0.5	0.5	0	0	0	0.5

Notes: The number of evaluation points $M(n) = \max(3, \ln(n))$. U comes from a mixed distribution with π_1 probability being a point mass at zero and $1 - \pi_1$ probability being a continuous random variable U_c with a pdf f_c. For the tuning parameter λ_n, we search with a grid $\log(\lambda_n) = -1.5, -1, -0.5, 0, 0.5, 1, 1.5$.

Table 2.5: DESIGN II, $\log Y = \alpha + U + V$, $V \sim N(0, 0.4^2)$, $\alpha = 1$

	$N = 1000$				$N = 3000$			
	$U_c \sim Exp(1)$		$U_c \sim \vert N(0,1)\vert$		$U_c \sim Exp(1)$		$U_c \sim \vert N(0,1)\vert$	
	$\pi_1 = 0.4$	$\pi_1 = 0.7$	$\pi_1 = 0.4$	$\pi_1 = 0.7$	$\pi_1 = 0.4$	$\pi_1 = 0.7$	$\pi_1 = 0.4$	$\pi_1 = 0.7$
$Bias^2(\hat{\alpha})$	0.0001	0.0048	0.0299	0.0002	0.0062	0.0174	0.0001	0.0001
$Var(\hat{\alpha})$	0.0195	0.0282	0.0210	0.0140	0.0187	0.0177	0.0518	0.0140
$MSE(\hat{\alpha})$	0.0196	0.0331	0.0508	0.0142	0.0249	0.0351	0.0519	0.0141
$\log(\lambda_{1n})$	0.5	-0.5	0	0	-0.5	-0.5	-0.5	0
$Bias^2(\hat{\pi}_1)$	0.0725	0.0049	0.1239	0.0233	0.0448	0.0009	0.0120	0.0171
$Var(\hat{\pi}_1)$	0.0133	0.0172	0.0375	0.0106	0.0153	0.0148	0.0897	0.0121
$MSE(\hat{\pi}_1)$	0.0858	0.0222	0.1614	0.0339	0.0600	0.0157	0.1017	0.0292
$\int Bias^2(\hat{f}_c(u))du$	0.0532	0.0281	0.0219	0.0101	0.0406	0.0196	0.0222	0.0073
$\int Var(\hat{f}_c(u))du$	0.0107	0.0302	0.0240	0.0112	0.0080	0.0210	0.0452	0.0118
$\int MSE(\hat{f}_c(u))du$	0.0639	0.0584	0.0459	0.0213	0.0486	0.0406	0.0675	0.0191
$\log(\lambda_{2n})$	0.5	0	-0.5	0	0.5	0	-0.5	0

Notes: The number of evaluation points $M(n) = \max(3, \ln(n))$. U comes from a mixed distribution with π_1 probability being a point mass at zero and $1 - \pi_1$ probability being a continuous random variable U_c with a pdf f_c. For the tuning parameter λ_n, we search with a grid $\log(\lambda_n) = -1, -0.5, 0, 0.5, 1$.

Table 2.6: Selected Penalty Parameters for Heteroskedastic Frontier Estimation

	$log\lambda_{1n}$	$log\lambda_{2n}$
$N = 500, \pi_1 = 0.4$	{-0.5,-1,-1,-1,-1}	{0.5,1,1,1,1}
$N = 500, \pi_1 = 0.7$	{-1,1,-1,1,0}	{1,0.5,0.5,0,1}
$N = 1000, \pi_1 = 0.4$	{-1,0,-0.5,-1,-0.5,-1,-1,-1,0.5,0.5}	{1,1,1,1,1,0,0,1,1,0.5}
$N = 1000, \pi_1 = 0.7$	{-0.5,-1,1,0.5,-1,0, 0.5,-1,0,-1}	{1,1,1,0.5,1,1,1,0,1,1}

Notes: $log\lambda_{1n}$ and $log\lambda_{2n}$ stand for the penalty parameter for the first step and the second step respectively. There are 5 quintiles for $N = 500$ and 10 quintiles for $N = 1000$. Each is chosen based on the bootstrap procedure with 50 replications.

Figure 2.1: Heteroskedastic Frontiers with N=500, $\pi_1 = 0.4$

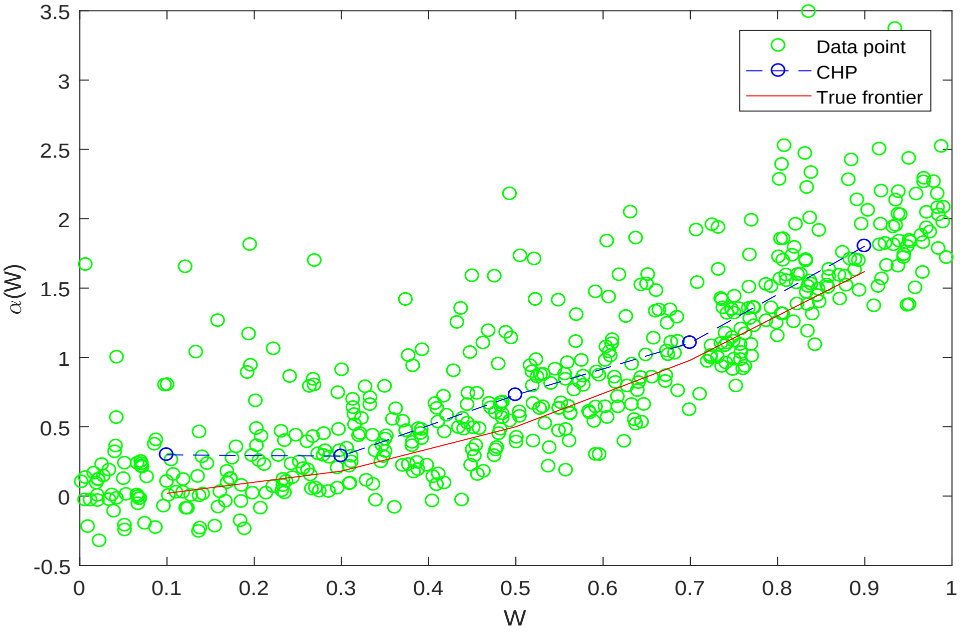


Figure 2.2: Heteroskedastic Frontiers with N=1000, $\pi_1 = 0.4$

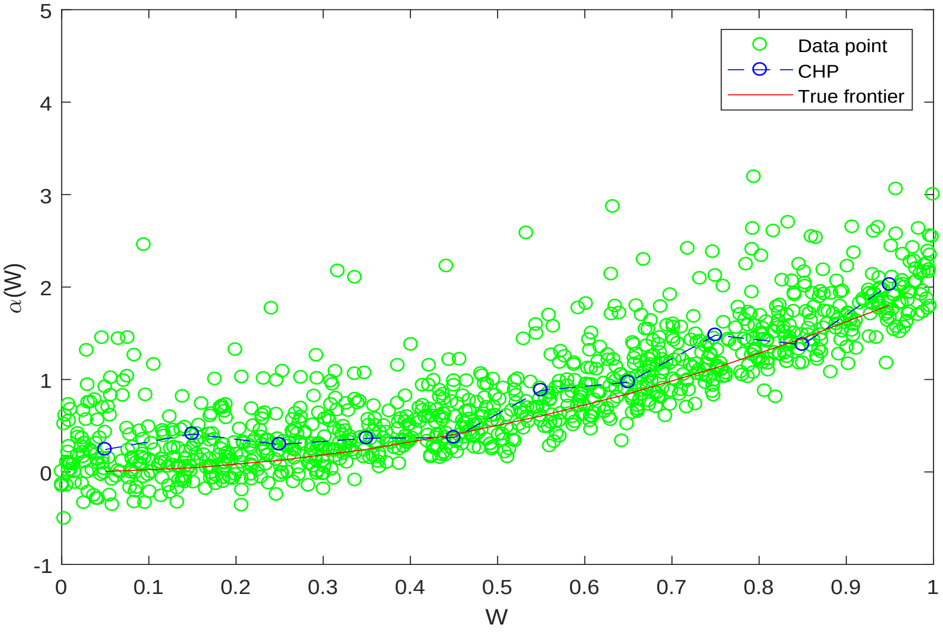


Figure 2.3: Heteroskedastic Frontiers with N=500, $\pi_1 = 0.7$

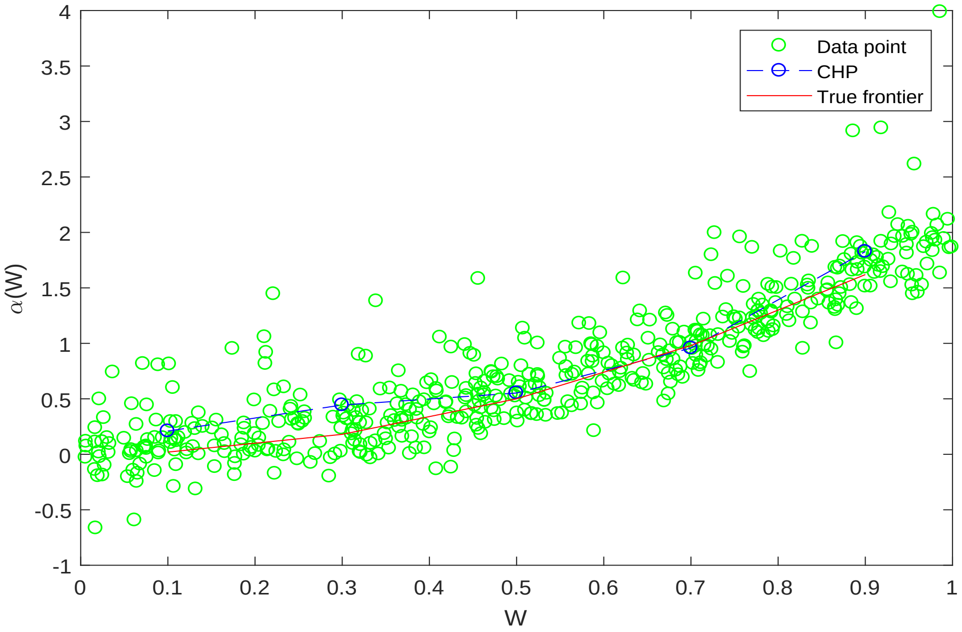


Figure 2.4: Heteroskedastic Frontiers with N=1000, $\pi_1 = 0.7$

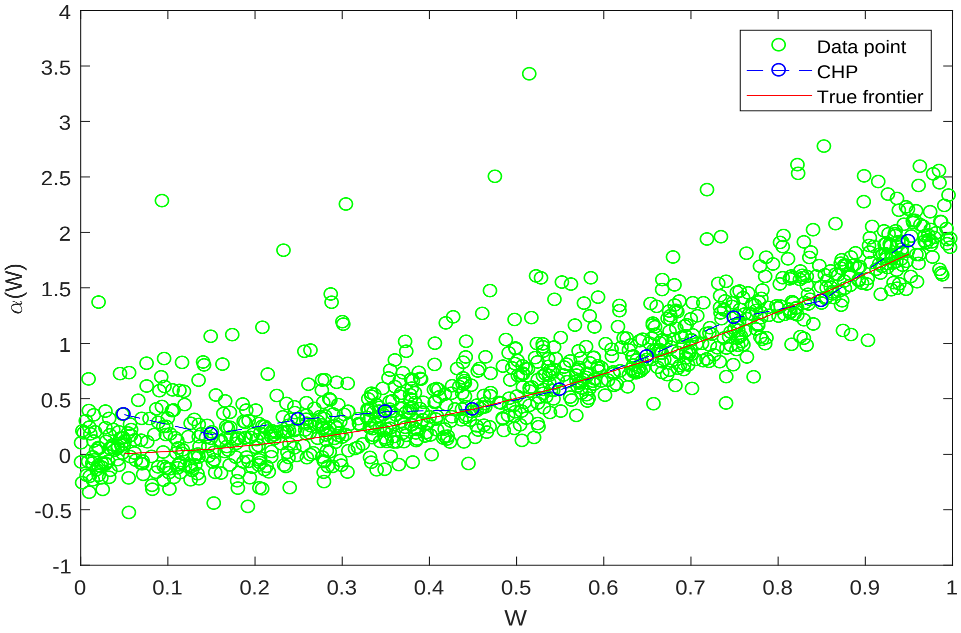


Figure 2.5: Estimate of Cost Frontiers With US Bank Data

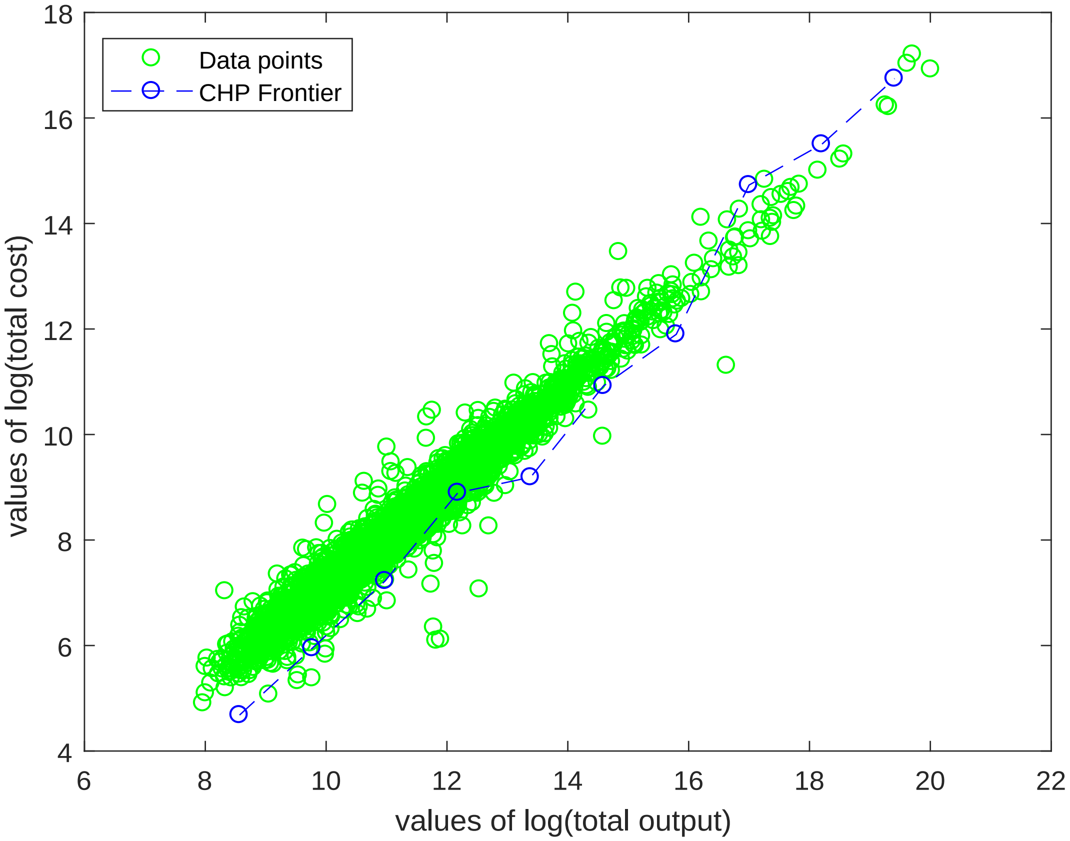


Table 2.7: ZISF Analysis for Subgroups on US Banks in 2005

Subgroups	1	2	3	4	5	6	7	8	9	10
$\log(W)$	8.57	9.77	10.98	12.18	13.39	14.59	15.79	17.00	18.20	19.41
$E(U\|W)$	1.38	1.01	0.75	0.16	0.99	0.49	0.63	-0.96	-0.74	-0.03
Inefficiency portion (%)	22.7	14.5	9.46	1.71	9.73	4.31	5.00	-6.97	-5.04	-0.16
QLR Test	0.829	0.0001	58.84	2.174	2.376	69.11	13.56	13.97	19.91	8.471
Number of Obs	222	1370	2230	1499	467	128	56	23	9	5

Notes: The inefficiency portion is calculated by $E(U|W)/E(Y)$ for each of the groups, measured in percentage. The negative inefficiency portion means efficiency gains from the random shocks. For a $0.5\chi_1^2 + 0.5\chi_0^2$ distribution, the 95% confidence interval is $(0.001, 2, 512)$.

Appendices

2.A Iterative EM Algorithm

We focus on the iterative EM algorithm used to solve the proposed LS sieve method based on characteristic functions (LS-ChF) in Section 2.1.2. The iterative EM algorithm is an extension of the iterative minimum algorithm in Lee et al. (2015). The main differences are that now we deal with a case with *known* zero point mass but *unknown* noise variance. For the LS sieve method based on cumulative probability functions (LS-CDF), we can derive similarly.

The goal is to estimate the unknown parameters: zero point mass probability π_1, the discretized continuous density θ and the unknown noise variance σ_v^2 by minimizing the objective function $S_{chf}(\pi, \theta, \sigma_v)$, which is defined by equation (2.7):

$$\begin{aligned}(\hat{\pi}, \hat{\theta}, \hat{\sigma_v}) &= \underset{\pi\in\Pi,\theta\in\Theta,\sigma\in\Sigma}{\arg\min}\; S_{chf}(\pi, \theta, \sigma_v) \\ &= \underset{\pi\in\Pi,\theta\in\Theta,\sigma\in\Sigma}{\arg\min} \int |\hat{\varphi}_n(t) - \tilde{\varphi}_Y(t|\pi, \theta, \sigma_v)|^2 w(t)dt\end{aligned}$$

where $\Pi = \{(\pi_1, \pi_2) : \pi_1 \geq 0, \pi_2 \geq 0, \pi_1 + \pi_2 = 1\}$, $\Theta = \{(\theta_1, \dots, \theta_M) : \theta_j \geq 0, j = 1, \dots, M(n), \sum_{j=1}^{M}(n)\theta_j = 1\}$, $\Sigma = \{\sigma_v : 0 < \sigma_v^2 \leq Var(logY)\}$. $\hat{\varphi}_n(t) = \frac{1}{n}\sum_{k=1}^{n} exp(it \cdot$

$logY_k$) is the empirical characteristic function of $logY$ and

$$\tilde{\varphi}_Y(t|\pi,\theta,\sigma_v) = \pi_1 e^{it\cdot 0}\varphi_V(t) + \pi_2 \sum_{j=1}^{M(n)} \theta_j e^{itu_j}\varphi_V(t) = \pi_1\varphi_V(t) + \pi_2 \sum_{j=1}^{M(n)} \theta_j e^{itu_j}\varphi_V(t)$$

where $\varphi_V(t) := E(e^{itV}) = \int e^{itv} f_v(v)dv = exp(-0.5\sigma_v^2 t^2)$ as $V \sim N(0,\sigma_v^2)$.

Suppose σ_v and θ are given. Then $S_{chf}(\pi|\theta,\sigma_v)$ is a quadratic form in π. In addition, the parameter space of π is given by

$$\Pi = \{(\pi_1,\pi_2) : \pi_1 \geq 0, \pi_2 \geq 0, \pi_1 + \pi_2 = 1\}$$

which is a compact set. From these facts, we can confirm the existence and uniqueness of the minimum of $S_{chf}(\pi|\theta,\sigma_v)$. The situation is exactly the same when we fixed σ_v and π. The differences are coefficients of the optimization problem, and the fact that θ is defined on

$$\Theta = \{(\theta_1,\ldots,\theta_M) : \theta_j \geq 0, j = 1,\ldots,M(n), \sum_{j=1}^{M}(n)\theta_j = 1\}$$

where $M(n) = max(3, \ln(n))$.

The situation is a little different when we fixed π and θ and try to derive an estimate of σ_v. As we can see from the above minimization problem, the objective function S_{chf} is quadratic in $\varphi_V(t)$ rather than σ_v. this is an ill-posed problem. The good news is that $\varphi_V(t) := E(e^{itV}) = exp(-0.5\sigma_v^2 t^2)$ (as $V \sim N(0,\sigma_v^2)$), which is a monotonic decreasing function of σ_v^2 (and σ_v as well if we restrict $\sigma_v > 0$). So still we can solve a semi-quadratic optimization problem by minimizing $S_{chf}(\sigma_v|\pi,\theta)$ where σ_v is defined on

$$\Sigma = \{\sigma_v : 0 < \sigma_v^2 \leq Var(logY)\}$$

which is a compact support. With some penalty term $\lambda P(\theta)$, we can get a smoother density estimates $\hat{\theta}$.

Hence, we suggest an iterative EM algorithm to achieve the global minimum of $S_{chf}(.,.,.)$ as follows:

Step 1. Initialization: Set $\sigma_v^{(0)}$, $\pi^{(0)}$ and $\theta^{(0)}$;

Step 2. Updating:

$$\begin{aligned}\pi^{(1)} &= \arg\min_{\pi\in\Pi} \int |\hat{\varphi}_n(t) - \tilde{\varphi}_Y(t|\pi, \theta^{(0)}, \sigma_v^{(0)})|^2 w(t)dt \\ &= \arg\min_{\pi\in\Pi} \pi^T \int \big(a_1(t)w(t)a_1^T(t) + a_2(t)w(t)a_2^T(t)\big)dt \times \pi \\ &- 2\int \big(b_1(t)w(t)a_1^T(t) + b_2(t)w(t)a_2^T(t)\big)dt \times \pi\end{aligned}$$

$$\begin{aligned}\theta^{(1)} &= \arg\min_{\theta\in\Theta} \int |\hat{\varphi}_n(t) - \tilde{\varphi}_Y(t|\pi^{(1)}, \theta, \sigma_v^{(0)})|^2 w(t)dt \\ &= \arg\min_{\theta\in\Theta} \theta^T \int \big(a_3(t)w(t)a_3^T(t) + a_4(t)w(t)a_4^T(t)\big)dt \times \theta \\ &- 2\int \big(b_3(t)w(t)a_3^T(t) + b_4(t)w(t)a_4^T(t)\big)dt \times \theta\end{aligned}$$

$$\begin{aligned}\sigma_v^{(1)} &= \arg\min_{\sigma\in\Sigma} \int |\hat{\varphi}_n(t) - \tilde{\varphi}_Y(t|\pi^{(1)}, \theta^{(1)}, \sigma_v)|^2 w(t)dt \\ &= \arg\min_{\sigma\in\Sigma} \int \varphi_V(t)\big(a_5(t)w(t)a_5^T(t) + a_6(t)w(t)a_6^T(t)\big)\varphi_V(t)dt \\ &- 2\int \big(b_1(t)a_5(t) + b_2(t)a_6(t)\big)\varphi_V(t)dt\end{aligned}$$

for some functions a_i and b_i defined below.

Step 3. Set $\pi^{(0)} = \pi^{(1)}$, $\theta^{(0)} = \theta^{(1)}$ and $\sigma_v^{(0)} = \sigma_v^{(1)}$ and repeat *Step* 2 until convergence. Since the random noise V is symmetric about zero (recall $V \sim N(0, \sigma_v^2)$), the characteristic function of V is real-valued, the coefficient term in *Step* 2 can be explicitly given as

$$a_1(t) = \begin{pmatrix} \varphi_V(t, \sigma_v^{(0)}) \\ \varphi_V(t, \sigma_v^{(0)}) \sum_{j=1}^{M}(n)\theta_j^{(0)} cos(tu_j) \end{pmatrix}$$

$$a_2(t) = \begin{pmatrix} 0 \\ \varphi_V(t, \sigma_v^{(0)}) \sum_{j=1}^{M}(n)\theta_j^{(0)} sin(tu_j) \end{pmatrix}$$

$$b_1(t) = \frac{1}{n}\sum_{k=1}^{n} cos(tY_k), \quad b_2(t) = \frac{1}{n}\sum_{k=1}^{n} sin(tY_k)$$

Similarly, for all $j = 1, \ldots, M(n)$

$$[a_3(t)]_j = \varphi_V(t, \sigma_v^{(0)})\pi_2^{(1)} cos(tu_j), \quad [a_4(t)]_j = \varphi_V(t, \sigma_v^{(0)})\pi_2^{(1)} sin(tu_j)$$

$$b_3(t) = \frac{1}{n}\sum_{k=1}^{n} cos(tY_k) - \varphi_V(t, \sigma_v^{(0)})\pi_1^{(1)}, \quad b_4(t) = \frac{1}{n}\sum_{k=1}^{n} sin(tY_k) - \varphi_V(t, \sigma_v^{(0)})\pi_1^{(1)} sin(0) = \frac{1}{n}\sum_{k=1}^{n} sin(tY_k)$$

$$a_5(t) = \pi_1^{(1)} + \pi_2^{(1)}\sum_{j=1}^{M}(n)\theta_j^{(1)} cos(tu_j), \quad a_6(t) = \pi_2^{(1)}\sum_{j=1}^{M}(n)\theta_j^{(1)} sin(tu_j).$$

2.B Proof of Theorem 1

Proof. The proof is straightforward following the Theorem 3.1 in Chen and Pouzo (2012). We just need to check that the conditions of their theorem are satisfied in the present context.

First, for each integer $k < \infty$, $dim(\mathcal{H}_k) < \infty$, $\mathcal{H}_k$ is bounded as

$$\mathcal{H}_n = \{h \in H : h(t) = [\pi_1 + \pi_2 \sum_{j=1}^{k(n)} \theta_j (i \sin(tx_j) + cos(tx_j))]\phi_v(t, \sigma)\}$$

$k(n) \to \infty$ slowly as $n \to \infty$. $E[||m(t,h)||_W^2] \equiv E[m(t,h)^T m(t,h)]$ is continuous on $(\mathcal{H}_k, ||\cdot||_s)$, hence, lower semicontinuous on $(\mathcal{H}_k, ||\cdot||_s)$.

Second, we need to verify the restriction $\max\{\eta_{0,n}, E[||m(t,\Pi_n h_0)||_W^2], \bar{\delta}_{m,n}^2, \lambda_n\} = o(g(k(n), \epsilon))$ for all $\epsilon > 0$ (*). We discuss it for two cases: $\lambda_n = 0$ (no penalty) and $\lambda_n \neq 0$ (penalty case).

For the no penalty case, $\lambda_n = 0$, $\liminf_{k(n)\to\infty} g(k(n), \epsilon) \equiv \inf_{h\in\mathcal{H}:||h-h_0||_s\geq\epsilon} E[||m(t,h)||^2_W]$. Thus, given Assumption 1 (ii), for all $\epsilon > 0$, $\liminf_{k(n)\to\infty} g(k(n), \epsilon) > 0$ if $(\mathcal{H}, ||\cdot||_s)$ is compact, restriction $(*)$ becomes $\max\{\eta_{0,n}, E[||m(t, \Pi_n h_0)||^2_W], \bar{\delta}^2_{m,n}, \lambda_n\} = o(1)$ and it is trivially satisfied.

For the penalty case, rewrite the sieve space as

$$\mathcal{H}_n = \{h \in H : h(t) = [\pi_1 + \pi_2 \sum_{j=1}^{k(n)} \theta_j (i \sin(tx_j) + cos(tx_j))]\phi_v(t, \sigma_v),$$

$$||\Delta^2 h(t)||^2 \leq log^2(n)\}$$

$m(t, h) = E[|\phi_n(t) - \phi_Y(t|\pi, \theta)|] = E[h_0(Y) - h(Y)|t]$. Under very mild regular conditions on the conditional density of Y given t, $E[\cdot|t]$ is a compact operator mapping from $\mathcal{H} \subseteq L^2(f_Y)$ to $L_2(f_t)$ (See, Blundell, Chen and Kristensen, 2007), which has a singular value decomposition $\{\mu_k; \varphi_{1k}, \varphi_{0k}\}^\infty_{k=1}$, where $\{\mu_k\}^\infty_{k=1}$ are the singular numbers arranged in nonincreasing order ($\mu_k \geq \mu_{k+1} \searrow 0$) and $\{\varphi_{1k}\}^\infty_{k=1}$ and $\{\varphi_{0k}\}^\infty_{k=1}$ are eigenfunctions in $L^2(f_Y)$ and $L^2(f_t)$ respectively.

Note that $E[||m(t, \Pi_n h_0)||^2_W$ is continuous ion $(\mathcal{H}, ||\cdot||_s)$ and by the contradiction deduction arguments in NPIV Example (2) in Chen and Pouzo (2012), we can derive the same conclusion:

$$\frac{E[||m(t, \Pi_n h_0)||^2_W}{g(k(n), \epsilon)} \leq const \times ||\Pi_n h_0 - h_0||^2_s = o(1).$$

By letting $\frac{\max\{\eta_{0,n}, \bar{\delta}^2_{m,n}, \lambda_n\}}{g(k(n),\epsilon)} = o(1)$, Theorem 3.1 in Chen and Pouzo (2012) is applicable here. Hence, we have $||\tilde{h}_0 - h_0||_s = o_p(1)$. The result follows. □

2.C Proof of Theorem 2

Proof. Proof for the first part of the theorem is straightforward based on Lemma 3 by substituting $M(n) = k(n)$ and note that

$$||\tilde{\varphi} - \varphi_0||_s \leq ||\hat{\varphi} - \varphi_0||_s + ||\tilde{\varphi} - \hat{\varphi}||_s = O(n^{-1/2}) + O_p\big(M(n)^{-\kappa} + \sqrt{\frac{M(n)}{n \times \varphi(k(n)^{-2})}}\big),$$

where $\tilde{\varphi}$ is the estimated characteristic function based on the proposed penalized LS sieve and $\hat{\varphi}$ is the empirical characteristic function from the sample. For the second step, the convergence rate for the second term is derived in Lemma 3. Following Chen and Pouzo (2012), when $\tilde{\varphi}$ and $\hat{\varphi}$ are measurable function of (discretized) Y, we have $\varphi(k(n)^{-2}) = const$ which results in the $\ln(n)$ convergence rate of $\tilde{\varphi}$. Explicitly, we have

$$||\tilde{\varphi}(t) - \varphi(t)||_s = O_p\big(n^{-1/2} + M(n)^{-\kappa} + \sqrt{\frac{M(n)}{n \times \varphi(M(n)^{-2})}}\big) = O_p([\ln n]^{-\kappa/\varsigma})$$

where $\varphi(\tau) = \exp(-\tau^{-\varsigma/2})$ for some $\varsigma > 0$ and $M(n) = c[\ln n]^{1/\varsigma}$ for some $c \in (0, 1)$.

To prove the second part of the theorem, we apply Lemma 1 with the severe ill-posed case, i.e., $\varphi(\tau) = \exp(-\tau^{-\varsigma/2})$ for some $\varsigma > 0$. Hence, we have

$$\begin{aligned} E||\tilde{f}_U - f_U||_2 &= O_p(||\tilde{f}_c(u|\hat{\theta}) - f_c||) \\ &= O_p(M(n)^{-1} + ||\hat{\theta}_j - \theta_j||) \\ &= O_p\big(\max([\ln n]^{-1/\varsigma}, \quad [\ln n]^{-\kappa/\varsigma}\big) \end{aligned}$$

That is, $H^2(\tilde{f}_U, f_U) = O_p\big(\max([\ln n]^{-1/\varsigma}, \quad [\ln n]^{-\kappa/\varsigma})\big)$.

The convergence rate of the estimated frontiers α can be derived similarly but with two steps. First, consider estimation of the constant frontier. Under the unimodality assumption, we have $\hat{\alpha} = u_j^*$ where $j^* = index(\max(\hat{\theta}_j))$, hence

$$\begin{aligned}
||\hat{\alpha} - \alpha|| &= ||u_j^* - u_j|| \\
&= O_p(\max(M(n)^{-1}, ||\hat{h}_n - h_0||_s)) \\
&= O_p\big(\max([\ln n]^{-1/\varsigma}, \quad [\ln n]^{-\kappa/\varsigma}\big) \\
&= O_p((\ln n)^{-const})
\end{aligned}$$

Then for the heteroskedastic frontier estimator $\alpha(W)$, a similar result can be derived: the convergence rate is of the same order as $\hat{\theta}_j$ (as well as $\hat{h}_n$) since $\ln n$ and $\ln(nb^d)$ are asymptotically equivalent under Assumptions 1-5 where b is the bandwidth in Section 2.2. The conclusion follows. □

2.D Proof of Theorem 3

To prove Theorem 3, we list the main theorem (e.g., Theorem 4.1) utilized here from Chen and Puozo (2015) first and then check that each assumption holds in our setting.

Theorem 4.1 (Chen and Puozo, 2015) Let $\hat{\alpha}_n$ be the PSMD estimator and Assumption 3.1-3.4 hold. If Assumption 3.5-3.6 hold. Then

$$\sqrt{n}\frac{\phi(\hat{h}_n) - \phi(h_0)}{||v_n^*||_{sd}} \to_d \mathcal{N}(0,1).$$

Then we check the assumptions. Assumption 3.1-3.3 in Chen and Pouzo (2015) are the identification assumptions, penalty assumptions and sample criteria, which correspond to the Assumption 1-3 in the present paper respectively. Assumption 3.4 concerns the local curvature of the population criterion $Q(h_0)$ at h_0. When $\hat{Q}_n(h)$ is computed using the series LS estimator, CP (2012) show it is automatically satisfied. Assumption 3.5 restricts the local behavior of $\phi(\cdot)$. Specifically, Assumption 3.5(ii) controls the nonlinear bias of ϕ which will be automatically satisfied for a quadratic functional. Assumption 3.5(i) places a restriction on how fast the sieve dimension k(n) can grow with the sample size n. Assumption 3.5(iii) controls the linear bias part due to the finite dimensional sieve approximation of $h_{0,n}$ to h_0, which corresponds to our Assumption 4 on sieve approximation error. Assumption 3.6 is a local quadratic assumption which corresponds to our Assumption 5 sieve link condition.

Note that $\hat{m}$ is the LS series estimator in (*): $\hat{m}(t,h) = p^{k(n)}(t)^T \cdot (P^T P)^- \sum p^{k(n)}(t_i)\rho(Y_i,h)$ for model $E(\rho(Y,\hat{h})|t_i) = 0$ a.s. any t. Moreover $\phi(h) = \int \hat{m}(t_i,h)'\hat{m}(t_i,h)dt_i$ is a regular functional of h. For any regular functional $\phi(\cdot)$[22], the above theorem implies that $\sqrt{n}(\phi(\hat{h}_n) - \phi(h_0)) \to N(0, \sigma^2_{v^*})$ with

$$\sigma^2_{v^*} = \lim_{n\to\infty} ||v^*_n||^2_{sd} = ||v^*||^2_{sd} = E[\frac{d\phi(h)}{dh}[p^{k(n)(.)}]' D^-_n \Phi_n D^-_n \frac{d\phi(h)}{dh}[p^{k(n)(.)}]]$$

and $\frac{d\phi(h)}{dh}[p^{k(n)(.)}] = 2\int \hat{m}(t_i,h)p^{k(n)}(t_i)dt_i$ (Chen and Pouzo, 2015). Consequently, we have $||v^*_n||_{sd} = 2\int \hat{m}(t_i,h_0)p^{k(n)}(t_i)dt_i D^-_n \Phi_n D^-_n 2\int \hat{m}(t_i,h_0)p^{k(n)}(t_i)dt_i$, $D^-_n = E\big(E[p^{k(n)}(Y)|t_i]E[p^{k(n)}(Y)|t_i]'\big)$, $\Phi_n = E\big(E[p^{k(n)}(Y)|t_i]\rho^2(t_i,h_0)E[p^{k(n)}(Y)|t_i]'\big)$ and $\rho(t_i,h_0) = |\phi_n(t_i) - \phi_Y(t_i|\pi,\theta)|$. Therefore the conclusion follows.[23]

[22] We call $\phi(\cdot)$ regular (or irregular) at h_0 whenever $\lim_{k(n)\to\infty} ||v^*_n|| \leq \infty$ (or ∞).

[23] One could also apply Remark 6.1 in Chen and Puozo (2015) to prove Theorem 3.

Bibliography

[1] Aigner, D.J., Lovell, C.A.K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier models. Journal of Econometrics. 6, 21–37.

[2] Akaike, H., 1974. A new look at the statistical model identification. IEEE transactions on automatic control 19, 716–723.

[3] Andrews, D.W.K., 2001. Testing when a parameter is on the boundary of the maintained hypothesis. Econometrica 69, 683?734.

[4] Blundell, R., Chen, X., Kristensen, D., 2007. Semi-nonparametric IV estimation of shape invariant Engel curves. Econometrica. 75: 1613–1669.

[5] Chen, X., Reiss, M., 2011. On rate optimality for ill-posed inverse problems in econometrics. Econometric Theory. 27(3), 497–521.

[6] Chen, X., 2007. Large Sample Sieve Estimation of Semi-Nonparametric Models, in The Handbook of Econometrics, Vol. 6B, ed. by J. J. Heckman and E. E. Leamer. Amsterdam: North-Holland. [1057]

[7] Chen, X., Pouzo, D., 2012. Estimation of Nonparametric Conditional Moment Models With Possibly Nonsmooth Generalized Residuals. Econometrica. 80: 277–321.

[8] Chen, X., Pouzo, D., 2015. Sieve Wald and QLR Inferences on Semi/Nonparametric Conditional Moment Models. Econometrica. 83: 1013–1079.

[9] Chung, K.L., 2001. A course in probability theory (3rd ed.). Academic Press.

[10] Fan, J. 1991. On the optimal rates of convergence for nonparametric deconvolution problems. Annals of Statistics. 19(3), 1257–1272.

[11] Fan, Y., Li, Q., Weersink, A., 1996. Semiparametric estimation of stochastic production frontier models. Journal of Business and Economics Statistics 14, 460–468.

[12] Feng, G.,H., Serletis, A., 2009. Efficiency and productivity of the US Banking Industry, 1998-2005: Evidence from the Fourier cost function satisfying global regularity conditions. Journal of Applied Econometrics. 24, 105–138.

[13] Florens, J.L., Simar, L., Van Keilegom, I., 2019. Estimation of the Boundary of a Variable Observed With Symmetric Error. Journal of the American Statistical Association. DOI: 10.1080/01621459.2018.1555093

[14] Greene, W.H., 2008. The econometric approach to efficiency analysis. In: Fried, Harold, Knox Lovell, C.A., Schmidt, Shelton (Eds.), The Measurement of Productive Efficiency, second Edition. Oxford University Press.

[15] Hall, P., Simar, L., 2002. Estimating a change point, boundary or frontier in the presence of observation error. Journal of American Statistic Association. 97, 523–534

[16] Horrace, W.C., Parmeter, C.F., 2011. Semiparametric Deconvolution with Unknown Error Variance. Journal of Productivity Analysis. 35, 129–141.

[17] Horrace, W.C., Parmeter, C.F., 2018. A Laplace Stochastic Frontier Model. Econometric Reviews. 37, 260–280.

[18] Kneip, A., Simar, L., Van Keilegom, I., 2015. Frontier estimation in the presence of measurement error with unknown variance. Journal of Econometrics. 184: 379–393.

[19] Kumbhakar, S.C., Park, B.U., Simar, L., Tsionas, E.G., 2007. Nonparametric stochastic frontiers: a local likelihood approach. Journal of Econometrics 137, 1–27.

[20] Kumbhakar, S.C., Parmeter, C.F., Tsionas, E.G., 2013. A zero inefficiency stochastic frontier model. Journal of Econometrics. 172: 66–76.

[21] Kumbhakar, S.C., Parmeter, C.F., Zelenyuk, V., 2017. Stochastic Frontier Analysis: Foundations and Advances. Working Papers 2017-10, University of Miami, Department of Economics.

[22] Lee, M., Hall, P., Shen, H.P., Marron, J.S., Tolle, J., Burch, C., 2013. Deconvolution estimation of mixture distributions with boundaries. Electronic Journal of Statistics. 7: 323–341.

[23] Lee, M., Wang, L., Hall, P., Shen, H.P., Marron, J.S., 2015. Least Squares Sieve Deconvolution of Mixture Distributions with Boundary Effects. Journal of the Korean Statistical Society. 44: 187–201.

[24] Madrid-Padilla, O.H., Polson, N.G., Scott J., 2018. A deconvolution path for mixtures. Electronic Journal of Statistics. 12: 1717–1751.

[25] Martins-Filho, C., Yao, F., 2015. Semiparametric Stochastic Frontier Estimation via Profile Likelihood, Econometric Reviews. 34: 413–451

[26] Meeusen, W., van den Broek, J., 1977. Efficiency estimation from Cobb-Douglas production function with composed error. International Economic Review. 8, 435–444.

[27] Parmeter, C.F., Wang, H.J., Kumbhakar, S.C., 2017. Nonparametric estimation of the determinants of inefficiency, Journal of Productivity Analysis 47, 205–221.

[28] Rho, S., Schmidt, P., 2015. Are all firms efficient? Journal of Productivity Analysis. 43, 327–349.

[29] Schwarz, M., Van Bellegem, S., 2010. Consistent density deconvolution under partially known error distribution. Statistics and Probability Letters 80, 236–241.

[30] Simar, L., Van Keilegom, I., Zelenyuk, V. 2017. Nonparametric least squares methods for stochastic frontier models. Journal of Productivity Analysis 47, 189–204.

[31] Tikhonov, A. 1963. On the solution of incorrectly formulated problems and the regularization method. Soviet Math. Doklady. 4(4), 1035–1038.

[32] Tran, K.C., Tsionas, E.G., 2016. Zero-inefficiency stochastic frontier models with varying mixing proportion: A semiparametric approach. European Journal of Operational Research. 249: 1113–1123.

Chapter 3

Panel Nonparametric Identification and Estimation of Conditional Heteroskedastic Frontiers with an Application to CO2 Emission Productivity Analysis

JUN CAI[1]

3.1 Introduction

Since Aigner et al. (1977), stochastic frontier analysis (SFA) has been an important tool in the analysis of productive efficiency. The original model includes a log-linear production or cost function with an additively separable noise term and an additively separable, time-invariant, inefficiency term, which decreases output in the production function or increases costs in the cost function. It is a leading case of the "composed error model" with two independent components, and it has been the workhorse for countless empirical investigations of firm-level cost or productive inefficiency. While estimation of the linear production or cost

[1] Department of Economics, Syracuse University, jcai106@syr.edu. I am extremely grateful for the invaluable support and guidance of my advisor, William C. Horrace, and committee members Yoonseok Lee, Alfonso Flores-Lagunes and Subal Kumbhakar from Binghamton University. Special thanks to Hugo Jales, Yulong Wang, and participants at 16th EWEPA conference, SU Econometrics lunch, SUNY-Albany Seminar, and Binghamton University Seminar, for insightful comments and suggestions.

function is important in these models, the primary concern is often characterization of the error components. In particular, interest centers on characterizing the inefficiency term in a meaningful way. That is, most regression-based empirical studies treat errors or error components as nuisance parameters and focus on the marginal effects in the conditional mean function. However, in the stochastic frontier literature estimation of the error components or features of the error component distributions is an important aspect of the model's specification: one can not specify a production or cost function with inefficiency without trying to understand the statistical and economic significance of that inefficiency.

While early models were designed for cross sectional data and were fully parametric, proliferation of panel data has facilitated relaxation of parametric assumptions. Pitt and Lee (1981), Schmidt and Sickles (1984), and Battese and Coelli (1988) consider fixed-effect and random-effect estimation of the stochastic frontier model with additively separable time-invariant technical or cost inefficiency. Here, the fixed- or random-effect embodies inefficiency. Cornwell et al. (1990), Kumbhakar (1990), Lee and Schmidt (1993), Han et al. (2005) and Ahn et al. (2007, 2013) propose time-varying inefficiency versions of the model. These models are more sensible because they allow firms to decrease their inefficiency over time. That is, any reasonable model specification should allow firms to move closer to the efficient frontier over time. More recently, Greene (2005a, 2005b) introduces a "true" fixed-effect/random-effect stochastic frontier which includes the additively separable noise and inefficiency terms but also features an additively separable, time-invariant fixed or random effect. That is, each firm in the sample has an idiosyncratic and persistent effect in addition to a time-varying effect and a draw from the noise distribution. Wang and Ho (2010), Chen, Schmidt and Wang (2014), Wikstrom (2015) and more recently Belotti and Ilardi (2018) study estimation techniques for these models. This paper presents a nonparametric version of the "true" fixed-effect/random-effect model, where (a) the production or cost function is a general function of the inputs (and some other factors) allowing non-separable fixed or random effects, (b) inefficiency is time-varying (or heteroskedastic), and (c) relax the distributional assumption on random noise term. To the best of our knowledge we are the first to consider panel nonparametric estimation of the stochastic frontier model with nonseparable fixed/random effects.

Studies on nonparametric or semi-parametric stochastic frontier models for cross-sectional data are Fan et al. (1996), Kumbhakar et al. (2007), Parmeter and Racine (2012), Noh (2014), Martins-Filho and Yao (2015), Parmeter et al. (2017). They tried to relax one or a few specification restrictions in the cross-sectional stochastic frontier model. For panel data, there are only a couple nonparametric stochastic frontier models. Kneip and Simar (1996) employ a Nadaraya-Watson estimator to estimate a nonparametric version of the Schmidt and Sickles (1984) model with time-invariant inefficiency. Yao et al. (2018) investigate a semi-parametric smooth coefficient stochastic frontier model for panel data. It should be noted that, however, these studies do not consider (non-separable) fixed/random effects.

In this paper, we identify and consistently estimate the variance parameters associated with the noise and inefficiency components in the panel stochastic frontier model with non-separable fixed/random effects and added time effects, using the results of Kotlarski (1967) and Evdokimov and White (2012).[2] These variance components are often the primary focus of the stochastic frontier literature and of this paper. In particular, we consider

$$Y_{it} = \lambda_t + m(X_{it}, \alpha_i) + \varepsilon_{it} \tag{3.1}$$

$$\varepsilon_{it} = U_{it} + V_{it}, \quad i = 1, ..., n, \quad t = 1, ..., T, \tag{3.2}$$

where $m(X_{it}, \alpha_i)$ is the unspecified cost function or production function which allows non-separable unobserved heterogeneity α_i; λ_t is an added time effects which is a constant for each period; $X_{it} \in \mathcal{R}^p$, α_i is the random effects (RE) or fixed effects (FE). U_{it} is the inefficiency term which could be conditionally heteroskedastic in some exogenous X_{it}. It is

[2] The results of Kotlarski (1967) have been applied in a variety of economic settings. See, for example, Li and Vuong (1998), Schennach (2004), Li, Perrigne, Vuong (2000), Krasnokutskaya (2011), Arellano and Bonhomme (2009), Bonhomme and Robin (2010) and Kennan and Walker (2011).

constrained positive in a cost function and negative in a production function. The V_{it} is the random noise or disturbance term from unspecified distributional family. We derive moment conditions, based on which the model can be identified and consistently estimated with two time periods.

The nearest neighbors to our model and contribution is the aforementioned approach of Evdokimov (2010) and Evdokimov and White (2012), which are concerned with estimation of the model in equation (3.1) with the restriction $\varepsilon_{it} = V_{it}$. Like Evdokimov (2010) our model is quite general and only requires that the noise component (V) be from a zero-mean, conditionally symmetric distribution with finite second moment. However, unlike Evdokimov (2010), our contribution is identification and estimation of the variances of the error components (U and V) in equation (3.2) allowing added time effects, instead of the distribution of α_i or $m(x, \alpha)$ itself, though the latter can be obtained with a straightforward application of Evdokimov (2010) or more recently Ju, Gan and Li (2017). Since the model is identified when inefficiency equals zero (i.e., $U = 0$), we show that the model with non-zero inefficiency is still identified when its distribution is known up to its variance (i.e., half-normal or exponential inefficiency). Showing this requires a "common support" assumption on X_{i1} and X_{i2}, which will often be met in empirical applications where inputs are relatively stable across (short) periods and which allows us to remove the m function with time-differencing of equation (3.1). It also requires independence between V_{it} and $(\alpha_i, V_{is}, X_{is})$ for any $t \neq s$ conditional on X_{it} which is standard in a non-linear panel data setting.

The paper is organized as follows. In Section 3.2 we introduce the assumptions needed to identify the model in equation (3.1) and derive moment conditions for estimation of the variance components. Section 3.3 discusses estimation issues and bandwidth selection methods. Section 3.4 derives the large sample properties of the estimator. We study the finite sample properties of the estimator with Monte Carlo simulations in Section 3.5. In Section 3.6 we apply the model to investigate CO2 emission productivity with a panel of 136 countries over 25 years and discuss the policy implications. Directions for future research and conclusions are in Section 3.7. Proofs of theorems are in the Appendix.

3.2 Identification

For stochastic frontier analysis, the inefficiency term U_{it} in equation (3.2) is of primary interest. We focus on the identification of the distribution of U_{it}, especially its conditional variance. In what follows, we consider $X_{it} \in R^1$ (i.e., $p = 1$) for simplicity of presentation and assume that m is a cost function so that $U_{it} \geq 0$. We let $T = 2$ and assume the following conditions:

Assumption 6 (Identification)**.** *(i)* $\{X_{it}, U_{it}, V_{it}, \alpha_i\}$ *are i.i.d. across* i *and stationary over* t*.* λ_t *is an added time effects and we normalize* $\lambda_1 = 0$*.*
(ii) The inefficiency satisfies $U_{it}|X_{it} = x$ *is distributed as either half-normal,* $|\mathcal{N}(0, \sigma_u^2(x))|$*; or Exponential,* $Exp(\sigma_u(x))$*, where* $0 < \sigma_u^2(.) < \infty$ *is time-invariant.*
(iii) Given X_{it}*, the random noise* V_{it} *is independent of* U_{it} *and its distribution is symmetric with* $E(V_{it}|X_{it} = x) = 0$ *and* $Var(V_{it}|X_{it} = x) = \sigma_v^2(x) < \infty$ *for all* x *and* $t = 1, 2$*.*
(iv)(Conditional Independence) $f_{V_{it}|X_{it},\alpha_i,X_{i\tau},V_{i\tau}}(v|x, \alpha, \tilde{x}, \tilde{v}) = f_{V_{it}|X_{it}}(v|x)$ *for all* $(v, x, \alpha, \tilde{x}, \tilde{v})$ *and* $t \neq \tau$*, where* $f_{V_{it}|\cdot}$ *is the conditional density function of* V_{it}*.*
(v)(Common Support) *The joint density of* $X_i = (X_{i1}, X_{i2})$ *satisfies* $f_{X_{i1},X_{i2}}(x, x) > 0$ *for all* $x \in \chi$*, where* χ *is the common support of* X_{i1} *and* X_{i2}*.*
(vi) The conditional characteristic function $\phi_{V_{it}|X_{it}}(s|x)$ *does not vanish for all* s, x *and* $t = 1, 2$*.*

Assumptions 6-(i) and (ii) are standard assumptions for a two-way panel data or for the panel stochastic frontier model. We allow for conditional heteroskedastic variances, $\sigma_v^2(x)$ and $\sigma_u^2(x)$, which may be a function of environmental variables, say Z_{it} which can be a subset of X_{it}. In that case, the model allows for the "scaling property" specification $U_{it} \sim G(Z_{it})|\mathcal{N}(0, \sigma_u^2)|$ of Wang and Ho (2010).[3] In addition to the half-normal and exponential

[3] The scaling property specification possesses some appealing features. See Wang and Ho (2010), Wang and Schmidt (2002), Alvarez et al. (2006) and some others.

distributions, Assumption 6-(ii) can be generalized to include single-parameter distributional families with bounded second moments. In Assumption 6-(iii), both U_{it} and V_{it} are related to X_{it} through the variance terms, but they are independent once X_{it} is given.

Assumption 6-(iv) is crucial for identification. It implies that conditional on the contemporaneous covariate X_{it}, the disturbance term V_{it} is independent of $\alpha_i, V_{i\tau}$, and $X_{i\tau}$ for any $t \neq \tau$. For example, we let $V_{it} = \sigma_v(X_{it})\eta_{it}$, where $\sigma_v(x)$ is a bounded positive function and η_{it} are i.i.d $\mathcal{N}(0,1)$ that are independent of $(\alpha_i, X_{i\tau})$. It rules out lagged dependent variables in X_{it} and serially correlated disturbances. The assumption is strong but necessary for the derivation of identification moments that follow and also the deconvolution techniques for recovering m function. For the case with serially correlated V_{it}, we need at least three periods to identify the model as Evdokimov (2010). To avoid an overload of current paper, we omit it here.

Assumption 6-(v) is also important for identification, but it generally holds in the stochastic frontier models since X_{it} includes input elements, prices, and some environment variables that are continuous. Assumption 6-(vi) is satisfied for most of the distributions, including normal, log-normal, Cauchy, Laplace, χ^2 and Student-t distributions.

If we only care about characterizing the inefficiency term U_{it} (which may be reasonable in frontier analysis), then only Assumption 6 is needed for identification. Under Assumption 6, we first introduce the following theorem, which is the key identification result of this paper. Proof of Theorem 4 is in the Appendix, which is based on proof of Lemma 1 in Evdokimov and White (2012).

Theorem 4. *Suppose Assumption 6 holds. Then the time effects, the distribution of inefficiency U_{it} and the elasticity of the mean inefficiency $\mu_E := E[U_{it}]$ with respect to covariates X_{it} are identified. That is, λ_t, $\sigma_u^2(x)$ and $\xi_{\mu X} := \frac{\partial \mu_E}{\partial x}\frac{x}{\mu_E}$ are identified for all $x \in \chi$ and $t = 1, 2$.*

Identification of the distribution of U_{it} is achieved by recovering $\sigma_u^2(x)$ under Assumption 6-(ii). In particular, it is identified based on the following two moment conditions, which is obtained when U_{it} is half-normal for instance:

$$E_x[Y_{it}(Y_{it} - Y_{i\tau})] - E_x[Y_{it}]E_x[(Y_{it} - Y_{i\tau})] = \left(1 - \frac{2}{\pi}\right)\sigma_u^2(x) + \sigma_v^2(x), \tag{3.3}$$

and

$$\begin{aligned} E_x[Y_{it}(Y_{it} - Y_{i\tau})^2] - E_x[Y_{it}]E_x[(Y_{it} - Y_{i\tau})^2] - 2E_x[(Y_{it} - Y_{i\tau})](E_x[Y_{it}(Y_{it} - Y_{i\tau})] - \\ E_x[Y_{it}]E_x[(Y_{it} - Y_{i\tau})]) = \frac{(4-\pi)\sqrt{2}}{\pi\sqrt{\pi}}\sigma_u^3(x), \end{aligned} \tag{3.4}$$

where E_x denotes expectation conditional on $X_{it} = X_{i\tau} = x$. These moment conditions are calculated similarly to the moment conditions in Simar et al. (2017).[4] When U_{it} has exponential distribution, we can obtain the same moment conditions by just replacing the scaling terms $(1 - 2/\pi)$ and $(4 - \pi)\sqrt{2}/(\pi\sqrt{\pi})$ by 1 and 2 in equations (3.3) and (3.4), respectively.

And as a byproduct, the added time effects λ_t can be identified as well

$$E[Y_{it} - Y_{i1}|X_{it} = X_{i1} = x] = E[\varepsilon_{it} + \lambda_t - \varepsilon_{i1}|X_{it} = X_{i1} = x] = \lambda_t$$

with the normalization $\lambda_1 = 0$. Moreover, the elasticity of the mean efficiency with respect to the covariate X_{it} can be identified based on the derived moment conditions. Define the elasticity as $\xi_{\mu X} = \frac{\partial\mu_E}{\partial x}\frac{x}{\mu_E}$, note that $\mu_E(x) := E[U_{it}(x)] = \frac{\sqrt{2}\sigma_u(x)}{\sqrt{\pi}}$ for the half-normal distribution ($\sigma_u(x)$ for exponential distribution), so we can easily derive:

$$\xi_{\mu X} = \frac{\partial\sigma_u(x)}{\partial x}\frac{x}{\sigma_u(x)} = \frac{1}{3}\frac{\partial(\sigma_u^3(x))}{\partial x}\frac{x}{(\sigma_u^3(x))}$$

in which $\sigma_u^3(x)$ is identified in equation (3.4).

In addition to the inefficiency term U_{it}, if the cost or production function m is also of interest, we need additional conditions as follows, similarly as Evdokimov (2010).

[4] If one would like to assume that higher central moments exist for U and V, more odd moment conditions can be used to identify the unknown variance parameters σ_u^2 in a similar way.

Assumption 7 (Identification of m). *(i) $m(x, \alpha)$ is weakly increasing in α for all $x \in \chi$.*
(ii) $\alpha_i|X_i$ is continuously distributed for all $X_i \in \chi \times \chi$.
(iii) $m(x, \alpha)$ and conditional densities $f_{V_{it}|X_{it}}(v|x)$, $f_{\alpha_i|X_{it}}(\alpha|x)$, and $f_{\alpha_i|X_{i1},X_{i2}}(\alpha|x_1, x_2)$ are almost everywhere continuous in the continuously distributed components of x, x_1, x_2 for all α and v.

Assumption 7-(i) is not too restrictive in the stochastic frontier models. It can be weakened to "monotonic in α". Assumption 7-(ii) and Assumption 7-(iii) are standard. For further discussions, see Evdokimov (2010).

Finally, in order to identify the idiosyncratic term α_i, we need to decide if it will be treated as a random-effect (RE) or a fixed-effect (FE) and use either set of the following assumptions.

Assumption 8 (RE). *(i) α_i and X_i are independent. (ii) α_i is Uniform on* $[0, 1]$.

Assumption 8-(i) defines the random effects model, while Assumption 8-(ii) is a standard normalization. This normalization is necessary because the function $m(x, \alpha)$ is modeled nonparametrically.

Assumption 9 (FE). *(i) $m(x, \alpha)$ is strictly increasing in α for all x.*
(ii) For some $\bar{x} \in \chi$, $m(\bar{x}, \alpha) = \alpha$ for all α.
(iii) Supp$\{\alpha_i|(X_{it}, X_{i\tau}) = (x, \bar{x})\}$ = Supp$\{\alpha_i|X_{it} = x\}$ for all $x \in \chi$ and $t \neq \tau$, where Supp$\{\alpha_i|\vartheta\}$ denotes the support of α_i conditional on an event ϑ.

Assumption 9-(i) is standard and guarantees invertibility of function $m(x, \alpha)$ in α. Assumption 9-(ii) is a normalization given Assumptions 7-(ii) and 9-(i). Assumption 9-(iii) requires that the "extra" conditioning on $X_{i\tau} = \bar{x}$ does not reduce the support of α_i. A conceptually similar support assumption is made by Altonji and Matzkin (2005).

Direct application of the results in Evdokimov (2010) leads to the following two theorems. We let F_{α_i} denote the distribution function of α_i.

Theorem 5. *Suppose Assumptions 6, 7, and 8 hold. Then $m(x,\alpha)$, $\sigma_u^2(x)$ and $\sigma_v^2(x)$ are identified for all $x \in \chi$, $\alpha \in (0,1)$and $t = 1,2$.*

Theorem 6. *Suppose Assumptions 6, 7, and 9-(i), (ii) hold. Then $m(x,\alpha)$, $\sigma_u^2(x)$, $\sigma_v^2(x)$, $F_{\alpha_i}(\alpha|X_{it} = x)$ and $F_{\alpha_i}(\alpha)$ are identified for all $x \in \chi$, $\alpha \in Supp\{\alpha_i|(X_{it}, X_{i(-t)}) = (x, \bar{x})\}$ and $t = 1,2$. If in addition Assumption 9-(iii) is satisfied, $\sigma_u^2(x)$, $\sigma_v^2(x)$, $F_{\alpha_i}(\alpha|X_{it} = x)$ and $F_{\alpha_i}(\alpha)$ are identified for all $x \in \chi$, $\alpha \in Supp\{\alpha_i|X_{it} = x\}$ and $t = 1,2$.*

The proofs of Theorems 5 and 6 directly follow from Evdokimov (2010), which we sketch in the Appendix. Specifically, we consider a cost stochastic frontier model with fixed effects α_i, which is a common case in the literature. The proof can be easily extend to the random effects setting.

Remark 1. *For Theorem 4-6, the identification results still hold even with an added functional time effects, namely, $\lambda_t = \lambda_t(X_{it})$. Indeed, normalize $\lambda_1(x) = 0$ for all $x \in \chi$, then for any $t > 1$ time effects $\lambda_t(x)$ are identified as follows:*

$$E[Y_{it} - Y_{i1}|X_{it} = X_{i1} = x] = E[\varepsilon_{it} + \lambda_t(x) - \varepsilon_{i1}|X_{it} = X_{i1} = x] = \lambda_t(x)$$

Once the time effects are identified, identification of the rest of the model proceeds as described above with the random variable Y_{it} replaced by $Y_{it} - \lambda_t(X_{it})$ and setting the constant time effects as zeros.

3.3 Estimation

Since we focus on nonparametric identification and estimation of the variance components in equation (3.2), we now derive their estimation strategy under $T = 2$. As a by-product, the elasticity of mean inefficiency with respect to the covariates X_{it}, $\xi_{\mu X}$, is also estimated. Specifically, consistent estimators for the heterogeneous variance of the inefficiency and random noise (i.e., σ_u^2 and σ_v^2) are obtained by taking advantage of the conditional covariance structure of the panel model. This is consistent with the literature on nonparametric panel model estimation as well as the panel stochastic frontier model. For instance, Wang (2003) proposes a novel method for estimating nonparametric panel data models that utilize the information contained in the covariance structure of the model?s disturbances, as do Henderson et al. (2008) and some others (see Wang et al., 2005).

Note that the variance of the inefficiency term and the random noise are important in the stochastic frontier model, both theoretically and empirically. Theoretically, technical or cost inefficiency, which is usually defined as $E(U_{it}|\varepsilon_{it})$, is a function of σ_u^2 and σ_v^2, and they are closely related to one another. Here, we allow $E(U_{it}|\varepsilon_{it})$ to be not only a function of σ_u^2 and σ_v^2 but also a function of X_{it} though the variance components. Wang and Schmidt (2009) point out that the variance of random noise matters and the technical efficiency estimate $E(U_{it}|\varepsilon_{it})$ is a shrinkage of U_{it} toward its mean in a probabilistic sense. Empirically, the expectation of time-varying inefficiency $E(U_{it})$ equals $\sqrt{2/\pi}\sigma_u$ under the half-normal assumption and σ_u under the exponential distributional assumption. Both are monotonic functions of the inefficiency variance. Alternatively, instead of using the formula for $E(U_{it}|\varepsilon_{it})$, we can use the best linear predictor of U_{it} given ε_{it} (i.e., $a + b\varepsilon_{it}$ for some finite constant a and b), which was analyzed in detail in Waldman (1984). In particular, a simple calculation leads to $b = Var(U_{it})/(Vau(U_{it}) + Var(V_{it}))$ and $a = E(U_{it})(1 + b)$.

To estimate $\sigma_u^2(x)$ and $\sigma_v^2(x)$, we first define $A(x)$ as the conditional covariance between Y_{it} and its first difference $(Y_{it} - Y_{i\tau})$ and $B(x)$ as the conditional covariance between Y_{it} and $(Y_{it} - Y_{i\tau})^2$. $E_x(Y_{i1} - Y_{i2}) = -\lambda_2$ as we already normalize $\lambda_1 = 0$. It follows that, for the half-normal U_{it} case, the moment conditions in (3.3) and (3.4) are written as

$$\begin{aligned} A(x) &= \left(1 - \frac{2}{\pi}\right)\sigma_u^2(x) + \sigma_v^2(x) \\ B(x) + 2 * \lambda_2 A(x) &= \frac{(4-\pi)\sqrt{2}}{\pi\sqrt{\pi}}\left\{\sigma_u^2(x)\right\}^{3/2}, \end{aligned}$$

from which $\sigma_u^2(x)$ and $\sigma_v^2(x)$ are identified as

$$\sigma_u^2(x) = \left\{\frac{\pi\sqrt{\pi}}{(4-\pi)\sqrt{2}}B(x)\right\}^{2/3}$$
$$\sigma_v^2(x) = A(x) - \left(1-\frac{2}{\pi}\right)\left\{\frac{\pi\sqrt{\pi}}{(4-\pi)\sqrt{2}}\big(B(x)+2\lambda_2 A(x)\big)\right\}^{2/3}.$$

For the exponential case, we similarly have

$$\sigma_u^2(x) = \left\{\frac{1}{2}B(x)\right\}^{2/3}$$
$$\sigma_v^2(x) = A(x) - \left\{\frac{1}{2}\big(B(x)+2\lambda_2 A(x)\big)\right\}^{2/3}.$$

Then, $\sigma_u^2(x)$ and $\sigma_v^2(x)$ can be estimated using the following nonparametric kernel (conditional mean) estimators:

$$\begin{aligned}
\hat{A}(x) &= \frac{1}{2}\left\{\sum_{i=1}^n Y_{i1}(Y_{i1}-Y_{i2})\omega_{i,A1}(x) - \sum_{i=1}^n Y_{i1}\omega_{i,A1}(x)\sum_{i=1}^n (Y_{i1}-Y_{i2})\omega_{i,A1}(x)\right\} \qquad (3.5)\\
&\quad +\frac{1}{2}\left\{\sum_{i=1}^n Y_{i2}(Y_{i2}-Y_{i1})\omega_{i,A2}(x) - \sum_{i=1}^n Y_{i2}\omega_{i,A2}(x)\sum_{i=1}^n (Y_{i2}-Y_{i1})\omega_{i,A2}(x)\right\},\\
\hat{B}(x) &= \frac{1}{2}\left\{\sum_{i=1}^n Y_{i1}(Y_{i1}-Y_{i2})^2\omega_{i,B1}(x) - \sum_{i=1}^n Y_{i1}\omega_{i,B1}(x)\sum_{i=1}^n (Y_{i1}-Y_{i2})^2\omega_{i,B1}(x)\right\} \qquad (3.6)\\
&\quad -\frac{1}{2}\left\{2\sum_{i=1}^n (Y_{i1}-Y_{i2})\omega_{i,A1}(x)\hat{A}(x)\right\}\\
&\quad +\frac{1}{2}\left\{\sum_{i=1}^n Y_{i2}(Y_{i2}-Y_{i1})^2\omega_{i,B2}(x) - \sum_{i=1}^n Y_{i2}\omega_{i,B2}(x)\sum_{i=1}^n (Y_{i2}-Y_{i1})^2\omega_{i,B2}(x)\right\}\\
&\quad -\frac{1}{2}\left\{2\sum_{i=1}^n (Y_{i2}-Y_{i1})\omega_{i,A2}(x)\hat{A}(x)\right\},
\end{aligned}$$

where

$$\omega_{i,j}(x) = \frac{K\left((X_{i1}-x)/h_j\right)K\left((X_{i2}-x)/h_j\right)}{\sum_{i=1}^n K\left((X_{i1}-x)/h_j\right)K\left((X_{i2}-x)/h_j\right)}$$

for $j = A1, A2, B1, B2$. The $A1$ and $A2$ denote the bandwidth index for $\omega_{i,A1}$ and $\omega_{i,A2}$ in

$A(x)$ respectively and $B1$ and $B2$ stand similar for $B(x)$. Note that K is a non-negative kernel function and h_j is an appropriate bandwidth that is common for $t = 1, 2$.[5] The obtained variance estimator based on $\hat{A}(x)$ and $B\hat{(x)}$ is essentially a local method of moments estimator (Lewbel, 2007).

In practice, we can readily apply the bandwidth selection methods for nonparametric conditional mean estimation, which we summarize as follows.

The first approach is the adapted rule of thumb method:

$$h = C \cdot X_{sd} n^{-1/(4+2p)}$$

where p is the dimension of X; X_{sd} is the sample standard deviation of $\{X_{it}\}_{i=1}^{n}$; C is a constant and in practise we choose $C = 1.06$. Note that though $X \in R^1$, we use two dimensions based on the conditional argument $E_x(\cdot) := E(\cdot|X_{it} = X_{i\tau} = x)$. Hence, the power of n is $-1/(4 + 2p)$ rather than $-1/(4 + p)$.

Alternatively, one can directly apply the least squares leave-one-out cross validation method to the covariance, following Li et al. (2007) and Diggle and Verbyla (1998). In choosing h_A, for instance, we suggest minimizing the following cross-validation criterion[6]

$$CV_{LS}(h) = \sum_{i=1}^{n} (A^0(X_i) - \hat{A}_h(X_{-i}))^2 W(X_i),$$

where $A^0(X_i) = Y_{it}^0(Y_{it}^0 - Y_{i\tau}^0)$ with $Y_{it}^0 = Y_{it} - n^{-1}\sum_{i=1}^{n} Y_{it}$ and $\hat{A}_h(X_{-i})$ is the leave-one-out estimator of $A(x)$ with the bandwidth h, defined in equation (3.5). The $W(x)$ is a weight

[5] For the multivariate case (i.e., $p > 1$), we let

$$\omega_{i,j}(x) = \frac{K\left(H_j^{-1}(X_{i1} - x)\right) K\left(H_j^{-1}(X_{i2} - x)\right)}{\sum_{i=1}^{n} K\left(H_j^{-1}(X_{i1} - x)\right) K\left(H_j^{-1}(X_{i2} - x)\right)}$$

for $j = A, B$, where K is a non-negative p-variate kernel function, and H_j is a $p \times p$ bandwidth matrix that is symmetric and positive definite. The rest of the discussion holds if we simply consider the product kernel $K(r) = \prod_{\ell=1}^{p} k(r_\ell)$ and $H_j = h_j I_p$ for some bandwidth parameter h_j, where I_p is the identity matrix of rank.

[6] Here A and following B are generic symbols for $A1$, $A2$ and $B1$, $B2$.

function, such as the probability density function of X. If n is large, above cross validation evaluation could be performed on a random subsample of n^b units, where $n^b < n$ reduces the computational burden. The h_B can be derived similarly by replacing $(Y_{it} - Y_{i\tau})$ with $(Y_{it} - Y_{i\tau})^2$.

Though both the identification theorem in Section 2 and the estimation strategy in Section 3 are derived based on the basic setting with $T = 2$, they can be extended to more general cases with $T > 2$ accordingly. A sequential conditional independence assumption and a corresponding common support assumption are needed to identify the models and we propose a consecutive-period estimation strategy. The basic idea is for each consecutive time periods, we implement a similar estimation strategy with $T = 2$. With this we could obtain λ_t and the variance parameter $\sigma^2_{u,t}(x)$ for $t = 2, ..., T$. If the variance parameter $\sigma^2_{u,t}(x)$ is assumed to be time-invariant, i.e.,$\sigma^2_{u,t}(x) = \sigma^2_u(x)$, multiple time periods would benefit the estimation precision with larger sample size. In the following application, we illustrate this by applying proposed method to a CO2 emission panel with $T = 3$.

3.4 Asymptotics

In this section, we drive the asymptotic properties of $\hat{\sigma}^2_u(x)$ and $\hat{\sigma}^2_v(x)$. For any $t \neq \tau(\tau = 1, 2)$ and $X_{it} \in R^1$, we let

$$\widetilde{Y}_{[i,t,\tau]} = \begin{pmatrix} Y_{it} \\ Y_{it} - Y_{i\tau} \\ (Y_{it} - Y_{i\tau})^2 \end{pmatrix} \tag{3.7}$$

and denote $\widetilde{Y}_{[i,t,\tau],j}$ as the jth element of $\widetilde{Y}_{[i,t,\tau]}$, $m_j(x) = E(\widetilde{Y}_{[i,t,\tau],j}|X_{it} = X_{i\tau} = x)$ for $j = 1, 2, 3$. We assume the following regularity conditions from Yin et al. (2010).

Assumption 10. *(i) X_{it} has compact support and probability density $f(x)$, which is bounded away from zero and has two continuous derivatives.*

(ii) For any $1 \leq j_1, j_2 \leq 3$, there exists $\delta \in [0, 1)$ such that $\sup_x E[|\widetilde{Y}_{[i,t,\tau],j_1}\widetilde{Y}_{[i,t,\tau],j_2}|^{2+\delta}|X_{it} =$

$X_{i\tau} = x] < \infty$.

(iii) The conditional mean $E[\widetilde{Y}_{[i,t,\tau],j}|X_{it} = X_{i\tau} = x]$ has two continuous derivatives.

(iv) $E[\widetilde{Y}^{k_1}_{[i,t,\tau],j_1}\widetilde{Y}^{k_2}_{[i,t,\tau],j_2}\widetilde{Y}^{k_3}_{[i,t,\tau],j_1}\widetilde{Y}^{k_4}_{[i,t,\tau],j_2}|X_{it} = X_{i\tau} = x]$ has two continuous derivatives in x for $k_1, k_2, k_3, k_4 \in \{0, 1\}$, where $1 \leq j_1, j_2, j_3, j_4 \leq 3$, and j_1, j_2, j_3, and j_4 are not necessarily different.

(v) The Bandwidth satisfies $h \to 0$ and $nh^5 \to c > 0$ for some $0 < c < \infty$.

(vi) Kernel function $K(v)$ is a bounded probability density function symmetric about 0. For the δ in (ii), $\int K^{2+\delta}(v)v^j dv < \infty$ for $j = 0, 1, 2$. For two arbitrary indices v_1 and v_2, $|K(v_1) - K(v_2)| \leq K_c|v_1 - v_2|$ for some $K_c > 0$.

(vii) $|A(x)| > 1/M_1 > 0$, $|B(x)| > 1/M_2 > 0$ for all $x \in \chi$, where χ is the support of x defined in Assumption 6-(v) and M_1 and M_2 are some positive constants.

Assumption 10-(i) restricts the density of the covariates X_{it} and it is slightly stronger than Assumptions 6-(v) and 7-(iii). Assumption 10-(ii) is a moment requirement for the dependent variables. Assumptions 10-(iii) and (iv) are smoothness constraints on the conditional mean and conditional variance, respectively (Fan, 1993 and Yao and Tong, 1996). Assumption 10-(v) holds with the optimal bandwidth choice, which yields the optimal convergence rate as optimal rate of convergence $n^{-1/5}$. When $p > 1$, we suppose $nh^{4+p} \to c > 0$. Assumption 10-(vi) is a standard requirement for the kernel function (Li and Racine, 2007), which is trivially satisfied by the Gaussian and Epanechnikov kernels. Assumption 10-(vii) ensures the reciprocal of $A(x)$ and $B(x)$ are bounded which is necessary for the convergence proof of target parameters $\sigma_u^2(x)$ and $\sigma_v^2(x)$.[7]

Define $\sigma_{j_1j_2}(x)$ as the (j_1, j_2)th element of the 3×3 matrix $\Sigma(x) = Var[\tilde{Y}_{[i,t,\tau]}|X_{it} = X_{i\tau} = x]$ for $1 \leq j_1, j_2 \leq 3$, and $\hat{\sigma}_{j_1j_2}(x)$ as its consistent estimator. Then, $A(x) = \sigma_{12}(x)$ and $B(x) = \sigma_{13}(x)$. The following lemma characterizes the joint asymptotic distribution of $(\hat{A}(x), \hat{B}(x))' = (\hat{\sigma}_{12}(x), \hat{\sigma}_{13}(x))'$, which extends Theorem 1 of Yin et al. (2010). We let $\nu_0 = \int K^2(u)du$ and $\mu_2 = \int u^2K(u)du$. For any continuously differentiable function $g(x)$,

[7] Recall the target (distribution) parameter $\sigma_u(x)$ and $\sigma_v(x)$ is a power function of the nonparametric covariance $A(x)$ and $B(x)$. This is also necessary for the uniform convergence of the target parameter.

we denote $\dot{g}_s(x)$ and $\ddot{g}_{ss}(x)$ be the first and the second derivative of $g(x)$ with respect to the sth dimensional element of X, respectively.

Lemma 4. *Under Assumption 10,*

$$\sqrt{nh^2}\begin{pmatrix}\hat{\sigma}_{12}(x)-\sigma_{12}(x)-h^2\gamma_{12}(x)\\ \hat{\sigma}_{13}(x)-\sigma_{13}(x)-h^2\gamma_{13}(x)\end{pmatrix}\to_d\mathcal{N}\left(0,\frac{\nu_0}{f(x)}\begin{pmatrix}\phi_{12}^2(x) & \phi_{12,13}(x)\\ \phi_{12,13}(x) & \phi_{13}^2(x)\end{pmatrix}\right)$$

as $n\to\infty$, *where* $\gamma_{j_1j_2}(x)=\frac{\mu_2}{2}\{(\ddot{\sigma}_{j_1j_2})_{11}(x)+2(\dot{\sigma}_{j_1j_2})_1(x)\frac{\dot{f}_1(x)}{f(x)}\}+\frac{\mu_2}{2}\{(\ddot{\sigma}_{j_1j_2})_{22}(x)+2(\dot{\sigma}_{j_1j_2})_2(x)\frac{\dot{f}_2(x)}{f(x)}\}$, $\phi_{j_1j_2}^2(x)=Var[\varepsilon_{j_1j_2}(x)|X_{it}=X_{i\tau}=x]$, $\phi_{j_1j_2,j_3j_4}(x)=Cov[\varepsilon_{j_1j_2}(x),\varepsilon_{j_3j_4}(x)|X_{it}=X_{i\tau}=x]$ *and* $\varepsilon_{j_1j_2}(X)=\{\tilde{Y}_{[i,t,\tau],j_1}-m_{j_1}(X)\}\{\tilde{Y}_{[i,t,\tau],j_2}-m_{j_2}(X)\}-\sigma_{j_1j_2}(X)$.

From Lemma 4 and by the delta method, we derive asymptotic distributions of $\hat{\sigma}_u^2(x)$ and $\hat{\sigma}_v^2(x)$. The proof is in the Appendix.

Theorem 7. *Suppose Assumptions 6 and 10 are satisfied. Then, as* $n\to\infty$,

$$\sqrt{nh^2}\left\{\hat{\sigma}_u^2(x)-\sigma_u^2(x)-h^2b(x)\left(\gamma_{13}(x)+2\lambda_\tau\gamma_{12}(x)\right)\right\}\to_d\mathcal{N}\left(0,\frac{\nu_0b^2(x)\left(\phi_{13}^2(x)+4\lambda_\tau\phi_{12,13}+4\lambda_\tau^2\phi_{13}^2\right)}{f(x)}\right)$$

and

$$\begin{aligned}&\sqrt{nh^2}\left\{\hat{\sigma}_v^2(x)-\sigma_v^2(x)-h^2\big(\gamma_{12}(x)-ab(x)(\gamma_{13}(x)+2\lambda_\tau\gamma_{12})\big)\right\}\\ &\to_d\mathcal{N}\left(0,\frac{\nu_0\left\{(1-2\lambda_\tau ab(x))^2\phi_{12}^2(x)-2(1-2\lambda_\tau ab(x))ab(x)\phi_{12,13}(x)+a^2b^2(x)\phi_{13}^2(x)\right\}}{f(x)}\right)\end{aligned}$$

where λ_τ *is the added time fixed effects for period* τ *and*

$$a=1-\frac{2}{\pi}\quad\text{and }b(x)=\frac{2}{3}\left(\frac{\pi\sqrt{\pi}}{(4-\pi)\sqrt{2}}\right)^{2/3}\left(\sigma_{13}(x)+2\lambda_\tau\sigma_{12}(x)\right)^{-1/3}$$

for half-normal U_{it}; *or* $a=1$ *and and* $b(x)=\frac{2}{3}(1/2)^{2/3}(\sigma_{13}(x)+2\lambda_\tau\sigma_{12}(x))^{-1/3}$ *for exponential* U_{it}.

Under the optimal bandwidth of $h \sim n^{-1/6}$, the optimal convergence rate of the conditional variance estimators $\hat{\sigma}_u^2(x)$ and $\hat{\sigma}_v^2(x)$ is obtained as $n^{-2/6}$, which is the standard result of the kernel estimator.[8] It hence can be conjectured that the uniform rate of convergence of these two estimators is also the standard one, specifically, $O_p((\ln n/(nh^2))^{1/2} + 2h^2)$. Note that, however, this result does not extend to the regression function estimator $\hat{m}(x, \alpha)$. For instance, Evdokimov (2010) shows that, depending on either ordinary smooth or super smooth U_{it}, the convergence rate of $\hat{m}(x, \alpha)$ is of order $(\ln n/n)^c$ and $(\ln n)^c$ for some $c > 0$.[9] The main reason for this difference lies in the different strategies for identification and estimation. Identification in the present paper uses properties of (conditional) characteristic functions and their (conditional) moments, while identification in Evdokimov (2010) is based on nonparametric deconvolution techniques.

3.5 Simulation

This section presents a Monte Carlo study of the finite sample properties of the proposed estimators $\hat{\sigma}_u^2$ and $\hat{\sigma}_v^2$ in the stochastic cost frontier model for both the fixed effects and random effects specifications (Greene 2005a, 2005b). We consider the following panel data

[8] Recall that we have two kernels for a univariate X_{it} as we consider two consecutive periods.

[9] Here c is a function of d_1, d_2 and p where d_1 is the maximum continuous derivative of conditional cumulative distribution function $F_m(t|x)$, d_2 is the maximum continuous derivative of the joint density $f(x, x)$ and p is the dimension of X. For details, see the Theorem 4-5,Theorem 7-8 in Evdokimov (2010).

model with non-separable unobserved heterogeneity and added time effects:

$$\begin{aligned}
Y_{it} &= \lambda_t + m(X_{it},\alpha_i) + U_{it} + V_{it} \quad \text{for } i=1,...,n, \quad t=1,...,T=2 \\
m(x,\alpha) &= \alpha + (1+0.5\alpha)(2x-1)^3 \\
U_{it} &\sim |\mathcal{N}(0,\sigma_u^2(X_{it}))| \quad \text{or} \quad Exp(\sigma_u(X_{it})) \\
V_{it} &\sim \mathcal{N}(0,\sigma_v^2(X_{it})) \\
X_{it} &\sim iid \quad \mathcal{U}[0,1], \quad \lambda_1 = 0, \lambda_2 = 1 \\
\alpha_i &= \begin{cases} (\rho/T)\sum_{t=1}^{T}\sqrt{12}(X_{it}-0.5) + \sqrt{1-\rho^2}\phi_i & \text{for FE} \\ \sqrt{1-\rho^2}\phi_i & \text{for RE} \end{cases} \\
&\text{with } \rho = 0.5 \text{ and } \phi_i \sim iid \quad \mathcal{N}(0,1)
\end{aligned}$$

for $n = 2500$ and 10000. The following specifications for the variance of inefficiency and noise are considered:

Specification I: $\sigma_u^2 = 2$, $\sigma_v^2 = 1$;

Specification II: $\sigma_u^2 = 2X_{it}^2$, $\sigma_v^2 = 1$;

Specification III: $\sigma_u^2 = 2$, $\sigma_v^2 = X_{it}^2$;

Specification IV: $\sigma_u^2 = 2X_{it}^2$, $\sigma_v^2 = X_{it}^2$;

Since we focus on identification and estimation of the variance components which hinges on the first three conditional moments of the compound error term $\varepsilon_{it} = U_{it} + V_{it}$ in equation (3.2), the signal to noise ratio defined by $\frac{Var(\varepsilon)}{Var(\varepsilon)+Var(m(X))}$ is important. In particular, for Specification I (with constant variance) the signal-to-noise ratio is $1.73/2.73 \approx 0.63$ in the half-normal case and $3/4 = 0.75$ in the exponential case. For the remaining specifications, the average signal-to-noise ratios are between 0.2 and 0.75 for different realizations of X_{it}. Each Monte Carlo experiment is based on 1,000 replications. We use the rule of thumb bandwidth $h = 1.06 \times std(X_{it})n^{-1/6}$ for simplicity and consistency. We can also use the proposed maximum likelihood or leave-one-out cross validation method to choose the unknown bandwidths. Recall that for univariate $X_{it} \in \mathcal{R}^1$ with the special conditional argument $X_{i1} = X_{i2} = x$, we need to choose two bandwidths for each of $\omega_{i,A1}$, $\omega_{i,A2}$, $\omega_{i,B1}$, $\omega_{i,B2}$ in equation (3.5) and (3.6) respectively.[10]

We report the root integrated mean squared error ($RIMSE$), the root integrated squared bias ($RIBIAS^2$), and the root integrated variance ($RIVAR$) of the estimated variances and the root mean squared error ($RMSE$) of the estimated time effects, which are calculated as

$$RIMSE = \sqrt{\frac{1}{100}\sum_{k=1}^{100}\frac{1}{R}\sum_{r=1}^{R}[\hat{\sigma}^2_{ur}(x_k) - \sigma^2_u(x_k)]^2},$$

$$RIBIAS^2 = \sqrt{\frac{1}{100}\sum_{k=1}^{100}\left[\frac{1}{R}\sum_{r=1}^{R}\hat{\sigma}^2_{ur}(x_k) - \sigma^2_u(x_k)\right]^2},$$

$$RIVAR = \sqrt{\frac{1}{100}\sum_{k=1}^{100}\left[\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\sigma}^2_{ur}(x_k)\right)^2 - \left\{\frac{1}{R}\sum_{r=1}^{R}\hat{\sigma}^2_{ur}(x_k)\right\}^2\right]},$$

$$RMSE = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\lambda}_{2r} - \lambda_2)^2},$$

where $x_k = 0.1 + 0.008k$ for $k = 1, 2..., 100$ is the kth grid point between the 10th and 90th percentiles of x; $\hat{\sigma}^2_{ur}$ and $\hat{\lambda}_2$ is the estimate of the cost inefficiency variance function and the time effects in the rth Monte Carlo replication with $R = 1000$.[11]

Table 3.1 contains the results for the design $n = 2500$ and $T = 2$. The rows of the table are divided into four panels for each of our four specifications: I, II, III and IV, respectively. For example, the first panel contains the results for Specification I ($\sigma^2_u = 2$ and $\sigma^2_v = 1$,). The first three rows of each panel contain the $RIMSE$, the $RIBIAS^2$ and the $RIVAR$ (respectively) for the proposed estimator. The last row contains the $RMSE$

[10] In the case of N=2500, T=2, one Monte Carlo simulation is less than 1 second with rule of thumb bandwidth but about one hour implementing the leave-one-out cross validation method to choose the bandwidth. In the application section, we use leave-one-out cross validation to choose the bandwidths.

[11] The support trimming procedure ensures that the joint density $f_{X_{i1},X_{i2}}(x, x) > 0$, which is a key identification assumption in the nonparametric panel setting.

of the estimated time effects. The columns of results are (left to right) for fixed effects with half-normal inefficiency, random effects with half-normal inefficiency, fixed effects with exponential inefficiency, and random effects with exponential inefficiency.

Table 3.1 suggests that both proposed estimators of the variance components perform reasonably well. Vertically, the proposed method performs best (compared to itself) for the both heteroskedastic specification (fourth row panel), then heteroskedastic u or heteroskedastic v (third and second row panel) and lastly both homoskedastic specification (first row panel). The rule of thumb bandwidth choice may be the driving force behind this result since the bandwidth should go to infinity in the homoskedastic specifications in light of the irrelevance of the covariates. The feasible estimators perform better (compared to itself) when inefficiency is exponentially distributed than when it is half-normal (first and second column VS third and fourth column). This is probably due to the fact that in all specifications the random noise v_{it} is normally distributed, and disentangling moments from the same distributional family is always more difficult than from different families.[12] Another interesting pattern is that with the proposed method the fixed effects models and the random effects models yield similar results in both the variance and time effects estimation. This corresponds to Theorem 4 in which the identification and estimation of σ_u^2 and σ_v^2 does not hinge on the fixed effects or random effects assumptions.[13] The slight difference between them is an artifact of finite sampling variability.

Continuing with Table 3.1, one may observe that the estimated time effects has a smaller $RMSE$ with half-normally distributed inefficiency when compared with its exponential counterpart. This may come from the rule of thumb bandwidth selection as we use the same bandwidth for all the conditional first-differences and conditional covariances. It is not a general rule.

[12] Half-normal and normal distribution are both in the super smooth distributional family while exponential distribution is in the ordinary smooth distributional family. It is always more difficult to disentangle inefficiency from the random noise in the former case than in the latter.

[13] Actually, it is also true for the time effects identification and estimation as we could observe in the proof of Theorems 5 and 6 .

Similar findings can be found when we increase the sample size to 10000 for each period. Table 3.2 reports the results for the larger sample size design $n = 10000$ and $T = 2$. The proposed estimator performs better in all specifications with either heteroskedastic inefficiency or heteroskedastic noise or both than the case with none of them (second, third and fourth panel vs first panel). Exponential stochastic frontier model yields more precise variance estimators (in terms of $RIMSE$) than the half-normal counterpart while the differences between fixed effects estimators and random effects estimators are negligible within the same inefficiency distribution. Both proposed estimators of the variance components and the time effects have decent identification power in terms of $RIMSE$ and $RMSE$. Another main finding is that in all specifications, the proposed estimators perform better in Table 3.2 than in Table 3.1. For instance, the first $RIMSE$ 0.609 and $RMSE$ 0.036 in Table 3.2 versus 0.847 and 0.065 in Table 3.1 for the half-normal fixed effects model. This demonstrates the consistency and decent rate of convergence of the proposed estimators.

3.6 Application

We apply the proposed method to study an environmental Kuznets curve (EKC) on CO2 emission and economic development. We are especially interested in the relationship between CO2 emission technology development and human capital represented by average schooling years and capital stock per capita across 136 countries from 1990 to 2014. The EKC has been a popular approach among economists to model ambient pollution concentrations and aggregate emissions since Grossman and Krueger (1991) introduced it almost thirty years ago. It is a hypothesis that states the environmental impacts or pollutant emissions are an inverse U-shaped function of income per capita. For details, please refer to Stern (2017) which provides an excellent review on EKC. One theory behind the EKC is that technology improvement embedded in the production and environmental friendly processing may drive the inverse-U shape relationship between emissions and income per capita. This is referred to as the technique effect (Copeland and Taylor, 2004). The proposed nonparametric panel model could help provide quantitative evidence on this theory. As far as we acknowledge, there are few trials on this except Bertinelli and Strolb (2005), Azomahou et al. (2006) and Lee et al. (2019).

3.6.1 Data and Model

The data comes from the World Bank which provides CO2 emission per capita (in metric kilograms) from 1990 to 2014 worldwide and Penn World Table version 9.1 from where capita stosck per capita (in 2011 US dollars) is obtained.[14] The average schooling years worldwide from 1990 to 2017 is scripted from the Human Development Reports under the United Nation Development Programme. We merge the data by country and year and obtain a balance panel of 136 countries and 25 years from 1990 to 2014. As the average schooling by country evolves very slowly and changes very little across years, we choose three equidistant periods to investigate the technique effects underlying the EKC: 1990, 2002 and 2014.[15] Summary statistics of the collected data are reported in Table 3.3 . The average schooling years and capita stock per capita increase slowly across the three periods.

Existing studies typically specify a log linear panel model with one-way or two-way fixed effects to test the inverse-U shape relationship between CO2 emission and GDP per capita, namely the EKC hypothesis (Bertinelli and Strolb, 2005; Azomahou et al., 2006 and Lee et al., 2019). Here we advance steps further. First, rather than merely exploring the relationship between the CO2 emission and GDP per capita, which is determined by human capital and capita stock in the Solow growth model, we focus on the evolving of CO2 emission technology (i.e., the CO2 emission productivity) based on the two fundamentals of GDP per capita: human capital represented by schooling years and capital stock per capita across different countries in the past 25 years. Second, a flexible relationship between

[14] The capita stock per capita equals to the ratio between capita stock (in million 2011 US dollars) and population (in millions) in the Penn World Table 9.1.

[15] We also tried several other choices such as five periods or twelve periods and similar results could be found.

the CO2 emission and the two fundamental inputs of GDP per capita is modeled with a nonparametric panel model with non-separable country fixed effects and added time effects. Specifically we consider a model as follows

$$C_{it} = \lambda_t + m(\alpha_i, x_{it}) + u_{it} + v_{it} \quad i = 1, 2..., N; \quad t = 1, 2..., T, \tag{3.8}$$

where $C_{it} = \ln(CO2_{it})$ is the logarithm of CO2 emission (in metric kilograms) per capita for country i in year t; λ_t is the time effects and α_i is the non-separable unobserved heterogeneity (i.e., non-separable fixed effects or random effects); $u_{it} > 0$ is a time-varying CO2 emission productivity emancipated from the technology development or technology adaption in country i at year t; v_{it} is random noise which is assumed to be (conditional) symmetric; and $x_{it} = \ln X_{it}$ where X_{it} includes the average schooling years and capital stock per capita. The $m(\alpha_i, X_{it})$ is a general nonparametric function which models the complicated production process of CO2 with a panel.

One thing worthy noting here is that though the CO2 emission productivity u_{it} is unobserved, we assume it follows a Half-Normal or Exponential distribution as that in a typical stochastic frontier model for efficiency or productivity analysis, i.e., $u_{it} \sim |\mathcal{N}(0, \sigma_u^2)|$ or $Exp(\sigma_u)$. Hence, the mean productivity is determined by the variance parameter, namely, $E(u_{it}) = (1 - \frac{2}{\pi})\sigma_u(x_{it})$ or $E(u_{it}) = \sigma_u(x_{it})$ which provides key insight on the evolution of CO2 emission technology or adaption worldwide during the past 25 years.[16] Understanding the nature of technology improvement related to environment is crucial for understanding the determinants of Green House Gas (GHS) emission like CO2 emission. The latter could provide practical and insightful directions for policy makers who care about GHS emission and more broadly, global warming and climate change.

[16] We assume $u_{it} \sim |\mathcal{N}(0, \sigma_u^2)|$ for the following analysis. For exponentially distributed u_{it}, similar analysis could be derived.

3.6.2 Model Estimation

To apply the proposed method, we first check the validity of common support assumptions by drawing probability distribution density (pdf) functions of all covariates. Specifically, the pdf of each covariate across three periods are depicted in one graph in Figure A: the upper one shows those for $\ln(CapitalStockPercapita)$ and the bottom for $\ln(AvgSchoolYear)$. We could observe that both of them share a very good common support, although their mean value shift slowly across the three periods. Another assumption is conditional independence assumption. It assumes that given the contemporaneous inputs (capital stock per capita and average school year), the random noise is independent with the past time period inputs and the country specific effect. This assumption is not unreasonable if the random noise mainly comes from the measurement error and the conditioning covairates contain the main factors that affect the output.[17]

We apply the proposed nonparametric kernel estimation procedure with a panel of three periods: 1990, 2002, 2014. Hence, there are two time effects and four conditional covariances to be estimated as follows:

$$E[Y_{it} - Y_{i1} | X_{it} = X_{i1} = x] = E[\varepsilon_{it} + \lambda_t - \varepsilon_{i1} | X_{it} = X_{i1} = x] = \lambda_t$$

where $t = 2$ or 3 for year 2002 and 2014 and year 1990 is set as the benchmark year with normalization $\lambda_1 = 0$, and

$$A_{12}(x) = Cov_x(Y_{i1}, Y_{i1} - Y_{i2}); \qquad A_{23}(x) = Cov_x(Y_{i2}, Y_{i2} - Y_{i3});$$
$$B_{12}(x) = Cov_x\big(Y_{i1}, (Y_{i1} - Y_{i2})^2\big); \quad B_{23}(x) = Cov_x\big(Y_{i2}, (Y_{i2} - Y_{i3})^2\big);$$

where $Cov_x(C, D) = E_x(CD) - E_x(C) * E_x(D)$ with $E_x(\cdot) = E(\cdot | X_t = X_{t+1} = x)$ and $C = Y_{i1}$ or Y_{i2}, $D = Y_{i1} - Y_{i2}$ or $(Y_{i1} - Y_{i2})^2$. The bandwidths are chosen by leave-one-out

[17] We would talk about this in detail when we explain the empirical results and talk about the implications.

Cross Validation and we choose the same bandwidth for one covariate in two consecutive periods.[18] Specifically, we search the grid of 0.2 to 10 times the rule of thumb bandwidth in each period: h_{rot1}, h_{rot2} and h_{rot3}, and obtain the optimal bandwidth by minimizing the cross validation objective function. The optimal bandwidths and rule-of-thumb bandwidths, as a benchmark, are reported in Table 3.4.

Based on the cross validation bandwidth, we obtain the variance parameters as a function of inputs which determines the mean CO2 emission productivity, and the constant time effects for each period. Specifically, $\lambda_2 = 0.2107$ and $\lambda_3 = 0.2989$. Note that these estimates indicate an yearly increase of 2.79% and 3.88% respectively concerning the fact that the mean logarithm of CO2 emission per capita in 2002 and 2014 are 7.5366 and 7.7134. It shows an accelerating trend of CO2 emission per capita from 2002 to 2014 than from 1990 to 2002. At the same time period, the world average GDP per capita increase from $4,290$ (current US) dollars in 1990 to $5,527$ dollars in 2002 to $10,934$ dollars in 2014.[19] Though both increase rapidly, the growth GDP per capita overwhelmingly dominates that of CO2 emission. This reflects the nonlinear relationship between the CO2 emission per capita and GDP per capita in which the environmental technology development may be a driving force.

3.6.3 Empirical Results

Figure 3.1 shows a 3D surface indicating the mean CO2 emission productivity (i.e., CO2 emission technology) as a function of two inputs: $\ln(CapitalStockPercapita)$ and $\ln(AvgSchoolYear)$. In general, the 3D surface representing the estimated nonparametric relationship between the unobserved CO2 emission productivity and the observed inputs is very smooth. Interestingly, the CO2 emission productivity is an increasing function of the average schooling year but shows an inverse U-shape relationship with capital stock per capita. We could observe this point more clearly in Figures 3.2 and 3.3 when we take out two median slices of the 3D surface and explore the heterogeneous effects of different inputs on the CO2 emission

[18] This is a convenient simplification as the covariates change very little in two consecutive periods.

[19] The GDP data comes from World Bank.

productivity. The dash lines shows the 95% confidence interval bands derived from 199 bootstrap simulations. Fixing the other covariate at its median, CO2 emission productivity demonstrates an obvious monotone increasing trend with average schooling year and an inverse U-shape relationship with capital stock per capita with a turning point around 9.7 (i.e., 16318 dollars per capita). Both these patterns are statistically significant. The results indicate that capita stock may be a driving force underlying the technique effects but human capital is not.[20] This has huge implications for environmental policy makers worldwide. Details would be laid out in the implication subsection.

In Figure 3.1 , we also locate and highlight the world average CO2 emission productivity in each of the three investigated periods by its average schooling year and average capital stock per capita. Overall, the world average CO2 emission productivity was increasing in the first period in the 1990s and then decreasing over the second period, which covers the post 21th century years. This pattern is very intuitive concerning that one of the earliest global climate agreement - the Kyoto Protocol - was adopted in Kyoto, Japan, in 1997 and entered into force in 2005. However, there is a long way to go for the world as a whole given the fact that CO2 emission productivity is still very high and imbalanced.

To see the imbalance among different countries on CO2 emission productivity, we obtain the CO2 emission productivity of China and US similarly with information on average schooling year and average capital stock per capita and highlight them in Figure 3.4 and Figure 3.5 , respectively. In Figure 3.4 , as the biggest developing country in the world (in terms of GDP), China overally followed the evolving path of the world average. Specifically China was better than world average in 1990 in terms of CO2 emission productivity (1.42 VS 1.66), was slightly worse than the world average in 2002 (1.824 VS 1.74), which it outperformed a little bit in 2014 (1.14 VS 1.29) mainly due to a smaller than average schooling year compared with the world average (2.03 VS 2.08).[21] However, as the biggest developed

[20] More education should make people more aware of environmental protection. However, we don't find that in the data.

[21] Recall that CO2 emission productivity is the negative side of CO2 emission technology development, so the smaller it is, the better it is.

country in the world, the US was better than the world average across all three periods. China's CO2 emission productivity is almost three times worse of that of US in 2014 (1.13 VS 0.41). Even in 2014, China was lagged behind the 1990 US in terms of CO2 emission productivity (1.14 VS 0.68). There is a huge gap between developing countries and developed countries on CO2 emission productivity(i.e., CO2 emission technology) in the world.

To get an overview of the CO2 emission productivity among all 136 countries in the collected data, we fit a liner model for the ridge line of the 3D surface in Figure 3.1 and project it onto the XY plane. The ridge line represents the inputs combinations which achieve the maximum value of the Z axis, i.e., the maximum value of CO2 emission productivity. The fitted ridge line is

$$\ln(AvgSchoolYear) + \underset{(0.02)}{2.15} * \ln(CapitalStockPercapita) - \underset{(0.21)}{21.76} = 0$$

and the coefficients are statistically significant at 99% confidence interval. For each of the three periods, a scatter plot with the fitted ridge line can be drawn using the information of $\ln(CapitalStockPercapita)$ and $\ln(AvgSchoolYear)$ in the data. We depict them in Figures 3.6 -3.8 with highlights of OECD countries and BRICS countries. In each figure, the horizontal and vertical red lines represent the mean value of corresponding variables and the yellow line represent the fitted ridge curve which stays the same over three periods. Note that as time goes on, both the OECD countries and the BRICS countries move away from the "bad" ridge line and the OECD countries become more concentrated on the right upper quadrant. India, another large developing country, followed the path of China in the reduction of CO2 emission though it was still the least developed in terms of CO2 emission technology among BRICS countries in 2014. Another finding is that in the bottom left quadrant, most of African countries are trapped in the less developed (in terms of capita stock) and less CO2 emission circle. They are in dire need of assistance on CO2 emission technology from the developed countries to climb and cross over the "bad" ridge curve of CO2 emission.

3.6.4 Implications and Discussion

The observed fact that CO2 emission productivity (i.e., CO2 emission technology development) monotonically increases with human capital (represented by average schooling years) and demonstrates an inverse U-shape relationship with the capital stock per capita from the empirical analysis has insightful implications in at least three aspects. First, increasing the capita stock in poor and heavy CO2 emission countries could be a useful and effective recipe to reduce the CO2 emission globally and enforce the Paris Climate Agreement. There are two underlying explanations to behind this: the needed CO2 emission technology, or environmental friendly technology, are capital-intensive and people would care about the overall well being such as climate change and global warming more when they are getting richer. Secondly,the education we received so far has little effect on reducing the CO2 emission per capital or is far from enough for people to put carbon reduce propaganda into daily action. One explanation is that some people just do not believe that it matters at all. Education policy makers should reflect on this. Another explanation is that as people receive more education, they accumulate wealth and tend to enjoy life in a luxury way: driving too much and alone, frequent and unnecessary travel (especially flights), excessive entertainment, ect. The education making us rich is not enough to make us environmentally self-conscious. Thirdly, the inverse U-shape relationship between CO2 emission productivity provides a possible mechanism for a similar correlation between CO2 emission and GDP per capita in a typical EKC. Capital stock could be the driving force behind the EKC hypothesis rather than human capital.

Nevertheless, one empirical study on CO2 emission is not enough to demonstrate all the stunting questions or challenges on climate change and global warming economics. One caution is on the usage of the proposed model. Even though the proposed nonparametric panel model incorporates an added time effects and non-separable country fixed effects, allows flexible CO2 production function and imposes less restrictive distributional assumption on random noise (i.e., conditional symmetry), we still specify a Half-Normal (or Exponential) distributional assumption on the unobserved CO2 emission productivity. Though typically used by researcher in stochastic frontier analysis, this distributional assumption may not be truth or even close to the truth.[22] It may just be a convenient way to model the unknowns. Another caveat comes from the vulnerability of the conditional independence assumption. Due to data limitation, we only get information of CO2 emission per capita, average schooling years and capita stock per capita with a balance panel of 136 countries over 25 years from 1990-2014. Other factors such as political freedom, output structure or even trade may also impact the CO2 emission per capita (Stern, 2017). Though we build our CO2 production on a Cobb-Douglas model with main inputs human capital and capital stock, there is still the possibility that the conditional independence assumption is violated due to omitted variables other than the main inputs. Also, information on 136 countries may not reflect the whole picture of world, but hopefully it contains the major information of it.

3.7 Conclusion

We propose a new methodology to identify and estimate a nonparamtric panel model with composed errors. A typical example is a stochastic production/cost frontier model for panel data. Specifically, we are interested in identifying and estimating the variance parameters of the time-varying inefficiency (or productivity) as a function of multiple inputs or environmental variables. Compared with existing methods, the proposed methodology (a) doesn't impose log-linearity for the cost/production function and incorporates a non-separable unobserved heterogeneity and added time effects; (b) allows for heteroskedastic inefficiency and noise which may be a function of environmental variables; and (c) relaxes the distribution assumption for the random noise which is typically assumed to be normal or Laplace in the literature. Identification and estimation of the unknown production/cost function is built

[22] The minimum assumption is to forge the distribution assumption on the unobserved productivity and just keep the conditional symmetry assumption on the random noise. But with that the best we could obtain is the odd cumulants (or moments) of the productivity rather than a whole distribution. For details, please refer to Florens. et al. (2019).

on the novel deconvolution methodology of Evdokimov (2010). Identification and estimation of the time effects and variance parameters does not require specific information about the unspecified cost/production function as they are built upon the conditional first-difference transformation of the model. The proposed method for estimating the variance parameters is straightforward and easy to implement as it requires no deconvolution techniques.

As a useful demonstration, we apply the proposed method to study the evolving of CO2 emission productivity on human capital and capital stock with a collected panel of 136 countries over 25 years, which yields useful insights. While CO2 emission productivity increases with human capital accumulation, we find an inverse U-shape relationship of CO2 emission productivity on capital stock which is similar to that of CO2 emission on GDP per capita in a typical Environmental Kuznets Curve (EKC).

For future research, some refinements of the nonparametric estimation can be pursued. Hall and Horowitz (2013) propose a new bootstrap method to construct more precise nonparametric confidence bands for estimated functions which can be directly applied on our proposed estimators. Also the distribution assumption on inefficiency (productivity) could be relaxed to one-side distributed and with just conditional symmetry assumption of the random noise. In that case, we could identify and estimate the odd cumulants (or moments) of the unobserved inefficiency (or productivity). For details, please refer to Florens et al. (2019). This can be a straightforward and useful extension based on the present results.

Table 3.1: DESIGN I, n=2,500, T=2

			$U_{it} \sim \lvert N(0,\sigma_u^2)\rvert$		$U_{it} \sim Exp(\sigma_u)$	
σ_u^2	σ_v^2		Fixed Effects	Random Effects	Fixed Effects	Random Effects
			$(\hat{\sigma}_u^2,\ \hat{\sigma}_v^2)$	$(\hat{\sigma}_u^2,\ \hat{\sigma}_v^2)$	$(\hat{\sigma}_u^2,\ \hat{\sigma}_v^2)$	$(\hat{\sigma}_u^2,\ \hat{\sigma}_v^2)$
		$RIMSE$	(0.864, 0.300)	(0.845, 0.296)	(0.535, 0.388)	(0.536, 0.386)
2	1	$RIBIAS^2$	(0.216, 0.055)	(0.205, 0.062)	(0.084, 0.114)	(0.152, 0.096)
		$RIVAR$	(0.831, 0.294)	(0.823, 0.292)	(0.540, 0.376)	(0.538, 0.373)
$\lambda_2 = 1$		$RMSE$	0.065	0.063	0.090	0.080
		$RIMSE$	(1.033, 0.345)	(1.006, 0.334)	(0.300, 0.248)	(0.289, 0.238)
$2x^2$	1	$RIBIAS^2$	(0.848, 0.246)	(0.817, 0.261)	(0.120, 0.123)	(0.118, 0.071)
		$RIVAR$	(0.588, 0.217)	(0.588, 0.216)	(0.272, 0.233)	(0.263, 0.227)
$\lambda_2 = 1$		$RMSE$	0.055	0.054	0.065	0.063
		$RIMSE$	(0.640, 0.206)	(0.612, 0.204)	(0.492, 0.310)	(0.489, 0.312)
2	x^2	$RIBIAS^2$	(0.203, 0.042)	(0.126, 0.039)	(0.192, 0.084)	(0.107, 0.081)
		$RIVAR$	(0.609, 0.201)	(0.600, 0.199)	(0.498, 0.298)	(0.498, 0.301)
$\lambda_2 = 1$		$RMSE$	0.051	0.050	0.074	0.075
		$RIMSE$	(0.673, 0.196)	(0.577, 0.172)	(0.245, 0.173)	(0.237, 0.165)
$2x^2$	x^2	$RIBIAS^2$	(0.525, 0.130)	(0.417, 0.098)	(0.090, 0.035)	(0.076, 0.016)
		$RIVAR$	(0.421, 0.146)	(0.398, 0.142)	(0.228, 0.169)	(0.225, 0.164)
$\lambda_2 = 1$		$RMSE$	0.037	0.037	0.047	0.048

Notes: $RIMSE$, $RIBIAS^2$, and $RIVAR$ are "Root of the Integrated Mean Squared Error," "Root of the Integrated Squared Bias," and "Root of the Integrated Variance," respectively. $RMSE$ refers to the Root Mean Squared Error of the time effects λ_2. "$Exp(b)$" is the exponential pdf: $f(x) = \frac{1}{b}e^{-\frac{x}{b}}$ for $x \geq 0$.

Table 3.2: DESIGN II, n=10,000, T=2

			$U_{it} \sim \|N(0, \sigma_u^2)\|$		$U_{it} \sim Exp(\sigma_u)$	
σ_u^2	σ_v^2		Fixed Effects	Random Effects	Fixed Effects	Random Effects
			$(\hat{\sigma}_u^2, \quad \hat{\sigma}_v^2)$	$(\hat{\sigma}_u^2, \quad \hat{\sigma}_v^2)$	$(\hat{\sigma}_u^2, \quad \hat{\sigma}_v^2)$	$(\hat{\sigma}_u^2, \quad \hat{\sigma}_v^2)$
		RIMSE	(0.601, 0.219)	(0.603, 0.218)	(0.351, 0.251)	(0.356, 0.253)
2	1	$RIBIAS^2$	(0.084, 0.015)	(0.062, 0.011)	(0.019, 0.044)	(0.033, 0.036)
		RIVAR	(0.597, 0.218)	(0.601, 0.218)	(0.353, 0.245)	(0.358, 0.249)
$\lambda_2 = 1$		RMSE	0.036	0.037	0.048	0.048
		RIMSE	(0.691, 0.235)	(0.669, 0.225)	(0.196, 0.161)	(0.188, 0.158)
$2x^2$	1	$RIBIAS^2$	(0.548, 0.156)	(0.518, 0.163)	(0.081, 0.044)	(0.072, 0.047)
		RIVAR	(0.421, 0.159)	(0.424, 0.159)	(0.179, 0.153)	(0.174, 0.153)
λ_2		RMSE	0.029	0.030	0.036	0.035
		RIMSE	(0.423, 0.139)	(0.407, 0.135)	(0.329, 0.214)	(0.323, 0.211)
2	x^2	$RIBIAS^2$	(0.143, 0.024)	(0.095, 0.019)	(0.026, 0.041)	(0.062, 0.040)
		RIVAR	(0.397, 0.137)	(0.400, 0.134)	(0.332, 0.210)	(0.326, 0.207)
λ_2		RMSE	0.028	0.029	0.043	0.041
		RIMSE	(0.466, 0.133)	(0.401, 0.121)	(0.167, 0.111)	(0.160, 0.108)
$2x^2$	x^2	$RIBIAS^2$	(0.364, 0.085)	(0.283, 0.062)	(0.074, 0.027)	(0.062, 0.014)
		RIVAR	(0.290, 0.102)	(0.284, 0.104)	(0.149, 0.108)	(0.148, 0.107)
λ_2		RMSE	0.020	0.021	0.026	0.026

Notes: $RIMSE$, $RIBIAS^2$, and $RIVAR$ are "Root of the Integrated Mean Squared Error," "Root of the Integrated Squared Bias," and "Root of the Integrated Variance," respectively. $RMSE$ refers to the Root Mean Squared Error of the time effects λ_2. "$Exp(b)$" is the exponential pdf: $f(x) = \frac{1}{b}e^{-\frac{x}{b}}$ for $x \geq 0$.

Table 3.3: Summary Statistics for CO2 Emission Data, 1990-2014

	Full Sample		Year 1990		Year 2002		Year 2014	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
ln(CO2_emission)	7.54	1.67	7.36	1.79	7.53	1.70	7.71	1.49
ln(AvgSchoolYear)	1.86	0.58	1.62	0.66	1.88	0.53	2.08	0.44
ln(CapitaStock)	9.96	1.57	9.37	1.58	9.82	1.47	10.69	1.37
N	408		136		136		136	

Notes: CO2 emission per capita is in metric kilograms. Average schooling years refer to Average number of years of education received by people ages 25 and older, converted from education attainment levels using official durations of each level. Capita Stock per capita is in 2011 US dollars.

Figure A. Common Support of Selected Covariates

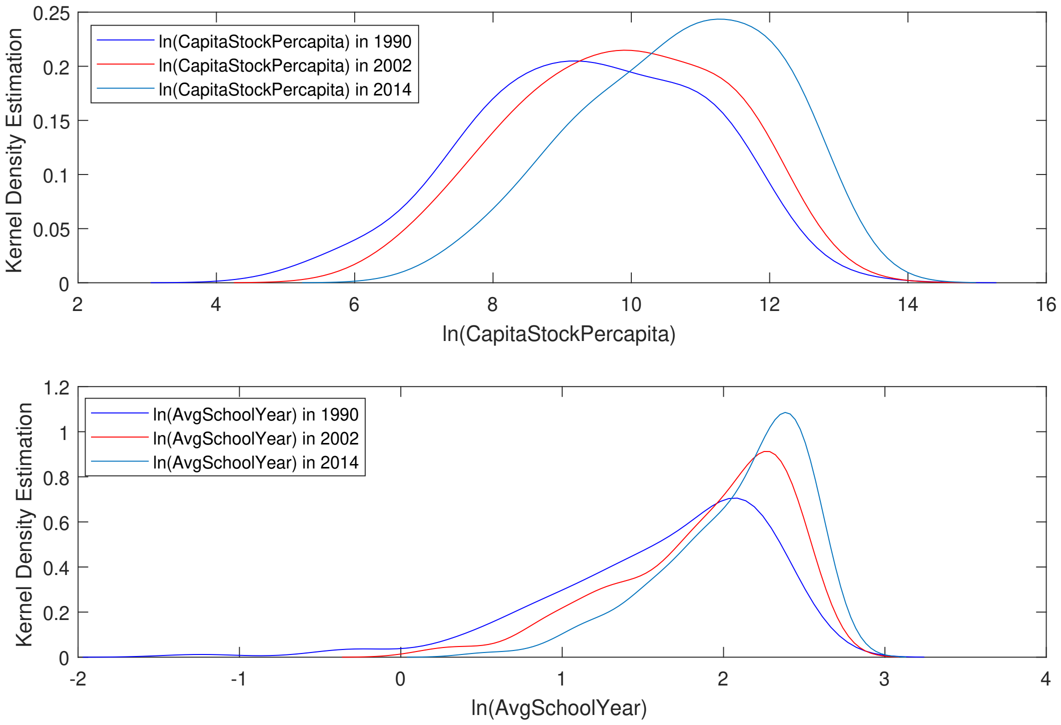


Table 3.4: Bandwidths for Nonparametric Covariance Estimators

	Cross-Validation				Rule-of-Thumb		
Covariates	$h_{A_{12}}$	$h_{A_{23}}$	$h_{B_{12}}$	$h_{B_{23}}$	h_{rot1}	h_{rot2}	h_{rot3}
ln(CapitaStockPercapita)	0.908	2.179	0.505	2.188	0.740	0.686	0.640
ln(AvgSchoolYear)	0.759	0.835	1.086	3.017	0.309	0.246	0.203

Notes: h_{A_j}, h_{B_j} are bandwidths for covariance $A_j(x)$ and $B_j(x)$ which are defined in the application section.

Figure 3.1: CO2 Emission Productivity as a Function of School Year and Capita Stock.

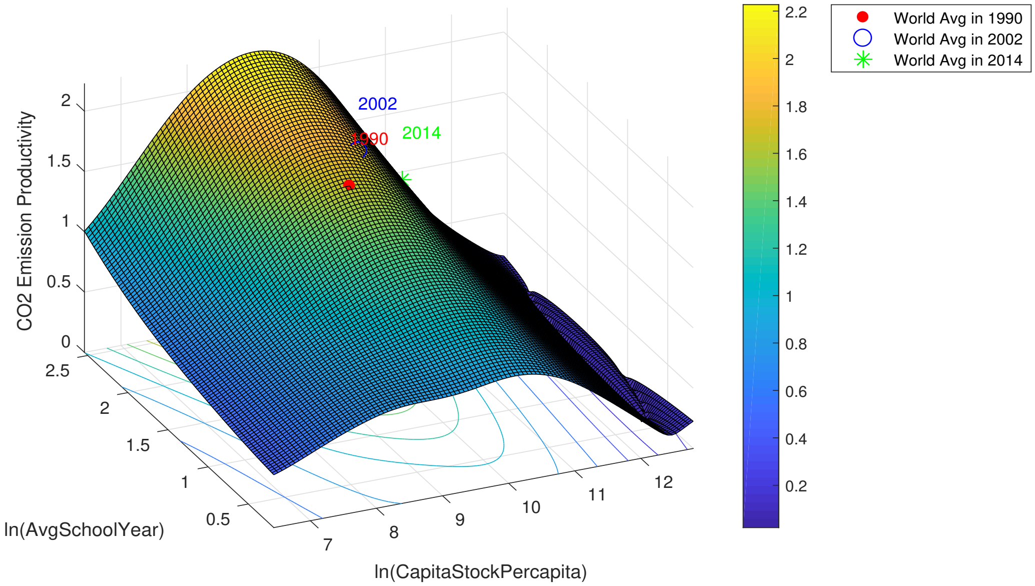


Figure 3.2: CO2 Emission Productivity as a Function of School Year at Median Capita Stock.

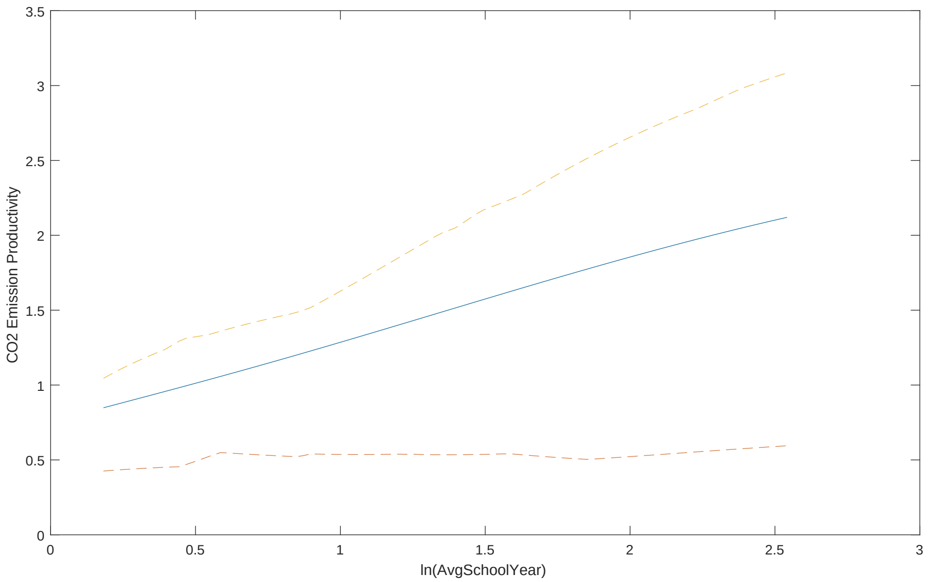


Figure 3.3: CO2 Emission Productivity as a Function of Capita Stock at Median School Year.

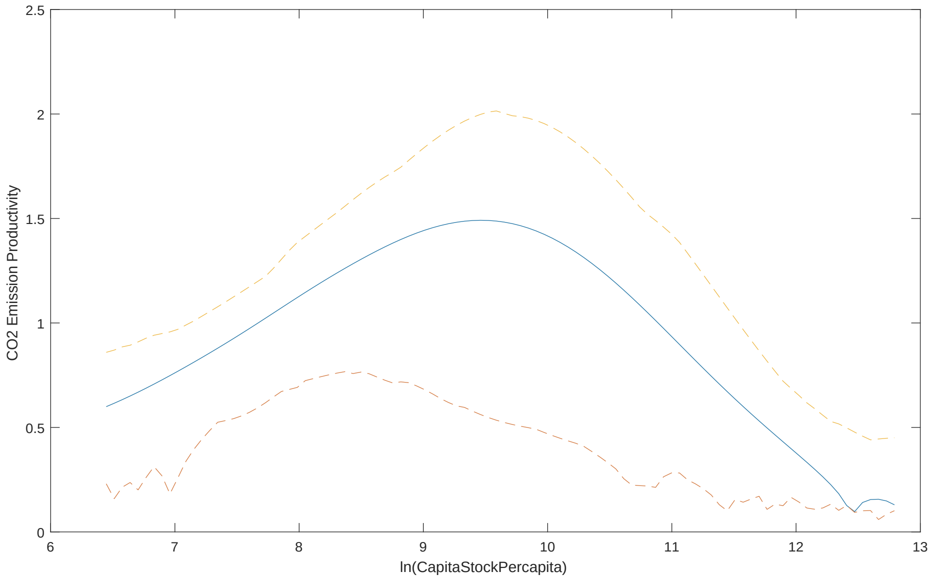


Figure 3.4: China's CO2 Emission Productivity Over Periods.

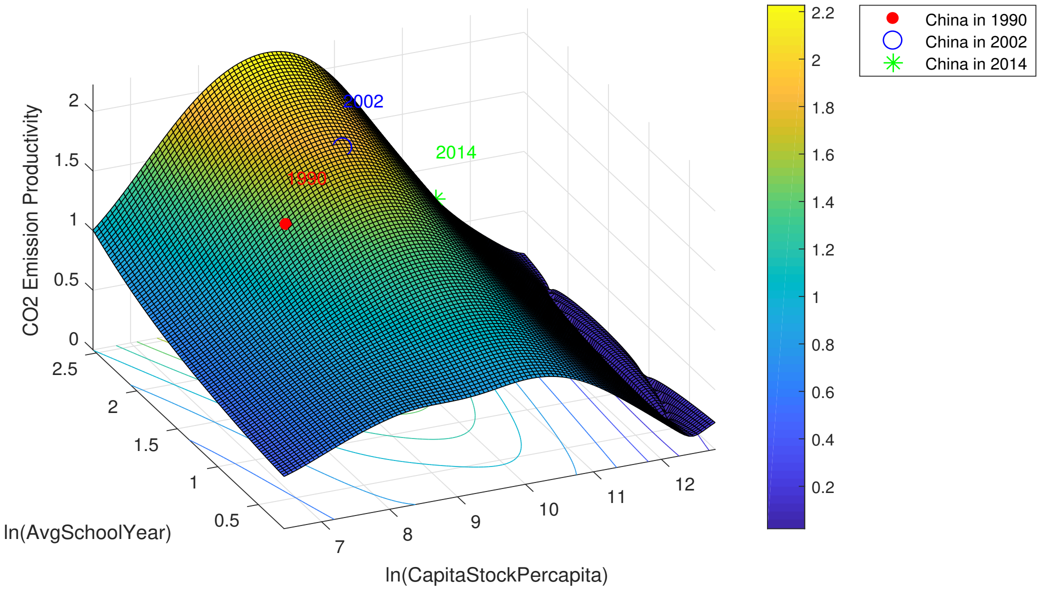


Figure 3.5: US's CO2 Emission Productivity Over Periods.

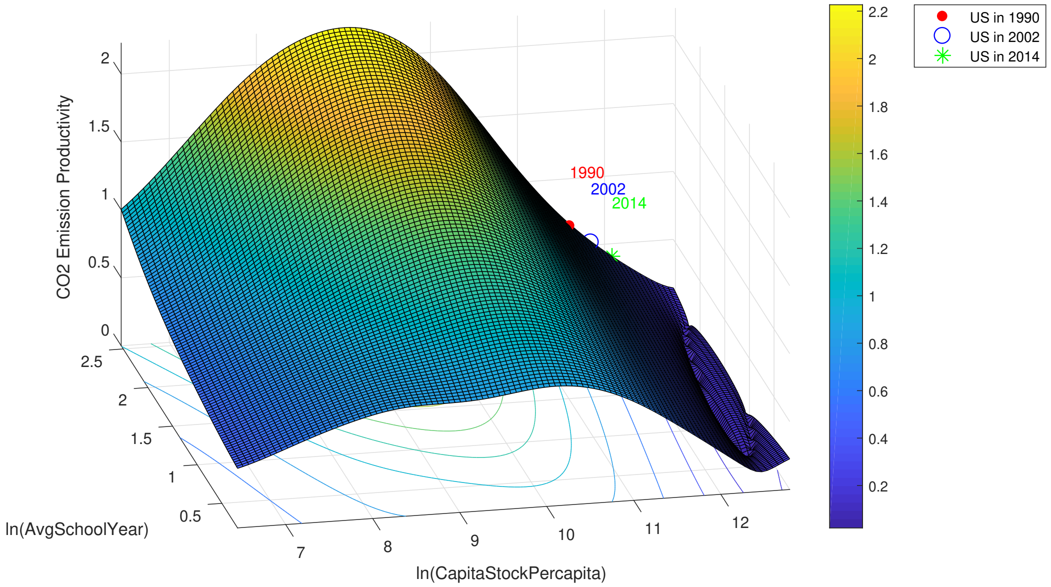


Figure 3.6: Overview of All Countries' CO2 Emission Productivity in 1990

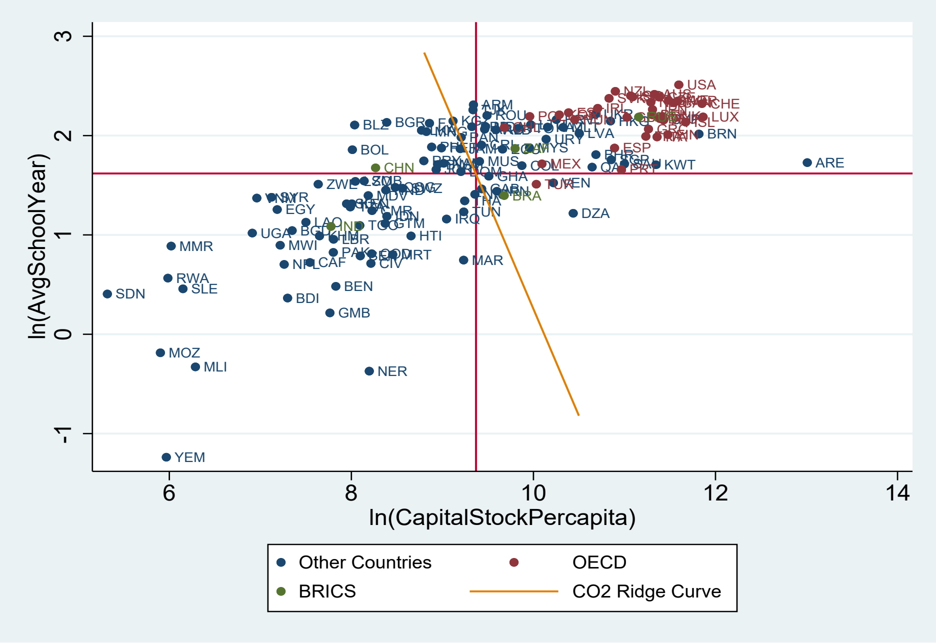


Figure 3.7: Overview of All Countries' CO2 Emission Productivity in 2002

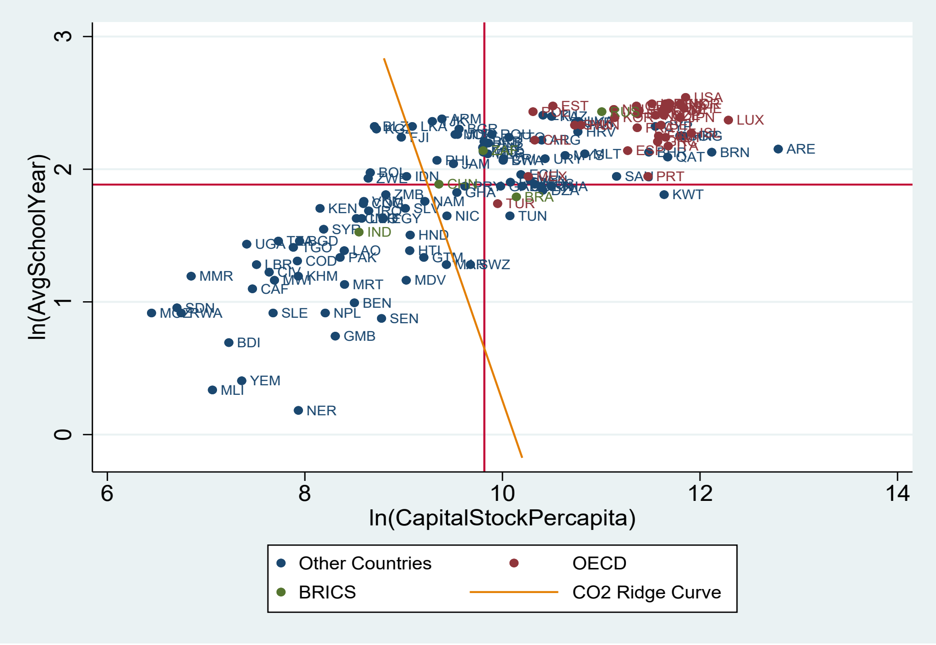
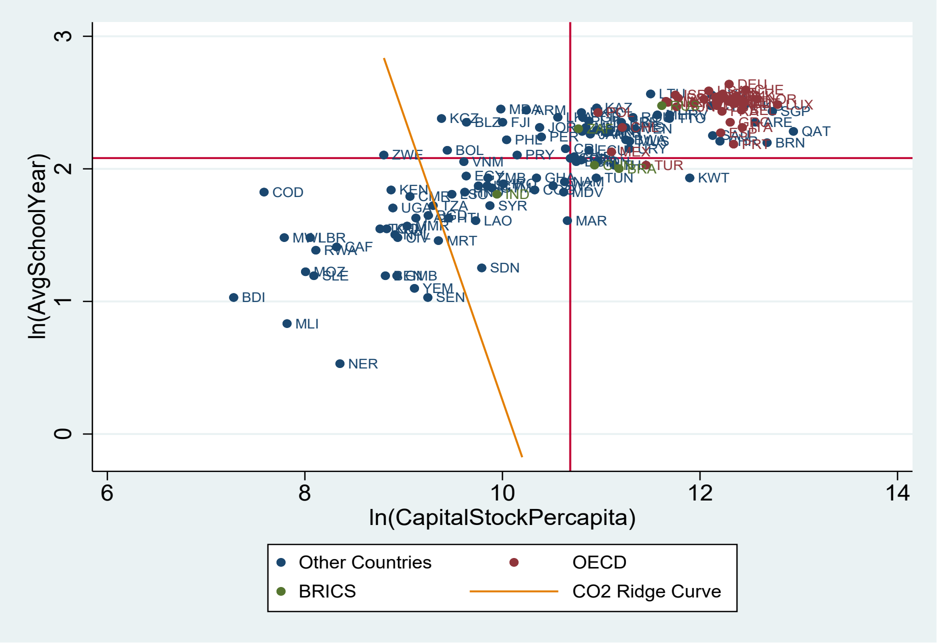


Figure 3.8: Overview of All Countries' CO2 Emission Productivity in 2014



Appendices

3.A Proof of Theorem 4

Proof. 1. Rewrite the model as:

$$Y_{it} = \tilde{m}(X_{it}, \alpha_i) + \tilde{\varepsilon}_{it} \tag{3.9}$$

$$\tilde{\varepsilon}_{it} = \lambda_t + U_{it} + V_{it} - E[U_{it}], \quad i = 1, ..., n, \quad t = 1, ..., T \tag{3.10}$$

Observe that $E[Y_{it} - Y_{i1}|X_{it} = X_{i1} = x] = E[\varepsilon_{it} + \lambda_t - \varepsilon_{i1}|X_{it} = X_{i1} = x] = \lambda_t$ with the normalization $\lambda_1 = 0$, then the time effects λ_t is identified.

Observe that $\tilde{m}(X_{i1}, \alpha_i) = \tilde{m}(X_{i2}, \alpha_i)$ when $X_{i1} = X_{i2} = x$. For any $x \in \chi$,

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} |\{X_{i1} = X_{i2} = x\} = \begin{pmatrix} \tilde{m}(x, \alpha) + \tilde{\varepsilon}_{i1} \\ \tilde{m}(x, \alpha) + \tilde{\varepsilon}_{i2} \end{pmatrix} |\{X_{i1} = X_{i2} = x\}. \tag{3.11}$$

There are four conditions (which come from the original Kotlarski's Lemma) to check before applying Lemma 1 on Evdokimov and White (2012): (1) $\tilde{m}$, $\tilde{\varepsilon}_{i1}$ and $\tilde{\varepsilon}_{i2}$ are mutually (conditional) independent; (2) $\tilde{m}$, $\tilde{\varepsilon}_{i1}$ and $\tilde{\varepsilon}_{i2}$ have at least one absolute moment; (3) $E(\tilde{\varepsilon}_{i1}) = 0$; (4) characteristic function $\phi_{\tilde{\varepsilon}_{it}}(s) \neq 0$ for all s and $t \in \{1, 2\}$.

For condition (1), condition on $X_{i1} = X_{i2} = x$, $\tilde{m}$ is independent with $\tilde{\varepsilon}_{i1}$ and $\tilde{\varepsilon}_{i2}$ respectively. The crucial part is to show $\tilde{\varepsilon}_{i1}$ are conditionally independent with $\tilde{\varepsilon}_{i2}$. Recall that $\varepsilon_{it} = U_{it} + V_{it}$. V_{i1} is conditional independent with V_{i2} by Assumption ID 4. Based on Assumption ID 2 U_{it} can be represented as $U_{it} = \sigma_u(X_{it})\eta_{it}$ where $\sigma_u(x)$ is a bounded positive function and η_{it} are i.i.d $|N(0,1)|$ which is independent of $(\alpha_i, X_i(-t))$ where $-t$ stands for other period. So U_{it} also satisfy the conditional independence in Assumption ID 4. As U_{it} and V_{it} are conditional independent with each other by Assumption ID 3, ε_{it} is also conditional independent with $\varepsilon_{i(-t)}$, so as its demeaned version $\tilde{\varepsilon}_{it} = U_{it} + V_{it} - E(U_{it})$ since $E(U_{it})$ is just a constant conditional on $X_{i1} = X_{i2} = x$.

For Condition (2), it is trivially satisfied since $m(\cdot|X_{i1} = X_{i2} = x)$ is a bounded function, $U_{it} \sim |N(0, \sigma_u^2(X_{it}))|$, and V_{it} is conditionally symmetrically distributed with finite variance. For condition (3), obviously $E(\tilde{\varepsilon}_{i1}|X_{i1} = X_{i2} = x) = 0$ due to the normalization $\lambda_1 = 0$. For condition (4), it holds since $U_{it} \sim |N(0, \sigma_u^2(X_{it}))|$ and conditional characteristic function $\phi_{V_{it}|X_{it}}(s|X_{it} = x)$ does not vanish for all s, x and $t = 1, 2$ by Assumption ID 6.

Assumption ID 1-6 ensures that the Lemma 1 on Evdokimov and White (2012) applies to (3.11), conditional on the event $X_{i1} = X_{i2} = x$ and identifies the conditional distributions (or characteristic functions) of $m(x, \alpha)$, $\tilde{\varepsilon}_{i1}$ and $\tilde{\varepsilon}_{i2}$, given that $X_{i1} = X_{i2} = x$, for all $x \in \chi$. By the conditional independence Assumption ID 4 and its above discussion, $f_{\tilde{\varepsilon}_{it}|X_{it},\alpha_i,X_{i(-t)},\tilde{\varepsilon}_{i(-t)}}(\epsilon_t|x, \alpha, x_{(-t)}, \tilde{\varepsilon}_{(-t)}) = f_{\tilde{\varepsilon}_{it}|X_{it}}(\tilde{\epsilon}_t|x)$ for $t \in \{1, 2\}$. That is the conditional density $f_{\tilde{\varepsilon}_{it}|X_{it}}(\tilde{\epsilon}|x)$ is identified for all $x \in \chi$, $\tilde{\varepsilon} \in R$ and $t \in \{1, 2\}$, as is the conditional characteristic function $\phi_{\tilde{\varepsilon}_{it}|X_{it}}(s|x)$ for all s.

$$\phi_{\tilde{\varepsilon}_{i1}|X_{i1}}(s|x) = \exp(\int_0^s \frac{\mathbf{i}E[Y_{i1}\exp(\mathbf{i}\xi(Y_{i1} - Y_{i2}))|X_{i1} = X_{i2} = x]}{E[\exp(\mathbf{i}\xi(Y_{i1} - Y_{i2}))|X_{i1} = X_{i2} = x]} d\xi - \mathbf{i}sE(Y_{i1}|X_{i1} = X_{i2} = x)) \tag{3.12}$$

$$\phi_{\tilde{\varepsilon}_{i2}|X_{i2}}(s|x) = \frac{E[\exp(\mathbf{i}s(Y_{i1} - Y_{i2}))|X_{i1} = X_{i2} = x]}{\phi_{\tilde{\varepsilon}_{i1}|X_{i1}}(-s|x)} \tag{3.13}$$

where $\mathbf{i} = \sqrt{-1}$.[23]

2. Note that, given $X_{it} = x$, $U_{it} \sim |N(0, \sigma_u^2(x))|$ by Assumption ID 2 which is a one-parameter asymmetric distribution and V_{it} is conditionally independent of U_{it} and symmetric

[23] Note the slight notational abuse: the subscript i is an index, while the bold $\mathbf{i}$ is the imaginary number.

with $E(V_{it}|X_{it} = x) = 0$ and finite variance $\sigma_v^2(x)$ by Assumption ID 3. Therefore, given $X_{it} = x$, for the first three moments of the demeaned disturbance $\tilde{\varepsilon}_{it}$ can be written as

$$E(\tilde{\varepsilon}_{it}|X_{it} = x) = 0$$

$$E(\tilde{\varepsilon}_{it}^2|X_{it} = x) = (1 - \frac{2}{\pi})\sigma_u^2(x) + \sigma_v^2(x)$$

$$E(\tilde{\varepsilon}_{it}^3|X_{it} = x) = \frac{(4-\pi)\sqrt{2}}{\pi\sqrt{\pi}}\sigma_u^3(x)$$

3. We can also calculate the first three moments of the demeaned disturbance $\tilde{\varepsilon}_{it}$ conditional on $X_{it} = x$ by taking derivative of the conditional characteristic function and setting $s = 0$: $E(\varepsilon^k) = (-\mathbf{i})^k \frac{\partial \phi_{\varepsilon|X}(s|x)}{\partial s^k}|_{s=0}$ for $k = 1, 2, 3$.

Plugging this into equation (3.12) and rearranging,

$$-\mathbf{i} * e^0 (\int_0^s \frac{\mathbf{i}E_x[Y_{it}exp(\mathbf{i}\xi(Y_{it} - Y_{i\tau}))]}{E_x[exp(\mathbf{i}\xi(Y_{it} - Y_{i\tau}))]} d\xi - \mathbf{i}sE_x(Y_{it}))|_{s=0} = 0 \tag{3.14}$$

$$E_x[Y_{it}(Y_{it} - Y_{i\tau})] - E_x[Y_{it}]E_x[(Y_{it} - Y_{i\tau})] = (1 - \frac{2}{\pi})\sigma_u^2(x) + \sigma_v^2(x) \tag{3.15}$$

$$\begin{aligned} E_x[Y_{it}(Y_{it} - Y_{i\tau})^2] - E_x[Y_{it}]E_x[(Y_{it} - Y_{i\tau})^2] - 2E_x[(Y_{it} - Y_{i\tau})](E_x[Y_{it}(Y_{it} - Y_{i\tau})] - \\ E_x[Y_{it}]E_x[(Y_{it} - Y_{i\tau})]) = \frac{(4-\pi)\sqrt{2}}{\pi\sqrt{\pi}}\sigma_u^3(x) \end{aligned} \tag{3.16}$$

where $E_x(.) = E(.|X_{it} = X_{i\tau} = x)$ and $\tau \in \{1, 2\}$ and $\tau \neq t$.

Given $X_{it} = X_{i\tau} = x$, the two unknowns $\sigma_u^2(x)$ and $\sigma_v^2(x)$ can be uniquely solved out by equation (3.15) and equation (3.16). The third term in equation (3.16) is a constant or a function of x as $E_x(Y_{it} - Y_{i\tau}) = E_x(-\lambda_\tau)$. Varying $x \in \chi$, we can identify the nonparametric function of $\sigma_u^2(.)$. Therefore, the distribution of inefficiency, U_{it}, is identified.

Consequently, the elasticity of the mean efficiency with respect to the covariate X_{it} can be identified following Simar et al. (2017). Define the elasticity as $\xi_{\mu X} = \frac{\partial \mu_E}{\partial x} \frac{x}{\mu_E}$, note that $\mu_E(x) := E[U_{it}(x)] = \frac{\sqrt{2}\sigma_u(x)}{\sqrt{\pi}}$ for the half-normal distribution ($\sigma_u(x)$ for exponential distribution), so we can easily derive:

$$\xi_{\mu X} = \frac{\partial \sigma_u(x)}{\partial x} \frac{x}{\sigma_u(x)} = \frac{1}{3} \frac{\partial E(\tilde{\varepsilon}_{it}^3 | X_{it} = x)}{\partial x} \frac{x}{E(\tilde{\varepsilon}_{it}^3 | X_{it} = x)} \tag{3.17}$$

in which $E(\tilde{\varepsilon}_{it}^3 | X_{it} = x)$ is identified in equation (3.16).

For $U_{it} \sim Exp(\sigma_u)$ where variance is $\sigma_u^2(x)$, just replace $(1 - \frac{2}{\pi})$ and $(\frac{(4-\pi)\sqrt{2}}{\pi\sqrt{\pi}})$ by 1 and 2 in equation (3.15) and (3.16).[24], and the result follows.

□

3.B Sketch Proof of Theorem 5 and 6

Proof. We consider a cost stochastic frontier model with fixed effects α_i which is a common case in the literature. The proof can be easily extended to the random effects setting. With assumption ID 1-9 and FE 1-3, we can sketch a procedure for identifying the $m(X_{it}, \alpha_i)$ which is the production function (or profit function) in the SFA context.

(I) Step one: Identifying the conditional distribution of $\tilde{\varepsilon}_{it}$ given X_{it} exactly follows the first step of the proof of Theorem 4 . In particular, the conditional characteristic functions $\phi_{\tilde{\varepsilon}_{it}|X_{it}}(s|x)$ are identified for all $t \in \{1, 2\}$.

(II) Step two: Identifying λ_t and the distribution of $\tilde{m}(x, \alpha)|\{X_{i1} = x, X_{i2} = \bar{x}\}$ and $\alpha|\{X_{i1} = x, X_{i2} = \bar{x}\}$.

$$E[Y_{it} - Y_{i1} | X_{it} = X_{i1} = x] = E[\varepsilon_{it} + \lambda_t - \varepsilon_{i1} | X_{it} = X_{i1} = x] = \lambda_t$$

with the normalization $\lambda_1 = 0$. Conditional on the event $\{(X_{i1}, X_{i2}) = (x, \bar{x})\}$, by the conditional independence Assumption ID 4 and the normalization Assumption FE 2, we

[24] $E(u) = \sigma_u$, $Var(u) = \sigma_u^2$, $Skewness(u) = 2$.

have

$$\phi_{Y_{i1}}(s|X_{i1} = x, X_{i2} = \bar{x}) = \phi_{\tilde{m}(X_{i1},\alpha_i)}(s|X_{i1} = x, X_{i2} = \bar{x})\phi_{\tilde{\varepsilon}_{i1}}(s|X_{i1} = x),$$

$$\phi_{Y_{i2}}(s|X_{i1} = x, X_{i2} = \bar{x}) = \phi_{\alpha_i}(s|X_{i1} = x, X_{i2} = \bar{x})\phi_{\tilde{\varepsilon}_{i2}}(s|X_{i2} = \bar{x}).$$

Then,

$$\phi_{\tilde{m}(X_{i1},\alpha_i)}(s|X_{i1} = x, X_{i2} = \bar{x}) = \frac{\phi_{Y_{i1}}(s|X_{i1} = x, X_{i2} = \bar{x})}{\phi_{\tilde{\varepsilon}_{i1}}(s|X_{i1} = x)}, \tag{3.18}$$

$$\phi_{\alpha_i}(s|X_{i1} = x, X_{i2} = \bar{x}) = \frac{\phi_{Y_{i2}}(s|X_{i1} = x, X_{i2} = \bar{x})}{\phi_{\tilde{\varepsilon}_{i2}}(s|X_{i2} = \bar{x})}. \tag{3.19}$$

The left-hand side of equation (3.18) and equation (3.19) can be identified since the numerators can be identified from the data and the denominators are already identified from the previous step. The conditional CDFs of $F_{\tilde{m}(x,\alpha_i)|X_{i1},X_{i2}}(w|x,\bar{x})$ and $F_{\alpha_i|X_{i1},X_{i2}}(a|x,\bar{x})$ can be obtained following (Gil-Pelaez 1951; Evdokimov 2010):

$$F_{\tilde{m}(x,\alpha_i)|X_{i1},X_{i2}}(w|x,\bar{x}) = \frac{1}{2} - \lim_{\chi\to\infty}\int_{-\chi}^{\chi}\frac{e^{-\mathbf{i}sw}}{2\pi\mathbf{i}s}\phi_{\tilde{m}(X_{i1},\alpha_i)|X_{it},X_{i\tau}}(s|x,\bar{x})ds, t,\tau = 1,2, t \neq \tau$$

$$F_{\alpha_i|X_{i1},X_{i2}}(a|x,\bar{x}) = \frac{1}{2} - \lim_{\chi\to\infty}\int_{-\chi}^{\chi}\frac{e^{-\mathbf{i}sa}}{2\pi\mathbf{i}s}\phi_{\alpha_i|X_{it},X_{i\tau}}(s|x,\bar{x})ds, t,\tau = 1,2, t \neq \tau$$

(III) Step three: Identifying the functional $m(x,.)$.

Inverting the conditional CDF $F_{\tilde{m}(x,\alpha_i)|X_{i1},X_{i2}}(a|x,\bar{x})$, we can obtain the conditional quantile function

$$Q_{\tilde{m}(x,\alpha_i)|X_{i1},X_{i2}}(q|x,\bar{x}) = \inf\{w : F_{\tilde{m}(x,\alpha_i)|X_{i1},X_{i2}}(w|x,\bar{x}) \geq q\}, q \in (0,1)$$

According to property of quantiles, we have

$$\tilde{m}(x,a) = Q_{\tilde{m}(x,\alpha_i)|X_{i1},X_{i2}}(F_{\alpha_i|X_{i1},X_{i2}}(a|x,\bar{x})|x,\bar{x})$$

for all x and a. And $m(x,\alpha)=\tilde{m}(x,\alpha)-E(u)$ where $E(u)$ is a function of $\sigma_u(x)$.

(IV) Step four: identifying $F_{\alpha_i}(a|X_{it}=x)$.

Similar to step 2, function $\phi_{Y_{it}}(s|X_{it}=x)$ is identified from the data and hence

$$\phi_{\tilde{m}(X_{it},\alpha_i)}(s|X_{it}=x)=\phi_{Y_{it}}(s|X_{it}=x)/\phi_{\tilde{\varepsilon}_{it}}(s|X_{it}=x)$$

is identified. Hence, the CDF $F_{\tilde{m}(x,\alpha_i)|X_{it}}(w|x)$ and the quantile function $Q_{\tilde{m}(x,\alpha_i)|X_{it}}(q|x)$ can be identified. By assumption FE 1, $\tilde{m}(x,\alpha)$ is strictly increasing in α, by the property of quantiles,

$$Q_{\alpha_i|X_{it}}(q|x)=\tilde{m}^{-1}(x,Q_{\tilde{m}(X_{it},\alpha_i)|X_{it}}(q|x))$$

Finally, one can identify the conditional cumulative distribution function $F_{\alpha_i}(a|X_{it}=x)$ by inverting the quantile function $Q_{\alpha_i|X_{it}}(q|x)$. □

3.C Proof of Lemma 4

Proof. Under Assumption 10, Theorem 1 in Yin et al. (2010) holds. That is

$$\sqrt{nh}\{\hat{\sigma}_{j_1j_2}(x)-\sigma_{j_1j_2}(x)-\theta_n\}\rightarrow_d \mathcal{N}(0,f^{-1}(x)\nu_0 w_{j_1j_2}(x))$$

as $n\rightarrow\infty$, where $\theta_n=\frac{h^2\mu_2}{2}\{\ddot{\sigma}_{j_1j_2}(x)+2\dot{\sigma}_{j_1j_2}(x)\frac{\dot{f}(x)}{f(x)}\}$, $\nu_0=\int K^2(u)du$, $\mu_2=\int u^2K(u)du$, $f(x)$ is the probability density function of X evaluated at $X=x$; $w_{j_1j_2}\equiv Var(\varepsilon_{j_1j_2}(i)|X_i)$, where $\varepsilon_{j_1j_2}(i)=\{\tilde{Y}_{[i,t,\tau],j_1}-m_{j_1}(X_i)\}\{\tilde{Y}_{[i,t,\tau],j_2}-m_{j_2}(X_i)\}-\sigma_{j_1j_2}(X_i)$. $\tilde{Y}_{[i,t,\tau]}$ is defined in equation (3.7) and $m_j(x)=E(\tilde{Y}_{[i,t,\tau],j}|X_{it}=x)$.

As we consider a panel model with T=2 periods, there are two kernels for the special conditional argument $E_x(.)=E(.|X_{it}=X_{i\tau}=x)$ with univariate X_{it}. Assume the same bandwidth are chosen for the two kernels and let $(\ddot{\sigma}_{j_1j_2})_s(x)$ (or $(\dot{\sigma}_{j_1j_2})_{ss}(x)$) denote the first (or second) order derivative with respect to the sth dimensional element of X, then we have

$$\sqrt{nh^2}\{\hat{\sigma}_{j_1j_2}(x)-\sigma_{j_1j_2}(x)-\theta_n\}\rightarrow_d \mathcal{N}(0,f^{-1}(x)\nu_0 w_{j_1j_2}(x))$$

as $n\rightarrow\infty$, where $\theta_n=\frac{h^2\mu_2}{2}\{(\ddot{\sigma}_{j_1j_2})_{11}(x)+2(\dot{\sigma}_{j_1j_2})_1(x)\frac{\dot{f}_1(x)}{f(x)}\}+\frac{h^2\mu_2}{2}\{(\ddot{\sigma}_{j_1j_2})_{22}(x)+2(\dot{\sigma}_{j_1j_2})_2(x)\frac{\dot{f}_2(x)}{f(x)}\}$, $\nu_0=\int K^2(u)du$, $\mu_2=\int u^2K(u)du$, $f(x)$ is the probability density function of X evaluated at $X=x$ ($X\in R^2$ here), $\dot{f}_s(x)$ denotes the first (or second) order derivative with respect to the sth dimensional element of X; $w_{j_1j_2}\equiv Var(\varepsilon_{j_1j_2}(i)|X_i)$, where $\varepsilon_{j_1j_2}(i)=\{\tilde{Y}_{[i,t,\tau],j_1}-m_{j_1}(X_i)\}\{\tilde{Y}_{[i,t,\tau],j_2}-m_{j_2}(X_i)\}-\sigma_{j_1j_2}(X_i)$. $\tilde{Y}_{[i,t,\tau]}$ is defined in equation (3.7) and $m_j(x)=E(\tilde{Y}_{[i,t,\tau],j}|X_{it}=X_{i\tau}=x)$.

Specifically, we have

$$\sqrt{nh^2}\{\hat{\sigma}_{12}(x)-\sigma_{12}(x)-h^2\gamma_{12}\}\rightarrow_d \mathcal{N}(0,f^{-1}(x)\nu_0 w_{12}(x))$$

where $\gamma_{12}=\frac{\mu_2}{2}\{(\ddot{\sigma}_{12})_{11}(x)+2(\dot{\sigma}_{12})_1(x)\frac{\dot{f}_1(x)}{f(x)}\}+\frac{\mu_2}{2}\{(\ddot{\sigma}_{12})_{22}(x)+2(\dot{\sigma}_{12})_2(x)\frac{\dot{f}_2(x)}{f(x)}\}$ and $w_{12}=Var(\varepsilon_{12}(i)|X_{it}=X_{i\tau}=x)$, and

$$\sqrt{nh}\{\hat{\sigma}_{13}(x)-\sigma_{13}(x)-h^2\gamma_{13}\}\rightarrow_d \mathcal{N}(0,f^{-1}(x)\nu_0 w_{13}(x))$$

where $\gamma_{13}=\frac{\mu_2}{2}\{(\ddot{\sigma}_{13})_{11}(x)+2(\dot{\sigma}_{13})_1(x)\frac{\dot{f}_1(x)}{f(x)}\}+\frac{\mu_2}{2}\{(\ddot{\sigma}_{13})_{22}(x)+2(\dot{\sigma}_{13})_2(x)\frac{\dot{f}_2(x)}{f(x)}\}$ and $w_{13}=Var(\varepsilon_{13}(i)|X_{it}=X_{i\tau}=x)$.

Jointly, we have

$$\sqrt{nh}\begin{pmatrix}\hat{\sigma}_{12}(x)-\sigma_{12}(x)-h^2\gamma_{12}(x)\\ \hat{\sigma}_{13}(x)-\sigma_{13}(x)-h^2\gamma_{13}(x)\end{pmatrix}\rightarrow_d \mathcal{N}\left(0,\frac{\nu_0}{f(x)}\begin{pmatrix}\phi_{12}^2(x) & \phi_{12,13}(x)\\ \phi_{12,13}(x) & \phi_{13}^2(x)\end{pmatrix}\right)$$

where $\phi_{12}^2(x)=w_{12}(x)=Var(\varepsilon_{12}(i)|X_{it}=X_{i\tau}=x)$, $\phi_{13}^2(x)=w_{13}(x)=Var(\varepsilon_{13}(i)|X_{it}=X_{i\tau}=x)$ and accordingly $\phi_{12,13}(x)=Cov(\varepsilon_{12},\varepsilon_{13}|X_{it}=X_{i\tau}=x)$[25] with $\varepsilon_{j_1j_2}(i)=\{\tilde{Y}_{[i,t,\tau],j_1}-$

[25] The second equality holds as $m_1(X)$, $m_2(X)$, $m_3(x)$ and $\sigma_{12}(X)$, $\sigma_{13}(X)$ are constants given $X_{it}=X_{i\tau}=x$.

$m_{j_1}(X_i)\}\{\tilde{Y}_{[i,t,\tau],j_2} - m_{j_2}(X_i)\} - \sigma_{j_1 j_2}(X_i)$.

Then the conclusion follows. □

3.D Proof of Theorem 7

Proof. By Lemma (4),

$$\sqrt{nh^2}\begin{pmatrix} \hat{\sigma}_{12}(x) - \sigma_{12}(x) - h^2\gamma_{12}(x) \\ \hat{\sigma}_{13}(x) - \sigma_{13}(x) - h^2\gamma_{13}(x) \end{pmatrix} \to_d \mathcal{N}\left(0, \frac{\nu_0}{f(x)}\begin{pmatrix} \phi_{12}^2(x) & \phi_{12,13}(x) \\ \phi_{12,13}(x) & \phi_{13}^2(x) \end{pmatrix}\right)$$

where $\phi_{12}^2(x) = w_{12}(x) = Var(\varepsilon_{12}(i)|X_{it} = X_{i\tau} = x)$, $\phi_{13}^2(x) = w_{13}(x) = Var(\varepsilon_{13}(i)|X_{it} = X_{i\tau} = x)$ and accordingly $\phi_{12,13}(x) = Cov(\hat{\sigma}_{12}(x), \hat{\sigma}_{13}(x)|X_{it} = X_{i\tau} = x) = Cov(\varepsilon_{12}, \varepsilon_{13}|X_{it} = X_{i\tau} = x)$.

As $A(x) = \sigma_{12}(x)$, $B(x) = \sigma_{13}(x)$ by definition, the identification strategy simplifies to

$$\sigma_u^2(x) = c^{-2/3}\left(\sigma_{13}(x) + 2\lambda_\tau\sigma_{12}(x)\right)^{2/3}$$

$$\sigma_v^2(x) = \sigma_{12}(x) - ac^{-2/3}\left(\sigma_{13}(x) + 2\lambda_\tau\sigma_{12}(x)\right)^{2/3}$$

where a and c are constants. For example, if the inefficiency term $U_{it} \sim |N(0, \sigma_u^2(X_{it}))|$, $a = 1 - \frac{2}{\pi}$, $c = \frac{(4-\pi)\sqrt{2}}{\pi\sqrt{\pi}}$; if $U_{it} \sim Exp(b)$ where $Var(U_{it}) = \sigma_u^2(X_{it})$, $a = 1$ and $c = 2$.

By the delta method, the asymptotic distribution of $\sigma_u^2(x)$ and $\sigma_v^2(x)$ follows

$$\sqrt{nh^2}\left\{\hat{\sigma}_u^2(x) - \sigma_u^2(x) - h^2 b(x)\left(\gamma_{13}(x) + 2\lambda_\tau\gamma_{12}(x)\right)\right\} \to_d \mathcal{N}\left(0, \frac{\nu_0 b^2(x)\left(\phi_{13}^2(x) + 4\lambda_\tau\phi_{12,13} + 4\lambda_\tau^2\phi_{13}^2\right)}{f(x)}\right)$$

where $b(x) = \frac{2}{3}\left(\sigma_{13}(x) + 2\lambda_\tau\sigma_{12}(x)\right)^{-1/3}(x)c^{-2/3}$ and

$$\sqrt{nh^2}\left\{\hat{\sigma}_v^2(x)-\sigma_v^2(x)-h^2\big(\gamma_{12}(x)-ab(x)(\gamma_{13}(x)+2\lambda_\tau\gamma_{12})\big)\right\}\to_d \mathcal{N}\left(0,D'\Sigma D\right)$$

where $D=\frac{\partial\sigma_v^2(x)}{\partial\sigma_{12,13}}=\binom{1-2\lambda_\tau ab(x)}{-ab(x)}$ where $\sigma_{12,13}=\binom{\sigma_{12}}{\sigma_{13}}$, $b(x)=\frac{2}{3}\left(\sigma_{13}(x)+2\lambda_\tau\sigma_{12}(x)\right)^{-1/3}(x)c^{-2/3}$ which is bounded constant by (vii) in Assumption 10;

$$\mathbf{\Sigma}=\frac{\nu_0}{f(x)}\begin{pmatrix}\phi_{12}^2(x) & \phi_{12,13}(x)\\ \phi_{12,13}(x) & \phi_{13}^2(x)\end{pmatrix}$$

Specifically, we have

$$\begin{aligned}&\sqrt{nh^2}\left\{\hat{\sigma}_v^2(x)-\sigma_v^2(x)-h^2\big(\gamma_{12}(x)-ab(x)(\gamma_{13}(x)+2\lambda_\tau\gamma_{12})\big)\right\}\\&\to_d\mathcal{N}\left(0,\frac{\nu_0\left\{(1-2\lambda_\tau ab(x))^2\phi_{12}^2(x)-2(1-2\lambda_\tau ab(x))ab(x)\phi_{12,13}(x)+a^2b^2(x)\phi_{13}^2(x)\right\}}{f(x)}\right)\end{aligned}$$

The results follow.

□

Bibliography

[1] Aigner, D.J., Lovell, C.A.K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier production functions. Journal of Econometrics 6(1), 21?37

[2] Alvarez, A., Amsler, C., Orea, L., Schmidt, P., 2006. Interpreting and testing the scaling property in models where inefficiency depends on firm characteristics. Journal of Productivity Analysis 25, 201-212.

[3] Arellano, M., Bonhomme,S., 2009. Identifying Distributional Characteristics in Random Coefficients Panel Data Models. Working paper, CEMFI.

[4] Azomahou, T., Laisney, F., Van Nguyen, P., 2006. Economic development and CO2 emissions: A nonparametric panel approach. Journal of Public Economics, 90(6?7), 1347?1363.

[5] Bertinelli, L., Strobl, E., 2005. The Environmental Kuznets Curve semi-parametrically revisited. Economic Letters, 88, 350-357.

[6] Bonhomme, S., Robin, J.M., 2010. Generalized nonparametric deconvolution with an application to earnings dynamics. Review of Economic Studies 77(2), 491?533.

[7] Battese, G.E., Coelli, T.J., 1988. Prediction of firm-level technical efficiencies with a generalized frontier production function and panel data. Journal of Econometrics 38: 387?399.

[8] Belotti, F., Daidone, S., Ilardi, G., Atella, V., 2013. Stochastic frontier analysis using Stata. Stata Journal 13, 719-758.

[9] Belotti, F., Ilardi, G., 2018. Consistent inference in fixed-effects stochastic frontier models. Journal of Econometrics. 202: 161-177

[10] Chen, Y., Schmidt, P., Wang, H., 2014. Consistent estimation of the fixed effects stochastic frontier model. Journal of Econometrics. 181, 65-76.

[11] Cornwell, C., Schmidt, P., Sickles, R.C., 1990. Production frontiers with cross sectional and time-series variation in efficiency levels. Journal of Econometrics. 46, 185?200.

[12] Copeland, B.R., Taylor, M.S., 2004. Trade, growth, and the environment. Journal of Economic Literature, 42, 7?71.

[13] Evdokimov, K., 2010. Identification and Estimation of a Nonparametric Panel Data Model With Unobserved Heterogeneity, unpublished manuscript,Princeton University, Department of Economics.

[14] Evdokimov, K., White., H., 2012. Some extensions of a Lemma of Kotlarski. Econometric Theory. 28, 925-932.

[15] Fan, J. 1993. Local linear regression smoothers and their minimax efficiencies. Annals of Statistics. 21, 196-216.

[16] Fan, Y., Li, Q., Weersink, A., 1996. Semiparametric estimation of stochastic production frontier models. Journal of Business and Economics Statistics 14, 460–468.

[17] Florens, J.L., Simar, L., Van Keilegom, I., 2019. Estimation of the Boundary of a Variable Observed With Symmetric Error. Journal of the American Statistical Association. DOI: 10.1080/01621459.2018.1555093

[18] Greene, W., 2005a. Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. Journal of Econometrics. 126, 269?303.

[19] Greene, W., 2005b. Fixed and random effects in stochastic frontier models. Journal of Productivity Analysis 23, 7?32.

[20] Grossman, G.M., Krueger, A.B., 1991. Environmental impacts of a North American Free Trade Agreement. NBER Working Papers, 3914

[21] Hall, P., Horowitz, J., 2013. A simple bootstrap method for constructing nonparametric confidence bands for functions. Annals of Statistics. 41(1), 1892-1921.

[22] Han, C., Orea, L., Schmidt, P., 2005. Estimation of a panel data model with parametric temporal variation in individual effects. Journal of Econometrics 126, 241-267.

[23] Henderson, D., Carroll, R.J., Li, Q., 2008. Nonparametric estimation and testing of fixed effects panel data models. Journal of Econometrics. 144(1), 257-275.

[24] Ju, G., Gan, L., Li, Q., 2017. Nonparametric panel estimation of labor supply. Journal of Business & Economic Statistics. DOI: 10.1080/07350015.2017.1321546

[25] Kennan, J., Walker, J.R., 2011. The effect of expected income on individual migration decisions. Econometrica 79(1), 211?251.

[26] Kotlarski, I., 1967. On characterizing the Gamma and the Normal Distribution. Pacific Journal of Mathematics, 20(1), 69-76.

[27] Krasnokutskaya, E., 2011. Identification and estimation of auction models with unobserved heterogeneity. The Review of Economic Studies 78(1), 293?327.

[28] Kumbhakar, S. C., 1990. Production frontiers, panel data, and time-varying technical inefficiency. Journal of Econometrics. 46, 201?211.

[29] Kumbhakar, S.C., Park, B.U., Simar, L., Tsionas, E.G., 2007. Nonparametric stochastic frontiers: a local likelihood approach. Journal of Econometrics 137, 1–27.

[30] Lee, Y.H., Schmidt, P., 1993. A production frontier model with flexible temporal variation in technical efficiency. In The Measurement of Productive Efficiency: Techniques and Applications, ed. H.O. Fried, C.A. Knox Lovell, and S.S. Schmidt, 237?255. New York: Oxford University Press.

[31] Lee, Y., Mukherjee, D., Ullah, A., 2019. Nonparametric estimation of the marginal effect in fixed-effect panel data models. Journal of Multivariate Analysis 171, 53-67.

[32] Lewbel, A., 2007. A local generalized method of moments estimator. Economics Letters. 94(1), 124-128.

[33] Li, T., Perrigne, I., Vuong, Q., 2000. Conditionally independent private information in OCS wildcat auctions. Journal of Econometrics 98(1), 129?161.

[34] Li, T., Vuong, Q., 1998. Nonparametric estimation of the measurement error model using multiple indicators. Journal of Multivariate Analysis 65, 139?165.

[35] Li, Q., Racine, J.S., 2007. Nonparametric Econometrics: Theory and Practice. Princeton University Press. ISBN-13: 978-0691121611.

[36] Li, Y., Wang, N., Hong, M., Turner, N.D., Lupton, J.R., Carroll, R.J., 2007. Nonparametric estimation of correlation functions in longitudinal and spatial data with application to coloncarcinogenesis experiments. Annals of Statistics. 35, 1608-1642

[37] Parmeter, C.F., Wang, H.J., Kumbhakar, S.C., 2017. Nonparametric estimation of the determinants of inefficiency, Journal of Productivity Analysis 47, 205?221.

[38] Pitt, M.M., Lee., L.F., 1981. The measurement and sources of technical inefficiency in the Indonesian weaving industry. Journal of Development Economics 9: 43?64.

[39] Schennach, S.M., 2004. Estimation of nonlinear models with measurement error. Econometrica 72(1), 33?75.

[40] Schmidt, P., Sickles, R.C., 1984. Production frontiers and panel data. Journal of Business and Economic Statistics 2: 367?374.

[41] Simar, L., Van Keilegom, I., Zelenyuk, V., 2017. Nonparametric least squares methods for stochastic frontier models. Journal of Productivity Analysis 47: 189.

[42] Stern, D., 2017. The environmental Kuznets curve after 25 years. Journal of Bioeconomics 19, 7-28.

[43] Waldman, D.M., 1984. Properties of technical efficiency estimators in the stochastic frontier model. Journal of Econometrics. 25: 353?364.

[44] Wang, H.J., Ho, C.W., 2010. Estimating fixed-effect panel stochastic frontier models by model transformation. Journal of Econometrics. 157(2), 286-296.

[45] Wang, H.J., Schmidt, P., 2002. One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. Journal of Productivity Analysis 18, 129-144.

[46] Wang, N., 2003. Marginal nonparametric kernel regression accounting for within-subject correlation. Biometrika 90, 43?52.

[47] Wang, N., Carroll, R.J., Lin, X., 2005. Efficient semiparametric marginal estimation for longitudinal/clustered data. Journal of the American Statistical Association 100, 147?157.

[48] Wang, W.S., Schmidt, P., 2009. On the distribution of estimated technical efficiency in the stochastic frontier models. Journal of Econometrics. 148(1), 36-45.

[49] Wikstrom, D., 2015. Consistent method of moments estimation of the true fixed effects model. Economic Letters, 137, 62-69.

[50] Yao, Q., Tong, H., 1996. Asymmetric least squares regression estimation: a nonparametric approach. Journal of Nonparametric Statistics 6, 273-292.

[51] Yao, F., Zhang, F., Kumbhakar, S.C., 2018. Semiparametric smooth coefficient stochastic frontier model with panel data.Journal of Business and Economic Statistics, forthcoming.

[52] Yin, J.X, Geng, Z, Li, R.Z., Wang, H.S, 2010. Nonparametric covariance model. Statistica Sinica 20(1), 469-479.

JUN CAI

Center for Policy Research, 426 Eggers Hall, Syracuse, NY 13244

+1 (315)560-2641, jcai106@syr.edu

Education

Ph.D. Economics, Syracuse University, to be completed in June, 2020 (Expected).

Dissertation: *Nonparametric Identification and Estimation of Stochastic Frontier Models*

M.Phil. Quantitative Economics, Shanghai University of Finance and Economics, 2015.

B.S. Mechanical Engineering and Automation, Shanghai Jiao Tong University, 2011.

Specification and Interests

Fields: Econometrics, Applied Microeconomics, Labor Economics

Research Interests: Nonparametrics, Productivity Analysis, Program Evaluation

Working Papers

"Panel Nonparametric Identification and Estimation of Conditional Heteroskedastic Frontiers with An application to CO2 Emission Productivity Analysis" won an early career researcher award at the *European Workshop on Efficiency and Productivity Analysis* (EWEPA) Conference, 2019

"An Endogenous Nonlinear Panel Stochastic Frontier Model without External Instruments", Solo authored

"Detrimental Effects of Residential Job Training Programs: Evidence from Job Corps" with Alfonso Flores-Lagunes

"Nonparametric Zero-Inefficiency Stochastic Frontier Estimation" with William C. Horrace and Christopher F. Parmeter

"Density Deconvolution with Laplacian Errors and Unknown Variance" with William C. Horrace and Christopher F. Parmeter, Revise & Resubmit at *Journal of Productivity Analysis*

Research in Progress

"Regulatory Thresholds and Compliance with Audit Requirements: Bunching Evidence from Nonprofits" with Yoon-Jung Choi (draft upon request)

Conferences and Seminars

Seminars	SUNY Albany (2019), Binghamton University (2019), Syracuse University Econometrics Lunch (2019)
Conferences	Midwest Economics Association Annual Conference/SOLE (scheduled), Midwest Econometrics Group meeting (Ohio, 2019), 16th EWEPA Conference (London, 2019), New York Camp Econometrics XIV (poster, Clayton, 2019), 10th NAPW Conference (Miami, 2018)

Teaching Experience

Research Assistant for Professor Horrace, Fall 2019

Instructor, Introduction to Statistics and Econometrics (Online), Summer 2019

Evaluation rating 4.33 out of 5

Teaching Assistant, Math Camp (PhD level), Summer 2018

Teaching Assistant, Economic Statistics (undergraduate), Spring 2018, Fall 2018, Spring 2019

Teaching Assistant, Intermediate Microeconomics (undergraduate), Fall 2017, Spring 2017

Honors

2018-2019	Graduate Student Organization Travel Grant, Syracuse University
2018-2019	Economic Department Travel Grant, Syracuse University
2015-2019	Syracuse University Teaching Assistantship
2015-2019	Syracuse University Summer Stipend
2015-2018	Honor Society Membership

Skills

Computer Skills:

Proficient in Stata, Matlab (parallel computing), R and LaTeX. Familiarity with Python, PowerPoint, Mathematica and Visual C++.

Language Skills:

Fluent in *English*.

Native speaker of *Chinese*.