Convex Non-Parametric Least Squares, Causal Structures and Productivity

Mike G. Tsionas*

February 15, 2022

Abstract

In this paper we consider Convex Nonparametric Least Squares (CNLS) when productivity is introduced. In modern treatments of production function estimation, the issue has gained great importance as when productivity shocks are known to the producers, input choices are endogenous and estimators of production function parameters become inconsistent. As CNLS has excellent properties in terms of approximating arbitrary monotone concave functions, we use it, along with flexible formulations of productivity, to estimate inefficiency and productivity growth in Chilean manufacturing plants. Inefficiency and productivity dynamics are explored in some detail along with marginal effects of contextual variables on productivity growth, inputs, and output. Additionally, we examine the causal structure between inefficiency and productivity as well as model validity based on a causal deconfounding approach. Unlike the Cobb-Douglas and translog production functions, the CNLS system is found to admit a causal interpretation.

Key Words: Productivity and Efficiency; Production Functions; Convex Non-Parametric Least Squares; Causal Models; Deconfounding.

Acknowledgments: The author is indebted to three anonymous reviewers who provided insightful comments on an earlier version.

*Lancaster University Management School, LA1 4YX, U.K., m.tsionas@lancaster.ac.uk.

1 Introduction

Since Kuosmanen (2008) introduced Convex Nonparametric Least Squares (CNLS) to the literature of productive efficiency analysis, the tool has been proven to be a useful item in the arsenal of applied studies. Kuosmanen (2008) and Kuosmanen and Johnson (2010) have shown that Convex Non-Parametric Least Squares can be shown to avoid the functional form assumption of stochastic frontier analysis building on the same axioms as Data Envelopment Analysis (DEA), but it also takes into account noise. However, the computational complexity of the method is considerable and Lee et al. (2013) have shown that, in fact, there are more efficient computational techniques to implement CNLS through linear programming and smart reduction techniques.[1]

CNLS is important as it provides a non-parametric technique to recover the technology and technical inefficiency. Alternatives include versions of bootstrapped DEA which, however, do not deal explicitly with noise (Simar and Wilson, 2008, 2011; see also the work of Johnson and Kuosmanen, 2011 on a new one-stage semi-nonparametric estimator that combines the nonparametric DEA-style frontier with a regression model of the contextual variables, and Johnson and Kuosmanen, 2012 for the estimation of the effects of environmental variables). In this paper we equip the original model with a productivity term which has been found essential in the modern literature on production functions [Olley and Pakes (1996) and further developed by Ackerberg et al. (2007, 2015), Doraszelski and Jaumandreu (2013), Levinsohn and Petrin (2003), Petrin and Sivadasan (2013), and Wooldridge (2009)].

To the extent that CNLS and DEA are closely related, our contribution is the introduction of productivity which can be parametrized as a flexible, semi-parametric Markov process with contextual or environmental variables. The same variables can be used to determine marginal effects on inefficiency and, additionally, investigate whether inefficiency and productivity are persistent and measure their degrees of persistence. This does not involve modeling inefficiency explicitly so, complicated and parametric formulations are avoided. On the other hand, this leaves open the question of a causal or structural interpretation of the model, which we address explicitly using a deconfounding approach. This persistence can be understood in the context of path dependence. Path dependence is a property of causal structures and it is related to technological and other resource capabilities (Kasy, 2011; Tsekouras et al., 2016, 2017).

We provide a Bayesian interpretation of CNLS and show that the posterior can be analyzed easily using Markov Chain Monte Carlo and Sequential Monte Carlo methods. The methods are easy to apply, they have the same performance as CNLS and can, effectively, extend CNLS - like formulations in much larger samples than before. The new techniques are applied to sectors of Chilean manufacturing, a data set that has received considerable attention in the literature of estimating production functions with endogeneity due to the correlation between variable input decisions and productivity shocks. The issue has received attention in other contexts of production economics relying on instrumental variables (e.g. Soytas et al., 2019) or other techniques (Shaw et al., 2017; Nadkarni and Shenoy, 2001; Anderson and Vastag, 2004). Of course, the issue of the existence of a causal structure is more involved, as a causal model's predictive ability should not deteriorate as we observe changes in its, broadly defined, environment or context. Additionally, we provide a methodology to examine policy analysis questions in the context of deconfounding.[2]

2 Convex non-parametric least squares

Suppose $\mathbf{x} \in \mathbb{R}^K$ is an input vector and the production function is

$$y = \mathfrak{f}(\mathbf{x}) + \varepsilon, \tag{1}$$

for some continuous, non-decreasing, concave function $\mathfrak{f} : \mathbb{R}^K_+ \to \mathbb{R}_+$, where y is output[3] and ε is a random variable for which $\mathbb{E}(\varepsilon|\mathbf{x}) = 0$. Kuosmanen (2008) proposed the use of CNLS:

$$\begin{gathered} \min_{\alpha,\boldsymbol{\beta},\boldsymbol{\varepsilon}} : \textstyle\sum_{i=1}^n \varepsilon_i^2; \\ y_i = \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i + \varepsilon_i,\ \forall i = 1,\dots,n, \\ \alpha_i + \boldsymbol{\beta}_i'\mathbf{x}_i \leq \alpha_h + \boldsymbol{\beta}_h'\mathbf{x}_h,\ \forall h = 1,\dots,n,\ h \neq i = 1,\dots,n, \\ \boldsymbol{\beta}_i \geq 0,\ \forall i = 1,\dots,n. \end{gathered} \tag{2}$$

The primary purpose of CNLS is to generate a monotone concave non-parametric approximation of the production function. Its difference with DEA, is that DEA assumes $\varepsilon_i \leq 0$ and can be attributed to inefficiency. This assumption is, in itself, quite strong, as productivity could be also part of the error term.

[1] Despite the close connections between CNLS and DEA, CNLS estimates the average production function, not the frontier. However, the shape of the frontier must be exactly the same as that of the average production function.

[2] Coupling CNLS with productivity is consistent with the modern literature on production function estimate (see above). Moreover, to the best of my knowledge, no methods are available when CNLS contains a latent autoregressive variable (productivity, ω_{it}).

[3] Extension to multiple output production through input or output distance functions belongs to the same set of problems, and it is straightforward.

Lee et al. (2013) proposed the use of elementary Afriat's theorem (Afriat, 1967, 1972). If we sort the data, in some sense, $\mathbf{x}_1 \preceq \mathbf{x}_2 \preceq \cdots \preceq \mathbf{x}_n$, then[4] we must have

$$\alpha_i \geq \alpha_{i-1},\ \boldsymbol{\beta}_i \leq \boldsymbol{\beta}_{i-1}, \forall i = 2, \ldots, n. \tag{3}$$

This motivated Lee at al. (2014) to solve the following relaxation problem:

$$\begin{gathered} \min_{\alpha,\boldsymbol{\beta},\boldsymbol{\varepsilon}} : \sum_{i=1}^{n} \varepsilon_i^2; \\ y_i = \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_i + \varepsilon_i,\ \forall i = 1, \cdots, n, \\ \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_i \leq \alpha_{i+1} + \boldsymbol{\beta}_{i+1}' \mathbf{x}_{i+1},\ \forall i = 1, \cdots, n-1, \\ \boldsymbol{\beta}_i \geq 0,\ \forall i = 1, \cdots, n. \end{gathered} \tag{4}$$

From this problem, they generate an initial set of constraints which can be used to solve the full problem. Then the violated constraints can be added one by one or in terms of groups.

3 CNLS and productivity

The full potential of CNLS has not yet been realized as it does not deal with productivity, an issue that has been central at least since Marschak and Andrews (1944) as well as in the modern literature on production functions [Olley and Pakes (1996) and further developed by Ackerberg et al. (2007, 2015), Doraszelski and Jaumandreu (2013), Levinsohn and Petrin (2003), Petrin and Sivadasan (2013), and Wooldridge (2009)]. Under the assumption that a productivity shock is known to the producer, variable input decisions will depend on the productivity shock and, therefore, variable inputs and the error term in the production function will be correlated, creating an endogeneity problem. The issue is not exhausted in its statistical implications but it has wider implications as inefficiency is likely to be mismeasured in the absence of productivity, and endogeneity will distort estimates of both inefficiency and productivity.

To fix ideas, suppose $\boldsymbol{k}_{it} \in \mathbb{R}_+^K$ is a vector of quasi-fixed factors of production (for DMU i and date t, where $i = 1, \ldots, n$, and $t = 1, \ldots, T$), $\boldsymbol{x}_{it} \in \mathbb{R}_+^M$ is the vector of variable inputs, and $\boldsymbol{z}_{it} \in \mathbb{R}^{d_z}$ is a vector of predetermined or contextual / environmental variables. The production function can be written as

$$y_{it} = f(\boldsymbol{k}_{it}, \boldsymbol{x}_{it}; \beta_{it}) + v_{it,1} + \omega_{it}, \tag{5}$$

where y_{it} represents output, $\beta_{it} \in \mathbb{R}^{d_\beta}$ is a vector of parameters, $v_{it,1}$ is an unknown shock, and ω_{it} represents the known (to the DMU) productivity shock. As the variable inputs are chosen after the productivity shock has been realized, a general form of first-order conditions[5] has the form:

$$\boldsymbol{x}_{it} = \boldsymbol{\Phi}(\boldsymbol{k}_{it}, \omega_{it}; \boldsymbol{\gamma}) + \boldsymbol{v}_{it,2}, \tag{6}$$

where $\Phi : \mathbb{R}_+^K \times \to \mathbb{R}_+^M$ represents the first-order conditions, $\boldsymbol{\gamma}$ is a vector of parameters, and $\boldsymbol{v}_{it,2}$ is an error term supported in $\mathbb{R}^M$. This vector function is taken in the literature to be a flexible approximation to an unknown functional form. *It is quite important to point out that there is an error term $\boldsymbol{v}_{it,2}$ in (6) and, more importantly perhaps, that this formulation allows for multiple variable inputs.* This error term is omitted in modern treatments of production functions (e.g. Olley and Pakes, 1996 and Levinsohn and Petrin, 2003) as the objective is to invert the function in terms of productivity, ω_{it}, and only one variable input is taken into account in (6).

The unknown functional forms in (5) can be represented using the method of sieves[6]:

$$\begin{aligned} \Phi_1(\boldsymbol{k}_{it}, \omega_{it}; \boldsymbol{\gamma}) &= \sum_{i_1=0}^{P} \cdots \sum_{i_K=0}^{P} \sum_{i_{K+1}=0}^{P} k_{it,1}^{i_1} \ldots k_{it,K}^{i_K} \omega_{it}^{i_{K+1}} \gamma_{i_1 \cdots i_K i_{K+1}}^{(1)}, \\ &\vdots \\ \Phi_M(\boldsymbol{k}_{it}, \omega_{it}; \boldsymbol{\gamma}) &= \sum_{i_1=0}^{P} \cdots \sum_{i_K=0}^{P} \sum_{i_{K+1}=0}^{P} k_{it,1}^{i_1} \ldots k_{it,K}^{i_K} \omega_{it}^{i_{K+1}} \gamma_{i_1 \cdots i_K i_{K+1}}^{(M)}, \end{aligned} \tag{7}$$

where P is the order of polynomials,[7] $\gamma_{i_1 \cdots i_K i_{K+1}}^{(1)}, \ldots, \gamma_{i_1 \cdots i_K i_{K+1}}^{(M)}$ are unknown coefficients which we collect in vector $\boldsymbol{\gamma} \in \mathbb{R}^{d_\gamma}$, and

$$\boldsymbol{v}_{it,2} = [v_{it,2,1}, \ldots, v_{it,2,M}]'.$$

[4] Here we use the Euclidean norm: $\mathbf{x} \preceq \mathbf{y} \Leftrightarrow ||\mathbf{x}|| \leq ||\mathbf{y}||$.

[5] These first-order conditions are compatible with cost minimization when $y_{it} - \omega_{it}$ is taken as given and productivity shocks are known to the producer. An alternative is to use multiplicative error terms as in subsection 6.2 of Kuosmanen and Kortelainen (2007, 2012).

[6] A reviewer suggested the use of local likelihood techniques. Although this is greatly complicated by the presence of a dynamic latent variable (ω_{it}), it is worth exploring in future research.

[7] In the sieve expansions, similar terms are omitted. Moreover, one could have different polynomial orders. For our preferred model we report Bayes factors (or posterior odds ratios in Table A.1 of Appendix A).

The introduction of error terms, $\boldsymbol{v}_{it,2}$, in (6) appears to be novel. In the case of one variable input, the literature so far has used (6) to solve for ω_{it} as a function of $\boldsymbol{k}_{it}$ and x_{it} (the single element of $\boldsymbol{x}_{it}$), substitute the resulting repression in (5), and, in turn, use a combination of least-squares and generalized-method-of-moments estimators to recover consistent estimators of parameter vector $\boldsymbol{\beta}_{it} = \boldsymbol{\beta} \in \mathbb{R}^{K+M}$ which is assumed to be the same for all DMUs, time-invariant, and corresponding to a Cobb-Douglas production function. Of course, these are highly restrictive assumptions which limit the scope not only of production function estimation, but also estimation of productivity and productivity growth.
From the point of view of CNLS, inputs are assumed to be exogenous, and productivity is ignored. For the development of these techniques these simplifying assumptions were, of course, quite natural, but in the presence of productivity we have to incorporate the facts that (i) variable inputs become correlated with productivity, and (ii) productivity itself becomes of independent interest. To complete the system of (5) and (6), we assume a productivity process of the form

$$\omega_{it} = \Phi_{\omega}(\omega_{i,t-1}, \boldsymbol{z}_{it}; \boldsymbol{\delta}) + v_{it,3} = \sum_{i_1=0}^{Q} \sum_{i_2=0}^{Q} \cdots \sum_{i_{d_z}=0}^{Q} \omega_{i,t-1}^{i_1} z_{it,1}^{i_2} \cdots z_{it,d_z}^{i_{d_z}} \delta_{i_1 i_2 \ldots i_{d_z}} + v_{it,3}, \tag{8}$$

where the $\delta_{i_1 i_2 \ldots i_{d_z}}$s are unknown parameters, collected in vector $\delta \in \mathbb{R}^{d_\delta}$, Q denotes the order of the polynomials, and $v_{it,3}$ is an error term. This is a sieve approximation to an unknown Markov process for $\{\omega_{it}\}$ including the forcing variables in $\boldsymbol{z}_{it}$. Moreover, Q is the order of the polynomial (for estimates see, e.g., Table A.1).
Clearly, the production function in (5) can be written as

$$y_{it} = \alpha_{it} + \boldsymbol{k}_{it}' \boldsymbol{\beta}_{k,it} + \boldsymbol{x}_{it}' \boldsymbol{\beta}_{x,it} + v_{it,1} + \omega_{it}. \tag{9}$$

As we allow for general DMU-specific and time-varying coefficients, this can approximate any functional form. In (5) we have a general functional form $f(\boldsymbol{k}_{it}, \boldsymbol{x}_{it})$ where we omit β_{it} for the sake of presentation. If we write

$$f(\boldsymbol{k}_{it}, \boldsymbol{x}_{it}) = \alpha_{it} + \boldsymbol{k}_{it}' \boldsymbol{\beta}_{k,it} + \boldsymbol{x}_{it}' \boldsymbol{\beta}_{x,it}, \tag{10}$$

then, this is an *identity* if we do *not* make any assumptions about $\boldsymbol{\beta}_{k,it}$ and $\boldsymbol{\beta}_{x,it}$. However, to make this useful it is natural to make further assumptions about $\boldsymbol{\beta}_{k,it}$ and $\boldsymbol{\beta}_{x,it}$.

For simplicity, let $\boldsymbol{\chi}_{it} = [\boldsymbol{k}_{it}', \boldsymbol{x}_{it}']'$ and $\boldsymbol{\beta}_{it} = [\boldsymbol{\beta}_{k,it}', \boldsymbol{\beta}_{x,it}']'$ so that we can write (9) as follows.

$$y_{it} = \alpha_{it} + \boldsymbol{\chi}_{it}' \boldsymbol{\beta}_{it} + v_{it,1} + \omega_{it}. \tag{11}$$

In effect, the CNLS problem becomes:

$$\begin{gathered} y_{it} = \alpha_{it} + \boldsymbol{\chi}_{it}' \boldsymbol{\beta}_{it} + v_{it,1} + \omega_{it}, \\ \alpha_{it} + \boldsymbol{\chi}_{it}' \boldsymbol{\beta}_{it} \leq \alpha_{\iota\tau} + \boldsymbol{\chi}_{it}' \boldsymbol{\beta}_{\iota\tau} \; \forall (i,t) \neq (\iota,\tau), \\ \boldsymbol{\beta}_{it} \geq 0 \; \forall i,t, \\ \boldsymbol{x}_{it} = \boldsymbol{\Phi}(\boldsymbol{k}_{it}, \omega_{it}; \boldsymbol{\gamma}) + \boldsymbol{v}_{it,2}, \\ \omega_{it} = \Phi_{\omega}(\omega_{i,t-1}, \boldsymbol{z}_{it}; \boldsymbol{\delta}) + v_{it,3}. \end{gathered} \tag{12}$$

Part of modeling endogeneity is that the error terms, which we collectively denote

$$\boldsymbol{v}_{it} = [v_{it,1}, \boldsymbol{v}_{it,2}', v_{it,3}]',$$

cannot be independent, although as a working approximation it is reasonable to assume that

$$\boldsymbol{v}_{it} \sim \mathcal{N}_{M+2}\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right), \tag{13}$$

where $\boldsymbol{\Sigma}$ is a covariance matrix and $\mathcal{N}_{M+2}$ denotes the multivariate normal distribution in $\mathbb{R}^{M+2}$. To handle the constraints in the second equation of (12) define

$$d_{it,\iota\tau}(\boldsymbol{\beta}) = (\alpha_{it} + \boldsymbol{\chi}_{it}' \boldsymbol{\beta}_{it}) - (\alpha_{\iota\tau} + \boldsymbol{\chi}_{it}' \boldsymbol{\beta}_{\iota\tau}) \; \forall (i,t) \neq (\iota,\tau), \tag{14}$$

for which we must have

$$d_{it,\iota\tau}(\boldsymbol{\beta}) \leq 0 \, \forall (i,t) \neq (\iota,\tau). \tag{15}$$

Collecting all the constraints we can write them in vector form

$$\boldsymbol{d}(\boldsymbol{\beta}) \leq \mathbf{0}_{N(N-1)}, \tag{16}$$

where $N = nT$. Alternatively, we can write these constraints in stochastic form as follows:

$$\boldsymbol{d}(\boldsymbol{\beta}) = \boldsymbol{V} - \boldsymbol{U}, \tag{17}$$

where $\boldsymbol{V} \sim \mathcal{N}_{N(N-1)}\left(0,\, \sigma_V^2\mathbf{I}\right)$, and $\boldsymbol{U} \sim \mathcal{N}^{+}_{N(N-1)}\left(0,\, \sigma_U^2\mathbf{I}\right)$, where $\mathbf{I}$ is the identity matrix, $\mathcal{N}^{+}_{N(N-1)}$ denotes the multivariate half-normal distribution in $\mathbb{R}^{N(N-1)}$,and σ_V, σ_U are unknown parameters. As $\sigma_V \to 0$ we can impose exactly the constraints in (15). So, it may be sensible to impose a small value for σ_V (like 10^{-5}) and leave σ_U as a free parameter.[8]

4 Likelihood and posterior

4.1 General

Suppose all parameters are denoted

$$\boldsymbol{\theta} = \left[\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}', \boldsymbol{\delta}'\right]',$$

and suppose $\mathcal{Y}$ denotes the data. Based on our distributional assumptions the likelihood function, augmented with latent productivity $\boldsymbol{\omega} = \{\omega_{it}\}$ and the slack variables $\boldsymbol{U} = \{U_{it}\}$ is derived as follows.
Define

$$\boldsymbol{R}_{it}(\boldsymbol{\theta}, \boldsymbol{\omega}) = \begin{bmatrix} y_{it} - \alpha_{it} - \boldsymbol{\chi}_{it}'\boldsymbol{\beta}_{it} - \omega_{it} \\ \boldsymbol{x}_{it} - \boldsymbol{\Phi}(\boldsymbol{k}_{it}, \omega_{it}; \boldsymbol{\gamma}) \\ \omega_{it} - \Phi_{\omega}(\omega_{i,t-1}, \boldsymbol{z}_{it}; \boldsymbol{\delta}) \end{bmatrix}, \tag{18}$$

as "residuals" from the first three equations in (12)

$$\begin{aligned} \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{U}, \sigma_U, \boldsymbol{\Sigma}; \mathcal{Y}) \propto |\boldsymbol{\Sigma}|^{-N/2} \cdot \exp\left\{-\tfrac{1}{2}\textstyle\sum_{i=1}^{n}\sum_{t=1}^{T} \boldsymbol{R}_{it}(\boldsymbol{\theta}, \boldsymbol{\omega})'\boldsymbol{\Sigma}^{-1}\boldsymbol{R}_{it}(\boldsymbol{\theta}, \boldsymbol{\omega})\right\} \cdot \\ \sigma_U^{-N(N-1)/2} \cdot \exp\left\{-\tfrac{1}{2\sigma_V^2}\left(\boldsymbol{d}(\boldsymbol{\beta}) + \boldsymbol{U}\right)'\left(\boldsymbol{d}(\boldsymbol{\beta}) + \boldsymbol{U}\right) - \tfrac{1}{2\sigma_U^2}\boldsymbol{U}'\boldsymbol{U}\right\} \cdot \mathbb{I}(\boldsymbol{\beta} \geq \mathbf{0}), \end{aligned} \tag{19}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. The likelihood function as a function of the parameters alone, is

$$L(\boldsymbol{\theta}; \mathcal{Y}) = \int_{\mathbb{R}_{+}^{N(N-1)}} \int_{\mathbb{R}^N} \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{U}; \mathcal{Y})\, \mathrm{d}\boldsymbol{\omega}\, \mathrm{d}\boldsymbol{U}. \tag{20}$$

This involves multivariate integrals that cannot be computed in closed form, particularly in view of the fact that $\boldsymbol{\omega}$ is given by a dynamic, nonlinear latent variable model as in (8). Our strategy is to use Bayes' theorem in the augmented likelihood in (19) to obtain

$$p(\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{U}, \sigma_U, \boldsymbol{\Sigma}; \mathcal{Y}) \propto \mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{U}, \sigma_U, \boldsymbol{\Sigma}; \mathcal{Y}) \cdot p(\boldsymbol{\theta}, \sigma_U, \boldsymbol{\Sigma}),$$

where $p(\boldsymbol{\theta}, \sigma_U, \boldsymbol{\Sigma})$ is a prior on the parameters. In all instances, we treat the initial values $\omega_{i,0}$ as unknown with a flat prior.

4.2 Benchmark prior

For $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ we use a flat prior subject to the constraints $\boldsymbol{\beta} \geq \mathbf{0}$ as required by monotonicity, viz.

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \mathbb{I}(\boldsymbol{\beta} \geq \mathbf{0}). \tag{21}$$

Our prior for σ_U is (Zellner, 1971, p. 371)

$$p(\sigma_U) \propto \sigma_U^{-(\overline{N}+1)} e^{-\overline{Q}/(2\sigma_U^2)}, \tag{22}$$

where $\overline{N}, \overline{Q} \geq 0$ are prior hyperparameters. A proper but diffuse prior results if we select $\overline{N} = 1$ and $\overline{Q} = 10^{-5}$. For the different elements of $\boldsymbol{\Sigma}$ we use an inverted Wishart prior of the form (Zellner, 1971, p. 395)

$$p(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-[(M+2)+\overline{d}+1)/2} e^{-tr\left(\overline{\mathbf{A}}\boldsymbol{\Sigma}^{-1}\right)}, \tag{23}$$

where $\overline{d}$ is a scalar, $tr(\cdot)$ denotes the trace of a matrix, and $\overline{\mathbf{A}}$ is a matrix conformable with $\boldsymbol{\Sigma}$. A proper bit diffuse prior

[8] In practice, we start from $\sigma_V = 10^{-5}$ and we decrease this value until posterior moments are stabilized within the tolerance provided by the numerical standard error (Geweke, 1992).

results if we choose $\overline{d}=1$ and $\overline{\mathbf{A}}=h\cdot\mathbf{I}$, where $h=10^{-5}$. All parameters are a priori independent. Therefore, the posterior of the model is as follows.

$$\begin{gathered}p(\boldsymbol{\theta},\boldsymbol{\omega},\boldsymbol{U},\sigma_U,\boldsymbol{\Sigma};\mathcal{Y})\propto|\boldsymbol{\Sigma}|^{-(N+\overline{d}+1)/2}\cdot\exp\left\{-\tfrac{1}{2}\left[\overline{\mathbf{A}}+\mathbf{A}(\boldsymbol{\theta},\boldsymbol{\omega})\right]\Sigma^{-1}\right\}\cdot\\ \sigma_U^{-\left[N(N-1)+\overline{N}+1\right]/2}\cdot\exp\left\{-\tfrac{1}{2\sigma_V^2}\left(\boldsymbol{d}(\boldsymbol{\beta})+\boldsymbol{U}\right)'\left(\boldsymbol{d}(\boldsymbol{\beta})+\boldsymbol{U}\right)-\tfrac{1}{2\sigma_U^2}\left(\overline{Q}+\boldsymbol{U}'\boldsymbol{U}\right)\right\}\cdot\\ \mathbb{I}(\boldsymbol{\beta}\geq\mathbf{0})\cdot\mathbb{I}(\boldsymbol{U}\geq\mathbf{0}),\end{gathered}\tag{24}$$

where

$$\mathbf{A}(\theta,\boldsymbol{\omega})\equiv\sum_{i=1}^{n}\sum_{t=1}^{T}\boldsymbol{R}_{it}(\boldsymbol{\theta},\boldsymbol{\omega})\boldsymbol{R}_{it}(\boldsymbol{\theta},\boldsymbol{\omega})'.$$

To access the posterior we use Markov Chain Monte Carlo (MCMC) methods organized around Sequential Monte Carlo (SMC) also known as Particle Filtering (PF) to obtain draws $\left\{\boldsymbol{\omega}_{it}^{(s)},\ 1\leq s\leq S\right\}$, where S denotes the number of MCMC samples. Drawing $\left\{\boldsymbol{U}^{(s)},\ 1\leq s\leq S\right\}$ is much simpler and, finally, producing draws for $\left\{\boldsymbol{\alpha}^{(s)},\boldsymbol{\beta}^{(s)},\sigma_U^{(s)},\boldsymbol{\Sigma}^{(s)},\ 1\leq s\leq S\right\}$ is also relatively simple as we explain in Appendix A.

4.3 Restricted model for panel data

We call the model in the previous section, **Model I**. With panel data instead of (12) it may be of interest to use the following **Model II**:

$$\begin{gathered}y_{it}=\alpha_{it}+\boldsymbol{\chi}_{it}'\boldsymbol{\beta}_i+v_{it,1}+\omega_{it},\\ \alpha_i+\boldsymbol{\chi}_{it}'\boldsymbol{\beta}_i\leq\alpha_\iota+\boldsymbol{\chi}_{it}'\boldsymbol{\beta}_\iota\ \forall(i,t)\neq(\iota,\tau),\\ \boldsymbol{\beta}_i\geq 0\ \forall i,\\ \boldsymbol{x}_{it}=\boldsymbol{\Phi}(\boldsymbol{k}_{it},\omega_{it};\boldsymbol{\gamma})+\boldsymbol{v}_{it,2},\\ \omega_{it}=\Phi_\omega(\omega_{i,t-1},\boldsymbol{z}_{it};\boldsymbol{\delta})+v_{it,3}.\end{gathered}\tag{25}$$

In this model we restrict the coefficients α_i and β_i are assumed to be time-invariant. Moreover, in **Model III** we replace the last equation with the following parametric alternative

$$\omega_{it}=\delta_0+\delta_1\omega_{i,t-1}+\boldsymbol{z}_{it}'\boldsymbol{\delta}_2+v_{it,3},\tag{26}$$

in the interest of simplifying the dynamics. Whether such simplification is empirically possible is, of course, an empirical matter. Relative to (12), the formulation in (25) loses the DEA interpretation and is closer to econometric formulations for panel data. Finally, we define **Model IV**, as a model where we omit the productivity component altogether from (25). This model is fully equivalent to the model of Johnson and Kuosmanen (2012).[9] As the number of parameters in (12) is large, we use a compression technique inspired by Guhaniyogi and Dunson (2015); see subsection A.4 for the technical details. Their compression technique, however, was applied to regressors in linear models with conjugate priors (so that the marginal likelihood is available in closed form) rather than to parameters of nonlinear models with dynamic latent variables.

5 Simulation evidence

5.1 Without productivity

For our simulation experiments we use the same specifications as in Kuosmanen (2008):
Cobb-Douglas: $f^{CD}(\mathbf{x})=x_1^{0.4}x_2^{0.5}$.

Generalized Leontief: $f^{GL}(\mathbf{x})=(0.2x_1^{0.5}+0.3x_2^{0.5}+0.4x_1^{0.5}x_2^{0.5})^{0.9}$.

Piece-wise linear: $f^{PWL}(\mathbf{x})=\min\{x_1+2x_2,2x_1+x_2,0.5x_1+x_2+225,x_1+0.5x_2+225\}$.

The values x_1 and x_2 were independently and randomly sampled from the uniform distribution $\mathscr{U}[100,200]$, and random

[9] For Models II, III, and IV we use the same priors as for Model I. The reader may have noticed the absence of input and output relative prices from our specifications and particularly in (25). Since there are quasi-fixed inputs, obviously, the first-order conditions should include these quasi-fixed factors and also relative prices. Unfortunately, more often than not price information is lacking, which is also the approach taken in Olley and Pakes (1996) and further developed by Ackerberg et al. (2007, 2015), Doraszelski and Jaumandreu (2013), Levinsohn and Petrin (2003), Petrin and Sivadasan (2013), and Wooldridge (2009) as well as Gnadhi et al. (2020). Of course, if price information is available, it should be included in the fourth equation of (25).

Table 1: Simulation evidence

		MSE (s.d. in parentheses)		
True function	Std. Dev.,σ_ε	CNLS$^{(a)}$	CD$^{(a)}$	Bayes
CD	2.5	0.79	0.22	0.70
		(0.34)	(0.16)	(0.29)
	5	2.71	0.89	2.70
		(1.23)	(0.64)	(1.20)
	10	9.48	3.99	9.45
		(4.47)	(2.97)	(4.50)
GL	2.5	0.66	7.58	0.65
		(0.30)	(1.04)	(0.31)
	5	2.33	38.46	2.31
		(1.10)	(8.97)	(1.12)
	10	8.19	46821	8.20
		(4.02)	(40176)	(4.01)
PWL	2.5	1.65	58.94	1.65
		(0.50)	(8.04)	(0.51)
	5	5.58	59.79	5.50
		(1.80)	(8.20)	(1.84)
	10	17.86	62.83	17.85
		(6.20)	(9.00)	(6.22)

Notes: Standard deviations in parentheses. $^{(a)}$ Taken from Table 2 of Kuosmanen (2008).

normal noise was added using $\sigma_\varepsilon = 2.5, 5$ and 10. Therefore, we have nine scenarios in total, and the number of observations is $N = 100$. We use 15,000 MCMC iterations the first 5,000 of which are discarded to mitigate possible start up effects.[10] We check convergence using Geweke's (1992) diagnostic. We start by treating σ_V as a parameter with the same prior as σ_U. In turn, if $\overline{\sigma}_V$ is the posterior mean, we set σ_V to a fixed value $\overline{\sigma}_V/b$ for $b = 2, 4, ...$ We stop when the results (in terms of posterior means of $\mathbf{U}$ converge to within 10% of maximum numerical standard error (NSE), see Geweke (1992). We have used 10,000 simulations (Kuosmanen, 2008 used 250 simulations) and the results are summarized in Table 1. As a benchmark to compare the performance of Bayesian approach, we use the results of CNLS from Kuosmanen (2008) and also his Cobb-Douglas fit (CD, Kuosmanen, 2008, Table 2).

From the results it turns out that the mean squared error (MSE) of fit is strikingly similar to Kuosmanen (2008) and the difference can, probably, be attributed to the larger number of simulations we used. This means that the Bayesian formulation of the problem, effectively, recovers the CNLS results, which are much harder to compute.

5.2 With productivity

We retain the true functional forms in the previous subsection but now we add productivity in the form

$$\omega_{it} = \rho\omega_{i,t-1} + \xi^{\omega}_{it}, \tag{27}$$

where $\rho = 0.5$ is an autoregressive parameter, $\xi^{\omega}_{it} \sim \mathcal{N}(0, \sigma^2_{\xi^\omega})$, and to maintain the total number of observations at $N = 100$ we set $n = 20$ and $T = 5$. We keep σ_ε the same as in the previous subsection and we set $\sigma^2_{\xi^\omega} = \frac{1}{2}\sigma^2_\varepsilon$. We use $P = Q = 2$ in (7) and (8) but results were nearly the same with $P = Q = 1$. The Monte Carlo results are reported in Table 2.

For the functional forms, the MSE results are nearly the same. Given the small sample size (especially in the time dimension) the correlations between actual and estimated productivity remain high even when $\sigma_\varepsilon = 10$. A possible reason is that part of the error in subsection Table 1, becomes signal in the form of (27) in Table 2.

[10] If convergence is rejected we take another 5,000 iterations and so on until we are unable to reject. Usually, 5,000 to 15,000 iterations were found enough. All computations are performed using Fortran 77 at High End Computing Cluster of Lancaster University. The cluster operating system is Scientific Linux, with job submission handled by Sun Grid Engine (SGE).

Table 2: Simulation evidence

True function	Std. Dev.,σ_ε	Bayes	correlation of actual and estimated productivity (ω_{it})
CD	2.5	0.71	0.97
		(0.28)	(0.04)
	5	2.70	0.93
		(1.22)	(0.05)
	10	9.44	0.90
		(4.52)	(0.07)
GL	2.5	0.65	0.98
		(0.30)	(0.03)
	5	2.30	0.95
		(1.12)	(0.04)
	10	8.21	0.90
		(4.00)	(0.10)
PWL	2.5	1.65	0.98
		(0.51)	(0.03)
	5	5.51	0.95
		(1.83)	(0.05)
	10	17.84	0.90
		(6.25)	(0.09)

Notes: Standard deviations in parentheses. The results corresponding to translog are omitted in the interest of saving space. To maintain the total number of observations at $N = 100$ we set $n = 20$ and $T = 5$. We keep σ_ε the same as in the previous subsection and we set $\sigma^2_{\xi^\omega} = \frac{1}{2}\sigma^2_\varepsilon$.

6 Empirical application

We use the data from Instituto Nacional de Estadistica which covers all Chilean manufacturing plants with more than ten employees during 1979 - 1996. These data have been used in Levinsohn and Petrin (2003), Gandhi et al. (2020), and Ackerberg et al. (2015). For each plant in the sample, the data include the age of firms, gross output, material inputs, capital stock and investments, fuels and electricity, and labor (measured in person-years, skilled as well as unskilled) converted where necessary into real values using industry-specific price deflators. A more detailed description of the data is available in Levinsohn and Petrin (2003).

We use the four largest industries (excluding petroleum and refining). The 3-digit level industries and their ISIC codes are Metals (381), Textiles (321), Food Products (311) and Wood Products (331). The data are observed annually and they include gross revenue (the output index), indices of labor and capital inputs, and a measure of the intermediate inputs electricity, materials, and fuels. Quasi-fixed inputs in our setting, are capital stock, skilled labor, and unskilled labor. Materials, fuels and electricity are assumed to be variable inputs.

To implement the Gibbs sampler, first we use repeated blocks of 5,000 iterations to attain convergence. After convergence, we take 15,000 passes of the MCMC procedure. For the data at hand, we needed 15,000 burn - in passes.

To compute inefficiency, we follow Kuosmanen and Johnson (2010). From the residuals

$$e_{it} = y_{it} - \alpha_{it} - \chi'_{it}\beta_{it} - \omega_{it}, \tag{28}$$

our inefficiency measure is computed as

$$u_{it} = -\left(e_{it} - \max_{i,t} e_{it}\right), \tag{29}$$

where e_{it} stands for $v_{it,1}$ in (17). In Kuosmanen and Johnson (2010, p. 154) this requires orthogonality with the regressors, which can be defended on the grounds that inefficiency is unknown to the DMU when decisions are made. See also Kuosmanen and Kortelainen (2012). If this is not the case, other methods should be used as the primary concern of CNLS is to estimate the unknown monotone concave functional form and not inefficiency, where it primarily differs from DEA. In our own context, orthogonality is clearly violated but it is accounted for by (6).

Our primary concern is estimating inefficiency and productivity at the plant level of Chilean manufacturing. We define

Table 3: Bayes factors

	Bayes factor
Model I	1,282.35
Model II	73.45
Model III	21.61
Model IV	1.000

Notes: Model I is defined in (12), Model II in (25), Model III has a productivity formulation as in (26), and Model IV is equivalent to traditional CNLS.

productivity growth (PG) as

$$\Delta\omega_{it} = \frac{\omega_{i,t}-\omega_{i,t-1}}{\omega_{i,t-1}}, \tag{30}$$

when the dependent variable is the level of output and

$$\Delta\omega_{it} = \omega_{i,t} - \omega_{i,t-1},$$

when the dependent variable is log output. In turn, we use (28) and (29) to obtain inefficiency. As we do not have $\boldsymbol{z}_{it}$ variables we use the lagged values of all inputs. The relationships between inefficiency and productivity growth are reported in Figure 1 (for sectors 311, 381, 321, and 331) in the form of joint densities and their contours, and in Figure 2 when we consider all data jointly. As CNLS and its formulation in (12) or (24) take full account of heterogeneity in the data, there is no reason to presume that the results in Figure 1 are biased relative to the results in Figure 2.

To examine Models II and III, we use marginal likelihood and Bayes factors[11] to compare them to Model I. The results are reported in Table 3, where we normalize the Bayes factor for Model IV to 1.00 (Model IV has no productivity component). Relative to Model IV, all other models perform better although the evidence in favor of Model I are overwhelming.

Although Model I appears to perform best according to Bayes factors, it is essential to examine its predictive ability in hold-out samples. For the four sectors as well as the entire data set, we use an estimation phase consisting of 75% of observations, leaving out the remaining 25% to predict inputs and output. We repeat the exercise 1,000 times (each one of which requires new application of MCMC) and we report the results in Table 4. We report relative mean absolute errors (RMAE) of each model relative to Model IV (specifically, their medians across all MCMC simulations for each hold-out sample). Numbers less than one indicate that a model performs better relative to Model IV. The reductions in RMAE for Model I are the most impressive, followed by Models II which also does a good job in terms of out-of-sample forecasting. Model III does not do as well as its productivity formulation in (26) is quite restrictive. Forecasting ability of different models can be tested formally using the Diebold and Mariano (1995) test which is asymptotically distributed as standard normal and its null is equal predictive ability. Equal predictive ability can be rejected when we compare Model I to II, III, or IV showing that the differences in RMAEs are statistically significant. Another quite useful deconfounding statistic was provided by Wang and Blei (2019). Suppose we partition the data into a set that we use, say $\mathcal{X} = \{x_i^{obs}\}_{i=1}^{n_1}$ and a hold-out sample. Therefore, we can compute the posterior $p(z_i|\mathcal{X})$. We also have the hold-out sample $\{x_i^{held}\}_{i=1}^{n_o}$ and we can compute the distribution $p(x_i^{held}|z_i)$. The posterior predictive distribution has density:

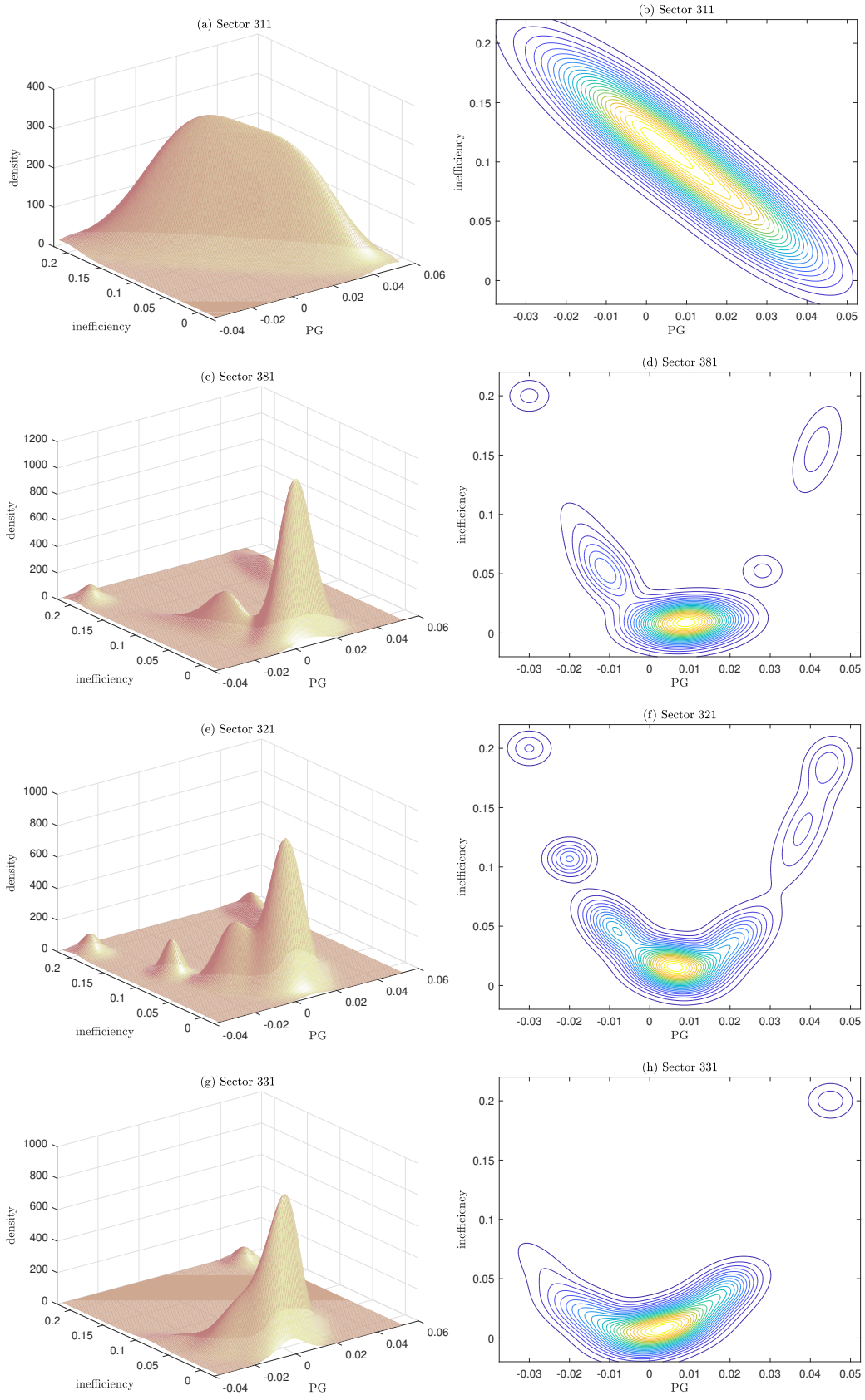
$$p(x_i^{held}|x_i^{obs}) = \int p(x_i^{held}|z_i)p(z_i|x_i^{obs})\,\mathrm{d}z_i. \tag{31}$$

To compute fitted and actual data, Wang and Blei (2019) suggest to use

$$\tau(x_i^{held}) = \mathbb{E}_z\left[\log p(x_i^{held}|z)|x_i^{obs}\right]. \tag{32}$$

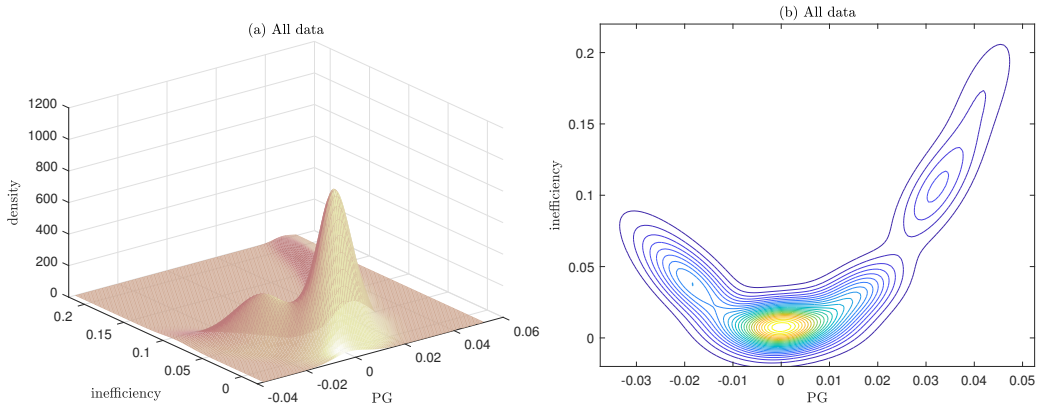
[11] For a model with parameters $\boldsymbol{\theta}$,data $\mathcal{Y}$, likelihood $L(\boldsymbol{\theta};\mathcal{Y})$ and prior $p(\boldsymbol{\theta})$, the marginal likelihood is $\mathcal{M}(\mathcal{Y}) = \int L(\boldsymbol{\theta};\mathcal{Y})p(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$, the integrating constant of the posterior. For any two models, say "1" and "2" (estimated on the same data with possibly different parameters and, therefore, likelihood functions and priors) the Bayes factor in favor of model "1" and against model "2" is $\mathcal{B}_{1:2} = \frac{\mathcal{M}_1(\mathcal{Y})}{\mathcal{M}_2(\mathcal{Y})}$.

Figure 1: Sectoral relationships between inefficiency and productivity growth



Notes: Results are from Model IV.

Figure 2: Relationship between inefficiency and productivity growth (All data)



Notes: Results are from Model IV.

If $\hat{x}_i^{held}$ is a draw from the distribution whose density is (31) then the posterior predictive score (PPS) is:

$$PPS = \Pr\left(\tau(x_i^{held}) < \tau(x_i^{held})\right). \tag{33}$$

This probability can be evaluated using the Monte Carlo method. In our application, we use both the linear and quadratic factor model and Markov Chain Monte Carlo (MCMC) methods to estimate the deconfounder (with, say, 10,000 replications). As Wang and Blei (2019, p. 1581) mention: "How to interpret the predictive score? A good model will produce values of the held-out causes that give similar log likelihoods to their real values—the predictive score will not be extreme. A mismatched model will produce an extremely small predictive score, often where the replicated data has much higher log-likelihood than the real data. An ideal predictive score is around 0.5. We consider predictive scores with predictive scores larger than 0.1 to be satisfactory; we do not have enough evidence to conclude significant mismatch of the assignment model. Note that the threshold of 0.1 is a subjective design choice. We find such assignment models that pass this threshold often yield satisfactory causal estimates in practice". Our posterior predictive scores are presented in Figure 3. Based on this evidence, Model IV passes the PPS test (as its PPS averages 0.4868) the other models do not. Therefore, Model IV is successfully deconfounded.

Regarding persistence of inefficiency and productivity growth, the results for productivity growth are reported in panel (a) of Figure 4, and for inefficiency in panel (b). In panel (a), reported are contours of marginal effects of the form $\frac{\partial \Phi_\omega}{\partial \omega_{i,t-1}}$. For inefficiency, in panel (b), we consider the joint distribution of the estimates in (29). Productivity growth is highly persistent (the average marginal effect is 0.710) while inefficiency is also persistent but the joint distribution of u_{it} and $u_{i,t-1}$ appears bimodal (with two modes near 0.03 and 0.14). For inefficiency levels higher than about 5%, the contours lie above the 45° line showing that $\mathbb{E}(u_{it}|u_{i,t-1}) > u_{i,t-1}$, implying that there is little effort exercised for improvements in inefficiency. Since this happens for relatively larger levels of inefficiency, its interpretation is consistent with the view that

Table 4: Relative Mean Absolute Errors in cross-validated samples

	Output	Materials	Fuels	Electricity
Model I	0.454	0.545	0.517	0.423
Model II	0.955	0.744	0.618	0.628
Model III	1.192	1.117	1.120	1.233
DM Model I vs. II	0.0042	0.0013	0.0010	0.0014
DM Model I vs. III	0.000	0.000	0.000	0.000
DM Model I vs. IV	0.000	0.000	0.000	0.000

Notes: Reported are relative mean absolute errors (RMAE) relative to Model IV. Numbers less than one indicate that a model performs better relative to Model IV. Model I is defined in (12), Model II in (25), Model III has a productivity formulation as in (26), and Model IV is equivalent to traditional CNLS. DM provides p-values of the Diebold and Mariano (1995) test of equal predictive ability. A p-value of 0.00 indicates that it is less than 10^{-4}. The null hypothesis is equal predictive ability.

Figure 3: Posterior predictive scores

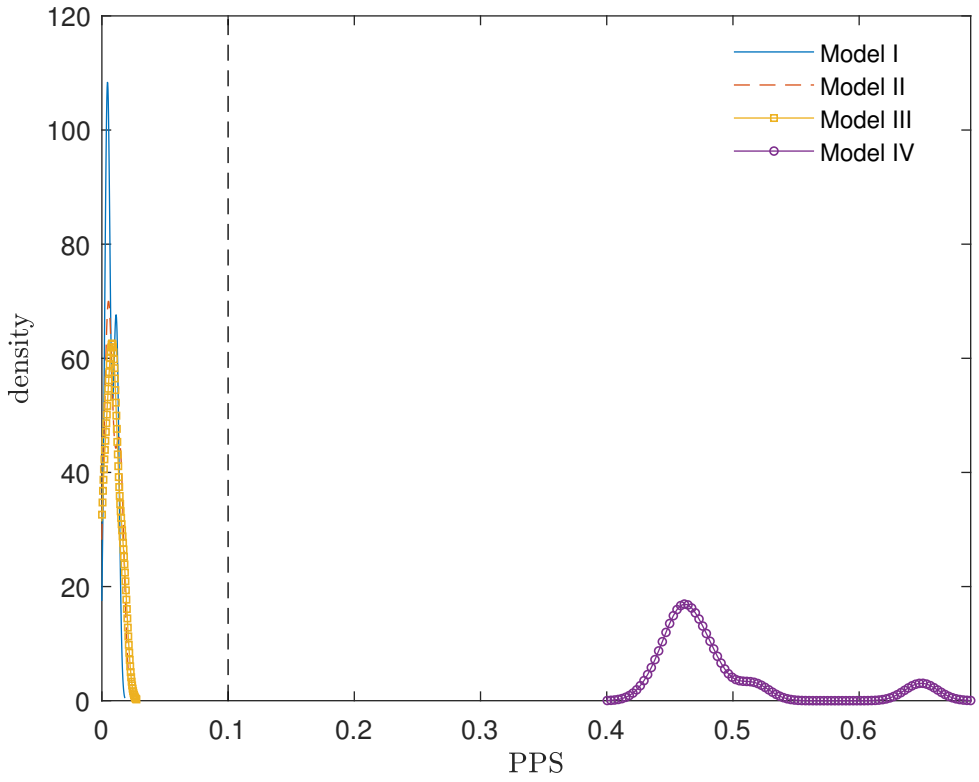
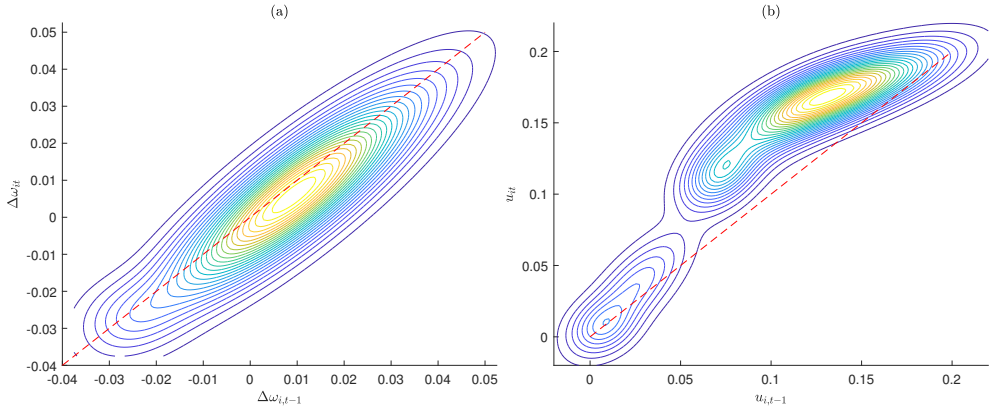


Figure 4: Persistence of inefficiency and productivity growth



Note: The dotted line represents the 45° line.

Table 5: Marginal effects

	marginal effects on		
lagged values of inputs and output	$\Delta\omega_{it}$	u_{it} (method 1)	u_{it} (method 2)
Materials	0.320	–0.042	–0.040
	(0.055)	(0.017)	(0.013)
Fuels	0.017	–0.035	–0.039
	(0.020)	(0.021)	(0.025)
Electricity	0.013	–0.037	–0.030
	(0.012)	(0.010)	(0.015)
Output	0.442	0.071	0.068
	(0.071)	(0.027)	(0.025)

Notes: Marginal effects for productivity are computed as $\frac{\partial \Phi_{\omega}(\cdot)}{\partial \boldsymbol{z}_{it}}$ and they are converted to elasticities. For inefficiency (method 1) we increase the observed values of a variable input by 1%. Given the existing MCMC parameter draws, we compute the residuals of the production function in (12) and we calculate new inefficiency estimates from (29). For method 2, we update MCMC draws using the Sampling-Importance-Resampling procedure of Rubin (1987, 1988) which involves only re-weighting the existing draws instead of performing again MCMC. All elasticities are computed for each MCMC draw and they are averaged to take full account of parameter uncertainty. Reported in parentheses are sample standard deviations.

inefficiency is costly to reduce and requires the use and access to resources (including managerial resources). Additionally, another factor at work is the negative association between inefficiency and productivity growth documented in Figure 2 and the sectoral evidence in Figure 1.

Persistence is also known as *path dependence* which is considered as a causal relationship within a structure in which current values are associated with past ones and jointly affect the future ones (Kasy, 2011; Tsekouras et al., 2016, 2017). In this context, path dependence is considered as a result of technological capabilities and resources which are, of course, highly persistent.
Our $\boldsymbol{z}_{it}$ variables include lagged values of inputs and output so, these are the determinants of productivity. As these variables proxy for other variables of the environment or context, they are likely to be correlated with and capture the effect of contextual variables (an issue that we examine in ore detail in the next section). The marginal effects are reported in Table 5 in the form of elasticities, that is they correspond to percentage increases in inefficiency or productivity growth when lagged inputs and lagged output are increased by 1%.

From the evidence in Table 5, variable inputs have a positive effect on productivity but a negative effect on inefficiency. Lagged output has an average marginal effect which is positive for both inefficiency and productivity. As inefficiency is not a "structural" part of the model, that is, we do not model explicitly inefficiency as a function of other variables or stochastic assumptions, we need to explain how marginal effects were computed. Specifically, ceteris paribus, we increase the observed values of a variable input (say fuels) by 1%. Given the existing MCMC parameter draws, we compute the

residuals of the production function in (12) and the new inefficiency estimates from (29). From a comparison of the existing and new estimates of inefficiency, it is then easy to compute marginal effects.
It is not clear whether the model should have been re-estimated in view of changes in one variable input, that is take the existing MCMC parameter draws as given when computing inefficiency marginal effects. To examine whether this issue is important, we update MCMC draws using the Sampling-Importance-Resampling procedure of Rubin (1987, 1988) which involves only re-weighting the existing draws instead of performing again MCMC. We call this approach "method 2" and the resulting estimates of marginal effects are reported in the last column of Table 5. Although there are some differences (particularly in standard deviations) the qualitative conclusions remain the same. The results suggest under-use of variable inputs, to different degrees, and different contributions to productivity growth.

7 On productivity and inefficiency

One may argue that productivity and inefficiency are correlated or, perhaps, causally related. For example, productive firms may be more efficient and vice versa. The argument has been made forcefully by Bandyopadhyay and Das (2006). As they write: "It has been argued that controllable errors may very well be affected by statistical noise even though it may not directly affect it. For example, in multiple cropping agricultural productions, natural calamity in one season may affect decision making in subsequent seasons and managerial decisions may be affected even by such random factors as weather. Moreover, model misspecification error resulting from exclusion of variables affecting inefficiency may make the component errors correlated. Thus, if firm level production efficiency depends upon the age and size of firms, exclusion of these variables from the model will make component errors correlated due to inclusion of these variables in statistical noise" (Bandyopadhyay & Das, 2006, p. 166). However, this argument depends on model misspecification which would invalidate the entire model. On the other hand, we can never be sure that we are free of omitted variables. Decounfing can, approximately, find these omitted variable so, the possibility that common factors are correlated with both noise and inefficiency cannot be excluded on prior grounds. Another reason is that we measure inefficiency using (28) and (29), so, necessarily noise and inefficiency are correlated. Therefore, there is thre potential of correlation between noise and inefficiency. A most important (as opposed to statistical reason) arises if the production is

$$y = f(\boldsymbol{x})e^{-u} + g(\boldsymbol{x})v,$$

with the convention that $\boldsymbol{x}, y$ are not in log terms. Under the assumption of expected utility of profit maximization, the producer solves

$$\max \mathbb{E}\mathscr{U}(\Pi) = \mathbb{E}\left[p\left\{f(\boldsymbol{x})e^{-u} + g(\boldsymbol{x})v\right\}\right], \tag{34}$$

where $\mathscr{U}(\cdot)$ denotes the utility function. Inefficiency u is known but noise v is not at the time of decision-making. After a little algebra (Kumbhakar, 2002), the first-order conditions are:

$$\frac{\partial f(\boldsymbol{x})}{\partial x_j}(1-u) = w_j - \vartheta\frac{\partial g(\boldsymbol{x})}{\partial x_j},$$

where $\vartheta = \frac{\mathbb{E}\{\mathscr{U}'(\Pi)v\}}{\mathbb{E}\{\mathscr{U}'(\Pi)\}}$. As u is given, factor demands will depend on u as well as the *distribution* of v since ϑ depends on the distribution of v (through its moments and functional form).

7.1 A causal relationship?

As inefficiency is not a "structural" part of the model (see, for example, Kuosmanen and Kortelainen, 2007, 2012) one has to wonder whether observed associations like the ones in Figures 1 and 2 mean anything at all. Of course, it is not necessary to mean something, in which case, as in Kuosmanen and Johnson (2010) one can take the inefficiency estimates as estimates of deviations from the frontier without further interpretations. If, however, such associations can be given a causal interpretation, this would imply that there features in the model that provide a logical link between productivity and inefficiency. If this is true, in the absence of a direct causal model that relates inefficiency to productivity, it would allow us to perform policy experiments in which we can examine the effect of reducing inefficiency on increasing productivity or productivity growth. In doing so, we should, additionally, document the channels through which such effects are possible, as inefficiency and productivity are only implicitly related (if at all) in the model.
In the absence of experimental data, establishing causal relationships is known to be extremely difficult (Pearl, 2009; Imbens and Rubin, 2015; Peters et al., 2013, 2017, Pfister et al., 2019). For causal models, it is known that changes in the environment should not adversely affect their prediction properties as the effect of confounding variables has been accounted for. The literature on deconfounding strategies is long. Here, we follow the principles in Wang and Blei (2019). The goal is to determine the causal effect of certain variables $\mathbb{X}$ on a dependent variable $\mathbb{Y}$. In the absence of decondounding variables, they can be approximated using factor models of the form:

$$\begin{aligned} \xi &\sim p(\cdot|\vartheta_1), \\ \mathbb{X}|\xi &\sim p(\cdot|\xi,\vartheta_2), \end{aligned} \tag{35}$$

where ξ is a set of unobserved variables, and $p(\cdot|\vartheta_1)$ and $p(\cdot|\xi,\vartheta_2)$ denote distributions that depend on parameters ϑ_1 and ϑ_2. Given estimates $\hat{\xi} = \mathbb{E}(\xi|\mathbb{X})$ which can be treated as substitute confounders, one can estimate $\mathbb{E}(\mathbb{Y}|\mathbb{X},\hat{\xi})$ instead of $\mathbb{E}(\mathbb{Y}|\mathbb{X})$. As Wang and Blei (2019) argue: "If we find a factor model that accurately represents the distribution of causes then that model can provide a variable that captures the unobserved multiple-cause confounders. The reason is that the multiple-cause confounders induce dependence among the causes; a good factor model provides a variable that renders the causes conditionally independent; thus, that variable captures the confounders. This is the blessing of multiple causes" (Wang and Blei, 2019, p. 1578). Clearly, the difference $\mathbb{E}(\mathbb{Y}|\mathbb{X},\hat{\xi}) - \mathbb{E}(\mathbb{Y}|\mathbb{X})$ is a measure of the confounding effect, and it should be zero if there is no confounding. For example, to investigate the relationship between productivity growth ($\Delta\omega_{it}$) and inefficiency (u_{it}) we assume that the potential confounders are related to inputs ($\boldsymbol{\chi}_{it} = [\boldsymbol{k}'_{it}, \boldsymbol{x}'_{it}]'$) through a factor model[12] of the form

$$\begin{aligned} \underset{(d_f\times 1)}{\boldsymbol{\xi}_{it}} &\sim \mathcal{N}_{d_f}(\mathbf{0},\, \mathbf{I}), \\ \underset{(K+M)\times 1}{\boldsymbol{\chi}_{it}} &= \underset{(K+M)\times d_f}{\boldsymbol{\Lambda}} \underset{(d_f\times 1)}{\boldsymbol{\xi}_{it}}, \end{aligned} \tag{36}$$

where d_f denotes the number of factors, $\boldsymbol{\Lambda}$ is a matrix containing unknown parameters (factor loadings), and $\boldsymbol{\xi}_{it}$ denote the factors of deconfounders. The factor model in (36) is estimated using maximum likelihood for each MCMC iteration. In turn, we focus attention on the joint distribution of $\Delta\omega_{it}$ and u_{it} conditional on $\hat{\boldsymbol{\xi}}_{it}$. For each MCMC draw[13], the estimates of $\Delta\omega_{it}$ and u_{it} will be different when we condition on the deconfounder which, in this instance, becomes an argument of (8). The joint distribution of the new estimates of $\Delta\omega_{it}$ and u_{it} is shown in Figure 5. Although the two distributions are different compared to Figure 2, the characteristic "smile" characteristic in panel (b) is still evident, showing that inefficiency is likely to increase when productivity increases but decreases when productivity growth is negative.

Based on this evidence, the relationship between inefficiency and productivity in Figure 2 (or its more accurate version in Figure 5) can be given a causal interpretation. Evidence from panel Granger-causality tests (Dumitrescu and Hurlin, 2012; not reported here but available on request) show that the direction of causality runs from productivity growth to inefficiency and there is no feedback in this relationship.

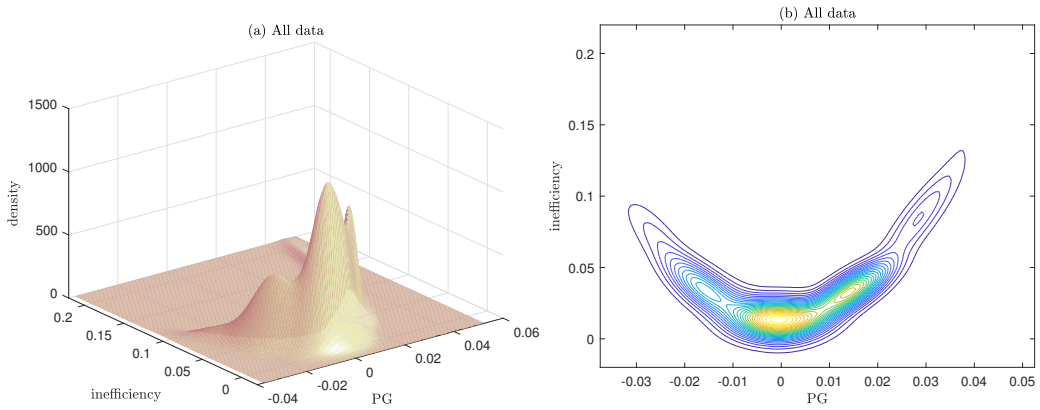
7.2 Model validity

The deconfounding approach in the previous subsection can be used to examine whether the model as a whole admits a causal interpretation in the absence of the randomization principle. The factor model in (36) delivers deconfounder estimates $\hat{\boldsymbol{\xi}}_{it}$ which can be used as (i) parts of $\boldsymbol{z}_{it}$, affecting productivity, (ii) an input in the production function (5) (without its coefficient being subject to the Afriat inequalities). We run an additional MCMC procedure in (12) using

[12] Although the model in (36) is a linear factor model, in principle, nonlinear factor models can be used if they represent better the distribution of $\mathbb{X}$. In this application, the fit of the linear model was excellent so we do not explore this possibility further.

[13] Initial experimentation with the data shows that a single principal component accounts for more than 90% of the total variation in outputs and inputs so, we set $d_f = 1$.

Figure 5: Deconfounded relationship between productivity growth and inefficiency



the deconfounder estimates[14] $\hat{\boldsymbol{\xi}}_{it}$ and we would like to compare the joint distributions of output, inputs, productivity growth and inefficiency, viz. $p(y_{it}, \boldsymbol{x}_{it}, \omega_{it}, u_{it})$ and $p(y_{it}, \boldsymbol{x}_{it}, \omega_{it}, u_{it}|\hat{\boldsymbol{\xi}}_{it})$ where averaging has been performed with respect to MCMC draws to account for parameter uncertainty. We compare the two distributions using a hold-out sample of 20% of our total number of observations and, instead of full MCMC, we update the posteriors in the estimation samples using Sampling-Importance-Resampling as in Rubin (1987, 1988). We use 1,000 randomly selected hold-out samples. For each hold-out sample, we predict $y_{it}, \boldsymbol{x}_{it}, \omega_{it}, u_{it}$ conditionally and unconditionally on $\hat{\boldsymbol{\xi}}_{it}$ and, in turn, we perform a posterior predictive check using the equal forecasting performance test of Diebold and Mariano (1995). As the test is asymptotically distributed as standard normal, it is easy to compute its p-value. Across the 1,000 different hold-out sub-samples, the distribution of p-values of the Diebold and Mariano (1995) test are reported in Figure 6. We report the same tests when (5) is Cobb-Douglas and translog instead of the CNLS approximation. For both these functional forms, the system in (12), other than the production function and the Afriat inequalities, is of the same form, and we use the same deconfounding approach as in (36).

Although in a relatively small number of sub-samples, the predictions of the two models are different, for the most part, the Diebold and Mariano (1995) test shows that the two models (with and without the decondounder) perform equally well. This provides direct evidence that CNLS coupled with general first-order conditions for the variable inputs and flexible productivity formulation, do not appear to be non-causal in the light of the data. On the contrary, the predictive performance of both the Cobb-Douglas and the translog deteriorate significantly under the deconfounding approach, showing that these specifications cannot have a causal interpretation.[15]
A further model validity test can be based on what we know about technologies, namely the fact that they do not tend to change abruptly from one period to the other. As CNLS does not impose such constraints on the β_{it}s in (12) this can be examined ex post. To this purpose. we fit regressions of the form

$$\bar{\beta}_{it} = a_i + r_i \bar{\beta}_{i,t-1} + \xi_{it}^{\bar{\beta}},\ i = 1, \ldots, n,\ t = 1, \ldots, T,$$

where $\bar{\beta}_{it}$ are estimated posterior means of β_{it} in (12), a_i and r_i are parameters and $\xi_{it}^{\bar{\beta}}$ are error terms. Parameters a_i and r_i are estimated by least-squares and the sample distribution (across banks) is presented in Figure 7.

From the evidence in Figure 7, without the deconfounder, the average persistence parameter is close to 0.35 and ranges from 0.3 up to 0.4. With the deconfounder, it increases sharply to almost 0.95 and ranges from slightly less than 0.9 to slightly above one. From these important differences, we understand how important it is to use the deconfounding approach as a reality check.

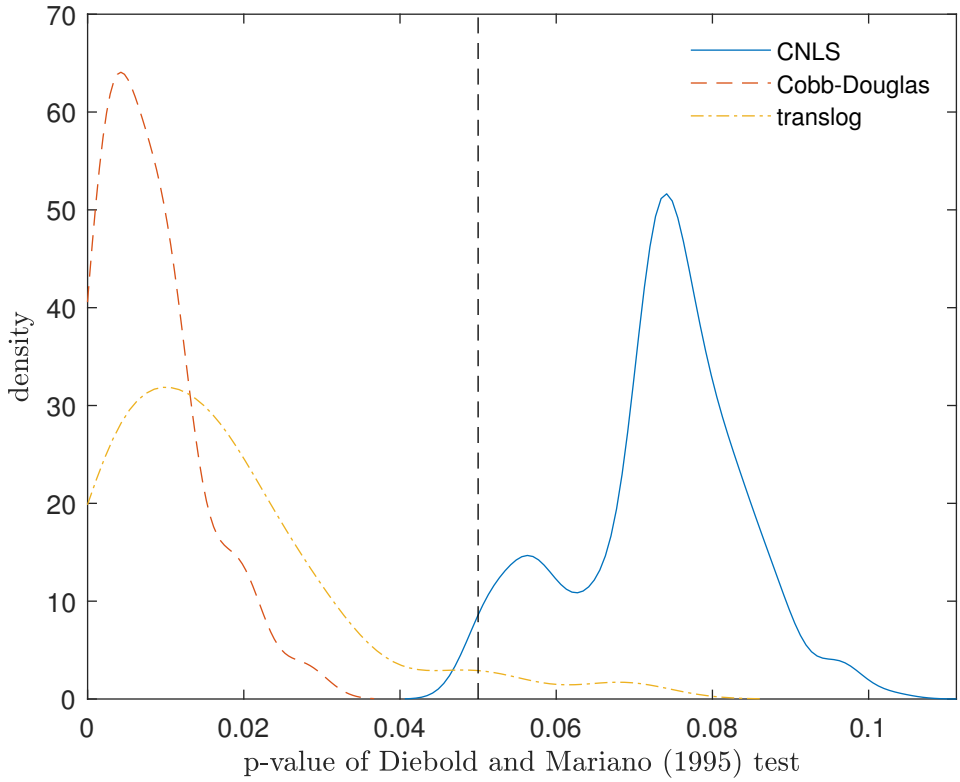
7.3 Policy analysis in the light of deconfounding

Given a factor model like (36), it is possible to use Bayes' theorem to recover the distribution $p(\xi|\mathbb{X}) \propto p(\boldsymbol{\xi}) \cdot p(\mathbb{X}|\boldsymbol{\xi})$ if (36) is estimated using Bayesian techniques. In turn, it is possible, to examine how changes in $\mathbb{X}$ (or the distribution of $\mathbb{X}$) affect the deconfounder (or its distribution) and, in turn, examine the effects on the dependent variable or the main variables of

[14] That is, we do not embed estimation of the deconfounders within MCMC.

[15] For the Cobb-Douglas production function, we impose the constraint that the slope coefficients are non-negative. For the translog, we impose monotonicity at the means of the data (in log form) and ten other randomly selected points so as not to interfere with the second-order approximation properties of the translog.

Figure 6: Model validity



Notes: The dotted line corresponds to the 5% level of statistical significance.

Figure 7: Persistence in estimated β_{it}s

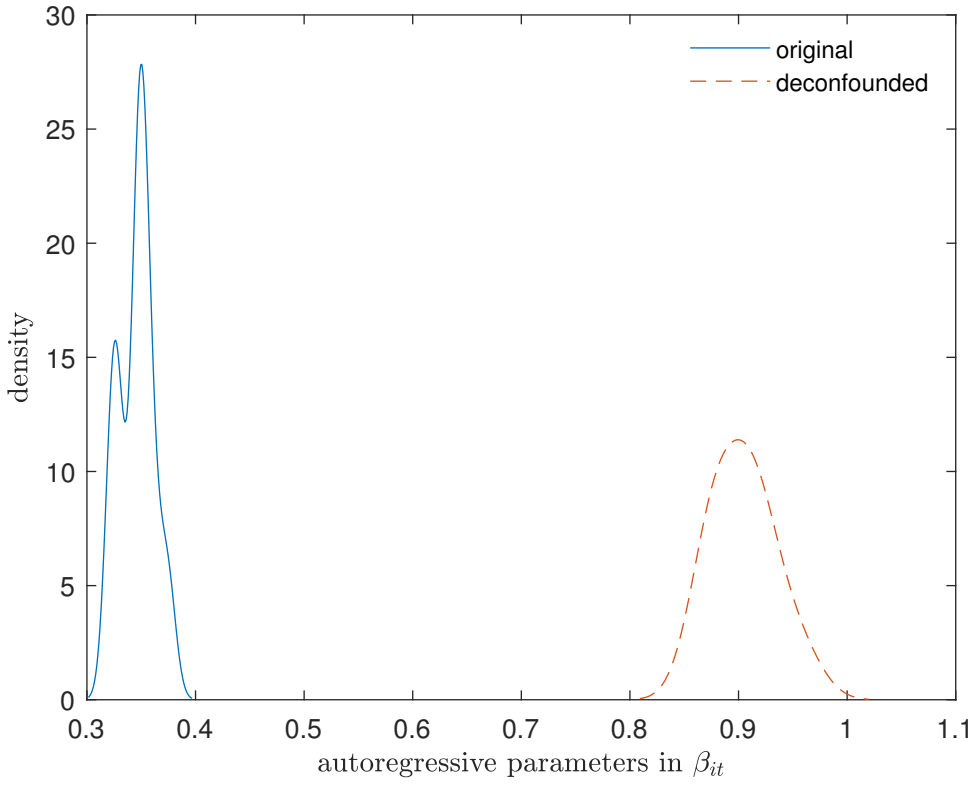


Table 6: Marginal effects

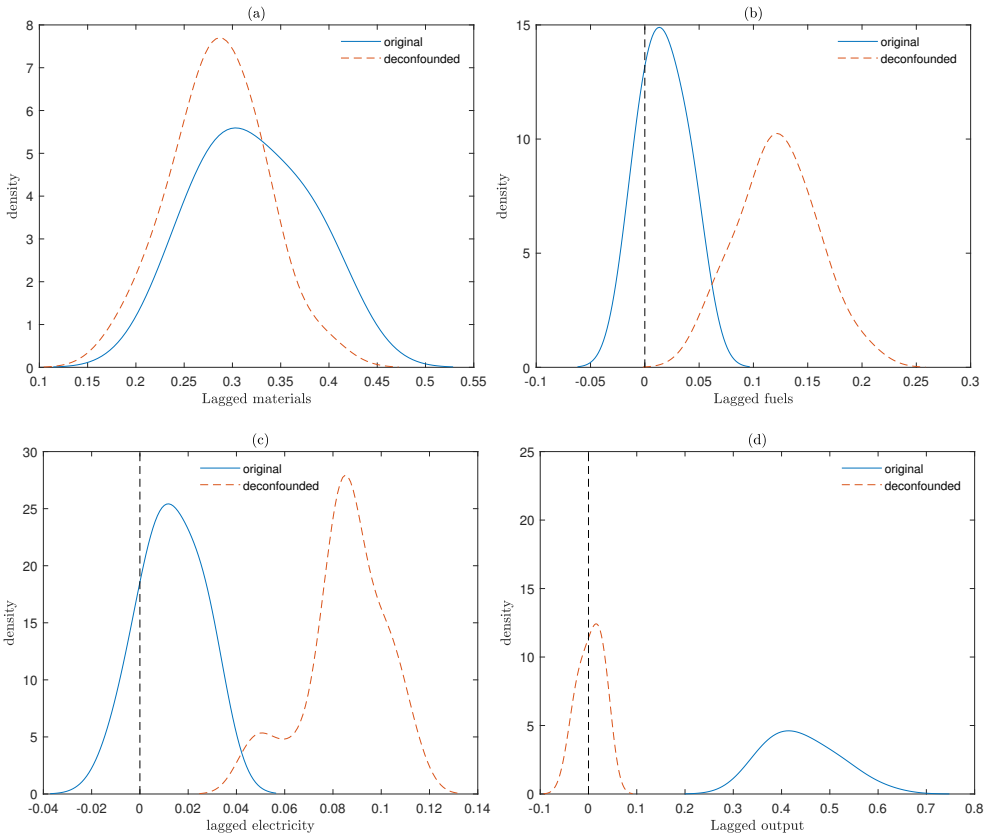
	marginal effects on	
lagged values of inputs and output	$\Delta\omega_{it}$	u_{it}
Materials	0.285	–0.034
	(0.047)	(0.022)
Fuels	0.122	–0.022
	(0.035)	(0.035)
Electricity	0.085	–0.029
	(0.017)	(0.024)
Output	0.005	0.067
	(0.025)	(0.015)

Notes: See Notes to Table 5. All marginal effects are elasticities.

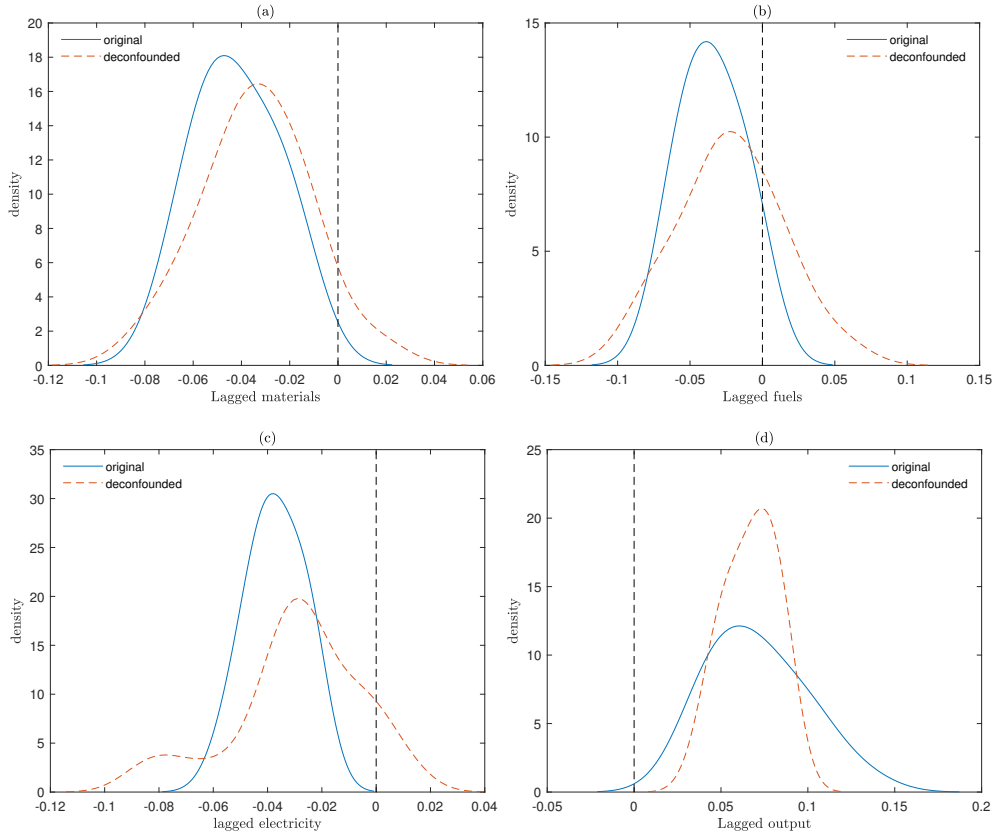
interest which can be efficiency or productivity growth. For this, it is not necessary to embed the deconfounder in a full Bayesian analysis of (36) using MCMC analysis of (12) which includes the deconfounder as we described in the previous subsection. As long as we have an acceptable approximation to $p(\boldsymbol{\xi}|\mathbb{X})$, changes in $\mathbb{X}$ can be used to infer the associated changes of the deconfounder $\boldsymbol{\xi}$ and, in turn, the effects on efficiency and / or productivity growth. This approach is, clearly, an alternative to the approach that was used to deliver the results in Table 5, the difference being that now these can have meaningful causal interpretations. The new marginal effects (elasticities) are reported in Table 6. Relative to Table 5, the deconfounded elasticities are different as can be seen from their sample distributions reported in Figure 5. The most notable difference is with respect to marginal elasticities of lagged output in both efficiency and productivity growth (panels (d) on Parts A and B of Figure 8). The results suggest, again, input under-use, as increasing the use of inputs reduces inefficiency. So, clearly, a managerial implication is that input use should become more intensive. However, from the point of increasing productivity, inputs should be decreased, as a result of the negative relationship between inefficiency and productivity. From the managerial point of view, one should solve an optimization problem to find feasible combinations of inputs and output that can decrease inefficiency and increase productivity, as it is impossible to perform both tasks.

Figure 8: Distributions of marginal effects (elasticities)

A. Productivity growth



B. Inefficiency



The effects of lagged inputs on productivity are positive and the spread of the distributions in Part A of Figure 8, indicates that there is considerable heterogeneity in terms of how different plants respond to productivity shocks. The negative elasticities of fuels and electricity on productivity from the original approach are hard to justify, as we would expect variable inputs to increase as a result of increases in productivity shocks. The deconfounded elasticities are more in line with theoretical considerations and it is easier to argue that these effects are causal. The effect of lagged output appears positive in the original approach (ranging from 0.2 to 0.8) but it averages close to zero in the deconfounding approach. In terms of inefficiency effects, the contribution of variable inputs is negative, and the contribution of lagged output is positive, at least for the most part. The result is consistent with input under-use and can be used to inform discussions on possible directions in directional distance functions or, more importantly, to inform policy analysis to improve efficiency in manufacturing. The policy recommendation would be to follow directions that reduce inputs and increase output which is, essentially, a recommendation on mobilizing managerial resources to reduce waste.

Concluding Remarks

In this paper we address the problem of incorporating productivity in the Convex Nonparametric Least Squares (CNLS) approximation of an arbitrary monotone concave production function. As productivity shocks are correlated with the inputs we address a major problem in CNLS, viz. input endogeneity. Productivity determines variable inputs, as the case should be, under standard behavioral assumptions like cost minimization and, additionally, it is given by a flexible dynamic specification that allows for the presence of predetermined or contextual variables. The model is estimated for four sectors of Chilean manufacturing, a data set that has been prominent in contributions to production function estimation under endogeneity. Although inefficiency is not a structural part of the model we are, nevertheless, able to document a causal relationship between inefficiency and productivity growth based on a deconfounding approach. The deconfounding approach also shows that, unlike more standard production functions like the Cobb-Douglas and the translog, CNLS along with the system of first-order conditions and dynamic productivity, admits a causal interpretation. We use the deconfounding approach to show how we can arrive at policy conclusions and inform policy discussions regarding waste in the sector.

Appendix A. Markov Chain Monte Carlo

APPENDIX. Markov Chain Monte Carlo

A.1 General

Given the posterior in (23) we can use Gibbs sampling to draw parameters and latent variables from their respective conditional posterior distributions. Specifically, we can draw $\boldsymbol{U}$ as follows.

$$\boldsymbol{U}|\theta,\sigma_U,\boldsymbol{\Sigma},\mathcal{Y}\sim\mathcal{N}^{+}_{N(N-1)}\left(-\tfrac{\sigma_U^2}{\sigma_V^2+\sigma_U^2}\boldsymbol{d}(\boldsymbol{\beta}),\ \tfrac{\sigma_V^2\sigma_U^2}{\sigma_V^2+\sigma_U^2}\cdot\mathbf{I}\right). \tag{A.1}$$

The draws are obtained one by one and the operation is easily parallelizable in computers with multiple nodes.

The conditional posterior of σ_U has the simple form:

$$\left.\tfrac{\overline{Q}+\boldsymbol{d}(\boldsymbol{\beta})'\boldsymbol{d}(\boldsymbol{\beta})}{\sigma_U^2}\right|\theta,\Sigma,\mathcal{Y}\sim\chi^2_{N+\overline{N}}, \tag{A.2}$$

$$p(\boldsymbol{\Sigma}|\theta,\sigma_U,\mathcal{Y})\propto|\boldsymbol{\Sigma}|^{-(N+\overline{d}+1)/2}\cdot\exp\left\{-\tfrac{1}{2}\left[\,\overline{\mathbf{A}}+\mathbf{A}(\theta,\boldsymbol{\omega})\right]\Sigma^{-1}\right\}, \tag{A.3}$$

which is in the form of an inverted Wishart distribution (Zellner, 1971, pp. 395–6). Random drawings from (A.1), (A.2), and (A.3) can be realized easily.

Parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are drawn using a Riemannian based Girolami and Calderhead (2011) update which is reviewed

below. To draw

$$\boldsymbol{\eta}_{it} = \begin{bmatrix} \alpha_{it} \\ \boldsymbol{\beta}_{it} \end{bmatrix}, \underset{(D\times 1)}{\boldsymbol{\eta}} = \{\boldsymbol{\eta}_{it},\, i = 1, \ldots, n,\, t = 1, \ldots, T\},$$

where $D = (K+M+1)N$ we recognize that their dimensionality is quite large. Suppose $\boldsymbol{\varpi}$ is a vector whose dimensionality is $m \times 1$ ($m \ll D$) and for some $D \times m$ matrix $\boldsymbol{\Psi}$ we can express

$$\boldsymbol{\eta} = \boldsymbol{\Psi}\boldsymbol{\varpi}. \tag{A.4}$$

To proceed, we use a compression method as in Guhaniyogi and Dunson (2015), which the elements of matrix Ψ are as follows:

$$\Psi_{ij} = \begin{cases} -1/\sqrt{\psi}, \text{ with probability } \psi^2, \\ 0, \text{ with probability } 2\psi(1-\psi), \\ 1/\sqrt{\psi}, \text{ with probability } (1-\psi)^2, \end{cases} \qquad i = 1, \ldots, D,\, j = 1, \ldots, m, \tag{A.5}$$

where $\psi \in (0, 1)$ is an unknown parameter. The order m is also unknown, so Guhaniyogi and Dunson (2015) recommend to draw random values from (A.5) for given values of m and ψ and choose the elements of $\boldsymbol{\Psi}$ and m and ψ using the marginal likelihood criterion. In turn, the elements of $\boldsymbol{\varpi}$ are updated using another application of the Girolami and Calderhead (2011) approach. The marginal likelihood is standard output of the Particle Filtering approach, see A.3.

A.2 Riemannian MCMC

We use the Girolami and Calderhead (2011, GC) algorithm to update draws for a parameter $\boldsymbol{\theta}$. The algorithm uses local information about the gradient and the Hessian of the log-posterior at the existing draw. A Metropolis test is again used for accepting the candidate so generated but the GC algorithm moves considerably faster relative to our naive scheme previously described. MCMC is run until convergence. It has been found that the GC algorithm performs vastly superior relative to the standard MH algorithm and autocorrelations are much smaller.

Suppose $\mathscr{L}(\boldsymbol{\theta}) = \log p(\boldsymbol{\theta}|\mathcal{Y})$ is used to denote for simplicity the log posterior of $\boldsymbol{\theta}$. Moreover, define

$$\boldsymbol{G}(\boldsymbol{\theta}) = \text{est.cov}\, \tfrac{\partial}{\partial \boldsymbol{\theta}} \log p(\boldsymbol{\mathcal{Y}}|\boldsymbol{\theta}), \tag{A.6}$$

the empirical counterpart of

$$\boldsymbol{G}_o(\boldsymbol{\theta}) = -E_{\mathcal{Y}|\boldsymbol{\theta}}\, \tfrac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \log p(\mathcal{Y}|\boldsymbol{\theta}) \tag{A.7}$$

The Langevin diffusion is given by the following stochastic differential equation:

$$d\boldsymbol{\theta}(t) = \tfrac{1}{2}\tilde{\nabla}_{\boldsymbol{\theta}}\mathscr{L}\{\boldsymbol{\theta}(t)\}\, dt + \mathrm{d}\mathbf{B}(t), \tag{A.8}$$

where

$$\tilde{\nabla}_{\boldsymbol{\theta}}\mathscr{L}\{\boldsymbol{\theta}(t)\} = -\boldsymbol{G}^{-1}\{\boldsymbol{\theta}(t)\} \cdot \tilde{\nabla}_{\boldsymbol{\theta}}\mathscr{L}\{\boldsymbol{\theta}(t)\}, \tag{A.9}$$

is the so called "natural gradient" of the Riemann manifold generated by the log posterior. The elements of the Brownian

motion are

$$\begin{aligned}\boldsymbol{G}^{-1}\left\{\boldsymbol{\theta}\left(t\right)\right\}d\mathbf{B}_{i}\left(t\right)=&\left|\boldsymbol{G}\left\{\boldsymbol{\theta}\left(t\right)\right\}\right|^{-1/2}\sum_{j=1}^{K_{\beta}}\frac{\partial}{\partial\boldsymbol{\theta}}\left[\boldsymbol{G}^{-1}\left\{\boldsymbol{\theta}\left(t\right)\right\}_{ij}\left|\boldsymbol{G}\left\{\boldsymbol{\theta}\left(t\right)\right\}\right|^{1/2}\right]\mathrm{d}t\\&+\left[\sqrt{\boldsymbol{G}\left\{\boldsymbol{\theta}\left(t\right)\right\}}\mathrm{d}\boldsymbol{B}\left(t\right)\right]_{i},\end{aligned}\tag{A.10}$$

The discrete form of the stochastic differential equation provides a proposal as follows:

$$\begin{aligned}\tilde{\boldsymbol{\theta}}_{i}=&\boldsymbol{\theta}_{i}^{o}+\tfrac{\varepsilon^{2}}{2}\left\{\boldsymbol{G}^{-1}\left(\boldsymbol{\theta}^{o}\right)\nabla_{\boldsymbol{\theta}}\mathscr{L}\left(\boldsymbol{\theta}^{o}\right)\right\}_{i}-\varepsilon^{2}\sum_{j=1}^{K_{\theta}}\left\{\boldsymbol{G}^{-1}\left(\boldsymbol{\theta}^{o}\right)\frac{\partial\boldsymbol{G}\left(\boldsymbol{\theta}^{o}\right)}{\partial\boldsymbol{\theta}_{j}}\boldsymbol{G}^{-1}\left(\boldsymbol{\theta}^{o}\right)\right\}_{ij}\\&+\tfrac{\varepsilon^{2}}{2}\sum_{j=1}^{K_{\theta}}\left\{\boldsymbol{G}^{-1}\left(\boldsymbol{\theta}^{o}\right)\right\}_{ij}\mathrm{tr}\left\{\boldsymbol{G}^{-1}\left(\boldsymbol{\theta}^{o}\right)\frac{\partial\boldsymbol{G}\left(\boldsymbol{\theta}^{o}\right)}{\partial\boldsymbol{\theta}_{j}}\right\}+\left\{\varepsilon\sqrt{\boldsymbol{G}^{-1}\left(\boldsymbol{\theta}^{o}\right)}\boldsymbol{\xi}^{o}\right\}_{i}\\=&\boldsymbol{\mu}\left(\boldsymbol{\theta}^{o},\varepsilon\right)_{i}+\left\{\varepsilon\sqrt{\boldsymbol{G}^{-1}\left(\boldsymbol{\theta}^{o}\right)}\boldsymbol{\xi}^{o}\right\}_{i},\end{aligned}\tag{A.11}$$

where $\boldsymbol{\theta}^{o}$ is the current draw. The proposal density is

$$q\left(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}^{o}\right)=\mathcal{N}_{d_{\theta}}\left(\tilde{\boldsymbol{\theta}},\varepsilon^{2}\boldsymbol{G}^{-1}\left(\boldsymbol{\theta}^{o}\right)\right).\tag{A.12}$$

The convergence to the invariant distribution is guaranteed by using the standard form Metropolis-Hastings probability

$$\min\left\{1,\ \frac{p\left(\tilde{\boldsymbol{\theta}}|\cdot,\mathcal{Y}\right)q\left(\boldsymbol{\theta}^{o}|\tilde{\boldsymbol{\theta}}\right)}{p\left(\boldsymbol{\theta}^{o}|\cdot,\mathcal{Y}\right)q\left(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}^{o}\right)}\right\}.\tag{A.13}$$

A.3 Sequential Monte Carlo

To draw $\{\omega_{it}\}$ in (8) the most computationally efficient procedure is to use Particle Filtering, also known as Sequential Monte Carlo. We use the particle Gibbs (PG) sampler, see Andrieu et al. (2010). This allows us to draw paths of the state variables en bloc. Particle filtering is a simulation-based algorithm that sequentially approximates continuous, marginal distributions using discrete approximations which is performed by using a set of support points called ''particles'' and probability masses (Creal, 2012) for a review.

The PG sampler draws a single path of the latent or state variables from this discrete approximation. As the number of particles $\mathbb{M}$ goes to infinity, the PG sampler draws from the exact full conditional distribution. As mentioned in Creal and Tsay (2015, p. 339): "The PG sampler 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, Andrieu et al. (2010) note that there must be positive probability of sampling the existing path of the state variables that were drawn at the previous iteration. The pre-existing path must survive the resampling steps of the particle filter. The conditional resampling step within the algorithm forces this path to be resampled at least once. We use the conditional multinomial

resampling algorithm from Andrieu et al. (2010), although other resampling algorithms exist, see Chopin and Singh (2013)."

We follow Creal and Tsay (2015). Suppose the posterior is $p(\boldsymbol{\theta}, \Lambda_{1:T}|\mathbf{y}_{1:T})$ where $\Lambda_{1:T}$ denotes the latent variables whose "prior" (in reality, its law of motion) can be described by $p(\Lambda_t|\Lambda_{t-1}, \boldsymbol{\theta})$ and we omit the index i for simplicity. In the PG sampler we can draw the structural parameters $\boldsymbol{\theta}|\Lambda_{1:T}, \mathbf{y}_{1:T}$ as usual, from their posterior conditional distributions. This is important because, in this way, we can avoid mixture approximations or other Monte Carlo procedures that need considerable tuning and may not have good convergence properties. As such posterior conditional distributions we omit the details and focus on drawing the latent variables.

Reintroducing panel data notation (for unit i) Suppose we have $\Lambda_{i,1:T}^{(1)}$ from the previous iteration. The particle filtering procedure consists of two phases.

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

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

$$w_{i,t}^{(m)} = \frac{p(y_{i,t}; \Lambda_{i,t}^{(m)}, \theta) p(\Lambda_{i,t}^{(m)}|\Lambda_{i,t-1}^{(m)}, \boldsymbol{\theta})}{q(\Lambda_{i,t}|\Lambda_{i,t-1}^{(m)}, \boldsymbol{\theta})}, m = 1, \ldots, \mathbb{M}. \tag{A.14}$$

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

Another improvement is 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 compute the normalized weights and particles and we draw a path of the latent variables as we describe 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 $\Lambda_{i,T}^{*} = \Lambda_{i,T}^{(m)}$.
- Compute the backward weights: $w_{t|T}^{(m)} \propto \tilde{w}_{t}^{(m)} p(\Lambda_{i,t+1}^{*}|\Lambda_{i,t}^{(m)}, \theta)$.
- Normalize the weights: $\tilde{w}_{t|T}^{(m)} = \frac{w_{t|T}^{(m)}}{\sum_{m'=1}^{M} w_{t|T}^{(m')}}, m = 1, \ldots, \mathbb{M}$.
- Draw a particle $\Lambda_{i,t}^{*} = \Lambda_{i,t}^{(m)}$ with probability $\tilde{w}_{t|T}^{(m)}$.

Therefore, $\Lambda_{i,1:T}^{*} = \{\Lambda_{i1}^{*}, \ldots, \Lambda_{iT}^{*}\}$ 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 $q(\Lambda_{i,t}|\Lambda_{i,t-1}, \theta)$. We use an importance density implicitly defined by $\Lambda_{i,t} = a_{i,t} + \sum_{p=1}^{P} b_{i,t}\Lambda_{i,t-1}^{p} + h_{i,t}\xi_{i,t}$ where $\xi_{i,t}$ follows a standard (zero location and unit scale)

Student-t distribution with $\nu = 5$ degrees of freedom. That is, we use polynomials in $\Lambda_{i,t-1}$ of order $\mathfrak{P}$. We select the parameters $a_{i,t}, b_{i,t}$ and $h_{i,t}$ during the burn-in phase (using $\mathfrak{P} = 1$ and $\mathfrak{P} = 2$) so that the weights $\{\tilde{w}_{i,t}^{(m)}, m = 1, \ldots, \mathbb{M}\}$ and $\{\tilde{w}_{t|T}^{(m)}, m = 1, \ldots, \mathbb{M}\}$ 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 PG sampler in terms of asymptotic efficiency.

Alternatively, when the dimension of the state vector is large, we can draw $\Lambda_{i,1:T}$, conditional on all other paths $\Lambda_{-i,1:T}$ that are not path i. Therefore, we can draw from the full conditional distribution $p(\Lambda_{i,1:T}|\Lambda_{-i,1:T}, \mathbf{y}_{1:T}, \theta)$.

A.4 Compression features

Another important issue is the compression we use for the parameters α_{it} and $\boldsymbol{\beta}_{it}$ in (A.4) and (A.5). The selection of ψ, m and the elements of Ψ according to (A.5) must be performed for fixed values of ψ and m over a number of draws for Ψ. For ψ we can consider values in $\{0.1, 0.2, \ldots, 0.9\}$ and for dimensionality, m, values in $\{1, 2, \ldots, \overline{m}\}$ for some upper bound $\overline{m}$. The number of observations varies by sector (6115 for sector 311, 1,394 for 381, 1,129 for 321, and 1,032 for 331) so, for effective dimensionality reduction we set $\overline{m} = 100$.

The whole procedure has to be repeated J times for each MCMC implementation to generate J different configurations for Ψ and, finally, select the best configuration (ψ, m, Ψ) that maximizes the marginal likelihood of the model.

As there are 900 different combinations of ψ and m for which we have to search $J = 10,000$, say, times for an optimal Ψ we would have to perform nine million MCMC procedures (each consisting of 15,000 passes or 10,000 after omitting the first 5,000 and we obtain good starting values). Although the number of time series observations is quite small and, therefore, the SMC steps can be executed extremely efficiently (and in parallel) the number of MCMC procedures is still quite large (although, again, they can be executed in parallel).

We experimented with an alternative scheme in the interest of reducing computational time. Specifically, we generate ψ uniformly in the interval (0.1, 0.9), m with equal probability in $\{1, \ldots, \overline{m}\}$ and Ψ as in (A.5). We perform 10,000 different MCMCs based on 10,000 *jointly generated* configurations of ψ, m and Ψ. In turn, we compare the performance of the two methods which we call, respectively, Procedure I and Procedure II. We report some statistics in Table A.1. As it turns out, the two procedures perform nearly the same as the differences in log marginal likelihoods (LML) are trivial, and the correlation coefficients (simple or rank-based) between inefficiency and productivity growth are quite high. This is, of course, evidence that the alternative ways of compressing the parameters, perform in much the same way and the results are not different. However, Procedure II can be implemented in a fraction of the time relative to Procedure I.

A.5 Technical issues

In this subsection we examine the numerical performance of MCMC. To save space, we report results only for the whole data set (for the four sectors) which contains 9,710 observations.

Table 7: Comparison of Procedures I and II

	Sector 311	Sector 381	Sector 321	Sector 331	All data
obs.	6,155	1,394	1,129	1,032	9,710
diff. LML	–0.0015	0.0002	–0.0001	0.0003	–0.0003
corr. ineff.	0.987	0.991	0.995	0.997	0.989
corr. PG	0.976	0.994	0.993	0.995	0.997
rank corr. ineff.	0.987	0.996	0.993	0.997	0.998
rank corr. PG	0.985	0.995	0.997	0.997	0.997

Notes: "obs." denotes number of observations, LML is log marginal likelihood, "corr." denotes correlation coefficient, "ineff." is inefficiency (see (**??**)) and PG is productivity growth (see (30)).

Autocorrelation functions (acf) are reported in Figure A.1. We show results for 50 randomly selected elements of ϖ (in panel (a)), ω (in panel (b)), u (in panel (c)), and β (in panel (d)). The acf show that although MCMC autocorrelation is a problem it does not, nevertheless, prevent us from a thorough exploration of the posterior in (24).

From Procedure II the optimal values of ψ and m and the corresponding LMLs are shown in Figure A.2. Evidently, the relationship is multimodal making it difficult to explore it by means other than the ones suggested here. The optimal configuration of (ψ, m) is (0.4, 175). This shows that parameter compression is highly effective as the problem reduces to 175 parameters ϖ in (A.4).

Corresponding to different configurations of ψ and m productivity growth results change and one interesting question is how well they correlate with the benchmark results reported in Figure A.3. In panel (a) of Figure A.3, we report contours of the correlation coefficients corresponding to different values of ψ and m for productivity growth (in a neighborhood of optimal m for visual clarity). In panel (b) we do the same for inefficiency.

Figure A.3: Correlation contours of ψ and m

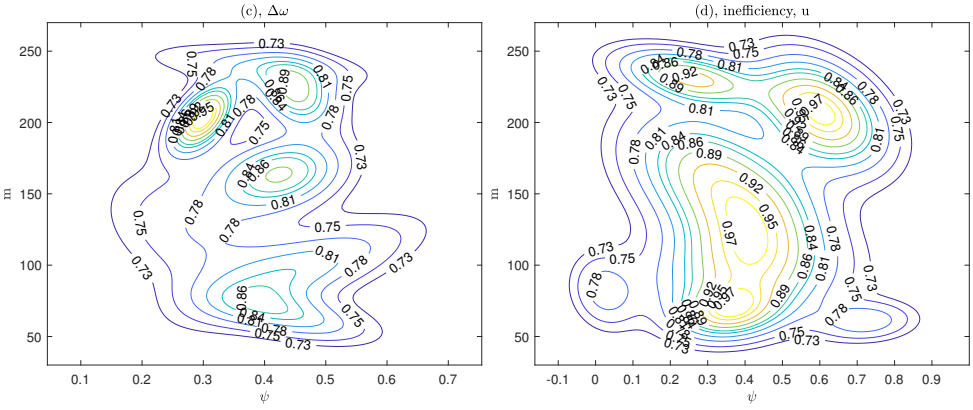


Figure A.1: Autocorrelation functions

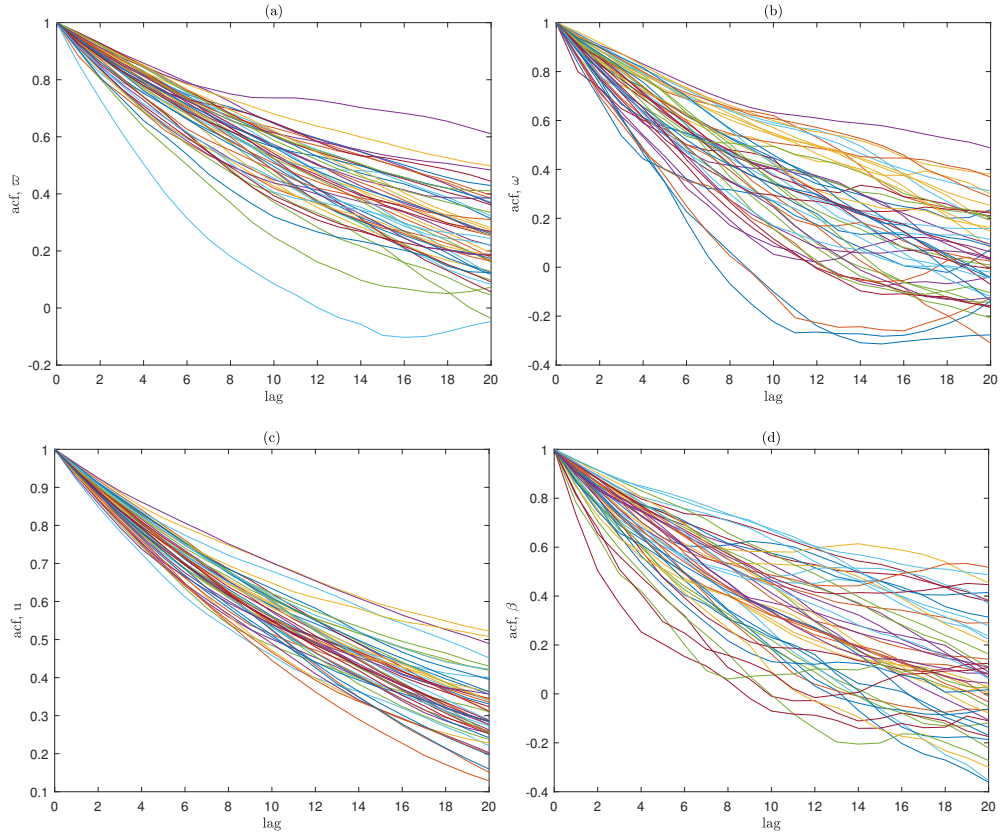
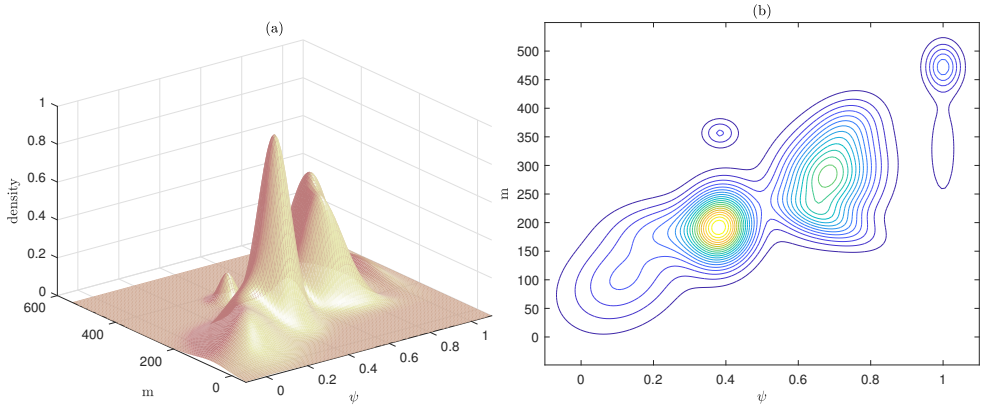
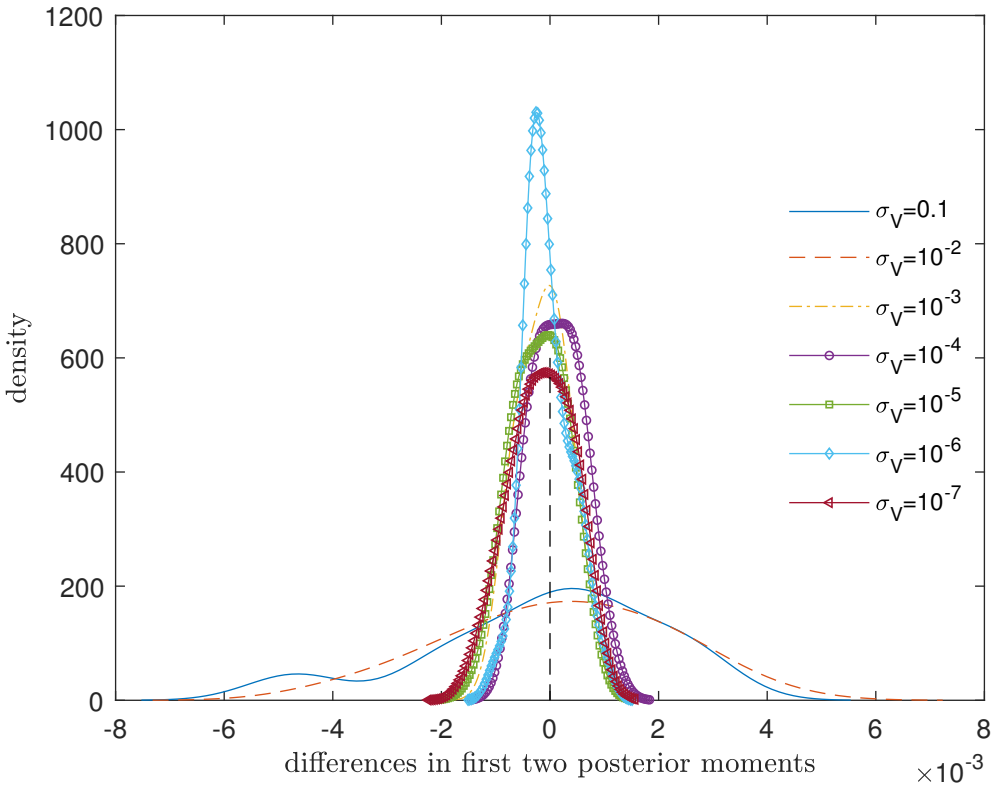


Figure A.2: Marginal likelihood as a function of ψ and m



Notes: The maximum of marginal likelihood is normalized to unity.

Figure A.4: Posterior sensitivity to values of σ_V



From this evidence, it turns out that the results are not extremely sensitive to the choices of ψ, m and Ψ although there are important differences relative to the optimal benchmark choice particularly when ψ and m are relatively far away from their optimal values. However, one can obtain a rough idea about the model's implications by simulating with a few values of ψ, m and Ψ, although accurate estimates of efficiency and productivity or productivity growth, require a careful selection, as we described.

A potential concern is the sensitivity of the posterior in (24) to values of σ_V which was taken to be 10^{-5}. As an overall sensitivity measure of the posterior we take absolute differences of the first two posterior moments for all parameters and latent productivity, including inefficiency. For the four Chilean manufacturing sectors we generate densities of these differences across the different values of ψ, m, and Ψ and we report the results in Figure A.4. From this evidence, the first two posterior moments stabilize quickly at values of σ_V near 10^{-3} and remain the same when σ_V is decreased to 10^{-7} .

Finally, in Table A.1 we report Bayes factors (or posterior odds ratios when the prior odds are 1:1) for our preferred deconfounded model for different values of the polynomial order, Q. For ease of comparison, we normalized the marginal likelihood of of the model with $Q = 1$ to unity. From the results it turns out that the optimal value of Q is 4.

Table A.1. Bayes factor for different polynomial order, Q

Q	Bayes factor
1	1.000
2	85.47
3	125.32
4	417.28
5	322.55
7	182.10
10	55.86

References

[1] Ackerberg, D. A., K. Caves, and G. Frazer (2015). Identification Properties of Recent Production Function Estimators. Econometrica 83 (6), 2411–2451.

[2] Afriat, S.N. (1967). The construction of a utility function from expenditure data. International Economic Review 8, 67–77.

[3] Afriat, S. (1972). Efficiency estimation of production functions. International Economic Review 13, 568–598.

[4] Anderson, R. D., and G. Vastag (2004). Causal modeling alternatives in operations research: Overview and application. European Journal of Operational Research 156 (1), 92–109.

[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] Bandyopadhyay, D., & Das, A. (2006). On measures of technical inefficiency and production uncertainty in stochastic frontier production model with correlated error components. Journal of Productivity Analysis 26, 165–180.

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

[8] Creal, D.D. (2012). A survey of sequential Monte Carlo methods for economics and finance. Econometric Reviews 31 (3), 245–296.

[9] Creal, D., and R. Tsay (2015). High dimensional dynamic stochastic copula models. Journal of Econometrics 189 (2), 335–345.

[10] Diebold, F.X. and R.S. Mariano (1995). Comparing Predictive Accuracy. Journal of Business and Economic Statistics 13, 253–263.

[11] Doraszelski, U., and J. Jaumandreu (2013). R&D and Productivity: Estimating Endogenous Productivity. Review of Economic Studies 80 (4), 1338–1383.

[12] Dumitrescu, E.-I., and C. Hurlin. (2012). Testing for Granger non-causality in heterogeneous panels. Economic Modelling 29: 1450–1460.

[13] Fernandez, C., J. Osiewalski, and M.F.J. Steel, 1997. On the use of panel data in stochastic frontier models with improper priors. Journal of Econometrics 79, 169–193.

[14] Gandhi, A., S. Navarro, & D. A. Rivers (2020). On the Identification of Gross Output Production Functions. Journal of Political Economy 128 (8), 2973–3016.

[15] 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, U.K.: Oxford University Press, pp. 169-193.

[16] Girolami, M., and B. Calderhead (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73 (2), 123–214.

[17] 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.

[18] Guhaniyogi, R., and Dunson, D. B. (2015). Bayesian Compressed Regression. Journal of the American Statistical Association 110 (512), 1500–1514.

[19] Imbens, G.W., and Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction, New York: Cambridge University Press.

[20] Johnson, A. L., and T. Kuosmanen (2011). One-stage estimation of the effects of operational conditions and practices on productive performance: asymptotically normal and efficient, root-n consistent StoNEZD method. Journal of Productivity Analysis 36, 219–230.

[21] Johnson, A. L., and Kuosmanen, T. (2012). One-stage and two-stage DEA estimation of the effects of contextual variables. European Journal of Operational Research 220 (2), 559-570, 2012.

[22] Kasy, M. (2011). Identification in triangular systems using control functions, Econometric Theory 27, 663–671.

[23] Kumbhakar, S.C. (2002). Specification and Estimation of Production Risk, Risk Preferences and Technical Efficiency. American Journal of Agricultural Economics 84 (1), 8–22.

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

[25] Kuosmanen, T. (2008). Representation theorem for convex nonparametric least squares. Econometrics Journal 11, 308–325.

[26] Kuosmanen, T., Johnson, A.L. (2010). Data envelopment analysis as nonparametric least-squares regression. Operations Research 58 (1), 149–160.

[27] Kuosmanen, T. and Kortelainen, M. (2007). Stochastic Nonparametric Envelopment of Data: Cross-Sectional Frontier Estimation Subject to Shape Constraints (May 1, 2007). Available at SSRN: https://ssrn.com/abstract=983882 or http://dx.doi.org/10.2139/ssrn.983882

[28] Kuosmanen, T., and M. Kortelainen (2012). Stochastic non-smooth envelopment of data: semiparametric frontier estimation subject to shape constraints, Journal of Productivity Analysis 38, 11–28.

[29] Lee, C.-Y., A.L. Johnson, E. Moreno-Centeno, T. Kuosmanen (2013). A more efficient algorithm for Convex Nonparametric Least Squares, European Journal of Operational Research 227, 391–400.

[30] Levinsohn, J., and A. Petrin (2003). Estimating Production Functions Using Inputs to Control for Unobservables. Review of Economic Studies 70 (2), 317–341.

[31] Marschak, J., and W. H. Andrews, Jr. (1944). Random Simultaneous Equations and the Theory of Production. Econometrica, 12 (3/4), 143–205.

[32] Nadkarni, S., and P. P. Shenoy (2001). A Bayesian network approach to making inferences in causal maps. European Journal of Operational Research 128 (3), 479–498.

[33] Olley, G., and A. Pakes (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. Econometrica 64 (6), 1263–1297.

[34] Pearl, J. (2009). Causality: Models, Reasoning, and Inference (2nd ed.), New York: Cambridge University Press.

[35] Peters, J., Janzing, D., and Schölkopf, B. (2013). Causal Inference on Time Series Using Structural Equation Models, in Advances in Neural Information Processing Systems (NIPS) (Vol. 26), Curran Associates, Inc., pp. 585–592.

[36] Peters, J., Janzing, D., and Schölkopf, B. (2017). Elements of Causal Inference: Foundations and Learning Algorithms. Cambridge, MA: MIT Press.

[37] Petrin, A., and J. Sivadasan (2013). Estimating Lost Output from Allocative Inefficiency, with an Application to Chile and Firing Costs. The Review of Economics and Statistics 95 (1), 286–301.

[38] Pfister, N., P. Bühlmann & J. Peters (2019). Invariant Causal Prediction for Sequential Data. Journal of the American Statistical Association, 114 (527), 1264–1276.

[39] Roberts, G.O., and A.F.M. Smith (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. Stochastic Processes and their Applications 49 (2), 207–216.

[40] Rubin, D. B. (1987). Comment on "The calculation of posterior distributions by data augmentation", by M. A. Tanner and W. H. Wong, Journal of the American Statistical Association 82, 543–546.

[41] Rubin, D. B. (1988), Using the SIR Algorithm to Simulate Posterior Distributions, in Bayesian Statistics 3, ed. J. M. Bernardo, M. H. DeGroot , D. V. Lindley, and A. F. M. Smith, 395–402. Oxford: Oxford University Press.

[42] Shaw, D., C. M. Smith, and J. Scully (2017). Why did Brexit happen? Using causal mapping to analyse secondary, longitudinal data. European Journal of Operational Research 263 (3), 1019–1032.

[43] Simar, L., Wilson, P.W. (2008). Statistical inference in nonparametric frontier models: recent developments and perspectives. In: Fried, H., Lovell, C.A. K., Schmidt, S. (Eds.), The Measurement of Productive Efficiency, 2nd ed. Oxford University Press.

[44] Simar, L., Wilson, P.W. (2011). Inference by the m out of n bootstrap in nonparametric frontier models. Journal of Productivity Analysis 36, 33–53.

[45] Soytas, M. A., Denizel, M., and Usar, D. D. (2019). Addressing endogeneity in the causal relationship between sustainability and financial performance. International Journal of Production Economics 210, 56–71.

[46] Tsekouras, K., N. Chatzistamoulou, K. Kounetas, & D.C. Broadstock (2016). Spillovers, path dependence and the productive performance of European transportation sectors in the presence of technology heterogeneity. Technological Forecasting and Social Change, 102, 261–274.

[47] Tsekouras, K., N. Chatzistamoulou, & K. Kounetas (2017). Productive performance, technology heterogeneity and hierarchies: Who to compare with whom. International Journal of Production Economics 193, 465–478.

[48] Tsionas, E.G. (2000). Full Likelihood Inference in Normal-Gamma Stochastic Frontier Models, Journal of Productivity Analysis 13 (3), 183-205.

[49] Wang, Y., and D. M. Blei (2019) The Blessings of Multiple Causes, Journal of the American Statistical Association, 114 (528), 1574–1596.

[50] Wooldridge, J. M. (2009). On estimating firm level production functions using proxy variables to control for unobservables. Economics Letters 104 (3), 112–114.

[51] Zellner, A. (1971). An introduction to Bayesian inference in econometrics. Wiley, New York.