A Bayesian Semiparametric Approach to Stochastic Frontiers and Productivity[1]

Mike G. Tsionas[(a)] and Sushanta K. Mallick[(b)]

Abstract

In this paper we take up the analysis of production functions / frontiers removing the assumptions of known functional form for the productivity equation, given the heterogeneity of productivity and the endogeneity of inputs at firm level. The assumption of exogenous regressors is removed through taking account of the first order conditions of profit maximization. We introduce latent dynamic stochastic productivity in our framework and perform Bayesian analysis using a Sequential Monte Carlo / Particle-Filtering approach. We investigate the performance of the new approach relative to alternative methods in the literature, in a substantive application to Indian non-financial firms, and find that total factor productivity (TFP) growth has remained stagnant at firm level in India despite rapid growth at the aggregate level, with technical efficiency or catching-up effect driving TFP growth in the recent years rather than technological progress or frontier shift.

JEL Classifications: C13, C14.

Keywords: Productivity and competitiveness; Stochastic frontier model; Endogenous Regressors; Sequential Monte Carlo; Particle-Filtering.

(a) Lancaster University Management School, Bailrigg, Lancaster, LA1 4YX, UK; Email: m.tsionas@lancaster.ac.uk .

(b) **Corresponding author:** School of Business and Management, Queen Mary University of London, Mile End Road, London E1 4NS, UK; Tel: +44 (0)20 7882 7447, Email: s.k.mallick@qmul.ac.uk

[1] The authors gratefully acknowledge the editor and the three anonymous reviewers of this journal for their very constructive comments that contributed to the improvement of the paper. We of course take sole responsibility for any possible errors and omissions that might yet remain.

1. Introduction

Several challenges have been noted in the literature in measuring total factor productivity at firm level, although it is well known that increasing productivity is one of the ways to achieve sustainable improvement in living standards. The unobservable productivity shocks can be correlated with the standard inputs making the input levels endogenous. When talking about "endogeneity", there are three sources. To be precise, the first source is from intuition, i.e., decisions about k and l depend on the overall productivity. The second source is the measurement errors in the right-hand-side variables. The third source is from the profit maximization, i.e., the firms choose inputs and output simultaneously to maximize profit.

Olley and Pakes [OP] (1996) use investment as a proxy for such unobservable shocks, while Levinsohn and Petrin [LP] (2003) use intermediate inputs as a better proxy that may respond more smoothly to unobserved productivity shocks. Both approaches which are widely used in the empirical firm-level productivity literature have been subject to criticism. This paper presents a new estimation method of firm-level productivity to deal with the endogeneity problem, which is pervasive in production function estimation, by relaxing functional form assumptions typically made in practice.

The measurement of TFP is always problematic. Neither OP nor LP are devoid of problems (see Gandhi, Navarro, and Rivers[GNR] (2017), Ackerberg, Caves and Fraser [ACF] (2015), and Doralzeski and Jaumendreu (2013)). Most prominently, GNR (2017) show that, besides perfect dependence problem pointed out by ACF (2015), both the OP and LP estimators suffer from the lack of relevant IVs for the endogenous static inputs in the model. Ackerberg, Caves and Frazer (2006) have shown that the OP and LP approaches to estimating TFP have a problem of collinearity if labor and intermediate inputs are a function of TFP just like investment. Although they suggest an alternative approach to tackling the endogeneity of the inputs, the estimated TFP via a parametric approach will always be subject to collinearity in the inputs. Ferrara and Vidoli (2017) and GNR (2017) therefore proposed a semiparametric/nonparametric treatment of the production function. There could also be non-linearity due to capital-labour substitution in the sense that when labour input is very high and costly, capital could be substituted to replace labour, making the relationship endogenous and non-linear. This raises questions

about whether the functional form should be parametric or should the entire distribution be taken into account for each variable in computing TFP in a semi-parametric sense. Such endogeneity problems can be dealt with by modeling the joint distribution of regressors and the error term without the use of instruments.

Flexible functional forms are useful in many fields of applied economics and finance (e.g., Lee (1983), Heckman and Honore (1989), Trivedi and Zimmer (2006), Park and Gupta (2012), Tzeremes (2015), Sun et al. (2015), Matousek and Tzeremes (2016), and Kevork et al. (2017)). Amsler, Prokhorov and Schmidt (2014, 2016) recently modeled time dependence in stochastic frontier panel data models to address the endogeneity problem (also see Tran and Tsionas, 2013; 2016). Their approach is different from the one proposed in this paper in the sense that they use a reduced form equation to construct the joint density of the errors, while we use a flexible functional approach with latent prices and persistent measurement error in the data, to directly model the correlation between the endogenous regressors and the errors, which neither the reduced form nor the instruments are needed to obtain consistent estimates of the model parameters.

In a substantive extension of the model, we introduce latent dynamic stochastic productivity shocks a la Olley and Pakes (1996) and Levinsohn and Petrin (2003) in our framework. This is essential as it is typically ignored in applied studies (see a survey in Fethi and Pasiouras, 2010; and Fukuyama and Matousek, 2017). Bayesian analysis is performed using a Sequential Monte Carlo / Particle-Filtering approach. In this paper, using nonparametric approach and firm-level data from India to address the endogeneity of regressors in a production function, our results reveal that a flexible functional form approach best describes our data in estimating productivity. Our results are indicative of the inappropriateness of deriving TFP estimates in the presence of endogenous regressors. In fact, differentiating firms by size reveals that the TFP estimates are much smaller for medium and large firms and even negative for these firms, whereas smaller firms tend to display higher productivity. From our econometric analysis, it is revealed that TFP growth has remained stagnant at firm level in India, with technical efficiency or catching-up effect driving TFP growth in the recent years rather than technological progress or frontier shift. Lower firm productivity gains using our methodology suggest

that the high firm growth in the last decade can be traced to higher capital accumulation and higher skilled workers rather than firm spending on innovation.

The remainder of the paper is organized as follows. In Section 2, we provide a summary of the existing methodologies to estimate productivity, and in section 3, we propose the methodology that extends the traditional stochastic frontier model, using a flexible functional form approach. In section 4, the Bayesian approach is discussed, along with the empirical application in Section 5. Conclusions and further discussion are given in Section 6.

2. Preliminaries

In this section we review previous contributions to the estimation of production functions and productivity. The basic model in logs is $y_{it} = \beta_k k_{it} + \beta_l l_{it} + \omega_{it} + v_{it}$ where y_{it}, k_{it}, l_{it} are the logs of output, capital and labor, v_{it} is a shock which is unobserved by the firm, and ω_{it} represents a shock potentially predictable or observable by the firm when making its decisions. This is often referred to as the productivity shock. As decisions about k_{it}, l_{it} depend on ω_{it}, it is clear that inputs are correlated with the error term giving rise to an endogeneity problem. It is commonly assumed that the conditional distribution of the productivity shock has a Markovian structure so that $p(\omega_{it} \mid \mathcal{I}_{it}) = p(\omega_{it} \mid \omega_{i,t-1})$ where $\mathcal{I}_{it}$ is an information set.

In Olley and Pakes (OP), the endogeneity problem for capital is solved using a deterministic rule for capital accumulation. It is solved using the timing assumption about investment decisions in the face of adjustment costs in capital. This rule takes the form: $k_{it} = \kappa\left(k_{i,t-1}, i_{it}\right)$ where i_{it} is related to investment. In original units, this may take the form:

$$K_{it} = (1-\delta)K_{i,t-1} + I_{it}, \tag{1}$$

where I_{it} represents investment. As in Pakes (1994) and Ericson & Pakes (1995), OP make the assumption that $i_{it} = g_t\left(\omega_{it}, k_{it}\right)$ and strictly increasing in the productivity shock. In turn this function can be inverted to yield: $\omega_{it} = g_t^{(-1)}\left(i_{it}, k_{it}\right)$. Therefore, we can write the production function as:

$$y_{it} = \beta_k k_{it} + \beta_l l_{it} + g_t^{(-1)}\left(i_{it}, k_{it}\right) + v_{it} \equiv \beta_l l_{it} + \Phi\left(i_{it}, k_{it}\right) + v_{it}\,. \tag{2}$$

Clearly, function $\Phi(i,k)$ can be treated non-parametrically and one can obtain an estimate of this function along with an estimate of β_l. In a second stage, OP estimate productivity shock using GMM.

A problem with OP is that investment is often zero and, therefore, the monotonicity assumption cannot be invoked. Levinsohn and Petrin (LP) use instead a variable input like intermediate inputs or raw-materials (denoted by m_{it}) to express the production function as: $y_{it} = \beta_k k_{it} + \beta_l l_{it} + \beta_m m_{it} + \omega_{it} + v_{it}$. Given the input demand function $m_{it} = g_t\left(\omega_{it}, k_{it}\right)$, under monotonicity this function can be inverted to yield: $\omega_{it} = g_t^{(-1)}\left(m_{it}, k_{it}\right)$. In turn, the production function takes the form:

$$y_{it} = \beta_k k_{it} + \beta_l l_{it} + \beta_m m_{it} + g_t^{(-1)}\left(m_{it}, k_{it}\right) + v_{it}\,. \tag{3}$$

This equation can be used to obtain an estimate of β_l but not of β_k and β_m. LP use the same orthogonality condition as in OP to obtain estimates of these parameters.

OP and LP rely crucially on several rules. The first one is monotonicity result that follows from neoclassical assumptions about the production technology. The second is that there are no measurement errors so that the only latent variable is indeed ω_{it}. Thirdly, k_{it} must be a quasi-fixed input otherwise the orthogonality conditions will no longer hold. Fourth, timing assumptions are important: If m_{it} was chosen before ω_{it} was observed, then it cannot be used to invert and obtain a functional form for ω_{it}. Even in the face of all of this, the models remain under-identified as shown in GNR (2017). Ackerberg, Caves and Fraser [ACF] (2015) point out that collinearity problems arise even in the first stage and identification of β_l would be doubtful when, for example, l_{it} was chosen before ω_{it} was observed. ACF suggest alternative estimation techniques using the basic ideas of OP and LP.

Gandhi et al. (2017) propose using information from the first-order conditions with respect to variable inputs and prove an important non-parametric identification theorem for the production function. Even when prices are missing, we can estimate the model.[2]

[2] In fact, the idea goes back to Altug and Miller (1998), where they assume that unobserved wages are noisy relations of marginal products. From there to assuming that all prices are unobserved –or can be related to noisy versions of both optimality first-order conditions and observed benchmarks (if they are available), the path is not too long.

First, assume $\omega_{it} = h(\omega_{i,t-1})$. When material input is variable, one can write: $\omega_{i,t-1} = g^{(-1)}\left(k_{i,t-1}, l_{i,t-1}, m_{i,t-1}\right)$ so that the production function becomes:

$$y_{it} = f(k_{it}, l_{it}, m_{it}) + h\left(g^{(-1)}(k_{i,t-1}, l_{i,t-1}, m_{i,t-1})\right) + v_{it}. \quad (4)$$

The first order condition with respect to materials is:

$$P_t \frac{\partial F\left(K_{it}, L_{it}, M_{it}\right)}{\partial M_{it}} e^{\omega_{it}} = \rho_t, \quad (5)$$

where ρ_t is the price of materials and P_t denotes output price. These prices are common for all firms under the assumption of perfect competition. The demand for the intermediate input can be written as $m_{it} = \tilde{\mathcal{M}}\left(k_{it}, l_{it}, \omega_{it}, w_{mt}, \rho_t, P_t\right) \equiv \mathcal{M}\left(k_{it}, l_{it}, \omega_{it}\right)$. As Gandhi et al. (2013) mention, they make "*this implicit relationship an explicit one by transforming the first order condition to identify the intermediate inputs elasticity and the ex-post shock non-parametrically, which fills the void left by the lack of an exclusion restriction*".

In this paper, we also make use of the first order conditions implied by profit maximization but in a different way. The first order conditions are explicitly incorporated into a system involving the production function and the unobserved productivity shock. However, we face three problems. First, and most importantly, not all prices are available so the completion of the system depends on the fact that not all relative prices are observed. Second, we wish to take the GNR criticism seriously in that we do not intend to have ω_{it} as the only unobservable. Most importantly, the measurement errors in inputs have been shown in Kim et al. (2016). Therefore, we allow for measurement errors in the inputs. Third, we wish to opt for a flexible or semi-parametric functional form for h in $\omega_{it} = h\left(\omega_{i,t-1}\right)$. Although the literature has focused on this, we address this in a more systematic manner in a Bayesian setting along with an empirical application. GNR also use price information to estimate their model which can be traced back to Marschak and Andrews (1944). However, prices are used differently here. GNR use a proxy equation to invert and obtain productivity (and use first order condition with respect to intermediate inputs only), whereas we use all first-order conditions directly to complete the system of production function along with endogenous inputs. GNR use a non-parametric framework whereas we use a flexible semi-parametric neural network formulation (that can approximate to arbitrary degree of accuracy for any given functional form). In summary, GNR use only one first order condition while we use all of them.

3. The model

a. Economics and econometrics of the new model

Suppose the production function is $Y = F(X)e^{\omega}$ where Y is output, $X \in \mathbb{R}^K$ is a $K \times 1$ vector of inputs and ω represents productivity. Suppose $W \in \mathbb{R}^+_K$ denotes input relative prices. Under the assumption that the firm maximizes profits the optimization problem is $\max_{X \in \mathbb{R}^+_K, Y>0} : F(X)e^{\omega} - \sum_{k=1}^{K} W_k X_k$. This suggests that firms maximize static profits in all inputs (i.e., all inputs are flexible) as in OP & LP, along with the auto-regressive-ness of omega.

The first-order conditions of this problem are the following:

$$\begin{aligned} &\frac{\partial F(X)}{\partial X_k} e^{\omega} = W_k, k = 1,...,K, \\ &Y = F(X)e^{\omega}. \end{aligned} \tag{6}$$

Multiplying both sides of the first set of conditions by X_k we obtain:

$$\begin{aligned} &\frac{\partial F(X)}{\partial X_k} X_k e^{\omega} = W_k X_k, k = 1,...,K, \\ &Y = F(X)e^{\omega}. \end{aligned} \tag{7}$$

We can write these conditions in alternative form as follows:

$$\begin{aligned} &F(X)\frac{\partial \ln F(X)}{\partial \ln X_k} e^{\omega} = W_k X_k, k = 1,...,K, \\ &\ln Y = \ln F(X) + \omega. \end{aligned} \tag{8}$$

Suppose lowercase letters denote logarithms so that $x = \ln X$, $y = \ln Y$ and $w = \ln W$. Moreover let $\ln F(X) = f(x)$. The conditions above can be written as:

$$\begin{aligned} &f(x) + \ln \frac{\partial f(x)}{\partial x_k} + \omega = w_k + x_k, k = 1,...,K, \\ &y = f(x) + \omega. \end{aligned} \tag{9}$$

This is a system of $K+1$ equations in the $K+1$ endogenous variables, viz. X and Y. For most purposes, a translog production function is adequate, while considering a non-parametric treatment of F(.). The functional form is:

$$f(x;\beta) = \beta_o + \sum_{k=1}^{K} \beta_k x_k + \tfrac{1}{2} \sum_{k=1}^{K} \sum_{k'=1}^{K} \beta_{kk'} x_k x_{k'}. \tag{10}$$

The derivatives are:

$$\frac{\partial f(x)}{\partial x_k} = \beta_k + \sum_{k'=1}^{K} \beta_{kk'} x_{k'}. \tag{11}$$

Therefore, the full system of equations arising from profit maximization is as follows:

$$\begin{aligned} & f(x) + \ln\left(\beta_k + \sum_{k'=1}^{K} \beta_{kk'} x_{k'}\right) + \omega - w_k - x_k = 0, k = 1,...,K, \\ & y - \left(\beta_o + \sum_{k=1}^{K} \beta_k x_k + \tfrac{1}{2} \sum_{k=1}^{K} \sum_{k'=1}^{K} \beta_{kk'} x_k x_{k'}\right) - \omega = 0. \end{aligned} \tag{12}$$

The endogenous variables are y and $x \in \mathbb{R}^K$. These endogenous variables will, of course, depend on unobserved productivity ω. As prices vary across firms, to keep the term (in *ln*) positive we use $\frac{1}{2}\ln\left(\beta_k + \sum_{k'=1}^{K} \beta_{kk'} x_{k'}\right)^2$.

Provided we have panel data, the system of equations can be written as:

$$\begin{aligned} & f(x_{it};\beta) + \ln\left(\beta_k + \sum_{k'=1}^{K} \beta_{kk'} x_{k',it}\right) + \omega_{it} - w_{k,it} - x_{k,it} = v_{it,k}, k = 1,...,K, \\ & y_{it} - \left(\beta_o + \sum_{k=1}^{K} \beta_k x_{k,it} + \tfrac{1}{2} \sum_{k=1}^{K} \sum_{k'=1}^{K} \beta_{kk'} x_{k,it} x_{k',it}\right) - \omega_{it} = v_{it,K+1}, \end{aligned} \tag{13}$$

where $V_{it} = \left[v_{1,it},...,v_{K+1,it}\right]'$ represents a vector error term, and $\beta_{o,i}$ allows for individual effects. The Jacobian of transformation from V_{it} to (y_{it}, x_{it}) can be shown to be the absolute value of the determinant of the matrix: $\mathcal{J} = [G_{kk'}] - I$ where $G_{kk'} = f_{k,it} + \dfrac{\beta_{kk'}}{f_{k,it}}$, $f_k = \beta_k + \sum_{k'=1}^{K} \beta_{kk'} x_{k',it}$. We denote this Jacobian by $J_{it} = \|\mathcal{J}_{it}\|$.

For unobserved productivity we assume[3]:

$$\omega_{it} = h\left(\omega_{i,t-1};\alpha\right) + \varepsilon_{it}, \tag{14}$$

[3] It might be tempting to assume that ω_{it} depends also on current or lagged inputs. However, it is made clear in OP and LP that input choices depend on productivity, not the other way round.

where $\mathbb{E}\left(\varepsilon_{it} \mid \omega_{i,t-1}\right) = 0$, h is a functional form and α is a vector of parameters. More specifically,

$$\varepsilon_{it} \sim N(0, \sigma_{\varepsilon}^2).$$

Provided we specify this functional form *and assume* that all input relative prices are available, it is possible to estimate the system of equations in (13) by maximum likelihood (ML) or Bayesian techniques by taking into account the Jacobian of transformation.

However, we face three problems. *First*, and most importantly, not all prices are available. *Second*, we wish to take the ACF criticism seriously in that we do not intend to have ω_{it} as the only unobservable. *Third*, we wish to opt for a flexible or semi-parametric functional form for h.

Regarding the functional form for h we think that a *flexible semi-parametric functional form* (as opposed to a full nonparametric specification) is enough for most practical purposes.

The claim is due to the fact that neural networks can approximate arbitrarily well any functional form; a fact well-known in the literature. Full nonparametric specifications are subject to the curse of dimensionality as the number of explanatory variables increases. We use the term "flexible semiparametric functional form" to refer to an artificial neural network: Although clearly parametric, this family is quite flexible as it can approximate any functional form, see for example Hornik et al. (1989). Besides, in the productivity literature most papers use a Cobb-Douglas functional form and focus on other issues, like endogeneity. Here, we do focus on endogeneity but we use a much more flexible functional form.

Evidence to that is provided by OP, LP and ACF when they use a polynomial for their inverse function which relates productivity shock to the observable variables. In this study we use a neural network which has well-known global approximation properties (Hornik et al., 1989). Therefore, we assume:

$$h\left(\omega_{i,t-1}; \alpha\right) = \sum_{g=1}^{G} \alpha_g \frac{1}{1 + \exp\left(-\alpha_{G+g} \omega_{i,t-1}\right)}. \tag{15}$$

In this neural network, there are G nodes, and a logistic sigmoidal activation function is used. This sigmoid is given by: $\varphi(z) = \frac{1}{1+\exp(-z)}, z \in \mathbb{R}$. Here, $\alpha \in \mathbb{R}^{2G}$ is an unknown parameter vector. The exact value of G will be chosen using the data.

We use the concept of marginal likelihood and Bayes factor to select the value of G. To summarize, for any value of G, there is a probability to observe the data, say P(Y;G) which results from the posterior: $P(Y;G)=\int p(\theta\,|\,Y;G)d\theta$. This can be approximated for any value of G and in turn we can compute Bayes factors relative to, say, G=1 as follows: $BF_g=\dfrac{P(Y;G=g)}{P(Y;G=1)}, g=2,3,...$ The Bayes factor takes into account both model fit as well as model complexity arising from more parameters as G is increased.

We next account for measurement error in the data which is, clearly, a major problem in all production and productivity studies. The actual data $\left(y_{it},x_{it}\right)$ are unobserved. Instead we observe $\mathcal{D}=\left(y_{it}^{o},x_{it}^{o}\right)$ which are related to the unobserved data as follows:

$$\begin{aligned} y_{it}^{o} &= y_{it}+\varepsilon_{y,it}, \\ x_{it}^{o} &= x_{it}+\varepsilon_{x,it}, \end{aligned} \tag{16}$$

where $\varepsilon_{it}=\left[\varepsilon_{y,it},\varepsilon_{x,it}'\right]'$ represents measurement error. Of course, $\varepsilon_{y,it}$ cannot be distinguished from $v_{it,K+1}$ so, effectively, we have measurement error only in the inputs and we assume $y_{it}^{o}=y_{it}$ without loss of generality. We do not assume that observed variables are random realizations centered on x_{it}. Instead *we assume persistence of measurement errors*:

$$\varepsilon_{x,it}=\Phi\varepsilon_{x,i,t-1}+\xi_{it},\ \xi_{it}\sim N_K\left(0,\Omega\right), \tag{17}$$

where Φ is a $K\times K$ matrix of unknown coefficients and Ω is a general covariance matrix. Both Φ and Ω are not necessarily diagonal.

The problem of missing prices is more important. In our application we have the nominal price of labor but not the nominal or relative prices of capital and intermediate inputs. The price of output is unavailable. Therefore, it is best to treat all relative prices as unobserved. We assume the following structure:

$$w_{k,it}=\mu_{ki}+\lambda_{kt}, k=1,...,K-1, \tag{18}$$

where μ_{ki},λ_{it} represent input-specific individual and time effects. Therefore, relative prices are not assumed to be constant over all firms, for a given time period. The last price, viz. $w_{K,it}$ is actually observed.

Our assumptions are as follows:

$$\mu_{ki} \sim N\left(\overline{\mu}_k, \sigma^2_{\mu,k}\right), k = 1,..,K, i = 1,...,n, \tag{19}$$

$$\lambda_{kt} \sim N\left(\overline{\lambda}_k, \sigma^2_{\lambda,k}\right), k = 1,..,K, t = 1,...,T, \tag{20}$$

These equations say that we have two components of input relative prices: one that is firm specific and another that is time specific. In this sense we assume that input relative prices can be separated into firm specific and time varying components. Finally, we assume $v_{it} \sim N_{K+1}\left(0, \Sigma\right)$. Given our focus on non-financial manufacturing-based firms, the relative price across different manufacturing sectors can be assumed to remain non-heterogeneous, suggesting a somewhat stable relative price with output.

3.2 Some notes on the new approach

The new approach relies on the exploitation of all first order conditions from profit maximization in system (13). The productivity shock, ω_{it}, appears in all equations of the system and is subject to the flexible specification in (14). Individual effects (accounting for heterogeneity in production) can be used in the translog production function in (13) that can be separately identified from the individual and time effects in the first order conditions. The assumption that all inputs are flexible can be removed easily by assuming that certain inputs (like capital and possibly labor) are chosen before observing the productivity shock). In this case, these inputs are quasi-fixed and their first order conditions can be removed from the system in (13). Relative prices will not be needed for such inputs so the number of individual and time effects is reduced.

From the system in (13) it is also clear that the inputs and the productivity shock are correlated (giving rise to the classical endogeneity problem in production function estimation) as they are chosen conditionally on the productivity shock. Other than that, (14) allows for a rich semi-parametric specification of its law of motion.

Finally, measurement errors can be identified through the nonlinearity implied by the system of production function and profit maximization first order conditions in (13).

The assumption of a common translog technology (with individual effects to capture heterogeneity) may be valid for firms in the same sector, but otherwise it may be questionable. We will examine this assumption in section 5.2.

3.3 Priors

For the parameters β of the translog production function in (10) we assume:

$$p(\beta) \propto \mathbb{I}_{\mathcal{S}}(\beta), \tag{21}$$

where $\mathcal{S}$ is the set where the monotonicity conditions are satisfied, viz. the first derivatives in (11) are positive at all unobserved data points x_{it} whose observed counterpart is x^o, and $\mathbb{I}$ denotes the indicator function. For the parameters α in the productivity equation (15) we assume:

$$\alpha \sim N_{2G}(\bar{\alpha}, \bar{V}_{\alpha}), \tag{22}$$

where $\bar{\alpha} = 0$ and $\bar{V}_{\alpha} = 10^3 I_{2G}$. This prior is relatively diffuse. For matrix Φ in (17) we impose the prior notion that it is close to diagonal. Specifically, we have:

$$\Phi = \left[\phi_{ij}\right],\ \phi_{ii} \sim N(\underline{\phi}, \underline{\sigma}^2_{\phi}),\ \phi_{ij} \sim N(0, \underline{\sigma}^2_{\phi}), i \neq j, \tag{23}$$

where $\underline{\phi} = 0.5, \underline{\sigma}^2_{\phi} = 0.1$. The prior implies that the diagonal elements range from 0.3 to 0.7 with prior probability 95%.

For Ω we use a Wishart prior of the form:

$$p(\Omega) \propto |\Omega|^{-(\bar{n}+1)} \exp\left\{-\tfrac{1}{2}\bar{A}\Omega^{-1}\right\}, \tag{24}$$

where $\bar{n} = 1$ and $\bar{A} = 10^{-3} I$. For parameters $\bar{\mu}_k, \sigma^2_{\mu,k}$ and $\bar{\lambda}_k, \sigma^2_{\lambda,k}$ in (19) and (20) we assume:

$$\bar{\mu}_k \sim N(\underline{\mu}, \underline{\sigma}^2_{\mu}),\ \frac{\bar{q}_{\sigma}}{\sigma^2_{\mu,k}} \sim \chi^2(\underline{n}_{\sigma}), \tag{25}$$

$$\bar{\lambda}_k \sim N(\underline{\mu}, \underline{\sigma}^2_{\lambda}),\ \frac{\bar{q}_{\lambda}}{\sigma^2_{\lambda,k}} \sim \chi^2(\underline{n}_{\lambda}), \tag{26}$$

where $\underline{\mu} = 0$, $\bar{q}_{\sigma} = \bar{q}_{\lambda} = 10^{-3}$, $\underline{n}_{\sigma} = \underline{n}_{\lambda} = 1$. The prior for σ^2_{ε} has the same form. Our prior for Σ is the same as the prior for Ω. For practical purposes, these priors are diffuse priors. To impose the monotonicity restrictions in set $\mathcal{S}$ we use rejection sampling after imposing monotonicity at the means of the data.

4. Bayesian analysis

We collect all parameters in the vector $\theta \in \Theta \subset \mathbb{R}^d$ where d denotes the dimensionality of the parameter vector. We denote $\omega = \{\omega_{it}, i = 1,...,n, t = 1,...,T\}$ and

$\omega_i = \{\omega_{it}, t = 1,...,T\}, i = 1,...,n$. The collection of relative prices is $w = \{w_{it}, i = 1,...,n, t = 1,...,T\}$. Suppose we write the system in (13) compactly as follows:

$$\mathcal{F}\left(\mathbf{x}_{it}; \theta, \omega_{it}, w_{it}\right) = V_{it}, \tag{27}$$

where $\mathbf{x}_{it} = \left[x_{it}', y_{it}\right]'$. Therefore, we can write the augmented posterior distribution as follows:

$$\begin{aligned} &p(\theta, \omega, w, x \mid \mathcal{D}) \propto \left|\Sigma\right|^{-nT/2} \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T} \mathcal{F}\left(\mathbf{x}_{it}; \theta, \omega_{it}, w_{it}\right)' \Sigma^{-1} \mathcal{F}\left(\mathbf{x}_{it}; \theta, \omega_{it}, w_{it}\right)\right\} \cdot \\ &\prod_{i=1}^{n}\prod_{t=1}^{T} J_{it}(\theta) \cdot \left|\Omega\right|^{-nT/2} \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T} [(x_{it} - x_{it}^{o}) - \Phi(x_{i,t-1} - x_{i,t-1}^{o})]' \Omega^{-1} [(x_{it} - x_{it}^{o}) - \Phi(x_{i,t-1} - x_{i,t-1}^{o})]\right\} \cdot \\ &\sigma_{\varepsilon}^{-nT} \exp\left\{-\tfrac{1}{2\sigma_{\varepsilon}^{2}}\sum_{i=1}^{n}\sum_{t=1}^{T} [\omega_{it} - h(\omega_{i,t-1}; \alpha)]^2\right\} \cdot p(\theta), \end{aligned} \tag{28}$$

where $x = \{x_{it}\}$ is the collection of latent input data, and $p(\theta)$ denotes the joint prior on the "structural" parameters. Moreover, $w_{k,it} = \mu_{ki} + \lambda_{kt}$. In the model we have three sets of latent variables, viz. productivity shocks ω, unobserved input data x whose empirical counterpart is x^{o}, and unobserved relative input prices, w.

We write the posterior, in terms of the latent productivity shocks, ω_{it}, as:

$$p(\{\omega_{it}\} \mid \theta, \cdot) \propto \prod_{i=1}^{n}\prod_{t=1}^{T} p(\omega_{it} \mid \omega_{i,t-1}, \theta)\, p(\theta; \omega_{it}, \cdot, \mathcal{D}). \tag{29}$$

To update the dynamic latent variables ω, we use Sequential Monte Carlo or Particle Filtering as developed in Creal and Tsay (2015). The same applies to other dynamic latent variables in the model, like x_{it}. We describe the algorithm here briefly. Suppose we have $\omega_{i,1:T}^{(1)}$ from the previous iteration. For $t = 1,...,T$ we proceed as follows:

- For $s = 2,...,S$ draw a proposal $\omega_{it}^{(s)} \sim q(\omega_{it} \mid \omega_{i,t-1}^{(s)}, .)$.
- For $s = 1,...,S$ compute the weights $p_t^{(s)} = \dfrac{p(\omega_{it}^{(s)} \mid \omega_{i,t-1}^{(s)}, \theta) \cdot p(\theta; \omega_{it}^{(s)}, \mathcal{D})}{q(\omega_{it}^{(s)} \mid \omega_{i,t-1}^{(s)}, .)}$, where $p(\theta; \omega_{it}^{(s)}, \mathcal{D})$ denotes the part of the posterior that excludes the prior part for ω_{it} as in (29).

- Normalize the weights: $\hat{p}_t^{(s)} = \frac{p_t^{(s)}}{\sum_{s'=1}^{S} p_t^{(s')}}, s = 1,...,S.$

- Conditionally resample the weights $\left\{\omega_{it}^{(s)}\right\}_{s=1}^{S}$ with probabilities $\{\hat{p}_t^{(s)}\}_{s=1}^{S}$.

As Creal and Tsay (2015, p. 339) mention, this "*is a standard Gibbs sampler but defined on an extended probability space that includes all the random variables that are generated by a particle filter. Implementation of the PG sampler is different than a standard particle filter due to the 'conditional' resampling algorithm used in the last step. Specifically, in order for draws from the particle filter to be a valid Markov transition kernel on the extended probability space*". Moreover, we use a backwards step proposed by Whiteley (2010) and Godsill et al. (2004) which improves dramatically in terms of performance. Specifically, given the normalized weights and particles $\{\hat{p}_t^{(s)}, \omega_{it}^{(s)}\}$ for $t = 1,...,T$ we draw, from this discrete distribution, a path of latent variables $\{\omega_{i,1:T}^{*}\}$.

- For $t = T$ we draw a particle $\omega_{iT}^{*} = \omega_{iT}^{(s)}$ with probability $\hat{p}_T^{(s)}$.
- For $t = T-1,....,1$ we run:
- Compute backward weights $p_{t|T}^{(s)} = \hat{p}_t^{(s)}(\omega_{i,t+1}^{*} \mid \omega_{it}^{(s)}, \theta), s = 1,...,S$.
- Renormalize the weights: $\hat{p}_{t|T}^{(s)} = \frac{p_{t|T}^{(s)}}{\sum_{s'=1}^{S} p_{t|T}^{(s')}}, s = 1,...,S$.
- Draw a particle $\omega_{it}^{*} = \omega_{it}^{(s)}$ with probability $\hat{p}_{t|T}^{(s)}$.

Then the draw $\omega_{i,1:T} = \{\omega_{i,1:T}^{*}\}$ is a draw from the full posterior conditional of the latent productivity growth variables. Chopin and Singh (2013) prove that the particle sample is uniformly ergodic and that backwards sampling strictly improves in terms of asymptotic efficiency. As in Creal and Tsay (2015) we have found that S=100 particles were adequate. The results were robust to taking S=500 and S=5,000.

We apply the same Particle Filtering technique to generate draws from the posterior conditional distribution of x_{it} which are also dynamic latent variables by construction.

To generate draws from the posterior conditional distribution of β we use a Metropolis-Hastings update. Given the current draw, $\beta^{(s)}$, a candidate draw is generated from $\beta^* \sim q(\beta)$. The next draw is $\beta^{(s+1)} = \beta^*$ with probability $\min\left\{1, \frac{p(\beta^* \mid \mathcal{D},.)/q(\beta^*)}{p(\beta^{(s)} \mid \mathcal{D},.)/q(\beta^{(s)})}\right\}$, else we set $\beta^{(s+1)} = \beta^{(s)}$. The proposal distribution, $q(\beta)$, is a multivariate Student-*t* with five degrees of freedom, with location parameter the least squares estimate from the translog production function and scale matrix hV, where V is the least squares covariance matrix and h is a positive constant which is adjusted during the burn-in phase to generate an acceptance rate close to 25% (the final acceptance rate was 27.5%). The remaining parameters are updated using their conditional posterior distributions using a Gibbs sampling step (all details are available on request).

We run our MCMC scheme using 60,000 iterations the first 10,000 of which are omitted to mitigate possible start up effects. The MCMC sampler, which was started from 100 random initial conditions and convergence, was assessed using Geweke's (1992) diagnostic.

5. Empirical results

a. The Dataset

The measurement of TFP has always been an area of active research. Any output effects that are not driven by capital, labour, and intermediate inputs, are generally accepted as a measure of technical efficiency, which is seen as the real driver of long-term growth and a forward-looking firm performance indicator. In that sense, TFP is a better measure of firm performance compared to profitability (return on assets) which is more of a backward-looking performance indicator that can change in line with the business cycles. We compare our productivity estimates obtained from a non-linear semi-parametric approach compared to the standard linear parametric TFP estimates normally used in the literature, using firm level data from India.

For the key production function variables, we use a firm-level annual dataset from India called Prowess, provided by the Centre for Monitoring Indian Economy (CMIE) over the time period 1989 to 2014, covering 5,680 (out of a total of 26,000) non-financial companies in India. The following variables are used, namely total sales, labour, capital (fixed assets), wages and salaries; intermediate inputs include expenditure on raw-materials, and energy and fuel consumption.

Data on employment is either underreported or not reported by firms, which give rise to further measurement problems, whereas wages and salaries are always reported more correctly and hence it is immaterial to know how many people are employed since there is significant variability in their skills and accordingly some workers are more productive than others and therefore get paid more. Besides, many firms tend to use contract workers (due to labour market rigidities) who do not get counted in the employment number (neither in the firm they work nor by the agent who sends these workers), but will be part of the reported total wage bill of a firm, which will better reflect the importance of labour input than just the under-reported number of employees that can inflate the TFP estimate. Moreover, in the TFP estimates, if all the variables involved are available in monetary terms, it will be immaterial whether one deflates or not, as both sides of the equation will be scaled downwards through deflating.

5.2 General Results

In this sub-section, we report the empirical results. As TFP is a residual measure encompassing the effect of technical progress, by observing the distribution of TFP within the production frontier approach we can distinguish between different categories of firms across the distribution in terms of their productivity and efficiency. As discussed earlier, different approaches have been proposed in the literature in order to derive TFP (including growth accounting, OP, LP, and SFA). A flourishing literature has examined various aspects of firm-level productivity measurement (see for example, Ackerberg et al., 2015; Bournakis and Mallick, 2018). Although the focus of the application in this paper is limited to the case of India, our methodology can be of much broader applicability.

Slower growth in this productivity is more of a global phenomenon. Also firm productivity in India remains significantly heterogeneous, despite over two decades of policy reforms including industrial and trade reforms (see Haidar, 2012; and Mallick and Yang, 2013). Lack of firm-level innovation can be the key underlying factor for any productivity puzzle in manufacturing industry. While Olley and Pakes (1996) use the investment decision to proxy for unobserved productivity, Levinsohn and Petrin (2003) make use of intermediate inputs as a proxy. This may have to do with how do we measure gains from productivity. In this context, the correct estimation of TFP becomes crucial.

We are using two benchmarks to compare our results. First, a translog production function is used without productivity; and second, we add a time trend, its square and interactions with all other inputs in the same production function. Third, our nonparametric model with a linear autoregressive process for Δw_{it}. Comparisons with the first model will help us understand the importance of unknown functional form, endogenous regressors and unobserved latent dynamic semi-parametric productivity. Comparisons with this Benchmark help us understand the importance of semi-parametric productivity *per se*.

Firm-specific TFP

In (15) it is assumed that the parameter vector α is common to all firms. Although this is reasonable as a point of departure or a working hypothesis, it is worth testing the assumption. Specifically, we assume an alternative model:

$$h\left(\omega_{i,t-1};\alpha_i\right)=\sum_{g=1}^{G}\alpha_{i,g}\frac{1}{1+\exp\left(-\alpha_{i,G+g}\omega_{i,t-1}\right)}, \tag{40}$$

where $\alpha_i \sim N_{2G}\left(\bar{\alpha},\Omega_\alpha\right)$, $\bar{\alpha} \sim N(0,10^2)$ and Ω_α follows a Wishart prior with parameters as stated previously in other instances. Draws from the posterior conditional distributions of α_is are realized using an independence Metropolis algorithm whose proposal distribution is a multivariate Student-t distribution with five degrees of freedom and location – scale parameters determined by the mode and Hessian of the log posterior conditional distributions. Some useful Bayes factors are reported in Table 1. We normalize the Bayes factor for G=1 to be equal to

1.000 and divide all other Bayes factors by its actual value. In this way we have relative Bayes factors which can be used easily for model comparison.

Table 1. Bayes factors for model with firm-specific productivity

model	Bayes factor
$G=1$	1.000
$G=2$	3.455
$G=3$	0.817
$G=4$	0.444
$G=5$	0.313
$G=6$	0.101
against $\Omega_\alpha = O_{2G\times 1}$ when $G=2$	1.473

The evidence against $\Omega_\alpha = O_{2G\times 1}$ is rather weak (the Bayes factor in favor of the hypothesis is only 1.473) so, in the light of the data, it may well be the case that the productivity equation has firm-specific coefficients α. The optimal number of terms in the neural network equation in this case is G=2, which is less than the number of terms with fixed coefficients.

We distinguish between different components of TFP namely input elasticities for each of the inputs (capital, labour and intermediate inputs) (Figure 1), returns to scale (Figure 2), technical inefficiency (Figure 3) and efficiency improvements or technical change (Figure 4). Our model captures higher 'returns to scale' estimates than from the traditional production functions that are commonly used in the literature.

In Figure 1 we report sampling distributions of posterior mean elasticities with respect to inputs, capital, labor and intermediate materials. These are sample distributions of the posterior means. All these values are positive suggesting that the monotonicity restrictions are satisfied at all (unobserved) data points. Further evidence is reported in Figure 2 where we provide sampling distributions of posterior mean returns to scale. We also report returns to scale from two other models, viz. a simple translog model without

inefficiency and productivity and a translog model with trend. In the model presented in this study, returns to scale average near unity and extend from about 0.7 to 1.3. The simple translog models produce very different estimates.

Figure 1. Sampling distributions of posterior means of input elasticities

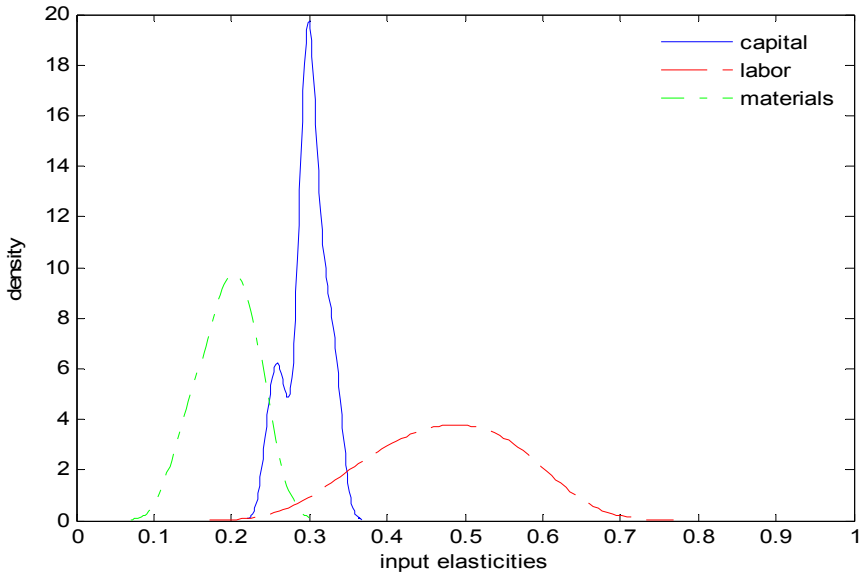


Figure 2. Sampling distributions of posterior means of returns to scale

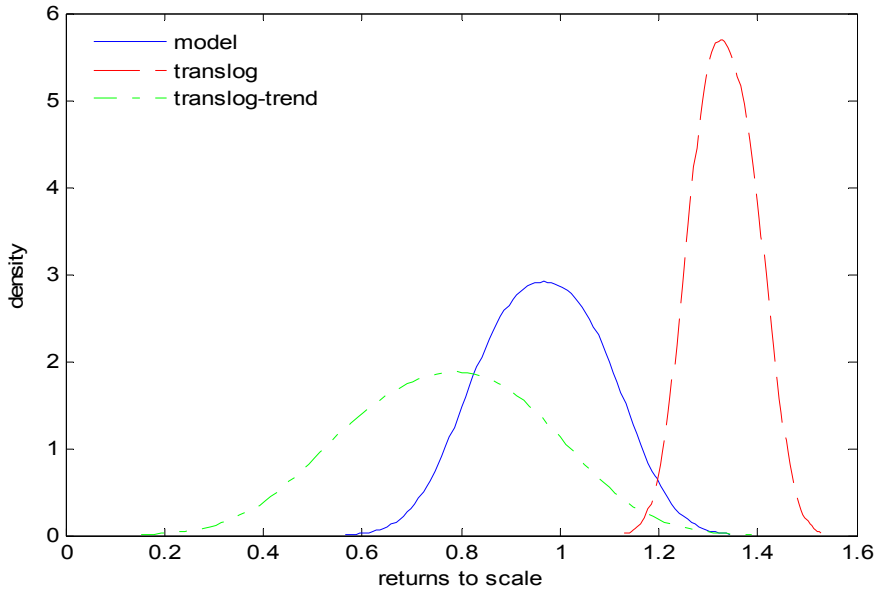


Figure 3. Sampling distributions of posterior means of technical inefficiency

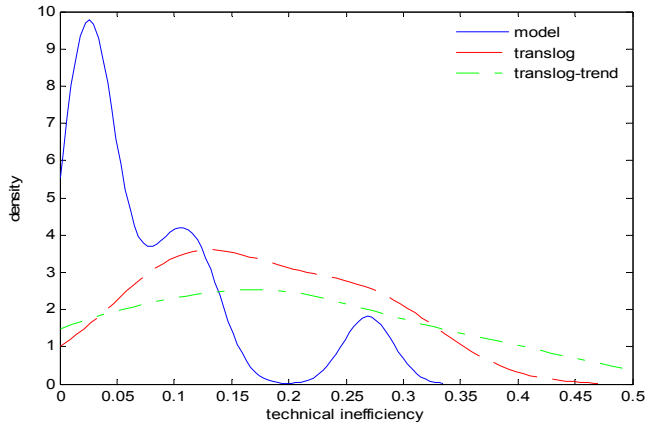
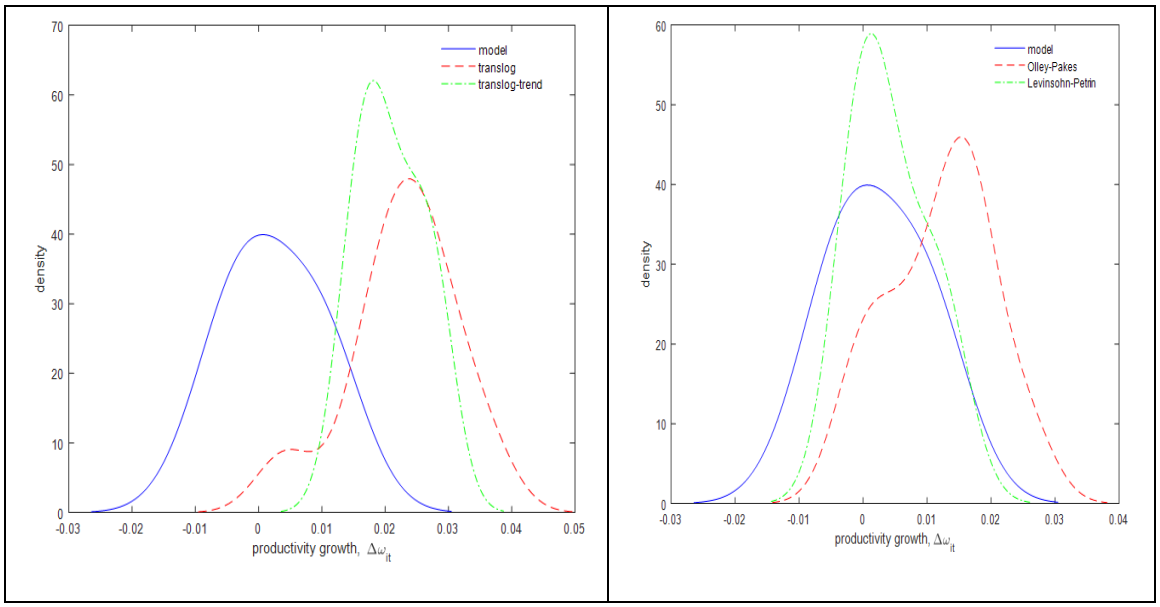


Figure 4. Sampling distributions of posterior means of productivity growth



a. Evidence on technical inefficiency

We can modify the system in (13) to allow for technical inefficiency in production. Specifically, we can modify (13) as follows:

$$\begin{aligned} & f(x_{it};\beta)+\ln\left(\beta_k+\sum_{k'=1}^{K}\beta_{kk'}x_{k',it}\right)+\omega_{it}-\tilde{u}_{it}-w_{k,it}-x_{k,it}=v_{it,k},k=1,...,K, \\ & y_{it}-\left(\beta_o+\sum_{k=1}^{K}\beta_k x_k+\tfrac{1}{2}\sum_{k=1}^{K}\sum_{k'=1}^{K}\beta_{kk'}x_k x_{k'}\right)-\omega_{it}+\tilde{u}_{it}=v_{it,K+1}, \end{aligned} \tag{30}$$

where $\tilde{u}_{it} \geq 0$ represents technical inefficiency. This corresponds to a production function of the form $Y = F(K,L,M)e^{\omega-\tilde{u}}$ in original units. According to much of the previous literature, technical inefficiency is assumed known to the producer and, therefore, it enters into each of the first order conditions.[4] Our modeling of technical inefficiency follows the seminal study of Cornwell, Schmidt and Sickles [CSS] (1990) and we parametrize it as:

$$u_{it} = \delta_{i1} + \delta_{i2}t + \delta_{i3}t^2 \ . \tag{31}$$

To impose the non-negativity constraint, we use the final estimate:

$$u_{it} = \tilde{u}_{it} - \min_{i,t} \tilde{u}_{it} \ . \tag{32}$$

For the coefficients $\delta_i = \left[\delta_{i1}, \delta_{i2}, \delta_{i3}\right]'$ we assume a random coefficient structure of the form:

$$\delta_i \sim N\left(\mathrm{O}, \Omega_\delta\right), i = 1,...,n, \tag{33}$$

where Ω_δ is a 3×3 covariance matrix. The formulation is novel due to the random coefficient specification as an anonymous referee pointed out.

Sampling distributions of posterior mean technical inefficiencies are reported in Figure 3. The sampling distribution from the model is clearly multimodal and averages 0.12 with estimates from the translog and the translog-trend models being, again, quite different.

b. Productivity growth

Productivity growth results are reported in Figure 4. According to our model, productivity growth averages near zero. According, however, to OP and LP[5], the estimates are concentrated around positive values with an average of 2% and 1% respectively, and extending from near zero to 4-5%.

In Figure 4, the estimated TFP from the semi-parametric method proposed in this study is compared with OP and LP methods as the benchmarks. It is clear that the obtained estimates of TFP growth are marginally different between the two benchmark methods.

[4] This can be seen if we replace ω by ω-u in the formulation of the original model.

[5] These models have been estimated using GMM using respective control functions suggested by the two different models. For LP we generate investment as $I_{it}=K_{i,t+1}-(1-\delta)K_{it}$. We assume a rate of depreciation δ equal to 4%.

While our semi-parametric method reveals lower productivity growth as shown in the distribution plot, the benchmark semi-parametric OP and LP methods overestimate productivity. Overall, when the LP suggests a TFP growth of around 1%, the semi-parametric method reveals little productivity growth in India in the recent decades, suggesting that firm growth has not been driven by technological innovation.

The posterior means of the functional form $h()$ relating ω_{it} and $\omega_{i,t-1}$ are reported in Figure 5 for the optimal value of G=2 and also for G=1 and G=4. The Bayes factors corresponding to different values of G (relative to G=1) are reported in Table 1. In Figure 6 we provide results related to sensitivity analysis with respect to the prior. We use 500 different priors with parameters randomly drawn using the baseline specification. For sensitivity analysis we examine the posterior means of ω_{it}, technical inefficiency and returns to scale (RTS).[6] As we can see, the posterior means are highly robust relative to the prior specification.

Another question related to the individual and time effects that have been used to model relative input prices in the model, see equations (19) and (20). Bayes factors for alternative specifications[7] are reported in Table 2.

[6] Given the baseline specification the parameters of the prior are changed in a random manner to produce a new prior. Specifically, given any parameter p whose prior involves a parameter a, we change a to aU where U is a uniform random number between 0.1 and 10 if the parameter is positive and a+Z where Z is a normal random variable with mean zero and standard deviation 10.

[7] All these Bayes factors are computed using the Verdinelli and Wasserman (1995) approach. For a possibly vector parameter δ, this approach evaluates the hypothesis H: $\delta=\delta_o$ using the Savage-Dickey ratio for the Bayes factor, viz.: $BF=p(\delta_o|Y)/p(\delta_o)$ where the expressions in the numerator and dominator are, respectively, the posterior and prior of δ evaluated at δ_o. The prior $p(\delta)$ should be proper.

Table 2. Bayes factors for alternative specifications of input relative prices

specification	Bayes factor
a) $\lambda_{ki} = \lambda_k, k = 1,..., K$	$3.12\ 10^{-7}$
b) $\lambda_{ki} = 0, k = 1,..., K$	$4.29\ 10^{-5}$
c) $\mu_{kt} = \mu_k, k = 1,..., K$	$5.14\ 10^{-4}$
d) $\mu_{kt} = 0, k = 1,..., K$	$8.71\ 10^{-5}$
e) $\lambda_{ki} = \mu_{kt} = 0$	$3.12\ 10^{-9}$
f) $\lambda_{ki} = \mu_{kt} = c_k$	$2.44\ 10^{-3}$

In specification (a), we assume individual effects are the same across firms. In specification (b) we examine whether they are actually zero. In specification (c) we examine whether time effects do not vary across time and in specification (d) whether they are actually zero. In specification (e) we examine whether both individual and time effects can be omitted and in (f) whether they can be omitted and an input-specific constant be used in their place. All Bayes factors are relative to the full model. All these hypotheses can be rejected in favor of the baseline model.

Figure 5. Posterior mean estimates of relation between $\Delta\omega_{it}$ and $\Delta\omega_{i,t-1}$

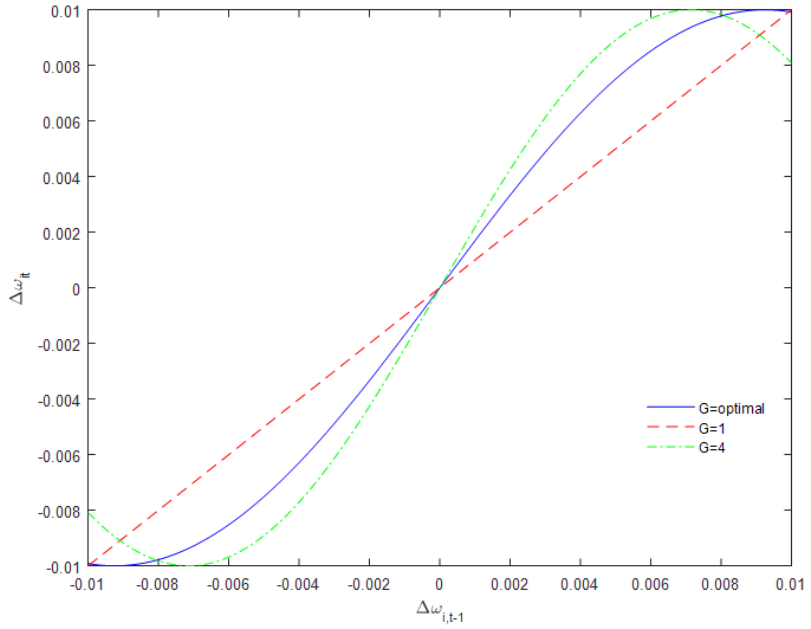
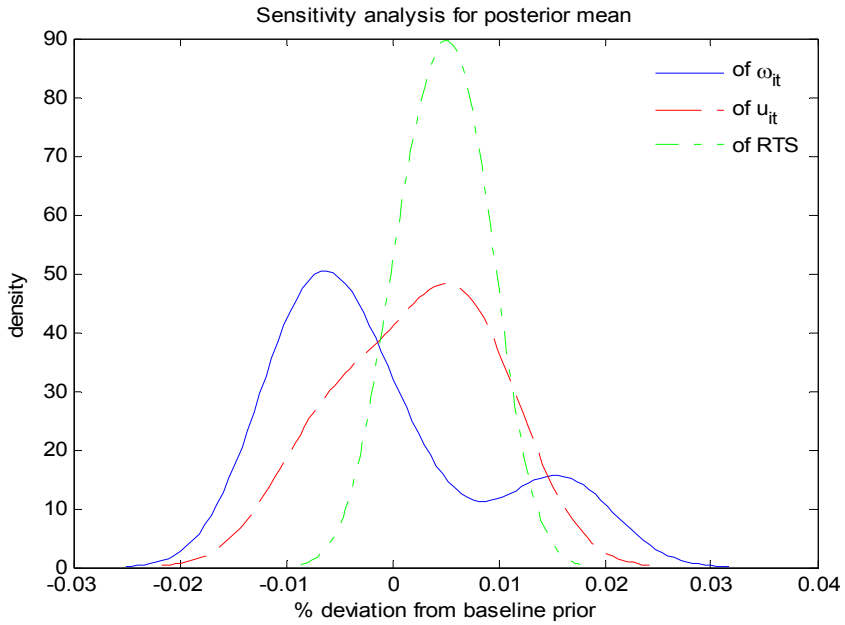


Figure 6. Posterior mean sensitivity analysis with respect to prior



Another question relates to whether we have, indeed, important measurement error in the model. The hypothesis, is, therefore, $H:\Omega = O_{K\times K}$ where $O_{K\times K}$ is a zero matrix.[8] We report the results across all 500 priors in Figure 7. Clearly, the hypothesis does not receive much support in the light of the data, unless we are willing to assume that there are no time effects (a hypothesis that has been previously rejected).

It is perhaps instructive to look at measurement error more closely. The estimated Φ matrix (estimates are posterior means) and posterior standard deviations are shown in Table 3. The eigenvalues at the posterior means are 0.4709 and 0.9516±0.163i suggesting that Φ corresponds to a stationary vector autoregressive process and there is substantial persistence in measurement errors.

Table 3. Posterior mean estimates of Φ and posterior standard deviations

	Labor	Capital	Intermediate inputs
Labor	0.772 (0.034)	-0.120 (0.037)	0.221 (0.015)
Capital	-0.225 (0.044)	0.813 (0.052)	-0.032 (0.013)
Intermediate inputs	0.152 (0.041)	0.331 (0.044)	0.789 (0.022)

The posterior density of maximum mod eigenvalue of Φ is reported in Figure 8. From the results it turns out that there is substantial persistence in measurement errors. Notice that the results in Table 3 do not have a structural or direct interpretation as they refer to a vector autoregressive model and attention should be focused on the eigenvalues of Φ, as we do here.

[8] Bayes factors are computed using the Verdinelli and Wasserman (1995) approach.

Figure 7. Log Bayes factors for $\Omega=O_{kxk}$ for different priors

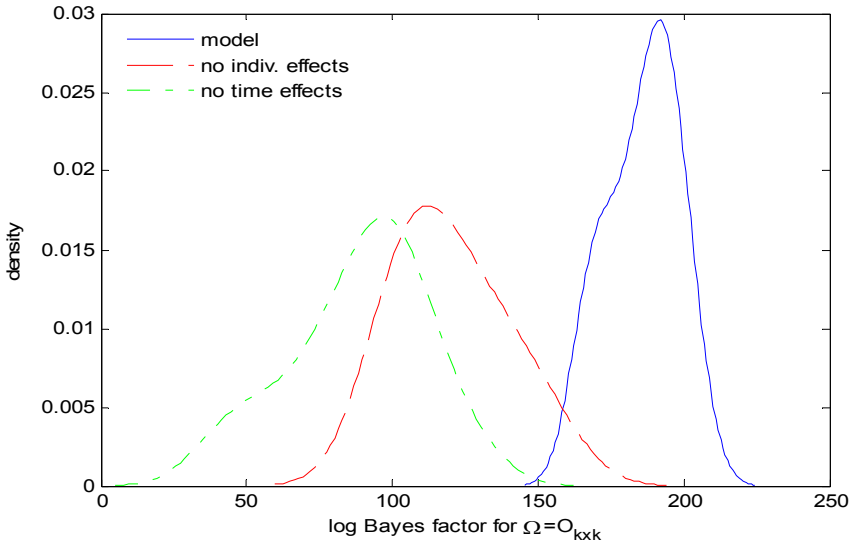
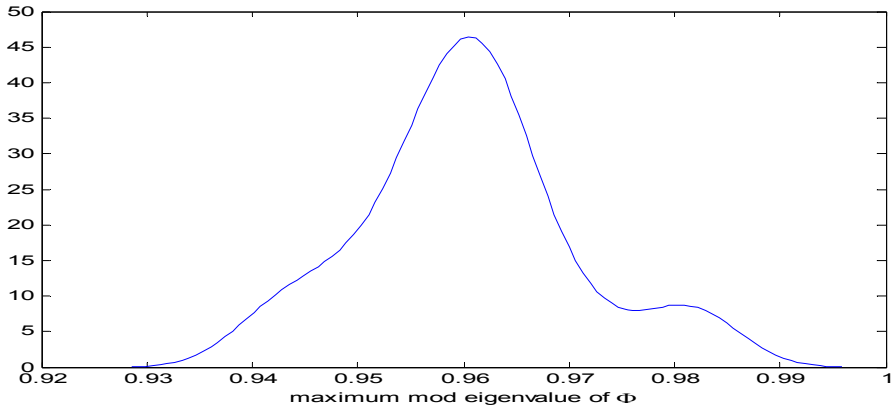


Figure 8. Posterior density of maximum mod eigenvalue of Φ



Apart from formal rejection of H based on Bayes factors, we wish to address the question of whether the presence of measurement error makes material economic difference in terms of returns to scale, technical inefficiency and productivity. The question is addressed in Figure 9. Clearly, ignoring measurement error yields drastically different estimates. For example, productivity is systematically higher and returns to scale are concentrated heavily around 0.70.

In Figure 10 we present productivity growth (averaged across industries) for the basic model (with 95% error bars) and also for Oley-Pakes and Levinsohn-Petrin. Clearly, we get different estimates. In Figure 10, we show the productivity growth over time, using

the functional relationship between w_{it} and $w_{i,t-1}$. Although there is evidence of catching up since 2008, the efficiency improvement seems to have been over-estimated by the conventional methods, which remains negligible with the approach undertaken in this study (see Figure 10).

Figure 9. Effect of measurement error

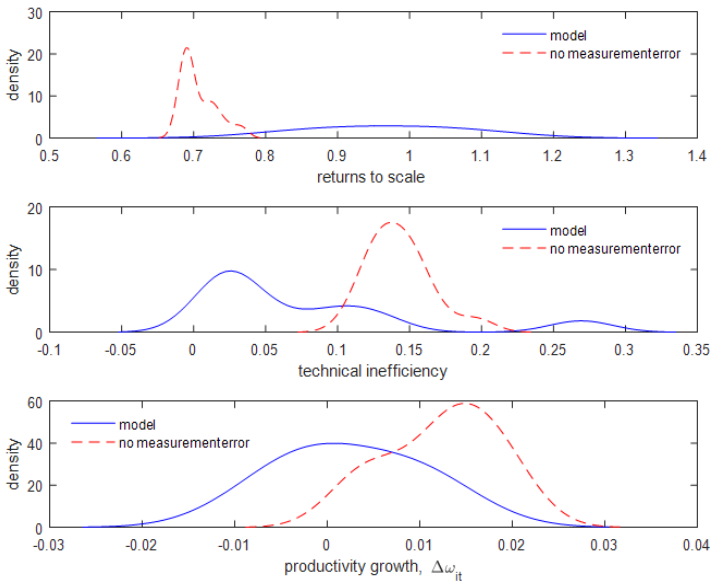
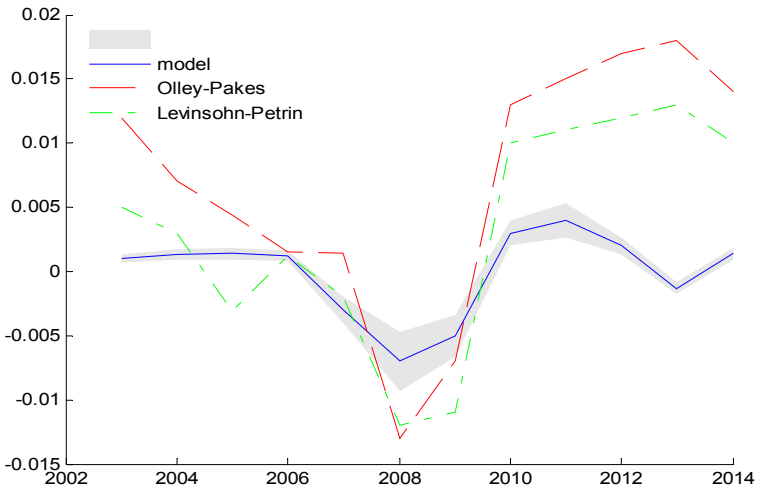


Figure 10. Productivity growth, $\Delta\omega_{it}$, over time



c. The question of a common translog production function and misspecification tests

Although a translog production function with individual effects is a fairly general representation of the technology, one may be right in raising the question of whether it is valid when firms from different sectors are pooled together. Let us write the translog production function in (13) as follows:

$$y_{it} = \beta_{oi} + z_{it}'\beta_i + v_{it,1} + \omega_{it} - u_{it}, \tag{34}$$

where $v_{it,1} \equiv v_{it}$ and we allow for possibly firm-specific coefficients. The translog coefficients do not admit a structural interpretation; the translog is simply a second-order approximation to an arbitrary production function. The issue we wish to address is whether the hypothesis $H: \beta_i = \beta, i = 1,...,n$. There are two ways to address this problem. First, we assume that:

$$\beta_i \sim N\left(\overline{\beta}, V_\beta\right), i = 1,...,n, \tag{35}$$

where $\overline{\beta}, V_\beta$ represent the prior mean and prior covariance matrix of the translog coefficients. Then, H is equivalent to evaluating the hypothesis $H: V_\beta = O$. The Bayes factor in favor of the hypothesis, using the Verdinelli and Wasserman (1995) approach, is 17.21 using our baseline prior. MCMC needs only minor modifications to apply under this random-coefficient structure.

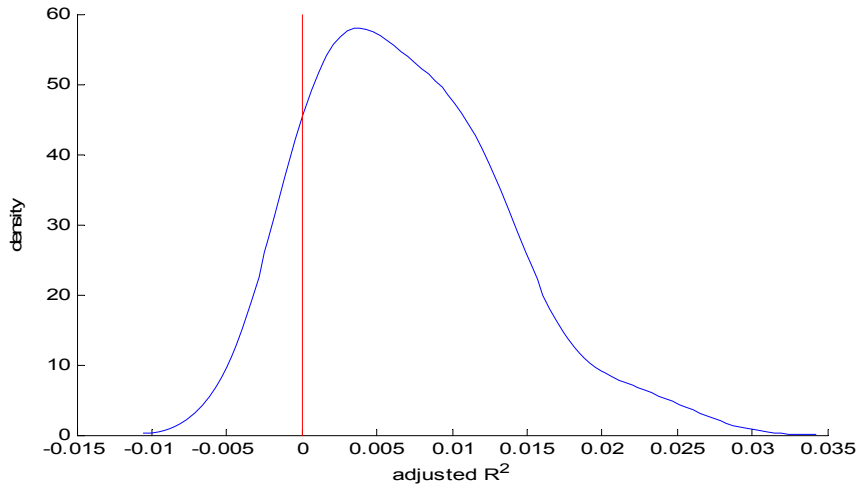
Another way to look at the problem is from the standpoint of model misspecification which addresses the more general issue of whether the translog induces misspecification. Following Ramsey's approach consider the model $y_{it} = \beta_{oi} + z_{it}'\beta + v_{it,1} + \omega_{it} - u_{it}$. At each MCMC iteration define $\hat{V}_{it}^{(s)} = y_{it} - \left(\beta_{oi}^{(s)} + z_{it}'\beta^{(s)} + \omega_{it}^{(s)} - u_{it}^{(s)}\right), s = 1,...,S$, and $\hat{y}_{it}^{(s)} = \beta_{oi}^{(s)} + z_{it}'\beta^{(s)} + \omega_{it}^{(s)} - u_{it}^{(s)}$. We fit the model:

$$\hat{V}_{it}^{(s)} = \sum_{j=2}^{J} \gamma_j^{(s)} \left(\hat{y}_{it}^{(s)}\right)^j, s = 1,...,S. \tag{36}$$

We choose J=4. Then the problem of misspecification boils down to testing whether the γ_j coefficients are zero. We save the adjusted coefficient of determination ($\overline{R}^2$) from these least squares regressions, say $\overline{R}^{2,(s)}$ and we present their posterior distribution in Figure 11. From the posterior distribution of $\overline{R}^2$ we do not obtain enough evidence in favor of misspecification. For the OP and LP specifications we can follow the same

approach after using GMM. The $\bar{R}^2$ was 0.314 and 0.303 for OP and LP, respectively. This, of course, indicates some form of misspecification.

Figure 11. Posterior distribution of adjusted R^2 for misspecification of translog



In the same spirit, we can examine residuals from the first order conditions in (13). Define:

$$\hat{v}_{it,k}^{(s)} = f(x_{it};\beta^{(s)}) + \ln\left(\beta_k^{(s)} + \sum_{k'=1}^{K}\beta_{kk'}^{(s)}x_{k',it}\right) + \omega_{it}^{(s)} - w_{k,it}^{(s)} - x_{k,it}, k = 1,...,K. \quad (37)$$

We determine, again, $\bar{R}^{2,(s)}$ for each of these residuals when they are regressed on powers of $\hat{y}_{it}^{(s)}$. The posterior densities are reported in Figure 12. Again, we obtain no evidence of misspecification.

For the productivity equation we examine possible misspecifications by using the residuals $\hat{r}_{it} = \omega_{it}^{(s)} - h\left(\omega_{i,t-1}^{(s)};\alpha^{(s)}\right)$ which are available for each MCMC iteration. We use two tests for misspecification. The first is based on the $\bar{R}^2$ from the model:

$$\hat{r}_{it}^{(s)} = \sum_{j=2}^{J}\gamma_j\left(\tilde{y}_{it}^{(s)}\right)^j, \quad (38)$$

where $\tilde{y}_{it}^{(s)} = \beta_{oi}^{(s)} + z_{it}'\beta^{(s)} - u_{it}^{(s)}$. The second is based on the model:

$$\sum_{j=2}^{J}\gamma_{j,1}\left(\tilde{y}_{it}^{(s)}\right)^j + \sum_{j'=1}^{J'}\gamma_{j',2}\left(\omega_{i,t-1}^{(s)}\right)^{j'}. \quad (39)$$

In the second model, we examine whether productivity depends on powers of the fitted values of output as well as on powers of lagged productivity. We choose, again,

$P = P' = 4$ and the results are presented in Figure 13. There is no evidence that the productivity equation is mis-specified. However, if we use G=2 then the $\overline{R}^2$ rises sharply to about 0.70 which implies that a "correct" selection of the value of G is essential.

Figure 12. Posterior distribution of adjusted R^2 for misspecification of first order conditions

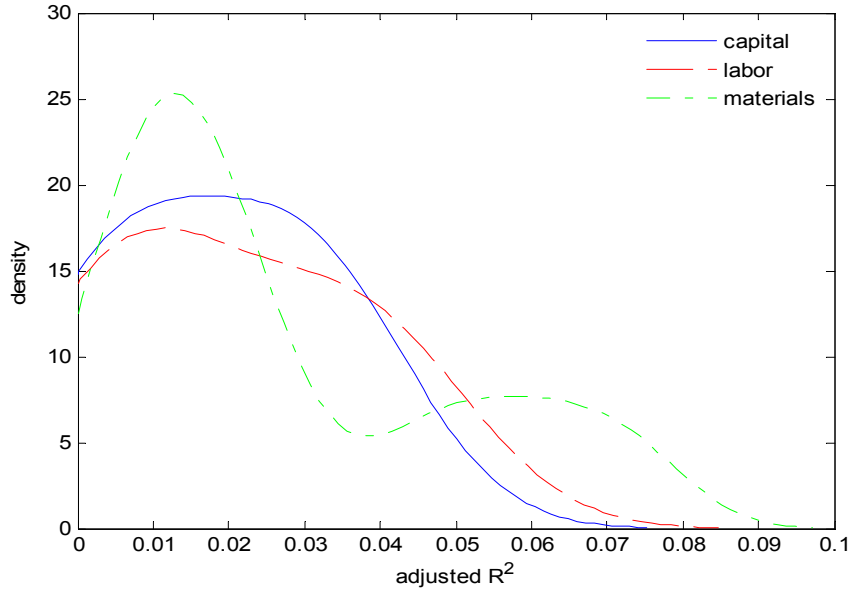
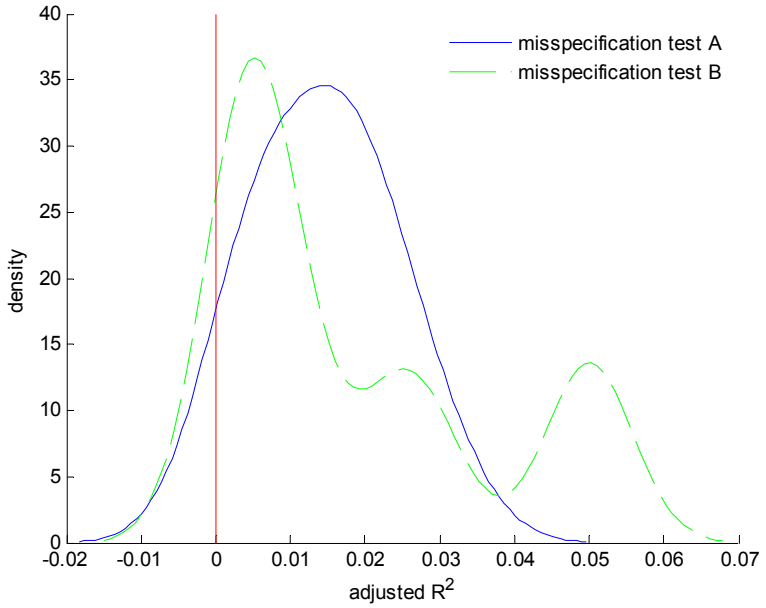


Figure 13. Misspecification tests (methods A and B)



In Figures 14 we report the sampling distributions of posterior means of productivity growth from the models with constant and firm-specific coefficients and, in Figure 15 we report the posterior means of functional forms $h(\omega_{it};\alpha)$ with constant and firm-specific coefficients. There is no evidence that the sample distributions of ω_{it} or the average (posterior mean) functional form are very different in the two cases so, at first, the introduction of firm-specific productivity growth does not seem to matter much. Finally, for further corroboration, we report the $h(\omega_{it};\alpha_i)$ functions for 50 randomly selected firms in Figure 16.

Figure 14. Sample distributions of productivity growth from models with constant and firm-specific α parameters

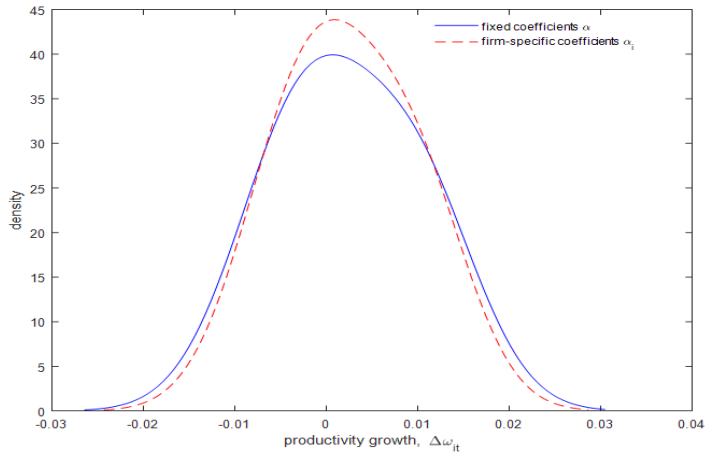


Figure 15. Posterior means of function h (ω$_{it}$; α)

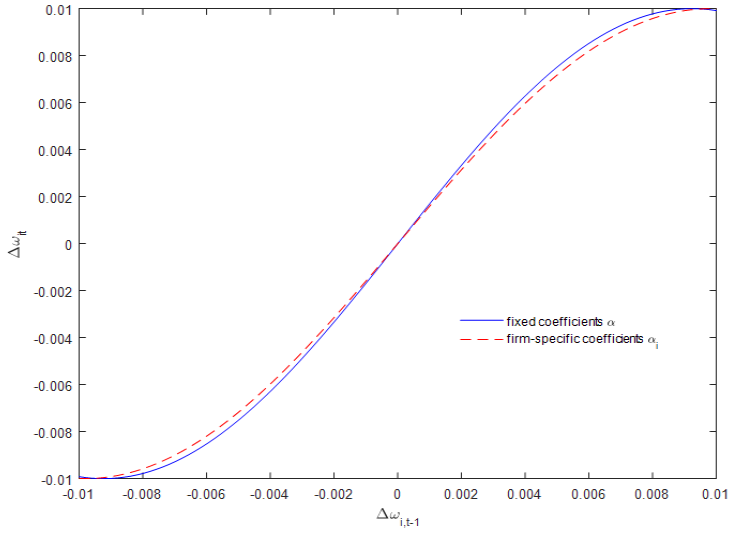
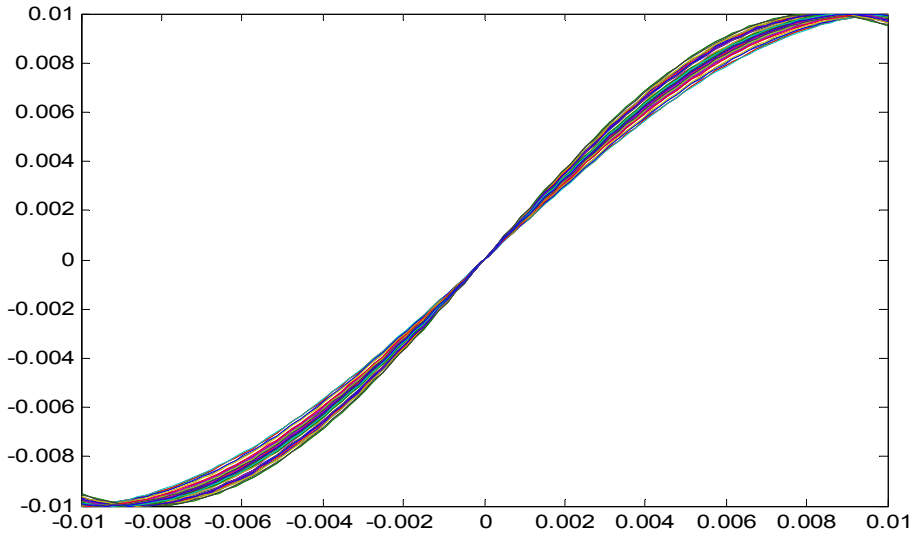


Figure 16. Different posterior means of function h(ω_{it};α) for randomly selected firms



d. Distributional and further robustness issues

Among the (possibly critical) assumptions of the model we list the following:

i) The distribution of $(K+1)\times 1$ vector V_{it} is multivariate normal.

ii) The distribution of error term ε_{it} in the productivity equation is normal.

iii) Lagged productivity $\omega_{i,t-1}$ and ε_{it} are orthogonal / independent.

iv) Lagged productivity $\omega_{i,t-1}$ and $v_{it} \equiv v_{it,1}$ are independent.

v) $(v_{it}, \varepsilon_{it})$ do not depend on x_{it}.

It is possible to provide estimates of certain quantities for each MCMC draw, say $V_{it}^{(s)}, v_{it}^{(s)}, \varepsilon_{it}^{(s)}, \omega_{i,t-1}^{(s)}$ and use various techniques to evaluate whether these assumptions hold, at least approximately. To evaluate (iii) we use a vector autoregression (VAR) model of the form:

$$\psi_{it}^{(s)} \equiv \begin{bmatrix} \varepsilon_{it}^{(s)} \\ \omega_{i,t-1}^{(s)} \end{bmatrix} = \sum_{l=1}^{L} A_l \psi_{i,t-l}^{(s)} + \varsigma_{it}^{(s)},$$

where $\varsigma_{it}^{(s)}$ is an error term and A_l is a 2×2 matrix of coefficients, $A_l = \begin{bmatrix} a_{l,11} & a_{l,12} \\ a_{l,21} & a_{l,22} \end{bmatrix}, l = 1,...,L.$ For each MCMC iteration (s) we determine L using the BIC and we record the p-value of the *F*-statistic for testing $H: a_{l,12} = a_{l,21} = 0, l = 1,...,L$. A similar construction is used to evaluate (iv) and (v) –although in (v) the dimensionality of the VAR is somewhat larger as it involves $K+2$ variables for each MCMC iteration.

Hypotheses (i) and (ii) are more cumbersome to evaluate as they require, for example, abandoning normality in favorite of, say, a mixture-of-normals and testing normality within a parametric framework. This complicates the MCMC procedure. Instead we evaluate (ii) as follows. Given draws $\varepsilon_{it}^{(s)}$ we use the "scores" $\tilde{e}_{it}^{(s)} = \Phi^{-1}(\tilde{\varepsilon}_{it}^{(s)})$ where Φ denotes the standard normal distribution function and $\tilde{\varepsilon}_{it}^{(s)}$ denotes that $\varepsilon_{it}^{(s)}$ have been standardized to mean zero and unit standard deviation. If normality is acceptable the scores $\tilde{e}_{it}^{(s)}$ must be approximately uniformly distributed for each MCMC iteration.

Similarly, to evaluate (i) we standardize $V_{it}^{(s)}$ to $\tilde{V}_{it}^{(s)} = \Phi_{K+1}^{-1}\left(C_{\Sigma}^{\prime(s)}(v_{it}^{(s)} - \frac{1}{nT}\sum_{i,t} v_{it}^{(s)}) \right)$, where $C_{\Sigma}^{\prime(s)} C_{\Sigma}^{(s)} = \Sigma$ and Φ_{K+1} denotes the standard multivariate normal distribution function in $\mathbb{R}^{K+1}$. Again, if normality is acceptable the scores $\tilde{v}_{it}^{(s)}$ must be approximately uniformly distributed, for each MCMC iteration. We summarize our evidence in Table 4. All uniformity tests use the Anderson-Darling statistic.

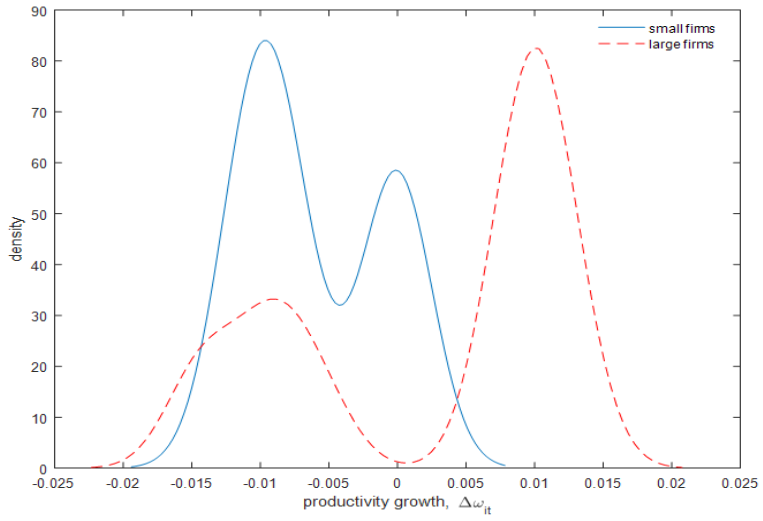
Table 4. Assessing distributional and other robustness issues

assumption	description	median p-value	Percentage of MCMC draws where p<0.01
(i)	V_{it} is multivariate normal	0.203	2.3%
(ii)	ε_{it} is normal	0.181	1.7%
(iii)	independent $\omega_{i,t-1}$ and ε_{it}	0.312	4.4%
(iv)	independent $\omega_{i,t-1}$ and v_{it}	0.144	3.02%
(v)	$(v_{it},\varepsilon_{it})$ do not depend on x_{it}	0.330	2.5%

From the evidence in Table 4, it seems that none of the assumptions (i) – (v) is not justified in the light of the data. Particularly convincing is, of course, the last column which reports the percentage of MCMC draws where the p-value is less than 0.01.

As technological progress displays heterogeneity at industry level (see Mallick and Sousa, 2017), firms at different quantiles of productivity improvements could converge differently as shown in Figure 17, where there is very clear evidence of multiple equilibria with multiple clusters of small and large firms at both ends of the distribution. While small firms can have the advantage of managerial efficiency, larger firms can have the advantages of economies of scale and better access to finance, thereby experiencing relatively higher productivity as in Figure 17.

Figure 17: Bimodal distribution of productivity growth for small and large firms



Also Figure 17 reveals that small firms at the top quintile perform better than the large firms at the bottom end of the distribution, because these small firms at the top-end can be in the growing industries or at a catching up stage. This also implies that there are inefficient firms at low ends of the size distribution which are making productivity growth stagnant, although large firms on average are more productive than small firms.

6. Concluding remarks

Given the methodological challenges in estimating total factor productivity at firm level, this paper adopts a new semi-parametric framework in estimating and decomposing TFP growth into technical efficiency change (or 'catching up'), and a technological progress (or 'frontier shift'). With a new model as a benchmark using Indian firm-level data, the

paper shows the superiority of our approach in capturing the temporal evolution of TFP growth. We show that the model can be estimated using SMC techniques, and a battery of specification tests can be provided.

Using the new approach and firm-level data from India to address the endogeneity of regressors in a production function, our results reveal that the new approach best describes our data in estimating productivity. Our results are indicative of the inappropriateness of deriving TFP estimates in the presence of endogenous regressors and a linear functional form. In fact, differentiating firms by size reveals that the TFP estimates are much smaller for low-end medium and large firms and even negative for these firms, whereas top-end smaller firms tend to exhibit higher productivity.

For most part of the sample, TFP growth has remained stagnant during the post-reform period in India with little significant differences observed across three different types of firms, although technical efficiency or catching-up effect appears to have driven TFP growth in the recent years. This suggests that lack of technological progress is indeed a cause of concern and therefore policy shift towards greater innovation should be prioritised in enhancing productivity. Applications of the same methodology to other countries and the resulting comparison with earlier studies can be pursued in future research.

References

Ackerberg, D. A., Caves, K. and Frazer, G. (2015). "Identification Properties of Recent Production Function Estimators". Econometrica, 83: 2411-2451.

Aigner, D.J., C.A.K. Lovell and P. Schmidt (1977). "Formulation and estimation of stochastic frontier production models." *Journal of Econometrics*, 6 (1), 21-27.

Altug, S. and Miller, R. A. (1998). The Effect of Work Experience on Female Wages and Labour Supply. *Review of Economic Studies* 65 (1), 48-85.

Amsler, C., A. Prokhorov and P. Schmidt (2014). "Using copula to model time dependence in stochastic frontier models." *Econometric Reviews*, 33(5-6), 497-522.

Amsler, C., A. Prokhorov and P. Schmidt (2016), "Endogeneity in stochastic frontier models." *Journal of Econometrics,* 190 (2): 280-288.

Battese. G.E. and T.J. Coelli (1995), "A model for technical inefficiency effects in a stochastic frontier production function for panel data." *Empirical Economics*, 20, 325-332.

Bournakis, I., and S. Mallick (2018) TFP estimation at firm level: The fiscal aspect of productivity convergence in the UK, *Economic Modelling*, 70: 579–590.

Caudill, S.B., J.M. Ford and D.M. Gropper (1993). "Frontier estimation and firm-specific inefficiency measures in the presence of heteroskedasticity." *Journal of Business & Economic Statistics*, 13, 105-111.

Chopin, N., Singh, S.S. (2013). On the particle Gibbs sampler. Working paper, ENSAE. http://arxiv.org/abs/1304.1887.

Cornwell, C., P. Schmidt and R.C. Sickles (1990). "Production frontiers with cross-sectional and time-series variation in efficiency levels". *Journal of Econometrics* 46, 185-200.

Creal, D.D., and R. Tsay (2015). "High dimensional dynamic stochastic copula models". *Journal of Econometrics* 189, 335-345.

Doraszelski, U., and J. Jaumandreu (2013), R&D and Productivity: Estimating Endogenous Productivity, *Review of Economic Studies*, 80, 1338 - 1383.

Ferrara, Giancarlo, and Francesco Vidoli (2017), Semiparametric stochastic frontier models: A generalized additive model approach, *European Journal of Operational Research*, 258 (2): 761-777.

Fethi, M.D., and F. Pasiouras (2010), Assessing bank efficiency and performance with operational research and artificial intelligence techniques: A survey, *European Journal of Operational Research*, 204 (2): 189-198.

Fukuyama, H., and R. Matousek (2017), Modelling bank performance: A network DEA approach, *European Journal of Operational Research*, 259 (2): 721-732.

Gandhi, Amit, Salvador Navarro, and David A. Rivers (2017), On the Identification of Gross Output Production Functions, Mimeo.

Geweke, J. (1992), "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments." In Bayesian Statistics 4 (eds. J.M. Bernardo, J. Berger, A.P. Dawid and A.F.M. Smith), Oxford: Oxford University Press, 169-193.

Godsill, S.J., Doucet, A., West, M., 2004. Monte Carlo smoothing for nonlinear time series. J. Amer. Statist. Assoc. 99 (465), 156–168.

Haidar, J.I. (2012) Trade and productivity: Self-selection or learning-by-exporting in India, *Economic Modelling*, 29 (5): 1766-1773.

Hausman, J.A., W.K. Newey, and J.L. Powell (1995), "Nonlinear errors in variables: Estimation of some Engel curves." *Journal of Econometrics*, 65, 205-233.

Heckman, J.J. and B.E. Honore (1989), "The identifiability of the competing risks model." *Biometrika*, 76 (2), 325-330.

Kevork, I. S., J. Pange, P. Tzeremes, and N.G. Tzeremes (2017) Estimating Malmquist productivity indexes using probabilistic directional distances: An application to the European banking sector, *European Journal of Operational Research*, 261 (3): 1125-1140.

Kim, Kyoo il, Amil Petrin, and Suyong Song (2016), Estimating production functions with control functions when capital is measured with error, *Journal of Econometrics*, 190 (2): 267-279.

Kumbhakar, S.C. (1997). Modeling allocative inefficiency in a translog cost function and cost share equations: An exact relationship. *Journal of Econometrics*, 76 (1-2), 351-356.

Kumbhakar, S.C., Parmeter, C.F. and E.G. Tsionas (2013). "A zero inefficiency stochastic frontier model." *Journal of Econometrics*, 172, 66-76.

Kutlu, L. (2010), "Battese-Coelli estimator with endogenous regressors." *Economics Letters*, 109, 79-81.

Lee, L.F. (1983), "Generalized econometric models with selectivity." *Econometrica*, 61, 381-428.

Lewbel, A. (1997), "Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R&D." *Econometrica*, 65 (5), 1201-1213.

Levinsohn, J. & Petrin, A. (2003), Estimating production functions using inputs to control for unobservables, *The Review of Economic Studies*, 70 (2): 317-341.

Li, Q. and J. Racine, (2007). *Nonparametric Econometrics*. Princeton University Press, Princeton, NJ.

Mallick, S. K., and Sousa, R. M. (2017) The Skill Premium Effect of Technological Change: New Evidence from United States Manufacturing, *International Labour Review,* 156 (1): 113–131.

Mallick, S. and Y. Yang (2013), Productivity performance of export market entry and exit: Evidence from Indian firms, *Review of International Economics*, 21 (4): 809-824.

Matousek, R., and N.G. Tzeremes (2016) CEO compensation and bank efficiency: An application of conditional nonparametric frontiers, *European Journal of Operational Research*, 251 (1): 264-273.

Meeusen, W. and J. van den Broeck, (1997). "Efficiency estimation from Cobb-Douglas production functions with composed error." *International Economic Review*, 18 (2), 435-444.

Olley, S. & Pakes, A. (1996), The dynamics of productivity in the telecommunications equipment industry, *Econometrica*, 64, 1263-1297.

Park, S. and S. Gupta (2012), "Handling endogenous regressors by joint estimation using Copulas." *Marketing Science*, 31 (4), 567-586.

Pitt, M.K., N. Shephard (1999). Filtering via simulation based on auxiliary particle filters. *Journal of the American Statistical Association,* 94 (446): 590-599.

Sun, K., S.C. Kumbhakar, and R. Tveterås (2015), Productivity and efficiency estimation: A semiparametric stochastic cost frontier approach, *European Journal of Operational Research*, 245 (1): 194-202.

Tran, K.C. and E.G. Tsionas (2013), "GMM estimation of stochastic frontier models with endogenous regressors." *Economics Letters*, 118, 233-236.

Tran, K.C., and M.G. Tsionas (2016), Zero-inefficiency stochastic frontier models with varying mixing proportion: A semiparametric approach, *European Journal of Operational Research*, 249 (3): 1113-1123.

Tzeremes, Nickolaos G. (2015), Efficiency dynamics in Indian banking: A conditional directional distance approach, *European Journal of Operational Research*, 240 (3): 807-818.

Whiteley, N. (2010). Discussion on particle Markov chain Monte Carlo methods. J. R. Stat. Soc. Ser. B 72 (3), 306–307.

Verdinelli, I., and L. Wasserman (1995). "Computing Bayes factors using a generalization of the Savage-Dickey ratio". *Journal of the American Statistical Association* 90 (430), 614-618.