A Dynamic Framework for Identification and Estimation of Nonseparable Production Functions

Justin Doty*†

October 28, 2021

Abstract

This paper studies identification and estimation of a nonseparable model for production functions with unobserved heterogeneity. Nonparametric 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 nonlinear 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 firms' heterogeneous responses to productivity shocks corresponding to the size of their input demand. This paper illustrates these results in an application to U.S. manufacturing firms where the proposed model is estimated using nonlinear quantile regression.

1 Introduction

The production function is a fundamental component of many economic models, and its estimates can be used to study patterns of productivity heterogeneity, returns to scale, and market power. Estimation of the production function is constrained by endogeneity bias from unobserved productivity. The most popular methods for correcting this source of bias impose strict structural assumptions on the functional form of production and restrictions on the number of unobservables in the model. This paper proposes a nonparametric estimation procedure that is robust to these unobservables and captures heterogeneity in firm behavior,

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

†All replication files for the estimation procedure and interactive plots can be found on the author's personal webpage.

which is not found in standard models. My model allows for nonseparability of productivity in the production function and input demand functions. It also allows for nonseparable unobservables beyond the productivity term, which I show to be an important determinant in heterogeneous firm-level estimates. Unlike previous approaches, the structural features of interest are all nonparametrically identified. This contribution is important because the parametric restrictions used in previous models rely on data that is often not available to researchers.

This paper uses the nonseparability of productivity to illustrate the importance of capturing the unobservable interactions between inputs and productivity. For example, this is used to show that estimates of output elasticities vary over the productivity levels of a firm. It is also used to provide empirical evidence on the non-Hicks neutral effects of productivity, which is shown to vary with respect to the size of the input demand of a firm. Nonseparability of unobservables in the input demand functions are also important and this paper reveals heterogeneous input adjustments with respect to productivity changes. In addition, a flexible productivity evolution process is used to show asymmetric persistence with respect to realizations in uncertainty and productivity history. In contrast, the standard production function approaches assume separability of the unobservables and place more emphasis on addressing the simultaneity bias from unobserved productivity under various timing assumptions on firm input decisions. 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 using observed variables. This is then substituted into the production function, which is estimated in a two-step approach.

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 unobserved productivity. They show that for positive investment levels this function is invertible in productivity. Their other contribution is correcting a sample 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 for 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 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) (hereafter LP) address the first challenge by providing conditions for which an intermediate input demand function, such as materials, energy, or fuels, is 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 labor in the first stage and state variables in the second stage. An 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 exist only specific data-generating processes that can break this functional dependence problem. The paper by Ackerberg *et al.* (2015) (hereafter ACF) provide 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. This precludes identification in the first stage. Instead, it is included in the second stage with the 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 are 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 adjustment costs, optimization error, or 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 proxy variable 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 equations nonparametrically. Kim *et al.* (2016) allow for measurement error in capital and other inputs using identification arguments from Hu and Schennach (2008) in the proxy variable framework. Hu *et al.* (2020) (hereafter HHS) take a similar identification approach, but propose an alternative 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. 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 proposed by Gandhi *et al.* (2020) also show that choice of the production function is important. The proxy variable approach is subject to under-identification due to an instrument-irrelevance problem using a gross-output production function. A value-added model may avoid this critique, however estimates recovered from value-added are fundamentally different from gross-output since the latter conditions on intermediate inputs. Estimates of TFP and its dispersion ratios will appear more variable using a value-added production function. Their model can be estimated nonparametrically if the productivity term is Hicks-neutral.

The assumption of a Hicks-neutral productivity term implies that technological improvements are not factor-specific. This assumption is difficult to justify empirically, as productivity can be biased towards favoring inputs like labor. Labor-augmenting productivity

[1]For example, Ackerberg *et al.* (2007) shows that when a demand shock enters the investment function, a firm's 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).

is an important component to economic models of growth. Therefore, understanding the sources of labor productivity and its heterogeneity, 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. 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 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 and nonparametric rank conditions to show validity of a proxy variable as an IV for another variable. It is important to note that the identification results of Hu and Schennach (2008) can be applied to nonseparable 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 the rank condition, which is difficult to verify in practice. In addition, their model trivially satisfies a normalization assumption on the error term, which for nonseparable 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 are 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 captured in previous approaches.

Unlike the GMM estimator proposed by HHS, I propose an estimator that can accommodate nonseparability of the production function in addition to 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 to 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.

My empirical results show that nonseparability of unobservables are an important determinant in heterogeneous firm-level estimates. I show that estimates of the output elasticities vary with respect to productivity levels and the size of input demands for capital, labor, and materials. For example, I find that capital elasticity exhibits more variation over productivity than it does over capital usage. For the non-Hicks neutral effects, I find that capital exhibits positive efficiency effects of productivity, while the materials effects are negative. I find that labor efficiency is positive for small firms and negative for large firms. I also show firms' heterogeneous input adjustments with respect to productivity changes. For investment and labor adjustments, I find some firms' positive or negative adjustments with respect to productivity levels for different sizes of investment and labor demand. Finally, I show asymmetric impacts of negative and positive innovation shocks on the full history of productivity and input demand. For example, I find that low investment firms decrease capital usage dramatically in response to a large negative productivity shock compared to high investment firms whose adjustments are minimal.

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 its implementation. In Section 6, I apply this estimator to U.S. 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 with Nonseparable Unobservables

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

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

where y_{it} is firm i's output at time t, k_{it} denotes 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, f_t, is assumed to be strictly increasing in η_{it} and can vary over time.

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 is important for developing the econometric restrictions of the model because they are based on conditional quantiles. This representation will also be used for the productivity equation and the input demand functions. Without loss of generality, I re-write the 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 nonseparably 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 i.i.d. over firms and time.

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

(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 which is additive in the innovation shocks are necessary to form conditional moment restrictions in the proxy variable framework. The error term in those models are the differences between realized productivity and the firm's expected productivity. The contribution of nonseparability 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 knowledge investment from 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, which I model 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. More specifically, I show that returns to productivity vary between firms subjected to different

shocks in the productivity and R&D processes. I also examine whether firms with large R&D expenditures, who experience low productivity shocks, have higher/lower persistence of productivity history depending on productivity levels.

Industries with substantial periods of restructuring are also characterized by entry and exit of firms due to changes in future expected productivity levels. Therefore, artificially balancing the data may lead to selection bias if firm's beliefs about future productivity is partially determined by their current productivity. OP show a particular form of bias in the production function estimates in the presence of non-random exit. It is not straightforward to characterize 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 Appendix C.4, I propose a strategy to correct for non-random exit using their framework. The main contribution of this extension is to show that sample selection may affect the entire distribution of productivity. To summarize the restrictions on productivity, I provide the following assumptions:

Assumption 2.2 *(Productivity)*

(a) The productivity innovation shocks ξ_{it} are i.i.d. 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 i.i.d. 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

these 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. 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 important for understanding how firm's hiring decisions are affected by technological developments such as 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, $\epsilon_{\ell,it}$ are i.i.d. and independent of current period state variables including previous labor decisions. In Appendix C.2, I show how this model can be used to capture employment decisions in response to adjustment shocks to labor. This is important from a policy perspective for examining 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 i.i.d. 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 $\epsilon_{\ell,it}$ and $\epsilon_{m,it}$ are i.i.d. across firms and time.

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

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

2.4 Investment

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

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

where $\Pi_t(\cdot)$ is current period profits as a function of the state variables. Current costs to investment are given by $c(I_t)$, β is the firm's discount factor, and $\mathcal{I}_t$ is the information available to the firm when making investment decisions. I introduce an empirical investment rule (in logs) for (9) given by

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

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 extending this methodology to other empirical applications. This is easily implemented in my quantile modelling due to the equivariance property of quantiles. Capital accumulates according to the following generalized law of motion

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

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 i.i.d. 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 nonlinear 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* $f(y_{t+1}|I_t, \omega_t, Z_t) = f(y_{t+1}|\omega_t, Z_t)$.

The first equality of Assumption 3.1 states that conditional on productivity ω_t and Z_t, future output y_{t+1} and current investment I_t do not provide any additional information about current output y_t. The second equality states that conditional on ω_t and Z_t, current investment I_t does not provide any additional information about future output y_{t+1}. These are satisfied by mutual independence assumptions on η_t and ζ_t conditional on $(k_t, l_t, m_t, \omega_t)$ 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 $L_{y_t|\omega_t,Z_t}$ and $L_{y_{t+1}|\omega_t,Z_t}$ are injective.*

This allows me to take inverses of the operators. Consider the operator $L_{y_t|\omega_t,Z_t}$. Following Hu and Schennach (2008), injectivity of this operator can be interpreted as its corresponding density $f_{y_t|\omega_t,Z_t}(y_t|\omega_t,Z_t)$ 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 $f_{y_t|\omega_t,Z_t}(y_t|\omega_t,Z_t)$. More formally, for a given $Z_t \in Supp(Z_t)$,

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

for all y_t implies $g(\omega_t) = 0$ for all ω_t. For injectivity of the second operator $L_{y_{t+1}|\omega_t,Z_t}$, 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 to utilize convolution type arguments, which require conditional independence assumptions and regularity conditions on conditional characteristic functions. I also require two additional assumptions.

Assumption 3.3 *(Uniqueness): For any $\bar{\omega}_t, \tilde{\omega}_t \in \Omega$, the set $\{f_{I_t|\omega_t,Z_t}(I_t|\bar{\omega}_t,Z_t) \neq f_{I_t|\omega_t,Z_t}(I_t|\tilde{\omega}_t,Z_t)\}$ 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 $f_{I_t|\omega_t,Z_t}(I_t|\omega_t,Z_t)$ 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 $\Gamma[f_{y_t|\omega_t,Z_t}(y_t|\omega_t,Z_t)] = \omega_t$.*

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 nonseparable model, this assumption is satisfied if $E[y_t|\omega_t,Z_t]$ is strictly increasing in ω_t. Then one could normalize $\omega_t = E[y_t|\omega_t,Z_t]$. In my empirical application, I use a nonseparable Translog production function. In this case, the normalization can be achieved by setting $E[y_t|\omega_t,0] = \omega_t$, which is

standard in the production function with separable productivity. In my model, this requires restrictions on a subset of parameters. With these assumptions, I can now state the first part of the identification results.

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 shown 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 auto-regressive process. These data requirements are similar to the control function approach, where the instrument set often includes secondary lags of inputs.

4 Econometric Procedure

This section presents the model specifications and econometric strategy that are used in the empirical 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 nonlinear 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 as $Q_t^y(k_{it}, l_{it}, m_{it}, \omega_{it}; \beta(\tau))$. 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

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

where τ_η denotes the rank of 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

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

In order to account for the unobserved productivity in the marginal effects, I propose a simulation-based method to construct productivity evolution using the estimated parameters of the model. This allows for a better visualization for heterogeneous estimates as opposed to reporting the individual coefficients.

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

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

This allows me to capture heterogeneous persistence of productivity, which can depend on the level of previous productivity and the size of the innovation shock. In Appendix C.4, I show that Equation (17) must be modified to account for the fact that unobserved selection alters the productivity distribution rankings. When I augment the productivity model with R&D activities in Appendix C.3, I consider the following specification, which is similar to Doraszelski and Jaumandreu (2013)

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

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

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

where $\phi_{r,j}$ is a nonlinear function approximated by a tensor product Hermite polynomial. This function can capture interaction effects between capital and productivity. I specify an initial condition for productivity as:

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

where $\phi_{\omega_1,j}$ is approximated by a second degree polynomial in initial capital k_{i1}.

I approximate the input demand functions using tensor product Hermite polynomials in the state variables. For example, I specify the labor input demand function as:

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

where $\phi_{\ell,j}$ is approximated 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:

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

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

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

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

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

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

It is important to note that the functional forms I consider do not guarantee monotonicity in τ, but the estimator discussed in the next section automatically re-arranges quantiles to enforce monotonicity similar to Chernozhukov *et al.* (2010). It would be interesting to consider other shape restrictions in the quantile modelling presented here. The Translog production function provides a preliminary illustration of the flexibility of the identification and estimation strategy, which can be modified to include returns to scale restrictions. For more general production functions, elasticity of substitution restrictions can also be imposed. This may be pursued similarly as Blundell *et al.* (2017), who provide a quantile regression

framework for nonseparable demand functions with shape constraints. This approach would provide more structure on the functional forms in this section and is left for future research agenda.

5 Implementation

The following conditional moment restrictions hold as an implication of Assumptions 2.1-2.4 (constant omitted in conditioning set). For the production function:

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

For the labor input demand function:

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

For the material input demand function:

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

For the investment demand function:

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

For the productivity process at $t \geq 2$:

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

and for initial productivity:

$$\mathbb{E}\left[\Psi_\tau\left(\omega_{i1} - \sum_{j=1}^{J} \rho_{\omega_1,j}(\tau)\phi_{\omega_1,j}(k_{i1})\right)\middle| k_{i1}\right] = 0, \tag{32}$$

where $\Psi_\tau(u) = \tau - \mathbb{1}\{u < 0\}$. Estimating the parameters from the 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 productivity. To fix ideas, consider the following unconditional moment restriction corresponding to the production function from Equation (27):

$$\mathbb{E}\left[\int \Psi_\tau\Big(y_{it} - Q_t^y(k_{it}, l_{it}, m_{it}, \omega_{it}; \beta(\tau))\Big) \otimes \begin{pmatrix} k_{it} \\ l_{it} \\ m_{it} \\ \omega_{it} \end{pmatrix} g_i(\omega_i^T; \theta(\cdot)) d\omega_i^T\right] = 0, \tag{33}$$

where $\omega_i^T = (\omega_{i1}, \ldots, \omega_{iT})$ and $\theta(\cdot) = (\beta(\cdot), \alpha_l(\cdot), \alpha_m(\cdot), \delta(\cdot), \rho(\cdot), \rho_{\omega_1}(\cdot))$ denotes a vector of all the model parameters. The posterior density $g_i(\omega_i^T; \theta(\cdot)) = f(\omega_i^T | y_i^T, k_i^T, l_i^T, m_i^T, i_i^T; \theta(\cdot))$ conditions on the entire set of model parameters. This is due to the equivalence between 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-linear 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 constant on $(0, \tau_1]$ and $(\tau_Q, 1)$. Therefore, I write for all $q = 1, \ldots, Q-1$:

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

The intercept coefficient, $\beta_0(\tau)$, is specified as the quantile of an exponential distribution on $(0, \tau_1]$ (indexed by λ_β^-) and $(\tau_{Q-1}, 1)$ (indexed by λ_β^+) 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 has a closed form expression and does not rely on strong distributional assumptions. For example, the density corresponding to the production

function can be written as:

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

The exponential parameters are updated using a likelihood approach:

$$\lambda_\beta^- = \frac{-\mathbb{E}[\int \mathbb{1}\{y_t \leq Q_t^y(\cdot;\beta(\tau_1))\} g_i(\omega_i^T;\theta(\cdot))d\omega_t]}{\mathbb{E}[\int (y_t - Q_t^y(\cdot;\beta(\tau_1)))\mathbb{1}\{y_t \leq Q_t^y(\cdot;\beta(\tau_1))\} g_i(\omega_i^T;\theta(\cdot))d\omega_t]},$$

and

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

To estimate the model, the integral inside the expectation of Equation (33) 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 proceeds 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

$$g_i(\omega_i^T;\hat{\theta}^{(s)}) = f(\omega_i^T|y_i^T,k_i^T,l_i^T,m_i^T,i_i^T;\hat{\theta}^{(s)}) \propto$$
$$\prod_{t=1}^{T} f(y_{it}|k_{it},l_{it},m_{it},\omega_{it};\hat{\beta}^{(s)}) f(l_{it}|k_{it},\omega_{it};\hat{\alpha}_l{}^{(s)}) f(m_{it}|k_{it},l_{it},\omega_{it};\hat{\alpha}_m^{(s)})$$
$$\times f(i_{it}|k_{it},\omega_{it};\hat{\delta}^{(s)}) \prod_{t=2}^{T} f(\omega_{it}|\omega_{it-1};\hat{\rho}^{(s)}) f(\omega_{i1}|k_{i1};\hat{\rho}_{\omega_1}^{(s)}).$$

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

$$\hat{\boldsymbol{\beta}}(\tau_q)^{(s+1)} = \underset{\boldsymbol{\beta}(\tau_q)}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{m=1}^{M} \psi_{\tau_q}\left(y_{it} - Q_t^y(k_{it}, l_{it}, m_{it}, \omega_{it}^{(m)}; \beta(\tau_q))\right),$$

$$\hat{\boldsymbol{\alpha}}_l(\tau_q)^{(s+1)} = \underset{\boldsymbol{\alpha}_\ell(\tau_q)}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{m=1}^{M} \psi_{\tau_q}\left(l_{it} - \sum_{j=1}^{J} \alpha_{\ell,j}(\tau_q)\phi_{l,j}(k_{it}, \omega_{it}^{(m)})\right),$$

$$\hat{\boldsymbol{\alpha}}_m(\tau_q)^{(s+1)} = \underset{\boldsymbol{\alpha}_m(\tau_q)}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{m=1}^{M} \psi_{\tau_q}\left(m_{it} - \sum_{j=1}^{J} \alpha_{m,j}(\tau_q)\phi_{m,j}(k_{it}, l_{it}, \omega_{it}^{(m)})\right),$$

$$\hat{\boldsymbol{\delta}}(\tau_q)^{(s+1)} = \underset{\boldsymbol{\delta}(\tau_q)}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{m=1}^{M} \psi_{\tau_q}\left(i_{it} - \sum_{j=1}^{J} \delta_j(\tau_q)\phi_{\iota,j}(k_{it}, \omega_{it}^{(m)})\right),$$

$$\hat{\boldsymbol{\rho}}(\tau_q)^{(s+1)} = \underset{\boldsymbol{\rho}(\tau_q)}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{m=1}^{M} \psi_{\tau_q}\left(\omega_{it}^{(m)} - \rho_0(\tau_q) - \rho_1(\tau_q)\omega_{it-1}^{(m)} - \rho_2(\tau_q)\omega_{it-1}^{(m)2} - \rho_3(\tau_q)\omega_{it-1}^{(m)3}\right),$$

$$\hat{\boldsymbol{\rho}}_{\omega_1}(\tau_q)^{(s+1)} = \underset{\boldsymbol{\rho}_{\omega_1}(\tau_q)}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{m=1}^{M} \psi_{\tau_q}\left(\omega_{i1}^{(m)} - \sum_{j=1}^{J} \rho_{\omega_1}(\tau_q)\phi_{\omega_1,j}(k_{i1})\right),$$

where $\psi_\tau(u) = (\tau - \mathbb{1}\{u < 0\})u$ is the "check" function from quantile regression. The exponential parameters for the intercept coefficients (e.g. the production function) are updated from:

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

and

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

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 provides 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 discuss asymptotic properties of the estimator in this paper.

6 Application

The estimator is applied to data on U.S. 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.

The estimation algorithm is ran with 500 random-walk Metropolis-Hastings steps and 200 EM steps, taking $M = 1$. In the sampling algorithm, productivity is drawn from a normal distribution centered at the current draw of productivity with variance equal to 0.01. This achieves an average acceptance rate of 10%. The final estimates are used to simulate productivity from its initial conditions and the decision rules for investment, labor, and materials. A capital accumulation process is needed to simulate these values. The process specified in Equation (11) is flexible and is used to accumulate capital following the perpetual inventory method with a constant depreciation rate set at 0.02. Estimates are similar using different types of capital accumulation processes and depreciation rates. All results in this paper are plotted and provided interactively on my personal webpage.

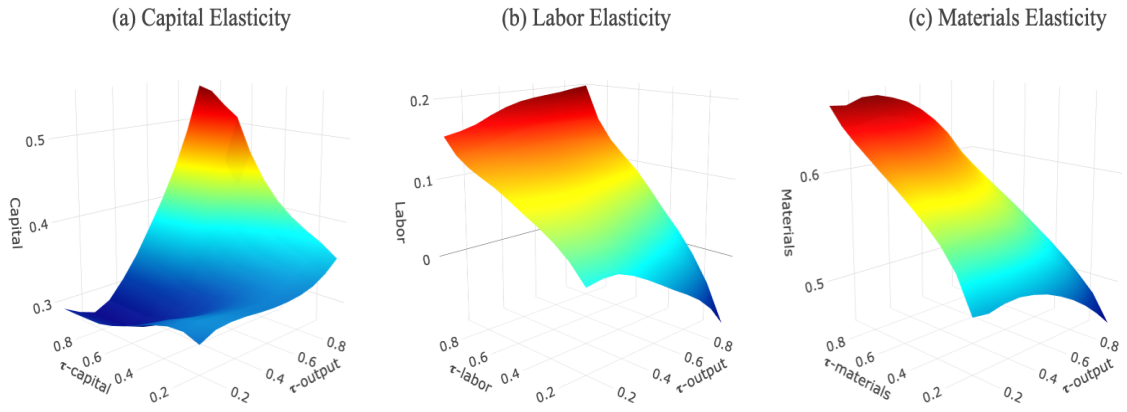
6.1 Empirical Results

6.1.1 Production Function Estimates

Estimates of the heterogeneous production function elasticities are shown in Figures 1 and 2. Panel (a) of Figure 1 reports the estimates of the average capital elasticity evaluated at percentiles of capital and output. The estimates range from 0.28 for firms at the lowest percentile of output and highest percentile of capital to 0.54 for firms at the highest percentile of output and capital. For firms with high levels of capital, there is more heterogeneity across the output distribution, contrasted to the low heterogeneity for firms at the lower percentiles

of capital. Panel (b) reports the average labor elasticity evaluated at percentiles of labor and output. The relationship is opposite to that of capital. The estimates are -0.08 for firms at the highest percentile of output and lowest percentile of labor, and 0.2 for firms in the highest percentile of output and labor. For these estimates, there is larger heterogeneity across firms with low levels of labor than firms who use more labor. Panel (c) shows the estimates of the average materials elasticity evaluated at percentiles of materials and output. The relationship is similar to labor. The estimates are lowest at 0.46 for firms at the highest percentile of output and lowest percentile of materials, and highest at 0.66 for firms who use high levels of materials at the bottom of the output distribution. Overall, these results suggest that elasticities vary over the size of the firm measured by the rank on the conditional output distribution and the amount of inputs a firm uses.

Figure 1: Output Elasticities



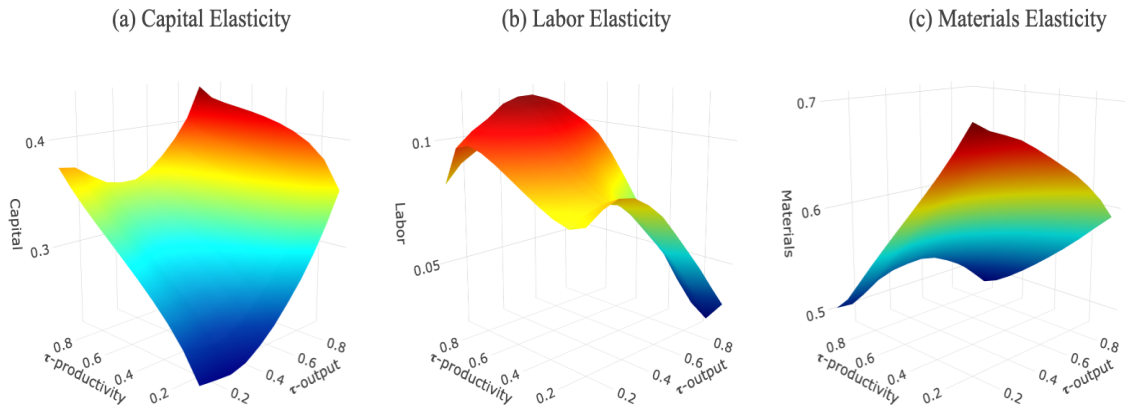
*Panel (a): Capital elasticity evaluated at τ_η and percentiles of capital τ_k averaged over values of ω_{it} and (l_{it}, m_{it}) that correspond to τ_k. Panel (b): Labor elasticity evaluated at τ_η and percentiles of labor τ_l averaged over values of ω_{it} and (k_{it}, m_{it}). Panel (c): Materials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

In Figure 2, I report similar estimates as Figure 1, instead evaluating the estimates at fixed percentiles of productivity. Panel (a) reports the capital elasticities over percentiles of output and productivity. These results are more heterogeneous for firms with different levels of productivity. For low output and productivity firms, the capital elasticity is 0.229. For high output and productivity firms, this is 0.435. For low to medium productivity firms, these estimates are increasing faster in the rank of the output distribution, but for high productivity firms this relationship is U-shaped. For low output firms, capital elasticity increases faster when the firm is more productive. For high output firms, the rate at

which the estimates increase is slower. These results imply that unobserved productivity is an important dimension of firm heterogeneity in capital elasticities. Panel (b) reports the estimates of labor elasticity. For low output and productivity firms, the estimate is 0.081 and for high output and productivity firms, the estimate is 0.07. Similar to panel (b) in Figure 1, estimates are decreasing in output size except for firms at the highest percentile of productivity, where the relationship is inverse U-shaped. Labor elasticity estimates are increasing in productivity except for the most productive firms, where the estimates begin to fall. Estimates rise faster for high output firms than low output firms. Panel (c) reports the material elasticities over percentiles of productivity. For low output and productivity firms, estimates are highest at 0.7, and for high output and productivity firms, estimates are lowest at 0.491. For low productivity firms, material elasticities are decreasing in output size, but inverse U-shaped for high productivity firms. Overall, these estimates are increasing in productivity size for fixed levels of output, although estimates increase faster for low output firms compared to higher output firms.

These exercises emphasize the importance of heterogeneity of output elasticities in non-separable models. Both the size of the input demand and productivity levels reveal dramatic differences in these estimates. Similar estimates can be illustrated for fixed percentiles of the other inputs. For example, evaluating the capital elasticity over fixed percentiles of labor may show some complementarities or substitution patterns between the two inputs. These illustrations are provided as an extension to the main results in Appendix C.1.

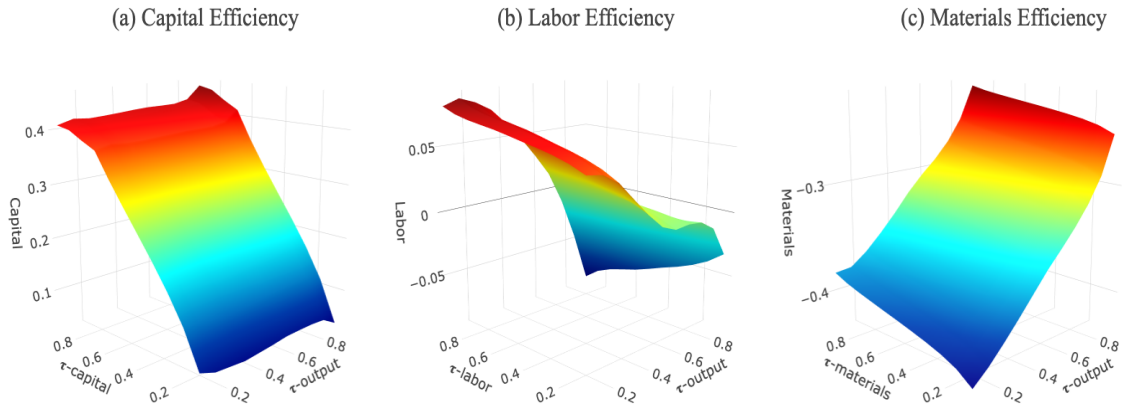
Figure 2: Output Elasticities



*Panel (a): Capital elasticity evaluated at τ_η and τ-productivity averaged over values of (k_{it}, l_{it}, m_{it}) that correspond to τ-productivity. Panel (b): Labor elasticity evaluated at evaluated at τ_η and τ-productivity averaged over values of (k_{it}, l_{it}, m_{it}) that correspond to τ-productivity. Panel (c): Materials elasticity evaluated at τ_η and τ-productivity averaged over values of (k_{it}, l_{it}, m_{it}) that correspond to τ-productivity.

Figure 3 presents estimates of the non-Hicks neutral effects of productivity on capital, labor, and material inputs. These effects can be interpreted as the marginal productivity of inputs with respect to small changes in productivity. The labor-augmenting aspect of the shock is of particular importance, since the empirical literature often points to labor productivity as sources of long-run economic growth. Despite its importance, there are relatively few papers that study these sources of productivity at the firm-level. This is because identification and estimation of these models are difficult due to the issues of endogeneity and multi-dimensional productivity.[3] 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, however the estimates under this alternative would be more suited for comparison to existing empirical work.

Figure 3: Non-Hicks Neutral Effects



*Panel (a): Capital efficiency 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 efficiency evaluated at τ_η and percentiles of labor τ_l averaged over values of (k_{it}, m_{it}). Panel (c): Materials efficiency evaluated at τ_η and percentiles of materials τ_m averaged over values of (k_{it}, l_{it}).

Panel (a) in Figure 3 shows the results for the capital-augmenting effect of 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 0.44 for firms at the highest percentiles of output and capital. Overall, the capital efficiency effects are increasing for

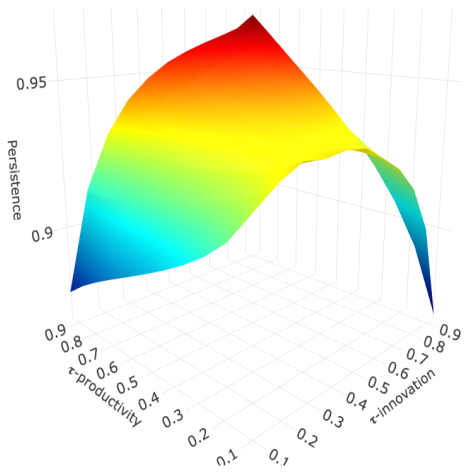
[3]As mentioned earlier, Doraszelski and Jaumandreu (2018) have made progress in this regard using rich firm-level data. Dermirer (2020) also studies nonparametric identification of these models and applies his estimates to U.S. public manufacturing firms.

firms who use more capital, but almost constant across the conditional output distribution. For the labor efficiency estimates in the panel (b), there is more heterogeneity between firms of different sizes who use varying amounts of labor. The estimates range from -0.08 for firms at the highest percentile of output and labor, to 0.08 for firms at the lower percentiles of output, but the highest percentile of labor. Interestingly, the labor estimates are decreasing in labor size for the smallest percentile of output but increasing at the largest percentile. These results seem consistent with empirical results that find large amounts of firm variation in labor-productivity, and suggests that smaller firms use labor more efficiently than larger firms in this sample. Panel (c) reports the material efficiency estimates. These range from -0.42 for firms at the lowest percentiles of output and materials, to -0.22 for firms at the highest percentiles of output and materials. These estimates reveal that for firms in this sample, an increase in productivity leads to a decrease in the marginal product of materials. This could suggest that either firms are inefficient in their usage of materials or that there are structural differences in materials productivity that are not captured by this model.

6.1.2 Persistence of Productivity

Next, I examine the estimates of the productivity process. Figure 4 reports the estimates of productivity persistence at various percentiles of the innovation shock and percentiles of last period productivity. Persistence 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 negative shock (0.92). This indicates that for high productivity firms, large unanticipated negative shocks can reduce the history of high productivity by more than firms with a history of low productivity. This relationship changes when firms are hit by large positive shocks. High productivity firms have higher persistency of productivity history (0.97) than low productivity firms (0.86) when hit by positive shocks.

Figure 4: 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 changes in productivity and innovation shocks. The input responses to productivity are calculated as the derivative of Equations (21), (23), and (24) with respect to productivity. These provide insights on how firms input demand changes in response to technological or organizational innovations measured by the unobserved productivity component. For example, whether a firm adjusts 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 at different percentiles of productivity and input demand.

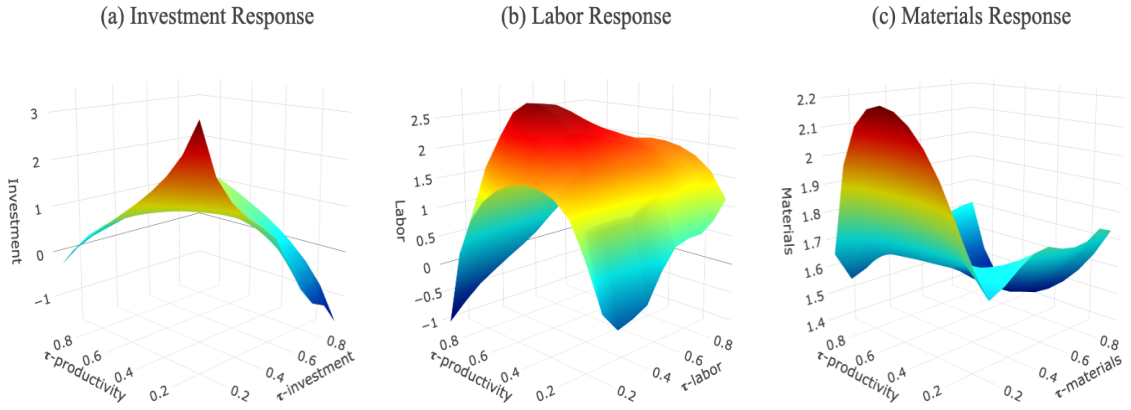
Panel (a) in Figure 5 shows the relationship between investment demand and productivity. Firms at the lowest percentile of investment and productivity have the largest productivity response at 3.2. As productivity increases for lower investment firms, this effect decreases to -0.28. For high investment firms at the lowest percentile of productivity, the effect is -1.58. For similar levels of investment, high productivity firms have an effect equal to 1.18. 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 demand and productivity. Firms at the lowest percentile of labor and productivity have an effect equal to 0.83. For firms at the

lowest percentile of labor and highest percentiles of productivity, the effect is equal to -1. For firms at the highest percentile of labor, but the lowest percentile of productivity, the effect approaches its maximum around 1.13, but for firms at highest percentile of productivity this effect diminishes to 0.67. The overall shape of the labor productivity response is an inverse U-shape. For firms who use the smallest and largest amounts of labor, the estimates are more heterogeneous across the output distribution than for firms at the median of labor demand. These results show that for firms who use less labor and are less productive, increases in productivity leads to a decrease in the amount of labor, whereas large labor firms increase labor in response to productivity changes at any level of productivity.

Panel (c) shows the relationship between material input demand and productivity. Firms at the lowest percentile of materials and productivity have a productivity effect equal to 1.7. For firms at the lowest percentile of materials and highest percentiles of productivity, this effect is equal to 1.64. For firms at the highest percentile of materials, but the lowest percentile of productivity, the effect is equal to 1.73, and for the highest percentile of productivity is 1.71. For firms who use the smallest amounts of materials, the productivity effect is strongly inverse U-shaped, but quickly decreases to a U-shaped curve for higher levels of material demand. Overall, firms respond to productivity increases by using more material inputs.

Figure 5: Input Demand Responses to Productivity



*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} and l_{it}.

6.1.4 Impulse Responses to Productivity Shocks

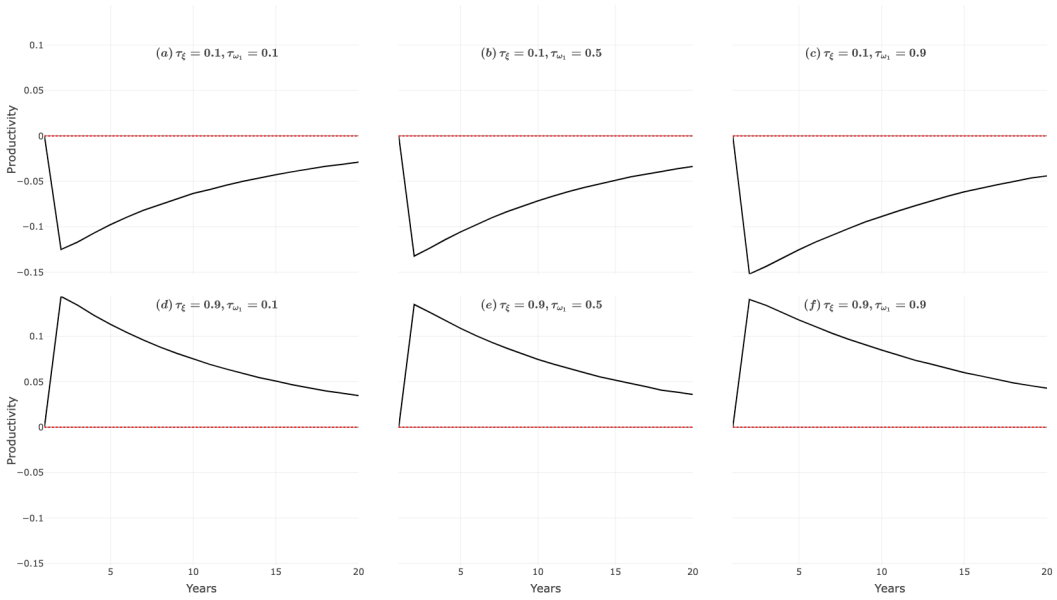
This section simulates the impact of innovation shocks to the productivity process and input demand functions using estimates from the model. Similar to HHS, I estimate how quickly firms respond to shocks to current productivity. This analysis will show whether input decision rules for capital, labor, and materials 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 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:

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

where $\partial Q_t^\omega(\omega_{it-1}, \tau_\xi)/\partial \xi_{it}$ 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.

Figures 6, 7, 8 and 9 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, capital, labor and materials. I simulate the model so that the impact of the shock occurs at $t = 2$. I examine the initial responses 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 HHS. In their paper, they study how quickly firms adjust inputs in response to the latest shocks in 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 estimates document the impact of differently sized innovation shocks and input demand functions beyond the mean, as well as the full history of the impact.

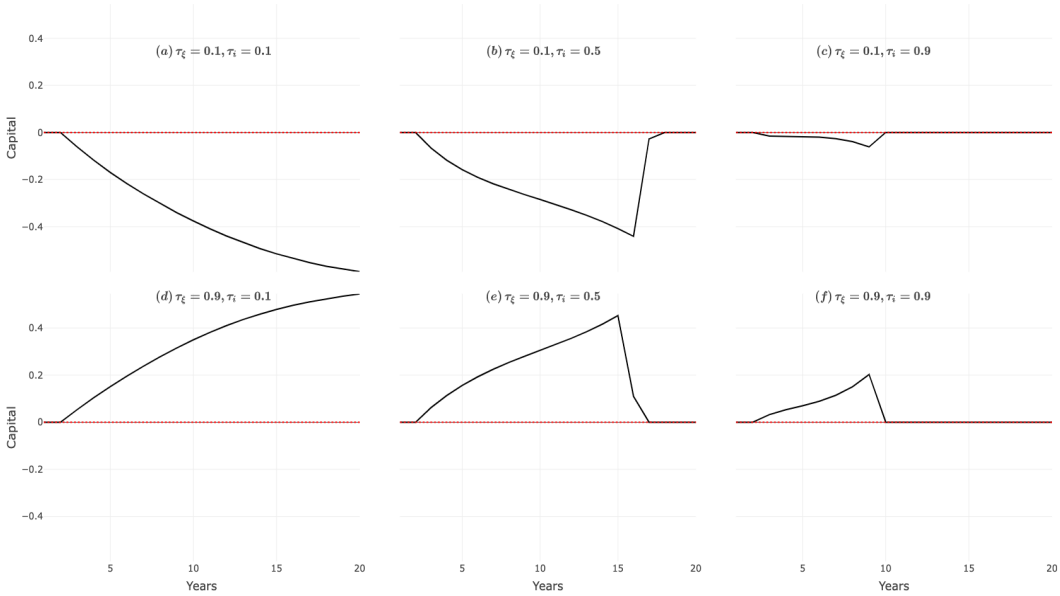
Figure 6: Impulse Response of an Innovation Shock to Productivity



*Top row: Differences in productivity 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: Differences in productivity 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 responses to innovation shocks are reported in Figure 6, 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 observable 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 hit by negative innovation shocks.

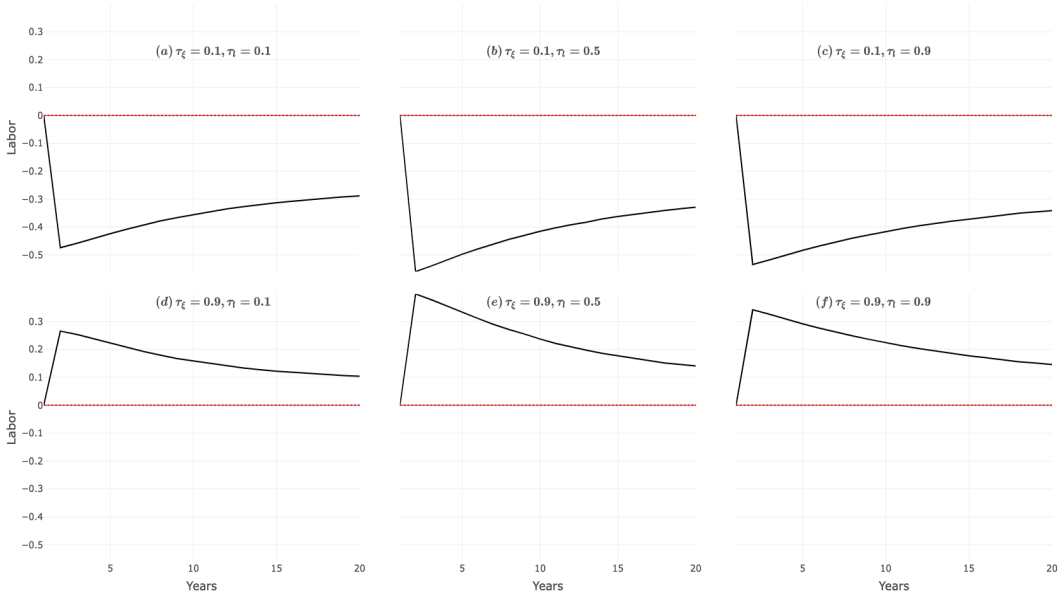
Figure 7: Impulse Response of an Innovation Shock to Capital



*Top row: Differences in capital 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: Differences in capital 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 capital responses to innovation shocks are reported in Figure 7, which shows the impact of 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 initially decreases capital by 6% and afterwards capital usage continues to fall. The opposite situation occurs for high investment firms hit by large productivity shocks. For both cases, the rate at which capital falls/rises decreases over time. For firms with large investment demand, a large negative productivity shock leads to a small decrease in capital usage (1%). When these firms are hit by a large positive productivity shock, capital increases by about 3%. Both capital responses decrease/increase until year 9 and stabilizes after year 10.

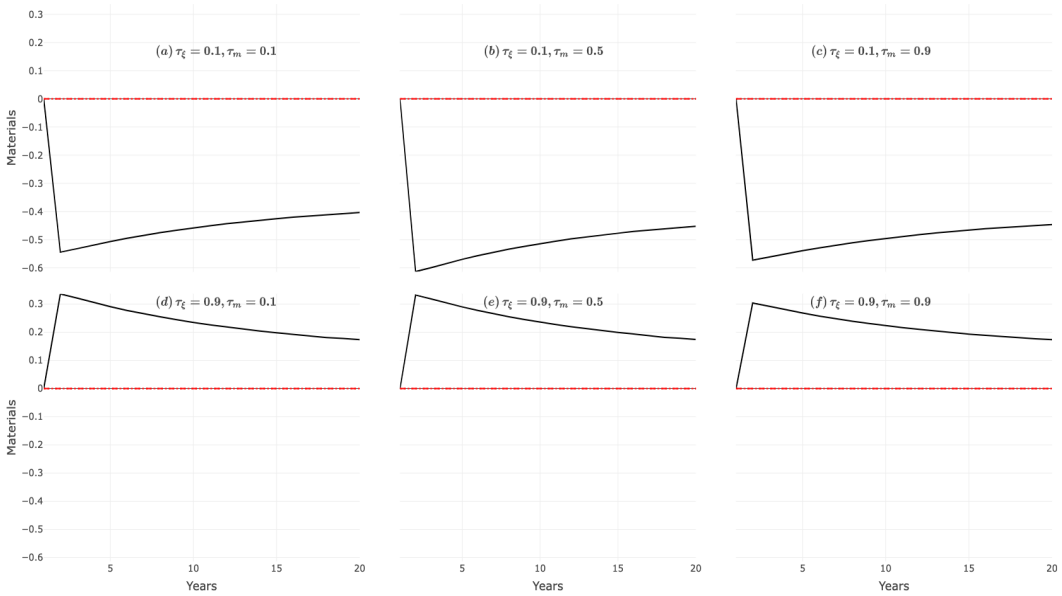
Figure 8: Impulse Response of an Innovation Shock to Labor



*Top row: Differences in labor 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: Differences in labor 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 responses to innovation shocks are reported in Figure 8, which shows the impact of a negative productivity shock in panel (a-c) and a positive productivity shock in panel (d-f) for various levels of labor demand $\tau_\ell = (0.1, 0.5, 0.9)$. For firms with the lowest labor demand, a large negative productivity shock decreases labor inputs by 47%, while a large positive shock increases labor inputs by 26%. For firms with the highest labor demand, a large negative productivity shock decreases labor inputs by 53%, and a large positive productivity shock increases labor inputs by about 34%.

Figure 9: Impulse Response of an Innovation Shock to Materials



*Top row: Differences in materials 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: Differences in materials 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 responses to innovation shocks are reported in Figure 9, which shows the impact of a negative productivity shock in panel (a-c) and a positive productivity shock 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 54%, while a large positive shock increases material inputs by 33%. For firms with the highest materials demand, a large negative productivity shock decreases material inputs by 57%, and a large positive productivity shock increases material inputs by 30%. These results are consistent with the overall finding that material inputs tend to be the most flexible, and hence respond more dramatically to productivity changes than other inputs such as labor and capital.

7 Conclusion

This paper proposes a nonseparable 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, which specify technology that is fixed across firms, and instead suggests that nonlinear, 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 productivity history can vary with respect to the latest innovation shocks, and that good or bad shocks have asymmetric impacts for both high and low productivity firms.

The production function, input demand functions, and productivity are nonparametrically identified in the presence of nonseparable unobservables. I show that under additional independence restrictions, conditional quantile restrictions can be imposed, and the quantile estimators can be used to capture firm-level heterogeneity. The estimator proposed in this paper is computationally tractable and involves quantile regression in each iteration of the simulation algorithm. This provides new results that have not been considered in the prior production function literature. For example, this paper shows that firms have asymmetric input adjustments in response to productivity changes. This type of analysis is useful from a policy perspective, as 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 impacts of good and bad 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. For example, I found that low investment firms with low productivity have a large continuous decrease in capital following a negative productivity shock.

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. 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 results from the standard production function model.

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. 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.

Blundell, R., Horowitz, J. and Parey, M. (2017). Nonparametric estimation of a nonseparable demand function under the slutsky inequality restriction. *The Review of Economics and Statistics*, **99** (2), 291–304.

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.

Chernozhukov, V., Fernndez-Val, I. and Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, **78** (3), 1093–1125.

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 Schwartz, J. T. (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*, **128** (8), 2973–3016.

Hu, Y., Huang, G. and Sasaki, Y. (2020). Estimating production functions with robustness against errors in the proxy variables. *Journal of Econometrics*, **215** (2), 375–398.

— 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), 457–489.

Olley, G. S. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, **64** (6), 1263–1297.

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), 279–300.

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: Deflated 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 of their decomposition technique can be found in their paper.

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

$$\begin{aligned} f_{y_t,I_t|y_{t+1},Z_t} &= \int f_{y_t,I_t,\omega_t|y_{t+1},Z_t}(y_t, I_t, \omega_t|y_{t+1}, Z_t)d\omega_t \\ &= \int f_{y_t|y_{t+1},I_t,\omega_t,Z_t}(y_t|y_{t+1}, I_t, \omega_t, Z_t) f_{I_t|y_{t+1},\omega_t,Z_t}(I_t|y_{t+1}, \omega_t, Z_t) f(\omega_t|y_{t+1}, Z_t)d\omega_t \\ &= \int f_{y_t|\omega_t,Z_t}(y_t|\omega_t, Z_t) f_{I_t|\omega_t,Z_t}(I_t|\omega_t, Z_t) f(\omega_t|y_{t+1}, Z_t)d\omega_t, \end{aligned} \tag{34}$$

where the third line follows from applying the conditional independence in Assumption 3.1. The goal of the identification strategy is to show that the conditional densities in Equation (34) 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, the following are equivalent:

$$L_{y_t,I_t|y_{t+1}} = L_{y_t|\omega_t}\Delta_{I_t|\omega_t}L_{\omega_t|y_{t+1}}. \tag{35}$$

Integrating (35) 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:

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

Plugging (36) into (35):

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

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:[4]

$$L_{y_t,I_t|y_{t+1}}L^{-1}_{y_t|y_{t+1}} = L_{y_t|\omega_t}\Delta_{I_t|\omega_t}L^{-1}_{y_t|\omega_t}. \tag{37}$$

The LHS of (37) 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. To identify the unobserved densities of interest, the representation in (37) and its decomposition must be unique. This is guaranteed by Theorem XV.4.5 in Dunford and Schwartz (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

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

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

[4]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:

$$L_{y_{t+1}|\omega_t} = L_{y_{t+1}|\omega_{t+1}} L_{\omega_{t+1}|\omega_t}. \tag{38}$$

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

$$L_{\omega_{t+1}|\omega_t} = L_{y_{t+1}|\omega_t} L^{-1}_{y_{t+1}|\omega_{t+1}}, \tag{39}$$

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 $(y_{t+2}, y_{t+1}, I_{t+1})$.

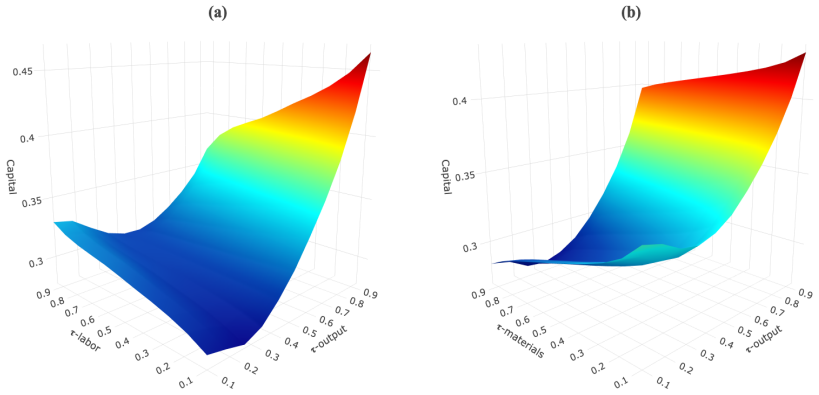
C Extensions

This section addresses the various extensions mentioned earlier in this paper. I consider the extension of the heterogeneous estimates from the main model in Section C.1. I apply the estimator to study labor adjustment frictions in Section C.2. In Section C.3, I compare estimates between R&D firms and non R&D firms. In Section C.4, I propose a correction to possible selection bias arising from non-random firm exit.

C.1 Extension of Heterogeneous Estimates from the Main Model

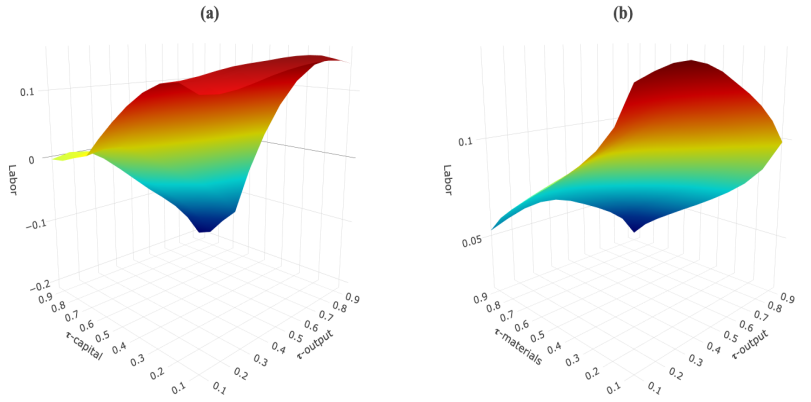
This section re-evaluates the heterogeneous estimates from Section 6.1. The estimates of the output elasticities, non-Hicks neutral effects, and marginal productivity of inputs are plotted over percentiles of the other inputs, instead of the primary input of consideration and productivity.

Figure 10: Capital Elasticities



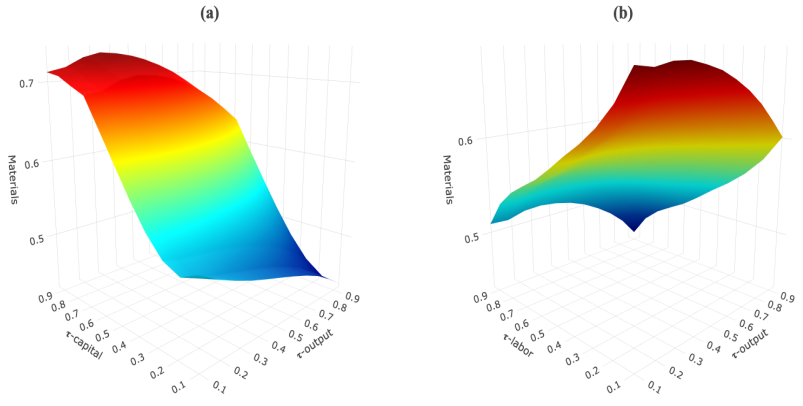
*Panel (a): Capital elasticity evaluated at τ_η and percentiles of labor τ_l averaged over values of $(k_{it}, m_{it}, \omega_{it})$. Panel (b): Capital elasticity evaluated at τ_η and percentiles of materials τ_m averaged over values of $(k_{it}, l_{it}, \omega_{it})$

Figure 11: Labor Elasticities



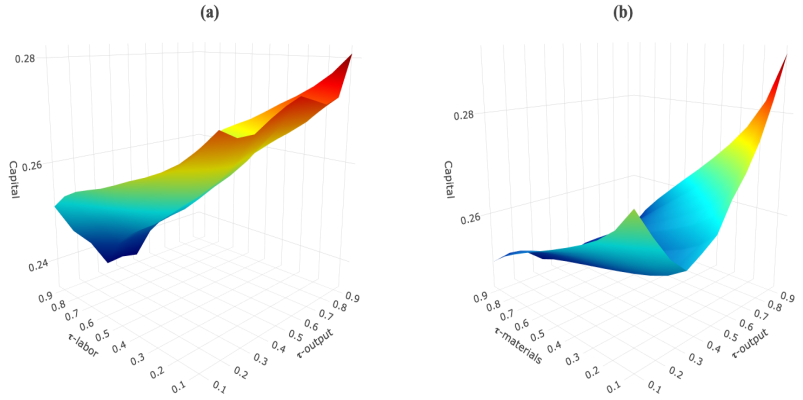
*Panel (a): Labor elasticity evaluated at τ_η and percentiles of capital τ_k averaged over values of $(l_{it}, m_{it}, \omega_{it})$. Panel (b): Labor elasticity evaluated at τ_η and percentiles of materials τ_m averaged over values of $(k_{it}, l_{it}, \omega_{it})$

Figure 12: Materials Elasticities



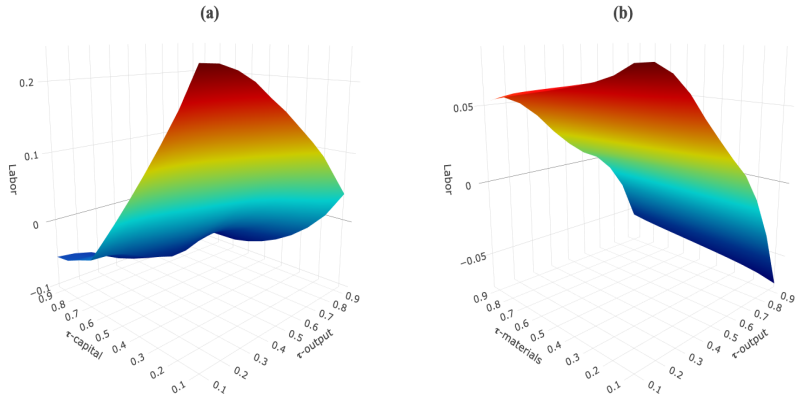
*Panel (a): Materials elasticity evaluated at τ_η and percentiles of capital τ_k averaged over values of $(l_{it}, m_{it}, \omega_{it})$. Panel (b): Materials elasticity evaluated at τ_η and percentiles of labor τ_l averaged over values of $(k_{it}, m_{it}, \omega_{it})$

Figure 13: Non-Hicks Neutral Effects of Capital



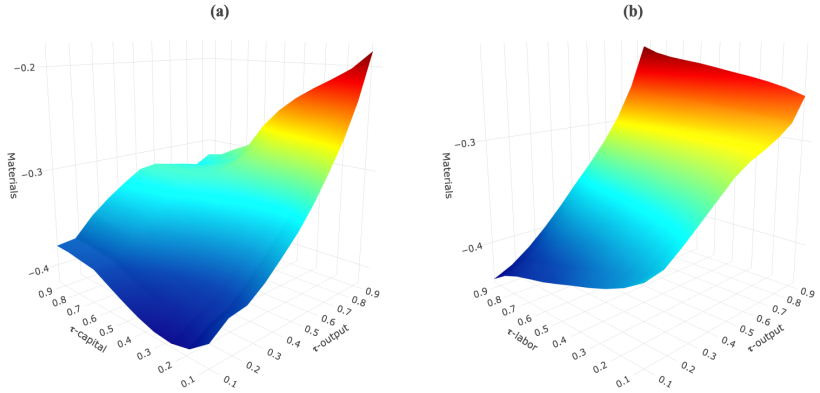
*Panel (a): Capital efficiency evaluated at τ_η and percentiles of labor τ_l averaged over values of (k_{it}, m_{it}). Panel (b): Capital efficiency evaluated at τ_η and percentiles of materials τ_m averaged over values of (k_{it}, l_{it})

Figure 14: Non-Hicks Neutral Effects of Labor



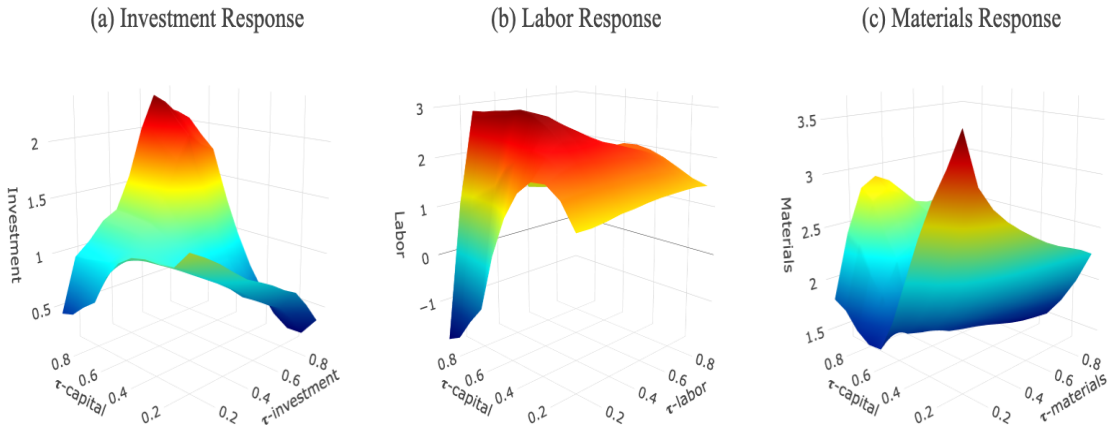
*Panel (a): Labor efficiency evaluated at τ_η and percentiles of capital τ_k averaged over values of (l_{it}, m_{it}). Panel (b): Labor efficiency evaluated at τ_η and percentiles of materials τ_m averaged over values of (k_{it}, l_{it})

Figure 15: Non-Hicks Neutral Effects of Materials



*Panel (a): Materials efficiency evaluated at τ_η and percentiles of capital τ_k averaged over values of (l_{it}, m_{it}). Panel (b): Materials efficiency evaluated at τ_η and percentiles of labor τ_l averaged over values of (k_{it}, m_{it})

Figure 16: Input Demand Response to Productivity

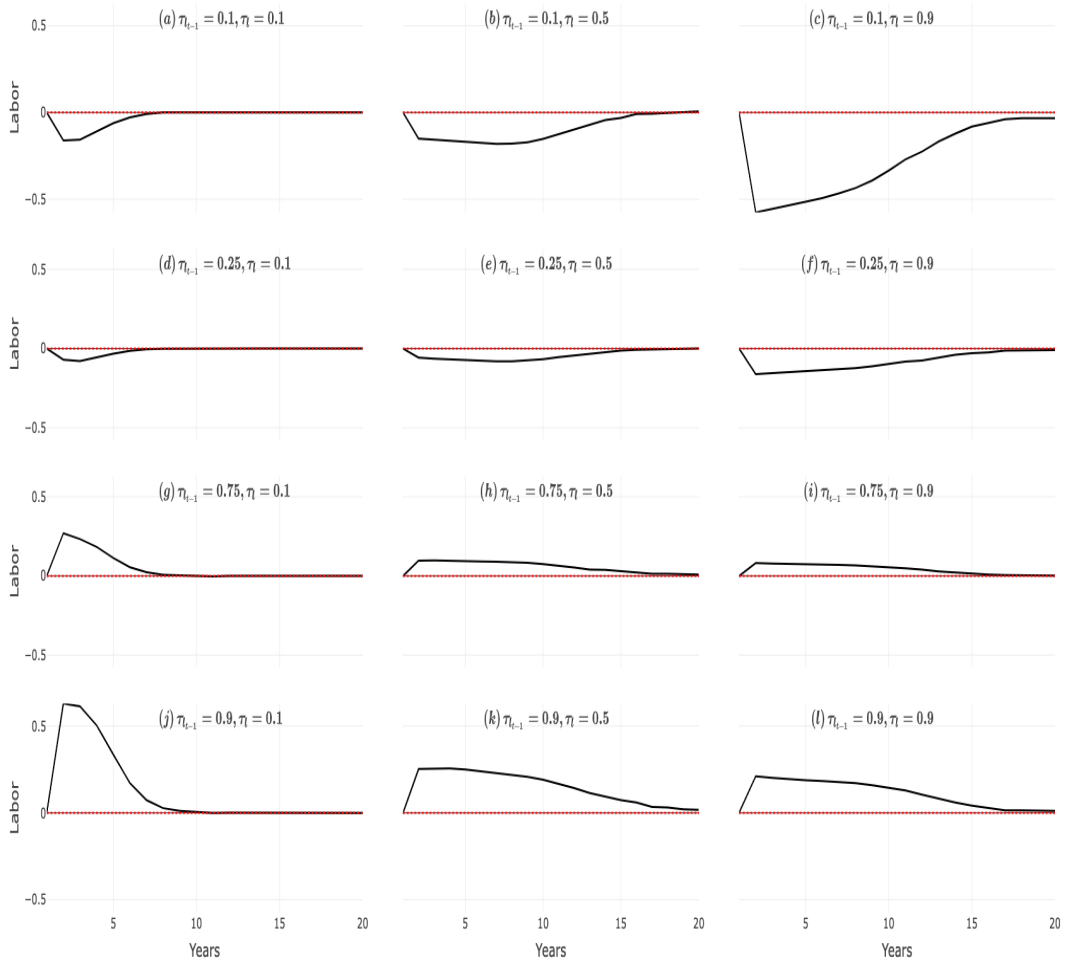


*Panel (a): Investment demand evaluated at τ_ζ and percentiles of capital τ_k averaged over productivity. Panel (b): Labor demand evaluated at τ_{ϵ_l} and percentiles of capital τ_k averaged over productivity. Panel (c): Material demand evaluated at τ_{ϵ_m} and percentiles of capital τ_k averaged over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

C.2 Labor Dynamics

This section extends the labor demand function to include lagged labor from Equation (22). In Figure 17, I report impulse response function for firms of different labor demand size who are hit with differently sized shocks to previous labor. The results show heterogeneous responses for high labor demand firms hit by low shocks and low labor demand firms hit by high shocks.

Figure 17: Impulse Response of Adjustment Shocks to Labor

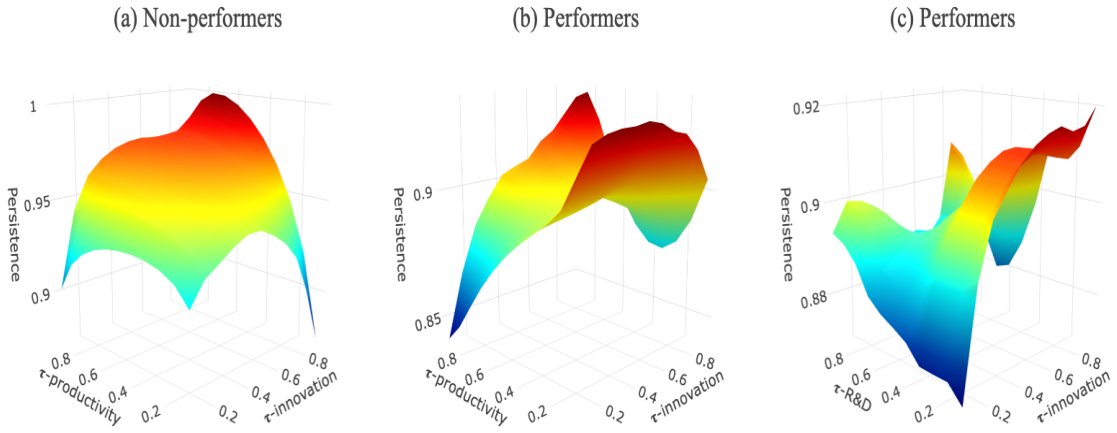


*Top row: Difference between firms hit with low labor shock $\tau_{l_{t-1}} = 0.1$ and medium shock $\tau_{l_{t-1}} = 0.5$ at different levels of labor demand. Second row: Difference between firms hit with labor shock $\tau_{l_{t-1}} = 0.25$ and medium shock $\tau_{l_{t-1}} = 0.5$ at different levels of labor demand. Third row: Difference between firms hit with labor shock $\tau_{l_{t-1}} = 0.75$ and medium shock $\tau_{l_{t-1}} = 0.5$ at different levels of labor demand. Bottom row: Difference between firms hit with high labor shock $\tau_{l_{t-1}} = 0.9$ and medium shock $\tau_{l_{t-1}} = 0.5$ at different levels of labor demand. Labor demand is evaluated at percentiles of lagged labor averaged over capital and productivity.

C.3 R&D Activities

The first set of results in Figure 18 compares the estimates of productivity persistence between firms that do not perform and those that perform R&D. In panel (a), productivity persistence is plotted at fixed percentiles of previous productivity and innovation shocks for non-R&D firms. For low productivity firms, a low shock to productivity has a higher persistency (0.91) than high shocks (0.87). For high productivity firms, large innovation shocks have higher persistency (0.97) than low shocks (0.9). In panel (b), low productivity R&D firms have higher persistence than non-R&D firms regardless of the size of the innovation shock. For high productivity R&D firms, a low and high innovation shock has lower persistence than non-R&D firms. Panel (c) reports persistence estimates for R&D firms evaluated at percentiles of the R&D distribution. Overall, the shape of persistence is quite lumpy, but it is noted that low R&D firms have higher persistence than high R&D firms regardless of the size of the innovation shock.

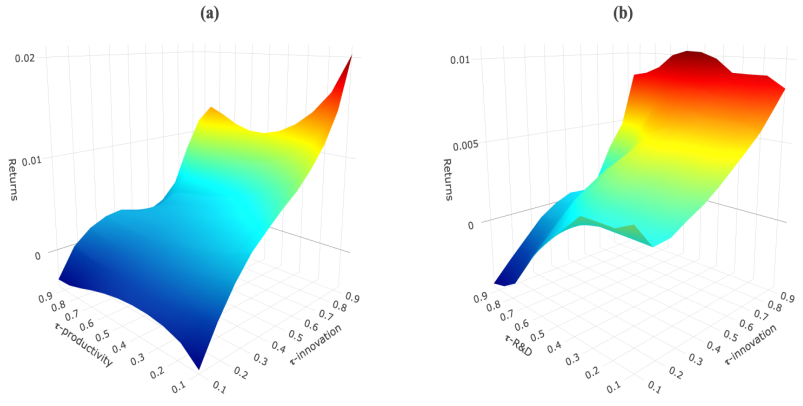
Figure 18: Productivity Persistence for Non-performing and R&D Performing Firms



*Panel (a): Estimates of average productivity persistence for non R&D firms evaluated at τ_ξ and percentiles of previous productivity. Panel (b): Estimates of productivity persistence for R&D firms evaluated at τ_ξ and percentiles of previous productivity averaged over R&D. Panel (c): Estimates of productivity persistence for R&D firms evaluated at τ_ξ and percentiles of R&D averaged over productivity.

The second set of results in Figure 19 plots the returns to R&D measured by the elasticity of productivity with respect to R&D expenditures, evaluated at various percentiles of productivity, innovation shocks, and R&D. In both figures, the returns are increasing in size of the innovation shock. In panel (a), the relationship between productivity and R&D returns is inverse U-shaped for low innovation shocks and U-shaped for high shocks. In panel (b), returns are decreasing in percentiles of R&D for low productivity shocks and somewhat flat for high shocks.

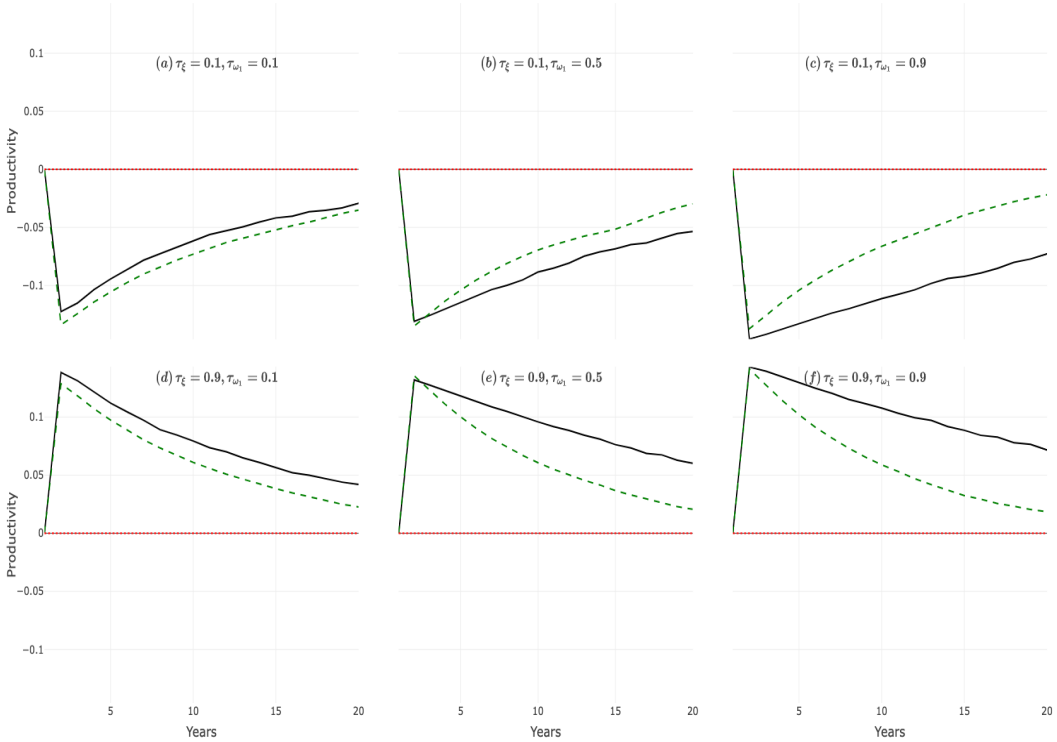
Figure 19: Returns to R&D



*Panel (a): Returns to R&D for firms evaluated at τ_ξ and percentiles of previous productivity averaged over R&D. Panel (b): Returns to R&D for firms evaluated at τ_ξ and percentiles of R&D averaged over productivity.

The productivity responses to innovation shocks are reported in Figure 20, which shows the impact of a large negative shock ($\tau_\xi = 0.1$) in panel (a-c) and a large positive shock ($\tau_\xi = 0.9$) in panel (d-f) at various levels of initial productivity $\tau_{\omega_1} = (0.1, 0.5, 0.9)$ for R&D firms (in green) and non-R&D firms (in black). The interesting result of this exercise is that both R&D and non-R&D firms experience similar decreases/increases in productivity following a negative/positive shock. However, R&D firms recover faster from negative shocks for higher levels of initial productivity.

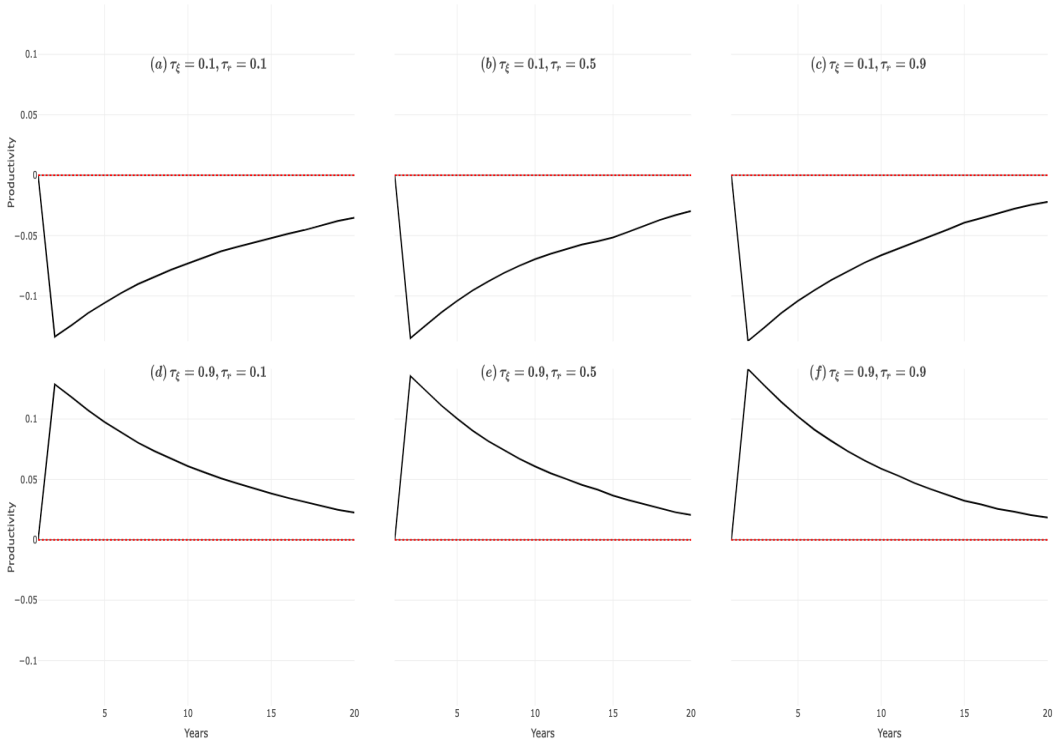
Figure 20: Impulse Response of an Innovation Shock to Productivity



*Top row: Differences in productivity 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: Differences in productivity between firms hit with high productivity shock $\tau_\xi = 0.9$ and medium shock $\tau_\xi = 0.5$ at different levels of initial productivity. Dashed green line denotes the R&D firms. The solid black line is non-R&D firms.

The productivity responses to innovation shocks for different levels of R&D are reported in Figure 21, which shows the impact of a large negative shock ($\tau_\xi = 0.1$) in panel (a-c) and a large positive shock ($\tau_\xi = 0.9$) in panel (d-f) for various levels of R&D expenditure $\tau_r = (0.1, 0.5, 0.9)$. For firms with the lowest R&D expenditure, a large negative innovation shock decreases productivity by 13%, while a large positive shock increases productivity by 12.8%. For firms with the highest R&D expenditure, a large negative innovation shock decreases productivity by 13%, and a large positive shock increases productivity by about 14%. Overall, there is not much heterogeneity in firm productivity responses to innovation shocks and R&D expenditure.

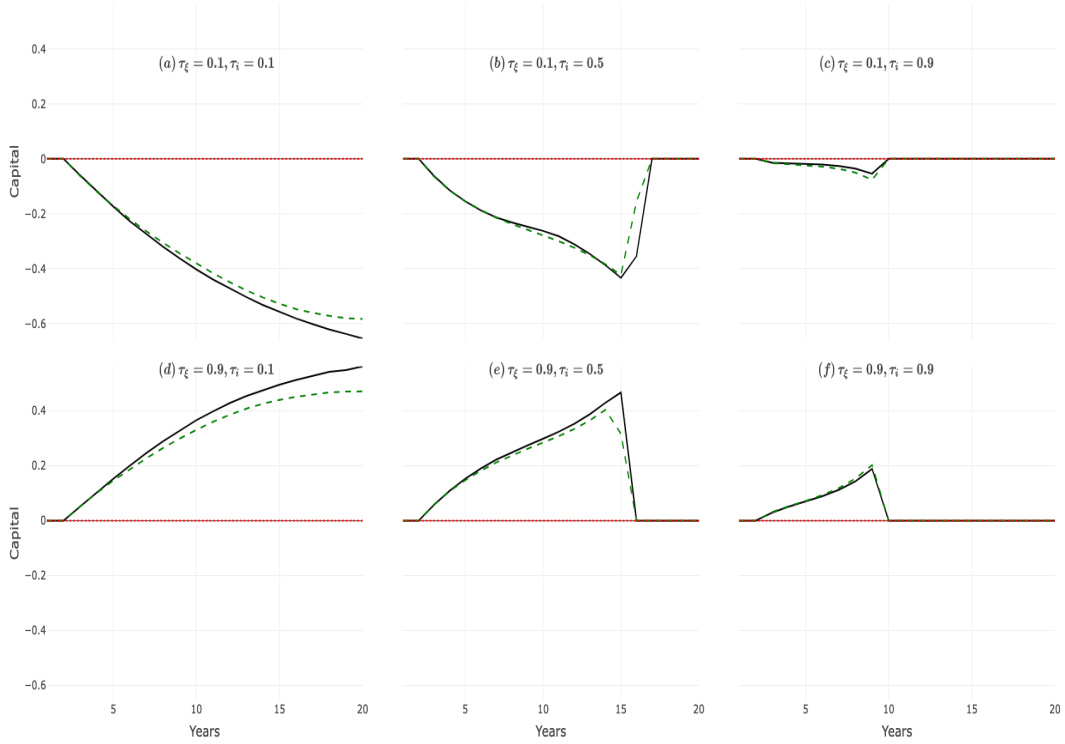
Figure 21: Impulse Response of an Innovation Shock to Productivity for Different Levels of R&D



*Top row: Differences in productivity between firms hit with low productivity shock $\tau_\xi = 0.1$ and medium shock $\tau_\xi = 0.5$ at different levels of R&D expenditure $\tau_r = (0.1, 0.5, 0.9)$. Bottom row: Differences in productivity between firms hit with high productivity shock $\tau_\xi = 0.9$ and medium shock $\tau_\xi = 0.5$ at different levels of R&D expenditure.

The capital responses to innovation shocks are reported in Figure 22, which shows the impact of a large negative shock ($\tau_\xi = 0.1$) in panel (a-c) and a large positive shock ($\tau_\xi = 0.9$) in panel (d-f) at various levels of investment $\tau_i = (0.1, 0.5, 0.9)$ for R&D firms (in green) and non-R&D firms (in black). The overall finding is that there are not significant heterogeneous capital responses between these types of firms.

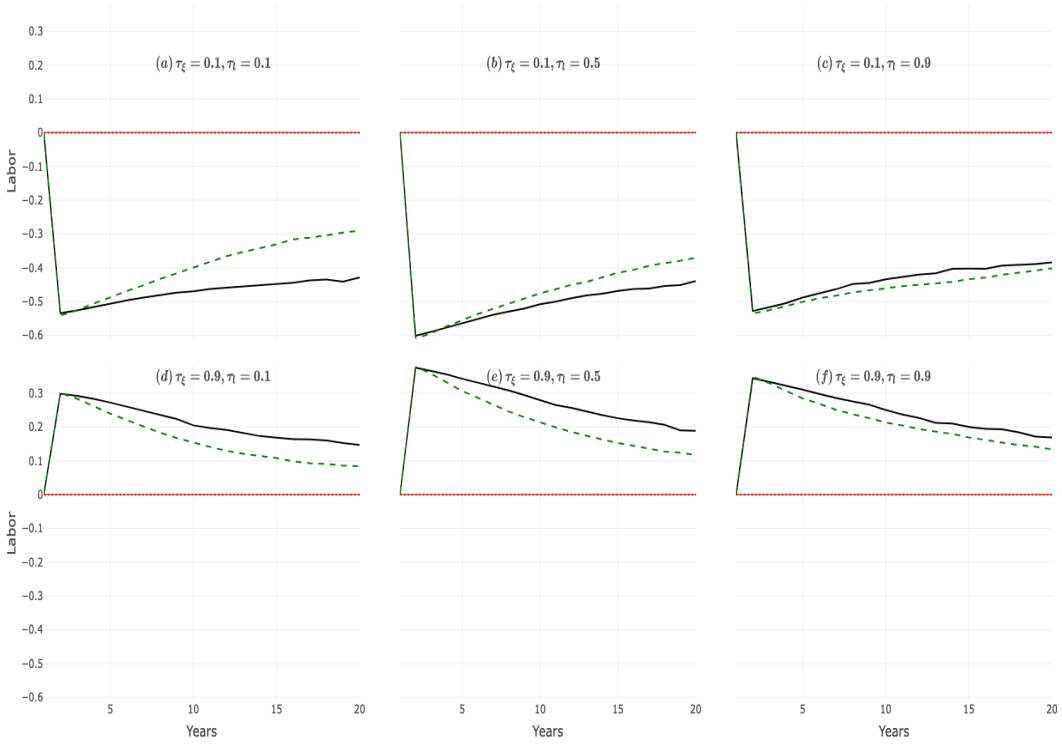
Figure 22: Impulse Response of an Innovation Shock to Capital



*Top row: Differences in capital 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: Differences in capital between firms hit with high productivity shock $\tau_\xi = 0.9$ and medium shock $\tau_\xi = 0.5$ at different levels of investment demand. Dashed green line denotes the R&D firms. The solid black line is non-R&D firms.

The labor responses to innovation shocks are reported in Figure 23, which shows the impact of a large negative shock ($\tau_\xi = 0.1$) in panel (a-c) and a large positive shock ($\tau_\xi = 0.9$) in panel (d-f) at various levels of labor demand $\tau_l = (0.1, 0.5, 0.9)$ for R&D firms (in green) and non-R&D firms (in black). These results contrast with those from the productivity analysis in Figure 20. Firms that perform R&D experience higher rates of recovery from bad productivity shocks at lower percentiles of labor demand. For the highest percentile of labor, there is not much difference in labor responses.

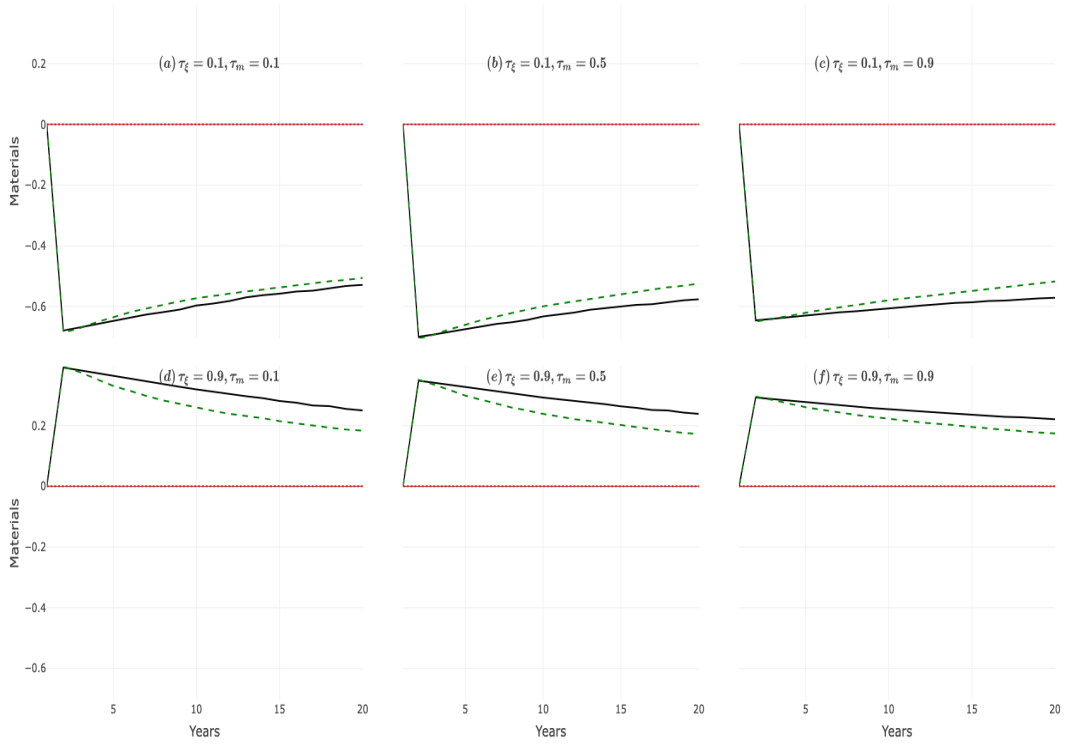
Figure 23: Impulse Response of an Innovation Shock to Labor



*Top row: Differences in labor 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: Differences in labor between firms hit with high productivity shock $\tau_\xi = 0.9$ and medium shock $\tau_\xi = 0.5$ at different levels of labor demand. Dashed green line denotes the R&D firms. The solid black line is non-R&D firms.

The materials responses to innovation shocks are reported in Figure 24, which shows the impact of a large negative shock ($\tau_\xi = 0.1$) in panel (a-c) and a large positive shock ($\tau_\xi = 0.9$) in panel (d-f) at various levels of materials demand $\tau_m = (0.1, 0.5, 0.9)$ for R&D firms (in green) and non-R&D firms (in black). The overall result is that there are not much heterogeneity in materials demand responses to productivity shocks between R&D and non-R&D firms.

Figure 24: Impulse Response of an Innovation Shock to Materials



*Top row: Differences in materials 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: Differences in materials between firms hit with high productivity shock $\tau_\xi = 0.9$ and medium shock $\tau_\xi = 0.5$ at different levels of materials demand. Dashed green line denotes the R&D firms. The solid black line is non-R&D firms.

C.4 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, which determines a threshold level of productivity for which firms will stay in operation. The decision to stay in operation or exit is given by:

$$\chi_{it} = \begin{cases} 1 & \text{if} \quad \omega_{it} \geq \underline{\omega}_t(k_{it}) \\ 0 & \text{otherwise.} \end{cases} \tag{40}$$

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 (17), the decision to stay in

operation can be written as:

$$\begin{aligned} Q_t^{\omega}(\omega_{it-1},\xi_{it}) &\geq \underline{\omega}_t(k_{it}), \\ \xi_{it} &\geq Q_t^{\omega^{-1}}(\omega_{it-1}, \underline{\omega}_t(k_{it})), \\ \xi_{it} &\geq Q_t^{\omega^{-1}}(\omega_{it-1}, k_{it}), \\ \xi_{it} &\geq \underline{\omega}_t(\omega_{it-1}, k_{it}), \end{aligned} \tag{41}$$

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 written as

$$\underline{\omega}_t(\omega_{it-1}, k_{it}) = \text{Prob}(\chi_{it} = 1|\omega_{it-1}, k_{it}) \equiv p(\omega_{it-1}, k_{it}). \tag{42}$$

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

$$\xi_{it}|(k_{it}, \omega_{it-1}, \chi_{it} = 1) \sim U(p(\omega_{it-1}, k_{it}), 1). \tag{43}$$

To see how this could be used to correct for selection bias, consider a simple linear random coefficient model for productivity: $\omega_{it} = \rho(\xi_{it})\omega_{it-1}$. The independence assumptions imply:

$$\begin{aligned} &\text{Prob}(\omega_{it} \leq \rho(\tau)\omega_{it-1}|\omega_{it-1}, k_{it}, \chi_{it} = 1) \\ &\quad = \text{Prob}(\xi_{it} \leq \tau|\omega_{it-1}, k_{it}, \chi_{it} = 1) \\ &\quad = \frac{\tau - p(\omega_{it-1}, k_{it})}{1 - p(\omega_{it-1}, k_{it})} \equiv G(\tau, p). \end{aligned} \tag{44}$$

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

$$\hat{\rho}(\tau_q)^{(s+1)} = \underset{\rho(\tau_q)}{\text{argmin}} \sum_{i=1}^{N}\sum_{t=2}^{T}\sum_{m=1}^{M} \chi_{it}\Big[G(\tau_q, p)(\omega_{it}^{(m)} - \rho\omega_{it-1}^{(m)})^{+} + (1 - G(\tau_q, p))(\omega_{it}^{(m)} - \rho\omega_{it-1}^{(m)})^{-}\Big], \tag{45}$$

where $a^+ = max(a, 0)$, $a^- = max(-a, 0)$, and $p(\omega_{it-1}, k_{it})$ can be estimated from a probit regression on $\omega_{it-1}^{(m)}$ and k_{it}. This estimator is similar to the one proposed by Arellano and Bonhomme (2017), although in my case the shift in the productivity rank is easier to characterize from the structural model used here. Implementing this selection correction is straight-forward in standard quantile regression packages. For example, in quantreg for R, this requires 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 of $\rho(\tau)$ may be biased. The amount of bias is likely to be larger at 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 in the tails. To illustrate, I use the simple AR(1) model for productivity at $\tau \leq \tau_1$:

$$\omega_{it} = \rho(\tau_1)\omega_{it-1} + v_{it} + u_{it}, \quad \omega_{it} \leq \rho(\tau_1)\omega_{it-1}, \tag{46}$$

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

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

Note that $\mathbb{E}[v_{it}|\omega_{it-1}, \chi_{it} = 1] \neq 0$ 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 propensity score $p = p(\omega_{it-1}, k_{it})$ and ω_{it-1}. Therefore, I have the following equation:

$$\mathbb{E}[\omega_{it}|\omega_{it-1}, \chi_{it} = 1] = \rho(\tau_1)\omega_{it-1} + s_1(p, \omega_{it-1}), \tag{48}$$

where $s_1(\cdot)$ 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

$$\hat{\lambda}_\rho^{-(s)} = \frac{-\sum_{n=1}^{N}\sum_{t=2}^{T}\sum_{m=1}^{M} \mathbb{1}\{\omega_t^{(m)} \leq \hat{\rho}(\tau_1)^{(s)}\omega_{t-1}^{(m)} + \hat{s}_1(p_t, \omega_{t-1}^{(m)})\}}{\sum_{n=1}^{N}\sum_{t=2}^{T}\sum_{m=1}^{M}(\omega_t^{(m)} - \hat{\rho}(\tau_1)^{(s)}\omega_{t-1}^{(m)} - \hat{s}_1(p_t, \omega_{t-1}^{(m)}))\mathbb{1}\{\omega_t^{(m)} \leq \hat{\rho}(\tau_1)^{(s)}\omega_{t-1}^{(m)} + \hat{s}_1(p_t, \omega_{t-1}^{(m)})\}}. \tag{49}$$

The algorithm proceeds similarly as before. 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

$$
\begin{aligned}
g_i(\omega_i^T; \hat{\theta}^{(s)}) = f(\omega_i^T | y_i^T, k_i^T, l_i^T, m_i^T, i_i^T, \chi_i^T; \hat{\theta}^{(s)}) \propto \\
\prod_{t=1}^{T} f(y_{it} | k_{it}, l_{it}, m_{it}, \omega_{it}, \chi_i^T; \hat{\beta}^{(s)}) f(l_{it} | k_{it}, \omega_{it}, \chi_i^T; \hat{\alpha_l}^{(s)}) f(m_{it} | k_{it}, l_{it}, \omega_{it}, \chi_i^T; \hat{\alpha}_m^{(s)}) \\
\times f(i_{it} | k_{it}, \omega_{it}, \chi_i^T; \hat{\delta}^{(s)}) \prod_{t=2}^{T} f(\omega_{it} | \omega_{it-1}, \chi_i^T; \hat{\rho}^{(s)}) p(k_{it}, \omega_{it-1}; \hat{\rho}_\chi^{(s)}) f(\omega_{i1} | k_{i1}; \hat{\rho}_{\omega_1}^{(s)}).
\end{aligned}
$$

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

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}(\tau_q)^{(s+1)} &= \underset{\boldsymbol{\beta}(\tau_q)}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{m=1}^{M} \psi_{\tau_q} \Big(y_{it} - Q_t^y(k_{it}, l_{it}, m_{it}, \omega_{it}^{(m)}; \beta(\tau_q)) \Big), \\
\hat{\boldsymbol{\alpha}}_l(\tau_q)^{(s+1)} &= \underset{\boldsymbol{\alpha}_\ell(\tau_q)}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{m=1}^{M} \psi_{\tau_q} \Big(l_{it} - \sum_{j=1}^{J} \alpha_{\ell,j}(\tau_q) \phi_{l,j}(k_{it}, \omega_{it}^{(m)}) \Big), \\
\hat{\boldsymbol{\alpha}}_m(\tau_q)^{(s+1)} &= \underset{\boldsymbol{\alpha}_m(\tau_q)}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{m=1}^{M} \psi_{\tau_q} \Big(m_{it} - \sum_{j=1}^{J} \alpha_{m,j}(\tau_q) \phi_{m,j}(k_{it}, l_{it}, \omega_{it}^{(m)}) \Big), \\
\hat{\boldsymbol{\delta}}(\tau_q)^{(s+1)} &= \underset{\boldsymbol{\delta}(\tau_q)}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{m=1}^{M} \psi_{\tau_q} \Big(i_{it} - \sum_{j=1}^{J} \delta_j(\tau_q) \phi_{\iota,j}(k_{it}, \omega_{it}^{(m)}) \Big), \\
\hat{\boldsymbol{\rho}}_\chi^{(s+1)} &= \underset{\boldsymbol{\rho}_\chi}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=2}^{T} \sum_{m=1}^{M} \Big(\chi_{it} \ln \Phi(w(k_{it}, \omega_{it-1}^{(m)}; \rho_\chi)) + (1 - \chi_{it}) \ln(1 - \Phi(w(k_{it}, \omega_{it-1}^{(m)}; \rho_\chi))) \Big), \\
\hat{\boldsymbol{\rho}}(\tau_q)^{(s+1)} &= \underset{\rho(\tau_q)}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=2}^{T} \sum_{m=1}^{M} \chi_{it} \Big(G(\tau_q, \hat{p})(\omega_{it}^{(m)} - \rho \omega_{it-1}^{(m)})^+ + (1 - G(\tau_q, \hat{p}))(\omega_{it}^{(m)} - \rho \omega_{it-1}^{(m)})^- \Big), \\
\hat{\boldsymbol{\rho}}_{\omega_1}(\tau_q)^{(s+1)} &= \underset{\boldsymbol{\rho}_{\omega_1}(\tau_q)}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{m=1}^{M} \psi_{\tau_q} \Big(\omega_{i1}^{(m)} - \sum_{j=1}^{J} \rho_{\omega_1}(\tau_q) \phi_{\omega_1,j}(k_{i1}) \Big),
\end{aligned}
$$

where $\psi_\tau(u) = (\tau - \mathbb{1}\{u < 0\})u$ is the "check" function from quantile regression. Here, ρ_χ are the parameters estimated from a probit regression of the exit decision on capital and lagged productivity from the third-to-last equation in the above M-step procedure. I approximate the function $w(k_{it}, \omega_{it-1}; \rho_\chi)$ by a second-order polynomial in k_{it} and ω_{it-1}. Kernel density estimators can also be employed to estimate the selection probabilities. The exponential

parameter for $\tau \leq \tau_1$ is updated using Equation (49) and for $\tau > \tau_Q$:

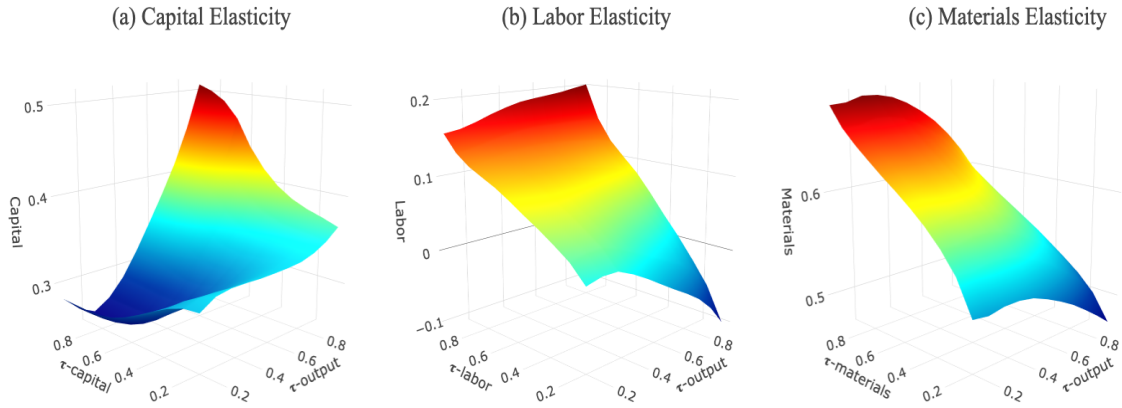
$$\hat{\lambda}_{\rho}^{+(s)} = \frac{\sum_{n=1}^{N}\sum_{t=2}^{T}\sum_{m=1}^{M} \mathbb{1}\{\omega_t^{(m)} > \hat{\rho}(\tau_Q)^{(s)}\omega_{t-1}^{(m)} + \hat{s}_2(p_t, \omega_{t-1}^{(m)})\}}{\sum_{n=1}^{N}\sum_{t=2}^{T}\sum_{m=1}^{M}(\omega_t^{(m)} - \hat{\rho}(\tau_Q)^{(s)}\omega_{t-1}^{(m)} - \hat{s}_2(p_t, \omega_{t-1}^{(m)}))\mathbb{1}\{\omega_t^{(m)} > \hat{\rho}(\tau_Q)^{(s)}\omega_{t-1}^{(m)} + \hat{s}_2(p_t, \omega_{t-1}^{(m)})\}}, \tag{50}$$

where $\hat{s}_2(\cdot)$ denotes another sample selection correction function. Selection correction methods for nonseparable 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 literatures.

To examine the extent of selection bias, I re-estimate the original model with the proposed correction. Figure 25 reports the estimates of the average capital elasticity corrected for selection. Compared to the uncorrected estimates from Figure 1, the corrected elasticities are larger for firms at the bottom percentile of capital usage and smaller for firms at the highest percentile. This seems consistent with OP's structural model of firm exit, which predicts that firms with smaller capital stock will exit at lower productivity realizations than firms with large capital stock. The downward bias is only relevant for firms who use less capital. The estimates for labor and materials are almost identical. Figure 26, which reports similar estimates evaluated over percentiles of productivity, shows that the selection correction leads to slightly smaller capital estimates. The selection corrected estimates for the non-Hicks neutral effects are reported in Figure 27. The capital efficiency effects are larger for high capital firms after correcting for selection and slightly lower for small capital firms.

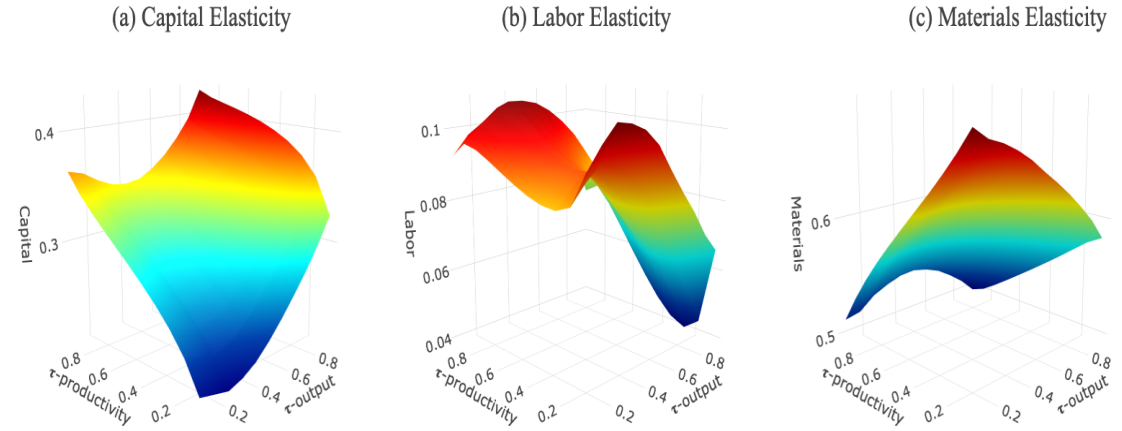
Figure 28 plots selection corrected productivity persistence over percentiles of innovation shocks and previous productivity levels. The difference between these estimates and the uncorrected estimates in Figure 4, is higher productivity persistence for low productivity-low innovation shock firms and higher persistence for low productivity-high innovation shock firms. Figure 29 reports estimates of the input responses to productivity after correcting for selection. Although there are some differences between these and the uncorrected estimates in Figure 5, there does not appear to be any noticeable relationship across the percentiles of productivity and size of input usage.

Figure 25: Output Elasticities (Selection Corrected)



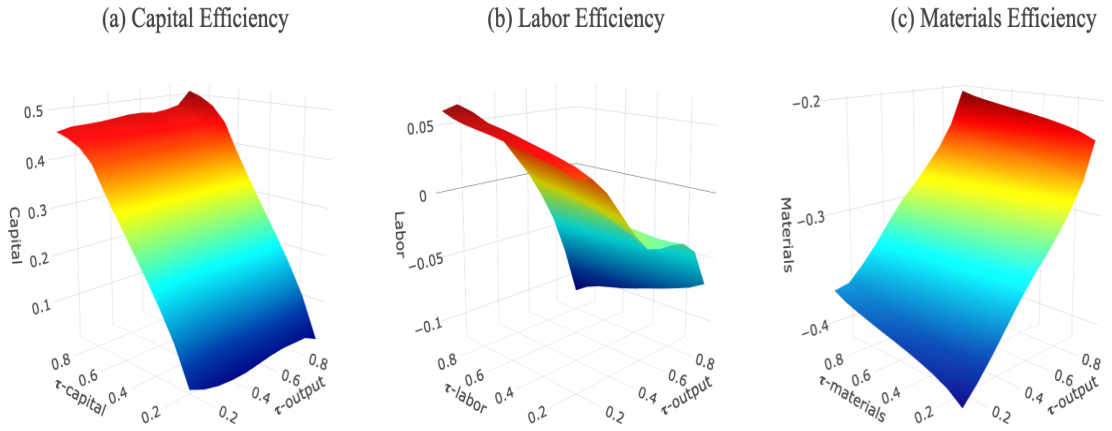
*Panel (a): Capital elasticity evaluated at τ_η and percentiles of capital τ_k averaged over values of ω_{it} and (l_{it}, m_{it}) that correspond to τ_k. Panel (b): Labor elasticity evaluated at τ_η and percentiles of labor τ_l averaged over values of ω_{it} and (k_{it}, m_{it}). Panel (c): Materials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

Figure 26: Output Elasticities (Selection Corrected)



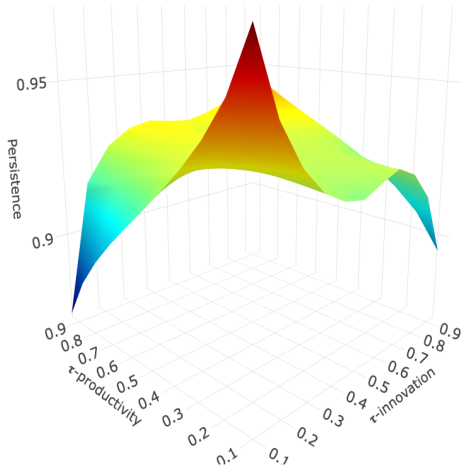
*Panel (a): Capital elasticity evaluated at τ_η and τ-productivity averaged over values of (k_{it}, l_{it}, m_{it}) that correspond to τ-productivity. Panel (b): Labor elasticity evaluated at evaluated at τ_η and τ-productivity averaged over values of (k_{it}, l_{it}, m_{it}) that correspond to τ-productivity. Panel (c): Materials elasticity evaluated at τ_η and τ-productivity averaged over values of (k_{it}, l_{it}, m_{it}) that correspond to τ-productivity.

Figure 27: Non-Hicks Neutral Effects (Selection Corrected)



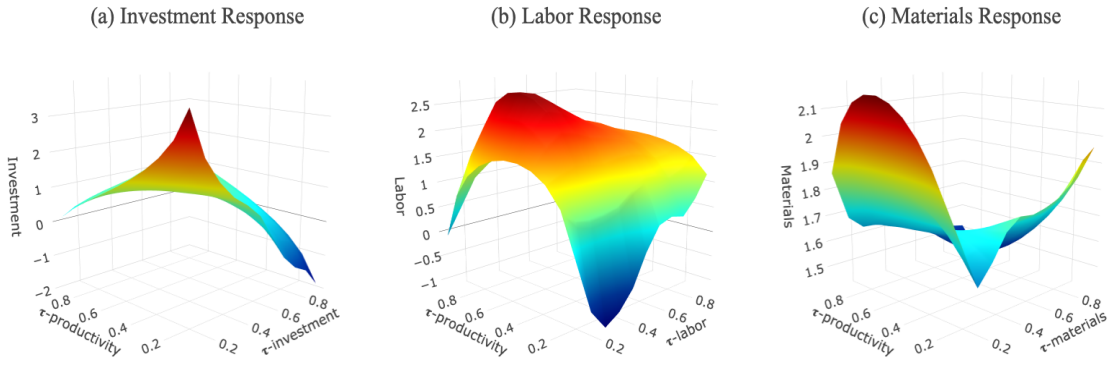
*Panel (a): Capital elasticity evaluated at τ_η and percentiles of capital τ_k averaged over values of (l_{it}, m_{it}) that correspond to τ_k. Panel (b): Labor elasticity evaluated at τ_η and percentiles of labor τ_l averaged over values of (k_{it}, m_{it}). Panel (c): Materials elasticity evaluated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

Figure 28: Productivity Persistence (Selection Corrected)



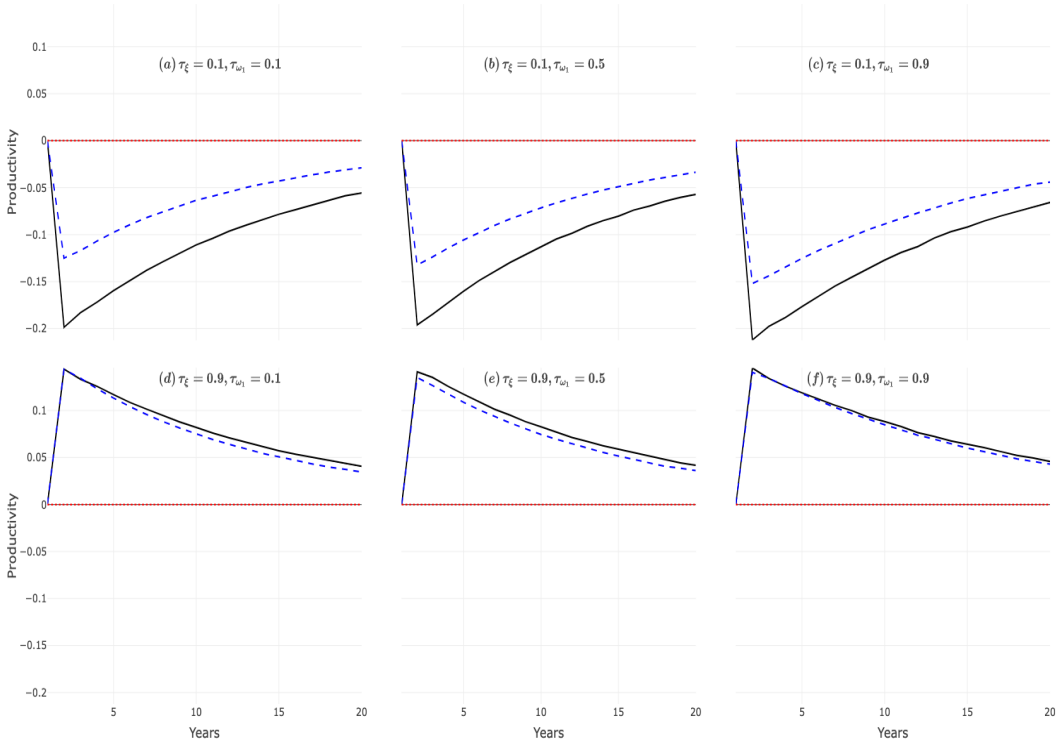
*Estimates of average productivity persistence evaluated at τ_ξ and percentiles of previous productivity.

Figure 29: Input Demand Responses to Productivity (Selection Corrected)



*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}.

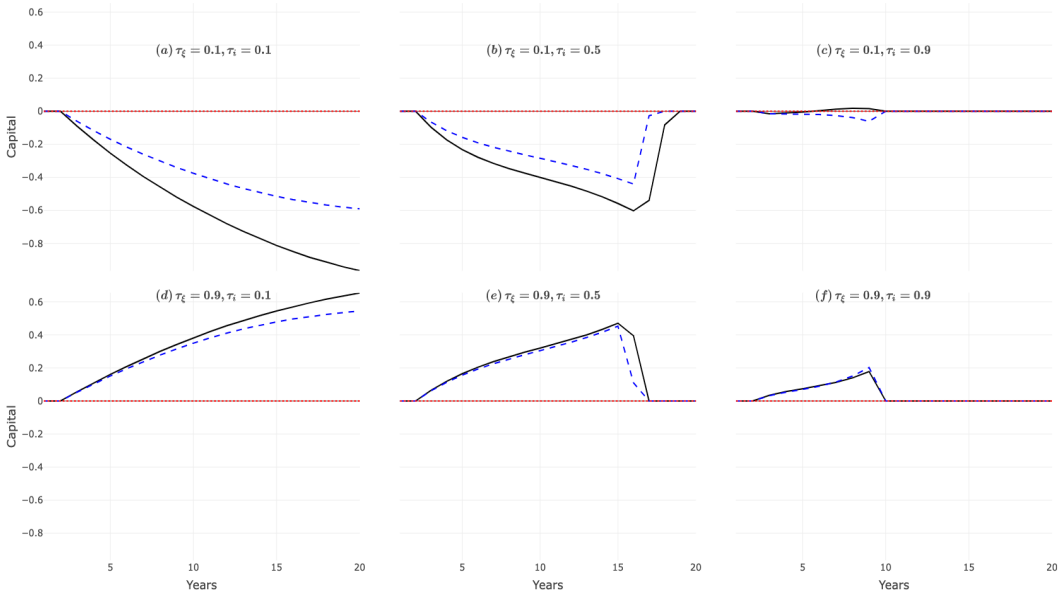
Figure 30: Impulse Response of an Innovation Shock to Productivity (Selection Corrected)



*Top row: Differences in productivity 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: Differences in productivity 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 6.

The productivity responses to innovation shocks are reported in Figure 30, which show the impact of a large negative shock ($\tau_\xi = 0.1$) in panel (a-c) and a large positive shock ($\tau_\xi = 0.9$) in panel (d-f) for various levels of initial productivity $\tau_{\omega_1} = (0.1, 0.5, 0.9)$. The solid black line denotes the estimates corrected for selection. The dashed blue line denotes estimates from the original model. Focusing on the estimates corrected for selection, firms with the lowest initial productivity, a large negative innovation shock decreases productivity by 20%, while a large positive shock increases productivity by 14%. For firms with the highest initial productivity, a large negative innovation shock decreases productivity by 21% and a large positive shock increases productivity by about 14%. The estimates correcting for selection have the most pronounced differences when firms are hit by low productivity shocks, as more firms are likely to exit at this level. After correcting for selection, productivity decreases by a larger amount. This is because conditional on staying in the market, industry productivity tends to appear higher than the unobserved distribution of productivity because of firm exits at low productivity realizations. That is, there is positive selection into staying in operation.

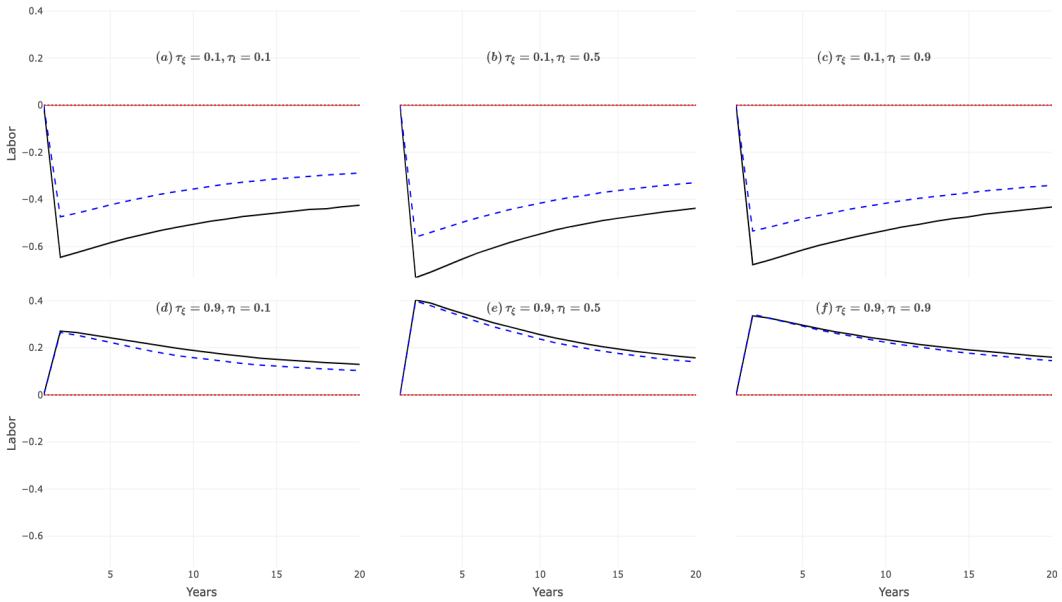
Figure 31: Impulse Response of an Innovation Shock to Capital
(Selection Corrected)



*Top row: Differences in capital 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: Differences in capital 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 7.

The capital responses to innovation shocks are reported in Figure 31, which shows the impact of a negative productivity shock in panel (a-c) and a positive productivity shock in panel (d-f) for various levels of investment demand $\tau_i = (0.1, 0.5, 0.9)$. Overall, the response functions have similar trends as the ones in Figure 7. One observation is that the differences between the estimates that correct for selection and the uncorrected estimates decreases for higher investment firms. Also, in panel (d), there are some differences between the capital responses. After correcting for selection, the capital response to productivity is higher.

Figure 32: Impulse Response of an Innovation Shock to Labor
(Selection Corrected)

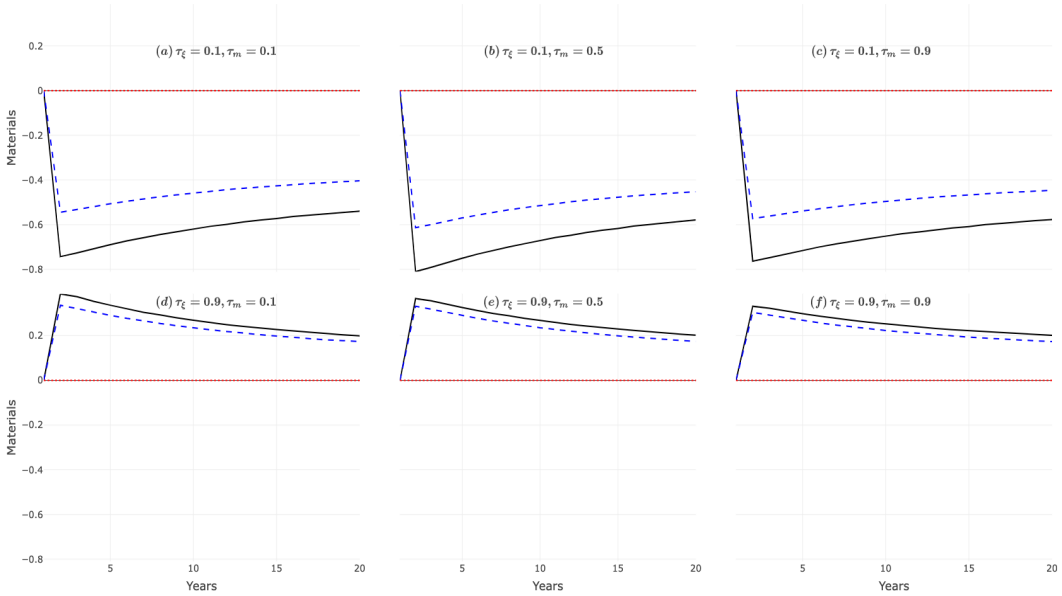


*Top row: Differences in labor between firms hit with low productivity shock $\tau_\xi = 0.1$ and medium shock $\tau_\xi = 0.5$ at different levels of labor demand. Bottom row: Differences in labor between firms hit with high productivity shock $\tau_\xi = 0.9$ and medium shock $\tau_\xi = 0.5$ at different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

The labor responses to innovation shocks are reported in Figure 32, which shows the impact of a negative productivity shock in panel (a-c) and a positive productivity shock 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 64%, while a large positive shock increases labor inputs by 26%. For firms with the highest labor demand, a large negative productivity shock decreases labor inputs by 67%, and a large positive productivity shock increases labor inputs by about 34%. Similar to the productivity estimates, there is almost no difference between the selection corrected method and the

original estimates for firms with high productivity shocks.

Figure 33: Impulse Response of an Innovation Shock to Materials
(Selection Corrected)



*Top row: Differences in materials 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: Differences in materials 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 9.

The materials responses to innovation shocks are reported in Figure 33, which shows the impact of a negative productivity shock in panel (a-c) and a positive productivity shock 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 74%, while a large positive shock increases material inputs by 38%. For firms with the highest materials demand, a large negative productivity shock decreases material inputs by 76%, and a large positive productivity shock increases material inputs by 33%. For firms hit by high productivity shocks, there are some differences between the selection corrected estimates and the original estimates, although these differences decrease with time.

In conclusion, the extensions in the Appendix suggest that heterogeneity can be pronounced due to labor adjustment frictions, R&D performance, and correcting for econometric issues such as selection bias.