

Productivity and Performance: A GMM approach*

July 11, 2022

Abstract

In this paper we propose a single step GMM approach to estimate a production function with multiple quasi-fixed and variable inputs as well as productivity and inefficiency. Our approach relies on the system consisting of the production function, the first order conditions of expected profit maximization with respect to the variable inputs, as well as general formulations for dynamic productivity and inefficiency. The estimation procedure takes care of correlations of both productivity and inefficiency with the variable inputs without using any distributional assumptions on the error terms (including inefficiency) in the system. We use Indonesian manufacturing census data to illustrate workings of our procedure.

Keywords: Productivity; Performance; Generalized Method of Moments; Compression; Inefficiency.

JEL Classification No: C40, C14, D24.

1 Introduction

Given the importance of productivity in almost every sphere of economic activities, it is important to measure it correctly. The tool that is widely used to measure it is the production function. Because an increase in productivity increases profit, every producer strive for an increase in productivity. Similarly, if a firm is inefficient its output and profit can be increased without increasing inputs and hence cost. Thus, both productivity and inefficiency are likely to be correlated with inputs which invalidates the direct estimation of the production function using OLS. In this paper we consider specification and estimation of a production function that includes multiple quasi-fixed inputs, multiple variable inputs, productivity, and inefficiency. It is well known, at least since Marschak and Andrews (1944), and from the modern literature on production functions [Olley and Pakes (hereafter OP, 1996) and further developed by Akerberg et al. (2007, 2015), Levinsohn and Petrin (hereafter LP, 2003), Petrin and Sivadasan (2013), Hu et al. (2020), and Wooldridge (2009)], that variable inputs are correlated with productivity. Similarly, from the efficiency literature, especially in stochastic frontiers, it is known that inefficiency is also correlated with the variable inputs (Amsler, Prokhorov, and Schmidt, 2016). For example, a positive productivity shock may induce the firm to purchase more variable inputs, and the same is true when inefficiency is reduced. So, productivity and inefficiency affect output in opposite ways. In this paper we break new ground in several dimensions, particularly in modeling and estimating both productivity and inefficiency.

To summarize, we propose a single-step approach to estimate the production function parameters as well as productivity and inefficiency, when they are correlated with the inputs. Our contribution relative to the existing literature, is that we do not assume that inefficiency is static or independent of the inputs. We also allow for

*We thank the editor and two anonymous referees for their detailed comments on an earlier version of the paper. We, alone, are responsible for any remaining errors and omissions.

multiple variable inputs. Since productivity and inefficiency are conceptually different they can be affected by the same or different factors. The productivity shock can increase or decrease outputs by shifting the production frontier, whereas inefficiency always decreases output. Both can influence input demand and that is why failure to include one or the other is likely to introduce endogeneity bias. Note that in Olley and Pakes (1996), LP, Akerberg et al. (2015), Gandhi et al. (2020), among many others, productivity is lumped with inefficiency. Here we show that the production frontier can be identified without any distributional assumptions on the error components. Furthermore, identification (separation) of inefficiency from productivity is done without any distributional assumptions. In sieve expansions for the first-order conditions, productivity and inefficiency, the number of polynomial terms is, usually, quite large, making econometric inferences a challenging task. To solve this problem we propose to compress the sieve regressors as in Guhaniyogi and Dunson (2015), albeit in a non-Bayesian framework. Empirically, we implement a new way of estimating the productivity gains from reducing tariffs on final goods and from reducing tariffs on intermediate inputs which is the main concern in Amiti and Konings (hereafter AK, 2007).

The rest of the paper is organized as follows. In section 2 we introduce the model with inefficiency and productivity, discuss their identification and estimation. Section 3 discusses the data and the empirical results. Section 4 concludes the paper.

2 Productivity and inefficiency model

Following the previous literature we assume a Cobb-Douglas functional form but add inefficiency to it¹ to represent the underlying technology, which is

$$Y_{i,t} = f(\mathcal{K}_{i,t,j}, X_{i,t})e^{v_{i,t}+\omega_{i,t}-u_{i,t}} = \prod_{j=1}^K \mathcal{K}_{i,t,j}^{\beta_{k,j}} \prod_{m=1}^M X_{i,t,m}^{\beta_{x,m}} e^{v_{i,t}+\omega_{i,t}-u_{i,t}}, \quad (1)$$

where $\mathcal{K}_{i,t,j}$ is the j th quasi- fixed input (predetermined at the point in time t) ($j = 1, \dots, K$), $X_{i,t,m}$ denotes the m th variable input (materials, electricity, other services, etc., $m = 1, \dots, M$), $v_{i,t}$ is an unpredictable productivity shock unknown to the firm when making variable input use decisions, $\omega_{i,t}$ is a persistent productivity shock, and $u_{i,t} \geq 0$ represents technical inefficiency. Both $\omega_{i,t}$ and $u_{i,t}$ are either known to the firm or can be predicted by the firm while making input use decisions. Using the lower case letters to denote logs of the upper case variables, i.e., $\mathbf{k}_{i,t} = [\log \mathcal{K}_{i,t,1}, \dots, \log \mathcal{K}_{i,t,K}]'$ and $\mathbf{x}_{i,t} = \log \mathbf{X}_{i,t}$, we can write the production function above as

$$y_{i,t} = \beta_0 + \beta'_k \mathbf{k}_{i,t} + \beta'_x \mathbf{x}_{i,t} + v_{i,t} + (\omega_{i,t} - u_{i,t}), \quad (2)$$

where $\mathbf{k}_{i,t} \in \mathbb{R}^K$ is the vector of logs of quasi-fixed variables, $\mathbf{x}_{i,t} \in \mathbb{R}^M$ denotes logs of the variable inputs, their respective coefficients are in $\beta_k \in \mathbb{R}^K$ and $\beta_x \in \mathbb{R}^M$.

Along with the production function in (2), we use the variable input demand functions derived from the first order conditions (FOCs) of expected profit maximization, given in (4) below. The complete system consists of the following equations:

$$y_{i,t} = \beta_o + \mathbf{k}'_{i,t} \beta_k + \mathbf{x}'_{i,t} \beta_x + v_{i,t,1} + \omega_{i,t} - e^{\ln u_{i,t}}, \quad (3)$$

$$\mathbf{x}_{i,t} = \Phi(\mathbf{k}_{i,t}, \omega_{i,t}, u_{i,t}; \gamma_o) + \mathbf{v}_{i,t,2}, \quad (4)$$

$$\omega_{i,t} = h_1(\omega_{i,t-1}, \mathbf{k}_{i,t}, \mathbf{x}_{i,t}, \mathbf{z}_{i,t}; \gamma_1) + v_{i,t,3}, \quad (5)$$

$$\ln u_{i,t} = h_2(\ln u_{i,t-1}, \mathbf{k}_{i,t}, \mathbf{x}_{i,t}, \mathbf{z}_{i,t}; \gamma_2) + v_{i,t,4}, \quad (6)$$

where $\mathbf{z}_{i,t} \in \mathbb{R}^{d_z}$ is a vector of predetermined variables (which includes lagged inputs and output and, possibly, other predetermined variables), and $\gamma_o, \gamma_1, \gamma_2$ are parameter vectors. **Note that (3) is the production function, (4) is the system of input demand functions (it can also be viewed as a system of reduced form equations for**

¹This can be easily generalized to a more flexible functional form such as the translog. One can also use a nonparametric function as in Gandhi et al. (2020).

the endogenous variable inputs), (5) is the productivity equation which describes the evolution of productivity (ω) and (6) describes the evolution of inefficiency. The specification (5) and is widely used in the productivity literature (Olley and Pakes (1996), Levinsohn and Petrin (2003), Akerberg et al. (2015), Gandhi et al. (2020), among many others). We extended it by including several other variables such as inefficiency. Similar approach is followed for specifying the evolution of inefficiency in (6) as in Tsionas (2006) but it is extended to include the input variables as well as some other exogenous variables. The specification (3) and (4) are structural but (5) and (6) are not.

We treat $\omega_{i,0}$ and $u_{i,0}$ as known and we set $\omega_{i,0} = 0$ and $\ln u_{i,0} = h_2(\ln u_{i,0}, \bar{\mathbf{k}}, \bar{\mathbf{x}}, \bar{\mathbf{z}}; \gamma_2)$, viz., starting values are obtained from the steady state of these variables where an overbar denotes sample means. Notice that in the case of $\ln u_{i,0}$ we have to solve a nonlinear equation to obtain the starting value.

The $\mathbf{z}_{i,t}$ vector includes a firm-level indicator of importing firms, denoted by FM, as well as input and output tariffs, inter alia. These variables are denoted $\tau_{i,t}$. Note that unlike AK who used the Olley and Pakes (1996) methodology to obtain TFP growth and, in turn, regress it on a number of explanatory variables (see the list in their Table 6), we are modeling and estimating the production function, productivity and inefficiency from the system above. Note that in (5) we do not have a simple Markov process. Instead ω in (5) depends on many other variables.²

In equation (3) we have the production function, in (4) we use the variable input demand functions derived from the FOCs for the variable inputs, specified as *flexible* functions of productivity and quasi-fixed inputs, in (5) we assume that productivity depends on certain variables, and in (6) we assume that inefficiency depends on the same set of variables.

The functional forms of $\Phi(\cdot)$ and $h_1(\cdot)$, $h_2(\cdot)$ will be discussed as we proceed. However, it should be clear that inefficiency and productivity depend on lagged values of output, as well as their own lags, (5) and (6) thereby providing a quite general representations of dynamics in inefficiency and productivity.

The vector function $\Phi(\cdot)$ is

$$\Phi(\cdot) = [\Phi_1(\cdot), \Phi_2(\cdot), \dots, \Phi_M(\cdot)]'$$

We do not make assumptions about the error terms in the system above. Both inefficiency and productivity are dynamic and they are treated as unobserved latent variables in (3)-(6). Moreover, the FOCs in (4) depend on both productivity and inefficiency, and inefficiency in (6) is dynamic and depends also on productivity.

Suppose $\mathbf{Z}_{i,t} = [\mathbf{k}'_{i,t}, \omega_{i,t}, u_{i,t}] \equiv [Z_{i,t,1}, \dots, Z_{i,t,D_Z}] \in \mathbb{R}^{D_Z}$. We follow the literature and use the method of sieves to approximate the elements of the vector function as follows.³

$$\begin{aligned} \Phi_1(\mathbf{k}_{i,t}, \omega_{i,t}, u_{i,t}; \gamma_o) &= \sum_{i_1=1}^P \sum_{i_2=1}^P \cdots \sum_{i_{D_Z}=1}^P Z_{i,t,1}^{i_1} \cdots Z_{i,t,D_Z} \gamma_{o1,i_1,i_2,\dots,i_{D_Z}}, \\ &\vdots \\ \Phi_2(\mathbf{k}_{i,t}, \omega_{i,t}, u_{i,t}; \gamma_o) &= \sum_{i_1=1}^P \sum_{i_2=1}^P \cdots \sum_{i_{D_Z}=1}^P Z_{i,t,1}^{i_1} \cdots Z_{i,t,D_Z} \gamma_{o2,i_1,i_2,\dots,i_{D_Z}}, \end{aligned} \tag{7}$$

where P is the common polynomial order in sieve expansions. We use the same specification for the $h(\cdot)$ functions in the inefficiency and productivity equation (assuming their lagged values are included as well)

Suppose $\mathcal{Z}_{i,t} = [\mathbf{k}'_{i,t}, \mathbf{z}_{i,t}, \mathbf{x}'_{i,t}] \equiv [\mathcal{Z}_{i,t,1}, \dots, \mathcal{Z}_{i,t,D_Z}] \in \mathbb{R}^{D_Z}$

$$h_1(\mathcal{Z}_{i,t}; \gamma_1) = \sum_{i_1=1}^Q \sum_{i_2=1}^Q \cdots \sum_{i_{D_Z}=1}^Q \mathcal{Z}_{i,t,1}^{i_1} \cdots \mathcal{Z}_{i,t,D_Z}^{i_{D_Z}} \gamma_{1,i_1,i_2,\dots,i_{D_Z}}, \tag{8}$$

²The two-stage approach used in AK is incorrect because the variables in $\tau_{i,t}$ which are assumed to affect TFP growth are excluded in the first stage of the Olley and Pakes (1996) methodology, resulting biased and inconsistent estimates of TFP growth reported in Table 6 of AK.

³Similar terms in these expansions are omitted.

$$h_2(\mathcal{Z}_{i,t}; \gamma_2) = \sum_{i_1=1}^Q \sum_{i_2=1}^Q \cdots \sum_{i_{D_Z}=1}^Q \mathcal{Z}_{i,t,1}^{i_1} \cdots \mathcal{Z}_{i,t,D_Z}^{i_{D_Z}} \gamma_{2,i_1,i_2,\dots,i_{D_Z}}, \quad (9)$$

where Q is the common polynomial order.

In both cases, the polynomial orders, viz., the P s and Q s, are selected using Andrews' (1999) BIC criterion.⁴ The moment conditions are as follows:

$$(nT)^{-1} \sum_{i=1}^n \sum_{t=1}^T \begin{bmatrix} y_{i,t} - \beta_o - \mathbf{k}'_{i,t} \boldsymbol{\beta}_k - \mathbf{x}'_{i,t} \boldsymbol{\beta}_x - \omega_{i,t} + e^{\ln u_{i,t}} \\ \mathbf{x}_{i,t} - \boldsymbol{\Phi}(\mathbf{k}_{i,t}, \omega_{i,t}, u_{i,t}; \gamma_o) \\ \omega_{i,t} - h_1(\omega_{i,t-1}, \mathbf{k}_{i,t}, \mathbf{x}_{i,t}, \mathbf{z}_{i,t}; \gamma_1) \\ \ln u_{i,t} - h_2(\ln u_{i,t-1}, \mathbf{k}_{i,t}, \mathbf{x}_{i,t}, \mathbf{z}_{i,t}; \gamma_2) \end{bmatrix} \otimes \mathbf{W}_{i,t} \quad , \quad (10)$$

where $\mathbf{W}_{i,t} \in \mathbb{R}^{d_W}$ are vector of instrumental variables. Following AK, “the set of instruments includes the 1991 levels of output tariffs, the 1991 levels of input tariffs, an interaction between the 1991 input tariffs and a firm-level indicator equal to one if the firm was an importer in all years, a dummy indicator for product codes that comprised at least one nine-digit code that was excluded from the commitment to reduce bound tariffs to 40 percent, and the proportion of skilled workers at the five-digit industry level” (AK p. 1631). We also uses lagged output, materials and labor as instruments. Since prices are available we also use them as instruments increasing the number of orthogonality conditions.

The model depicted above accounts for a principled approach to estimate parameters, efficiency and productivity by allowing for (i) dynamics of both inefficiency and productivity (as (8) and (9) depend on lagged inputs and output in a rather flexible way), (ii) endogeneity via a method of sieves, and (iii) a general process of both efficiency and productivity. Since the model involves dynamic latent variables, it can be estimated using frequentist Markov Chain Monte Carlo (MCMC) GMM estimation as in Chernozhukov and Hong (2003), (see also Gallant et al., 2017).

However, the problem is that the sieves expansions involve a large number of parameters. For example, with $M = 3$ variable inputs, $K = 3$ quasi-fixed inputs and a single output, we have $D_Z = 5$ and $D_Z = 13$. If we choose third-order polynomials ($P = 3$) then there will be 56 different parameters per equation in (7) giving a total of 168 parameters. With the 13 elements of \mathcal{Z} the total number of parameters is highly likely to exceed the number of moment conditions which is $(M + 3)d_W$ (which is 78 when $M = 3$ and $d_W = 13$).

To tackle this problem, we use a compression method in Guhaniyogi and Dunson (2015). The objective is to reduce the dimensionality of the predictors used in (7), (8), and (9). To present the methodology, let us focus on (8), say, where the predictors are denoted $\mathfrak{Z}_{i,t}$ and include all different combinations of powers of $\mathcal{Z}_{i,t}$ s. Suppose there are, in total, d such predictors in $\mathfrak{Z}_{i,t}$ (which is, therefore, a $d \times 1$ vector), and our objective is to reduce them to $\delta \ll d$. Let

$$\underset{(\delta \times 1)}{\mathbb{Z}_{i,t}} = \underset{(\delta \times d)}{\mathbb{F}} \underset{(d \times 1)}{\mathfrak{Z}_{i,t}} \quad , \quad (11)$$

where $\mathbb{Z}_{i,t}$ is the compressed version of $\mathfrak{Z}_{i,t}$, and $\mathbb{F} = [\mathbb{F}_{ij}]$ is a $\delta \times d$ matrix whose elements are

$$\mathbb{F}_{ij} = \begin{cases} -1/\sqrt{\psi}, & \text{with probability } \psi^2, \\ 0, & \text{with probability } 2\psi(1 - \psi), \\ 1/\sqrt{\psi}, & \text{with probability } (1 - \psi)^2, \end{cases} \quad i = 1, \dots, \delta, j = 1, \dots, d, \quad (12)$$

where $\psi \in (0, 1)$ is an unknown parameter. The order δ is also unknown, so Guhaniyogi and Dunson (2015) recommend to draw random values from (12) for given values of δ and ψ and choose the elements of \mathbb{F} and δ

⁴If c denotes the number of moment conditions and p the dimensionality of the parameter vector, the criterion is $J - (c - p) \ln(nT)$ where J is the J -test for over-identifying restrictions. In the test all moment conditions are used. If only some conditions are used, then c is a vector of zeros and ones (ones selecting the moment conditions used and $|c|$ is used instead in the definition of the criterion).

and ψ using the marginal likelihood criterion. We use instead the BIC criterion proposed by Andrews (1999). Write the moment conditions as

$$\mathbf{F}(\boldsymbol{\theta}; \boldsymbol{\Lambda}, \mathcal{Y}) \equiv (nT)^{-1} \sum_{i=1}^n \sum_{t=1}^T \mathbf{f}(\boldsymbol{\theta}, \boldsymbol{\Lambda}_{i,t}; \mathcal{Y}_{i,t}), \quad (13)$$

where $\boldsymbol{\Lambda}_{i,t} = (\omega_{it}, u_{it})$ denotes the collection of all latent variables for unit i and firm t , $\boldsymbol{\Lambda} = (\boldsymbol{\Lambda}_{i,t}, i = 1, \dots, n, t = 1, \dots, T)$, \mathcal{Y} denotes the data, $\mathcal{Y}_{i,t}$ denotes the data of period t of firm i , $\boldsymbol{\theta}$ is the vector of unknown parameters and $\mathbf{f}(\cdot)$ is a vector function.

Our objective is to minimize

$$Q(\boldsymbol{\theta}, \boldsymbol{\Lambda}; \mathcal{Y}) \equiv \mathbf{F}(\boldsymbol{\theta}; \boldsymbol{\Lambda}, \mathcal{Y})' \boldsymbol{\Omega} \mathbf{F}(\boldsymbol{\theta}; \boldsymbol{\Lambda}, \mathcal{Y}), \quad (14)$$

for some weighting matrix $\boldsymbol{\Omega}$. The optimal choice of $\boldsymbol{\Omega}$ depends on both parameters, $\boldsymbol{\theta}$, latent variables, and the data, and is given by

$$\boldsymbol{\Omega} = \left[(nT)^{-1} \sum_{i=1}^n \sum_{t=1}^T \mathbf{f}(\boldsymbol{\theta}; \boldsymbol{\Lambda}_{i,t}, \mathcal{Y}_{i,t}) \cdot \sum_{i=1}^n \sum_{t=1}^T \mathbf{f}(\boldsymbol{\theta}; \boldsymbol{\Lambda}_{i,t}, \mathcal{Y}_{i,t})' \right]^{-1}. \quad (15)$$

Denote the parameters as $\boldsymbol{\theta} = (\beta_o, \boldsymbol{\beta}_k, \boldsymbol{\beta}_x, \gamma_o, \gamma_1, \gamma_2)$. We follow Gallant et al. (2017) to define a pseudo-posterior and then apply MCMC for statistical inferences. Let $\mathbf{g}(\boldsymbol{\theta}, \boldsymbol{\Lambda}, \mathcal{Y}) = \frac{1}{\sqrt{nT}} \sum_{i=1}^n \sum_{t=1}^T \mathbf{f}(\boldsymbol{\theta}; \boldsymbol{\Lambda}_{i,t}, \mathcal{Y}_{i,t})$ with weighting matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}, \boldsymbol{\Lambda}, \mathcal{Y}) = \frac{1}{T} \sum_{i=1}^n \sum_{t=1}^T [\tilde{\mathbf{f}}(\boldsymbol{\theta}; \boldsymbol{\Lambda}_{i,t}, \mathcal{Y}_{i,t})][\tilde{\mathbf{f}}(\boldsymbol{\theta}; \boldsymbol{\Lambda}_{i,t}, \mathcal{Y}_{i,t})]'$, where $\tilde{\mathbf{f}}(\boldsymbol{\theta}; \boldsymbol{\Lambda}_{i,t}, \mathcal{Y}_{i,t}) = \mathbf{f}(\boldsymbol{\theta}; \boldsymbol{\Lambda}_{i,t}, \mathcal{Y}_{i,t}) - \frac{1}{\sqrt{nT}} \mathbf{F}(\boldsymbol{\theta}; \boldsymbol{\Lambda}, \mathcal{Y})$. In turn, one can define a pseudo-posterior consistent with the moment conditions as follows.

$$p(\boldsymbol{\theta}, \boldsymbol{\Lambda} | \mathcal{Y}) \propto |\boldsymbol{\Sigma}(\boldsymbol{\theta}, \boldsymbol{\Lambda}, \mathcal{Y})|^{-1/2} \exp \left\{ -\frac{1}{2} \mathbf{F}(\boldsymbol{\theta}; \boldsymbol{\Lambda}, \mathcal{Y})' \boldsymbol{\Sigma}(\boldsymbol{\theta}, \boldsymbol{\Lambda}, \mathcal{Y})^{-1} \mathbf{F}(\boldsymbol{\theta}; \boldsymbol{\Lambda}, \mathcal{Y}) \right\}. \quad (16)$$

We then set up a MCMC scheme by drawing the latent variables $\boldsymbol{\Lambda}$ conditional on $\boldsymbol{\theta}$ and the data, and then drawing $\boldsymbol{\theta} | \boldsymbol{\Lambda}, \mathcal{Y}$. The second step is easy and can be performed using any standard MCMC algorithm, for example the Girolami and Calderhead (2011) Langevin diffusion. The first step is more involved and requires particle filtering (PF) also known as Sequential Monte Carlo (SMC). One version is described in Algorithm 1 of Gallant et al. (2017). Suppose the posterior is $p(\boldsymbol{\theta}, \boldsymbol{\Lambda}_{1:T} | \mathcal{Y}_{1:T})$ where $\boldsymbol{\Lambda}_{1:T}$ denotes the latent variables whose prior can be described by $p(\boldsymbol{\Lambda}_t | \boldsymbol{\Lambda}_{t-1}, \boldsymbol{\theta})$ and $\mathcal{Y}_{1:T}$ denotes all data. Also suppose that we have $\boldsymbol{\Lambda}_{1:T}^{(1)}$ from the previous iteration. The particle filtering procedure consists of two phases.

Phase I: Forward filtering (Andrieu et al., 2010).

- Draw a proposal $\boldsymbol{\Lambda}_{i,t}^{(m)}$ from an importance density $q(\boldsymbol{\Lambda}_{i,t} | \boldsymbol{\Lambda}_{i,t-1}^{(m)}, \boldsymbol{\theta})$, $m = 2, \dots, M$.
- Compute the importance weights:

$$w_{i,t}^{(m)} = \frac{p(y_{i,t}; \boldsymbol{\Lambda}_{i,t}^{(m)}, \boldsymbol{\theta}) p(\boldsymbol{\Lambda}_{i,t}^{(m)} | \boldsymbol{\Lambda}_{i,t-1}^{(m)}, \boldsymbol{\theta})}{q(\boldsymbol{\Lambda}_{i,t} | \boldsymbol{\Lambda}_{i,t-1}^{(m)}, \boldsymbol{\theta})}, m = 1, \dots, M. \quad (17)$$

- Normalize the weights: $\tilde{w}_{i,t}^{(m)} = \frac{w_{i,t}^{(m)}}{\sum_{m'=1}^M w_{i,t}^{(m')}}$, $m = 1, \dots, M$.
- Resample the particles $\{\boldsymbol{\Lambda}_{i,t}^{(m)}, m = 1, \dots, M\}$ with probabilities $\{\tilde{w}_{i,t}^{(m)}, m = 1, \dots, M\}$.

In the original PF sampler, the particles are stored for $t = 1, \dots, T$ and a single trajectory is sampled using the probabilities from the last iteration. An improvement upon the original PF sampler (PG sample) was proposed by Whiteley et al. (2010), who suggested drawing the path of the latent variables from the particle

approximation using the backwards sampling algorithm of Godsill et al. (2004). In the forwards pass, we store the normalized weights and particles and we draw a path of the latent variables as we detail below (the draws are from a discrete distribution).

Phase II: Backward filtering (Chopin and Singh, 2013, Godsill et al., 2004).

- At time $t = T$ draw a particle $\mathbf{\Lambda}_{i,T}^* = \mathbf{\Lambda}_{i,T}^{(m)}$.
- Compute the backward weights: $w_{t|T}^{(m)} \propto \tilde{w}_t^{(m)} p(\mathbf{\Lambda}_{i,t+1}^* | \mathbf{\Lambda}_{i,t}^{(m)}, \theta)$.
- Normalize the weights: $\tilde{w}_{t|T}^{(m)} = \frac{w_{t|T}^{(m)}}{\sum_{m'=1}^M w_{t|T}^{(m')}}$, $m = 1, \dots, M$.
- Draw a particle $\mathbf{\Lambda}_{i,t}^* = \mathbf{\Lambda}_{i,t}^{(m)}$ with probability $\tilde{w}_{t|T}^{(m)}$.

Therefore, $\mathbf{\Lambda}_{i,1:T}^* = \{\mathbf{\Lambda}_{i,1}^*, \dots, \mathbf{\Lambda}_{i,T}^*\}$ is a draw from the full conditional distribution. The backwards step often

results in dramatic improvements in computational efficiency. For example, Creal and Tsay (2015) find that $M = 100$ particles is enough. There remains the problem of selecting an importance density $q(\mathbf{\Lambda}_{i,t} | \mathbf{\Lambda}_{i,t-1}, \theta)$. We use an importance density implicitly defined by $\mathbf{\Lambda}_{i,t} = \mathbf{a}_{i,t} + \sum_{p=1}^P \mathbf{b}_{i,t} \odot \mathbf{\Lambda}_{i,t-1}^p + \mathbf{h}_{i,t} \odot \boldsymbol{\xi}_{i,t}$ where $\boldsymbol{\xi}_{i,t}$ follows a standard (zero location and unit scale) Student-t distribution with $\nu = 5$ degrees of freedom. That is, we use polynomials in $\mathbf{\Lambda}_{i,t-1}$ of order P . We select the parameters $\mathbf{a}_{i,t}$, $\mathbf{b}_{i,t}$ and $\mathbf{h}_{i,t}$ during the burn-in phase (using $P = 1$ and $P = 2$) so that the weights $\{\tilde{w}_{i,t}^{(m)}, m = 1, \dots, M\}$ and $\{\tilde{w}_{t|T}^{(m)}, m = 1, \dots, M\}$ are approximately not too far from a uniform distribution. Chopin and Singh (2013) have analyzed the theoretical properties of the PG sampler, and proved that the sampler is uniformly ergodic. They also prove that the PG sampler with backwards sampling strictly dominates the original PF sampler in terms of asymptotic efficiency. Alternatively, when the dimension of the state vector is large, we can draw $\mathbf{\Lambda}_{i,1:T}$, conditional on all other paths $\mathbf{\Lambda}_{-i,1:T}$ that are not path i . Therefore, we can draw from the full conditional distribution $p(\mathbf{\Lambda}_{i,1:T} | \mathbf{\Lambda}_{-i,1:T}, \mathcal{Y}_{1:T}, \theta)$.

3 Data and empirical results

3.1 Data

We use the data from the Indonesian manufacturing census covering all firms with more than 20 employees during 1991 - 2001 (taken from AK). There are 170,741 observations in the sample, and the data include imports, exports, output, labor, material inputs, and capital stock. Details on the construction/description of the data/variables can be found in AK. Following AK the only quasi-fixed input is capital stock. Labor and materials are assumed to be variable inputs.

Once the parameters are estimated, we use the predicted values of ω and u from (5) and (6). That is, $\hat{\omega}_{i,t} = \hat{h}_1(\hat{\omega}_{i,t-1}, \mathbf{k}_{i,t}, \mathbf{x}_{i,t}, \mathbf{z}_{i,t}; \hat{\gamma}_1)$ and $\hat{u}_{i,t} = \exp\{\hat{h}_2(\ln \hat{u}_{i,t-1}, \mathbf{k}_{i,t}, \mathbf{x}_{i,t}, \mathbf{z}_{i,t}; \hat{\gamma}_2)\}$. Note that since we specify ω and u in (5) and (6) as two separate equations, the intercept terms in both ω and u are identified.

3.2 Empirical results

We use a translog production function (see also AK, p. 1633).⁵ The selection of compression parameters ψ and δ for the different functions is reported in Table 1.⁶

⁵We implement the method of Chernozhukov and Hong (2003) using 150,000 iterations omitting the first 50,000 in the interest of mitigating possible starting value effects. The PF approach is implemented using 10,000 particles per MCMC iteration. Technical details on MCMC performance, convergence and particle behavior are available from the authors upon request.

⁶The p -value of the Hansen-Sargan J -statistic was 0.34 for the final specification in Table 1.

Table 1: Selection of compression parameters

	ψ	$\frac{\delta}{d}$
Ψ_1	0.71	0.44
Ψ_2	0.45	0.30
h_1	0.38	0.21
h_2	0.41	0.25

Note: In this Table, we report optimal values of parameters ψ and δ (as a fraction of the actual dimensionality d) from the compression procedure in (11) and (12), using Andrews' (1999) BIC statistic.

Since ω is identified up to a constant in OP (1996), i.e., it cannot be separated from the intercept term in the production function. This is why productivity growth, defined as $\Delta\omega_{i,t} = \hat{\omega}_{i,t} - \hat{\omega}_{i,t-1}$, is often reported. This is not the case in our model. In panel (a) of Figure 1, we report sample distributions of efficiency ($e^{-\hat{u}_{i,t}}$). It can be seen from the figure that efficiency of firms in sectors 311, 312, 313 and 314 are quite high – the mean efficiency ranging from about 92% to 95%. The efficiency distributions are also quite tight. In panel (b) we report $\Delta\omega_{i,t}$. Here we see large variations ranging from -6% to 10%. Almost all the firms in sector 313 show positive productivity growth – the mean being around 3.7%. In panel (c) we report sample distributions of efficiency change defined as $\Delta r_{i,t} = r_{i,t} - r_{i,t-1}$ where $r_{i,t} \equiv e^{-\hat{u}_{i,t}}$. Firms in sector 311 have on average zero efficiency change. Almost half of these firms have positive efficiency change. Firms in sector 314 have on average 1% efficiency change with a range of -3% to 3.4%.

Next, we report sample distributions of input elasticities in Figure 2. We compare estimated elasticities of capital, labor and materials from our model with those of AK. It can be seen that, on average, capital elasticities from AK are almost half of the estimates from our model. These elasticities (based on all data) are unbelievably low, ranging from zero to 0.27, with a mean of 0.12. In contrast, capital elasticities range from 0.19 to 0.28 with a mean of .024 in our model. These numbers are more realistic than those from AK. The same pattern is observed for firms in Sectors 311. Technology in Sector 312 is however different from Sector 311 in terms of our model as well as in AK, in which elasticity of labor is much smaller. In contrast labor elasticities in our model are much higher. Materials elasticities are the highest in AK (almost double of our model). The extremely low elasticities of capital and labor in AK are compensated by much larger elasticities materials in our estimates. The general consensus is that the technology in these two Sectors are found to be quite different no matter whether we look at it from our model or AK's model. Thus although returns to scale (sum of these elasticities) are quite similar in our model as well as in AK, the individual elasticities are very different. This difference is likely due to the fact that we use a system approach which is more flexible and comprehensive. This is reflected by the fact that the capital elasticities are more realistic (not too low as evidenced by other papers that use an approach similar to AK).

In Figure 3, we show the relationship between inefficiency and productivity growth. Note that in our model both inefficiency and productivity are functions of $\mathbf{k}_{i,t}, \mathbf{x}_{i,t}, \mathbf{z}_{i,t}$ (see equation (5) and (6)) as well as their lagged values. Thus, it is expected that they will be correlated via $\mathbf{k}_{i,t}, \mathbf{x}_{i,t}, \mathbf{z}_{i,t}$. But the question is about the general relationship – not just the correlation which measures a linear relationship.

To explain the relationship between trade liberalization and plant level productivity (inefficiency), we report the marginal effect of output tariff, input tariff, import status, etc., on productivity ($\omega_{i,t}$) and inefficiency (u_{it}) in Tables 2 and 3. Because both productivity and inefficiency are latent variables the R^2 is computed as the average of squared correlation coefficient between “actual” (drawn through MCMC) and predicted (from right-hand-side) values from the two equations in (5) and (6). From Table 2 it can be seen that both output and input tariffs are negatively related to productivity change. The effect of input tariff is much stronger (negative) for the firms with positive import share. Given everything else, firms with positive foreign share enjoyed higher productivity growth.

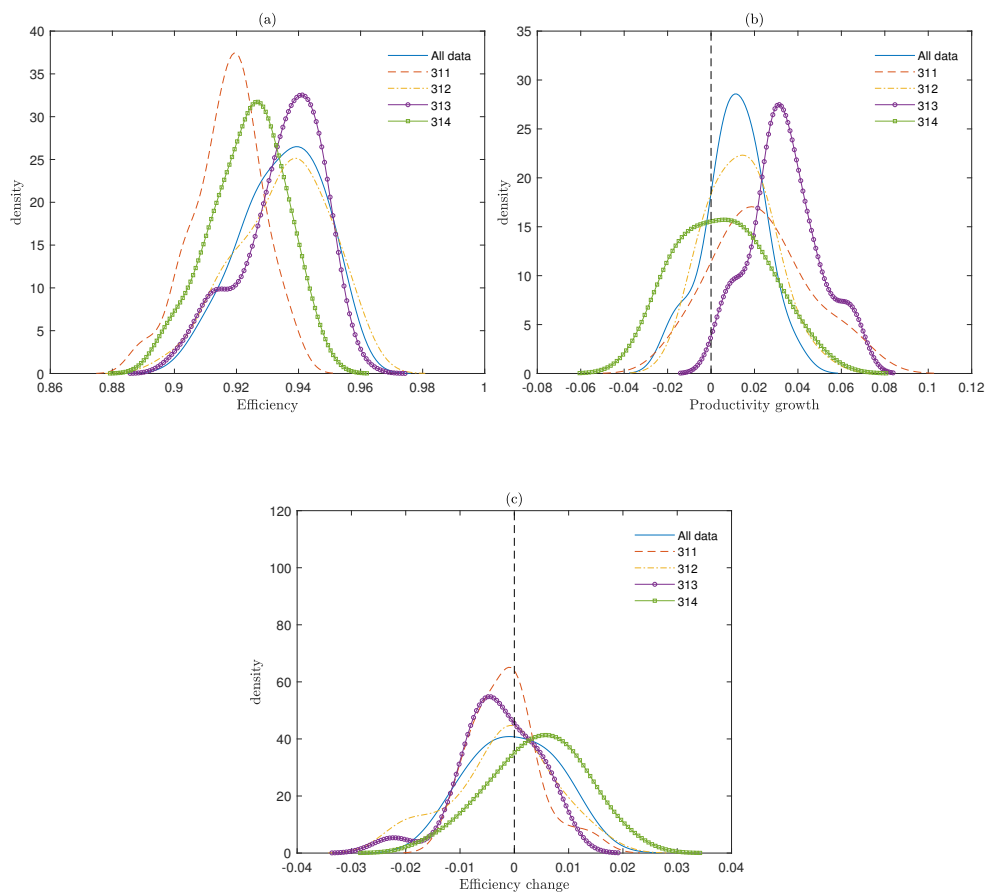
Similar to Table 2, in Table 3 we report marginal effect of tariffs on inefficiency ($\log u_{it}$). These marginal effects are in percentages, i.e., the effect of a one percent change in output tariff, for example, will increase

Table 2: Productivity and tariffs (variable to explain $\omega_{i,t}$)

variable	1991-1996 (Asian crisis)	1991-2001 (Full sample)
Output tariff	-0.031 (0.004)	-0.037 (0.002)
Input tariff	-0.077 (0.005)	-0.062 (0.003)
Input tariff×FM	-0.057 (0.007)	-0.043 (0.004)
FM=1 (import share >0)	0.017 (0.003)	0.006 (0.002)
FX=1 (export share >0)	0.008 (0.001)	0.006 (0.001)
Foreign share ≥ 0.1	0.034 (0.007)	0.029 (0.004)
Exit in $t+1$	-0.032 (0.004)	-0.021 (0.003)
Herfindahl index	-0.007 (0.002)	-0.009 (0.002)
R -squared	0.64	0.64

Note: In this Table we report marginal effects (with sample standard deviations in parentheses) of productivity growth, defined as $\Delta\omega_{i,t} = \omega_{i,t} - \omega_{i,t-1}$ on selected variables from Table 4 of AK. We include Island \times year effects and firm individual effects. Because both productivity and inefficiency are latent variables the R^2 is computed as the average of squared correlation coefficient between “actual” (drawn through MCMC) and predicted (from right-hand-side) values from the two equations in (5) and (6).

Figure 1: Plots of efficiency and productivity changes



Note: In panel (a) we report sample distributions of efficiency. In panel (b) we report sample distributions of productivity growth, and in panel (c) we report sample distributions of efficiency change. The vertical lines correspond to zero.

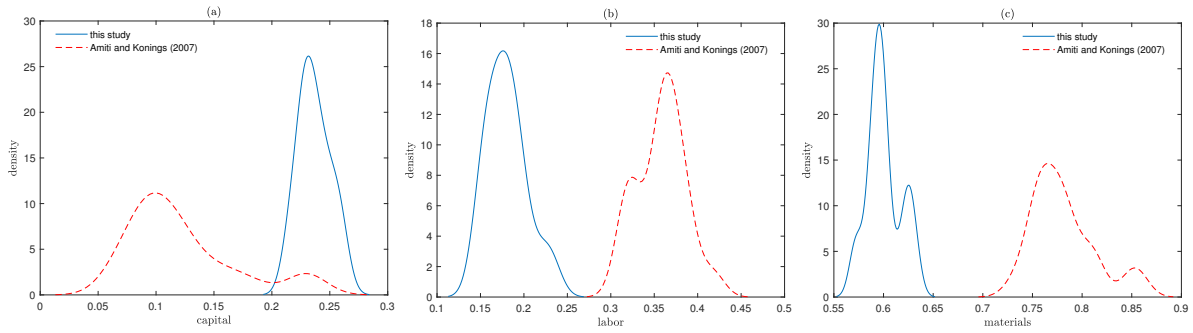
inefficiency (and hence output) by 0.025%. The effect of input tariffs is much higher (0.044%). Further, the effects are much bigger for firms with positive import share.

4 Concluding remarks

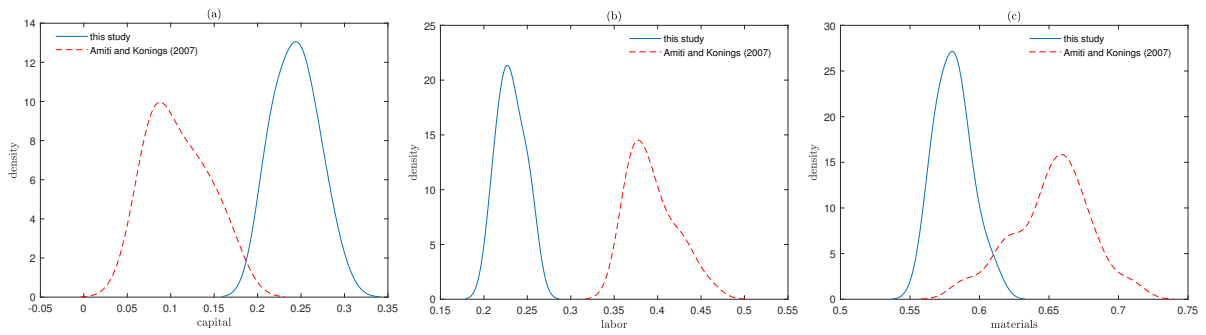
In this paper we have proposed and implemented a single-step GMM approach for estimating production functions with both productivity and inefficiency. Inefficiency and productivity depend on quasi-fixed and variable inputs semi-parametrically, and the inputs are also related semi-parametrically to inefficiency and productivity. We avoid using distributional assumptions, which are standard in stochastic frontier analysis, to separate inefficiency from productivity. We apply the new techniques to a large panel of Indonesian manufacturing plants (1991-2001). We find strong evidence that trade liberalization had a positive effect on productivity growth and a negative effect on inefficiency. Our results differ significantly from those in AK.

Figure 2: Sample distributions of input elasticities

I. All Data



II. Sector 311



III. Sector 312

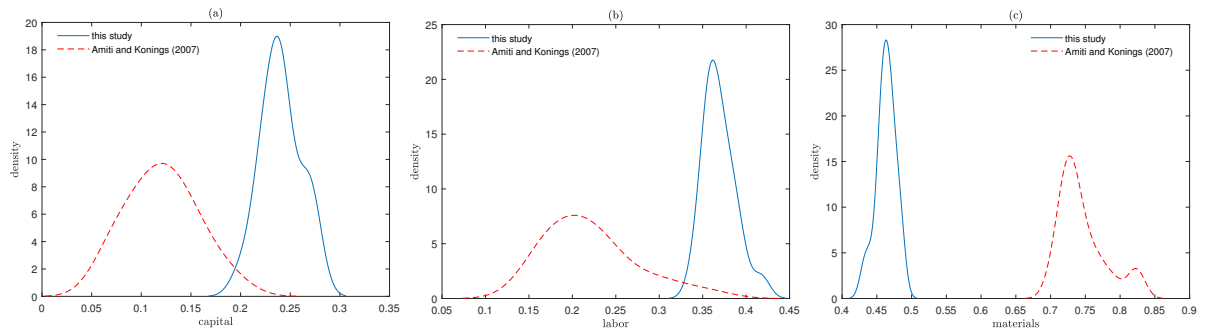
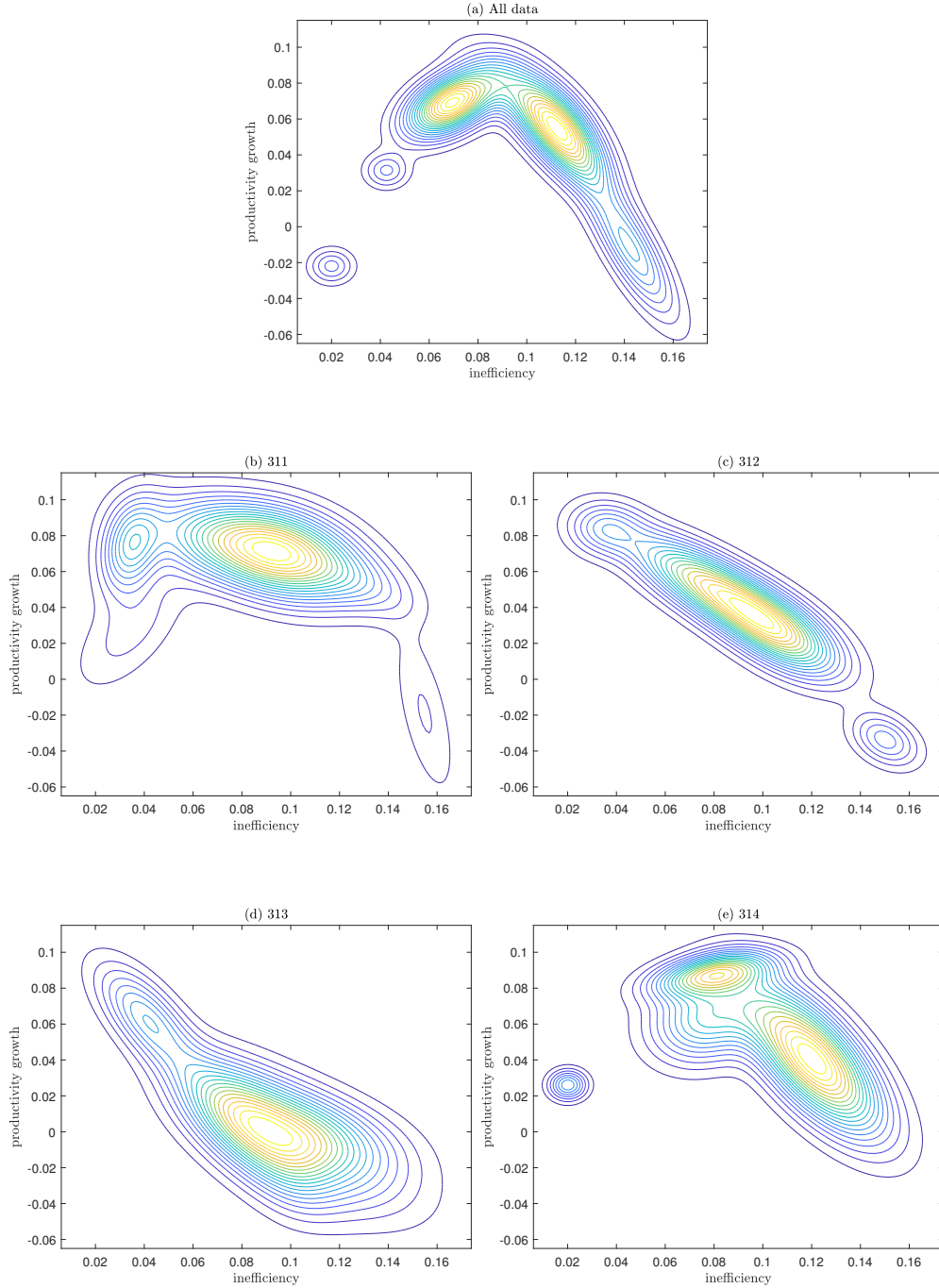


Figure 3: Relation between inefficiency and productivity growth



Note: In these figures we show the relationship between inefficiency ($u_{i,t}$) and productivity growth defined as $\Delta\omega_{i,t} = \omega_{i,t} - \omega_{i,t-1}$. In panel (a) we use all data. In panels (b) through (d) we show the relationships for sectors 311, 312, 313, and 314.

Table 3: Inefficiency and tariffs (variable to explain $\log u_{it}$)

variable	1991-1996 (Asian crisis)	1991-2001 (Full sample)
Output tariff	0.025 (0.007)	0.030 (0.005)
Input tariff	0.044 (0.011)	0.077 (0.013)
Input tariff \times FM	0.032 (0.015)	0.057 (0.012)
FM=1 (import share >0)	-0.007 (0.003)	-0.036 (0.006)
FX=1 (export share >0)	-0.054 (0.015)	-0.144 (0.006)
Foreign share ≥ 0.1	-0.034 (0.014)	-0.065 (0.008)
Exit in $t+1$	0.125 (0.017)	0.221 (0.013)
Herfindahl index	0.117 (0.026)	0.139 (0.019)
R -squared	0.55	0.61

Note: In this Table we report marginal effects (with sample standard deviations in parentheses) of log inefficiency, on selected variables from Table 4 of AK. We include Island \times year effects and firm individual effects. Because both productivity and inefficiency are latent variables the R^2 is computed as the average of squared correlation coefficient between “actual” (drawn through MCMC) and predicted (from right-hand-side) values from the two equations in (5) and (6).

References

- [1] Akerberg, D. A., K. Caves, and G. Frazer (2015). Identification Properties of Recent Production Function Estimators. *Econometrica* 83 (6), 2411–2451.
- [2] Amiti, M. and Konings, J. (2007). Trade Liberalization, Intermediate Inputs, and Productivity: Evidence from Indonesia. *American Economic Review* 97 (5), 1611–1638.
- [3] Amsler, C., Prokhorov, A., Schmidt, P. (2016). Endogeneity in stochastic frontier models. *Journal of Econometrics* 190 (2), 280–288.
- [4] Andrews, D. W. B. (1999). Consistent Moment Selection Procedures for Generalized Method of Moments Estimation. *Econometrica* 67 (3), 543–564.
- [5] Andrieu, C., Doucet, A., Holenstein, R., 2010. Particle Markov chain Monte Carlo methods (with discussion). *Journal of the Royal Statistical Society Series B* 72 (2), 1–33.
- [6] Chernozhukov, V., H. Hong (2003). An MCMC approach to classical estimation. *Journal of Econometrics* 115 (2), 293–346.
- [7] Chopin, N., Singh, S.S., 2013. On the particle Gibbs sampler. Working paper, ENSAE. <http://arxiv.org/abs/1304.1887>.
- [8] Creal, D., and R. Tsay (2015). High dimensional dynamic stochastic copula models. *Journal of Econometrics* 189 (2), 335–345.
- [9] Gallant, A. R., R. Giacomini, G. Ragusa (2017). Bayesian estimation of state space models using moment conditions. *Journal of Econometrics* 201 (2), 198–211.
- [10] Gandhi, A., S. Navarro, and D. A. Rivers (2020). On the Identification of Gross Output Production Functions. *Journal of Political Economy* 128 (8), 2973–3016.
- [11] Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Clarendon Press, Oxford, UK, 169–193.
- [12] Godsill, S.J., Doucet, A., West, M., 2004. Monte Carlo smoothing for nonlinear time series. *Journal of the American Statistical Association* 99 (465), 156–168.
- [13] Guhaniyogi, R., and Dunson, D. B. (2015). Bayesian Compressed Regression. *Journal of the American Statistical Association* 110 (512), 1500–1514.
- [14] Levinsohn, J., and A. Petrin (2003). Estimating Production Functions Using Inputs to Control for Unobservables. *Review of Economic Studies* 70 (2), 317–341.
- [15] Olley, G., and A. Pakes (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica* 64 (6), 1263–1297.
- [16] Petrin, A., and J. Sivadasan (2013). Estimating Lost Output from Allocative Inefficiency, with an Application to Chile and Firing Costs. *Review of Economics and Statistics* 95 (1), 286–301.
- [17] Tsionas, M. G. (2006). Inference in dynamic stochastic frontier models. *Journal of Applied Econometrics*, 21(5), 669–676.
- [18] Wooldridge, J. M. (2009). On estimating firm level production functions using proxy variables to control for unobservables. *Economics Letters* 104 (3), 112–114.

- [19] Whiteley, N., Sumeetpal, S., Godsill, S., 2010. Auxiliary Particle Implementation of Probability Hypothesis Density Filter. *IEEE Transactions on Aerospace and Electronic Systems* 46 (3), 1437–1454.