Journal of Econometrics 184 (2015) 379–393

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Frontier estimation in the presence of measurement error with unknown variance

Alois Kneip [a], Léopold Simar [b,*], Ingrid Van Keilegom [b]

[a] Department of Economics, University of Bonn, Bonn, Germany

[b] Institute of Statistics, Biostatistics and Actuarial Sciences, Université catholique de Louvain, Louvain-la-Neuve, Belgium

ARTICLE INFO

Article history:
Received 29 August 2012
Received in revised form
6 February 2014
Accepted 26 September 2014
Available online 18 October 2014

JEL classification:
primary C13
secondary C14
C49
D24

Keywords:
Deconvolution
Stochastic frontier estimation
Nonparametric estimation
Penalized likelihood

ABSTRACT

Frontier estimation appears in productivity analysis. Firm's performance is measured by the distance between its output and an optimal production frontier. Frontier estimation becomes difficult if outputs are measured with noise and most approaches rely on restrictive parametric assumptions. This paper contributes to nonparametric approaches, with unknown frontier and unknown variance of a normally distributed error. We propose a nonparametric method identifying and estimating both quantities simultaneously. Consistency and rate of convergence of our estimators are established, and simulations verify the performance of the estimators for small samples. We illustrate our method with data on American electricity companies.

1. Introduction

Frontier estimation problems arise naturally in economics, in the context of productivity analysis. When analyzing the productivity of firms, one may compare how the firms transform their inputs W (labor, energy, capital, etc.) into an output X (the quantity of goods produced). In this context, the set of technically possible outputs is determined by a production frontier $\tau(W)$ which is the geometric locus of optimal production plans. The economic efficiency of the firm operating at the level (W_0, X_0) is then measured in terms of the distance between its production level X_0 and the boundary level $\tau(W_0)$.

Efficiency and productivity analysis have been applied in many different fields of economic activity, including industry, hospitals, transportation, schools, banks, public services, etc. Frontier models were even introduced to measure the performance of portfolios in finance, in the line with the seminal work of Markovitz (1959) using Capital Assets Pricing Models (CAPM), where W measures the risk of a portfolio and X its average return. Gattoufi et al. (2004) cite more than 1800 published articles on efficiency analysis, appearing in more than 400 journals in business and economics.

In *deterministic* frontier models it is assumed that $\tau(W)$ corresponds to the boundary of the support of X. For a random sample (W_i, X_i) one then has $P(X_i \leq \tau(W_i)) = 1$. Most nonparametric approaches are then based on the idea of enveloping the data. Farrell (1957) introduced Data Envelopment Analysis (DEA), based on either the conical hull or the convex hull of the data. Deprins et al. (1984) extended the idea to non convex sets and suggested the Free Disposal Hull (FDH) estimator, equal to the smallest free disposal set containing all the data. Statistical properties of these estimators are well known (see Banker, 1993; Korostelev et al., 1995a,b; Kneip et al., 1998; Gijbels et al., 1999; Park et al., 2000; Jeong, 2004; Jeong and Park, 2006; Kneip et al., 2008; Park et al., 2010; Daouia et al., 2010). However all these methods rely on the unrealistic assumption of deterministic frontier models that the outputs X_i are observed without noise. In the presence of noise, the envelopment methods will be biased and not consistent.

More realistic *stochastic* frontier models assume that observed outputs Y_i represent underlying, "true" outputs X_i contaminated

* Correspondence to: Institute of Statistics, Biostatistics and Actuarial Sciences, Université catholique de Louvain, Voie du Roman Pays 20, 1348 Louvain-la-Neuve, Belgium.

E-mail addresses: akneip@uni-bonn.de (A. Kneip), leopold.simar@uclouvain.be (L. Simar), ingrid.vankeilegom@uclouvain.be (I. Van Keilegom).

http://dx.doi.org/10.1016/j.jeconom.2014.09.012
0304-4076/

with some additional noise. In most of the stochastic frontier approaches developed in the econometric literature, a fully parametric model is assumed. For instance, in the pioneering work of Aigner et al. (1977) and Meeusen and van den Broek (1977), we have an iid sample of (W_i, Y_i) of inputs and outputs generated by the basic model

$$Y_i = \tau(W_i)\exp(-U_i)\exp(V_i), \tag{1}$$

where $\tau(W_i)$ is a parametric production function (e.g. a Cobb–Douglas) quantifying the optimal attainable output for a given input level W_i. Moreover, $U_i > 0$ is a positive random variable having a jump at the origin that represents the inefficiency; in parametric models, U_i has a known density depending on one or two unknown parameters (often a half normal, truncated normal or exponential). So the latent unobserved output is $X_i = \tau(W_i)\exp(-U_i)$. The noise term is $Z_i = \exp(V_i)$, where $V_i \in \mathbb{R}$ has usually a normal density with mean zero and unknown variance. Finally, U_i is supposed to be conditionally independent of V_i, given W_i. These approaches have been very popular in the econometric literature and estimation is based on standard parametric techniques, like maximum likelihood or modified least squares methods (see Greene, 2008, for a survey).

However, these approaches rely on very restrictive assumptions on both the frontier function and on the stochastic part of the model. A crucial issue is the specification of the distribution of the inefficiencies U_i. While some central limit arguments can be advocated for the Gaussian noise, there does usually not exist any information justifying particular distributional assumptions on U_i.

Recent attempts have been made to attack the problem from a non- or semi-parametric point of view. Using nonparametric techniques it is possible to avoid any parametric assumptions on the structure of $\tau(W_i)$. Important contributions in this direction are Fan et al. (1996) and Kumbhakar et al. (2007). They, however, still rely on parametric specifications for the density of U_i.

Even when assuming Gaussian noise, dropping parametric assumptions on the structure of the distribution of U_i greatly complicates the problem and enforces to develop completely new methods. Estimation of the boundary $\tau(W)$ of X then necessitates to solve a complicated, non-standard deconvolution problem.

In order to concentrate on the core of the problem, we will start by analyzing a slightly simplified version of the general model which assumes that the boundary $\tau(\cdot)$ is constant, i.e. $\tau(W) \equiv \tau$ for all W and some fixed, but unknown $\tau > 0$. With $X = \tau\exp(-U)$ and $Z = \exp(V)$ the general setup then reduces to the following situation: There are i.i.d. observations $Y_1, \ldots, Y_n$ with a density g on $\mathbb{R}_+$, generated by the model

$$Y_i = X_i \cdot Z_i, \tag{2}$$

where X_i is a latent unobserved true signal having a density f on the support $[0, \tau]$, with $f(\tau) > 0$ for some unknown $\tau > 0$, and Z_i is the noise. We assume that Z_i is independent of X_i and is log-normally distributed. More precisely, $\log Z_i \sim N(0, \sigma^2)$, where $\sigma^2 > 0$ is an unknown variance. The problem then consists in estimating τ as well as σ, when only the Y_i's are observable.

Our estimation procedure for the simplified model (2) is based on the maximization of a penalized profile likelihood. Based on local constant or local linear approximation techniques this approach is then generalized to define estimators for the stochastic frontier model (1). Precise descriptions of estimators and a corresponding asymptotic theory are given in Sections 2 and 3.

Our basic approach is similar to the setup described in Hall and Simar (2002) and Simar (2007). They propose a nonparametric approach where the noise has an unspecified symmetric density with variance σ^2 converging to zero when the sample size increases. Different from their approach we avoid the restriction of having the noise converging to zero when the sample size increases. We want to note, however, that a lognormal distribution of Z is crucial to ensure identifiability in our context, while Hall and Simar (2002) rely on unspecified error distributions.

As already mentioned above, (2) with unknown τ and σ leads to a non-standard deconvolution problem. The novelty of our approach consists in the simultaneous estimation of both parameters and the derivation of resulting convergence rates.

The problem of estimating an unknown boundary τ when the error variance σ^2 is *known*, has already been studied in a number of papers, see e.g. Goldenshluger and Tsybakov (2004), Delaigle and Gijbels (2006), Meister (2006a), or Aarts et al. (2007). Under (2) the value of σ^2 is explicitly required in order to construct either one of these estimators.

Another related problem is the deconvolution problem with unknown error variance, but without assuming the existence of a finite boundary. Butucea and Matias (2005), Meister (2006b, 2007), Butucea et al. (2008), as well as Schwarz and Van Bellegem (2010) proposed estimators under this model, and they proved (among others) the identifiability and consistency of their estimators.

The paper is organized as follows. Sections 2 and 3 describe our estimation procedure and corresponding asymptotic properties, respectively. Numerical illustrations are presented in Section 4. We first begin with a simulation study to verify the performance of the estimators in (2) for small samples. We then compare the performance of our estimator of a production frontier with the procedure proposed in Hall and Simar (2002). We also apply our procedure to analyze the production outputs of American electricity utility companies. Proofs of some core results can be found in Appendix A.

2. Estimation procedure

2.1. Estimation under the simplified model

Recall that under model (2), the latent variable X is defined on $[0, \tau]$ and its density f satisfies $f(\tau) > 0$. In addition, let g be the density of the observed variable Y. Also note that the model can equivalently be written as $Y^* = X^* + Z^*$, where $Y^* = \log Y$, $X^* = \log X$ and where $Z^* \sim N(0, \sigma^2)$ is independent of X^*, and σ^2 is unknown.

Whenever confusion is possible, we will add a subindex 0 to indicate the true quantities (e.g. $f_0, g_0, \tau_0, \ldots$ stand for the true densities f and g and the true value of τ). Let $\phi(z)$ denote the standard normal density, and recall that the density ρ_σ of a log-normal random variable with parameters $\mu = 0, \sigma^2 > 0$ is given by $\rho_\sigma(z) = \frac{1}{\sigma z}\phi(\frac{\log z}{\sigma})$ for $z > 0$. For all $y > 0$ we can then write

$$\begin{aligned} g_0(y) &= \int_0^{\tau_0} f_0(x)\frac{1}{x}\rho_{\sigma_0}\left(\frac{y}{x}\right)dx = \int_0^1 h_0(t)\frac{1}{t\tau_0}\rho_{\sigma_0}\left(\frac{y}{t\tau_0}\right)dt \\ &= \frac{1}{\sigma_0 y}\int_0^1 h_0(t)\phi\left(\frac{1}{\sigma_0}\log\frac{y}{t\tau_0}\right)dt, \end{aligned} \tag{3}$$

where

$$h_0(t) = \tau_0 f_0(t\tau_0) \quad \text{for } 0 \le t \le 1.$$

For an arbitrary density h defined on $[0, 1]$ and for arbitrary values of $\tau > 0$ and $\sigma > 0$, define

$$g_{h,\tau,\sigma}(y) = \frac{1}{\sigma y}\int_0^1 h(t)\phi\left(\frac{1}{\sigma}\log\frac{y}{t\tau}\right)dt.$$

Obviously, $g_0 \equiv g_{h_0,\tau_0,\sigma_0}$.

Since our model does not suppose the variance of the error to be known, it is important and even crucial to verify whether our model is identifiable. The answer is given in the next theorem.