A Dynamic Framework for Identification and Estimation of Nonseparable Production Functions

Justin Doty*

October 7, 2021

Abstract

This paper studies identification and estimation of a nonseparable model for production functions with unobserved heterogeneity. Non-parametric identification results are established for the production function and productivity process under stationarity conditions. This framework allows for heterogeneous effects of output elasticities and factor efficiencies in addition to non-linear productivity persistence. It also allows for additional unobservables in the input demand functions, which would violate the scalar unobservability requirement in proxy variables under previous approaches. This extension is used to show firm's heterogeneous responses to productivity shocks corresponding to the size of their input demand. This paper illustrates these results in an application to US manufacturing firms where the proposed model is estimated using non-linear quantile regression.

1 Introduction

Recent advancements in production function estimation have sought to address the simultaneity bias from unobserved productivity under various timing assumptions on firm input decisions. Consistent estimates of the production function are necessary for studying patterns of productivity heterogeneity, returns to scale, and market power. These proxy variable approaches use a firm's input demand function, which is assumed to be strictly increasing in unobserved productivity. The function is inverted so that productivity can be expressed as a function of observed variables. This is then substituted into the production function which is estimated in a two-step approach.

*Department of Economics, University of Iowa, S321 Pappajohn Business Building, 21 E Market St, Iowa City, IA 52242. Email: justin-doty@uiowa.edu

This approach was introduced by Olley and Pakes (1996) (hereafter OP) who consider a dynamic optimization problem of a firm who chooses investment to maximize long-run expected profits and an exit rule which depends on its sell-off value. The investment demand function depends on state variables such as capital stock and productivity, which is unobserved to the econometrician. They show that for positive investment levels, the function is invertible in productivity. Their other contribution is to correct a selection problem which is generated by the firm's optimal exit rule. They characterize an equilibrium in which a firm exits the market if their productivity drops below a threshold value determined by its state variables. The selection problem biases estimates of the elasticities corresponding to the state variables. The correction for this is to include the survival probabilities, estimated from a probit regression, as an additional argument in the productivity process in the second stage estimation procedure.

There are two disadvantages to this approach. First, the monotonicity assumption requires discarding observations for which investment is zero. In many plant-level datasets such as the manufacturing census conducted by Chile, investment levels are often truncated at zero due to high adjustment costs. Second is a violation of the scalar unobservability assumption which requires that productivity be the only unobservable in the investment demand function. This particular violation is not unique to their approach and is a common source of identification failure in the proxy variable literature. Intuitively, if there were additional unobservables in the investment demand function, then productivity cannot be expressed as a function of observed variables alone.

Levinsohn and Petrin (2003) address the first challenge by providing conditions for which an intermediate input demand function, such as materials, energy, or fuels are strictly increasing in productivity. This function is used to express productivity as a function of the observed variables. Since many plants report positive use of intermediate inputs, this eliminates the need to discard observations with zero investment levels. Their approach thereafter is similar to OP. They estimate the parameters corresponding to variable inputs such as labor in the first stage and state variables in the second stage. One issue with this approach is that if labor is a variable input (chosen to maximize short-run profits) then it is a function of the state variables capital and productivity. This is problematic because productivity is inverted as a function of the same conditioning variables. There exists only specific data-generating processes that can break this functional dependence problem. The paper by Ackerberg *et al.* (2015) provide these scenarios in which labor can be identified in the first stage. They propose conditioning on labor in the intermediate input demand function to avoid non-identification of the labor coefficient altogether. This precludes identification of labor in the first stage so it is included in the second stage with the other state variables. This alternative procedure suggests that labor can be chosen prior to or simultaneously as the intermediate inputs. For example, firms will only use certain amounts of material inputs if they know there will be enough workers to utilize them.

The appeal of the control function approaches is its computational simplicity and interpretable timing conditions on input decisions. First stage estimates can be obtained by a polynomial regression and the second stage consists of a nonlinear Generalized Method of Moments (GMM) estimator. The current direction in this literature addresses identification of the model when the input demand functions contain additional unobservables as well as the issue of model specification and its implications for production function estimates.

Invertibility of productivity from the proxy variables is not possible if there are unobserved variables such as demand shocks, input prices, or measurement error. In the OP approach, if the investment demand function contained other unobservables researchers would not be able to infer values of productivity from different levels of investment. Examples of shocks affecting investment demand include, but are not limited to adjustment costs, optimization error, and shocks to product demand. Inversion of multi-dimensional unobservables may be possible if one observes additional proxies, but data on suitable proxies is often not available.[1] The same issue is encountered when intermediate inputs are used as the control function in the LP and ACF framework. Other unobservables, such as measurement error in capital, also poses a serious identification problem since the measurement error appears in both the first and second stage estimates non-parametrically. Kim *et al.* (2016) allow for measurement error in capital and other inputs using identification arguments from Hu and Schennach (2008) (hereafter HS) in the framework of the control function approaches. Hu *et al.* (2019) take a similar identification approach, but propose a GMM estimator.

Controlling for additional unobservables may reduce some of the unexplained heterogeneity across firms, however, there is still a large amount of variation that is left unmodeled. Some part of this variation can be accounted for by model specification. The proxy variable approaches typically use a Cobb-Douglas production function with Hicks-neutral productivity. One implication of this specification is that capital shares are assumed constant across firms, which is often rejected by empirical evidence. Some researchers have addressed this by augmenting the parametric specification using firm-specific production functions in a random-coefficient framework.[2] Nonparametric estimation, such as the procedure pro-

[1]For example, Ackerberg *et al.* (2007) shows that when a demand shock enters the investment function, a firms pricing decision would be needed to proxy for the additional unobservable and productivity.

[2]See for example Kasahara *et al.* (2017), Balat *et al.* (2018) and Li and Sasaki (2017).

posed by Gandhi *et al.* (2020) also show that choice of the production function is important. Estimates from a value-added production function will be different from a gross-output production function since the latter conditions on intermediate inputs. Estimates of TFP and its dispersion ratios will appear more variable using a value-added production function. These models can be estimated non-parametrically if the productivity term is Hicks-neutral.

The assumption of a Hicks-neutral productivity term implies that technological change does not change the factor shares of a firm. This ignores the reality that some technological improvements improve the efficiency of inputs like labor without changing the efficiency of others. Labor-augmenting productivity is an important component to economic models of growth. Therefore, understanding the contributions to firm-level growth in labor productivity can help explain recent patterns of economic growth as well as the phenomenon of decreasing labor's share of GDP. Despite its importance, recovering estimates of labor-augmenting productivity is an econometric challenge. In order to obtain consistent estimates of the production function, the econometrician must be able to correct endogeneity bias with multi-dimensional productivity. Doraszelski and Jaumandreu (2018) suggest an approach that uses the input mix of a firm to invert for factor-augmenting productivity. They use the ratio of material to labor inputs to proxy for the labor-augmenting term, then solve the remaining endogeneity from the Hicks-neutral term by an extension of the proxy variable approach. Their empirical strategy relies on a parametric specification for the production function so that the decision rules of labor and materials can be expressed as a known function of the data, which include wages, input prices and output prices. Data at this level is often not available to researchers. It remains to be seen whether similar factor-augmenting estimates can be captured in applications with fewer data requirements while also considering the econometric issues of simultaneity bias and unobservables in the proxy variables.

In this paper, my goal is to address both the issue of unobservables in the proxy variables as well as the non-Hicks neutrality of productivity. These dimensions will allow me to examine heterogeneous effects in firm technology, productivity, and input usage. I propose an identification strategy that is an extension of Hu *et al.* (2019)(hereafter HHS) which uses inputs as instrumental variables (IVs) in the framework of the non-classical measurement error model developed by Hu and Schennach (2008). Their approach uses conditional independence arguments which provides conditions for which a proxy variable instruments for another. It is important to note that the identification results of Hu and Schennach (2008) applies to non-separable models, however HHS pursue an alternative strategy by assuming a Cobb-Douglas production function and input demand functions that are additive in unobservables (productivity plus demand shocks). This facilitates less restrictive conditions for identification such as a non-parametric rank condition on the IVs, which is difficult to verify in practice. In addition, their model trivially satisfies a normalization assumption on the error term, which for non-separable models, would require centering a subset of parameters. Their assumptions motivate the construction of a GMM estimator which relies crucially on the separability of error terms in the model. However, one could question the structural conditions for which the input demand functions is additive in their unobservables. Therefore, a more flexible specification may alleviate these concerns although at the cost of higher-level econometric assumptions. In my paper, these assumptions are needed, however the advantage is that I can consider a richer set of estimates for the production function that has not been considered in previous approaches.

Unlike the GMM estimator proposed by Hu *et al.* (2019), I propose an estimator that can accommodate non-separability of the production function as well as unobservables in the proxy variables. The first extension allows me to capture the non-Hicks neutral effects of productivity. The nonparametric specification I consider does not require a parametric inversion strategy to capture these effects. Since I use inputs as IVs, prices are not needed to invert for productivity. I interpret the interactions between productivity and the inputs as a factor-efficiency effect which is calculated as average derivatives of the production function with respect to inputs and productivity. The second extension allows for heterogeneity in firm input responses with respect changes in their productivity. For example, firms may have heterogeneous responses in their hiring decisions due to an increase in automation. I also examine how firms adjust their inputs in response to latest changes in their productivity across the entire distribution of input demand. In order to capture the full extent of these heterogeneous responses, I adopt a quantile regression framework using the estimation procedure proposed by Arellano and Bonhomme (2016) for nonlinear panel data models.

I introduce the economic model and its restrictions in Section 2. In Section 3, I discuss nonparametric identification. In Section 4 and 5, I discuss estimation based on the econometric restrictions and practical implementation. In Section 6, I apply this estimator to US publicly-listed manufacturing firms. Section 7 concludes and provides direction for future research.

2 The Model of Firm Production

In this section I outline the model for the production function, productivity process, flexible inputs, and investment decisions.

2.1 Production Function

Consider a nonlinear model for a firm's gross-output production function (in logs)

$$y_{it} = f_t(k_{it}, l_{it}, m_{it}, \omega_{it}, \eta_{it}), \tag{1}$$

where y_{it} is firm i's output at time t, k_{it} denotes the capital, l_{it} denotes labor which can be flexibly chosen or dynamic and m_{it} denotes material inputs. The unobserved productivity is denoted by ω_{it} which is correlated to input choices of the firm. The unobserved production shocks are denoted by η_{it} which are assumed to be independent of input choices and productivity. The production function is assumed to be strictly increasing in η_{it} and can vary over time.

Other than monotonicity, I do not place any additional functional form assumptions on the production function. The rank of the unobservable production shock η_{it}, determines the ranking of a firm on the conditional distribution of output. This provides a Skorohod representation of the production function which will be important for developing the econometric restrictions of the model based on conditional quantiles instead of conditional means. This representation will also be used for the productivity equation and the input demand functions. Without loss of generality, I re-write specification for the production function as

$$y_{it} = Q_t^y(k_{it}, l_{it}, m_{it}, \omega_{it}, \eta_{it}), \quad \eta_{it} \sim Uniform(0,1), \tag{2}$$

where Q_t^y denotes the conditional quantile function of output. The productivity term enters the production function non-separably so that interactions between this term and the inputs capture non-Hicks neutral factor efficiency effects. These will be estimated as average derivatives of the production function which can be interpreted as the increase/decrease in marginal product when there is a small change in productivity levels. I summarize the restrictions on the production function with the following assumptions:

Assumption 2.1 *(Production Function)*

(a) The unanticipated production shocks η_{it} are iid over firms and time.

(b) The unanticipated production shock η_{it} follows a standard uniform distribution independent of $(\omega_{it}, k_{it}, l_{it}, m_{it})$.

(c) $\tau \to Q_t^y(k_{it}, l_{it}, m_{it}, \omega_{it}, \tau)$ is strictly increasing on $(0,1)$.

2.2 Productivity

Productivity, ω_{it}, is assumed to evolve according to a first-order Markov process:

$$\omega_{it} = Q_t^{\omega}(\omega_{it-1}, \xi_{it}), \quad \xi_{it} \sim Uniform(0,1), \tag{3}$$

where $\xi_{i1}, \ldots, \xi_{iT}$ are independent uniform random variables which represent innovation shocks to productivity. This specification is not standard in the proxy variable literature. A productivity process that is additive in the innovation shock is necessary to form conditional moment restrictions which identify parameters in the second stage of OP, LP, and ACF. The error term in their model is the difference between realized productivity and the firm's expected/predicted productivity. The contribution of non-separability in the innovation shock is that firm's expectations of future productivity can vary with the size of unanticipated shocks. For example, this allows for bad shocks to erase a history of high productivity and good shocks to erase a history of low productivity. This specification may better capture the nature of heterogeneous productivity evolution. I also consider an extension to endogenous productivity evolution by considering firm investment in knowledge through R&D activities similar to Doraszelski and Jaumandreu (2013). To this end, I consider the alternative specification for productivity

$$\omega_{it} = Q_t^{\omega}(\omega_{it-1}, r_{it-1}, \xi_{it}), \quad \xi_{it} \sim Uniform(0,1), \tag{4}$$

where r_{it-1} denotes R&D expenditures. In this model, ξ_{it} captures the uncertainties in productivity and the R&D process. I model the R&D process as

$$r_{it} = Q_t^{r}(k_{it}, \omega_{it}, \varrho_{it}), \quad \varrho_{it} \sim Uniform(0,1), \tag{5}$$

where ϱ_{it} captures unobserved factors affecting R&D. This extension allows me to examine productivity heterogeneity between firms that perform R&D and those who do not. In addition, I show that returns to productivity vary between firms who have different levels of uncertainty in the productivity and R&D process. This flexible model allows for nonlinear interactions of productivity, R&D, and innovation shocks. Therefore, an important empirical question is whether firms with large R&D expenditures who experience low productivity shocks have a higher/lower persistence of productivity history dependent on their productivity levels.

Industries with substantial periods of restructuring are characterized by entry and exit of firms due to changes in future expected productivity levels. Therefore, artificially balancing the data made lead to selection bias if firm's beliefs about future productivity is partially determined by their current productivity. Olley and Pakes (1996) show a particular form of bias in the production function estimates in the presence of non-random exit. It is not as straight-forward to provide a sign for the bias in a nonseparable quantile model and the tools for correcting selection in these models are still in development. Arellano and Bonhomme (2017) have made significant progress in this regard and propose a selection correction with cross-sectional data. In the Appendix, I propose a strategy to correct for non-random exit in non-linear panel data models using their framework. The main contribution of this extension is to show that sample selection may affect the entire distribution of productivity and hence the distribution of firm output. To summarize the restrictions on productivity, I provide the following assumptions:

Assumption 2.2 *(Productivity)*

(a) The productivity innovation shocks ξ_{it} are iid across firms and time.

(b) ξ_{it} follows a standard uniform distribution independent of previous period productivity ω_{it-1}.

(c) $\tau \to Q_t^{\omega}(\omega_{it-1}, \tau)$ is strictly increasing on $(0, 1)$.

2.3 Flexible Inputs

The firm chooses labor and intermediate inputs to maximize short-term profits. Since I do not restrict the functional form of the production function, it is not necessary to characterize the input decisions as a parametric function of the state variables. Accordingly, I specify the following labor decision rule:

$$l_{it} = Q_t^{\ell}(k_{it}, \omega_{it}, \epsilon_{\ell,it}), \quad \epsilon_{\ell,it} \sim Uniform(0, 1), \tag{6}$$

where $\epsilon_{\ell,it}$ are iid and independent of current period state variables. The additional unobservable captures sources of labor demand variation across firms. Using this representation, it is not necessary to describe the distinct sources of heterogeneity across firms; although this can include wages, labor adjustment costs, and other demand shocks to labor. Instead, I interpret it as the ranking index of the firm on the conditional labor distribution. Therefore, a higher $\tau \in (0, 1)$ corresponds to a firm who uses more labor conditional on capital and productivity than a firm with low τ index. With this representation, I can estimate the effects of productivity on labor usage. This is importance for understanding how firms hiring decisions are affected by technological developments such an increase in automation. I can also consider the case where labor is a dynamic decision variable. This can arise when there are significant hiring/firing costs or industries with high turn-over and employment contracts. A dynamic decision rule for labor can be written as:

$$l_{it} = Q_t^{\ell}(k_{it}, l_{it-1}, \omega_{it}, \epsilon_{\ell,it}), \quad \epsilon_{\ell,it} \sim Uniform(0,1), \tag{7}$$

where again $\epsilon_{\ell,it}$ is iid and independent of current period state variables including previous labor decisions. In the Appendix, I show how this model can be used to capture employment decisions in response to adjustment shocks to previous labor. This is important from a policy perspective to examine unemployment responses to structural changes which can depend on the magnitude of the shock as well as the size of the firm's labor force.

The firm chooses intermediate inputs to maximize profits. The decision rule is given by:

$$m_{it} = Q_t^{m}(k_{it}, l_{it}, \omega_{it}, \epsilon_{m,it}), \quad \epsilon_{m,it} \sim Uniform(0,1), \tag{8}$$

where $\epsilon_{m,it}$ are iid and independent of current period state variables. I assume material inputs are chosen simultaneously or after labor decisions are made. This is to be consistent with the specification for dynamic labor mentioned earlier. I summarize the assumption on the flexible inputs below:

Assumption 2.3 *(Flexible Inputs)*

(a) The unobserved input demand shocks $\epsilon_{l,it}$ and $\epsilon_{m,it}$ are iid across firms and time.

(b) $\epsilon_{l,it}$ and $\epsilon_{m,it}$ follow a standard uniform distribution independent of (k_{it}, ω_{it}).

(c) $\tau \to Q_t^{\ell}(k_{it}, \omega_{it}, \tau)$ and $Q_t^{m}(k_{it}, \omega_{it}, \tau)$ are strictly increasing on $(0,1)$.

2.4 Investment

Investment decisions on the other hand are the solution to a long-run expected profit maximization problem:

$$I_{it}^* = \iota_t(K_{it}, \omega_{it}) = \underset{I_t \geq 0}{\operatorname{argmax}} \left[\Pi_t(K_{it}, \omega_{it}) - c(I_{it}) + \beta \mathbb{E}\big[V_{t+1}(K_{it+1}, \omega_{it+1}) | \mathcal{I}_t\big] \right], \tag{9}$$

where $\pi_t(\cdot)$ is current period profits as a function of the state variables. Current costs to investment are given by $c(I_t)$ and β is the firm's discount factor. I introduce an empirical investment rule for (9) given by

$$i_{it}^* = Q_t^i(k_{it}, \omega_{it}, \zeta_{it}), \quad \zeta_{it} \sim Uniform(0,1). \tag{10}$$

One possible interpretation for ζ_{it} is a shock to investment demand that increases the marginal productivity of capital. In the case where there are many zero observations of investment, I can write a censored version as $i_{it} = \max\{0, i_{it}^*\}$. Although this is not the case in the data considered in this paper, allowing for censoring in investment would be crucial for other extending this methodology to other empirical applications. This is easily implemented in my quantile modelling due to the equivariance property of quantiles. Capital accumulates to the following generalized law of motion

$$K_{it} = \kappa(K_{it-1}, I_{it-1}, \upsilon_{it-1}), \tag{11}$$

where I_{it-1} denotes firm investment in the prior period. Under this specification, capital is determined in period $t-1$. I introduce a random error term, υ_{it-1}, which eliminates the deterministic relationship of the capital accumulation process. This specification is also used by HHS. To summarize the restrictions on the capital process and investment, I assume the following:

Assumption 2.4 *(Capital Accumulation and Investment)*

(a) The unobserved investment demand shocks ζ_{it} is iid across firms and time.

(b) ζ_{it} follows a standard uniform distribution independent of (k_{it}, ω_{it}).

(c) The production shock η_{it} and ζ_{it} are independent conditional on $(k_{it}, l_{it}, m_{it}, \omega_{it})$. In addition, υ_{it} is independent of η_{it} conditional on $(k_{it}, l_{it}, m_{it}, \omega_{it})$

(d) $\tau \to Q_t^i(k_{it}, \omega_{it}, \tau)$ is strictly increasing on $(0,1)$

The next section uses the assumptions on the production function, productivity, flexible inputs, and investment to show that the model is nonparametrically identified. In addition, the assumptions also form econometric restrictions on the model which I use to estimate firm heterogeneity using quantile regression.

3 Identification

In this section I show that the conditional densities corresponding to the production function, productivity, input decisions and investment are nonparametrically identified using Hu and Schennach (2008). To show this, I introduce notation. Let $Z_t = (l_t, k_t, m_t, k_{t+1})$ denote conditioning variables where I have dropped the i subscript for convenience. Assume the following:

Assumption 3.1 *(Conditional Independence):*
$f(y_t|y_{t+1}, I_t, \omega_t, Z_t) = f(y_t|\omega_t, Z_t)$ *and* $f(I_t|y_{t+1}, \omega_t, Z_t) = f(I_t|\omega_t, Z_t)$

The first equality of Assumption 3.1 states that conditional on ω_t and Z_t, y_{t+1} and I_t do not provide any additional information about y_t. The second equality states that conditional on ω_t and Z_t, y_{t+1} does not provide any additional information about I_t. These are satisfied by mutual independence assumptions on η_t and ζ_t conditional on $(\omega_t, k_t, l_t, m_t)$ and the fact that η_{it} is assumed to be conditionally independent over time. The next assumption is more technical and requires the following preliminary definition:

Definition 3.1 *(Integral Operator) Let a and b denote random variables with supports $\mathcal{A}$ and $\mathcal{B}$. Given two corresponding spaces $\mathcal{G}(\mathcal{A})$ and $\mathcal{G}(\mathcal{B})$ of functions with domains $\mathcal{A}$ and $\mathcal{B}$, let $L_{b|a}$ denote the operator mapping $g \in \mathcal{G}(\mathcal{A})$ to $L_{b|a}g \in \mathcal{G}(\mathcal{B})$ defined by*

$$[L_{b|a}g](b) \equiv \int_{\mathcal{A}} f_{b|a}(b|a)g(a)da,$$

where $f_{b|a}$ denotes the conditional density of b given a.

With this definition, the uniqueness of an operator mapping can be defined by the next assumption.

Assumption 3.2 *(Injectivity): The operators $L_{y_t|Z_t,\omega_t}$ and $L_{y_{t+1}|Z_t,\omega_t}$ are injective*

This allows me to take inverses of the operators. Consider the operator $L_{y_t|Z_t,\omega_t}$. Following HS, injectivity of this operator can be interpreted as it's corresponding density $f_{y_t|Z_t,\omega_t}(y_t|Z_t, \omega_t)$ having sufficient variation in ω_t given Z_t. This assumption is often phrased as a completeness condition in the nonparametric IV literature on the density $f_{y_t|k_t,\omega_t}(y_t|Z_t, \omega_t)$. More formally, for a given $Z_t \in Supp(Z_t)$

$$\int f_{y_t|Z_t,\omega_t}(y_t|Z_t, \omega_t)g(\omega_t)d\omega_t = 0 \tag{12}$$

for all y_t implies $g(\omega_t) = 0$ for all ω_t. For injectivity of the second operator $L_{y_{t+1}|Z_t,\omega_t}$, one can consider y_{t+1} having sufficient variation for different values of ω_t given Z_t. Since productivity is specified as a Markov process and is highly persistent over time, this assumption is intuitive.

This assumption is more restrictive than that of HHS. Since their model is separable in ω_t they are able utilize convolution type arguments which require conditional independence assumptions as well as regularity conditions on the conditional characteristic functions. Injectivity is the cost of specifying a more general production function. I also require two additional assumptions.

Assumption 3.3 *(Uniqueness): For any $\bar{\omega}_t, \tilde{\omega}_t \in \Omega$, the set $\{f_{I|\omega,Z}(I_t|\bar{\omega}_t, Z_t) \neq f_{I|\omega,Z}(I_t|\tilde{\omega}_t, Z_t)\}$ has positive probability whenever $\bar{\omega}_t \neq \tilde{\omega}_t$*

This assumption is relatively weak and is satisfied if there is conditional heteroskedasticity in $f_{I|\omega,Z}$ or if any functional of its distribution is strictly increasing in ω_t. For example, this assumption is satisfied if $E[I_t|\omega_t, Z_t]$ is strictly increasing in ω_t which is similar to the invertibility conditions required in Olley and Pakes (1996). The flexible accumulation process for capital specified by (11) is necessary for this condition to hold, otherwise investment would be completely determined by k_{t+1} and k_t. In my empirical application, the average investment response to productivity is positive which supports using the monotonicity restrictions for identification.

Assumption 3.4 *(Normalization): There exists a functional Γ such that $\Gamma[f_{y|\omega,Z}(y_t|\omega_t, Z_t)] = \omega_t$*

This functional does not need to be known. It is sufficient to consider a known function of the data distribution as shown by Arellano and Bonhomme (2016). For a general nonseparable panel model this assumption is satisfied if $E[y_t|\omega_t, Z_t]$ is strictly increasing in ω_{it}. Then, one could normalize $\omega_t = E[y_t|\omega_t, Z_t]$. In my empirical application, I use a non-separable translog production function. In this case, the normalization can be achieved by $E[y_t|\omega_t, 0] = \omega_t$ which is standard in the production functions with a separable ex-post shock. In my model this requires restrictions on a subset of parameters in the model. With these assumptions, I can now state the first part of the identification result.

Theorem 3.1 *Under Assumptions 3.1, 3.2, 3.3, and 3.4, given the observed density* $f_{y_t,I_t|y_{t+1},Z_t}$*, the equation*

$$f_{y_t,I_t|y_{t+1},Z_t}(y_t,I_t|y_{t+1},Z_t) = \int f_{y_t|\omega_t,Z_t}(y_t|\omega_t,Z_t) f_{I_t|\omega_t,Z_t}(I_t|\omega_t,Z_t) f_{\omega_t|y_{t+1},Z_t}(\omega_t|y_{t+1},Z_t) d\omega_t \tag{13}$$

admits a unique solution for $f_{y_t|\omega_t,Z_t}$, $f_{I_t|\omega_t,Z_t}$ *and* $f_{\omega_t|y_{t+1},Z_t}$

Proof: See Appendix B

This result identifies the conditional density of output and investment. It also identifies the marginal distribution for productivity and the input decision rules as I show in Appendix B. Additional assumptions are needed to identify the Markov transition function for productivity, $f_{\omega_{t+1}|\omega_t}(\omega_{t+1}|\omega_t)$. The requirements for identification of this density are different under two cases involving stationarity and non-stationarity of the density $f_{y_t|\omega_t,Z_t}(y_t|\omega_t,Z_t)$.

Corollary 3.1 *(Stationarity): Suppose that the production function is stationary i.e.* $f_{y_t|\omega_t,Z_t} = f_{y_1|\omega_1,Z_1} \forall t \in \{1,\cdots,T\}$*. Then, under Assumptions 3.1, 3.2, 3.3, and 3.4, the observed density* $f_{y_t,I_t|y_{t+1},Z_t}$ *uniquely determines the density* $f_{\omega_{t+1}|\omega_t}$ *for any* $t \in \{1,\ldots,T-1\}$

Proof: See Appendix B

Corollary 3.2 *(Non-Stationary): Under Assumptions 3.1, 3.2, 3.3, and 3.4, the observed density* $f_{y_{t+1},I_{t+1}|y_{t+2},Z_{t+1}}$ *uniquely determines the density* $f_{\omega_{t+1}|\omega_t}$ *for any* $t \in \{1,\ldots,T-2\}$

Proof: See Appendix B.

The main conclusion of these two corollaries is that under the condition of stationarity, the productivity process can be identified with $T = 2$ observations per firms whereas under non-stationarity, the productivity process is identified with $T = 3$ observations per firm. The number of time periods required for identification increases with the length of the autoregressive process. These data requirements are similar to the control function approach where the instrument set often includes secondary lags of inputs.

4 Estimation Strategy

This section presents the model specifications and econometric strategy that is used in the application. I consider the following functional form for the production function:

$$\begin{aligned}
&Q_t^y(k_{it}, l_{it}, m_{it}, \omega_{it}, \tau) = \\
&\quad \gamma_0(\tau) + (\gamma_k(\tau) + \sigma_k(\tau)\omega_{it})k_{it} + (\gamma_l(\tau) + \sigma_l(\tau)\omega_{it})l_{it} + (\gamma_m(\tau) + \sigma_m(\tau)\omega_{it})m_{it} \\
&\quad + (\gamma_{kl}(\tau) + \sigma_{kl}(\tau)\omega_{it})k_{it}l_{it} + (\gamma_{lm}(\tau) + \sigma_{lm}(\tau)\omega_{it})l_{it}m_{it} + (\gamma_{km}(\tau) + \sigma_{km}(\tau)\omega_{it})k_{it}m_{it} \\
&\quad + (\gamma_{kk}(\tau) + \sigma_{kk}(\tau)\omega_{it})k_{it}^2 + (\gamma_{ll}(\tau) + \sigma_{ll}(\tau)\omega_{it})l_{it}^2 + (\gamma_{mm}(\tau) + \sigma_{mm}(\tau)\omega_{it})m_{it}^2 + \sigma_\omega(\tau)\omega_{it}.
\end{aligned} \tag{14}$$

This corresponds to a Translog production function with first-order interactions of productivity. The coefficients $\gamma(\tau) = (\gamma_k(\tau), \gamma_l(\tau), \gamma_m(\tau), \gamma_{kl}(\tau), \gamma_{lm}(\tau), \gamma_{km}(\tau), \gamma_{kk}(\tau), \gamma_{ll}(\tau), \gamma_{mm}(\tau))$ are the output elasticities which capture the non-linear effects of inputs on production. The terms $\sigma(\tau) = (\sigma_k(\tau), \sigma_l(\tau), \sigma_m(\tau), \sigma_{kl}(\tau), \sigma_{lm}(\tau), \sigma_{km}(\tau), \sigma_{kk}(\tau), \sigma_{ll}(\tau), \sigma_{mm}(\tau), \sigma_\omega(\tau))$ capture the non-Hicks neutral effects of productivity on inputs and a Hicks-neutral effect that varies across quantiles measured by $\sigma_\omega(\tau)$. To simplify notation, I denote the vector of production function parameters as $\beta(\tau) = (\gamma(\tau), \sigma(\tau))$ and write the conditional quantile indexed by these parameters by $Q_t^y(k_{it}, l_{it}, m_{it}, \omega_{it}; \beta(\tau))$. A similar specification of (14) was considered by Ackerberg and Hahn (2015) for a conditional mean model. This more flexible specification allows for a variety of heterogeneous estimates. I calculate output elasticities of inputs as *individual* quantile marginal effects which can vary over the conditional distribution of output and the distribution of input demand. For example, consider the quantile marginal effect of capital as

$$\boldsymbol{\beta}_k(\tau_\eta, \tau_k) = \mathbb{E}\left[\frac{\partial Q^y(Q(k_{it}; \tau_k), l_{it}, m_{it}, \omega_{it}; \beta(\tau_\eta))}{\partial k_{it}}\right], \tag{15}$$

where τ_η denotes the rank on the conditional output distribution and τ_k denotes the rank of the unconditional capital distribution. This effect is calculated by averaging over ω_{it} as well as l_{it} and m_{it} evaluated at the fixed percentiles of capital. The non-Hicks neutral effects can be captured by average partial derivatives of the production function with respect to inputs and productivity. For example, the non-Hicks neutral effect of capital can be captured by

$$\boldsymbol{\sigma}_k(\tau_\eta, \tau_k) = \mathbb{E}\left[\frac{\partial^2 Q^y(Q(k_{it}; \tau_k), l_{it}, m_{it}, \omega_{it}; \beta(\tau_\eta))}{\partial k_{it}\omega_{it}}\right]. \tag{16}$$

In order to account for the unobserved productivity in the marginal effect, I propose a simulation based method which uses the estimated parameters of the model in the next section. This allows a better visualization for heterogeneous estimates as opposed to reporting the individual coefficients in the model.

The Markov process for productivity is estimated using a third-order polynomial:

$$Q^{\omega}(\omega_{it-1}, \tau) = \rho_0(\tau) + \rho_1(\tau)\omega_{it-1} + \rho_2(\tau)\omega_{it-1}^2 + \rho_3(\tau)\omega_{it-1}^3. \tag{17}$$

This allows me to capture heterogeneous persistence of productivity which can depend on the level of previous productivity and the size of its innovation shock. When I augment the productivity model with R&D activities, I consider the following specification which is similar to Doraszelski and Jaumandreu (2013)

$$Q^{\omega}(\omega_{it-1}, r_{it-1}, \tau) = \mathbb{1}\{R_{it-1} = 0\}Q^{\omega}(\omega_{it-1}, \tau) + \mathbb{1}\{R_{it-1} > 0\}Q^{\omega,r}(\omega_{it-1}, r_{it-1}, \tau). \tag{18}$$

This allows a firm to adopt corner solutions to R&D expenditure represented by the different functions corresponding to positive or zero R&D expenditures. The quantile function $Q^{\omega,r}(\omega_{it-1}, r_{it-1}, \tau)$ can be expressed as some non-linear function. In addition, the R&D process is specified as

$$Q^{r}(k_{it}, \omega_{it}, \tau) = \sum_{j=1}^{J} \rho_j^r(\tau)\phi_{r,j}(k_{it}, \omega_{it}), \tag{19}$$

where $\phi_{r,j}$ is a non-linear function that is approximated by a tensor product Hermite polynomial. This function can capture interaction effects between capital and productivity. In the Appendix, I show that the equation in (20) must be modified to account for the fact that unobserved selection alters the productivity distribution rankings. I specify an initial condition for productivity by

$$Q^{\omega_1}(k_{i1}, \tau) = \sum_{j=1}^{J} \rho_{\omega_1,j}(\tau)\phi_{\omega_1,j}(k_{i1}), \tag{20}$$

where $\phi_{\omega_{i1},j}$ is approximated by a second degree polynomial in initial capital k_{i1}. I specify the labor input demand equation as follows:

$$Q_t^{\ell}(k_{it}, \omega_{it}, \tau) = \sum_{j=1}^{J} \alpha_{\ell,j}(\tau)\phi_{\ell,j}(k_{it}, \omega_{it}), \tag{21}$$

where $\phi_{\ell,j}$ is another nonlinear function. In practice, I approximate this function by a tensor product Hermite polynomial of degree $(3,3)$. In the case where I consider labor adjustment frictions, I specify the following model for labor

$$Q_t^{\ell}(k_{it}, l_{it-1}, \omega_{it}, \tau) = \sum_{j=1}^{J} \alpha_{\ell,j}(\tau)\phi_{\ell,j}(k_{it}, \ell_{it-1}, \omega_{it}), \tag{22}$$

which is approximated by another Hermite polynomial of degree $(3,3,3)$. I specify the material input demand equation as follows:

$$Q_t^{m}(k_{it}, l_{it}, \omega_{it}, \tau) = \sum_{j=1}^{J} \alpha_{m,j}(\tau)\phi_{m,j}(k_{it}, l_{it}, \omega_{it}), \tag{23}$$

where $\phi_{m,j}$ is another Hermite polynomial of degree $(2,2,2)$. I specify the investment demand equation (in logs) as:

$$i_{it}^{*} = Q_t^{i}(k_{it}, \omega_{it}, \tau) = \sum_{j=1}^{J} \delta_j(\tau)\phi_{i,j}(k_{it}, \omega_{it}), \tag{24}$$

where $\phi_{\iota,j}$ is specified similarly as the labor decision rule. In the case where investment is censored, I can write

$$i_{it} = \max\{0, i_{it}^{*}\} = \max\{0, Q_t^{i}(k_{it}, \omega_{it}; \zeta_{it})\}, \tag{25}$$

in which case the conditional quantiles can be written as

$$Q_{\tau}(i_{it}|k_{it}, \omega_{it}) = \max\{0, \sum_{j=1}^{J} \delta_j(\tau)\phi_{i,j}(k_{it}, \omega_{it})\}, \tag{26}$$

due to the equivariance property of quantiles. The censored quantile regression model avoids distributional assumptions in estimation at the cost of computational complexity.

5 Implementation

The following conditional moment restrictions hold as an implication of Assumptions 2.1-2.4. For the production function:

$$\mathbb{E}\left[\Psi_\tau\left(y_{it} - Q_t(y_{it}|k_{it}, l_{it}, m_{it}, \omega_{it}; \beta(\tau))\right)\middle| k_{it}, l_{it}, m_{it}\right] = 0. \tag{27}$$

For the labor input demand:

$$\mathbb{E}\left[\Psi_\tau\left(l_{it} - \sum_{j=1}^{J} \alpha_{l,j}(\tau)\phi_{l,j}(k_{it}, \omega_{it})\right)\middle| k_{it}, \omega_{it}\right] = 0. \tag{28}$$

For the material input demand:

$$\mathbb{E}\left[\Psi_\tau\left(m_{it} - \sum_{j=1}^{J} \alpha_{m,j}(\tau)\phi_{m,j}(k_{it}, \omega_{it})\right)\middle| k_{it}, \omega_{it}\right] = 0. \tag{29}$$

For the investment demand function:

$$\mathbb{E}\left[\Psi_\tau\left(i_{it} - \sum_{j=1}^{J} \delta_j(\tau)\phi_{\iota,j}(k_{it}, \omega_{it})\right)\middle| k_{it}, \omega_{it}\right] = 0. \tag{30}$$

and for the productivity process for $t \geq 2$:

$$\mathbb{E}\left[\Psi_\tau\left(\omega_{it} - \rho_0(\tau) - \rho_1(\tau)\omega_{it-1} - \rho_2(\tau)\omega_{it-1}^2 - \rho_3(\tau)\omega_{it-1}^3\right)\middle| \omega_{it-1}\right] = 0, \tag{31}$$

where $\Psi_\tau(u) = \tau - \mathbb{1}\{u < 0\}$. Initial productivity is assumed to follow a normal distribution, $\omega_{i1} \sim N(\mu_{\omega_1}, \sigma_{\omega_1}^2)$. Estimating the above conditional moment restrictions is infeasible due to the unobserved productivity component. Therefore, I use the following unconditional moment restrictions and posterior distributions for ω_{it} to integrate out the unobserved productivity. To fix ideas, consider the following unconditional moment restriction corresponding to the production function from Equation (27):

$$\mathbb{E}\left[\int_\Omega \Psi_\tau\left(y_{it} - Q_t(y_{it}|k_{it}, l_{it}, m_{it}, \omega_{it}; \beta(\tau))\right) \otimes \begin{pmatrix} 1 \\ k_{it} \\ l_{it} \\ m_{it} \\ \omega_{it} \end{pmatrix} g_i(\omega_{it}; \theta(\cdot)) d\omega_{it}\right] = 0, \tag{32}$$

where $\theta(\cdot) = (\beta(\cdot), \alpha_l(\cdot), \alpha_m(\cdot), \delta(\cdot), \rho(\cdot))$ denotes the vector of all of the model parameters.

The posterior density $g_i(\omega_{it};\theta(\cdot)) = f(\omega_{it}|y_{it}, k_{it}, l_{it}, m_{it}, i_t, \theta(\cdot))$ conditions on the entire set of model parameters. This is due to the equivalence of the density of a random variable and the inverse of the derivative of its quantile function. Therefore, it is impossible to estimate the model parameters in a τ-by-τ procedure. To eliminate the intractability of this problem, the continuous model parameters are approximated by spline functions. For example, the function $\beta(\tau)$ is approximated by a piecewise-polynomial interpolating spline on a grid $[\tau_1, \tau_2], [\tau_3, \tau_4], \ldots, [\tau_{Q-1}, \tau_Q]$, contained in the unit interval and is constant on $(0, \tau_1]$ and $[\tau_Q, 1)$ except for the intercept parameter which is modeled on the tail intervals in the next paragraph. The function is approximated by:

$$\beta(\tau) = \beta(\tau_q) + \frac{\tau - \tau_q}{\tau_{q+1} - \tau_q}\big(\beta(\tau_{q+1}) - \beta(\tau_q)\big), \qquad \tau_q < \tau \leq \tau_{q+1}$$

The intercept coefficient β_0 is specified as the quantile of an exponential distribution on $(0, \tau_1]$ (indexed by λ^-) and $[\tau_{Q-1}, 1)$ (indexed by λ^+). For example, the intercepts for the production function are given by:

$$\beta_0(\tau) = \beta_0(\tau_1) + \frac{\ln(\tau/\tau_1)}{\lambda_\beta^-}, \quad \tau \leq \tau_1$$

and

$$\beta_0(\tau) = \beta_0(\tau_Q) + \frac{\ln(1 - \tau/1 - \tau_Q)}{\lambda_\beta^+}, \quad \tau > \tau_Q.$$

The remaining functional parameters are modeled similarly. The usefulness of the piece-wise linear spline is that the posterior density now has a closed form expression without relying on strong distributional assumptions. For example, the density corresponding to the production function can be written as:

$$\begin{aligned} f_{y_t|k_t,l_t,m_t,\omega_t}(y_t|k_t, l_t, m_t, \omega_t; \beta) = &\sum_{q=1}^{Q-1} \frac{\tau_{q+1} - \tau_q}{Q_t(\cdot|\cdot;\beta(\tau_{q+1})) - Q_t(\cdot|\cdot;\beta(\tau_q))} \mathbb{1}\{Q_t(\cdot|\cdot;\beta(\tau_q)) < y_t \leq Q_t(\cdot|\cdot;\beta(\tau_{q+1}))\} \\ &+ \tau_1\lambda_\beta^- \exp\big(\lambda_\beta^-(y_t - Q_t(\cdot|\cdot;\beta(\tau_1)))\big)\mathbb{1}\{y_t \leq Q_t(\cdot|\cdot;\beta(\tau_1))\} \\ &+ (1-\tau_Q)\lambda_\beta^+ \exp\big(-\lambda_\beta^+(y_t - Q_t(\cdot|\cdot;\beta(\tau_Q)))\big)\mathbb{1}\{y_t > Q_t(\cdot|\cdot;\beta(\tau_Q))\}. \end{aligned}$$

The exponential parameters are updated using a likelihood approach:

$$\lambda_\beta^- = \frac{-\mathbb{E}[\int \mathbb{1}\{y_t \leq Q_t(\cdot|\cdot;\beta(\tau_1))\}g(\omega_t;\theta(\cdot))d\omega_t]}{\mathbb{E}[\int (y_t - Q_t(\cdot|\cdot;\beta(\tau_1)))\mathbb{1}\{y_t \leq Q_t(\cdot|\cdot;\beta(\tau_1))\}g(\omega_t;\theta(\cdot))d\omega_t]}$$

and

$$\lambda_\beta^+ = \frac{\mathbb{E}[\int \mathbb{1}\{y_t > Q_t(\cdot|\cdot;\beta(\tau_Q))\}g(\omega_t;\theta(\cdot))d\omega_t]}{\mathbb{E}[\int (y_t - Q_t(\cdot|\cdot;\beta(\tau_Q)))\mathbb{1}\{y_t > Q_t(\cdot|\cdot;\beta(\tau_Q))\}g(\omega_t;\theta(\cdot))d\omega_t]}.$$

In order to estimate the model, the integral inside the expectation of Equation (32) needs to be approximated. This can be done using quadrature methods or Monte Carlo integration and converting the problem into a weighted quantile regression. Due to the high-dimensionality of my application, I choose to use a random-walk Metropolis Hastings algorithm to compute the integral. This becomes a Monte Carlo Expectation Maximization (MCEM) procedure where the maximization step is performed using quantile regression. The algorithm procedures as follows. Given an initial parameter value $\hat{\theta}^0$. Iterate on $s = 0, 1, 2, \ldots$ in the following two-step procedure until convergence to a stationary distribution:

1. *Stochastic E-Step*: Draw M values $\omega_i^{(m)} = (\omega_{i1}^{(m)}, \omega_{i2}^{(m)}, \ldots, \omega_{iT}^{(m)})$ from

$$\begin{aligned} g_i(\omega_{it};\hat{\theta}^{(s)}) = f(\omega_{it}|y_{it}, k_{it}, l_{it}, m_{it}, i_t, ;\hat{\theta}^{(s)}) \propto \\ \prod_{t=1}^{T} f(y_{it}|k_{it}, l_{it}, m_{it}, \omega_{it};\hat{\beta}^{(s)}) f(l_{it}|k_{it}, \omega_{it};\hat{\alpha}_l^{(s)}) f(m_{it}|k_{it}, \omega_{it};\hat{\alpha}_m^{(s)}) \\ \times f(i_{it}|k_{it}, \omega_{it};\hat{\delta}^{(s)}) \prod_{t=2}^{T} f(\omega_{it}|\omega_{it-1};\hat{\rho}^{(s)}) f(\omega_{i1};\hat{\rho}_1^{(s)}) \end{aligned}$$

2. *Maximization Step*: For $q = 1, \ldots, Q$, solve

$$\begin{aligned}
\hat{\boldsymbol{\beta}}(\tau_q)^{(s+1)} &= \underset{\boldsymbol{\beta}(\tau_q)}{\text{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{m=1}^{M} \psi_{\tau_q}\left(y_{it} - Q_t(y_{it}|k_{it}, l_{it}, m_{it}, \omega_{it}^{(m)};\beta(\tau_q))\right) \\
\hat{\boldsymbol{\alpha}}_l(\tau_q)^{(s+1)} &= \underset{\boldsymbol{\alpha}_l(\tau_q)}{\text{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{m=1}^{M} \psi_{\tau_q}\left(l_{it} - \sum_{j=1}^{J} \alpha_{l,j}(\tau_q)\phi_{l,j}(k_{it}, \omega_{it}^{(m)})\right) \\
\hat{\boldsymbol{\alpha}}_m(\tau_q)^{(s+1)} &= \underset{\boldsymbol{\alpha}_m(\tau_q)}{\text{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{m=1}^{M} \psi_{\tau_q}\left(m_{it} - \sum_{j=1}^{J} \alpha_{m,j}(\tau_q)\phi_{m,j}(k_{it}, \omega_{it}^{(m)})\right) \\
\hat{\boldsymbol{\delta}}(\tau_q)^{(s+1)} &= \underset{\boldsymbol{\delta}(\tau_q)}{\text{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{m=1}^{M} \psi_{\tau_q}\left(i_{it} - \sum_{j=1}^{J} \delta_j(\tau_q)\phi_{\iota,j}(k_{it}, \omega_{it}^{(m)})\right) \\
\hat{\boldsymbol{\rho}}(\tau_q)^{(s+1)} &= \underset{\boldsymbol{\rho}(\tau_q)}{\text{argmin}} \sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{m=1}^{M} \psi_{\tau_q}\left(\omega_{it}^{(m)} - \rho_0(\tau_q) - \rho_1(\tau_q)\omega_{it-1}^{(m)} - \rho_2(\tau_q)\omega_{it-1}^{(m)2} - \rho_3(\tau_q)\omega_{it-1}^{(m)3}\right),
\end{aligned}$$

where $\psi_\tau(u) = (\tau - \mathbb{1}\{u < 0\})u$ is the "check" function from quantile regression. Since initial productivity is specified as normally distributed, its parameters can be estimated as $\hat{\rho}_1 = (\hat{\mu}_{\omega_1}, \hat{\sigma}^2_{\omega_1})$. The exponential parameters for the intercept coefficients (e.g. the production function) are updated from:

$$\hat{\lambda}_\beta^{-(s)} = \frac{-\sum_{n=1}^N \sum_{t=1}^T \sum_{m=1}^M \mathbb{1}\{y_t \leq Q_t(\cdot|\cdot, \omega_{it}^{(m)}; \hat{\beta}(\tau_1)^{(s)})\}}{\sum_{n=1}^N \sum_{t=1}^T \sum_{m=1}^M (y_t - Q_t(\cdot|\cdot, \omega_{it}^{(m)}; \hat{\beta}(\tau_1)^{(s)})\mathbb{1}\{y_t \leq Q_t(\cdot|\cdot, \omega_{it}^{(m)}; \hat{\beta}(\tau_1)^{(s)})\}}$$

and

$$\hat{\lambda}_\beta^{+(s)} = \frac{\sum_{n=1}^N \sum_{t=1}^T \sum_{m=1}^M \mathbb{1}\{y_t > Q_t(\cdot|\cdot, \omega_{it}^{(m)}; \hat{\beta}(\tau_Q)^{(s)})\}}{\sum_{n=1}^N \sum_{t=1}^T \sum_{m=1}^M (y_t - Q_t(\cdot|\cdot, \omega_{it}^{(m)}; \hat{\beta}(\tau_Q)^{(s)})\mathbb{1}\{y_t > Q_t(\cdot|\cdot, \omega_{it}^{(m)}; \hat{\beta}(\tau_Q)^{(s)})\}}.$$

In this setting, it is computationally efficient to take $M = 1$ in the MCEM algorithm and report estimates of the average $\tilde{S} = S/2$ draws. This is known as the stochastic EM algorithm (stEM) of Celeux and Diebolt (1985). The sequence of maximizers $\hat{\theta}^{(s)}$ is a time-homogeneous Markov chain which, if ergodic, will converge to its stationary distribution. Nielsen (2000) provides sufficient conditions for ergodicity and asymptotic properties of the estimator when the "M-step" is solved using maximum likelihood. Arellano and Bonhomme (2016) discusses the asymptotic properties of the estimator when the M-step is solved using quantile regression. There are many components to this model that complicates asymptotic inference such as the dimension of the series approximations and the number of knots in the interpolating spline. Therefore, I do not pursue asymptotic properties of the estimator in this paper.

6 Application

I apply the estimator to data on US manufacturing firms from the Standard and Poors Compustat database. The sample covers publicly traded firms and contains data from their financial statements. I collect a sample between 1997 and 2016 on sales, capital expenditures, property, plant, and equipment, employees, and other expenses to construct output, investment, capital stock, labor, and material inputs. The financial data is deflated using 3-digit deflators from the NBER-CES Manufacturing Industry Database. After data cleaning, there are a total of $N = 2961$ firms with an average of 1545 firms per year. Summary statistics are provided in Appendix A.

I use the stochastic EM algorithm presented in this paper. I run the estimation procedure with 500 random walk Metropolis-Hastings steps and 200 EM steps with $M = 1$. I experimented with many different types of proposal distributions for the Metropolis-Hastings step.

I use a uniform distribution centered at the current draw of productivity with step size equal to 0.5. The final estimators are used to simulate productivity from its initial conditions and the decision rules for investment, labor and materials. In order to simulate output, I need to specify an accumulation process for capital. The process specified in in Equation 11 is flexible and I use it to accumulate capital using the perpetual inventory method with the depreciation rates fixed at 0.1. I have found that estimates are similar using different types of capital accumulation processes and depreciation rates.

Using the simulated data, I construct a number of structural estimates. For example consider the *average quantile marginal effect* of capital on output:

$$\bar{\beta}_k(\tau) = \mathbb{E}\left[\frac{\partial Q_t(y_{it}|k_{it}, l_{it}, m_{it}, \omega_{it}; \beta(\tau))}{\partial k_{it}}\right],$$

where the expectation is taken with respect to the distributions of $(l_{it}, m_{it}, \omega_{it})$ conditional on k_{it} at the average level of capital stock. I compute this estimator for each input and the non-Hicks neutral effects. I use the same formulation to compute estimates of average quantile productivity persistence in addition to computing *individual quantile effects* of the production function with respect to inputs (capital) by:

$$\beta(\tau_\eta, \tau_k) = \mathbb{E}\left[\frac{\partial Q_{\tau_\eta}(y_{it}|Q_{\tau_k}(k_{it}), l_{it}, m_{it}, \omega_{it}; \beta(\tau_\eta))}{\partial k_{it}}\right],$$

where the expectation is taken with respect to the distributions of $(l_{it}, m_{it}, \omega_{it})$ conditional on k_{it} evaluated at different percentiles of capital. I also calculate the individual quantile effects of productivity given by:

$$\hat{\rho}(\omega_{it-1}, \tau) = \hat{\rho}_1(\tau) + 2\hat{\rho}_2(\tau)\omega_{it-1} + 3\hat{\rho}_3(\tau)\omega_{it-1}^2.$$

Also of interest are the dynamic effects of productivity innovation shocks on inputs. Similar to Hu *et al.* (2019) I estimate how quickly firms respond to shocks to current productivity. This analysis would show whether input decision rules for labor, materials and capital are subject to substantial adjustment frictions. For example, if the finding is that labor responds positively to increases in productivity, then policies designed to increase productivity may have a faster effect depending on how quickly the firm is able to adjust its work force which has implications for some labor market outcomes. My model allows me to examine this effect on two different dimensions: the size of the labor demand across firms and the size of the productivity shock. This estimator can be given by:

$$\hat{l}(\tau_\xi, \tau_{\epsilon_l}) = \mathbb{E}\left[\frac{\partial Q_{\tau_{\epsilon_l}}\big(l_{it}|k_{it}, Q_{\tau_\xi}(\omega_{it}|\omega_{it-1})\big)}{\partial \omega_{it}} \times \left(\frac{\partial Q_{\tau_\xi}(\omega_{it}|\omega_{it-1})}{\partial \xi_{it}}\right)\right],$$

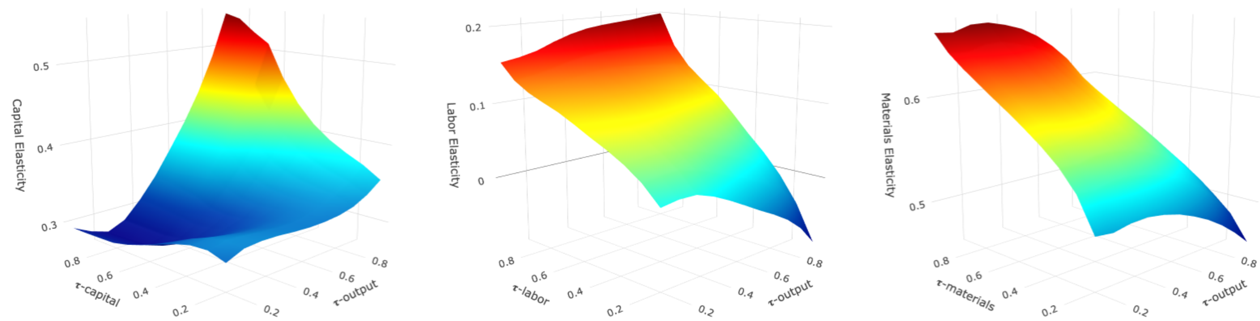
where $\partial Q_{\tau_\xi}(\omega_{it}|\omega_{it-1})/\partial \xi_{it}$ can be approximated by finite differences. In practice, I simulate impulse response functions under various innovation shocks to productivity and input demand functions under some initial conditions which I will discuss in the next section.

6.1 Empirical Results

6.1.1 Production Function Estimates

Estimates of the heterogeneous production function elasticities are shown in Figure 1 and 2. In the first panel of Figure 1, I report the estimates of the average capital elasticity evaluated at percentiles of capital and percentiles of output. The estimates range from 0.31 for firms at the lowest percentiles of output and capital to 0.54 for firms at the highest percentiles of output and capital. The second panel in Figure 1 reports the average labor elasticity evaluated at percentiles of labor and percentiles of output. The relationship is opposite to that of capital. The estimates are lowest just below -0.7 for firms at the highest percentiles of output and the lowest percentiles of labor to 0.42 for firms in the highest percentile of output and the highest percentile of labor. The third panel shows the estimates of the average materials elasticity evaluated at percentiles of materials and output. The relationship is similar to that of labor. The estimates are lowest at -0.22 for firms at the highest percentiles of output and lowest percentiles of materials. The materials estimates are highest at 1.24 for firms at the lowest percentile of output and highest percentile of materials. The estimates for labor and material elasticities are unusually small and large for the lowest and highest percentiles of their respective distribution. However at the median the estimates are more reasonable. For example, labor elasticities fall between 0.11 and 0.12 while material elasticities fall between 0.67 and 0.77.

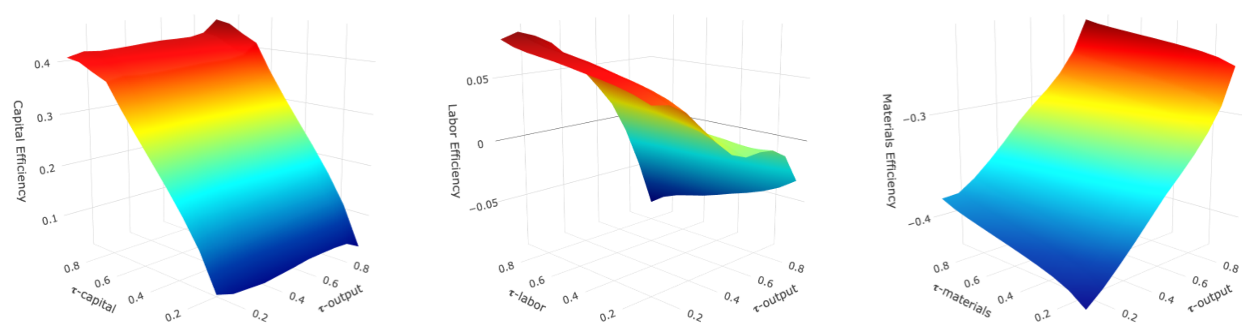
Figure 1: Individual Output Elasticities



*Panel (a): Capital elasticity evaluated at τ_η and percentiles of capital τ_k averaged over values of (l_{it}, m_{it}) that correspond to τ_k. Panel (b): Labor elasticity evaluated at τ_η and percentiles of labor τ_l averaged over values of (k_{it}, m_{it}). Panel (c): Materials elasticity evaluated at τ_η and percentiles of materials τ_m averaged over values of (k_{it}, l_{it}).

Figure 2 show estimates of the non-Hicks neutral effects of productivity on capital, labor, and material inputs. These effects can be interpreted as the degree of which the Hicks-neutral productivity have non-Hicksian neutral effects. This is in contrast to the standard Hicks-neutral production function technologies, which assume the productivity term scales input usage by an equal amount. The labor-augmenting aspect of the shock is of particular importance since the empirical literature often points to labor-augmenting technological change as sources of productivity growth over time. Despite its importance, there are relatively few papers in the empirical IO literature that study these sources of productivity. This is because identification and estimation of these models is difficult since controlling for the endogeneity of multi-dimensional productivity in some situations require observing firm-specific input prices (Doraszelski and Jaumandreu, 2018) or certain functional form assumptions (Dermirer, 2020). My estimates provide complementary evidence to these productivity differences between different inputs. It is worth noting that the identification arguments presented here may accommodate multi-dimensional unobservables such as Hicks-neutral and labor-augmenting productivity. Extra unobservables require additional proxies which increases the data requirements in my approach, but the estimates under this alternative would be more suited for comparison to existing empirical work.

Figure 2: Individual Non-Hicks Neutral Elasticities



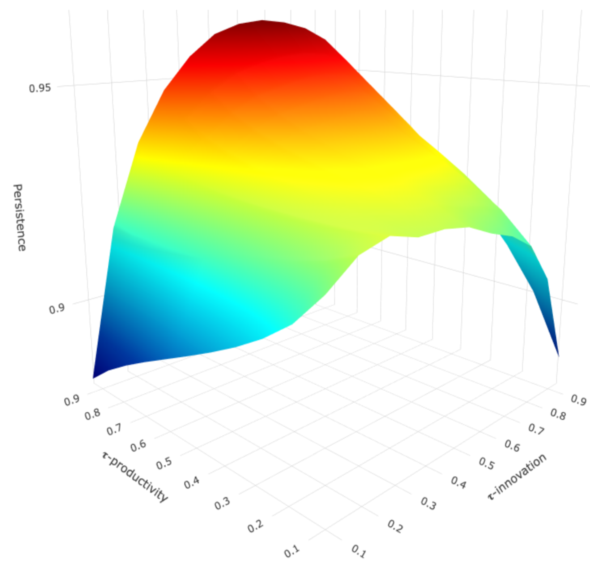
*Panel (a): Capital elasticity evaluated at τ_η and percentiles of capital τ_k averaged over values of (l_{it}, m_{it}) that correspond to τ_k. Panel (b): Labor elasticity evaluated at τ_η and percentiles of labor τ_l averaged over values of (k_{it}, m_{it}). Panel (c): Materials elasticity evaluated at τ_η and percentiles of materials τ_m averaged over values of (k_{it}, l_{it}).

Panel (a) in Figure 2 shows the results for the capital-augmenting effect of Hicks-neutral productivity. These estimates are computed at various percentiles of output and capital to examine how the capital efficiency effect varies over firms. The estimates range from 0.03 for firms at the highest percentiles of output and lowest percentile of capital to 0.46 for firms at the the highest percentiles of output and capital. Overall, the capital efficiency effects are increasing for firms who use more capital, but almost constant across the conditional output distribution. For the labor efficiency estimates in the panel (b), there is heterogeneity between firms of different sizes who use varying amounts of labor. The estimates range from -0.13 for firms at the highest percentile of output and labor to 0.11 for firms at the lower percentiles of output, but the highest percentile of labor. This is consistent with empirical findings that labor-productivity exhibits high amounts of firm variation. Panel (c) reports the material efficiency estimates. These range from -0.75 for firms at the lowest percentiles of output and materials to -0.12 or firms at the highest percentiles of output and materials. These estimates suggest that for firms in this sample, increases in productivity lead to decreases in the marginal product of materials or that firms are material inefficient.

6.1.2 Persistence of Productivity

I next examine the estimates of the productivity process. Figure 3 reports the estimates of productivity persistence at various percentiles of the innovation shock and percentiles of last period productivity. Persistency exhibits a significant asymmetric relationship. These results suggest that high productivity firms (τ-productivity$= 0.9$) hit by negative shocks have a lower persistence of productivity history (0.88) than low productivity firms (τ-productivity$=$ 0.1) hit by the same sized shock (0.92). This indicates that for high productivity firms, unanticipated negative shocks can reduce the history of periods of high productivity, whereas this effect for low productivity firms is less severe. This relationship changes when firms are hit by large positive shocks. High productivity firms have higher persistency of productivity history (0.95) than low productivity firms (0.89) when hit by positive shocks.

Figure 3: Productivity Persistence



*Estimates of average productivity persistence evaluated at τ_ξ and percentiles of previous productivity.

6.1.3 Input Demand and Productivity

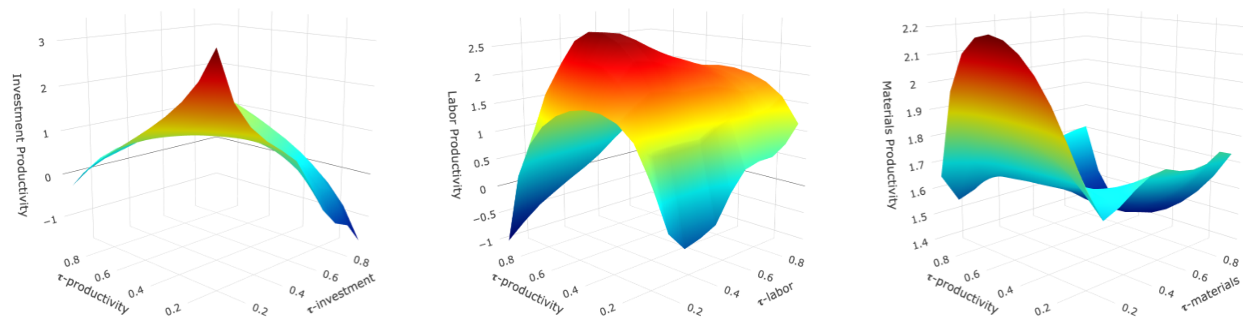
I also examine how firms adjust inputs with respect to productivity and innovation shocks. These estimates provide insight on how firms input demand changes in response to technological or organizational innovations measured by the unobserved productivity component. For example, whether a firm changes its labor demand in response to innovations in automation has important consequences for employment displacement and its public policy responses. My estimates show that there is heterogeneity at the firm-level in these productivity responses for different percentiles of productivity and input demand sizes.

First looking at the investment response to productivity in Panel (a) in Figure 4, firms at the lowest percentile of investment and the lowest percentile of productivity have the largest response at 5.6. As productivity increases for lower investment firms, this effect decreases to -0.38 for firms at the highest percentile of productivity. For high investment firms at the lowest percentile of productivity, the effect is -1.04 and for similar levels of investment at the highest percentile of productivity is 2.18. Overall, these results suggest there is significant heterogeneity in firms investment adjustments with respect to changes in productivity levels.

Panel (b) shows the relationship between labor input and productivity. Firms at the lowest percentile of labor and the lowest percentile of productivity have a productivity effect equal to -2.11. For firms at the lowest percentile of labor and the highest percentiles of productivity this effect is equal to -0.36. For firms at the highest percentile of labor, but the lowest percentile of productivity the effect approaches its maximum around to 1.93, but for firms at highest percentile of productivity this effect diminishes to 0.53. These results show that for firms who use less labor and relatively less productive, increases in productivity leads to a decrease in the amount of labor.

Panel (c) shows the relationship between materials input and productivity. Firms at the lowest percentile of materials and the lowest percentile of productivity have a productivity effect equal to 2.23. For firms at the lowest percentile of materials and the highest percentiles of productivity this effect is highest and equal to 4.05. For firms at the highest percentile of materials, but the lowest percentile of productivity the effect is equal to 1.93 and for the highest percentile of productivity is 1.79. Overall, there is not much heterogeneity in materials productivity except for high productivity firms who do not use much material inputs. However, firms respond to increases in productivity by using more materials.

Figure 4: Marginal Productivity of Inputs

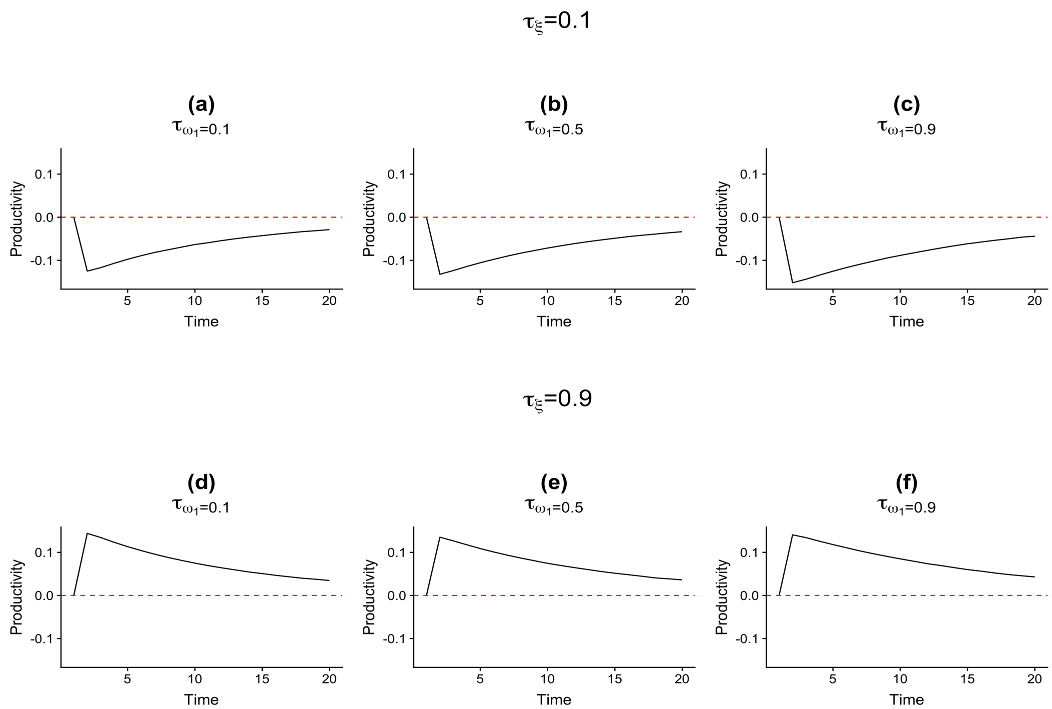


*Panel (a): Investment demand evaluated at τ_ζ and percentiles of productivity τ_ω averaged over values of k_{it}. Panel (b): Labor demand evaluated at τ_{ϵ_l} and percentiles of productivity τ_ω averaged over values of k_{it}. Panel (c): Material demand evaluated at τ_{ϵ_m} and percentiles of productivity τ_ω averaged over values of k_{it}

6.1.4 Impulse Responses to Productivity Shocks

This section simulates the impact of innovation shocks to the productivity process and input demand functions from the model estimates. Figures 5, 7, 8 and 6 report median differences in low innovation shocks $\tau_\xi = 0.1$ and high innovation shocks $\tau_\xi = 0.9$ and firms hit by medium innovation shocks at $\tau_\xi = 0.5$ for productivity, investment, labor and materials. I simulate the model so that the impact of the shock occurs at $t = 2$. I examine the corresponding response to productivity and inputs as well as the length of time it takes for firms to recover from negative productivity shocks. This analysis is somewhat similar to Hu *et al.* (2019) since in their paper, they study how quickly firms adjust inputs in response to the latest shocks to productivity. Their GMM estimator allows them to estimate the covariance between inputs, productivity and its shocks. This is useful in their context as it provides guidance for choosing proxies for the latent productivity. These estimates can also identify industry efficiency and frictions in the input markets. Unlike their GMM estimator, my simulations document the impact of different size innovation shocks and input demand functions beyond the mean as well as the full history of the impact. Similar to Arellano *et al.* (2017) I refer to productivity and input demand paths as impulse response functions.

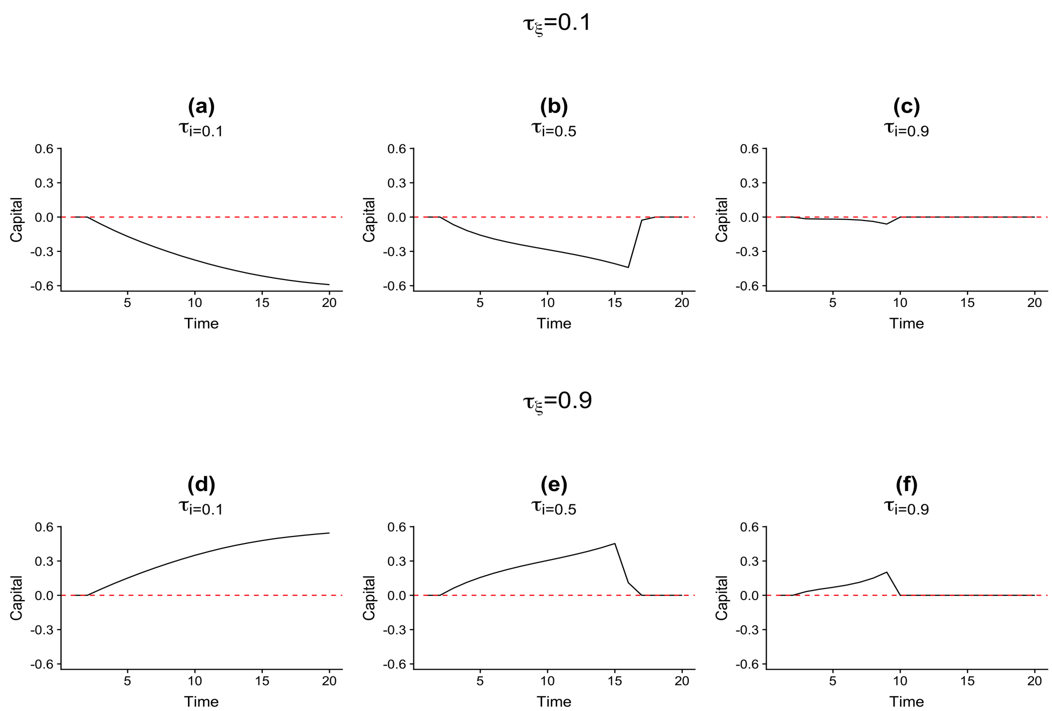
Figure 5: Impulse Response of an Innovation Shock to Productivity



*Top row: Difference between firms hit with low productivity shock $\tau_\xi = 0.1$ and medium shock $\tau_\xi = 0.5$ at different levels of initial productivity. Bottom row: Difference between firms hit with high productivity shock $\tau_\xi = 0.9$ and medium shock $\tau_\xi = 0.5$ at different levels of initial productivity.

The productivity response to innovation shocks are reported in Figure 5 which shows the impact of a large negative shock ($\tau_\xi = 0.1$) in panel (a-c) and large positive shock ($\tau_\xi = 0.9$) in panel(d-f) for various levels of initial productivity $\tau_{\omega_1} = (0.1, 0.5, 0.9)$. For firms with the lowest initial productivity, a large negative innovation shock decreases productivity by 12.5% while a large positive shock increases productivity by 14%. For firms with the highest initial productivity, a large negative innovation shock decreases productivity by 15% and a large positive shock increases productivity by about 14%. There is no discernible difference in the length of time required to recover from negative productivity shocks which is consistent with the small difference in productivity persistence for high and low productivity firms (0.88 vs 0.92) hit by negative innovation shocks.

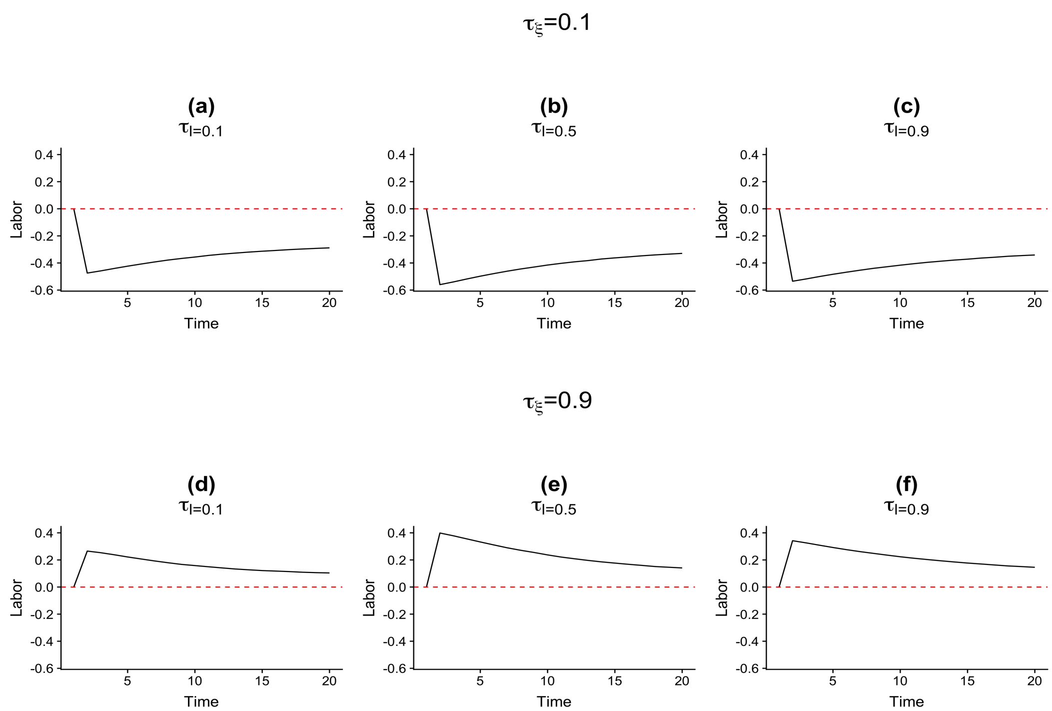
Figure 6: Impulse Response of an Innovation Shock to Investment



*Top row: Difference between firms hit with low productivity shock $\tau_\xi = 0.1$ and medium shock $\tau_\xi = 0.5$ at different levels of investment demand. Bottom row: Difference between firms hit with high productivity shock $\tau_\xi = 0.9$ and medium shock $\tau_\xi = 0.5$ at different levels of investment demand

The investment response to innovation shocks are reported in Figure 6 which shows the impact of a negative productivity shocks in panel (a-c) and positive productivity shocks in panel(d-f) for various levels of investment demand $\tau_i = (0.1, 0.5, 0.9)$. For firms with the lowest investment demand, a large negative productivity shock decreases investment by 22% while a large positive shock increases investment by 17%. In this scenario, a firm with low investment demand does not recover from negative productivity shocks. For firms with the highest investment demand, a large negative productivity shock decreases investment by 2.3% and a large positive productivity shock increases investment by 4.6%. As firms increase in the size of investment demand, the length of time to recover from negative productivity shocks decreases. In addition, there is a large lag between the time the firm is hit by the negative shock and when investment demand is lowest. For example, in panel (b), demand for medium investment demand firms continues to decrease for almost 15 years until there is any recovery.

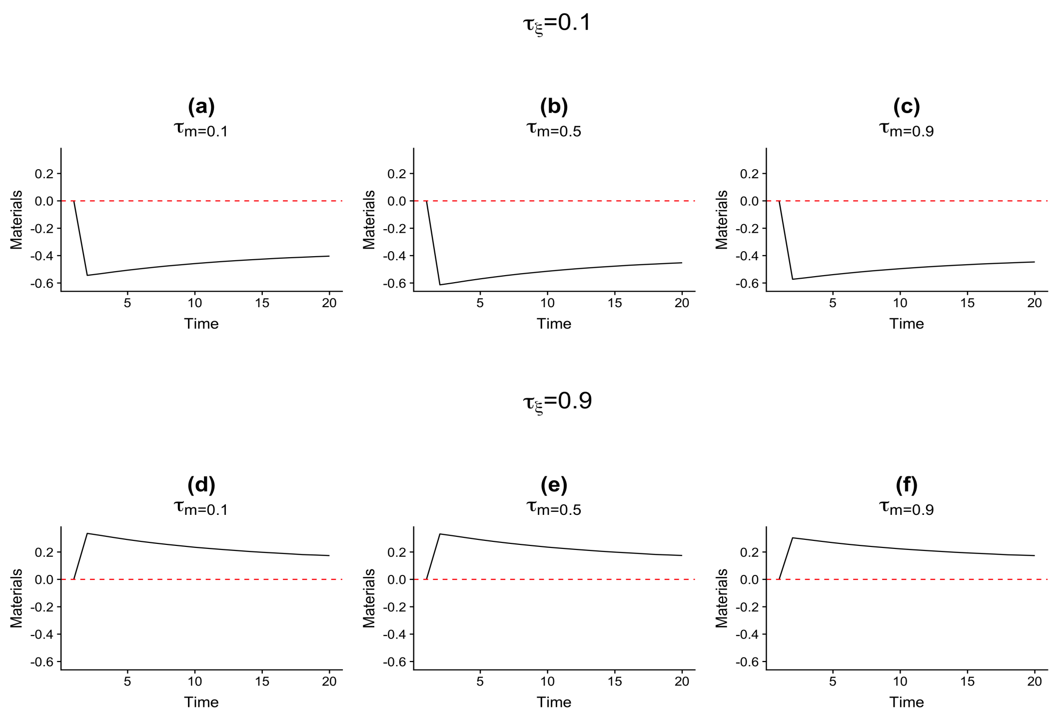
Figure 7: Impulse Response of an Innovation Shock to Labor



*Top row: Difference between firms hit with low productivity shock $\tau_\xi = 0.1$ and medium shock $\tau_\xi = 0.5$ at different levels of labor demand. Bottom row: Difference between firms hit with high productivity shock $\tau_\xi = 0.9$ and medium shock $\tau_\xi = 0.5$ at different levels of labor demand.

The labor response to innovation shocks are reported in Figure 7 which shows the impact of a negative productivity shocks in panel (a-c) and positive productivity shocks in panel(d-f) for various levels of labor demand $\tau_l = (0.1, 0.5, 0.9)$. For firms with the lowest labor demand, a large negative productivity shock decreases labor inputs by 22% while a large positive shock increases labor inputs by 16%. For firms with the highest labor demand, a large negative productivity shock decreases labor inputs by 22% and a large positive productivity shock increases labor inputs by about 17%. For all firms hit by negative productivity shocks, the recovery trends are U-shaped. Firms with large labor demand overall have the smallest decrease in labor after a negative productivity shock, but they also stay at these levels for much longer than low labor firms. However, the speed of recovery for low labor firms happens more gradually than high labor firms

Figure 8: Impulse Response of an Innovation Shock to Materials



*Top row: Difference between firms hit with low productivity shock $\tau_\xi = 0.1$ and medium shock $\tau_\xi = 0.5$ at different levels of materials demand. Bottom row: Difference between firms hit with high productivity shock $\tau_\xi = 0.9$ and medium shock $\tau_\xi = 0.5$ at different levels of materials demand.

The materials response to innovation shocks are reported in Figure 8 which shows the impact of a negative productivity shocks in panel (a-c) and positive productivity shocks in panel(d-f) for various levels of materials demand $\tau_m = (0.1, 0.5, 0.9)$. For firms with the lowest materials demand, a large negative productivity shock decreases material inputs by 50% while a large positive shock increases material inputs by 44%. For firms with the highest materials demand, a large negative productivity shock decreases material inputs by 30% and a large positive productivity shock increases material inputs by 29%. Low material demand firms hit by negative productivity shocks recover faster than medium and high material demand firms. For high material firms, demand continues to fall for 10 years until there is a rapid recovery. For medium firms, the time of recovery is the longest.

7 Conclusion

This paper proposed a non-linear model for firm production which allows for elasticities and non-Hicks neutral effects of productivity to vary over the conditional distribution of output. The estimates reveal substantial heterogeneity across this distribution as well as across different percentiles of input demand. This challenges the standard approach of estimating production functions that specify technology that is fixed across firms and instead suggests that non-linear, firm-specific models are more suitable when heterogeneity is prevalent in the data. The approach considered here also allows for a more flexible productivity process where persistence in history of productivity can vary with respect to the latest innovation shocks and that good or bad innovation shocks reduces persistence for both high and low productivity firms.

The identification approach in this paper is an extension of Hu *et al.* (2019) which identifies the production function in the presence of unobservables in the input demand functions. I show that when these functions are strictly increasing in the unobservables, quantile regression can be used to document the heterogeneity in which firms adjust their inputs in response to productivity. This type of analysis is useful from a policy perspective, as policy proposals aimed to increase productivity may have different outcomes for firms with different input demand functions and productivity levels. This paper also studies the adjustment frictions of input demand functions in response to innovation shocks to productivity and finds asymmetries in the impact of good and bad shocks as well as the length of time for recovery from bad productivity shocks. The overall finding is that firms with the highest input-productivity adjustments also have the largest drop in input demand following a bad productivity shock. However for these types of firms, the length of recovery is the shortest.

There are many interesting extensions that can be considered in the framework proposed in this paper. The first would be to include additional unobservables beyond the productivity term. For example, fixed effects can be included in the production function and productivity process to account for firm-specific unobservables, However, additional proxies would need to be added to the model to guarantee identification. The current model assumes productivity is scalar and that its interactions with inputs measure the magnitude of non-Hicks neutral effects. It would be interesting to consider multi-dimensional productivity shocks, for example a Hicks-neutral and a labor-augmenting term to capture productivity effects that are biased towards labor. Extending the identification arguments to this case would be more demanding since labor-augmenting productivity is typically serially correlated. Lastly, the results presented here are often used to estimate other aspects of firm technology and market power. Further analysis of total factor productivity and markup estimates would provide an interesting comparison with the standard production function estimates.

References

Ackerberg, D., Benkard, C. L., Berry, S. and Pakes, A. (2007). Chapter 63 econometric tools for analyzing market outcomes. In *Handbook of Econometrics*, Elsevier, pp. 4171–4276.

— and Hahn, J. (2015). Some non-parametric identification results using timing and information set assumptions. Working Paper.

Ackerberg, D. A., Caves, K. and Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, **83** (6), 2411–2451.

Arellano, M., Blundell, R. and Bonhomme, S. (2017). Earnings and consumption dynamics: A nonlinear panel data framework. *Econometrica*, **85** (3), 693–734.

— and Bonhomme, S. (2016). Nonlinear panel data estimation via quantile regressions. *The Econometrics Journal*, **19** (3), C61–C94.

— and — (2017). Quantile selection models with an application to understanding changes in wage inequality. *Econometrica*, **85** (1), 1–28.

Balat, J., Brambilla, I. and Sasaki, Y. (2018). Heterogeneous firms: skilled-labor productivity and the destination of exports, Working paper.

Celeux, G. and Diebolt, J. (1985). The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational Satistics Quarterly*, (2), 73–82.

Dermirer, M. (2020). Production function estimation with factor augmenting technology: An application to markups. Working paper.

Doraszelski, U. and Jaumandreu, J. (2013). R&d and productivity: Estimating endogenous productivity. *The Review of Economic Studies*, **80** (4), 1338–1383.

— and — (2018). Measuring the bias of technological change. *Journal of Political Economy*, **126** (3), 1027–1084.

Dunford, N. and a, J. T. S. (1971). *Linear Operators*, vol. 3. Wiley.

Gandhi, A., Navarro, S. and Rivers, D. A. (2020). On the identification of gross output production functions. *Journal of Political Economy*, pp. 000–000.

Hu, Y., Huang, G. and Sasaki, Y. (2019). Estimating production functions with robustness against errors in the proxy variables. *Journal of Econometrics*.

— and Schennach, S. M. (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, **76** (1), 195–216.

Kasahara, H., Schrimpf, P. and Suzuki, M. (2017). Identification and estimation of production function with unobserved heterogeneity, Working paper.

Kim, K. I., Petrin, A. and Song, S. (2016). Estimating production functions with control functions when capital is measured with error. *Journal of Econometrics*, **190** (2), 267–279.

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

Li, T. and Sasaki, Y. (2017). Constructive identification of heterogeneous elasticities in the cobb-douglas production function, Working paper.

Nielsen, S. F. (2000). The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli*, **6** (3), 457.

Olley, G. S. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, **64** (6), 1263.

Portnoy, S. and Koenker, R. (1997). The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators. *Statistical Science*, **12** (4).

Appendix

A Data Appendix

Table 1: Summary Statistics (in logs) for U.S. Manufacturing Firms

	1st Qu.	Median	3rd Qu.	Mean	sd
Output	4.24	5.79	7.27	5.79	2.14
Capital	3.12	4.84	6.45	4.81	2.35
Labor	-1.23	0.22	1.62	0.21	1.95
Materials	3.95	5.47	6.95	5.46	2.15
Investment	0.57	2.40	3.94	2.23	2.49

Variable Construction:

- Output: Deflated Net Sales from Compustat (SALE).
- Capital: Deflated Property Plant and Equipment Net of Depreciation (PPENT).
- Labor: Number of Workers (EMPLOY).
- Labor Expense: EMPLOY times average industry wage calculated from the ratio of PAY and EMP in the NBER-CES Manufacturing Industry Database.
- Materials: Sales (SALE)-Operating Income Before Depreciation (OIBDP)-labor expense.
- R&D: XRD in Compustat.

B Identification

In this section I show how the results of Hu and Schennach (2008) can be applied to identify the production function, input demand functions, and the marginal distribution of productivity. Technical details for the proof for their decomposition technique can be found in their main paper.

Proof of Theorem 3.1 :
First, a conditional density constructed from observed data can be written as a product of the unknown unconditional densities of interest:

$$\begin{aligned}
f_{y_t,I_t|y_{t+1},Z_t} &= \int f_{y_t,I_t,\omega_t|y_{t+1},Z_t}(y_t,I_t,\omega_t|y_{t+1},Z_t)d\omega_t \\
&= \int f_{y_t|y_{t+1},I_t,\omega_t,Z_t}(y_t|y_{t+1},I_t,\omega_t,Z_t)f_{I_t|y_{t+1},\omega_t,Z_t}(I_t|y_{t+1},\omega_t,Z_t)f(\omega_t|y_{t+1},Z_t)d\omega_t \\
&= \int f_{y_t|\omega_t,Z_t}(y_t|\omega_t,Z_t)f_{I_t|Z_t}(I_t|\omega_t,Z_t)f(\omega_t|y_{t+1},Z_t)d\omega_t,
\end{aligned} \tag{33}$$

where the third line follows from applying the conditional independence assumption in 3.1. The goal of the identification strategy is to show that the conditional densities in Equation (33) can be written into its corresponding integral operators which can be shown to admit a unique decomposition. Using Definition 3.1 and omitting the conditioning on Z_t for notational convenience:

$$\begin{aligned}
[L_{y_t,I_t|y_{t+1}}g](y_t) &= \int f_{y_t,I_t|y_{t+1}}(y_t,I_t|y_{t+1})g(y_{t+1})dy_{t+1} \\
&= \int\int f_{y_t,I_t,\omega_t|y_{t+1}}(y_t,I_t,\omega_t|y_{t+1})d\omega_t g(y_{t+1})dy_{t+1} \\
&= \int\int f_{y_t|I_t,y_{t+1},\omega_t}(y_t|I_t,y_{t+1},\omega_t)f_{I_t|y_{t+1},\omega_t}(I_t|y_{t+1},\omega_t)f_{\omega_t|y_{t+1}}(\omega_t|y_{t+1})g(y_{t+1})dy_{t+1}d\omega_t \\
&= \int f_{y_t|,\omega_t}(y_t|\omega_t)f_{I_t|\omega_t}(I_t|\omega_t)\int f_{\omega_t|y_{t+1}}(\omega_t|y_{t+1})g(y_{t+1})dy_{t+1}d\omega_t \\
&= \int f_{y_t|\omega_t}(y_t|\omega_t)f_{I_t|\omega_t}(I_t|\omega_t)[L_{\omega_t|y_{t-1}}g](\omega_t)d\omega_t \\
&= \int f_{y_t|\omega_t}(y_t|\omega_t)[\Delta_{I_t|\omega_t}L_{\omega_t|y_{t-1}}g](\omega_t)d\omega_t \\
&= [L_{y_t|\omega_t}\Delta_{I_t|\omega_t}L_{\omega_t|y_{t-1}}g](\omega_t),
\end{aligned}$$

where $\Delta_{I_t|\omega_t}$ is the diagonal operator mapping $g(\omega_t)$ to the function $f_{I_t|\omega_t}(I_t|\omega_t)g(\omega_t)$. Therefore, we have the following equivalence:

$$L_{y_t,I_t|y_{t+1}} = L_{y_t|\omega_t}\Delta_{I_t|\omega_t}L_{\omega_t|y_{t-1}}. \tag{34}$$

Integrating (34) over I_t yields $L_{y_t|y_{t+1}} = L_{y_t|\omega_t}L_{\omega_t|y_{t+1}}$. Then using Assumption 3.2:

$$L_{\omega_t|y_{t+1}} = L^{-1}_{y_t|\omega_t}L_{y_t|y_{t+1}}. \tag{35}$$

Plugging (35) into (34):

$$L_{y_t,I_t|y_{t+1}} = L_{y_t|\omega_t}\Delta_{I_t|\omega_t}(L^{-1}_{y_t|\omega_t}L_{y_t|y_{t+1}}).$$

Note that the operator $L_{y_t|y_{t+1}} = L_{y_t|\omega_t}L_{\omega_t|y_{t+1}}$ is injective due to Assumption 3.2. Then we have the following:[3]

$$L_{y_t,I_t|y_{t+1}}L^{-1}_{y_t|y_{t+1}} = L_{y_t|\omega_t}\Delta_{I_t|\omega_t}L^{-1}_{y_t|\omega_t}. \tag{36}$$

The LHS of (36) is a function of observed data which can be considered as known. This expression states that the LHS admits a spectral decomposition that takes the form of an eigenvalue-eigenfunction decomposition. In order to identify the unobserved densities of interest, the representation in 36 and its decomposition must be unique. This is guaranteed by Theorem XV.4.5 in Dunford and a (1971) and Assumptions 3.3 and 3.4. Then applying Theorem 1 in Hu and Schennach (2008) identifies $f_{y_t|\omega_t,Z_t}$, $f_{I_t|\omega_t,Z_t}$ and $f_{\omega_t|y_{t+1},Z_t}$.

The marginal distribution of productivity is identified from

$$f_{\omega_t} = \int f_{y_{t+1},\omega_t}d\omega_t = \int f_{\omega_t|y_{t+1}}f_{y_{t+1}}d\omega_t,$$

since $f_{y_{t+1}}$ is observed and $f_{\omega_t|y_{t+1}}$ was identified from Theorem 3.1. The input demand functions for m_t and l_t are identified since f_{ω_t} is known. The next step is identification of the Markov process $f_{\omega_{t+1}|\omega_t}$ using Corollary 3.1 and 3.2.

Proof of Corollary 3.1:
Note that the integral operator corresponding to the density $f_{y_{t+1}|\omega_t}(y_{t+1}|\omega_t)$ can be written

[3] The fact that injectivity of $L_{y_{t+1}|\omega_t}$ implies injectivity of $L_{\omega_t|y_{t+1}}$ is non-trivial, but is guaranteed from Lemma 1 in Hu and Schennach (2008).

as:

$$
\begin{aligned}
[L_{y_{t+1}|\omega_t} g](y_{t+1}) &= \int f_{y_{t+1}|\omega_t}(y_{t+1}|\omega_t) g(\omega_t) d\omega_t \\
&= \int \int f_{y_{t+1},\omega_{t+1}|\omega_t}(y_{t+1}, \omega_{t+1}|\omega_t) d\omega_{t+1} g(\omega_t) d\omega_t \\
&= \int f_{y_{t+1}|\omega_{t+1}}(y_{t+1}|\omega_{t+1}) f_{\omega_{t+1}|\omega_t}(\omega_{t+1}|\omega_t) d\omega_{t+1} g(\omega_t) d\omega_t \\
&= \int \left[f_{y_{t+1}|\omega_{t+1}}(y_{t+1}|\omega_{t+1}) \int f_{\omega_{t+1}|\omega_t}(\omega_{t+1}|\omega_t) g(\omega_t) d\omega_t \right] d\omega_{t+1} \\
&= \int \left[f_{y_{t+1}|\omega_{t+1}}(y_{t+1}|\omega_{t+1}) [L_{\omega_{t+1}|\omega_t} g](\omega_{t+1}) \right] d\omega_{t+1} \\
&= [L_{y_{t+1}|\omega_{t+1}} L_{\omega_{t+1}|\omega_t} g](\omega_t).
\end{aligned}
$$

Hence:

$$
L_{y_{t+1}|\omega_t} = L_{y_{t+1}|\omega_{t+1}} L_{\omega_{t+1}|\omega_t}. \tag{37}
$$

Under stationarity, injectivity of $L_{y_t|\omega_t}$ is equivalent to injectivity of $L_{y_{t+1}|\omega_{t+1}}$, so that the Markov law of motion $f_{\omega_{t+1}|\omega_t}(\omega_{t+1}|\omega_t)$ is identified using

$$
L_{\omega_{t+1}|\omega_t} = L_{y_{t+1}|\omega_t} L^{-1}_{y_{t+1}|\omega_{t+1}}, \tag{38}
$$

since $f_{y_{t+1}|\omega_{t+1}}(y_{t+1}|\omega_{t+1})$ is equivalent to $f_{y_t|\omega_t}(y_t|\omega_t)$ under stationarity, $f_{\omega_{t+1}|\omega_t}(\omega_{t+1}|\omega_t)$ is identified since the densities $f_{y_t|\omega_t}(y_t|\omega_t)$ and $f_{y_{t+1}|\omega_t}(y_{t+1}|\omega_t)$ are identified from Theorem 3.1.

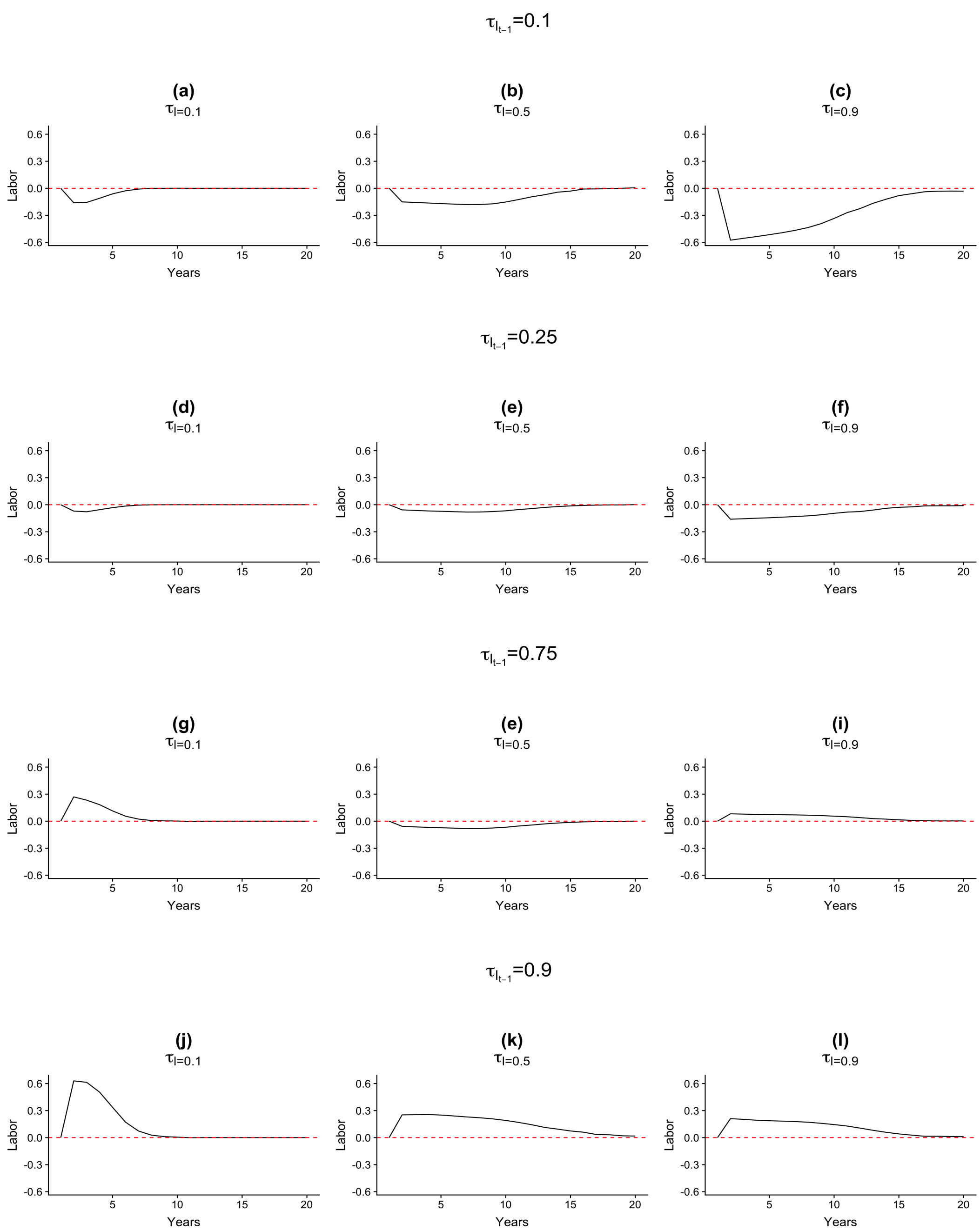
Proof of Corollary 3.2 :

In the absence of stationarity, the density $f_{y_{t+1}|\omega_{t+1}}$ is not the same as $f_{y_t|\omega_t}$. However, in this case, the identification strategy and result from Theorem 3.1 can be reapplied using observations $(y_{t+2}, y_{t+1}, I_{t+1})$.

C Extensions

C.1 Labor Dynamics

Figure 9: Impulse Response of Adjustment Shocks to Labor



C.2 R&D Activities

C.3 Correcting for Selection Bias

The estimation procedure presented here can be adapted to correct for non-random firm exit in the framework of Olley and Pakes (1996) and Dermirer (2020). An exit rule is part of a Markov perfect Nash equilibrium that determines a threshold level of productivity for which firms will stay in operation if productivity realizations are higher than this minimum level. The decision to stay in operation or exit is given by:

$$\chi_{it} = \begin{cases} 1 & \text{if} \quad \omega_{it} \geq \underline{\omega}_t(k_{it}) \\ 0 & \text{otherwise.} \end{cases} \tag{39}$$

The productivity threshold is determined by a firm's current capital stock. Firms with larger capital stocks can expect larger future returns for any given level of current productivity. Using the specification for the productivity process in Equation (20), the exit rule can be written as:

$$\begin{aligned} h_t(\omega_{it-1}, \xi_{it}) &\geq \underline{\omega}_t(k_{it}), \\ \xi_{it} &\geq h_t^{-1}(\omega_{it-1}, \underline{\omega}_t(k_{it})), \\ \xi_{it} &\geq h_t^{-1}(\omega_{it-1}, k_{it}), \\ \xi_{it} &\geq \underline{\omega}_t(\omega_{it-1}, k_{it}), \end{aligned} \tag{40}$$

where the second inequality follows from the monotonicity restriction in Assumption 2.2. Provided that the Markov process for productivity is exogenous $\Pr(\omega_{it}|\omega_{it-1}, \mathcal{I}_{it-1}) = \Pr(\omega_{it}|\omega_{it-1})$, the innovation shocks to productivity will be independent of current capital stock since $k_{it} \in \mathcal{I}_{it-1}$. This allows me to characterize the conditional distribution of innovation shocks as

$$\xi_{it}|(k_{it}, \omega_{it-1}) \sim U(0, 1).$$

The cutoff for which firms stay in operation can be estimated from

$$\underline{\omega}_t(\omega_{it-1}, k_{it}) = \text{Prob}(\chi_{it} = 1|\omega_{it-1}, k_{it}) \equiv p(\omega_{it-1}, k_{it}). \tag{41}$$

Therefore, firms that receive an innovation shock greater than $p(\omega_{it-1}, k_{it})$ will continue to operate. The distribution of productivity innovations conditional on (k_{it}, ω_{it-1}) and $\chi_{it} = 1$

is

$$\xi_{it}|(k_{it}, \omega_{it-1}, \chi_{it} = 1) \sim U(p(\omega_{it-1}, k_{it}), 1). \tag{42}$$

To see how this could be used to correct for selection bias in the framework of Arellano and Bonhomme (2016), consider a simple linear random coefficient model for productivity: $\omega_{it} = \rho(\xi_{it})\omega_{it-1}$. Then the independence assumptions implies:

$$\begin{aligned}
\text{Prob}(\omega_{it} \leq \rho(\tau)\omega_{it-1}|\omega_{it-1}, k_{it}, \chi_{it} = 1) \\
&= \text{Prob}(\xi_{it} \leq \tau|\omega_{it-1}, k_{it}, \chi_{it} = 1) \\
&= \frac{\tau - p(\omega_{it-1}, k_{it})}{1 - p(\omega_{it-1}, k_{it})} \equiv G(\tau, p).
\end{aligned} \tag{43}$$

This implies that for a current draw of productivity $\omega_{it}^{(m)}$, the persistency parameter $\rho(\tau)$ can be updated from the *rotated* quantile regression:

$$\hat{\rho}(\tau_q)^{(s+1)} = \underset{\rho(\tau_q)}{\text{argmin}} \sum_{i=1}^{N} \sum_{t=2}^{T} \sum_{m=1}^{M} \chi_{it} \Big[G(\tau, p)(\omega_{it}^{(m)} - \rho\omega_{it-1}^{(m)})^{+} + (1 - G(\tau, p))(\omega_{it}^{(m)} - \rho\omega_{it-1}^{(m)})^{-} \Big]. \tag{44}$$

where where $a^{+} = max(a, 0)$ and $a^{-} = max(-a, 0)$ and $p(\omega_{it-1}, k_{it})$ can be estimated from a probit regression on $\omega_{it-1}^{(m)}$ and k_{it}. Implementing this selection correction in standard quantile regression packages. For example, in quantreg for R, this amounts to using an individual-specific τ in the dual equality constraints. The estimator uses the Frisch-Newton linear programming algorithm in Portnoy and Koenker (1997) which can be implemented using **rq.fit.fnb**. The consequence of selection bias in this setting is the entire process for the productivity estimates may be biased. The amount of bias is likely to be larger in the bottom of the productivity distribution where the probability of exit is higher. Therefore, I must also control for selection bias for $\tau \leq \tau_1$ and $\tau > \tau_Q$ in the original model. I do this by adopting a control function approach. To illustrate, I use the simple AR(1) model for productivity at $\tau \leq \tau_1$

$$\omega_{it} = \rho(\tau_1)\omega_{it-1} + \xi_{it} + u_{it}, \quad \omega_{it} \leq \rho(\tau_1)\omega_{it-1}, \tag{45}$$

where ξ_{it} denotes to unobservable component of productivity that is correlated to the firm's exit decision and u_{it} denotes an iid shock that is assumed to be exponentially distributed. The issue of selection arises because

$$\mathbb{E}[\omega_{it}|\omega_{it-1}, \chi_{it} = 1] = \rho(\tau_1)\omega_{it-1} + \mathbb{E}[\xi_{it}|\omega_{it-1}, \chi_{it} = 1]. \tag{46}$$

Note that $\mathbb{E}[\xi_{it}|\omega_{it-1}, \chi_{it} = 1] \neq 0$ causes selection bias for productivity estimates at $\tau \leq \tau_1$. Provided that the density of ω_{it} conditional on ω_{it-1} is positive in a region about $\underline{\omega}_{it}$, following Olley and Pakes (1996), I invert the selection equation as a function of the selection probability p and ω_{it-1}. Therefore, I have the following equation

$$\mathbb{E}[\omega_{it}|\omega_{it-1}, \chi_{it} = 1] = \rho(\tau_1)\omega_{it-1} + s(p, \omega_{it-1}), \tag{47}$$

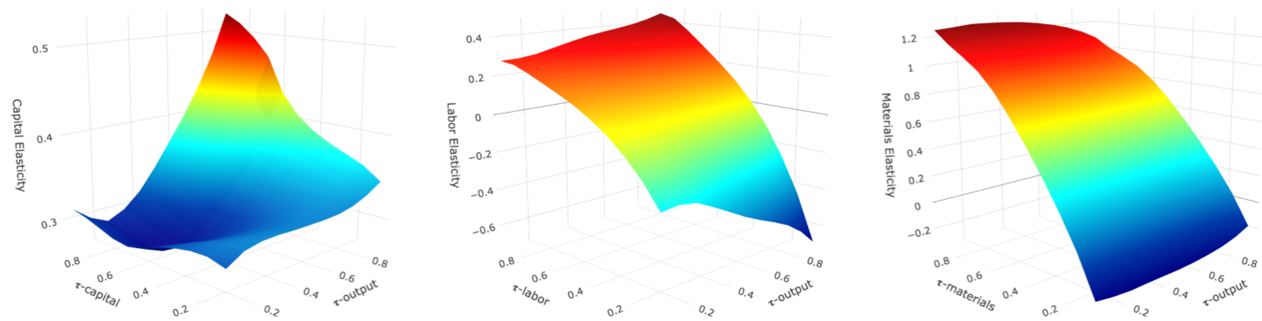
where $s(\cdot)$ denotes the sample selection correction function. I approximate this function by a second degree polynomial in p and ω_{it-1}. Then, an estimate for the exponential parameter is updated from

$$\hat{\lambda}_{\rho}^{-(s)} = \frac{-\sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{m=1}^{M} \mathbb{1}\{\omega_t^{(m)} \leq \hat{\rho}(\tau_1)^{(s)}\omega_{t-1}^{(m)} + \hat{s}(p_t, \omega_{t-1}^{(m)})\}}{\sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{m=1}^{M}(\omega_t^{(m)} \leq \hat{\rho}(\tau_1)^{(s)}\omega_{t-1}^{(m)} + \hat{s}(p_t, \omega_{t-1}^{(m)}))\mathbb{1}\{\omega_t^{(m)} \leq \hat{\rho}(\tau_1)^{(s)}\omega_{t-1}^{(m)} + \hat{s}(p_t, \omega_{t-1}^{(m)})\}}.$$

Selection correction methods for non-separable quantile models are studied by Arellano and Bonhomme (2017), but to my knowledge, has not been applied to non-linear panel data models. This extension may provide a useful starting point for combining the two methodologies.

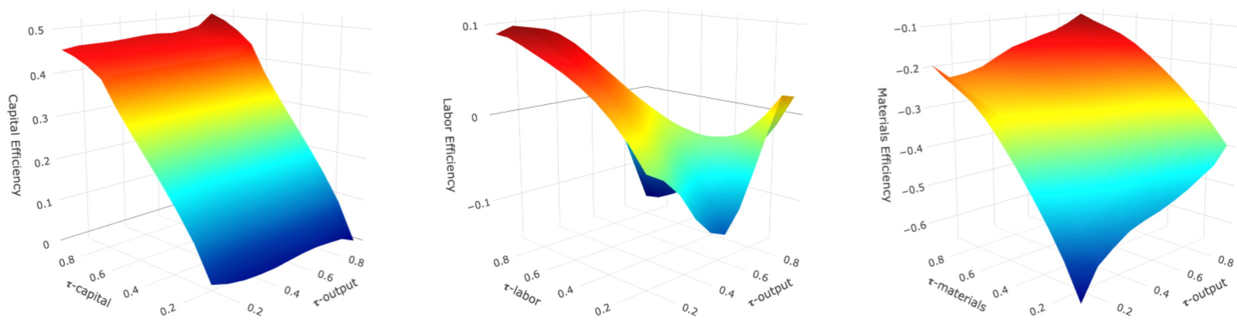
To examine the extent of selection bias, I re-estimate the model with the proposed correction in the next pages

Figure 10: Individual Output Elasticities



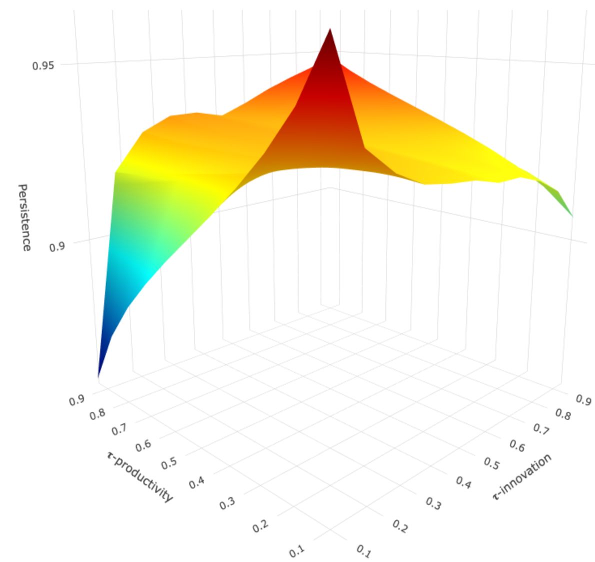
*Panel (a): Capital elasticity evaluated at τ_η and percentiles of capital τ_k averaged over values of (l_{it}, m_{it}) that correspond to τ_k. Panel (b): Labor elasticity evaluated at τ_η and percentiles of labor τ_l averaged over values of (k_{it}, m_{it}). Panel (c): Materials elasticity evaluated at τ_η and percentiles of materials τ_m averaged over values of (k_{it}, l_{it}).

Figure 11: Individual Non-Hicks Neutral Elasticities



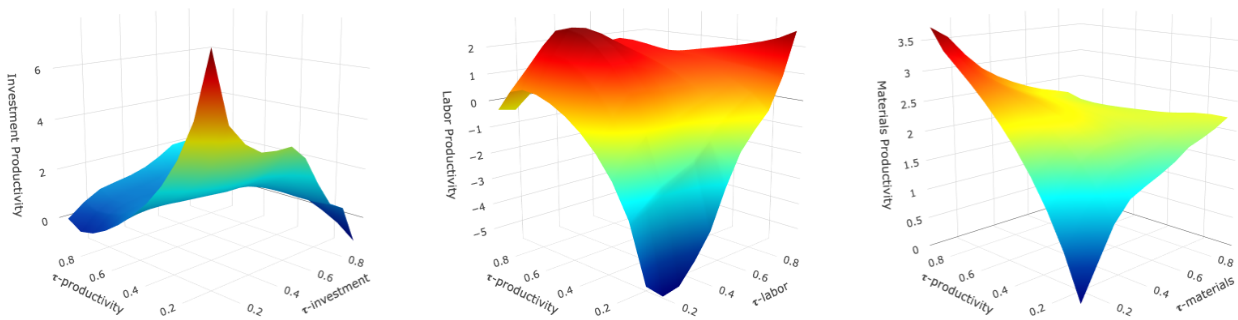
*Panel (a): Capital elasticity evaluated at τ_η and percentiles of capital τ_k averaged over values of (l_{it}, m_{it}) that correspond to τ_k. Panel (b): Labor elasticity evaluated at τ_η and percentiles of labor τ_l averaged over values of (k_{it}, m_{it}). Panel (c): Materials elasticity evaluated at τ_η and percentiles of materials τ_m averaged over values of (k_{it}, l_{it}).

Figure 12: Productivity Persistence



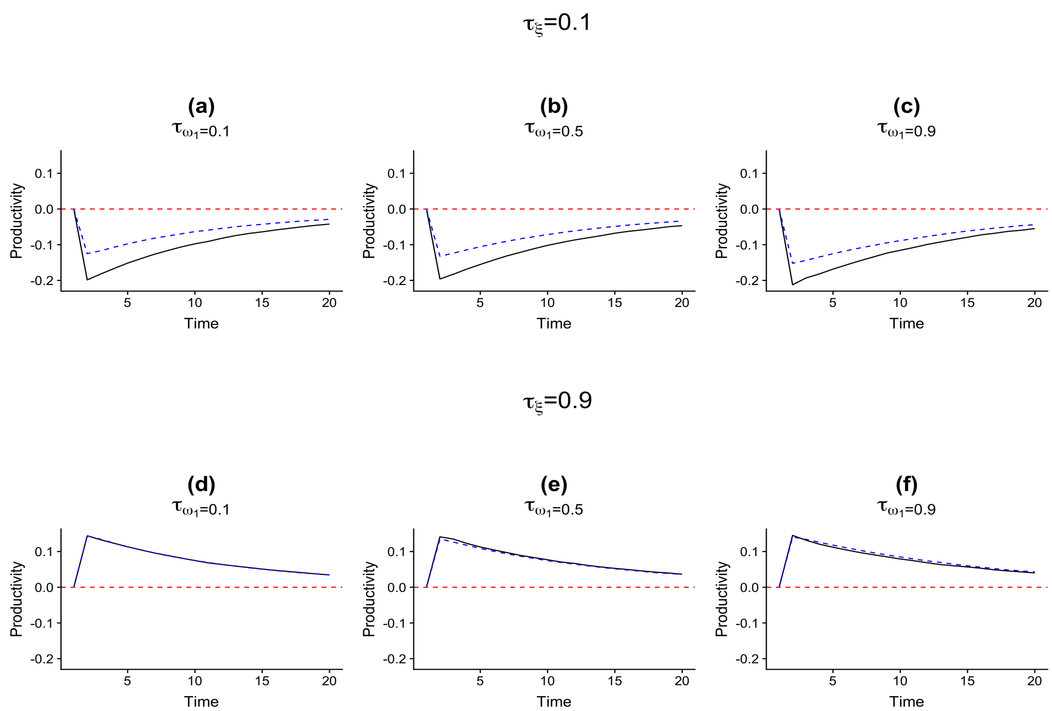
*Estimates of average productivity persistence evaluated at τ_ξ and percentiles of previous productivity.

Figure 13: Marginal Productivity of Inputs



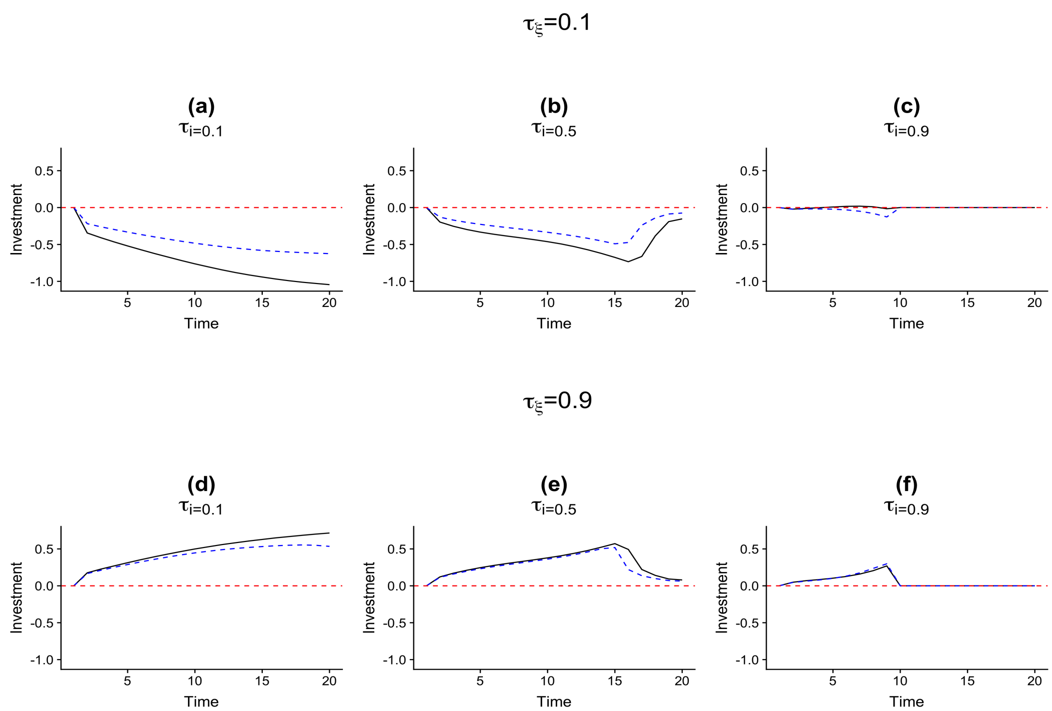
*Panel (a): Investment demand evaluated at τ_ζ and percentiles of productivity τ_ω averaged over values of k_{it}. Panel (b): Labor demand evaluated at τ_{ϵ_l} and percentiles of productivity τ_ω averaged over values of k_{it}. Panel (c): Material demand evaluated at τ_{ϵ_m} and percentiles of productivity τ_ω averaged over values of k_{it}

Figure 14: Impulse Response of an Innovation Shock to Productivity



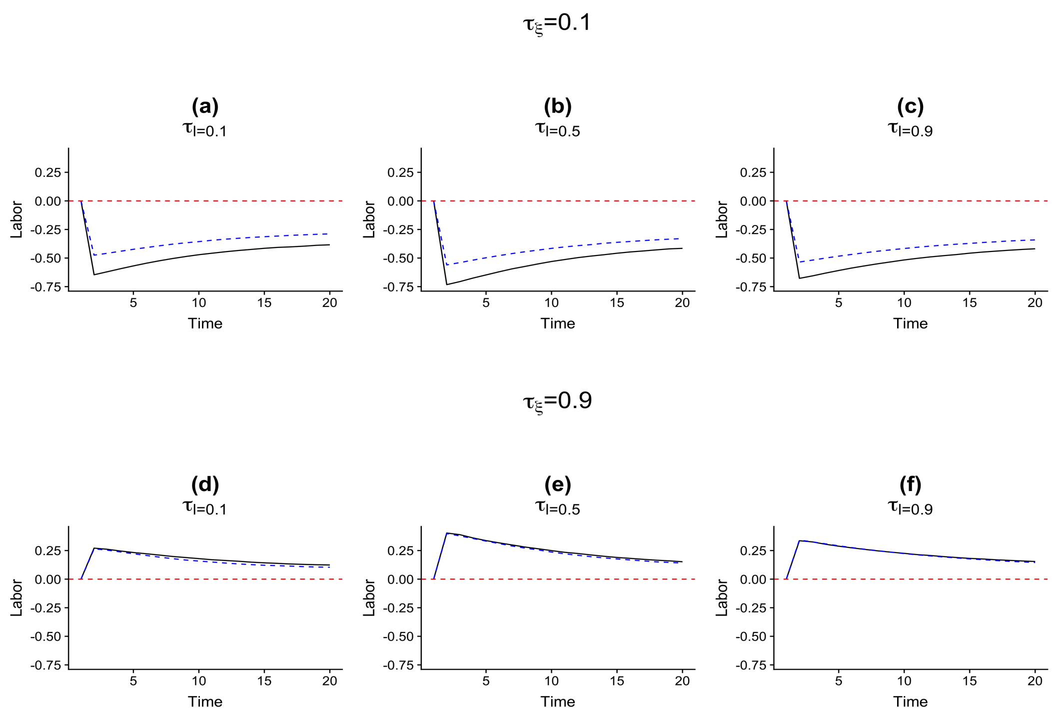
*Top row: Difference between firms hit with low productivity shock $\tau_\xi = 0.1$ and medium shock $\tau_\xi = 0.5$ at different levels of initial productivity. Bottom row: Difference between firms hit with high productivity shock $\tau_\xi = 0.9$ and medium shock $\tau_\xi = 0.5$ at different levels of initial productivity. The dashed blue line denotes the estimates from the original model without selection correction in Figure 5.

Figure 15: Impulse Response of an Innovation Shock to Investment



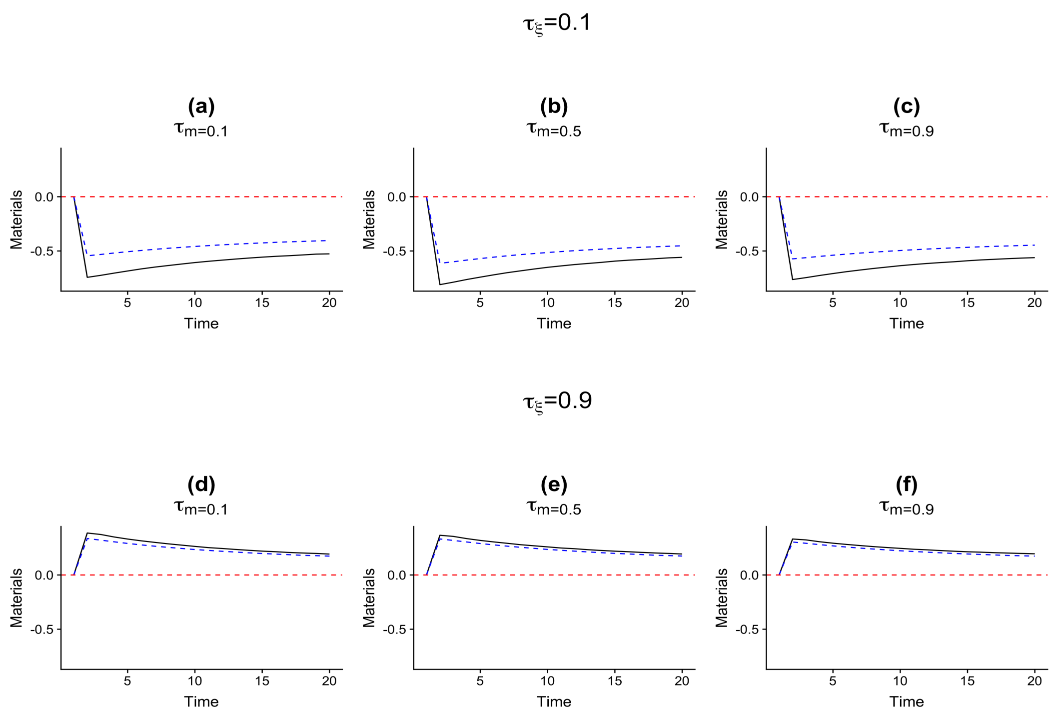
*Top row: Difference between firms hit with low productivity shock $\tau_\xi = 0.1$ and medium shock $\tau_\xi = 0.5$ at different levels of investment demand. Bottom row: Difference between firms hit with high productivity shock $\tau_\xi = 0.9$ and medium shock $\tau_\xi = 0.5$ at different levels of investment demand. The dashed blue line denotes the estimates from the original model without selection correction in Figure 6.

Figure 16: Impulse Response of an Innovation Shock to Labor



*Top row: Difference between firms hit with low productivity shock $\tau_\xi = 0.1$ and medium shock $\tau_\xi = 0.5$ at different levels of labor demand. Bottom row: Difference between firms hit with high productivity shock $\tau_\xi = 0.9$ and medium shock $\tau_\xi = 0.5$ at different levels of labor demand. The dashed blue line denotes the estimates from the original model without selection correction in Figure 7.

Figure 17: Impulse Response of an Innovation Shock to Materials



*Top row: Difference between firms hit with low productivity shock $\tau_\xi = 0.1$ and medium shock $\tau_\xi = 0.5$ at different levels of materials demand. Bottom row: Difference between firms hit with high productivity shock $\tau_\xi = 0.9$ and medium shock $\tau_\xi = 0.5$ at different levels of materials demand. The solid black line denotes the estimates correcting for selection bias. The dashed blue line denotes the estimates from the original model without selection correction in Figure 8.