Empirical Economics (2020) 58:2733–2760
https://doi.org/10.1007/s00181-018-01618-9

Why fully efficient banks matter? A nonparametric stochastic frontier approach in the presence of fully efficient banks

Kien C. Tran[1] · Mike G. Tsionas[2] · Emmanuel Mamatzakis[3]

Received: 10 May 2018 / Accepted: 5 December 2018 / Published online: 19 December 2018

Abstract
A common assumption in the banking stochastic performance literature refers to the non-existence of fully efficient banks. This paper relaxes this strong assumption and proposes an alternative semiparametric zero-inefficiency stochastic frontier model. Specifically, we consider a nonparametric specification of the frontier whilst maintaining the parametric specification of the probability of fully efficient bank. We propose an iterative local maximum likelihood procedure that achieves the optimal convergence rates of both nonparametric frontier and the parameters contained in the probability of fully efficient bank. In an empirical application, we apply the proposed model and the estimation procedure to a global banking data set to derive new corrected measures of bank performance and productivity growth across the world. The results show that there is variability across regions, and the probability of fully efficient bank is mostly affected by bank-specific variables that are related to bank's risk-taking attitude, whereas country-specific variables, such as inflation, also have an effect.

Keywords Backfitting local maximum likelihood · Mixture models · Probability of fully efficient banks · Global banking

JEL Classification C13 · C14 · G20 · G21

✉ Kien C. Tran
kien.tran@uleth.ca

Mike G. Tsionas
m.tsionas@lancaster.ac.uk

Emmanuel Mamatzakis
e.mamatzakis@sussex.ac.uk

[1] Department of Economics, University of Lethbridge, 4401 University Drive W, Lethbridge, AB T1K 7L1, Canada

[2] Lancaster University Management School, Lancaster LA1 4YX, UK

[3] University of Sussex Business School, Jubilee Building, Brighton BN1 9SL, UK

1 Introduction

One of the main assumptions in stochastic frontier analysis (e.g., Aigner et al. 1977; Meeusen and van den Broeck 1977) is that all firms are inefficient, and their inefficiency is modelled with a continuous density. However, when some firms are fully efficient (a fact that cannot be excluded on a priori grounds), applying standard stochastic frontier analysis has been shown to have serious implications on the inefficiency estimates (Kumbhakar et al. 2013). Thus, to account for the possibility of fully efficient firms, Kumbhakar et al. (2013) propose a special class of two-component mixture model which they call "*zero-inefficiency stochastic frontier model*" (ZISF) that allows inefficiency to have a mass at zero with certain probability π and a continuous distribution with probability $1 - \pi$. They further extend the model to allow for the probability of fully efficient firm to depend on a set of covariates via a logit or a probit function. For a review of parametric ZISF models, see Parmeter and Kumbhakar (2014). Tran and Tsionas (2016a) suggest a semiparametric version of the ZISF model by using nonparametric formulation of the probability function and propose an iterative backfitting local maximum likelihood procedure to estimate the frontier parameters and the nonparametric function.

In this paper, we propose an alternative semiparametric ZISF model, which is different from the one suggested by Tran and Tsionas (2016a). Specifically, we consider a nonparametric specification of the frontier whilst maintaining the parametric specification (e.g., logit or probit function) of the probability of fully efficient firms. Unlike Tran and Tsionas (2016a), by maintaining the parametric assumption of the probability of fully efficient firm, there is no need for imposing local restrictions to ensure that the estimated probability lies in the interval [0, 1]. To estimate the unknown function of the frontier and the parameters of the probability of a fully efficient firm, we modify the iterative backfitting local maximum likelihood procedure developed by Tran and Tsionas (2016a), which is fairly simple to compute in practice. We also provide the necessary asymptotic properties of the modified proposed estimator. Specifically, we show that the estimator for the parameter vector in the probability of fully efficient firm is $\sqrt{n}$-consistent and follows the asymptotic normal distribution. Moreover, based upon this $\sqrt{n}$-consistent estimate, the nonparametric estimates for the unknown frontier function have the same first-order asymptotic bias and variance as the nonparametric estimates with the true values of the parameter vector in the probability function.

Next, we apply the proposed model and the estimation procedure to a global banking data set. We follow the IMF's World Economic Outlook classification to examine the productivity growth and efficiency across banks in advanced, emerging and developing countries. Our application differentiates and contributes to the literature in several ways. First, there are several papers on bank productivity (see Allen and Rai 1996; Mester 1996; Berger and Mester 2003; DeYoung and Hasan 1998; Feng and Serletis 2010; Feng and Zhang 2014; Berg et al. 1992; Alam 2001; Orea 2002; Canhoto and Dermine 2003; Barros et al. 2009; Tortosa-Ausina et al. 2008; Delis et al. 2011). However, the majority of these papers are based on the approach of data envelopment analysis (DEA). When it comes to parametric measurement of productivity through stochastic frontier analysis, the evidence is scarce (see Kumbhakar et al. 2001;

Koutsomanoli-Filippaki et al. 2009; Assaf et al. 2011). Thus, from the methodological stand point, we provide, in this paper, a novel nonparametric stochastic frontier approach to measure both bank efficiency and productivity, allowing banks to be fully efficient. Second, to the best of our knowledge, this is the first study that presents large bank productivity and efficiency at a global level, aiming to examine cross-country variability, whilst controlling the impact of various control variables, whether bank- or country-specific. Finally, we examine the effect of the credit crunch in 2008 and estimate what control variables affect the probability of having a fully efficient bank prior and ex post the crisis. This is of the utmost importance in terms of bank performance, particularly over periods of high financial distress that could lead to a shift of the whole frontier.

Overall, the results also show that there is variability across countries, and the probability of having a fully efficient bank is mostly affected by bank-specific variables that are related to bank's risk-taking, whereas country-specific variables, such as inflation, also have an effect.

The rest of the paper is structured as follows. Section 2 develops the model and the estimation procedure. Also, in this section, limited Monte Carlo simulations are performed to examine the finite sample performance of the proposed estimators. Section 3 provides an empirical analysis of global banking system. Possible extensions of the model are discussed in Sect. 4, and concluding remarks are given in Sect. 5. The proofs of the theorems are presented in "Appendix A", whilst extension of the proposed model to the fully localized (or fully nonparametric) case is given in "Appendix B".

2 The model and estimation procedure

2.1 The model

Suppose that we have a random sample $\{(Y_i,\ X_i,\ Z_i):\ \ i=1,\ldots,n\}$ from the population $(X,\ Y,\ Z)$ where $Y_i \in \mathbb{R}$ is a scalar random variable representing output of firm i, $X_i \in \mathbb{R}^d$ is a vector of continuous regressors representing inputs of firm i and $Z_i \in \mathbb{R}^r$ is a vector of continuous covariates which may or may not have common elements with X. Let ξ be a binary latent class variable, and assume that for $c=0,\ \ 1$, ξ has a conditional discrete distribution $P(\xi=0|Z=z)=\pi(z)$ and $P(\xi=1|Z=z)=1-\pi(z)$. A nonparametric version of the zero-inefficiency stochastic frontier (NP-ZISF) model proposed by Kumbhakar et al. (2013) can be written as

$$Y_i=\begin{cases} m(X_i)+v_i & \text{with probability } \ \pi(Z_i) \\ m(X_i)+v_i-u_i & \text{with probability } 1-\pi(Z_i) \end{cases}, \tag{1}$$

where $m(X_i)$ is the frontier function, $v_i|X_i=x\sim N(0,\ \sigma_v^2(x))$ and $u_i|X_i=x\sim \left|N(0,\ \sigma_u^2(x))\right|$. Conditioning on $X_i=x$, the functions $m(x)$, $\ \sigma_v^2(x)$ and $\sigma_u^2(x)$ are unknown but assumed to be smooth. Note that model (1) is special case of a two-component mixture model as well as latent class stochastic frontier models (e.g., Greene 2005) with the (technology) function $m(x)$ being restricted to be the same

for both regimes, and the composed error is $v_i - u_i(1 - I\{u_i = 0\})$ where $I\{A\}$ is an indicator function such that $I(A) = 1$ if A holds, and zero otherwise. Model (1) has several interesting features. First, when $\pi(z) = 1$, model (1) reduces to a nonparametric regression model. Second, when $\pi(z) = 0$, it becomes a nonparametric stochastic frontier model (e.g., Fan et al. 1996; Kumbhakar et al. 2007). Third, when $m(x)$ is linear in x and $\sigma_u^2(.) = \sigma_u^2$ and $\sigma_v^2(.) = \sigma_v^2$, it becomes the semiparametric ZISF model of Tran and Tsionas (2016a). Finally, when $m(x)$ is linear in x and $\pi(.) = \pi$, $\sigma_u^2(.) = \sigma_u^2$ and $\sigma_v^2(.) = \sigma_v^2$, model (1) reduces to the parametric ZISF model of Kumbhakar et al. (2013). Consequently, model (1) can be viewed as a generalization of semiparametric partially linear stochastic frontier regression models as well as the ZISF models. Thus, model (1) provides a general framework for ZISF models.

2.2 Identification

We now turn our attention to the model identification. Under the standard stochastic frontier framework regardless of parametric or nonparametric specification of the frontier, the parameter σ_u^2, the variance of u_i is identified through the moment restrictions on the composed errors $\varepsilon_i = v_i - u_i$, when u_i is left unspecified. However, when the inefficiency term u_i is modelled in a flexible manner along with parametric specification the frontier, there are possible identification problems between the intercept and the inefficiency term. For more discussion on this identification issues, see, for example, Griffin and Steel (2004). In the context of model (1), we have an additional parameter $\pi(.)$, which can be identified only if there are nonzero observations in each class. As Kumbhakar et al. (2013) and Rho and Schmidt (2015) point out, when $\sigma_u^2 \to 0$, $\pi(.)$ is not identified since the two classes become indistinguishable. Conversely, when $\pi(.) \to 1$ for a given z, σ_u^2 is not identified. In fact, when a data set contains little inefficiency, one might expect σ_u^2 and $\pi(.)$ to be imprecisely estimated, since it is difficult to identify whether little inefficiency is because $\pi(.)$ is close to 1 or σ_u^2 is close to zero. For the present discussion, we will assume that $\sigma_u^2 > 0$, and $0 < \pi(.) < 1$ so that all the parameters in model (1) are identified.

To complete the specification of the model, first given $Z = z$, we assume that $\pi(z)$ takes a form of logistic function:

$$\pi(z) = \frac{\exp(z'\alpha)}{1 + \exp(z'\alpha)}, \tag{2}$$

so as to ensure that $0 < \pi(z) < 1$. Let $f(Y, \theta(x))$ denote the conditional density of Y given $X = x$, $Z = z$ where $\theta(x) = (\alpha', \gamma(x)')'$ and $\gamma(x) = (m(x), \sigma^2(x), \lambda(x))'$. Given the distributional assumptions of v and u, the conditional pdf of Y given $X = x$ and $Z = z$ is

$$f(Y|\theta(x)) = \left(\frac{\pi(z)}{\sigma_v(x)}\right)\phi\left(\frac{Y - m(x))}{\sigma_v(x)}\right) + (1 - \pi(z)) \times \left[\frac{2}{\sigma(x)}\phi\left(\frac{Y - m(x)}{\sigma(x)}\right)\Phi\left(-(Y - m(x))\frac{\lambda(x)}{\sigma_v(x)}\right)\right], \tag{3}$$

where $\pi(z)$ is defined in (2), $\sigma^2(x) = \sigma_u^2(x)+\sigma_v^2(x)$, $\lambda(x) = \sigma_u(x)/\sigma_v(x)$, $\phi(.)$ and $\Phi(.)$ are the probability density function (pdf) and cumulative distribution function (cdf) of a standard normal variable, respectively. It follows that the conditional log-likelihood is then given by

$$L_{1n}(\alpha, \gamma(x)) = \sum_{i=1}^{n} \log f(Y_i|\alpha, \gamma(x)). \tag{4}$$

2.3 Estimation: backfitting local maximum likelihood

From (4), we note that the vector $\theta(x)$ contains both finite-dimensional and non-parametric functions which makes the direct maximization of (4) over $\theta(x)$ in an infinite-dimensional function space intractable and generally suffers from over-fitting problem. To make (4) tractable in practice, we will employ local linear regression for model (1), albeit one could consider higher orders of local polynomials. However, general order of local polynomial fitting requires additional notational complexity, but the approach is the same. In local linear fitting, we first approximate $\gamma(x)$ by taking the first-order Taylor series expansion of $\gamma(x)$ at a given set point x_0. That is, for a given x_0 and x in the neighbourhood of x_0,

$$\gamma(x) \approx \gamma_0(x_0) + \Gamma_1(x_0)(x - x_0), \tag{5}$$

where $\gamma_0(x_0)$ is a (3×1) vector and $\Gamma_1(x_0)$ is a $(3 \times d)$ matrix of the first-order derivatives.

For the kernel function, we use a multivariate product kernel which is given by:

$$K\left(h^{-1}(X_i - x_0)\right) = \prod_{j=1}^{d} k\left(h_j^{-1}\left(X_{ij} - x_{0j}\right)\right),$$

where $k(.)$ is a symmetric univariate probability density function and h_j is the bandwidth associated with X_j. Then the corresponding conditional *local* log-likelihood function for data $\{(Y_i,\ X_i,\ Z_i):\ i = 1, \ldots, n\}$ can be written as

$$L_{2n}(\alpha, \gamma_0(x_0), \Gamma_1(x_0)) = \sum_{i=1}^{n} \{\log f(Y_i; \alpha, \gamma_0(x_0) + \Gamma_1(x_0)(X_i - x_0))\} K\left(h^{-1}(X_i - x_0)\right). \tag{6}$$

Thus, the conditional local log-likelihood depends on x. Notice, however, that the global parameter α does not depend on x, and by maximizing (6), α will be estimated *locally*, and hence, it does not possess the usual parametric $\sqrt{n}$-consistency rate. To preserve the $\sqrt{n}$-consistency property of the estimator of α, we use a backfitting approach similar to Tran and Tsionas (2016a), which is motivated by Huang and Yao (2012). To do this, let $\tilde{\gamma}(x_0) = \{\tilde{m}(x_0),\ \tilde{\sigma}^2(x_0),\ \tilde{\lambda}(x_0)\}$ and $\tilde{\alpha}(x_0)$ be the maximizer of the local log-likelihood function (6); then, the initial local linear estimators of $\gamma(x)$ and $\alpha(x)$ are given by $\tilde{\gamma}(x_0) = \tilde{\gamma}_0(x_0)$ and $\tilde{\alpha} = \tilde{\alpha}(x_0)$. Given the initial estimator

$\tilde{\gamma}(x_0)$, the parameter vector α can be estimated *globally* by maximizing the following *global* log-likelihood function where we replace $\gamma(x)$ with its initial estimate $\tilde{\gamma}(x_0)$ in (4):

$$L_{3n}(\alpha, \tilde{\gamma}(x_i)) = \sum_{i=1}^{n} \log f(Y_i|\alpha, \tilde{\gamma}(x_i)). \tag{7}$$

Let $\hat{\alpha}$ be the solution of maximizing (7). In Sect. 3, we will show that under certain regularity conditions $\hat{\alpha}$ will retain its $\sqrt{n}$-consistency property. Given the estimates of $\hat{\alpha}$, $\gamma(x)$ can be estimated by maximizing the following conditional *local* log-likelihood function:

$$L_{4n}\big(\hat{\alpha}, \gamma_0(x_0), \Gamma_1(x_0)\big) = \sum_{i=1}^{n} \big\{\log f\big(Y_i; \hat{\alpha}, \gamma_0(x_0) + \Gamma_1(x_0)(X_i - x_0)\big)\big\} K\Big(h^{-1}(X_i - x_0)\Big). \tag{8}$$

Let $\hat{\Gamma}_0(x_0)$ and $\hat{\Gamma}_1(x_0)$ be the maximizer of (8); then, the local linear estimator of $\gamma(x)$ is given by $\hat{\gamma}(x) = \hat{\gamma}_0(x)$. Finally, $\hat{\alpha}$ and $\hat{\gamma}(x)$ can be further be improved by iterating until convergence. The final estimates of $\hat{\alpha}$ and $\hat{\gamma}(x)$ will be denoted as backfitting local maximum likelihood (BLML). The final estimate of $\pi(z)$ can be obtained via $\hat{\pi}(z) = \frac{\exp(z'\hat{\alpha})}{1+\exp(z'\hat{\alpha})}$.

We summarize the above estimation procedure with the following computational algorithm:

Step 1: For each z_i, $\quad i = 1, \ldots, n$, in the sample, maximize the conditional *local* log-likelihood (6) to obtain the estimate of $\tilde{\gamma}(x_i)$. Note that if the sample size n is large the maximization could be performed on a random subsample N_s, where $N_s << n$ so as to reduce the computational burden.

Step 2: From step 1, conditional on $\tilde{\gamma}(x_i)$, maximize the *global* log-likelihood function (7) to obtain $\hat{\alpha}$.

Step 3: Conditional on $\hat{\alpha}$ from step 2, maximize the conditional local log-likelihood function (8) to obtain $\hat{\gamma}(x_i)$.

Step 4: Using $\hat{\gamma}(x_i)$, repeat step 2 and then step 3 until the estimate of $\hat{\alpha}$ converges.

Note that to implement the estimation algorithm described above, specifications of the kernel function $K(.)$ as well as bandwidth matrix H are required. For the kernel function, we use a product of univariate kernel where Epanechnikov or Gaussian function is a popular choice for each kernel. As for the bandwidth selection, we follow Kumbhakar et al. (2007) and use a d-dimensional vector of bandwidth $h = (h_1, \ldots, h_d)'$ such that $h = h_b s_X n^{-1/(2+d)}$ where h_b is a scalar and $s_X = \left(s_{X_1}, \ldots, s_{X_d}\right)'$ is the vector of empirical standard deviations of the d components of X. This choice of bandwidth vector is adjusted for different scales of the regressors and different sample sizes. Then data-driven methods such as cross-validation (CV) can be used (see, for

example, Li and Racine 2007) to evaluate a grid of values for h_b. In our context, we use a likelihood-based version of CV, which is given by

$$\mathrm{CV}(h_b) = \frac{1}{n}\sum_{i=1}^{n} \log f\Big(Y_i; \hat{\alpha}^{(i)}, \hat{\gamma}^{(i)}(x_i)|x, z\Big), \tag{9}$$

where $\hat{\alpha}^{(i)}$ and $\hat{\gamma}^{(i)}(x_i)$ are the leave-one-out version of the backfitting local MLE described above. However, it is important to note that, in semiparametric modelling, under-smoothing conditions (see Theorem 1) are typically required in order to obtain $\sqrt{n}$-consistency for the global parameters. The optimal bandwidth vector $\hat{h} = \hat{h}_b s_X n^{-1/(2+d)}$ selected by CV will be in the order of $n^{-1/(2+d)}$ which does not satisfy the required under-smoothing conditions. However, a reasonable adjusted bandwidth, which suggested by Li and Liang (2008) that satisfies the under-smoothing condition, can be used, and it is given by $\tilde{h} = \hat{h} \times n^{-2/15}$. We will apply this adjusted bandwidth vector in our simulations and empirical application below.

2.4 Estimation of bank-specific inefficiency

Following the discussion of Kumbhakar et al. (2013), we can similarly consider several approaches to estimate firm-specific inefficiency. The first approach is based on the popular estimator of Jondrow et al. (1982) where under our setting, the conditional density of u given $\varepsilon(x)$ is

$$f(u|\varepsilon(x)) = \begin{cases} 0, & \text{with probability } \pi(z) \\ N_+(\mu_*(x),\, \sigma_*^2(x)), & \text{with probability } (1-\pi(z)), \end{cases}$$

where $N_+(.)$ denotes the truncated normal distribution, $\mu_*(x) = -\varepsilon(x)\sigma_u^2(x)/\sigma^2(x)$ and $\sigma_*^2(x) = \sigma_u^2(x)\sigma_v^2(x)/\sigma^2(x)$. Thus, the conditional mean of u given $\varepsilon(x) = Y - m(x)$ is:

$$E(u|\varepsilon(x)) = (1-\pi(z))\frac{\sigma(x)\lambda(x)}{1+\lambda^2(x)}\left[\frac{\phi(\lambda(x)\varepsilon(x)/\sigma(x))}{\Phi(-\lambda(x)\varepsilon(x)/\sigma(x))} - \frac{\lambda(x)\varepsilon(x)}{\sigma(x)}\right]. \tag{10}$$

A point estimator of individual inefficiency score may be obtained by replacing the unknown parameters in (7) by their estimates discussed above and $\varepsilon(x)$ by $\hat{\varepsilon}(x) = Y - \hat{m}(x)$.

Another approach is to construct the posterior estimates of inefficiency $\tilde{u}_i$. To do this, let π_i^* denote the "posterior" estimate of the probability of being fully efficient where

$$\pi_i^* = \frac{(\hat{\pi}(z)/\hat{\sigma}_v(x))\phi(\hat{\varepsilon}_i(x)/\hat{\sigma}_v(x))}{(\hat{\pi}(z)/\hat{\sigma}_v(x))(\phi(\hat{\varepsilon}_i(x)/\hat{\sigma}_v(x)) + (1-\hat{\pi}(z))(2/\hat{\sigma}(x))\,\phi(\hat{\varepsilon}_i(x)/\hat{\sigma}(x))\,\Phi(-\hat{\varepsilon}_i(x)/\hat{\sigma}(x))}. \tag{11}$$

Then the "posterior" estimate of inefficiency can be defined as $\tilde{u}_i = (1-\pi_i^*)\hat{u}_i$ where $\hat{u}_i$ is the estimate of inefficiency based on (11).

2.5 Asymptotic theory

In this section, we derive the sampling property of the proposed backfitting local MLE $\hat{\alpha}$ and $\hat{\gamma}(x) = (\hat{\beta}^{'}(x), \hat{\sigma}^2(x), \hat{\lambda}(x))^{'}$. In particular, we will show that the backfitting estimator $\hat{\alpha}$ is $\sqrt{n}$-consistent and follows an asymptotic normal distribution. In addition, we also provide the asymptotic bias and variance of the estimator $\hat{\gamma}(x)$, and show that asymptotically, it has smaller variance compared to $\tilde{\gamma}(x)$. To this end, let us define the following additional notations.

Let $\theta(x) = (\alpha^{'}, \gamma(z))^{'}$, and $\ell(\theta(x), z, y) = \log f(y|\theta(x), z)$. Define $q_\theta(\theta(x), z, y) = \frac{\partial \ell(\theta(x), z, y)}{\partial \theta}$, $q_{\theta\theta}(\theta(x), z, y) = \frac{\partial^2 \ell(\theta(x), z, y)}{\partial \theta \partial \theta^{'}}$, and the terms q_α, q_γ, $q_{\alpha\alpha}$, $q_{\alpha\gamma}$ and $q_{\gamma\gamma}$ can be defined similarly. In addition, let $\Psi(w|x) = E[q_\gamma(\theta(x), z, y)|x = w]$,

$$I_{\theta\theta}(x) = -E[q_{\theta\theta}(\theta(x), z, y)|x] = \begin{bmatrix} I_{\alpha\alpha}(x) & I_{\alpha\gamma}(x) \\ I_{\alpha\gamma}(x) & I_{\gamma\gamma}(x) \end{bmatrix},$$

where

$$I_{\alpha\alpha}(x) = -E[q_{\alpha\alpha}(\theta(x), z, y)|x]$$
$$I_{\gamma\gamma}(x) = -E\big[q_{\gamma\gamma}(\theta(x), z, y)|x\big]$$
$$I_{\alpha\gamma}(x) = -E\big[q_{\alpha\gamma}(\theta(x), z, y)|x\big].$$

Finally, let $\mu_j = \big(\int u^j K(u)\mathrm{d}u\big) I_d, \kappa_j = \big(\int u^j K^2(u)\mathrm{d}u\big) I_d$ and $|H| = h_1 h_2 \ldots h_d$. We make the following assumptions:

Assumption 1 The sample $\{(X_i, Y_i, Z_i), \quad i = 1, \ldots, n\}$ is independently and identically distributed from the joint density $f(x, y, z)$, which has continuous first derivative and positive in is support. The support for X, denoted by χ, is a compact subset of $\mathbb{R}^d$ and $f(X) > 0$ for all $X \in \chi$.

Assumption 2 The unknown functions $\gamma(x) = (m(x), \sigma^2(x), \lambda(x))^{'}$ are twice partially continuously differentiable in its argument.

Assumption 3 The matrixes $I_{\theta\theta}(x)$ and $I_{\alpha\alpha}$ are positive definite.

Assumption 4 The kernel density function $K(.)$ is symmetric, is continuous and has bounded support.

Assumption 5 For some, $\zeta < 1 - r^{-1}$, $n^{2\zeta - 1}|H| \to \infty$ and $E(X^{2r}) < \infty$.

All the above assumptions are relatively mild and have been used in the mixture models and local likelihood estimation literature. Given the above assumptions, we now ready to state our main results in the following theorems.

Theorem 1 *Under Assumptions 1–5 and in addition,* $n|H|^4 \to 0$ *and* $n|H|^2 \log(|H|^{-1}) \to \infty$, *we have*

$$\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{D} N\Big(0, A^{-1}\Sigma A^{-1}\Big),$$

where $A = E\{I_{\alpha\alpha}(x)\}$ *and* $\Sigma = \text{Var}\left\{\frac{\partial \ell(\alpha,\theta(x),z,y)}{\partial \alpha} - I_{\alpha\gamma}(x)d(x,\ y,\ z)\right\}$ *with* $d(x,\ y,\ z)$ *being the first* $(r \times r)$ *submatrix of* $I_{\theta\theta}^{-1}(x)q_\theta(\theta(x),\ z,\ y)$.

Theorem 2 *Under Assumptions* 1–5 *and in addition, as* $n \to \infty$, $|H| \to 0$, *and* $n|H| \to \infty$ *we have*

$$(n|H|)^{1/2}\left\{\hat{\gamma}(x) - \gamma(x) - B(x) + O_p\left(\sum_{i=1}^{d} h_i^2\right)\right\} \xrightarrow{D} N\left\{0,\ \kappa_0 f^{-1}(x) I_{\gamma\gamma}^{-1}\right\},$$

where $B(x) = \frac{1}{2}\mu_2|H|^2 I_{\gamma\gamma}^{-1}(z)\Psi''(x|x)$.

The proofs of Theorems 1 and 2 are given in "Appendix A". The proofs are straightforward extension of the proofs of Theorems 1 and 2 in Tran and Tsionas (2016a) to the multivariate case, and therefore, we only provide the key steps of the proofs. Note that the result from Theorem 2 shows that, as for common semiparametric model, the estimate of α has no effect on the first-order asymptotic since the rate of convergence of $\hat{\gamma}(x)$ is slower than that of $\sqrt{n}$. Consequently, it is fairly straightforward to see that $\hat{\gamma}(x)$ is more efficient than the initial estimate of $\tilde{\gamma}(x)$.

2.6 Monte Carlo simulations

In this section, we use simulations to study the finite sample performance of the proposed estimator. To this end, we consider the following data generating process (DGP) for the specification of $m(x_i)$, $\sigma_u^2(x)$ and $\sigma_v^2(x)$:

$$\begin{aligned} m(x_1,\ x_2) &= 1 + x_1 + x_2 + 0.5x_1^2 + 0.5x_2^2 + x_1x_2, \\ \sigma_v^2(x_1,\ x_2) &= 0.5\exp(0.2x_1 + 0.5x_2), \\ \sigma_u^2(x_1,\ x_2) &= 0.5\exp(0.5x_1 + 0.2x_2), \\ \pi(z) &= \exp(0.5z)\big/\left[1 + \exp(0.5z)\right]. \end{aligned}$$

The covariates $x = (x_1,\ x_2)$ and z are generated independently from a uniform distribution on $[0,\ 1]$. The random error term v is generated as $N(0,\ \sigma_v^2(x))$, and the one-sided error u is generated as $\left|N(0,\ \sigma_u^2(x)\right|$. For all our simulations, we set $\lambda = (1,\ 2.5,\ 5\}$, and let the sample sizes vary over $n = 1000$ and $n = 2000$. For each experimental design, 1000 replications are performed.

To implement our approach, we use the Gaussian kernel function and the bandwidth vector is chosen according to $\tilde{h} = \hat{h} \times n^{-2/15}$ where $\hat{h}$ is the optimal bandwidth vector based on CV approach discussed earlier in Sect. 2.3. To assess the performance of the estimators of the unknown functions $m(x_i)$, $\sigma_v^2(x)$ and $\sigma_u^2(x)$, we consider the mean average square error (MASE) for each experimental design:

$$\text{MASE} = \frac{1}{1000}\sum_{r=1}^{1000}\left\{\frac{1}{n}\sum_{j=1}^{n}\left[\hat{\xi}_r(x_j) - \xi_r(x_j)\right]^2\right\},$$

Table 1 MASE of $(\hat{m}(x), \hat{\sigma}_v^2(x), \hat{\sigma}_u^2(x))$ and MSE of $\hat{\alpha}$

	$n = 1000$	$n = 2000$
MASE		
$m(.)$	0.129	0.106
$\sigma_v^2(.)$	0.124	0.102
$\sigma_u^2(.)$	0.136	0.111
MSE		
α	0.008	0.004

Authors' estimations. Mean square error (MSE), mean average square error (MASE)

where $\hat{\xi}(.) = \hat{m}(.)$, $\hat{\sigma}_v^2(.)$ or $\hat{\sigma}_u^2(.)$, and $\{x_{ji} : j = 1, 2; \ \ i = 1, \ldots, N\}$ are the set of evenly space grid points distributed on the support of $x = (x_1, x_2)$.

To assess the performance of the estimator of the unknown parameter in the probability function, we use the mean square error (MSE):

$$\text{MSE} = \frac{1}{1000} \sum_{r=1}^{1000} (\hat{\alpha}_r - \alpha)^2.$$

We use a bootstrap procedure to estimate the standard errors and construct pointwise confidence intervals for the unknown functions as well as the unknown parameters of the probability function. To do this, for a given x_i and z_i, generate the bootstrap sample Y_i^* from a given distribution of Y specified in (1) with $\{m(.), \sigma_v^2(.), \sigma_u^2(.), \alpha\}$ being replaced by their estimates. By applying the proposed estimation procedure, for each of the bootstrap samples, we obtain the standard errors and confidence intervals.

Finally, in addition to the assessment of the above properties, we also examine the average biases, standard deviations and MSEs of technical inefficiency and returns to scale measures. For comparison purposes, we also include these results for the parametric ZISF model of Kumbhakar et al. (2013) in which the frontier is estimated by:

$$m(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Before reporting the simulation results, we notice that the wrong skewness problem did occur in our simulation exercises. In order to obtain more accurate results, we only report the results for samples with correct skewness and discard samples which have wrong skewness. Table 1 displays the simulation results for the estimated MASE of $\hat{\xi}(x_j)$ and the estimated MSE of $\hat{\alpha}$ for two different sample sizes. From Table 1, first, we observe that as the sample size increases, both the estimated MSE for parameter estimates $\hat{\alpha}$ and MASE become lower. Second, we observe that as the sample size doubles, the estimated MSE of $\hat{\alpha}$ reduces to about one half of the original values; this is consistent with the fact that the backfitting local ML estimator of $\hat{\alpha}$ is $\sqrt{n}$-consistent as predicted by Theorem 1.

Table 2 Bootstrap standard errors, standard deviations and coverage probabilities

Parameter	STD	SE (STD)	95% coverage
$n = 1000,\ h = 0.08$			
α	0.024	0.026 (0.005)	94.8
$n = 1000,\ h = 0.16$			
α	0.029	0.028 (0.006)	93.9
$n = 2000,\ h = 0.07$			
α	0.016	0.017 (0.003)	94.9
$n = 2000,\ h = 0.14$			
α	0.018	0.019 (0.004)	94.5

Estimations based on 1000 estimated standard errors using bootstrap
STD standard deviations of estimated parameters, *SE* estimated standard errors using bootstrap procedure

We next examine the accuracy of standard error estimation via a bootstrap approach. Table 2 summarizes the performance of the bootstrap approach for standard errors of the estimated functions $(\hat{m}(x),\ \hat{\sigma}_v^2(x),\ \hat{\sigma}_u^2(x))$ evaluated at $x = \{0.1, 0.2, \ldots, 0.9\}$, for two different samples and two different bandwidths which correspond to under-smoothing $\tilde{h} = \hat{h} \times n^{-2/15}$ and appropriate amount $\hat{h}$. In the table, the standard deviation of 1000 estimates is denoted by STD which can be viewed as the true standard error, whilst the average bootstrap standard errors are denoted by SE along with their standard deviations given the parentheses. The SEs are calculated as the average of 1000 estimated standard errors. The coverage probabilities for all parameters are given in the last column, and they are obtained based on the estimated standard errors. The results from Table 3 show that the suggested bootstrap procedure approximates the true standard deviations quite well, and the coverage probabilities are close to the nominal levels for almost all cases.

Note that the bootstrap procedure also allows us to compute the pointwise coverage probabilities for the probability functions. Table 3 provides the 95% coverage probabilities of $m(.)$, $\sigma_v^2(.)$ and $\sigma_u^2(.)$ for a set of evenly space grid points distributed on the support of x using under-smoothing and appropriate smoothing bandwidths. In the table, the rows labelled $m_{(\hat{\alpha})}(x)$, $\sigma^2_{v(\hat{\alpha})}(x)$ and $\sigma^2_{u(\hat{\alpha})}(x)$ provide the results using the proposed approach, whilst $m_{(\alpha)}(x)$, $\sigma^2_{v(\alpha)}(x)$ and $\sigma^2_{u(\alpha)}(x)$ gives the results assuming α is known. For most cases, the coverage probabilities are close to the nominal level. However, the coverage levels are slightly low for points 0.1, 0.2 and 0.3 when the right amount smoothing is used. This is consistent with the expectation that under-smoothing is required. Note that we also conducted experiments where over-smoothing of the bandwidth parameters is used, and the results (not reported here but available upon request) show that over-smoothing does not provide satisfactory performance in the sense that the coverage probabilities are much larger than the nominal level (i.e., ranging from 0.97 to 1.00 for almost all points of x).

Finally, Table 4 provides comparisons of the average biases and MASEs of the estimated returns to scale (RTS) and technical inefficiency (TI) of our proposed model against the (incorrectly specified) parametric ZISF proposed by Kumbhakar

Table 3 Pointwise coverage probabilities for $\{m(x), \sigma_v^2(x), \sigma_u^2(x)\}$

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$n = 1000, \quad h = 0.16$									
$m_{(\hat{\alpha})}(x)$	0.90	0.92	0.92	0.94	0.95	0.96	0.95	0.94	0.94
$m_{(\alpha)}(x)$	0.92	0.93	0.93	0.95	0.95	0.96	0.95	0.95	0.94
$\sigma^2_{v(\hat{\alpha})}(x)$	0.89	0.89	0.91	0.92	0.94	0.95	0.95	0.95	0.93
$\sigma^2_{v(\alpha)}(x)$	0.91	0.92	0.92	0.95	0.95	0.95	0.97	0.95	0.96
$\sigma^2_{u(\hat{\alpha})}(x)$	0.84	0.88	0.90	0.91	0.94	0.95	0.95	0.95	0.95
$\sigma^2_{u(\alpha)}(x)$	0.89	0.91	0.93	0.95	0.95	0.95	0.95	0.94	0.92
$n = 2000, \quad h = 0.08$									
$m_{(\hat{\alpha})}(x)$	0.91	0.93	0.94	0.94	0.94	0.94	0.95	0.95	0.95
$m_{(\alpha)}(x)$	0.93	0.94	0.95	0.95	0.95	0.96	0.96	0.95	0.95
$\sigma^2_{v(\hat{\alpha})}(x)$	0.90	0.92	0.92	0.92	0.93	0.94	0.95	0.95	0.94
$\sigma^2_{v(\alpha)}(x)$	0.92	0.94	0.93	0.95	0.95	0.95	0.95	0.95	0.95
$\sigma^2_{u(\hat{\alpha})}(x)$	0.93	0.93	0.93	0.94	0.94	0.93	0.93	0.94	0.94
$\sigma^2_{u(\alpha)}(x)$	0.93	0.94	0.95	0.95	0.95	0.95	0.96	0.95	0.95

$m_{(\hat{\alpha})}(x)$: when $\hat{\alpha}$ is estimated and $m_{(\alpha)}(x)$: when α is assumed to be known. $\sigma^2_{v(.)}(x)$ and $\sigma^2_{u(.)}(x)$ are defined similarly

Table 4 Average biases and MASEs of estimated returns to scale (RTS) and technical inefficiency (TI): NP-ZISF versus ZISF

	$n = 1000$		$n = 2000$	
	Bias	MASE	Bias	MASE
NP-ZISF				
RTS	0.009	0.021	0.0062	0.009
TI	0.020	0.029	0.013	0.018
ZISF				
RTS	0.152	0.187	0.129	0.177
TI	0.111	0.178	0.110	0.165

NP-ZISF is our proposed nonparametric zero-inefficiency stochastic frontier model, whilst ZISF is the parametric zero-inefficiency stochastic frontier model of Kumbhakar et al. (2013)

et al. (2013). The results in Table 4 clearly show that when the parametric frontier is incorrectly specified, the average biases for both returns to scale (RTS) and technical inefficiency (TI) can be quite sizable, even in large samples. In another experiment (not reported here), we examine the performance of RTS and TI where the parametric ZISF model is correctly specified, and the results show that our proposed model and method perform as well as the parametric model in terms of biases and MASEs.

3 Empirical application: global banking analysis

3.1 The data set

In this section, we provide an application of the global banking system to illustrate the usefulness and merit of our proposed model and approach. The data we use are taken from Bankscope and World Bank Indicators database which consist of only *large* banks around the world from 2000 to 2014.[1] The banks are categorized into groups of advanced, emerging and developing countries. In addition, the banks are further classified into regional groups of Advanced, EU, Europe (outside EU), Asia–Pacific and the rest of the world which includes Latin America and the Caribbean, Middle East and North Africa, Common Wealth of Independence States and sub-Saharan Africa. The classification of country groups is based on IMF World Economic Outlook, April 2015. The data include all the bank-specific financial variables (in thousand euros), as well as other specific country-level variables. Detailed descriptions of these variables are discussed below. After removing errors and inconsistencies, we obtain an unbalanced panel that includes 3679 observations for 31 advanced countries, 2165 observations for 35 emerging countries and 1461 observations for 40 developing countries.

We consider the following nonparametric cost stochastic frontier model:

$$\mathrm{TC}_{it} = \begin{cases} f(P_{it}, Y_{it}, N_{it}, t) + v_{it}, & \text{with probablity } \pi(Z_{it}) \\ f(P_{it}, Y_{it}, N_{it}, t) + v_{it} + u_{it}, & \text{with probability } 1 - \pi(Z_{it}), \end{cases} \tag{12}$$

where TC_{it} is total cost for firm (bank) i at year t, P_{it} is a vector of input prices, Y_{it} is a vector of outputs, N_{it} is a vector of quasi-fixed netputs and Z_{it} is a vector of country-specific environmental variables. Since we are using panel data, whereas our model is designed for cross-sectional data, we need to make assumptions on the temporal behaviour of inefficiency, u_{it} and noise, v_{it} in (12). For simplicity, we assume that both u_{it} and v_{it} are independent and identically distributed (albeit our model can be easily extended to accommodate heteroscedasticity as discussed in Sect. 2), and we do not enforce any specific temporal behaviour on inefficiency, which implies that a bank can be fully efficient in 1 year but not in others. For comparison purposes, we also estimated the parametric ZISF model using the standard translog form for the frontier.

Inputs, input prices and outputs are selected based on the intermediation approach and follow Koutsomanoli-Filippaki and Mamatzakis (2009) and Tanna et al. (2011). The cost function includes two outputs: net loans (which include securities) and other earning assets. The inputs are financial capital (deposits and short-term funding), labour (personnel expenses) and physical capital (fixed assets). The price of financial capital is computed as interest expenses on deposits divided by total interest-bearing borrowed funds. The price of labour is the ratio of personnel expenses to total assets,

[1] We exclude banks for which: (i) we had less than three observations over time; (ii) we had no information on the country-level control variables; (iii) we had no information of nonperforming loans. Details of construction of the data are available from the authors upon request.

and the price of physical capital is the ratio of overhead expenses (excluding personnel expenses) to fixed assets. Total bank cost is then calculated as the sum of overheads, such as personnel and administrative expenses, interest income, fees and commission expenses. Furthermore, we include equity as a quasi-fixed netput (Berger and Mester 2003; Koutsomanoli-Filippaki and Mamatzakis 2009) to capture the effect of alternative sources of funding on the bank cost structure. If such effects are ignored, then this might cause bias in measuring efficiency, in particular for banks with high equity capital. If a bank issues more equity capital that it would imply that bank management leans towards risk aversion. In addition, we include nonperforming loans (NPL) as a bad output (Hughes and Mester 2010) to proxy banks' risk-taking activities and bank fixed assets to proxy physical capital (Berger and Mester 2003). Finally, to take into account heterogeneity in bank technology, we use the logarithm of total assets as a proxy for bank's size.

The following variables are used as determinants in the probability of being fully efficient function. First, since there have been episodes of high risk during the period of our sample (i.e., bank risk), we employ the z-score as a bank-specific measure insolvency risk. This is defined as $z\text{-score} = \frac{\text{ROE} - \left(\frac{\text{Equity}}{\text{Assets}}\right)}{\sigma_{\text{ROE}}}$, where ROE is the return on equity and σ_{ROE} is the estimate of standard deviation of ROE (as in Koutsomanoli-Filippaki and Mamatzakis 2009; Delis and Staikouras 2011; Staikouras et al. 2008).

Second, to account for liquidity and capital risk, we use the ratio of liquid assets to total assets and the ratio of equity to total assets, respectively (Koutsomanoli-Filippaki and Mamatzakis 2009).[2] High capital ratio implies low capital risk, viz. equity is a buffer against financial instability.

Third, we include GDP per capita and inflation to account for the effects of different macroeconomic environments.

Finally, to capture possible size's effects in the banking industry, we include population density and market size.[3]

We estimate the model using the procedure described in Sects. 2.3, 2.4, and the results are discussed below.

3.2 Results for bank efficiency in the presence of fully efficient banks

In Table 5, we report bank efficiency for each country group. There is some variability in efficiency across the world, notably in the Middle East and sub-Saharan Africa. Surprisingly, there is also variability in bank performance as measured by bank efficiency among economies in the EU. This is surprising because of the required convergence process that economies must go through prior to their accession to the EU. Clearly, when it comes to bank efficiency, we do not observe convergence in the EU. However, in the eurozone, the variability in efficiency is less pronounced, whereas for some economies, like Greece and Slovakia, bank efficiency is quite low. Economies in Latin America and the Caribbean show a rather low level of average efficiency at

[2] Liquid assets are the sum of trading assets, loans and advances with maturity less than 3 months. Liquidity ratio reports bank's liquid assets. If the ratio takes low values, it would imply high liquidity risk.

[3] For conservation of space and given the plethora of countries in our sample, we do not report the summary statistics of all the variables used in estimation, but they are available from the authors upon request.

Table 5 Global bank efficiency in the presence of fully efficient banks

Advanced economies outside Europe					
Australia	0.82	Japan	0.82	Singapore	0.90
Canada	0.87	Korea	0.81	Switzerland	0.87
Hong Kong	0.79	New Zealand	0.89	Taiwan	0.82
Iceland	0.81	Norway	0.84	USA	0.87
Israel	0.80	San Marino	0.81		
Average: 0.84					
EU					
Austria	0.88	Germany	0.82	Poland	0.73
Belgium	0.87	Greece	0.79	Portugal	0.83
Bulgaria	0.75	Hungary	0.86	Romania	0.80
Cyprus	0.82	Ireland	0.82	Slovakia	0.77
Czech	0.81	Italy	0.85	Slovenia	0.79
Denmark	0.86	Lithuania	0.79	Latvia	0.82
Estonia	0.71	Luxembourg	0.85	Sweden	0.87
Finland	0.76	Malta	0.81	Spain	0.85
France	0.87	Netherlands	0.85	UK	0.84
Average: 0.82					
Europe, except EU					
Albania	0.73	Croatia	0.87	Serbia	0.81
Andorra	0.89	FYROM	0.82	Turkey	0.84
Bosnia and Herzegovina	0.79				
Average: 0.82					
Latin America and the Caribbean					
Argentina	0.83	Colombia	0.75	Jamaica	0.80
Bahamas	0.80	Costa Rica	0.72	Panama	0.72
Bermuda	0.82	Dominican Rep.	0.72	Peru	0.79
Bolivia	0.77	Ecuador	0.70	Trinidad & Tobago	0.73
Brazil	0.82	El Salvador	0.79	Uruguay	0.77
Chile	0.83	Honduras	0.87	Venezuela	0.75
Average: 0.78					
Asia/Pacific					
Bangladesh	0.75	Malaysia	0.83	Taiwan	0.81
Cambodia	0.71	Nepal	0.70	Thailand	0.78
China	0.72	Pakistan	0.81	Vietnam	0.72
India	0.87	Philippines	0.84		
Indonesia	0.82	Sri Lanka	0.82		

Table 5 continued

Average: 0.78					
Middle East, North Africa					
Bahrain	0.81	Kuwait	0.81	Qatar	0.71
Egypt	0.67	Lebanon	0.83	Saudi Arabia	0.80
Jordan	0.73	Oman	0.77	UAE	0.83
Average: 0.77					
Commonwealth of independent states					
Armenia	0.72	Georgia	0.72	Russian	0.80
Azerbaijan	0.71	Kazakhstan	0.77	Ukraine	0.81
Belarus	0.75	Moldova Rep.	0.70		
Average: 0.75					
Sub-Saharan Africa					
Angola	0.72	Mauritius	0.62	South Africa	0.88
Benin	0.77	Mozambique	0.65	Swaziland	0.64
Botswana	0.75	Namibia	0.72	Tanzania	0.62
Ethiopia	0.73	Nigeria	0.70	Uganda	0.71
Ghana	0.71	Senegal	0.65	Zambia	0.62
Kenya	0.67	Senegal	0.69	Zambia	0.70
Average: 0.70					

The table reports average bank efficiency for each country according to geographic region. The classification is based on IMF World Economic Outlook April 2014

0.78, as well as economies in sub-Saharan Africa at 0.70, but there is some variability. Economies in Asia/Pacific and Common Wealth of Independent States have average efficiency scores around 0.78 and 0.75, respectively.

3.3 Estimated densities of bank efficiency

One of the advantages of the proposed methodology is that it allows deriving density functions of bank-level efficiency, allowing explicitly for the possibility of fully efficiency. In the previous section, we reported that there is considerable variation in efficiency scores across the world, but also within selected group of countries, most notably the EU. There are many reasons that can explain this variability. Given the time period covered by our sample, the importance of the financial crisis in 2008 cannot be underrated. It is undoubtedly the case that bank efficiency changes over time due to the impact of the financial crisis. We document this in our analysis, by presenting the estimated densities of bank efficiency, before and after the crisis.

Top panels in Fig. 1 present the estimated densities of bank efficiency before and after the crisis for all country groups using the model and approach discussed in Sect. 2. In the bottom panels, we present the estimated densities of bank-level efficiency using the parametric approach proposed by Kumbhakar et al. (2013). The results in Fig. 1 show that fully efficient banks are present in both models (see Advanced and EU

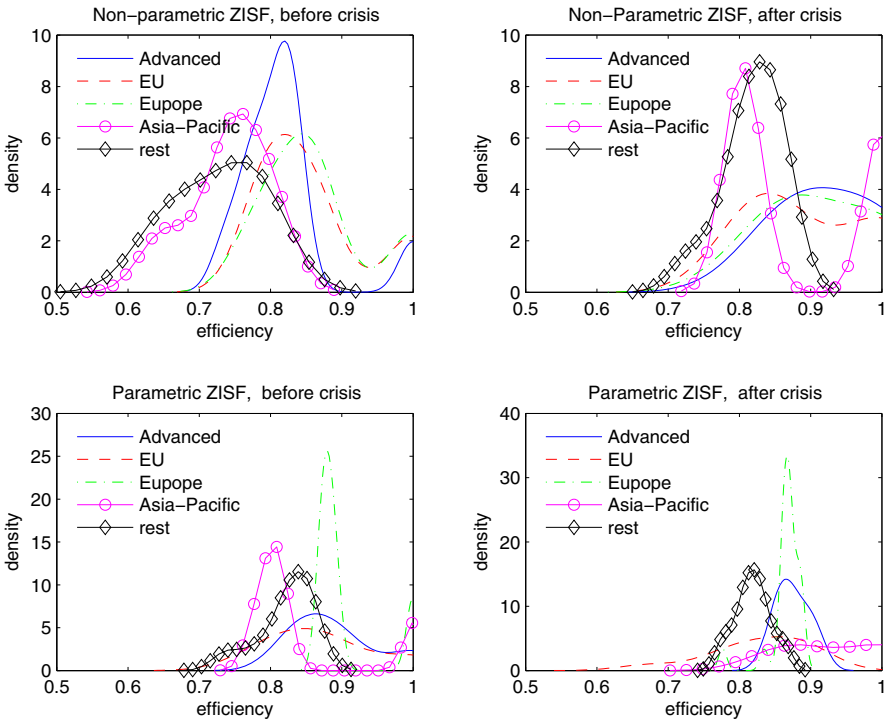

Fig. 1 Technical efficiency distributions; prior and ex post the credit crisis. *Note* To facilitate the presentation and comparison purpose, we present the estimated densities of efficiency for both nonparametric (top panel) and parametric ZISF (bottom panel) models

countries before and after the crisis). A standard stochastic frontier analysis would miss this finding. Before the crisis, efficiency scores range from 0.7 to 1 with an average of 0.83 for Advanced countries and the EU. In addition, there is evidence of a bimodal efficiency density in both groups of countries. These results are of interest as they confirm the importance of allowing for full efficiency. They also show that, despite the fact that average efficiency is around 0.83, there are also fully efficient banks in the sample. These findings change dramatically after the subprime crisis in 2008, as efficiency scores show greater variability, and range from 0.65 to 1. Also, it is worth noting that in the EU, the density after the crisis remains bimodal but exhibits higher presence of fully efficient banks, whilst for the Advance economies, the density of efficiency becomes unimodal and also displays higher frequency of fully efficient banks compared to the periods before the crisis. Thus, *our results suggest that the financial crisis was a catalyst for some banks in these two groups of countries to become fully efficient*.

The densities for Eurogroup countries display similar patterns with the density for the EU banks before and after the crisis. The densities for Asia–Pacific countries and the rest of the world display higher variations in the efficiency scores compared to the other groups before the crisis, with zero frequency of fully efficient banks. But note that after the crisis, the densities shift to the right towards higher efficiency scores for

these groups. In addition, for Asia–Pacific countries, the density after the crisis shows the presence of fully efficient banks, whilst for the rest of the world, the frequency of fully efficient bank remains zero. Interestingly, Fig. 1 shows that the crisis has not undermined bank efficiency in most countries. In fact, densities for most countries shift to higher efficiency scores. *This result is plausible in view of the fact that the subprime crisis leads to reduced liquidity and therefore adoption of more cautious cost reduction measures*.

In contrast to our proposed model, the parametric ZISF models produce the opposite results before and after the crisis. The estimated densities of bank efficiency are unimodal for both periods (with the exception of Asia–Pacific banks), although there is some small variation in the efficiency scores after the crisis compared to before the crisis period. Moreover, except for Asia–Pacific, the absence of fully efficient banks post-crisis seems to suggest that the crisis reduced bank's efficiency in these countries. Given that our approach proposes no functional form for the frontier, the contradictory results from the parametric ZISF models suggest that *misspecification of the frontier is the main factor responsible for these findings*.

To provide more information on efficiency when fully efficient banks are present, we present the density of changes in efficiency that captures the underlying dynamics around the financial crisis. The efficiency change is calculated as the difference between efficiency of bank i at time t and efficiency of bank i at time $t-1$. These results are displayed in Fig. 2. As expected, there is some variability in changes in efficiency before and after the crisis. During the second period, bank efficiency changes for most countries are slightly above zero. One of the main concerns that have been raised since the credit crunch is the low degree of alertness of financial systems prior to the crisis (Allen and Carletti 2010; Brunnermeier 2009; Covitz and Suarez 2013). Following our modelling, *our results show that signs of the crisis could have been identified well in advance*, thereby allowing for a better response to the crisis. For parametric ZISF models, there is higher variation in bank efficiency changes for most countries, and these changes tend to move towards negative values after the crisis, confirming our previous finding that the financial crisis is responsible for the reduction in bank efficiency for most of the countries considered.

In sum, the results from our analysis on efficiency and change in efficiency scores suggest that misspecification of the functional of the frontier in the parametric ZISF model is more likely the factor that provides contradictory results before and after the crisis, since our proposed nonparametric ZISF model is robust to misspecification of the frontier.

3.4 Productivity growth

Our model allows deriving not only the density of efficiency, but also densities of efficiency change. In turn, productivity growth is computed as technical change plus efficiency change. This is the first time that bank productivity growth is estimated at

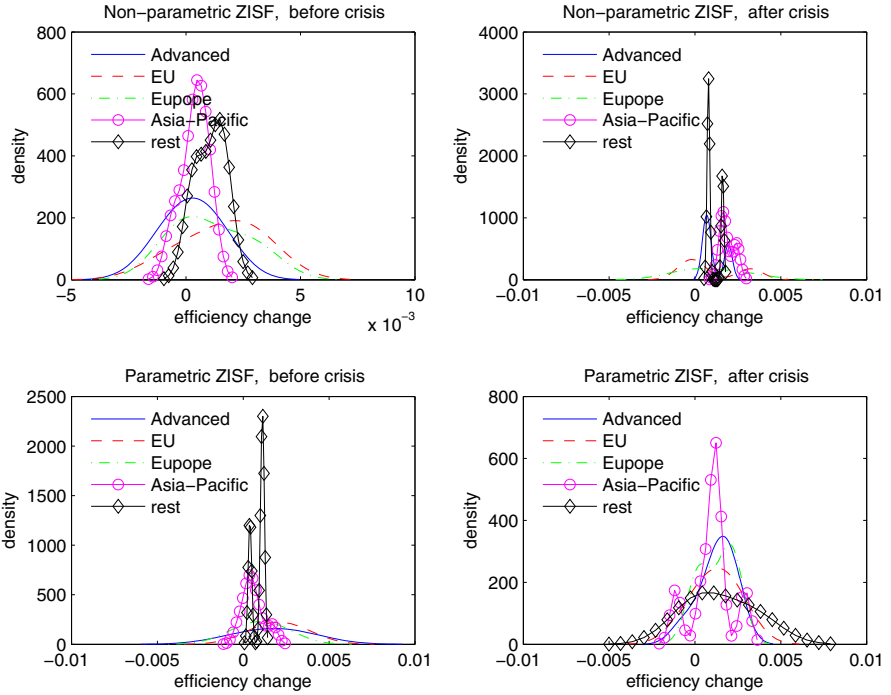

Fig. 2 Technical efficiency changes distributions; prior and ex post the credit crisis

global level, accommodating the possibility of fully efficient banks.[4] In Fig. 3, we present the density of productivity growth for the different regional classifications. Apparently, the financial crisis has been detrimental for productivity growth of large banks, across the world. Moreover, prior to the crisis, productivity growth of banks in Advanced, EU, and European countries exhibits lower kurtosis compared to post the financial crisis. It appears to converge to lower levels of productivity growth, and densities exhibit different shapes after the crisis (with the exception of the EU). *The crisis has, clearly, led to lower levels of productivity growth compared to prior the crisis, whereas variability in productivity growth is also lower*. A similar pattern is observed for bank productivity growth in Asian–Pacific countries. For the banks in the rest of the world, they have improved their productivity growth after the crisis, compared to prior the crisis.

For the parametric ZISF models, there is little variation in banking productivity growth before and after the crisis, albeit the densities of productivity growth after the

[4] Some bank productivity studies exist, but they focus mostly on a single country (e.g. Barros et al. 2009; Assaf et al. 2011) or for a certain group of countries, i.e. in EU (Koutsomanoli-Filippaki and Mamatzakis 2009; Delis et al. 2011).

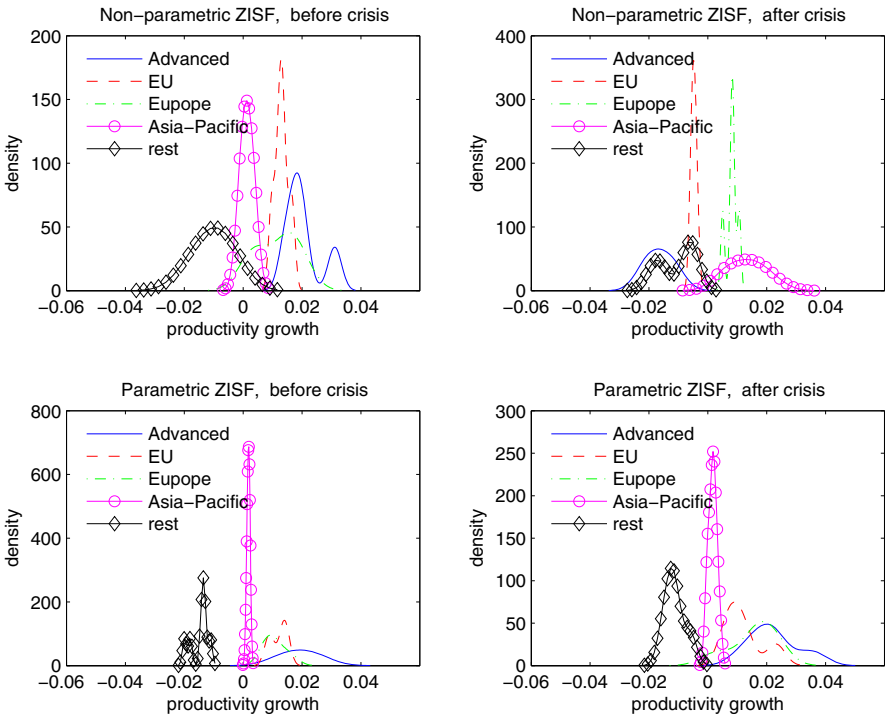

Fig. 3 Productivity growth distributions; prior and ex post the credit crisis

crisis show somewhat higher variation. These results suggest that the crisis seems to have had no effect in banking productivity growth which is counterintuitive.

3.5 Marginal effects of probability of full efficiency

Having derived bank efficiency at country and global level, it would be of interest to examine their underlying association with key bank- and country-specific covariates. Table 6 presents (average) marginal effects of probability of fully efficient banks, $\pi(z)$, with respect to the bank-specific and country-specific control variables.

The results show that not only bank-specific but also country-specific variables are important for full efficiency. All variables, apart from inflation, increase the probability of full efficiency. Note that economic as well as statistical significance is higher for bank-specific variables, and particularly for those that are related to risk, such as z-score and liquidity ratio. Improving the bank-specific risk profile appears to increase the probability of being a fully efficient bank. Alas, country-specific uncertainty, such as inflation, has a negative effect on this probability. Intuitively, the effects of country-specific variables are beyond the banks' control and they constitute one of the risk factors that affect bank performances. Therefore, we expect that these variables would make it harder for banks to become fully efficient.

Table 6 Marginal effects of probability of full efficiency, $\pi(z)$

	Adv.	EU	EUR	As-Pac.	Rest
Z-score	0.217 (0.044)	0.225 (0.002)	0.012 (0.017)	0.018 (0.005)	0.201 (0.018)
Capital ratio	0.032 (0.107)	0.025 (0.002)	0.022 (0.002)	0.044 (0.003)	0.103 (0.089)
Fees	0.032 (0.024)	0.021 (0.002)	0.035 (0.011)	0.032 (0.011)	0.044 (0.016)
Liquidity ratio	0.027 (0.005)	0.016 (0.007)	0.032 (0.012)	0.021 (0.007)	0.047 (0.011)
Securities	0.032 (0.011)	0.021 (0.005)	0.027 (0.003)	0.028 (0.009)	0.044 (0.018)
GDP per capita	0.014 (0.002)	0.017 (0.024)	0.021 (0.017)	0.015 (0.025)	0.022 (0.003)
Inflation	−0.001 (0.001)	−0.022 (0.017)	−0.028 (0.022)	−0.022 (0.001)	−0.019 (0.018)
Population density	0.0019 (0.031)	0.0021 (0.007)	0.0015 (0.014)	0.022 (0.019)	0.023 (0.030)
Market size	0.014 (0.005)	0.012 (0.003)	0.015 (0.002)	0.021 (0.003)	0.003 (0.001)

The table provides average elasticities of probability of full efficiency, $\pi(z)$, with respect to the bank- and country-specific control variables. Standard errors are reported in parentheses. Z-score = (ROE + (Equity/Assets)/(Standard Deviation of ROE); capital ratio = equity over total assets; liquidity ratio = liquid assets over total assets; fees = net fees, commission and trading income over total assets; securities = total securities over total assets. Country-specific variables: GDP per capita; inflation; population density is the number of people per square km; market size = value of total shares traded on the stock market exchange. Adv refers to Advanced countries, EU, EUR to Europe (except EU), As-Pac. to Asia Pacific

4 Extension

The proposed model in this paper can be extended in two ways. First, notice that, in our setting, one could model $\pi(z)$ nonparametrically, which makes model (1) fully nonparametric. "Appendix B" provides a brief discussion on how to estimate model (1) when all parameters are fully localized. However, as noted by Martins-Filho and Yao (2015), the main drawback of this approach is that, since all parameters are localized, the rate of convergence becomes slow when the number of conditioning variables is large (a case frequently encountered in practice), implying that the accuracy of the asymptotic approximation can be poor (i.e., the curse of dimensionality problem). Another possible extension of the model is to allow for endogenous regressors as in Tran and Tsionas (2016b) for the parametric ZISF model. However, allowing for endogeneity in nonparametric frontier can be quite complex and challenging. Finally, the proposed approach in this paper can be easily modified and extended to allow for the distribution of u_{it} to depend on a set of covariates either parametrically or nonparametrically, without affecting the proposed estimation algorithm.

5 Conclusions

This paper, first, provides an alternative semiparametric approach for estimating the ZISF model by allowing the frontier to have an unknown smooth function of explanatory variables whilst maintaining the parametric assumption on the probability of fully efficient firms. In particular, we suggest a modified version of the iterative backfitting local maximum likelihood estimator developed in Tran and Tsionas (2016a). We show

that the proposed estimator achieves the optimal convergence rates for both parameters of the probability of fully efficient bank and the nonparametric function of the frontier. We provide the asymptotic properties of the proposed estimator. The finite sample performance of the proposed estimator is examined via Monte Carlo simulations.

Next, we use the proposed method to examine productivity growth and efficiency of the global banking. Overall, our analysis demonstrates that the financial crisis has provided a valuable lesson that allowed large banks to cope, and hence increase the probability of full efficiency, particularly in Advanced countries and in the EU. To our knowledge, this is the first time that such results see the light of day, as most studies focus on the level of bank efficiency post-crisis.

Finally, in the interest of brevity, we did not consider hypothesis testing of parametric versus nonparametric frontier and/or whether all banks are fully inefficient/efficient in this paper because they are beyond the scope of this paper. However, these topics are of interest in their own right and deserve attention for future research.

Acknowledgements We would like to express our gratitude to the Editor, the Associate Editor and two anonymous referees for invaluable comments and suggestions that led to substantial improvement of the paper. We also appreciate the help of Marwan Izzeldin and the GOLCER centre at Lancaster University for their generous assistance in terms of computational support. The usual caveats apply.

Compliance with ethical standards

Conflict of interest We declare that there is no conflict of interest of any kind.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Appendix A: proofs of the theorems

Let $\tilde{\gamma}(.) = \{\tilde{m}(.), \tilde{\sigma}^2(.), \tilde{\lambda}(.)\}'$. Also let $\gamma(.) = \{m(.), \sigma^2(.), \lambda(.)\}'$ and α denote the true values.

Proof of Theorem 1 The proof of this theorem follows similarly to the proof of Theorem 1 of Tran and Tsionas (2016a) and Huang and Yao (2012). Thus, we only outline the key steps of the proof.

To derive the asymptotic properties of $\hat{\alpha}$, we first let

$$\begin{aligned}\hat{\alpha}^* &= \sqrt{n}(\hat{\alpha} - \alpha),\\ \ell(\tilde{\gamma}(X_i),\, \alpha,\, Z_i,\, Y_i) &= \log f(Y_i|\tilde{\gamma}(X_i),\, \alpha,\, Z_i)\\ \ell\Big(\tilde{\gamma}(X_i),\, \hat{\alpha} + n^{-1/2}\alpha^*,\, Z_i,\, Y_i\Big) &= \log f\Big(Y_i|\tilde{\gamma}(x_i),\, \hat{\alpha} + n^{-1/2}\alpha^*,\, Z_i\Big)\end{aligned}$$

Then $\hat{\alpha}^*$ is the maximization of

$$L_n(\alpha^*) = \sum_{i=1}^{n}\Big\{\ell\Big(\tilde{\gamma}(X_i), \alpha + n^{-1/2}\alpha^*, Z_i, Y_i\Big) - \ell(\tilde{\gamma}(X_i), \alpha, Z_i, Y_i)\Big\}. \qquad \text{(A.1)}$$

By using a Taylor series expansion and after some calculation, it yields

$$L_n(\alpha^*) = A_n\alpha^* + \frac{1}{2}\alpha^{*'} B_n\alpha^* + o_p(1), \tag{A.2}$$

where

$$A_n = n^{-1/2}\sum_{i=1}^{n}\frac{\partial\ell(\tilde{\gamma}(X_i),\,\alpha,\,Z_i,\,Y_i)}{\partial\alpha}$$

$$B_n = n^{-1}\sum_{i=1}^{n}\frac{\partial^2\ell(\tilde{\gamma}(X_i),\,\alpha,\,Z_i,\,Y_i)}{\partial\alpha\partial\alpha'}.$$

Next we evaluate the terms A_n and B_n. First, expanding A_n around $\gamma(X_i)$, we obtain

$$\begin{aligned}A_n &= n^{-1/2}\sum_{i=1}^{n}\frac{\partial\ell(\gamma(X_i),\,\alpha,\,Z_i,\,Y_i)}{\partial\alpha} + n^{-1/2}\sum_{i=1}^{n}\frac{\partial^2\ell(\gamma(X_i),\,\alpha,\,Z_i,\,Y_i)}{\partial\alpha\partial\gamma'}\big[\tilde{\gamma}(X_i)-\gamma(X_i)\big]\\ &\quad + O_p\Big(n^{-1/2}||\tilde{\gamma}(.)-\gamma(.)||_\infty^2\Big)\\ &= n^{-1/2}\sum_{i=1}^{n}\frac{\partial\ell(\gamma(X_i),\,\alpha,\,Z_i,\,Y_i)}{\partial\alpha} + D_{1n} + O_p\Big(n^{-1/2}||\tilde{\gamma}(.)-\gamma(.)||_\infty^2\Big),\end{aligned}$$

where the definition of D_{1n} should be apparent. Following Tran and Tsionas (2016a, b), it can be shown that

$$A_n = n^{-1/2}\sum_{i=1}^{n}\left\{\frac{\partial\ell(\gamma(X_i),\alpha,Z_i,Y_i)}{\partial\alpha} - I_{\alpha\gamma}(X_i)d(X_i,Y_i,Z_i)\right\} + o_p(1), \tag{A.3}$$

where $d(X,\,Y,\,Z)$ is the first $r\times r$ submatrix of $I_{\theta\theta}^{-1}(X)q_\theta(\theta(X),\,Z,\,Y)$. Similarly, for B_n, it can be shown that

$$B_n = -E[I_{\alpha\alpha}(X)] + o_p(1) = B + o_p(1). \tag{A.4}$$

Thus, from (A.2) in conjunction with (A.4), an application of quadratic approximation lemma [see, for example, Fan and Gijbels (1996, p. 210)] leads to

$$\hat{\alpha}^* = B^{-1}A_n + o_p(1) \tag{A.5}$$

if A_n is a sequence of stochastically bounded vectors. Consequently, the asymptotic normality of $\hat{\alpha}^*$ follows from that of A_n. Note that since A_n is the sum of i.i.d. random vectors, it suffices to compute the mean and covariance matrix of A_n and evoke the central limit theorem. To this end, from (A.3), we have

$$E(A_n) = n^{1/2}E\left\{\frac{\partial\ell(\gamma(X),\alpha,Z,Y)}{\partial\alpha} - I_{\alpha\gamma}(X)d(X,Y,Z)\right\}. \tag{A.6}$$

The expectation of each element of the first term on the right-hand side can be shown to be equal to 0, and further calculation shows that $E\{I_{\alpha\gamma}(X)d(X, Y, Z)\} = 0$. Thus, $E(A_n) = 0$. The variance of A_n is $\text{Var}(A_n) = \text{Var}\left\{\frac{\partial \ell(\gamma(x),\alpha,Z,y)}{\partial\alpha} - I_{\alpha\gamma}(X)d(X, Y, Z)\right\} = \Sigma$. By the central limit theorem, we obtain the desired result. □

Proof of Theorem 2 Recall that, given the estimate of $\hat{\alpha}$, $\hat{\gamma}(x)$ maximizes (7). By redefining appropriate notations:

$$\eta(x_0, X) = \gamma_0(x_0) + \Gamma_1(x_0)(X - x_0),$$
$$\gamma^* = (n|H|)^{1/2}\left\{\gamma - \gamma_0(x_0), |H|(\gamma' - \Gamma_1(x_0))\right\}',$$

then the proof follows directly from the proof of Theorem 2 of Tran and Tsionas (2016a, b). Thus, we omit it here for brevity. □

Appendix B: fully localized model

The discussion in Sect. 2 has been limited to the case where the probability of fully efficient firm $\pi(z)$ is assumed to have a logistic function. In this appendix, we extend the model to allow for a nonparametric function $\pi(z)$. We will consider two cases. In the first case, we assume that $Z = X$ and show how to estimate this model as well as discuss the asymptotic properties of the local MLE. In the second case where in general $Z \neq X$, we will briefly discuss only the estimation procedure but not the asymptotic properties since they are more complicated and beyond the scope of this paper.

Case 1: When $Z = X$

In this case, we first redefine the vector function $\theta(x) = (\pi(x), \gamma(x)')'$ and for a given set point x_0 and x in the neighbourhood of x, we approximate the function $\theta(x)$ by a linear function similar to (5),

$$\theta(x_0) \approx \theta_0(x_0) + \Theta_1(x_0)(x - x_0),$$

where $\theta_0(x_0)$ is a (4×1) vector and $\Phi_1(x_0)$ is a $(4 \times d)$ matrix of the first-order derivatives. Then the conditional local log-likelihood function is:

$$L_{5n}(\theta_0(x_0), \Theta_1(x_0)) = \sum_{i=1}^{n} \{\log f(Y_i; \theta_0(x_0) + \Theta_1(x_0)(X_i - x_0))\} K_H(X_i - x_0), \tag{B.1}$$

where the kernel function $K_H(X_i - x_0)$ is defined as before. Let $\hat{\theta}_0(x_0)$ denote the local maximizer of (B.1). Then the local MLE of $\theta(x)$ is given by $\hat{\theta}(x) = \hat{\theta}_0(x_0)$. To obtain the asymptotic property of $\hat{\theta}(.)$, we modify the following notations:

$$q_1(\theta(x),\, Y) = \partial L_5(\theta(x),\, Y)/\partial\theta, \quad q_2(\theta(x),\, Y) = \partial^2 L_5(\theta(x),\, Y)/\partial\theta\partial\theta^{'},$$

$$I(x) = -E\{q_2(\theta(X),\, Y)|X = x\}, \text{ and } \Psi(u|x) = \int_Y q_1(\theta(x),\, Y) f(Y|\theta(u))\mathrm{d}Y.$$

Assumptions

B1: The support for X, denoted by $\mathcal{X}$, is compact subset of $\mathbb{R}^d$. Furthermore, the marginal density $f(x)$ of X is twice continuously differentiable and positive for $x \in \mathcal{X}$.

B2: The unknown function $\theta(x)$ has continuous second derivatives, and in addition, $\sigma^2(x) > 0$ and $0 < \pi(x) < 1$ hold for all $x \in \mathcal{X}$.

B3: There exists a function $\mathcal{M}(y)$, with $E[\mathcal{M}(y)] < \infty$ such that for all Y, and all $\theta \in nbhd$ of $\theta(x)$, $|\partial L_5(\theta,\, Y)/\partial\theta_j\partial\theta_k\partial\theta_l| < \mathcal{M}(y)$.

B4: The following conditions hold for all i and j:$E\{|\partial L_5(\theta(x),\, Y)/\partial\theta_j|^3\} < \infty$, $E\{(\partial^2 L_5(\theta(x),\, Y)/\partial\theta_i\partial\theta_j)^2\} < \infty$.

B5: The kernel function $K(.)$ has bounded support and satisfies:

$$\left(\int K(u)\mathrm{d}u\right)I_d = 1, \quad \left(\int uK(u)\mathrm{d}u\right)I_d = 0, \quad \left(\int u^2K(u)\mathrm{d}u\right)I_d < \infty,$$

$$\left(\int K^2(u)\mathrm{d}u\right)I_d < \infty, \quad \left(\int |K(u)|^3\mathrm{d}u\right)I_d < \infty.$$

B6: $|H| \to 0$, $n|H| \to \infty$, and $n|H|^5 = O(1)$ as $n \to \infty$.

Proposition 1 *Suppose that conditions (B1)–(B6) hold. Then it follows that*

$$(n|H|)^{1/2}\left\{\hat{\theta}(x) - \theta(x) - \mathcal{B}(x) + o(|H|)^2\right\} \xrightarrow{D} N\left(0,\, \kappa_0 f^{-1}(x) I_{\theta\theta}^{-1}\right),$$

where $\mathcal{B}(x) = \frac{1}{2}\mu_2|H|^2 I_{\theta\theta}^{-1}(z)\Psi^{''}(x|x)$ *with* κ_0 *and* μ_2 *being defined as in Sect. 2.*

The proof of Proposition 1 is a straightforward extension of the proof of Theorem 2 in Huang et al. (2013) to the multivariate case, and hence, it is omitted.

Case 2: When $Z \neq X$

In this case, the local MLE is similar to case 1, albeit it is more complicated. To see this, let us once again redefine the vector function $\theta(z,\, x) = (\pi(z),\, \gamma(x)^{'})^{'}$; then, for a given set points z_0 and x_0, approximate $\theta(z,\, x)$ linearly as before. Also, define the kernel function for z as $W_A(Z_i,\, z_0) = |A|^{-1}W(A^{-1}(Z_i - z_0))$ where $W(v) = \prod_{j=1}^{r} w(v_j)$ with $w(.)$ being a univariate probability function, A being a

bandwidth matrix and $|A| = a_1 a_2 \ldots a_r$. Then the modified conditional local log-likelihood function can be written as:

$$\begin{aligned} & L_{6n}(\theta_0(z_0, x_0), \Theta_1(z_0, x_0)) \\ & = \sum_{i=1}^{n} \{\log f(Y_i; \theta_0(z_0, x_0) + \Theta_1(z_0, x_0)(Z_i - z_0)(X_i - x_0))\} \\ & \quad \times W_{A_1}(Z_i - z_0) K_{H_1}(X_i - x_0). \end{aligned} \tag{B.2}$$

Let $\theta^*(z_0, x_0)$ be the maximizer of (14) where $\theta^*(z_0, x_0) = (\pi^*(z_0, x_0), \gamma(z_0, x_0)')'$; then, the local MLE of $\theta(., .) = (\pi(., .), \gamma(., .)')'$ is given by $\tilde{\pi}(z, x) = \pi^*(., .)$ and $\tilde{\gamma}(z, x) = \gamma^*(., .)$. Note that, however, since the $\pi(z)$ do not depend on x and $\gamma(x)$ do not depend on z, the improved estimators of $\pi(z)$ and $\gamma(x)$ can be obtained using integrated backfitting approach. Thus, given the estimates $\tilde{\pi}(z, x)$ and $\tilde{\gamma}(z, x)$, the initial estimates of $\pi(z)$ and $\gamma(x)$ (up to additive constants) are given by

$$\begin{aligned} \tilde{\pi}(z) &= \int \tilde{\pi}(z, x) f_X(x) \mathrm{d}x \\ \gamma(x) &= \int \tilde{\gamma}(z, x) f_Z(z) \mathrm{d}z, \end{aligned}$$

where $f_X(x)$ and $f_Z(z)$ are marginal densities of X and Z, respectively. Now given the initial estimator of $\tilde{\pi}(z)$, for every fixed set points x_0 within the closed support of X, the improved estimator of $\gamma(x_0)$ is defined as $\hat{\gamma}(x_0) = \hat{\gamma}_0(x_0) = \hat{\gamma}_0$ where $\hat{\gamma}_0$ is the first minimizer of the following plug-in conditional local log-likelihood function:

$$\begin{aligned} & L_{7n}\left(\tilde{\pi}(z_i), \gamma_0(x_0), \Gamma_1(x_0)\right) \\ & = \sum_{i=1}^{n} \{\log f(Y_i; \tilde{\pi}(z_i), \gamma_0(x_0) + \Gamma_1(x_0)(X_i - x_0))\} K_{H_2}(X_i - x_0). \end{aligned} \tag{B.3}$$

Given the estimates of $\hat{\gamma}(x_i)$, we can obtain the improved estimator of $\pi(z_i)$, denote by $\hat{\pi}(z_0) = \hat{\pi}_0(z_0) = \hat{\pi}_0$ where $\hat{\pi}_0$ is the first maximizer the following plug-in conditional local log-likelihood function:

$$L_{8n}\big(\pi(z_0), \hat{\gamma}(x_i)\big) = \sum_{i=1}^{n} \big\{\log f(Y_i; \hat{\gamma}(x_i), \pi_0(z_0) + \Pi_1(z_0)(Z_i - z_0))\big\} W_{A_2}(Z_i - z_0). \tag{B.4}$$

References

Aigner DJ, Lovell CAK, Schmidt P (1977) Formulation and estimation of stochastic frontier production models. J Econom 6(1):21–27

Alam IMS (2001) A nonparametric approach for assessing productivity dynamics of large US banks. J Money Credit Bank 33:121–139

Allen F, Carletti E (2010) An overview of the crisis: causes, consequences and solutions. Int Rev Finance 10(1):1–27
Allen L, Rai A (1996) Operational efficiency in banking: an international comparison. J Bank Finance 20(4):655–672
Assaf AG, Barros CP, Matousek R (2011) Productivity and efficiency analysis of Shinkin banks: evidence from bootstrap and Bayesian approaches. J Bank Finance 35:331–342
Barros CP, Managi S, Matousek R (2009) Productivity growth and biased technological change: credit banks in Japan. J Int Financ Mark Inst Money 19:924–936
Berg SA, Førsund FR, Jansen ES (1992) Malmquist indices of productivity growth during the deregulation of Norwegian banking, 1980–89. Scand J Econom 94:S211–S228
Berger AN, Mester LJ (2003) Explaining the dramatic changes in performance of US banks: technological change, deregulation, and dynamic changes in competition. J Financ Intermediat 12(1):57–95
Brunnermeier K (2009) Deciphering the liquidity and credit crunch 2007–2008. J Econ Perspect 23(1):77–100
Canhoto A, Dermine J (2003) A note on banking efficiency in Portugal, new vs. old banks. J Bank Finance 27:2087–2098
Covitz N Liang, Suarez G (2013) The evolution of a financial crisis: collapse of the asset-backed commercial paper market. J Finance 68:815–848
Delis MD, Staikouras PK (2011) Supervisory effectiveness and bank risk. Rev Finance 15(3):511–543
Delis MD, Molyneux P, Pasiouras F (2011) Regulations and productivity growth in banking: evidence from transition economies. J Money Credit Bank 43(4):735–764
DeYoung R, Hasan I (1998) The performance of Denovo commercial banks: a profit efficiency approach. J Bank Finance 22:565–587
Fan J, Gijbels I (1996) Local polynomial modelling and its applications. Chapman and Hall, London
Fan Y, Li Q, Weersink A (1996) Semiparametric estimation of stochastic production frontier models. J Bus Econ Stat 14(4):460–468
Feng G, Serletis A (2010) Efficiency, technical change, and returns to scale in large US banks: panel data evidence from an output distance function satisfying theoretical regularity. J Bank Finance 34(1):127–138
Feng G, Zhang X (2014) Returns to scale at large banks in the US: a random coefficient stochastic frontier approach. J Bank Finance 39:135–145
Greene WH (2005) Reconsidering heterogeneity in panel data estimators of the stochastic frontier models. J Econom 126:269–303
Griffin J, Steel MFJ (2004) Semiparametric Bayesian inference for stochastic frontier models. J Econom 123:121–152
Huang M, Yao W (2012) Mixture of regression model with varying mixing proportions: a semiparametric approach. J Am Stat Assoc 107:711–724
Huang M, Li R, Wang S (2013) Nonparametric mixture regression models. J Am Stat Assoc 108:929–941
Hughes JP, Mester LJ (2010) Efficiency in banking: theory and evidence. In: Berger AN, Molyneux P, Wilson JOS (eds) Oxford handbook of banking. Oxford University Press, Oxford, pp 463–485
Jondrow J, Lovell CAK, Materov IS, Schmidt P (1982) On the estimation of technical inefficiency in the stochastic frontier production function model. J Econom 19(2/3):233–238
Koutsomanoli-Filippaki A, Mamatzakis E (2009) Performance and merton-type default risk of listed banks in the EU: a panel VAR approach. J Bank Finance 33:2050–2061
Koutsomanoli-Filippaki A, Margaritis D, Staikouras C (2009) Efficiency and productivity growth in the banking industry of central and eastern Europe. J Bank Finance 33:557–567
Kumbhakar SC, Lozano-Vivas A, Lovell CK, Hasan I (2001) The effects of deregulation on the performance of financial institutions: the case of Spanish savings banks. J Money Credit Bank 33:101–120
Kumbhakar SC, Park BU, Simar L, Tsionas EG (2007) Nonparametric stochastic frontiers: a local likelihood approach. J Econom 137(1):1–27
Kumbhakar SC, Parmeter CF, Tsionas EG (2013) A zero-inefficiency stochastic frontier model. J Econom 172:66–76
Li R, Liang H (2008) Variable selection in semiparametric modeling. Ann Stat 36:261–286
Li Q, Racine J (2007) Nonparametric econometrics. Princeton University Press, Princeton
Martins-Filho C, Yao F (2015) Semiparametric stochastic frontier estimation via profile likelihood. Econom Rev 34(4):413–451
Meeusen W, van den Broeck J (1977) Efficiency estimation from Cobb–Douglas production functions with composed error. Int Econ Rev 18(2):435–444

Mester L (1996) A study of bank efficiency taking into account risk-preferences. J Bank Finance 20:1025–1045
Orea L (2002) Parametric decomposition of a generalized Malmquist productivity index. J Prod Anal 18(1):5–22
Parmeter C, Kumbhakar SC (2014) Efficiency analysis: a primer on recent advances. Found Trends Econom 7(3–4):191–385
Rho S, Schmidt P (2015) Are all firms inefficient? J Prod Anal 43:327–349
Staikouras C, Mamatzakis E, Koutsomanoli-Filippaki A (2008) Cost efficiency of the banking industry in the South Eastern European region. J Int Financ Mark Inst Money 18(5):483–497
Tanna S, Pasiouras F, Nnadi M (2011) The effect of board size and composition on the efficiency of UK banks. Int J Econ Bus 18(3):441–462
Tortosa-Ausina E, Grifell-Tatjé E, Armero C, Conesa D (2008) Sensitivity analysis of efficiency and Malmquist productivity indices: an application to Spanish savings banks. Eur J Oper Res 184(3):1062–1084
Tran KC, Tsionas EG (2016a) Zero-inefficiency stochastic frontier models with varying mixing proportion: a semiparametric approach. Eur J Oper Res 249:1113–1123
Tran KC, Tsionas EG (2016b) On the estimation of zero-inefficiency stochastic frontier models with endogenous regressors. Econ Lett 147:19–22