Nonlinear Productivity Dynamics

José Alberto Ferreira[*] Marius Grünewald[†] David McCarthy[‡]

September 2026

Abstract

Researchers modelling firms usually describe productivity through a linear AR(1) process with Gaussian innovations. Using 15 years of Portuguese manufacturing data with firm-level prices and quantities, we document economically relevant features of the productivity distribution that no AR(1) can reproduce. First, the tails of the productivity distribution are rather wide. Second, firms at the tails face up to 3.7 times more volatility than the median firm. Third, large shocks are more frequent than predicted by an AR1 process, and some of them manage to erase a firm's accumulated history of very high (or low) productivity (*memory-wiping*). Given this, we estimate a flexible process for productivity that accommodates nonlinearities. To do so, we develop an algorithm to iteratively estimate the production function and the demand system, based on the control-function approach. We then embed it in a model of labour demand with adjustment costs. The tail dynamics brought about by the nonlinear process greatly reduce the magnitude of frictions and invert the relationship between hiring and firing costs. Our findings also stress the importance of separating technical efficiency (TFPQ) from revenue-based measures of productivity (TFPR): the two follow distinct processes and rank firms differently.

Keywords: Productivity, TFPQ, TFPR, Firm dynamics, Nonlinear panel methods.

[*]José Alberto Ferreira, European University Institute, alberto.ferreira@eui.eu

[†]Marius Grünewald, European University Institute and European Central Bank, marius.gruenewald@eui.eu. The views expressed are those of the authors and do not necessarily reflect those of the European Central Bank or the Eurosystem.

[‡]David McCarthy, European University Institute, david.mccarthy@eui.eu

We are indebted to Russell Cooper, Giancarlo Corsetti, Tom Crossley, and Jesús Bueren for invaluable guidance and support throughout. We also thank Micole de Vera, Siyu Shi, Jan De Loecker, Andrea Caggese, Luca Dedola, and Gonzalo Paz-Pardo for thoughtful comments and useful discussions. We are also grateful to seminar participants at UPF-CREI, DG-Research (ECB), and the EUI. All errors are our own. The authors declare that they have no relevant or material financial interests that relate to the research described in this paper. No funding was received for conducting this study.

1 Introduction

Firm-level productivity is key to understanding firm behaviour, business dynamism and macro aggregates. While broadly accepted as a dynamic object, it is however less clear *how* exactly productivity evolves over time. To address this, our paper poses three questions: (i) what does the firm-level productivity process look like, in the data?, (ii) how far off is it from a linear-Gaussian AR(1) benchmark?, and (iii) what does it mean for canonical models of firm dynamics?

When faced with modelling productivity, quantitative researchers commonly resort to a linear process with Gaussian innovations - a tractable but, we argue, not harmless avenue. It implies, among other things, that firms face the same uncertainty and are equally exposed to extreme shocks regardless of where they sit in the productivity distribution. It also means that those extreme shocks are very rare (thin tails) and that persistence of productivity is constant. As such, *memory-wiping events*, where a history of good (bad) productivity is reversed, do not occur. These features - usually assumed, not tested - carry over to how firms make decisions over time, be it investment, employment or pricing. In this work, we take a step back and test such features, drawing from a long panel of Portuguese manufacturing firms, in a diverse set of industries, where we observe prices and quantities for 15 years.

We show that productivity, when defined and measured as technical efficiency, has richer dynamics than what a linear-Gaussian AR(1) would entail. That richness is best understood through the contrast between the body and the tails of the productivity distribution. On the one hand, we find that most firms live in a narrower range of productivity than what the benchmark would entail (i.e. the gap between the top and bottom 10% of firms is smaller than conventionally assumed); in that range, productivity is fairly stable, remaining anchored to the inherited level regardless of the sign and size of the innovation (broadly unchanged persistence). On the other hand, the tails of the productivity distribution are much more spread out, and extreme events *do occur*. In fact, the most and least productive firms (the tails) are subject to volatility up to 3.7 times greater than those in the middle. Within this turbulence at the tails, we find evidence of *memory-wiping* shocks: very adverse (positive) shocks that make the firm carry over little of its past record of very high (low) productivity - as if *wiping out* part of the firm's history.

Implications for models of firm dynamics. To capture such dynamics - with fat tails, state-dependent higher-order moments, and memory-wiping shocks -, we borrow a process from the literature on household earnings: the conditional quantile framework by Arellano and Bonhomme (2016). This keeps the first-order Markovian nature of the productivity process and can be discretised for the purposes of quantitative exercises (with a larger and finer grid). After estimating such a nonlinear process in our sample, we study two economies - one facing a textbook AR(1) process for productivity, and the other facing our nonlinear process - in a standard partial equilibrium model of labour demand. Three results from this exercise are particularly revealing. Firstly, the nonlinear process induces significant inaction even absent any form of adjustment cost. Secondly,

once we bring in adjustment costs and estimate them to match a common set of first-order moments of the employment growth distribution - which both economies attain in a similar fashion -, the estimated adjustment costs majorly disagree between the two economies: convex adjustment costs under the nonlinear process are 23% of their AR(1) counterpart, fixed costs 8.5%, and the ranking of hiring and firing costs reverses, from firing costing 1.9 times more than hiring to costing a third of it. Thirdly, the economy with nonlinear productivity and (much) lower adjustment costs is able to match a broad set of untargeted higher-order moments of the employment growth distribution, generating the rare - but relevant - growth bursts and major contractions (i.e. growth rates above 50% in absolute value) that characterise business dynamism (Decker et al., 2016).

TFPQ, not TFPR. Our results pertain to productivity defined as TFPQ (Z_{it}^{Q}), in the spirit of Foster et al. (2008): a technical efficiency primitive, defined as the units of output a firm produces over its bundle of inputs. Revenue productivity[1], namely TFPR (Z_{it}^{R}), is a different object. Since revenue is price times quantity, the two are related by

$$Z_{it}^{R} = P_{it} \cdot Z_{it}^{Q}, \tag{1}$$

so TFPR is technical efficiency scaled by the price the firm charges, which in turn is set against the demand curve the firm faces - specific to the market structure at hand. Importantly, the two are not, in general, monotone transformations of one another: a firm with lower Z^{Q} ranks above a rival in Z^{R} whenever its demand appeal is large enough to charge a higher price. In fact, the two rank firms alike only loosely: their contemporaneous correlation is 0.49, firms around the median of TFPQ are spread almost evenly across the TFPR distribution, and even at the extremes, where the ranks agree most, the agreement is partial — 72% of the least efficient firms fall in the bottom quarter of TFPR and 80% of the most efficient in the top quarter. This suggests that pricing outcomes, regardless of their source, can either mask or strengthen the firm's technical efficiency. TFPQ and TFPR should thus be read as the distinct economic objects that they are. Any 'productivity' analysis based on TFPR — its spread, its persistence, its tails — mixes, by construction, technology and pricing, with no way of discerning the role of each. In this regard, ours is the first work to study the dynamics of TFPQ with nonlinear panel methods.

Measuring TFPQ: a novel algorithm. Productivity is never observed: it is a residual from an estimated production function. Hence, we set out to estimate *output* (not revenue) *elasticities*, such that our residual (TFPQ) does not carry contamination from demand-side variables. While the literature has offered some ways of doing so, we build on the existing control function approach and change it in three meaningful ways. First, by controlling for a latent demand advantage in the first-stage of the procedure; the rationale is simple: if input choices respond to some unobserved dimension of demand, as well as to productivity, then a control function that does not condition

[1]The most commonly used metric of revenue productivity is, perhaps, the residual of a revenue function (that we term REVRES) - often used when quantities are not observed; it takes as output nominal revenue deflated by an industry price index, returning *revenue elasticities*. In general, this residual does not equal TFPR. Further insights on how different productivity metrics relate can be found in Appendix A.

on that latent demand - that we term *demand advantages* - is bound to return a residual mixing the two. Second, we recover those demand advantages from a demand system (featuring imperfect competition) where prices are instrumented with the TFPQ estimates from the production function; because the identification of both demand and production coefficients depends on each other's residuals, we estimate them jointly, iterating until the estimates converge. Third, within the production function estimation block, we model TFPQ as a flexible higher-order polynomial, instead of imposing ex ante an AR(1) formulation, as is usual in the literature.

The remainder of the paper is organised as follows: Section 1.1 revises the literature. Section 2 describes the dataset used and Section 3 provides an overview of the productivity estimation procedure. Section 4 puts forward a series of stylised facts on TFPQ, explaining in what way TFPQ deviates from the linear-Gaussian benchmark, while in Section 5 we delve into the relationship between TFPQ and TFPR. In Section 6, we bring the nonlinear productivity model estimated from our data into a standard model of labour demand and assess its implications vis-à-vis the standard AR(1) assumption. In Section 7, we conclude with an overview of findings and a prelude to future iterations of this project.

1.1 Related Literature

Productivity differences across firms are large and persistent (Syverson, 2011), and how they evolve shapes what firms do. A firm that expects a good draw to persist invests and hires against it, while one that expects it to fade does not, and when adjustment is costly the uncertainty about that path itself delays action (Cooper and Haltiwanger, 2006; Bloom, 2009). Constrained firms accumulate internal funds in response to persistent productivity gains, so persistence also governs how quickly they escape financial constraints (Banerjee and Moll, 2010; Midrigan and Xu, 2014). Whether firms face the same risk regardless of where they sit in the productivity distribution, and whether a history of good draws can be undone by a single shock, are therefore questions with economic content. Until recently, however, the process was almost always taken to be a linear AR(1) with Gaussian innovations.

Jaimovich et al. (2023) and Salgado et al. (2019) work with revenue-based productivity and consistently find that a Gaussian AR(1) does not match its observed dynamics: the former document fat tails, the latter skewness that varies systematically over the business cycle. Bloom et al. (2018) depart from the linear-Gaussian benchmark only through time-varying volatility, leaving persistence linear. Melcangi and Sarpietro (2024), Ruiz-Garcıa (2021), and Fella et al. (2022) apply the quantile-based estimator of Arellano et al. (2017) to firm-level revenue-based productivity and estimate flexible first-order Markov processes with nonlinear persistence and non-normal shocks, which fit micro-data significantly better than the canonical AR(1).

All of this evidence is on revenue productivity. TFPR is a residual of a revenue function, so it conflates technical efficiency (TFPQ) with the price the firm charges, and hence with demand

(Foster et al., 2008; Haltiwanger, 2016). Klette and Griliches (1996) show that, when output prices are unobserved and demand is imperfectly elastic, deflating revenue with an industry price index yields inconsistent production function estimates and a residual that absorbs firm-level price variation. Separating the two requires firm-level prices and quantities (De Loecker, 2011). Cooper and Ozturk (2022) show theoretically that endogenous pricing drives a wedge between the cyclical properties of TFPR and TFPQ. The distinction is not merely conceptual: in the misallocation literature following Hsieh and Klenow (2009), dispersion in revenue productivity is read as evidence of distortions in the allocation of resources across firms, so when the object in hand is a revenue residual rather than a structurally defined TFPR, measured dispersion may be recording heterogeneity in demand conditions instead. We recover TFPQ and demand separately and ask which nonlinearities in TFPR originate in TFPQ.

Methodologically, our work builds on the "control function" approach: Olley and Pakes (1996) use investment to proxy for unobserved productivity, Levinsohn and Petrin (2003) use intermediate inputs, and Ackerberg et al. (2015) show that the variable-input coefficient is not identified in the first stage and move identification of all production parameters to the second stage. Within this tradition, De Loecker (2011) and De Loecker et al. (2016) expanded the control function approach to account for unobserved price variation by introducing observable demand shifters into the proxy equation to condition out both productivity and demand shocks. An alternative route, the dynamic panel estimator of Blundell and Bond (2000), obtains identification by imposing a linear AR(1) on productivity, which rules out from the start the nonlinearities we want to estimate. We propose a joint estimation of demand and TFPQ via an iterative procedure. Our work is most closely related to Eslava et al. (2024), who use price and quantity information to recover establishment-level technical efficiency and quality/appeal, estimating all structural parameters simultaneously by GMM. We instead iterate between the production function and the demand system until the estimates converge, so that the two remain mutually consistent throughout the control-function estimation. Finally, De Vera (2022) also estimates production and demand jointly, on Portuguese data, but to measure the pass-through of productivity and demand shocks to wages. Our object is the productivity process itself.

Like Foster et al. (2008), Eslava et al. (2013), and Eslava et al. (2024), we use firm-level prices and quantities to measure TFPQ. They study its role in firm entry, exit, and growth, whereas we document the dynamics of TFPQ itself, how they map into TFPR, and what they imply for firms' hiring decisions. Our data also separate quantities produced from quantities sold, which matters once inventories are a margin the firm uses: the production function is a statement about output and the demand system about sales, and the two coincide only period by period if stocks never move. On the latter, Cooper and Haltiwanger (2006) and Cooper et al. (2015) estimate capital and labour adjustment costs by simulated method of moments on plant-level data, and Cooper et al. (2024) use the same approach to attribute the decline in establishment-level responsiveness documented by Decker et al. (2020) to rising employment adjustment costs, all under a linear AR(1) for productivity. We estimate a dynamic labour demand model under our nonlinear process and under

an AR(1) calibrated to match its linear moments, and compare the implied adjustment costs and distribution of employment growth.

2 Data

2.1 Sources and dataset construction

Studying the dynamics of productivity, namely TFPQ, requires a long panel of firm data with information on outputs, inputs, and prices. We use a set of Portuguese administrative and survey datasets, spanning for 15 years (2008-2022), provided by Statistics Portugal (INE):

1. **Administrative data: SCIE and QP (annual)** - (*Sistema de Contas Integradas das Empresas* and *Quadros de Pessoal*). SCIE has universal coverage of private non-financial firms operating in Portugal from 2004 onwards (retrieved from mandatory tax filings), while QP has near-universal coverage of Portuguese workers and their linked employers since the 1980s, for firms with at least 1 employee [2]. We use SCIE to obtain information on firm-level inputs (capital, labour, and intermediate consumption) and QP to obtain information on firm characteristics (age, number of employees, NACE sector).

2. **Survey data: PRODCOM/IAPI (annual).** In the *Inquérito Anual à Produção Industrial* (IAPI) - the Portuguese version of Eurostat's PRODCOM survey -, a sample of manufacturing firms report the prices and quantities of output produced and sold. Data is collected at the firm (not plant) level. The survey design has both an exhaustive component - every firm with turnover above €10 million -, and a non-exhaustive component built by ranking all remaining enterprises by turnover, from largest to smallest, and adding them in descending order until a coverage target is met[3].

The inner merge of these data sources is our main dataset, after some cleaning steps described in Section B.4. It is a long panel of Portuguese manufacturers, between 2008 and 2022, with 88,386 firm-year observations on 11,398 firms over 23 products (which we will term industries).

Products as industries. We estimate production functions at the *industry* level, which we define as the 6-digit PRODCOM product code paired with a consistent unit of measurement of output[4]; 23 industries result from our cleaning proccedure, meeting demands for data quality as well as size and length of the panel. The five largest industries are described in Table B.3.

[2]The QP data refer to a reference week in October of each year, while all other source datasets refer to the end of the fiscal year, which in Portugal coincides with the calendar year for the vast majority of firms.

[3]The target itself is adaptive, and is set to ensure that roughly $80-90\%$ of each 2-digit NACE revenue is covered. Also note that the survey design underwent a reform in 2013 towards an smaller but more revenue-relevant sample (i.e. biased towards firms above 20 employees).

[4]In choosing the level at which to estimate the production function we face a tradeoff between product homogeneity and sample size. On the one hand, a too broadly defined industry can render uninformative estimates of the factor

Entry and exit. Due to survey design, larger firms are over-represented in our sample. This ensures a fairly stable share of manufacturing's value added being captured every year - around 25% - at the expense of not observing entry and exit margin, which the survey rarely captures [5]. Indeed, we find only 1,816 firm births throughout in our sample. Our analysis therefore characterizes best the productivity dynamics of incumbent firms. Because of its extensive time coverage and overall scale, our dataset enables an investigation of TFPQ that is seldom possible in the existing literature.

Time concordances. To ensure the consistency of our dataset over time, we carefully build time concordance routines that ensure a harmonisation over time of: (i) PCC8 product codes, that change annually; (ii) CAE sectors (the Portuguese NACE classification, *Classificação de Atividades Económicas*), which we harmonise to a single revision across sources; (iii) accounting reporting standards (with a major reform of reporting standards in 2008[6]).

Multiproduct firms. To obtain a single firm-year entry, we reduce every firm to its main product sold (6-digit level), i.e. the one with the highest share in revenue, and keep the firm-year only if that product accounts for at least 90% of the firm's revenue. This filtering is not particularly restrictive in our case [7] In the final sample, the main product accounts for a revenue share of 0.999 for the average firm, and 1.000 for the median firm. Nevertheless, we follow De Loecker et al. (2016) and apportion inputs to the main product sold according to the firm-year specific revenue share of that product.

Multi-establishment firms. In contrast to other work measuring TFPQ, we do not observe quantities and prices at the establishment (plant) level. Our prices and quantities should therefore be seen as averages across the plants. However, about 91% of our observations are from single-establishment firms, and they generate 72% of the revenue in the sample.[8]

2.2 Descriptive Statistics

Our sample is comprised of 88,386 firm-year observations, between 2008 and 2022. These are predominantly large and established manufacturing firms, with 22 years of age on average and

elasticities (as some average elasticity across multiple, potentially very different, technologies). On the other hand, too narrowly defined industries will cause us to estimate the elasticities with a potentially too small sample, thus imprecise. While the former compels us to resort to highly specific product codes (e.g. Foster et al. (2008) use eleven 7-digit-SIC- defined *products*, highly homogenous), the latter prompts us to consider broader definitions (e.g. Eslava et al. (2013) consider 4-digit ISIC *sector* codes). To address this tension between reducing technology/quality variation and granting sufficient sample size, given the survey nature of our data, we take the 6-digit route.

[5]While our dataset is an unbalanced panel, most of the attrition comes from either the survey redesign of 2012 or from the rank-based non-exhaustive component of the survey

[6]Namely, the transition from the *Plano Oficial de Contas* to the *Sistema de Normalização Contabilística*. For instance, the definition of fixed tangible assets changed, and so did the handling of depreciation flows. We were guided by the Bank of Portugal's harmonization practices for their BPLIM's dataset *Central Harmonizada de Balanços*

[7]Most firms in the survey are close to single-product. Across all IAPI firm-years, the main product accounts for a revenue share of 0.905 on average and 1.000 at the median, and 72.7% of firm-years, covering 63.4% of surveyed revenue, pass the threshold.

[8]Establishment data comes from the *Quadros de Pessoal* - it only exists for headcount and revenue (not at the product level).

about 20 employees (versus 10 in the population of manufacturers). Key summary statistics of our sample are reported in Table 1. Descriptives of the universe of the Portuguese manufacturing sector can be found in the appendix - Table B.1.

TABLE 1. Summary statistics for the estimation sample.

	Count	Min	P_{25}	P_{50}	P_{75}	Max	Mean	Std. Dev.
Age	66519	0.00	12.00	20.00	29.00	182.00	21.87	14.23
Number of Employees	88386	1.00	4.00	8.00	19.00	1495.00	20.18	46.17
Turnover (in Thd.)	88386	0.00	139.21	337.39	1041.06	284928.78	1576.20	6042.32
Capital (in Thd.)	88386	0.00	20.97	91.22	390.51	289603.08	884.81	5557.56

Notes. Firm-year observations in the estimation sample, 2008–2022, after the merge with the PRODCOM survey and all sample restrictions described in Section B.4. Turnover and capital are deflated and expressed in thousands of euros.

Consistent with the IAPI/PRODCOM survey design, our sample is skewed to large firms. All metrics of firm size - headcount, value added, and total assets - exhibit a larger mean than median. Note, however, that the difference between the two values is only a fraction of the standard deviation. A firm is observed for an average of 7.8 years.

Our sample accounts for about one-fifth of Portuguese manufacturing value-added and employment. In broader terms, it reflects 1.38% of total employment and and 2.72% of GDP. The yearly breakdown is available in Table B.2, in the appendix; the post-2012 increase in coverage is, as mentioned before, due to the survey redesign, which increased the number of large firms surveyed.

3 Estimation strategy

We estimate the production function via a control function (Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2015), where demand-related variables are allowed to impact input demand. As De Loecker (2011), one of those variables is market share - observed, yet only a valid demand control under constant elasticity of demand. Since we adopt a more generic demand formulation and let the elasticity vary with firm size, in the tradition of Atkeson and Burstein (2008), we add a new control variable for input deman: a latent demand advantage, recovered from a demand system estimation. Relative to Foster et al. (2008), cost-share elasticities are thus replaced by a control function and their CES demand by a variable-markup one (that nests the CES case).

Each block, which we first present separately, requires the other's output, so we estimate them by alternating between the two until the estimates converge: GMM updates of the demand system, instrumented by the current production function residual, alternate with re-estimation of the production function conditioning on the lagged demand advantage, until the production

elasticities cease to move (Sections 3.1–3.3). In turn, the instrument identifying demand is the residual of a production function that itself conditions on demand. Our iterative approach keeps every step a somewhat standard estimator and makes convergence an observable property rather than an assumption.

Notation. Firms are indexed by i, industries by s and years by t; lowercase letters denote logs. Two latent states drive the model. Physical productivity z^q_{ist}, or TFPQ, is the Hicks-neutral efficiency with which the firm converts inputs — labour l_{ist}, capital k_{ist} and materials m_{ist} — into physical units of output q_{ist}. The demand advantage D_{ist} is the firm-specific shifter of the quantity it can sell at a given price p_{ist}, and stands for quality, appeal or reputation; it is latent and only partly summarised by the observable market share ms_{ist}, the firm's share of industry-year revenue. We write ξ^q_{ist} and ξ^d_{ist} for the structural innovations to the two states, and Ω_{ist} for the information set on which the firm conditions its period-t input choices.

Assumptions. Identification of both blocks rests on the following, which adapt the framework of Olley and Pakes (1996) and Ackerberg et al. (2015) to a setting with demand advantages.

Assumption 1 (Information set)**.** *The information set of firm i at the beginning of period t, denoted Ω_{ist}, includes its current and past productivity and its past demand advantages:*

$$\Omega_{ist} := \{k_{ist},\, l_{ist},\, \{z^q_{is\tau}\}_{\tau\le t},\, \{D_{is\tau}\}_{\tau\le t-1}\}\,.$$

Labour l_{ist} belongs to Ω_{ist} because it is chosen at $t-b$, with $b\in[0,1]$ (Assumption 3).

Assumption 2 (First-order Markov dynamics)**.** *Productivity follows a first-order Markov process, left otherwise unrestricted:*

$$p\left(z^q_{ist+1} \mid \Omega_{ist}\right) = p\left(z^q_{ist+1} \mid z^q_{ist}\right).$$

The demand advantage is likewise first-order Markov, $p(D_{ist+1} \mid \Omega_{ist}) = p(D_{ist+1} \mid D_{ist})$, but take the form:

$$\mathbb{E}\left[D_{ist} \mid D_{ist-1}\right] = \rho_{0,s} + \rho_{1,s} D_{ist-1} + \rho_{2,s} D^2_{ist-1},$$

We reserve the richness in dynamics to TFPQ, and keep demand more tractable - the demand coefficients coefficients enter the demand moment conditions directly while the ones on the productivity process will be concentrated out (Section 3.2), and we only estimate the TFPQ process ex-post. The structural innovations to TFPQ (ξ^q_{ist}) and demand advantages (ξ^d_{ist}) are mutually independent; this is what allows physical productivity to serve as a supply-side instrument for prices in the demand block. The consequences to our algorithm of breaking this independence are explored in Appendix D.

Assumption 3 (Timing of input choices)**.** *Capital k_{ist} is predetermined: it is chosen at $t-1$, subject to time-to-build, and is therefore orthogonal to the innovation in productivity. Labour l_{ist} is chosen at $t-b$, with $b\in[0,1]$, and may thus be set within period t, after part of the productivity innovation has been observed, as in Ackerberg et al. (2015). Productivity z^q_{ist} is realized at the beginning of period t. Materials m_{ist} are a free variable, chosen within period t after z^q_{ist} is observed. The*

demand advantage D_{ist} is realized only at the end of period t, so that input choices in period t can condition on D_{ist-1} at most.

Assumption 4 (Materials demand and strict monotonicity)**.** *Materials demand takes the form $m_{ist} = m_{st}\left(z^q_{ist}, k_{ist}, l_{ist}, D_{ist-1}, ms_{ist}\right)$, where ms_{ist} is the firm's market share, and is strictly increasing in z^q_{ist} conditional on its remaining arguments.*

Assumptions 1–3 correspond to the information-set, first-order Markov, and timing conditions of Ackerberg et al. (2015). The timing of capital follows the standard time-to-build argument, while the timing of labour nests the cases discussed by Ackerberg et al. (2015) and De Loecker and Warzynski (2012).[9] Assumption 4 is the scalar-unobservable and strict-monotonicity condition of Ackerberg et al. (2015), with the lagged demand advantage and the market share as additional state variables.

3.1 Production Block

For firm i in industry s and year t, we consider a log gross-output Cobb-Douglas production function,

$$q_{ist} = \beta^0_s + \beta^L_s l_{ist} + \beta^K_s k_{ist} + \beta^M_s m_{ist} + z^q_{ist} + \epsilon_{ist},$$

where ϵ_{ist} stands for classical measurement error. Our primary object of interest is z^q_{ist}.

We estimate the production function using the *control function approach*, but departing from it in two ways: (i) the inversion of materials demand conditions on the firm's (latent) lagged demand advantage as well as on the other inputs, and (ii)productivity follows a nonlinear first-order Markov process rather than a linear AR(1), so that persistence can vary with the productivity level.

Implementation. Estimation proceeds in two stages, industry by industry; Appendix C gives the details. The first stage projects output on a flexible function of inputs and demand controls,

$$q_{ist} = \Phi_s\left(l_{ist}, k_{ist}, m_{ist}, \tilde{d}_{ist-1}, ms_{ist}\right) + \delta_{st} + \delta_{rs} + \epsilon_{ist},$$

where $\tilde{d}_{ist-1}$ is the lagged transient demand advantage of Section 3.2 and ms_{ist} the firm's market share. This purges measurement error but does not identify β, which is subsumed in Φ_s; we denote the fitted values $\widehat{\Phi}_{ist}$. The second stage recovers the elasticities from the dynamics of productivity. For a candidate β, trial productivity is $\hat{z}^q_{ist}(\beta) = \widehat{\Phi}_{ist} - \beta^0_s - \beta^L_s l_{ist} - \beta^K_s k_{ist} - \beta^M_s m_{ist}$, which we project on a cubic polynomial in its own lag,

$$\hat{z}^q_{ist} = \tilde{g}\left(\hat{z}^q_{ist-1}\right) + \tilde{\xi}_{ist}(\beta), \qquad \mathbb{E}\left[\tilde{\xi}_{ist} \mid \hat{z}^q_{ist-1}\right] = 0 \text{ by construction}, \tag{2}$$

and choose β to satisfy $\mathbb{E}[\tilde{\xi}_{ist}(\beta)\, Z_{ist}] = 0$, where Z_{ist} collects capital and lagged flexible inputs.

[9] Whenever $b < 1$, labour is set within period t and may be correlated with the productivity innovation, so the second-stage instruments use lagged labour; Appendix C gives the full instrument vector.

Assumption 2 leaves productivity unrestricted while (2) projects on a cubic, so $\tilde{\xi}_{ist}$ is not the structural TFPQ innovation and a *misspecification gap* generally remains. Identification does not require that gap to vanish, only that it be uncorrelated with the instruments, which are dated $t-1$ or earlier. Everything that could mark a nonlinear process — heteroskedastic loadings, skewed innovations, persistence varying with the size of the shock — enters $\tilde{\xi}_{ist}$ through higher moments that the moment condition never uses, and therefore costs efficiency rather than consistency. Section C.1.2 elaborates this argument and the following Section 3.3.1 reports the behaviour of our algorithm under different extents of this misspecification.

Variables. Output q_{ist} is the log quantity produced of the firm's main product, in the standard unit of its industry (Section 2). Labour l_{ist} is the firm's headcount. Capital k_{ist} is the book value of fixed tangible assets, deflated by a capital-goods price deflator. Materials m_{ist} are raw-material expenditures, deflated by intermediate-input price deflators; both input deflators vary by NACE industry and year. All inputs are apportioned according to the main product's share in the firm's revenue. The price $\hat{p}_{ist}$ is a unit value: the revenue from the main product divided by the physical quantity sold, in the same standard unit, deflated by the industry-year price index [10]

Since our focus lies in the dynamics of idiosyncratic objects, the (log) TFPQ measure used henceforth is $\hat{z}^q_{ist}$, i.e. z^q_{ist} demeaned of industry-year fixed effects. With $\hat{p}_{ist}$ being the log relative price defined above, we build (log) TFPR as $\hat{z}^r_{ist} = \hat{z}^q_{ist} + \hat{p}_{ist}$. For convenience, in the remainder of the paper refer to $\hat{z}^r_{ist}$, $\hat{z}^q_{ist}$ and $\hat{p}_{ist}$ simply as TFPR, TFPQ and prices, respectively.

3.2 Demand Block

To recover the latent demand advantages D_{ist}, we estimate for each industry

$$y_{ist} = \alpha_{p,s}\, p_{ist} + \alpha_{pm,s}\, p_{ist} \ln MS_{ist} + D_{ist} + \epsilon^d_{ist}, \tag{3}$$

where y_{ist} is the log quantity sold, MS_{ist} the firm's share of industry-year revenue and ϵ^d_{ist} i.i.d. demand noise. The price elasticity, $\alpha_{p,s} + \alpha_{pm,s} \ln MS_{ist}$, varies with market share and nests CES at $\alpha_{pm,s} = 0$. The demand advantage follows the quadratic Markov process of Assumption 2.

We estimate $\theta_s = \{\alpha_{p,s}, \alpha_{pm,s}, \rho_{0,s}, \rho_{1,s}, \rho_{2,s}\}$ by GMM. Price is endogenous, since the firm sets it knowing ξ^d_{ist}; physical productivity shifts marginal cost and hence price, and is independent of the demand innovation by Assumption 2. As instrument we use physical productivity itself, $\hat{z}^q_{ist}$, evaluated at the current elasticities and demeaned by year within industry.[11] Assumptions 1 and 2 additionally justify lagged variables, giving

$$\Xi_{ist} = \{1,\ \hat{z}^q_{ist},\ \hat{z}^q_{ist} \ln MS_{ist},\ y_{is,t-1},\ p_{is,t-1},\ y_{is,t-2},\ p_{is,t-2},\ k_{is,t-1},\ k_{is,t-2}\}, \tag{4}$$

[10] We compute this as the revenue-weighted geometric mean of the unit values of *all surveyed producers* in the industry.

[11] In practice $\hat{z}^q_{ist}$ is recomputed here from log output rather than taken from the first-stage fit $\widehat{\Phi}_{ist}$, which would carry the fitted demand controls into the instrument.

where the interaction with market share identifies $\alpha_{pm,s}$. Substituting the demand equation into the law of motion yields

$$\mathbb{E}\left[\left(\tilde{y}_{ist} - \rho_{0,s} - \rho_{1,s}\tilde{y}_{is,t-1} - \rho_{2,s}\tilde{y}^2_{is,t-1}\right)\Xi_{ist}\right] = 0, \qquad \tilde{y}_{ist} \equiv y_{ist} - \alpha_{p,s}p_{ist} - \alpha_{pm,s}p_{ist}\ln MS_{ist}, \quad (5)$$

with the nine moments weighted equally and $\alpha_{p,s}$ bounded above at -0.05 to rule out upward-sloping demand. Section C.2 states the moment conditions in full and gives the numerical detail.

From the GMM, we obtain the demand elasticities and recover the latent demand advantages as $\hat{D}_{ist} = y_{ist} - \hat{\alpha}_{p,s}p_{ist} - \hat{\alpha}_{pm,s}p_{ist}\ln MS_{ist}$. We then split $\hat{D}_{ist}$ into its firm mean $\hat{a}_{is}$ and end up with a transient component of firm-demand $\tilde{d}_{ist} = \hat{D}_{ist} - \hat{a}_{is}$. We estimate $(\rho_{0,s}, \rho_{1,s}, \rho_{2,s})$ by OLS of $\tilde{d}_{ist}$ on $\tilde{d}_{ist-1}$ and $\tilde{d}^2_{ist-1}$, seeking to remove firm fixed effects. The lagged transient component $\tilde{d}_{ist-1}$ is the demand control in the first stage of the production block.

3.3 Joint estimation: an iterative control-function algorithm for production and demand functions

The two blocks are mutually dependent: the demand elasticities are identified using physical productivity, while the first stage of the production block conditions on lagged demand advantages, which depend on those elasticities. We estimate the two blocks by alternating between them until the estimates converge.

Fixed-Point algorithm. The demand block maps production elasticities, through the instrument $\hat{z}^q_{ist}$, into demand parameters and demand advantages,

$$\mathcal{D}: \ \hat{\beta}_s \ \longrightarrow \ \left(\hat{\alpha}_{p,s}, \hat{\alpha}_{pm,s}, \hat{\rho}_{0,s}, \hat{\rho}_{1,s}, \hat{\rho}_{2,s}, \{\hat{D}_{ist}, \tilde{d}_{ist}\}\right),$$

while the production block maps lagged transient demand advantages into production elasticities and productivity,

$$\mathcal{P}: \ \{\tilde{d}_{is,t-1}\} \ \longrightarrow \ \left(\hat{\beta}_s, \{\hat{z}^q_{ist}\}\right),$$

as implied by the first stage of the control function under our timing assumptions. Hence, a jointly consistent estimator is one that is invariant to a further application of the composition $\mathcal{P} \circ \mathcal{D}$. Consistently with our industry-level estimation, we run the iterative algorithm for each industry separately.

Our implementation is as follows:

Step 0 (Initialization). We estimate the production function as in Section 3.1, with a first stage that conditions on market share but not on demand advantages. This is close to the *standard* technique for productivity computation in most of the literature. It yields initial production elasticities $\hat{\beta}^{(0)}_s$.

Step 1 (Demand update). Given $\hat{\beta}_s^{(r)}$ from iteration r, we build the instrument $\hat{z}_{ist}^{q,(r)}$, estimate the demand system by GMM, and recover updated elasticities $\hat{\alpha}_{p,s}^{(r+1)}$ and $\hat{\alpha}_{pm,s}^{(r+1)}$. We then construct the demand advantages $\hat{D}_{ist}^{(r+1)}$, their transient component $\tilde{d}_{ist}^{(r+1)}$, and the persistence parameters. Consistent with the timing assumptions laid out above, the object entering the production block is the lagged transient demand advantage $\tilde{d}_{is,t-1}^{(r+1)}$.

Step 2 (Production update). We re-estimate the production function using the procedure of Section 3.1, with $\tilde{d}_{is,t-1}^{(r+1)}$ in the first stage. The second-stage optimisation starts from $\hat{\beta}_s^{(r)}$ and perturbations around it, and returns updated production elasticities $\hat{\beta}_s^{(r+1)}$ and a new productivity series $\{\hat{z}_{ist}^{q,(r+1)}\}$.

Step 3 (Iteration). Steps 1 and 2 are repeated until the production elasticities converge. Writing $\Delta_\beta^{(r+1)} = \hat{\beta}_s^{(r+1)} - \hat{\beta}_s^{(r)}$ for the change in $(\beta_s^L, \beta_s^K, \beta_s^M)$, the algorithm stops at the first iteration $r^\star$ with $\max_j |\Delta_{\beta,j}^{(r^\star)}| < \varepsilon$ and $\|\Delta_\beta^{(r^\star)}\|_2 < \varepsilon$, where $\varepsilon = 0.03$. The demand elasticities are updated in every iteration but do not enter the stopping rule: they rest on a noisier instrument than the production elasticities, and we do not want noise in the demand block to decide when the algorithm stops. Upon convergence,[12] the procedure yields: (i) industry-level demand elasticities and demand dynamics parameters; (ii) industry-level production elasticities; and, consequently, (iii) firm-level series for demand advantages $\{\hat{D}_{ist}\}$ and physical productivity $\{\hat{z}_{ist}^q\}$. These objects are then used to construct TFPR, markups, and the subsequent decomposition analyses presented in the remainder of the paper.

3.3.1 Algorithm performance under nonlinearities: some Monte Carlo evidence

To assess the behaviour of the algorithm, namely the ability to return the true parameters of the production function, we resort to a series of Monte Carlo experiments. Firstly, we assess how the resulting estimates depend on the extent of nonlinearities in the TFPQ process. Secondly, we assess their sensitivity to the assumption of independent demand and supply.

Concerning the first experiment, we find that our estimator, even though not unbiased in small samples, performs better than the traditional ACF approach (i.e. no iteration and no control for latent demand advantages). We consider a menu of stochastic processes for TFPQ, with different forms of nonlinearities (detailed in Appendix D), and compare them; 3 of those are reported in Table 2.[13] We note that the first iteration (i.e. traditional ACF) is always biased: it overstates β^M by 0.12 to 0.19 in the three DGPs shown, because materials also respond to expected demand, which a control function without demand controls attributes to productivity; this is the main bias we wish to address with our methodology. The converged estimator removes most of this bias.

[12] We cap the procedure at 20 iterations; industries that reach the cap keep the estimates of the last iteration.

[13] Section D.3 lists where the simulation code departs from the procedure above.

Under DGP1 and DGP6 the production elasticities lie within 0.03 of the truth, and $\hat{\alpha}_p$ is -2.44 and -2.53 against a true value of -2.50. DGP7, the most nonlinear of the three, is the hardest case: $\hat{\beta}^L$ is 0.03 too low, $\hat{\beta}^K$ is 0.05 too high, and $\hat{\alpha}_p = -2.17$. The remaining biases are small in economic terms, but some exceed twice their bootstrap standard errors, for instance β^M under DGP6. The same pattern holds across all seven processes (Table D.2).

TABLE 2. Monte Carlo: production function estimates across selected DGPs

DGP	Stage	β^L (0.30)	β^M (0.50)	β^K (0.20)	α_p (-2.50)
DGP1 (AR(1), Gaussian)	Iter. 1	0.216	0.620	0.215	—
$z^q_{it} = \rho z^q_{i,t-1} + \xi^q_{it}$, $\rho = 0.7$		(0.012)	(0.025)	(0.007)	
	Final	**0.300**	**0.497**	**0.203**	**−2.44**
		(0.010)	(0.008)	(0.009)	
DGP6 (quadratic-mean loc.-scale)	Iter. 1	0.282	0.621	0.204	—
$z^q_{it} = \gamma_0 z^q_{i,t-1} + \gamma_3 (z^q_{i,t-1})^2 + (\gamma_1 + \gamma_2 z^q_{i,t-1})\varepsilon_{it}$		(0.006)	(0.020)	(0.011)	
	Final	**0.308**	**0.473**	**0.220**	**−2.53**
		(0.008)	(0.010)	(0.009)	
DGP7 (Hermite quantile)	Iter. 1	0.211	0.694	0.289	—
$Q_\tau(z^q_{it} \mid z^q_{i,t-1}) = \sum_{k=0}^{3} a_k(\tau) H_k(\tilde{z}_{i,t-1})$		(0.018)	(0.039)	(0.022)	
	Final	**0.270**	**0.497**	**0.248**	**−2.17**
		(0.014)	(0.016)	(0.037)	

Notes: True values in parentheses in the column headers. *Iter. 1* is the initialisation without demand controls; *Final* is the converged estimate of the iterative procedure (Section D.3). Panel block bootstrap standard errors in parentheses below each estimate, from 30 resamples of firms drawn with replacement; each resample re-estimates the production block at the converged demand elasticity, so the standard errors do not reflect uncertainty in $\hat{\alpha}_p$, for which none is reported. α_p is not estimated at iteration 1, which conditions on no demand information. One simulated panel of $N = 2{,}500$ firms and $T = 10$ years, single industry, per process. DGP numbering follows Section D.2, which reports the full menu of seven productivity processes; Appendix D describes the simulated economy and the sensitivity of the estimates to violations of the identifying assumption.

A second exercise asks what happens when the identifying assumption fails. Identification of the production block requires the current demand innovation to be absent from the materials proxy; the empirically relevant violation is not contemporaneous correlation but the case in which *lagged* demand innovations predict current productivity innovations, so that the proxy itself becomes informative about ξ^q_{it}. We re-simulate the AR(1) economy loading the productivity innovation on the previous period's demand innovation, and trace the bias as that loading rises (Section D.5). The damage is concentrated in the flexible inputs and is asymmetric (Figure D.1). Against a benchmark bias of -0.04 at zero loading, $\hat{\beta}^L$ falls a further 0.06 below its true 0.30 at a loading of 0.4 and a further 0.19 at 0.8 — implied correlations with the previous year's demand innovation of 0.41 and 0.70 — while materials move the other way, to 0.07 above a true 0.50. Capital and the demand elasticity are untouched, $\hat{\beta}^K$ staying within 0.01 of 0.20 and $\hat{\alpha}_p$ within 0.02 of -2.50 across the whole range. The pattern follows the timing assumptions: capital is predetermined and the demand elasticity rests on a supply-side instrument, whereas labour and materials are chosen after the productivity innovation is observed, and it is their identification that a contaminated proxy

undermines. The estimator is therefore robust to mild dependence, and degrades only when lagged demand shocks carry most of the information in current productivity innovations.

4 The dynamics of TFPQ: evidence of nonlinearities

After estimating the industry-specific production functions following the methodology detailed in Section 3.1, we construct firm-level physical productivity estimates ($\hat{z}^q$). To analyze the aggregate properties of these measures, we pool all observations across industries.

The algorithm is run separately on each of the 23 industries. The iteration between the two blocks reaches its fixed point in all but 3 of them, usually within three passes, and the production block is well behaved throughout: the observation-weighted output elasticities are 0.33 for labour, 0.55 for materials and 0.04 for capital, implying returns to scale of 0.93 on average and between 0.51 and 1.13 across industries. On the demand side the median price coefficient is -1.94, so the median industry faces demand somewhat more elastic than unit-elastic. Table C.1 reports both blocks industry by industry.

The properties of the resulting TFPQ estimates are studied in this section. We show they deviate from the linear-Gaussian AR(1) benchmark. To that end, it is perhaps useful to remember that if TFPQ were to follow such process (the usually assumed $z^q_{it} = \rho z^q_{i,t-1} + v_{it}, \quad v_{it} \sim N(0, \sigma^2)$, with $|\rho| < 1$), these properties (among others) would hold:

1. **Thin Tails:** Gaussian innovations preclude fat tails. Extreme productivity changes—whether "micro-disasters" or "lottery shocks"—occur with exponentially declining probability.

2. **State-Independent Risk:** All higher-order moments are constant across the distribution. High-TFPQ "superstars" and low-TFPQ "laggards" face identical volatility and tail risk

3. **Constant Persistence:** The scalar ρ implies that productivity histories stick equally to all firms. This precludes "memory-wiping" events where certain shocks erase past performance, regardless of a firm's position in the distribution.

Each of the following subsections shows that the 3 properties above fail to hold in the TFPQ distribution of Portuguese manufacturers. This will motivate our use of flexible, nonparametric methods to model TFPQ dynamics rather instead. In particular, a method that allows for what the literature has called *memory-wiping shocks* — shocks that sharply reverse a track record of a previously persistent high or low state —, associated with extreme tail events for particular sections of the distribution, that the classical AR(1) would render impossible.

4.1 Evidence from the unconditional distribution

Table 3 reports the unconditional distribution of TFPQ, set against the stationary Gaussian distribution of an AR(1) fitted to the same series.[14] The empirical TFPQ departs from the conventionally fitted AR(1) in a particular way: it is more single-peaked (i.e. more observations around the mean), while showing wider and fatter tails, at the expense of the shoulders of the distribution. Accordingly, the recovered TFPQ distribution is tighter between P90 and P10 (0.67 log points against 0.94), but wider between P01 and P99 (2.22 against 1.71). The mass at the tails is thus captured by a quantile-based kurtosis of 2.81, against 2.13. In comparison, the moment-based statistics (more sensitive to the extremes of the distribution) show much larger departures, signalling that the asymmetry and the tail thickness both sit in the outer regions of each tail. Figure E.1 in Appendix E illustrates the two distributions.

The departure from the classical AR(1) benchmark is, therefore, one of rare large realisations, not of a systematically tilted or simply a fat-tailed distribution.

TABLE 3. Summary statistics: TFPQ

				Skewness		Kurtosis		Interquantile ranges				Percentiles	
Variable	N	Mean	SD	Quantile	Moment	Quantile	Moment	$P_{75}-P_{25}$	$P_{90}-P_{10}$	$P_{95}-P_{05}$	$P_{99}-P_{01}$	P_{01}	P_{99}
TFPQ	88,386	-0.000	0.376	0.006	0.566	2.806	21.641	0.286	0.670	1.008	2.222	-1.031	1.191
				(0.009)	(0.247)	(0.026)	(1.671)						
AR(1) benchmark	—	0.000	0.367	0.000	0.000	2.134	3.000	0.495	0.940	1.206	1.706	-0.853	0.853

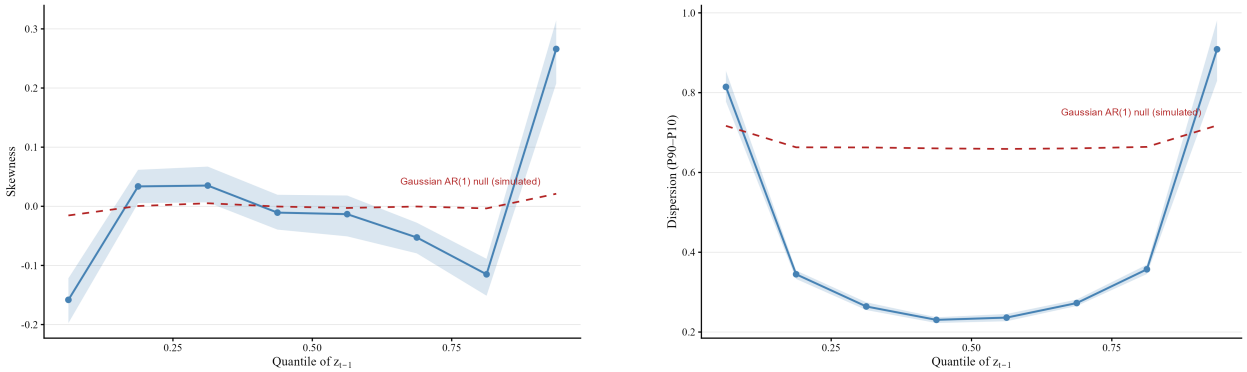
Notes. All statistics are defined in the text. Bootstrap standard errors in parentheses, from 1000 firm-block resamples drawn with replacement within industry, preserving each firm's full time series and the industry composition; they are reported for the four shape statistics only. The AR(1) benchmark row is not an estimate: it is each statistic evaluated in closed form on the stationary Gaussian distribution implied by an AR(1) fitted to TFPQ, and so carries no sampling error.

4.2 Evidence from the conditional distribution

Figure 1 shows how property (2) of the classical AR(1) previously mentioned fails to hold: splitting the sample into quantiles (8), state-dependence of higher-order conditional moments emerges. In particular, we find a U-shaped relationship between (lagged) TFPQ and conditional dispersion (P90-P10). Firms at the tails—both low-productivity and high-productivity types —face substantially higher volatility than those near the median, with dispersion at the extremes roughly 3.7 times as high as at the center. This contradicts the constant-volatility assumption of the Gaussian benchmark and suggests that where one firm sits in the TFPQ distribution matters for the shock

[14] The benchmark is the stationary distribution of $z_t^q = 0.72 \cdot z_{t-1}^q + \varepsilon_t$ with $\varepsilon_t \sim N(0, 0.26^2)$, where the persistence and the innovation variance come from an OLS regression of TFPQ on its own lag. The quantile statistics are skewness $= [(P_{90} - P_{50}) - (P_{50} - P_{10})]/(P_{90} - P_{10})$, which is 0 for any symmetric distribution, and kurtosis $= (P_{92.5} - P_{7.5})/(P_{75} - P_{25})$, which is 2.13 for a Gaussian; both remain well defined however heavy the tails are, and we refer to them simply as skewness and kurtosis in the rest of the paper. The moment statistics are the conventional $\mathbb{E}[(z^q - \mu)^3]/\sigma^3$ and $\mathbb{E}[(z^q - \mu)^4]/\sigma^4$, and are reported alongside too.

FIGURE 1. Higher-order Moments of TFPQ by Productivity Decile



(a) Conditional skewness

(b) Conditional dispersion

Notes: Observations are ordered by deciles of the distribution of $t-1$ residualized TFPQ (x-axis). Panel (a) reports the Kelley skewness. Panel (b) reports the P90–P10 dispersion

distribution it faces.

Panel (a) shows that also conditional skewness varies systematically with the inherited state, and that its sign carries economic content. Through the body of the distribution the asymmetry is mean-reverting: firms in the lower-middle octiles face innovations with a longer right tail (a skewness of about 0.03), while firms in the upper-middle octiles face a longer left one, falling monotonically to -0.12. A firm's relative position industry therefore governs the direction of its risk and not merely its magnitude — an above-average position is more easily lost than improved upon. The most interesting feature, however, lies at the extremes: the sign flip. In the bottom octile skewness is -0.2: a firm already far below its industry mean is more exposed to falling further than to recovering. In the top octile it is 0.3, the longer tail pointing further upward. The extremes are thus not merely more volatile than the centre; their risk seems to point away from it, making this very low (high) status more persistent than what would be entailed by a Gaussian AR(1).

Together, these patterns confirm that both the scale (dispersion) and shape (skewness) of idiosyncratic risk vary systematically with the inherited productivity state. The Gaussian AR(1) benchmark, which restricts all higher-order conditional moments to be constant, is fundamentally incompatible with these empirical regularities. This motivates our flexible quantile-based approach, that we now present, allowing the full conditional distribution of innovations to vary nonparametrically with lagged TFPQ.

4.3 Evidence from a quantile autoregression framework

A quantile-based framework for the TFPQ process. We now turn our attention to property (3): constant persistence of the TFPQ process. To test this, we adopt the quantile-based panel method of Arellano et al. (2017), which models a given variable (TFPQ, in our case) through a

conditional quantile function (CQF) [15]:

$$z_{it}^q = Q_z\left(z_{i,t-1}^q, u_{it}\right), \quad \left(u_{it} \mid z_{i,t-1}^q, z_{i,t-2}^q, \ldots\right) \sim U(0,1), \quad t > 1 \tag{6}$$

where $Q_z(z_{i,t-1}^q, \tau)$ maps a firm with lagged productivity $z_{i,t-1}^q$ and a shock of rank τ into its current TFPQ. The innovation u_{it} is uniformly distributed on $[0,1]$, allowing the quantile function to translate shock ranks into productivity outcomes without imposing distributional assumptions. We estimate Q_z nonparametrically using Hermite polynomials evaluated at quantile points $\tau \in \{0.11, \ldots, 0.88\}$.

This framework nests the Gaussian AR(1) as a special case. Under linearity and normality, the CQF collapses to:

$$z_{it}^q = \rho z_{i,t-1}^q + \Phi^{-1}(u_{it}; \sigma) \tag{7}$$

where Φ^{-1} is the inverse CDF of $N(0, \sigma^2)$. More generally, this framework allows for *nonlinear persistence*, defined as the partial derivative of the quantile function:

$$\rho\left(z_{i,t-1}^q, \tau\right) = \frac{\partial Q_z\left(z_{i,t-1}^q, \tau\right)}{\partial z_{i,t-1}^q} \tag{8}$$

Persistence is thus no longer a constant but varies with both the inherited state $z_{i,t-1}^q$ and the shock rank τ. A low value of $\rho(z, \tau)$ indicates a "memory-wiping" event where the shock partially or fully erases past productivity, while a high value indicates strong persistence. This flexibility allows i.i.d. innovations to have heterogeneous effects on productivity dynamics depending on the firm's inherited state.

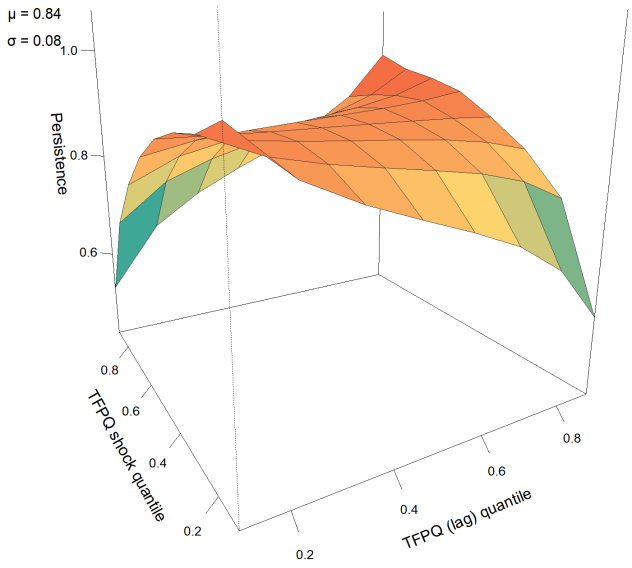
Estimates. Figure 2 displays the estimated persistence surface $\rho(z_{i,t-1}^q, \tau)$ as a function of both lagged TFPQ and the shock rank. The three-dimensional representation reveals rich nonlinearities that are entirely absent from the constant-ρ implied by a AR(1).

For low-TFPQ firms, persistence *declines* with the shock rank: large positive shocks (high τ) act as history-erasing events. A laggard firm experiencing a favorable innovation sees its poor productivity record partially wiped out—the shock allows the firm to break free from its low-productivity state, though not necessarily to reach the productivity frontier.

Symmetrically, for high-TFPQ firms, persistence declines with *negative* shocks (low τ). Large adverse innovations, albeit rare (as seen in the previou subsection), erode established productivity advantages, partially erasing the firm's strong track record. This asymmetry implies that productivity states at the extremes are less durable than those near the center: large shocks—whether

[15] At this stage, we do not split productivity into a permanent and a transitory component as the original paper does. While customary in work studying household income processes, this split is rarely done in the firm dynamics literature.

FIGURE 2. Nonlinear Persistence of TFPQ



Notes: The surface plots the estimated persistence $\rho(z^q_{i,t-1}, \tau) = \partial Q_z / \partial z^q_{i,t-1}$ as a function of lagged TFPQ (x-axis) and the shock rank τ (y-axis). Persistence varies substantially across both dimensions, indicating that the impact of past productivity depends critically on both the inherited state and the magnitude of current innovations.

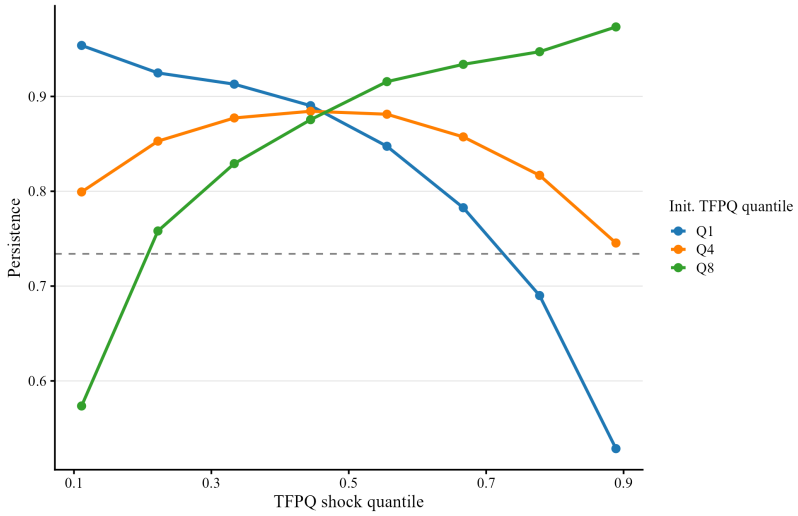
positive or negative—can fundamentally reset a firm's position in the distribution.

Firms near the median exhibit elevated and stable persistence across shock realizations, as shown in the slice-wise representation of Figure 3. For these middle- productivity firms, shocks have more uniform effects: productivity remains anchored to its inherited level regardless of whether the innovation is positive or negative; in other words, they are somewhat closer to the linear-Gaussian AR(1) bechmark. In contrast, firms at the extremes experience asymmetric responses where sufficiently large shocks can fundamentally alter their productivity trajectory.

Note that this nonlinear persistence structure could only be obtained through a methodology that allows for (a) richer dynamics in productivity and (b) controls for demand-side variables together. Otherwise, a constant persistence would have been imposed (a) and the first-stage of the control function would be contaminated (b). Focusing on the latter, as Figure F.2 of the Appendix shows, controlling for demand advantages changes the resulting persistence structure in two meaningful ways, compared to the standard one-shot ACF. First, it works as level shifter - reducing persistence across the board. Second, it reveals memory-wiping features for high-productivity firms (lower persistence).

These findings confirm that TFPQ dynamics exhibit substantial nonlinearities that are fully subsumed under the constant-persistence assumption of Gaussian AR(1) models. The magnitude

FIGURE 3. TFPQ Persistence by Shock Quantile



Notes: Persistence is plotted against lagged TFPQ for three representative shock quantiles: $\tau = 0.11$ (bottom decile, blue), $\tau = 0.50$ (median, orange), and $\tau = 0.88$ (top decile, green). The divergence across shock ranks confirms that memory-wiping effects vary systematically with both the inherited state and the innovation magnitude.

and direction of shocks interact with the inherited state to determine whether productivity states persist or get reset. In short, the tails of the TFPQ distribution contain particularly rich dynamics that would go unnoticed in standard empirical approaches to productivity. In the following sections we explore what this means for TFPR – empirically – and for the insights of standard models of firm dynamics – quantitatively.

5 From TFPQ to TFPR

TFPQ measures how efficiently a firm turns inputs into physical output. TFPR adds the price the firm charges, and so also reflects its pricing decision. In our framework, log revenue productivity is

$$z^r_{ist} = z^q_{ist} + p_{ist} \tag{9}$$

where p_{ist} is the log price charged by firm i in industry s at time t. The mapping from TFPQ to TFPR matters for three reasons. It shows whether a productivity advantage is passed on as a lower price or retained as a higher margin. It bears on the misallocation literature, where TFPR dispersion is read as evidence of distortions but may instead reflect differences in demand. And if the pass-through varies with the firm's state, the same productivity shock has different revenue consequences for firms in different states.

5.1 Prices, productivity and demand

The price is what separates TFPR from TFPQ, so we first ask what determines it. Table 4 reports the correlations between prices, TFPQ and demand advantages. TFPQ and prices are negatively correlated in the pooled sample (-0.110), as cost pass-through predicts: high-TFPQ firms have lower marginal costs and charge lower prices on average. The correlation is stronger in industries producing homogeneous goods, such as precast concrete (-0.35) and cut stone[16] (-0.32), exceeding the -0.19 that Foster et al. (2008) report for US homogeneous-goods industries. Table C.2 gives the industry-by-industry figures.

Prices co-move far more strongly with demand advantages than with TFPQ. The correlation between our estimated demand advantages and prices is close to 0.7, against -0.110 for TFPQ: firms in favourable demand states charge higher prices on average. Table C.2 shows that this holds across quite different market structures: the price–demand correlation is strong in relatively homogeneous-good industries such as cork products, and also in differentiated segments such as cotton fabrics and furniture (wooden and kitchen furniture). Wine and fresh cheese are the exceptions, with demand advantages barely reflected in prices.

TFPQ and demand advantages are themselves positively correlated, though weakly. The pooled correlation is 0.045, and it is positive in twelve of the twenty industries for which the demand block is identified. In many of these, we indeed observe that high-TFPQ firms not only produce more efficiently but also offer products that command greater demand — whether through superior quality, stronger brands, or better product-market fit. This complementarity has important implications for our decomposition of markups later in this section: when we observe high-TFPQ firms charging high markups, part of this relationship operates through the demand channel rather than solely through cost advantages.

Table 4. Correlations: price, TFPQ, and demand

	Price	TFPQ	Demand
Price	1.000		
TFPQ	-0.110	1.000	
Demand	0.698	0.045	1.000

5.2 The TFPR distribution

TFPR is more dispersed, more skewed and more heavy-tailed than TFPQ. Table 5 compares the two. The gap between the tenth and ninetieth percentiles is 1.11 log points against 0.67, skewness is 0.14 against 0.01, and the quantile kurtosis of 3.47, against 2.13 for a Gaussian, points to more

[16] This is comprised of natural stones simply cut prior to any transformation into end-use goods.

frequent extreme realisations than for TFPQ. These patterns are consistent with pricing and demand factors amplifying productivity-driven heterogeneity: firms with high technical efficiency can leverage their cost advantages through markup variation, while demand shocks introduce additional dispersion orthogonal to the production side. Figure F.1 in Appendix F plots the unconditional density against its Gaussian benchmark.

Table 5. Summary statistics: TFPQ, TFPR, and prices

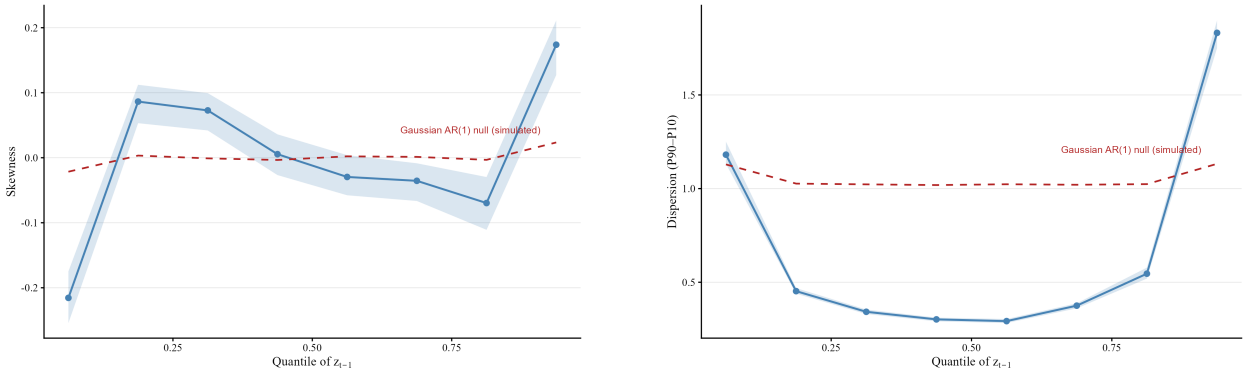
				Skewness		Kurtosis		Interquantile ranges				Percentiles	
Variable	N	Mean	SD	Quantile	Moment	Quantile	Moment	$P_{75}-P_{25}$	$P_{90}-P_{10}$	$P_{95}-P_{05}$	$P_{99}-P_{01}$	P_{01}	P_{99}
TFPQ	88,386	-0.000	0.376	0.006	0.566	2.806	21.641	0.286	0.670	1.008	2.222	-1.031	1.191
				(0.009)	(0.247)	(0.026)	(1.671)						
TFPR	88,386	0.006	0.637	0.144	0.683	3.474	15.302	0.411	1.114	1.896	3.723	-1.599	2.124
				(0.014)	(0.151)	(0.050)	(0.624)						
Log price	88,386	0.006	0.557	0.336	1.085	5.902	21.024	0.191	0.770	1.576	3.370	-1.333	2.037
				(0.023)	(0.208)	(0.120)	(1.058)						

Notes. All statistics are defined in the notes to Table 3. Bootstrap standard errors in parentheses, from 1000 firm-block resamples drawn with replacement within industry, preserving each firm's full time series and the industry composition; they are reported for the four shape statistics only.

Figure 4 turns to the conditional moments. Like TFPQ, TFPR exhibits U-shaped conditional dispersion (Panel b): firms at both tails of the lagged TFPR distribution face greater uncertainty about future revenue outcomes than those near the median. Conditional skewness (Panel a) varies with the inherited state, though the pattern is less systematic than for TFPQ. The conditional kurtosis, reported in Figure F.1 alongside the unconditional density, stays elevated across the entire distribution of lagged TFPR — roughly 2.4–3.1 at every quantile — indicating persistent tail risk regardless of a firm's inherited revenue position. That uniformly high kurtosis suggests large revenue shocks can strike any firm, not only those at the extremes.

TFPR is also more persistent than TFPQ. Estimating the same quantile autoregression as in Section 4 on TFPR gives a mean persistence of 0.86 across states and shocks, against 0.84 for TFPQ, and the surface varies less with the inherited state. Appendix G reports it.

FIGURE 4. Higher-order Moments of TFPR by Revenue Productivity Quantile



(a) Conditional skewness

(b) Conditional dispersion

Notes: Observations are ordered by quantiles of the distribution of $t-1$ residualized TFPR (x-axis). Panel (a) reports the Kelley skewness. Panel (b) reports the P90–P10 dispersion. The blue solid lines are the data moments; the dashed red lines are the moments implied by an $AR(1)$ estimated on the empirical TFPR.

5.3 The TFPQ–TFPR Correlation

The contemporaneous correlation between TFPQ and TFPR is 0.49, against the 0.75 that Foster et al. (2008) report for a set of homogeneous-good US products. TFPR tracks TFPQ moderately. Since $z^r = z^q + p$, a correlation of this size means that price variation is of the same order as variation in TFPQ, which is what the larger dispersion of TFPR already suggested.

TABLE 6. TFPQ to TFPR Transition Matrix

	$z_t^{r,1}$	$z_t^{r,2}$	$z_t^{r,3}$	$z_t^{r,4}$	$z_t^{r,5}$	$z_t^{r,6}$	$z_t^{r,7}$	$z_t^{r,8}$
$z_t^{q,1}$	0.53	0.19	0.03	0.02	0.02	0.02	0.05	0.14
$z_t^{q,2}$	0.16	0.36	0.26	0.04	0.03	0.03	0.04	0.09
$z_t^{q,3}$	0.08	0.22	0.30	0.22	0.04	0.04	0.04	0.07
$z_t^{q,4}$	0.05	0.12	0.17	0.37	0.16	0.05	0.04	0.05
$z_t^{q,5}$	0.05	0.05	0.13	0.16	0.40	0.09	0.07	0.05
$z_t^{q,6}$	0.04	0.03	0.06	0.12	0.21	0.37	0.10	0.07
$z_t^{q,7}$	0.05	0.03	0.03	0.05	0.13	0.30	0.31	0.11
$z_t^{q,8}$	0.03	0.02	0.02	0.02	0.02	0.09	0.37	0.43

Notes: Each cell reports the fraction of firm-year observations in a given TFPQ quantile group (row) that falls into each TFPR quantile group (column). Rows sum to one.

Table 6 looks at the same relationship through rank mobility. Firms are grouped into TFPQ deciles, and each row reports the share of firms in each TFPR decile. Ranks line up at the tails: 72% of firms in the bottom TFPQ decile are in the bottom two TFPR deciles, and 80% of top-decile TFPQ firms are in the top two TFPR deciles. However, firms in the 5th TFPQ decile are distributed more widely across TFPR deciles. For these firms, the price does most of the work in

determining where they rank in revenue productivity.

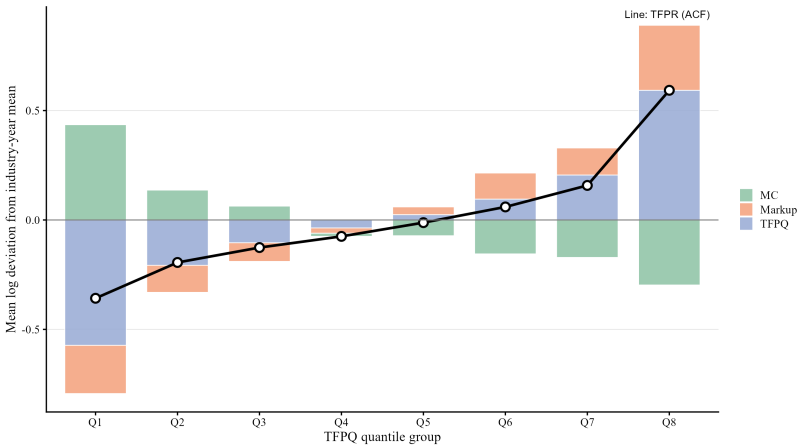
5.4 Decomposing TFPR into TFPQ, markups and marginal costs

The price is the product of the markup and marginal cost, so TFPR decomposes as

$$z^r_{ist} = z^q_{ist} + p_{ist} = z^q_{ist} + \mu_{ist} + mc_{ist} \tag{10}$$

where p_{ist} is the log price, μ_{ist} the log markup, and mc_{ist} the log marginal cost. Markups are estimated following De Loecker and Warzynski (2012), using the materials elasticity from Section 3. Appendix H gives the details. Figure 5 plots the decomposition by TFPQ decile. Between the lowest and the highest decile, average TFPQ differs by 1.17 log points, but TFPR by only 0.95. The gap is the price: markups are 0.52 log points higher at the top, but marginal costs are 0.73 lower, so prices are 0.21 lower. Prices thus offset part of the TFPQ difference and compress the TFPR distribution relative to TFPQ. The markup contribution turns positive only above the median decile. Below it, markups are below average regardless of TFPQ.

FIGURE 5. Decomposition of TFPR per TFPQ quantile



6 Model

The facts laid out in Section 4 refer mostly to the stochastic process for TFPQ — a primitive in most models of firm dynamics — and not to what that means for firm behaviour. This section addresses that question by taking a standard partial-equilibrium model of labour demand with adjustment costs, feeding it our estimated process, and comparing the outcomes with those obtained when the model is fed a conventional AR(1). In both cases, the model is estimated to match the same set of conventional moments in the Portuguese data.

The exercise has three steps. First, we discretise the two competing processes into Markov

chains (NL — our estimated specification — versus a standard AR(1)), *calibrated to be observationally equivalent in linear terms*[17]. Second, we estimate a model of dynamic labour demand with adjustment costs, under the two alternative processes for TFPQ. The estimation is done by simulated method of moments (SMM), targeting the same set of moments in both cases. The idea is to see how much the estimated adjustment costs differ when the underlying productivity process is misspecified. Thirdly, we evaluate the implications of each estimated model for untargeted moments — namely the distribution of employment growth rates.

6.1 The firm's problem

A firm enters the period with employment e inherited from last period's decision and observes its productivity state z, which follows the Markov chain described above.[18] Output is produced with the inherited stock, so there is one period of time to build, and operating profit is

$$\pi(z, e) = \exp(z)\, e^{\alpha}. \tag{11}$$

Choosing next period's employment e' costs

$$C(e', e) \;=\; \underbrace{\frac{\nu}{2}\left(\frac{e' - e}{e}\right)^{2} e}_{\text{convex}} \;+\; \underbrace{\left[\gamma_H\, \mathbf{1}\{e' > e\} + \gamma_F\, \mathbf{1}\{e' < e\}\right] \left|e' - e\right|}_{\text{kinked linear}} \;+\; \underbrace{F\, e}_{\text{fixed}}, \tag{12}$$

paid only if the firm adjusts. The three terms are the standard menu (Cooper and Haltiwanger, 2006): a smooth quadratic cost, a linear cost that is allowed to differ between hiring and firing, and a lumpy component proportional to the size of the workforce that prompts a region of inaction. The value function is $V(z, e) = \max\{V^a(z, e),\, V^i(z, e)\}$ with

$$V^i(z, e) = \pi(z, e) - we + \beta\, \mathbb{E}\big[V(z', e) \mid z\big], \tag{13}$$

$$V^a(z, e) = \max_{e'}\; \pi(z, e) - we - C(e', e) + \beta\, \mathbb{E}\big[V(z', e') \mid z\big]. \tag{14}$$

There is no entry, no exit and no aggregate state. The parameter space to be estimated is the adjustment cost block:

$$\Theta = (\nu,\ \gamma_H,\ \gamma_F,\ F,\ w),$$

We also impose two parameters: $\alpha = 0.558$ and the discount factor $\beta = 0.95$, common in the literature. The wage is estimated rather than imposed because it is the only parameter that pins

[17] Both chains live on the same 150-node grid, and both have a first-order autocorrelation of 0.751 and an unconditional standard deviation of 0.378. What separates them are the higher order moments and nonlinear persistence structure. Furthermore, given that the object of interest lives in the tails, the support for our discretisation runs from the empirical 0.1st to the 99.9th percentiles of z (-2.25 to 2.76) rather than to the usual ± 2–3 standard deviation box. Further details on the discretisation are provided in Appendix J.

[18] In this exercise, we interpret $\pi(\cdot)$ as an optimised production function: materials are chosen statically and capital is held fixed, so both are substituted out and the curvature that remains in labour is α.

down the scale of the firm, and the median firm size is one of the targets.

6.2 Estimation

We estimate Θ by simulated method of moments, separately for each process, following Cooper et al. (2024). For a candidate Θ we solve (13)–(14) by value function iteration on a grid of 150 productivity nodes and up to 900 employment nodes, simulate a panel of 5,000 firms for 50 years after a 60-period burn-in, and compute on the simulated panel the same statistics we compute on the data. The criterion is the sum of squared percentage deviations,

$$\mathcal{L}(\Theta) \;=\; \frac{\left(M^d - M^s(\Theta)\right)'}{M^d}\, W\, \frac{M^d - M^s(\Theta)}{M^d}, \tag{15}$$

where M^d stacks the $J = 7$ moments measured in the data, $M^s(\Theta)$ their simulated counterparts, the division is taken element by element, and W is the weighting matrix, which we set to the identity. The seven targets, reported in Table 7, are all related to the distribution of year-on-year growth rates in employment: the share of observations with no adjustment, the shares hiring and firing, the shares undertaking a positive and a negative employment spike, the autocorrelation of log employment, and median employment (i.e. size). Adjustment is defined as a midpoint (DHS) growth rate exceeding 2.5 percent in absolute value and a spike as one exceeding 20 percent, in line with the literature.

6.3 Results

The exercise targets only moments pertaining to the distribution of employment growth rates, so they are independent of how one estimates the underlying productivity. In our baseline economy with the frictions described above, the fit is similar on both the AR1 and NL productivity processes, with broadly all moments close to the data throughout — see Table 7. Excluding the autocorrelation of log employment, the two criteria are practically indistinguishable.[19]. This means that, at first sight, the two processes can deliver similar moments of the distribution of employment growth rates. The adjustment cost landscape, however, is very different.

Table 8 shows the estimated adjustment cost parameters. The AR(1) process for productivity needs a convex cost 4.3 times larger ($\nu = 0.209$ against 0.048) and a fixed cost an order of magnitude larger (8.2×10^{-4} against 7.0×10^{-5}). This is not an accident of the estimator: an AR(1) delivers a homoskedastic, symmetric flow of news, so the only way to generate the observed mass of very large employment changes alongside a 43 percent inaction rate is to make the frictions themselves do the sorting — a large fixed cost to keep most firms still, and a large convex cost to smooth the

[19]Employment in the data is very persistent (0.984) and neither process reaches it, because a model whose only endogenous state is last period's employment and whose only exogenous state is a first-order Markov productivity cannot manufacture that much inertia at any setting of the frictions.

TABLE 7. Targeted moments: data, the estimated models, and their frictionless counterparts

		With frictions		*No frictions*	
Moment	Data	AR(1)	NL	AR(1)	NL
Inaction rate	0.434	0.436	0.440	0.035	0.109
Hiring rate	0.301	0.287	0.284	0.483	0.449
Firing rate	0.265	0.277	0.277	0.482	0.442
Positive spike rate	0.093	0.091	0.095	0.331	0.190
Negative spike rate	0.088	0.092	0.091	0.330	0.189
Autocorrelation of $\log e$	0.984	0.927	0.879	0.750	0.769
Median employment	10.000	9.941	10.010	9.629	9.822
Criterion $\mathcal{L}$		0.010	0.018	15.971	3.673
excluding $\rho(\log e)$		0.006	0.007	15.915	3.625

Notes. Targeted moments in the data and in the model with and without frictions. NL denotes our non-linear productivity process and AR(1) the conventional AR(1) process. The criterion $\mathcal{L}$ is the loss in (15).

TABLE 8. SMM estimates of the adjustment-cost parameters

Parameter	AR(1)	Nonlinear	NL/AR(1)
Convex cost ν	0.209	0.048	0.230
Linear hiring cost γ_H	0.013	0.028	2.182
Linear firing cost γ_F	0.023	0.008	0.324
Fixed cost F	8.225×10^{-4}	6.992×10^{-5}	0.085
Wage w	0.213	0.207	0.974
γ_F/γ_H	1.858	0.276	0.149

Notes. Estimated parameters behind the model with frictions in Table 7. See (12)–(14) for the meaning of each parameter. The last column is the ratio of the NL to the AR(1) estimate.

rest. The NL process, instead, supplies large moves from the tails of the shock distribution itself. Indeed, even when all adjustment costs are set to zero (a frictionless benchmark), the NL process delivers a 10.9 percent inaction rate, three times the 3.5 percent of the AR(1).

Apart from changing the scale of the frictions, the two processes diverge on the asymmetry between hiring and firing costs: under the AR(1) the estimated firing cost is 1.86 times that of hiring; under NL, it becomes 0.28 times. This is a reflection of the dynamics on the left tail of the distribution of productivity (including, for instance, memory-wiping shocks) which, once accounted for, significantly downplay the role of firing frictions.

Looking at the higher order moments of the employment growth rate distribution — left

completely untargeted — is when the advantages of the NL process become clear. Looking at the adjusters (those who actually change their employment levels), the AR(1) process produces a distribution that is too narrow (s.d. of 0.197 against 0.268), too symmetric (skewness of −0.046 against −0.123) and too thin-tailed (the AR(1) delivers −0.81 kurtosis — thinner tails than a Gaussian). The NL process, instead, delivers a distribution that is very close to the data, only with a slight overstatement of kurtosis — as shown in Table 9.

TABLE 9. The distribution of employment growth among adjusters (untargeted)

Statistic	Data	AR(1)	Nonlinear
Standard deviation	0.268	0.197	0.290
Skewness	-0.123	-0.046	-0.164
Excess kurtosis	5.316	-0.806	6.258
$P_{90} - P_{50}$	0.246	0.181	0.250
$P_{50} - P_{10}$	0.325	0.303	0.317
$P_{90} - P_{10}$	0.571	0.484	0.567
$(P_{90} - P_{50})/(P_{50} - P_{10})$	0.759	0.599	0.788

Notes. Untargeted moments of employment growth, conditional on adjusting ($|g| > 0.025$). Data refers to our sample. Definitions follow Decker et al. (2016).

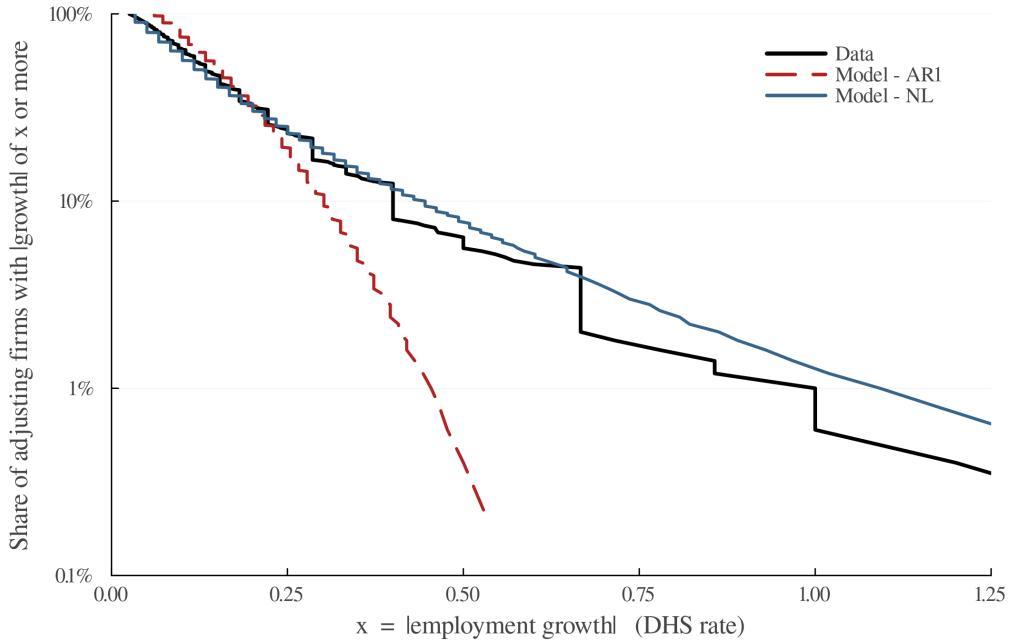
The linear process for productivity is unable to reproduce a set of tail adjustments; in particular, conditional on adjusting, it delivers essentially no large employment increases (i.e. above 50%). This is evident from Figure 6, which shows that up to small adjustments — 0.20 — the two models and the data align; however, from there onwards, they diverge, with the AR(1) model not being able to generate DHS growth rates above 0.53. The NL process, instead, is able to reproduce the empirical distribution of large adjustments, with the tails of the distribution being very close to the data, up to $|g| = 1.25$, which for a DHS growth rate means a firm more than quadrupling, or shrinking to under a quarter of its size.

Another way of looking at the ability of the two processes to reproduce the tails of the distribution is through the percentile differentials: following Decker et al. (2016), we look at the gaps between the median firm and (i) the P90 firm (high growth) and (ii) the P10 (low growth). While the lower tail is matched tolerably by both processes (0.325 in the data, 0.303 and 0.317 in the models), the upper tail is not, and it is where the AR(1) loses 26 percent of the empirical spread (0.181 against 0.246 in the data, with the NL process at 0.250). The overall $P_{90} - P_{10}$ gap — Decker et al.'s headline dispersion measure — is 0.571 in the data, 0.567 under our process and 0.484 under the AR(1). The ratio of the two halves, which equals one for any symmetric distribution whatever its shape, is 0.759 in the data and 0.788 under our process, but 0.599 under the AR(1): the linear model gets the asymmetry of firm growth wrong in the same direction and by roughly the same amount as it gets the tails wrong. All of these results are delivered by a simple model with a single

endogenous variable, estimated to match *only* the frequencies of adjustment, differing only in their productivity processes — which are themselves indistinguishable up to their first two autoregressive moments.

The extent to which this matters is, of course, dependent on the research question at hand. For instance, the literature on high-growth firms, such as Decker et al. (2016), sees bursts of firm-level growth as a key element of business dynamism. Relatedly, in the study of small young firms, reproducing high growth rates is critical — e.g. going from 2 to 4 employees is essentially unattainable under a linear process for productivity.

FIGURE 6. Frequency of employment changes, by size: data and model



Notes. Survival function of the absolute growth rate of employment among adjusters, on a log vertical scale. The horizontal axis is the absolute change in employment, so it captures both hiring and firing. Model curves come from data simulated under the model with frictions; Data is the Portuguese firm data.

7 Conclusion

This paper asks how firm-level productivity evolves, how far it is from the linear-Gaussian AR(1) that quantitative models assume, and what the difference means for those models. Because the question concerns technical efficiency and not revenue, measurement comes first. We develop an iterative procedure that jointly estimates a production function and a demand system on Portuguese manufacturing data with firm-level prices and quantities, so that the productivity residual is purged of demand contamination, without imposing linear dynamics on productivity. The result is a firm-level measure of technical efficiency (TFPQ) that the remainder of the paper studies.

Our estimated TFPQ departs from the linear-Gaussian benchmark, in both unconditional and conditional distribution. The body of the TFPQ distribution is tighter than an AR(1) implies and the tails wider. Conditional dispersion at the extremes of the lagged TFPQ distribution is roughly 3.7 times as high as at the centre, conditional skewness flips sign at the two ends, and persistence varies with both the inherited state and the size of the shock. We document a memory-wiping property: a large adverse shock erases much of a productive firm's record, and a large favourable shock much of an unproductive firm's record. The median firm, by contrast, is more likely to retain its inherited TFPQ level, like an AR(1) would predict. The nonlinearity lives largely in the tails - as the higher order moments reflect. Taken together, we can state that where a firm sits in the distribution determines both the risk it faces and how much of its history survives a shock.

We take these non-standard features of a productivity process to standard model of labour demand with adjustment costs. We estimate two versions of it, on the same moments of the employment growth distribution: one with the nonlinear process and another with a benchmark AR(1); both processes are undistinguishable up to the first two autoregressive moments and both fit the targete moments equally well. However, they cary starkly different economic interprerations in terms of adjustment costs: convex costs a quarter of their AR(1) value, fixed costs a tenth, and firing costs sit below hiring costs rather than almost twice. Only the nonlinear process reproduces the untargeted high-order moments and tails of the employment growth distribution. Most remarkbly, the sort of bursts and contractions that an AR(1) cannot generate at much much stronger frictions. Where the shape of the productivity process is left to the frictions to explain, the frictions are misestimated.

Because we observe prices, TFPR follows directly from TFPQ, and it does not track technical efficiency closely. The two are correlated at 0.49, TFPR is more dispersed, more skewed and more heavy-tailed, and it is more persistent (0.86 against 0.84). The gap is the price, and prices co-move with demand advantages far more than with efficiency. Along the TFPQ distribution prices fall as efficiency rises, so a 1.17 log-point difference in TFPQ between the bottom and top deciles shows up as a 0.95 log-point difference in TFPR. Revenue-based measures thus compress efficiency differences, add the dispersion of prices, and make technical efficiency look more persistent than it is.

References

Ackerberg, D. A., K. Caves, and G. Frazer (2015): "Identification Properties of Recent Production Function Estimators," *Econometrica*, 83, 2411–2451. (Cited on pages 4, 7, 8, 9, and 40.)

Arellano, M., R. Blundell, and S. Bonhomme (2017): "Earnings and Consumption Dynamics: A Nonlinear Panel Data Framework," *Econometrica*, 85, 693–734. (Cited on pages 3 and 16.)

Arellano, M. and S. Bonhomme (2016): "Nonlinear Panel Data Estimation via Quantile Regressions," *The Econometrics Journal*, 19, C61–C94. (Cited on pages 1 and 41.)

Atkeson, A. and A. Burstein (2008): "Pricing-to-Market, Trade Costs, and International Relative Prices," *American Economic Review*, 98, 1998–2031. (Cited on page 7.)

Banerjee, A. V. and B. Moll (2010): "Why Does Misallocation Persist?" *American Economic Journal: Macroeconomics*, 2, 189–206. (Cited on page 3.)

Bloom, N. (2009): "The Impact of Uncertainty Shocks," *Econometrica*, 77, 623–685. (Cited on page 3.)

Bloom, N., M. Floetotto, N. Jaimovich, I. Saporta-Eksten, and S. J. Terry (2018): "Really Uncertain Business Cycles," *Econometrica*, 86, 1031–1065. (Cited on page 3.)

Blundell, R. and S. Bond (2000): "GMM Estimation with Persistent Panel Data: An Application to Production Functions," *Econometric Reviews*, 19, 321–340. (Cited on page 4.)

Bond, S., A. Hashemi, G. Kaplan, and P. Zoch (2021): "Some Unpleasant Markup Arithmetic: Production Function Elasticities and Their Estimation from Production Data," *Journal of Monetary Economics*, 121, 1–14. (Cited on page 56.)

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010): "Quantile and Probability Curves Without Crossing," *Econometrica*, 78, 1093–1125. (Cited on page 59.)

Cooper, R., J. Haltiwanger, and J. L. Willis (2015): "Dynamics of Labor Demand: Evidence from Plant-Level Observations and Aggregate Implications," *Research in Economics*, 69, 37–50. (Cited on page 4.)

Cooper, R., J. C. Haltiwanger, and J. Willis (2024): "Declining Responsiveness at the Establishment Level: Sources and Productivity Implications," Tech. Rep. 32130, National Bureau of Economic Research. (Cited on pages 4 and 25.)

Cooper, R. and O. Ozturk (2022): "TFPR: Dispersion and Cyclicality," *NBER Working Paper*. (Cited on page 4.)

Cooper, R. W. and J. C. Haltiwanger (2006): "On the Nature of Capital Adjustment Costs," *The Review of Economic Studies*, 73, 611–633. (Cited on pages 3, 4, and 24.)

De Loecker, J. (2011): "Product Differentiation, Multiproduct Firms, and Estimating the Impact of Trade Liberalization on Productivity," *Econometrica*, 79, 1407–1451. (Cited on pages 4 and 7.)

De Loecker, J., P. K. Goldberg, A. K. Khandelwal, and N. Pavcnik (2016): "Prices, Markups, and Trade Reform," *Econometrica*, 84, 445–510. (Cited on pages 4, 6, and 56.)

De Loecker, J. and F. Warzynski (2012): "Markups and Firm-Level Export Status," *American Economic Review*, 102, 2437–2471. (Cited on pages 9, 23, 56, and 57.)

De Vera, M. (2022): "Firm-Level Productivity and Demand Shocks in Imperfectly Competitive Labor Markets: Implications for Wage Dynamics," *mimeo*. (Cited on page 4.)

Decker, R. A., J. Haltiwanger, R. S. Jarmin, and J. Miranda (2016): "Where Has All the Skewness Gone? The Decline in High-Growth (Young) Firms in the U.S." *European Economic Review*, 86, 4–23. (Cited on pages 2, 27, and 28.)

——— (2020): "Changing Business Dynamism and Productivity: Shocks versus Responsiveness," *American Economic Review*, 110, 3952–3990. (Cited on page 4.)

Eslava, M., J. Haltiwanger, A. Kugler, and M. Kugler (2013): "Trade and Market Selection: Evidence from Manufacturing Plants in Colombia," *Review of Economic Dynamics*, 16, 135–158. (Cited on pages 4 and 6.)

Eslava, M., J. Haltiwanger, and N. Urdaneta (2024): "The Size and Life-Cycle Growth of Plants: The Role of Productivity, Demand, and Wedges," *Review of Economic Studies*, 91, 259–300. (Cited on page 4.)

Fella, G., J. Galvez, B. Gonzalez, J. C. Ruiz-Garcia, and T. Senga (2022): "Non-Linear Productivity and Investment Dynamics," . (Cited on page 3.)

Foster, L., J. Haltiwanger, and C. Syverson (2008): "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?" *American Economic Review*, 98, 394–425. (Cited on pages 2, 4, 6, 7, 20, 22, and 33.)

Haltiwanger, J. (2016): "Firm Dynamics and Productivity: TFPQ, TFPR, and Demand-Side Factors," *Economía LACEA Journal*, 17. (Cited on page 4.)

Hsieh, C.-T. and P. J. Klenow (2009): "Misallocation and Manufacturing TFP in China and India*," *The Quarterly Journal of Economics*, 124, 1403–1448. (Cited on page 4.)

Jaimovich, N., S. J. Terry, and N. Vincent (2023): "The Empirical Distribution of Firm Dynamics and Its Macro Implications," . (Cited on page 3.)

Klette, T. J. and Z. Griliches (1996): "The Inconsistency of Common Scale Estimators When Output Prices Are Unobserved and Endogenous," *Journal of Applied Econometrics*, 11, 343–361. (Cited on page 4.)

Levinsohn, J. and A. Petrin (2003): "Estimating Production Functions Using Inputs to Control for Unobservables," *The Review of Economic Studies*, 70, 317–341. (Cited on pages 4 and 7.)

Melcangi, D. and S. Sarpietro (2024): "Nonlinear Firm Dynamics," . (Cited on page 3.)

Midrigan, V. and D. Y. Xu (2014): "Finance and Misallocation: Evidence from Plant-Level Data," *American Economic Review*, 104, 422–458. (Cited on page 3.)

Olley, G. S. and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64, 1263–1297. (Cited on pages 4, 7, and 8.)

Ruiz-Garcia, J. C. (2021): "Financial Frictions, Firm Dynamics and the Aggregate Economy:," . (Cited on page 3.)

Salgado, S., F. Guvenen, and N. Bloom (2019): "Skewed Business Cycles," Tech. Rep. w26565, National Bureau of Economic Research, Cambridge, MA. (Cited on page 3.)

Syverson, C. (2011): "What Determines Productivity?" *Journal of Economic Literature*, 49, 326–365. (Cited on page 3.)

Appendix A: A framework for TFPQ and TFPR

This appendix sets out the accounting framework behind the definitions used in Section 1. Nothing here is estimated; the assumptions that identify the latent objects are stated in Section 3.

TFPQ. Let Q_{it} denote the quantity produced by firm i in period t using capital K_{it} and labour L_{it} according to a Cobb–Douglas technology:

$$Q_{it} = Z_{it}^{Q} K_{it}^{\beta_K} L_{it}^{\beta_L}. \tag{A.1}$$

The term Z_{it}^{Q} captures the firm's physical productivity (TFPQ), that is, the efficiency with which inputs are converted into units of output. Z_{it}^{Q} is a latent variable whose identification rests on a set of assumptions laid out in Section 3.

Demand. Firms sell output in imperfectly competitive markets at price P_{it}, with quantities sold determined by a reduced-form demand relationship:

$$Q_{it} = f(P_{it}, D_{it}), \qquad \frac{\partial f}{\partial P} < 0, \tag{A.2}$$

where D_{it} is a firm- and time-specific demand advantage capturing factors such as product quality, brand reputation or customer base — a latent demand shifter. At this stage we impose no structural functional form on $f(\cdot)$, nor do we specify the dynamics of D_{it}.

TFPR. Revenue is $R_{it} = P_{it}Q_{it}$, so revenue productivity, termed TFPR following Foster et al. (2008), is

$$Z_{it}^{R} = P_{it} \cdot Z_{it}^{Q}.$$

TFPR is physical productivity scaled by the price at which output is sold, and hence an economic rather than a technological measure of performance. It is not a monotone transformation of TFPQ: a firm with lower Z^Q ranks higher in Z^R whenever its demand advantage is large enough. In logs,

$$\operatorname{Var}(\log Z^R) = \operatorname{Var}(\log Z^Q) + \operatorname{Var}(\log P) + 2\operatorname{Cov}(\log Z^Q, \log P), \tag{A.3}$$

so the dispersion of revenue productivity mixes technology, pricing, and the covariance between them. That covariance is what makes the two objects behave differently, and it is negative in our data.

REVRES. Absent price and quantity data, a researcher observes only R_{it} and estimates a revenue function,

$$R_{it} = \Omega_{it} K_{it}^{\alpha_K} L_{it}^{\alpha_L}, \qquad \Omega_{it} = \frac{R_{it}}{K_{it}^{\alpha_K} L_{it}^{\alpha_L}} = Z_{it}^{R} K_{it}^{\beta_K - \alpha_K} L_{it}^{\beta_L - \alpha_L}. \tag{A.4}$$

We term Ω_{it} the revenue function residual (REVRES). It coincides with TFPR only if the revenue elasticities α equal the output elasticities β, which requires demand to be perfectly elastic. Under imperfect competition the two wedges $\beta_K - \alpha_K$ and $\beta_L - \alpha_L$ are non-zero, so REVRES carries a capital and labour term of its own and cannot be mapped cleanly into Equation 1. This is why we keep the label distinct rather than calling revenue residuals TFPR.

Produced versus sold. Equation A.1 is about quantities *produced* while Equation A.2 is about quantities *sold*, and the two need not coincide within a period when inventories are a margin the firm uses. Our data record both, so the production function is estimated on output and the demand system on sales, rather than on a single quantity series standing in for both.

Remaining measurement difficulties. Even with output quantities and firm-level prices in hand, measuring TFPQ is not mechanical. Input heterogeneity — differences in the quality of materials or labour — and demand shifters, observed and unobserved, both affect the relationship between input use and units produced without reflecting technical efficiency. They must therefore be accounted for explicitly, which is what the estimation strategy of Section 3 does.

Appendix B: Data

B.1 Descriptive statistics: the Portuguese manufacturing sector

Table B.1 reports descriptive statistics for the Portuguese manufacturing sector as a whole over 2008–2022: every firm registered with at least one employee, measured before the merge with the survey data and before any of the restrictions that define our estimation sample. It is the benchmark against which the coverage of our sample is assessed in Section 2.2.

TABLE B.1. Summary statistics for the full sample of Portuguese manufacturing firms.

	Count	Min	P_{25}	P_{50}	P_{75}	Max	Mean	Std. Dev.
Age	413740	0.00	8.00	17.00	26.00	193.00	18.77	14.18
Number of Employees	1023077	1.00	1.00	1.00	6.00	5884.00	9.82	44.23
Turnover (in Thd.)	1023077	0.00	7.32	43.54	211.77	3347656.74	1012.08	13186.89
Capital (in Thd.)	577388	0.00	4.64	33.44	187.26	3523835.13	932.21	13712.76

Notes. All Portuguese manufacturing firms (NACE Rev. 2 section C) registered with at least one employee, 2008–2022, before the merge with the PRODCOM survey and before any of the filters that define the estimation sample. Turnover and capital are deflated and expressed in thousands of euros. Counts differ across rows because balance-sheet coverage is incomplete for the smallest registered firms.

B.2 Sample over time

Table B.2 describes the estimation sample on an annual basis and sets it against the Portuguese manufacturing sector as a whole. Coverage is not uniform over time: it rises sharply partway through the period, when the survey design changed, and from then on the sample accounts for roughly a fifth of manufacturing employment and a quarter of manufacturing value added.

TABLE B.2. Annual breakdown of the estimation sample and its contribution to Portuguese manufacturing.

Year	# Firms	Total # Firms	# Employees	Share, %	$\frac{\text{V.A.}}{\text{GDP}}$, %	V. A. to Manufact., %
2008	2244	79586	71419	9.75	1.63	13.66
2009	2185	75567	69216	10.16	1.59	14.26
2010	2161	70661	69119	10.53	1.61	13.88
2011	2101	69077	68066	10.56	1.64	14.48
2012	6655	66026	123458	20.10	2.76	24.21
2013	7123	64992	129637	21.47	3.00	25.97
2014	7729	64814	136787	22.19	3.27	27.73
2015	7757	65344	140127	22.09	3.30	27.20
2016	7941	65588	145218	22.36	3.41	28.03
2017	7876	66169	148751	22.12	3.35	27.11
2018	7813	66830	150600	21.67	3.32	27.07
2019	6723	67705	123893	17.45	2.58	21.68
2020	7480	65952	143227	20.38	3.10	25.68
2021	6643	66785	135334	19.03	3.07	24.64
2022	5955	67981	128897	17.77	2.88	23.11

Notes. '# Firms' and '# Employees' are the estimation sample. 'Total # Firms' is every manufacturing firm (NACE Rev. 2 section C) in the business register with at least one employee, and 'Share, %' is sample employees as a percentage of employment at those firms — both measured against manufacturing, not the whole economy. Value added is turnover net of raw materials, nominal, divided by nominal GDP and by nominal manufacturing gross value added respectively.

B.3 Sector overview

Table B.3 describes the five largest industries – sector-product-unit of measurement combinations – resulting from the classification procedure described in Section 2.

B.4 Data construction

The sample is built around PRODCOM (IAPI), which reports the physical quantity and the value of output at the firm-product level. Balance-sheet and further business statistics are merged onto it from administrative sources (QP and SCIE). Starting from 2,262,999 firm-product (8-digit PCC8)-year lines, the steps are as follows.

1. Quantities are reported in whatever unit each good is conventionally sold in, so we harmonise them into six physical families — mass, volume, energy, length, area and piece counts — each with a common unit. Dropping lines that report no revenue (only quantities) or vice-versa leaves 2,219,925 product lines.

2. We aggregate rows to firm-product (6-digit PCC8)-year level, so that closely related codes sold in the same unit are treated as a single good. At this stage, we compute revenue share on the firm's full product mix 1,059,677 lines remain.

3. We keep each firm's largest product by revenue share, and only where that share is at least 0.90. This restricts the sample to firms for which a single-product technology is a reasonable approximation, and leaves one observation per firm-year: 454,583 firm-years.

4. Matching these firms to their administrative data leaves 358,699 observations and confines the panel to 2008–2022. Capital is measured by gross fixed assets. Inputs are deflated with sector (NACE 2-digit)-year deflators provided by INE, and output prices both with and aggregate deflator (for cross-time comparison) and with the product-level price index described below (for cross-firm comparison).

5. We require every input entering the estimation — the wage bill, employment, capital, materials, output quantity and the output price — to be strictly positive. This is the most demanding restriction imposed, and 150,861 observations remain.

6. Since our goal is to study dynamics of productivity by incumbents, we impose that firms must be observed for at least 3 consecutive years, leaving 126,243 observations.

7. Since production functions are estimated separately by industry, we require each industry to have at least 75 observations per year. 89,762 observations remain.

8. Each industry must be observed in at least 5 distinct years, leaving 88,386 observations.

The resulting panel contains 88,386 firm-year observations on 11,398 firms in 23 industries, covering 2008–2022.

Output deflator . The product-level price index used above is a revenue-weighted geometric mean of firm unit values within each product-unit-year cell, constructed on all producers (with valid PRODCOM answers)rather than on the estimation sample. Market shares are computed on the same cross-section.

Table B.3. Sector overview: the five largest sectors

Product code	Description	Product composition	Obs	Firms	Unit(s)
107101	Bread and pastry	Bread (65%) and pastry products (35%).	20984	2674	kg
251201	Metal doors and windows	Doors, windows and their frames of cast iron, iron or steel.	12517	1655	pc
237001	Cut stone	Paving stones, kerbs and flagstones of natural stone, excluding slate.	8261	933	kg
152011	Footwear	Leather-upper footwear.	6946	812	pc
110201	Wine	Wines with protected designation of origin, chiefly still white wine.	5311	608	l

Notes. A sector is a product code paired with a unit of measurement of output. The five sectors shown are the largest by number of observations and together account for the majority of the estimation sample. Shares are given only where no single eight-digit product line reaches 80 per cent of sector revenue.

Appendix C: Estimation: Implementation and Identification

C.1 Production block

This appendix gives the implementation of the production block summarised in Section 3.1. Estimation runs industry by industry, in two stages.

C.1.1 The two stages

Stage 1. We regress output on a flexible function of inputs and demand controls,

$$q_{ist} = \Phi_s\left(l_{ist}, k_{ist}, m_{ist}, \tilde{d}_{ist-1}, ms_{ist}\right) + \delta_{st} + \delta_{rs} + \epsilon_{ist},$$

where Φ_s is a translog in $(l_{ist}, k_{ist}, m_{ist})$ (linear, squared and pairwise interaction terms) plus linear and squared terms in ms_{ist} and $\tilde{d}_{ist-1}$, and δ_{st} and δ_{rs} are year and NUTS-2 region effects. Here ms_{ist} is the firm's market share demeaned within industry-year, and $\tilde{d}_{ist-1}$ is the lagged transient demand advantage: the demand advantage of Section 3.2 net of the firm's mean over the sample. Only the transient component enters because the firm mean is close to a firm fixed effect in output and would absorb the between-firm variation that identifies the elasticities. Because that mean is computed over all years, $\tilde{d}_{ist-1}$ uses information beyond Ω_{ist} in Assumption 1; we accept this to keep the permanent component of demand out of the control function. In the initialisation of the iterative procedure (Section 3.3) no demand advantage is available yet, and the first stage conditions on market share only.

The year effects shift Φ_s additively, so the first stage is industry-specific but not year-specific. As in Ackerberg et al. (2015), the β coefficients are not identified at this stage, since they are subsumed into Φ_s. The first stage instead cleanses output of measurement error,

$$\mathbb{E}\left[q_{ist} - \Phi_s(\cdot) - \delta_{st} - \delta_{rs} \mid l_{ist}, k_{ist}, m_{ist}, \tilde{d}_{ist-1}, ms_{ist}\right] = 0,$$

and we denote the fitted values by $\widehat{\Phi}_{ist}$.

Stage 2. For a candidate $\beta \equiv (\beta_s^0, \beta_s^L, \beta_s^K, \beta_s^M)$, trial productivity is

$$\hat{z}_{ist}^q(\beta) = \widehat{\Phi}_{ist} - \beta_s^0 - \beta_s^L l_{ist} - \beta_s^K k_{ist} - \beta_s^M m_{ist}.$$

Step 1. For each β we regress $\hat{z}_{ist}^q(\beta)$ by OLS on a constant and on $\hat{z}_{ist-1}^q(\beta)$, its square and its cube, where the lag is built from $\widehat{\Phi}_{ist-1}$ and lagged inputs, and keep the residual $\tilde{\xi}_{ist}(\beta)$. This concentrates out $\tilde{g}(\cdot)$, as in Ackerberg et al. (2015).

Step 2. Under Assumptions 1–3 the innovation is orthogonal to capital and to lagged flexible

inputs. We use

$$\mathbb{E}\left[\tilde{\xi}_{ist}(\beta)\, Z_{ist}\right] = 0,$$
$$Z_{ist} = \left(1,\ l_{ist-1},\ m_{ist-1},\ k_{ist},\ k_{ist}^2,\ l_{ist}k_{ist},\ l_{ist-1}^2,\ m_{ist-1}^2,\ k_{ist-1}m_{ist-1}\right)',$$

and choose β to minimise $\left\| N_s^{-1} \sum_{i,t} \tilde{\xi}_{ist}(\beta)\, Z_{ist} \right\|^2$, where N_s is the number of observations in industry s. The interaction $l_{ist}k_{ist}$ is a valid instrument only if labour is chosen before the innovation is realised ($b = 1$ in Assumption 3). We minimise with a bounded quasi-Newton routine (L-BFGS-B) constraining each elasticity to $(0, 1)$, start it from the OLS estimates and from a grid of perturbations around them, and keep the solution with the lowest criterion. Because the intercept of $\tilde{g}(\cdot)$ absorbs any constant, β_s^0 is not separately identified; this is immaterial, since we only use TFPQ demeaned within industry-year.

Step 3. With the second-stage estimates, firm-level productivity is

$$\hat{z}_{ist}^q = \widehat{\Phi}_{ist} - \hat{\beta}_s^0 - \hat{\beta}_s^L l_{ist} - \hat{\beta}_s^K k_{ist} - \hat{\beta}_s^M m_{ist}.$$

C.1.2 Why the cubic projection does not restrict the process

Assumption 2 leaves the law of motion of productivity unrestricted beyond first-order Markov, while the second stage projects on a cubic. The two are consistent, because $\tilde{\xi}_{ist}$ is not the structural TFPQ innovation. Writing $\mu(\cdot) \equiv \mathbb{E}[\hat{z}_{ist}^q \mid \hat{z}_{ist-1}^q]$ for the true conditional mean,

$$\tilde{\xi}_{ist} = \underbrace{\left[\mu\left(\hat{z}_{ist-1}^q\right) - \tilde{g}\left(\hat{z}_{ist-1}^q\right)\right]}_{\text{misspecification gap}} + \underbrace{\left[\hat{z}_{ist}^q - \mu\left(\hat{z}_{ist-1}^q\right)\right]}_{\text{true innovation}}.$$

Every instrument in Z_{ist} is a function of variables dated $t-1$ or earlier, and hence measurable with respect to $\sigma(\hat{z}_{ist-1}^q)$; the second term is orthogonal to it by construction. The moment condition therefore holds if and only if the misspecification gap is uncorrelated with Z_{ist}, which requires only that $\tilde{g}$ span μ — not that productivity be cubic, homoskedastic or Gaussian. Distributional features of the true innovation propagate into $\tilde{\xi}_{ist}$ transformed, but only through its higher moments, which the moment condition does not use.

The three processes of Section 3.3.1 illustrate the three cases. Under a Gaussian AR(1), μ is linear, a linear $\tilde{g}$ matches it exactly, the gap is zero and $\tilde{\xi}_{ist}$ coincides with the structural innovation. Under a quadratic-mean location-scale process, μ is quadratic and a quadratic $\tilde{g}$ absorbs it, so the gap is again zero; the heteroskedastic loading has zero conditional mean and inflates only the higher moments of $\tilde{\xi}_{ist}$, at a cost in efficiency. Under the Hermite quantile process of Arellano and Bonhomme (2016), μ remains polynomial in the lag — its coefficients are expectations taken over the rank of the shock — so a polynomial $\tilde{g}$ of matching order absorbs it asymptotically, while shock-dependent persistence again appears only as richer heteroskedasticity.

The qualification is therefore finite-sample rather than asymptotic: when μ is of higher order than the polynomial, $\tilde{g}$ absorbs it only as the sample grows, and an under-parameterised $\tilde{g}$ leaves a bias in β. This is what Section 3.3.1 quantifies, and it is why the hardest of the three cases is the Hermite process. In a nutshell, the moment condition tolerates everything that lives in the higher moments of $\tilde{\xi}_{ist}$; the only real threat is misspecifying the conditional mean, which polynomial flexibility addresses directly.

C.2 Demand block

This appendix gives the moment conditions and the numerical detail behind Section 3.2. Estimation runs industry by industry.

Moment conditions. For a candidate $\theta_s = (\alpha_{p,s}, \alpha_{pm,s}, \rho_{0,s}, \rho_{1,s}, \rho_{2,s})$, strip the price terms from the log quantity sold,

$$\tilde{y}_{ist}(\theta_s) \;=\; y_{ist} - \alpha_{p,s}\, p_{ist} - \alpha_{pm,s}\, p_{ist} \ln MS_{ist},$$

so that $\tilde{y}_{ist}$ is the demand advantage implied by θ_s. Its innovation under the law of motion of Assumption 2 is

$$\xi^d_{ist}(\theta_s) \;=\; \tilde{y}_{ist} - \rho_{0,s} - \rho_{1,s}\, \tilde{y}_{ist-1} - \rho_{2,s}\, \tilde{y}^2_{ist-1},$$

and the estimator solves $\mathbb{E}\left[\xi^d_{ist}(\theta_s)\, \Xi_{ist}\right] = 0$ with the nine instruments

$$\Xi_{ist} = \Big\{ \underbrace{1}_{\text{constant}} ,\; \underbrace{\hat{z}^q_{ist},\; \hat{z}^q_{ist} \ln MS_{ist}}_{\text{supply-side}},\; \underbrace{y_{ist-1},\, p_{ist-1},\, y_{ist-2},\, p_{ist-2},\, k_{ist-1},\, k_{ist-2}}_{\text{predetermined}} \Big\}.$$

Five parameters against nine moments leaves the system over-identified by four. The supply-side pair does the work on the price elasticities: $\hat{z}^q_{ist}$ identifies $\alpha_{p,s}$ and its interaction with the market share identifies $\alpha_{pm,s}$, which is why a constant-elasticity specification needs only the first. The predetermined block is dated $t-1$ and earlier and is valid by Assumptions 1 and 2. Both lags of each variable are always included, so the moment count does not change if the instrument set is pushed one period deeper.[20]

Numerical implementation. The nine moments are weighted equally, by an identity weighting matrix: the alternative two-step optimal weights would be estimated industry by industry on samples of a few thousand observations, where the weight matrix is itself noisy. The criterion is minimised by a bounded quasi-Newton routine (L-BFGS-B) with $\alpha_{p,s}$ constrained above at -0.05, which rules out upward-sloping demand; the remaining parameters are unconstrained. Starting values

[20] Measurement error in sold quantities makes the demand innovation an MA(1) process, which in principle invalidates instruments dated $t-1$. Shifting the whole block to $\{t-2, t-3\}$ leaves the moment count unchanged but requires four consecutive firm-years per observation rather than three, at a cost of 21–41 per cent of the estimation sample by industry. We retain the $t-1$ specification and report the deeper one as a robustness check.

are the previous iteration's elasticities where available, and otherwise $(\alpha_{p,s}, \alpha_{pm,s}, \rho_{0,s}, \rho_{1,s}, \rho_{2,s}) = (\hat{\alpha}^{\mathrm{OLS}}_{p,s}, 0, 0, 0.5, 0)$. An industry is dropped from the demand block if the GMM fails to converge or returns $\alpha_{p,s}$ at the boundary.

From the GMM to the demand control. Only $(\hat{\alpha}_{p,s}, \hat{\alpha}_{pm,s})$ are retained; the ρ's enter the moment conditions but are re-estimated afterwards, for the reason given in Section 3.2: in levels the law of motion confounds a permanent firm component of demand with genuine persistence and drives $\rho_{1,s}$ towards one. Splitting $\hat{D}_{ist}$ into a firm mean and a transient component and re-estimating on the latter by OLS, with $|\rho_{1,s}| \leq 0.99$, separates the two. It is the lagged transient component that enters the production block as a control.

C.3 Industry-by-industry estimates

Table C.1 collects the parameters behind every estimate reported in the paper, one row per industry, identified by its six-digit PRODCOM product code and a short name. The production block is stable across industries: materials carry most of the output elasticity, capital little, and returns to scale cluster below one. The demand block is noisier, as one would expect of a system identified off within-industry price variation in samples of a few hundred to a few thousand firm-years, and three industries return a price coefficient on the admissible bound; those are excluded from the demand-side results but retain their production-function estimates.

C.4 Within-industry correlations

Table C.2 reports, for the industries discussed in Section 5.1, the within-industry correlations between price, physical productivity and the estimated demand advantage. The within-industry price–productivity correlation is stronger than the pooled one in almost every case: pooling across industries with different price levels and different productivity dispersion attenuates the relationship the supply-side argument is about.

TABLE C.1. Production and demand parameters, industry by industry

				Production block				Demand block	
Code	Industry	N	Conv.	β_l	β_m	β_k	RTS	α_p	First-stage F
107101_mass_kg	Bread and pastry	21,036	Yes	0.499	0.489	0.028	1.017	-1.93	2293.6
251201_count_pc	Metal doors and windows	12,754	Yes	0.320	0.295	0.059	0.674	-1.95	26.1
237001_mass_kg	Cut stone	8,339	Yes	0.410	0.538	0.089	1.037	-3.07	1926.0
152011_count_pc	Footwear	6,974	Yes	0.574	0.498	0.040	1.112	-0.05†	521.6
110201_volume_l	Wine	5,360	Yes	0.274	0.727	0.007	1.009	-1.90	1054.3
162902_mass_kg	Cork products	4,589	Yes	0.085	0.629	0.079	0.793	-11.03	493.2
257345_count_pc	Moulds and patterns	4,537	Yes	0.089	0.392	0.029	0.510	-7.72	68.9
181201_mass_kg	Printing	4,403	Yes	0.358	0.564	0.028	0.951	-1.88	38.4
251102_mass_kg	Metal structures	3,830	Yes	0.247	0.753	0.134	1.133	-0.05†	98.5
310901_count_pc	Wooden furniture	2,521	No	0.099	0.836	-0.043	0.892	-6.24	584.1
310201_count_pc	Kitchen furniture	2,045	Yes	0.078	0.893	0.042	1.013	-2.72	137.3
101301_mass_kg	Processed meat	1,894	Yes	0.322	0.679	0.039	1.040	-0.05†	65.7
234101_mass_kg	Ceramic tableware	1,427	Yes	0.557	0.432	0.046	1.035	-2.29	202.8
172121_mass_kg	Carton packaging	1,258	No	0.089	0.894	-0.021	0.962	-1.97	46.0
105104_mass_kg	Fresh cheese	1,225	Yes	0.128	0.834	0.034	0.995	-1.97	340.6
236101_mass_kg	Precast concrete	1,184	Yes	0.067	0.718	0.073	0.858	-1.64	42.6
139101_mass_kg	Knitted fabrics	1,152	No	0.034	0.907	-0.015	0.926	-1.81	38.3
143101_count_pc	Hosiery	1,077	Yes	0.216	0.735	0.019	0.970	-1.74	242.4
143901_count_pc	Knitwear	900	Yes	0.228	0.631	0.018	0.877	-1.78	40.0
289923_count_pc	Mixing machinery	786	Yes	0.213	0.452	0.055	0.721	-1.06	3.1
329965_mass_kg	Candles	713	Yes	0.159	0.689	0.076	0.924	-1.69	2.7
132002_area_m2	Cotton fabrics	677	Yes	0.140	0.775	0.022	0.937	-6.26	33.8
222201_mass_kg	Plastic packaging	599	Yes	0.110	0.698	0.054	0.862	-1.90	35.2

Notes. Estimates from the joint fixed point of Section 3.1 and Section 3.2, one industry per row, ordered by sample size. Industries are identified by their six-digit PRODCOM product code and the physical unit in which quantities are measured. β_l, β_m and β_k are output elasticities and RTS their sum; Conv. records whether the iteration between the two blocks reached its fixed point before the iteration cap. α_p is the price coefficient of the demand system and First-stage F that of the demand instruments. † marks industries where α_p sits on the admissible bound of -0.05 and the demand block is therefore not identified; their production-side estimates are retained.

TABLE C.2. Within-industry correlations between price, physical productivity and demand advantage

Code	Industry	N	$r(P, z^q)$	$r(P, D)$	$r(z^q, D)$
236101_mass_kg	Precast concrete	1,174	-0.349	0.544	-0.041
237001_mass_kg	Cut stone	8,261	-0.316	0.702	-0.161
110201_volume_l	Wine	5,311	-0.258	0.209	-0.171
310901_count_pc	Wooden furniture	2,476	-0.257	0.958	-0.213
162902_mass_kg	Cork products	4,552	-0.226	0.978	-0.176
132002_area_m2	Cotton fabrics	674	-0.140	0.670	0.054
105104_mass_kg	Fresh cheese	1,220	-0.118	0.051	-0.054
310201_count_pc	Kitchen furniture	2,026	0.286	0.781	0.189
	Pooled sample	88,386	-0.110	0.698	0.045

Notes. The industries discussed in Section 5.1; the pooled row covers the full sample. All three variables are demeaned by industry and year, so each correlation is a within-industry-year object. D is the estimated demand advantage; because it is recovered as $y - \hat{\alpha}_p p - \hat{\alpha}_{pm} p \ln MS$, price enters its construction directly and $r(P, D)$ is therefore in part mechanical. Correlations are computed on complete cases and reported without standard errors. Three industries are absent because α_p sits on the admissible bound, which leaves the demand advantage undefined for them.

Appendix D: Monte Carlo Simulation

D.1 The simulated economy

There is a single industry, so the industry index s is dropped. We simulate $N = 2{,}500$ firms for 100 periods and keep the last ten; the lag of the demand shifter removes the first of these, leaving 22,500 firm-year observations. $\Phi(\cdot)$ denotes the standard normal c.d.f. and lowercase letters denote logs.

Technology and demand. Physical output is Cobb–Douglas in labour, capital and materials,

$$q_{it} = \beta^0 + \beta^L l_{it} + \beta^K k_{it} + \beta^M m_{it} + z^q_{it} + \epsilon_{it}, \qquad \epsilon_{it} \sim N(0, \sigma^2_\epsilon), \tag{D.1}$$

where ϵ_{it} is classical measurement error in output. The price is read off the inverse of the demand curve of Section 3.2, with $q^{\text{sold}}_{it} = q_{it}$,

$$q_{it} = \alpha_p\, p_{it} + D_{it} \qquad \Longleftrightarrow \qquad p_{it} = \frac{q_{it} - D_{it}}{\alpha_p} + \epsilon^p_{it}, \qquad \epsilon^p_{it} \sim N(0, \sigma^2_p). \tag{D.2}$$

Latent states. The demand advantage follows the quadratic Markov process of Section 3.2,

$$D_{it} = \rho_0 + \rho_1 D_{i,t-1} + \rho_2 D^2_{i,t-1} + \xi^d_{it}, \qquad \sigma_{\xi^d} = \sqrt{\left(1 - \rho_1^2\right)\sigma^2_D}, \tag{D.3}$$

and z^q_{it} follows one of the seven processes of Section D.2. The two innovations are drawn as

$$\xi^d_{it} = \sigma_{\xi^d}\left[\sigma^{\text{cross}}\,\frac{\xi^q_{it}}{\sigma_{\xi^q}} + \sqrt{1 - \left(\sigma^{\text{cross}}\right)^2}\,\nu_{it}\right], \qquad \nu_{it} \sim N(0,1), \tag{D.4}$$

so that $\sigma^{\text{cross}} = \text{corr}\left(\xi^q_{it}, \xi^d_{it}\right)$. The baseline sets $\sigma^{\text{cross}} = 0$; Section D.5 relaxes it.

Input prices. Wages and materials prices are aggregate AR(1) processes common to all firms,

$$\ln w_t = (1 - \rho_w)\,\bar{w} + \rho_w \ln w_{t-1} + \varsigma^w_t, \qquad \ln p^M_t = (1 - \rho_m)\,\bar{p}^M + \rho_m \ln p^M_{t-1} + \varsigma^m_t, \tag{D.5}$$

where $\bar{w}$ and $\bar{p}^M$ are the long-run means.

Input choices. Writing $\sigma \equiv (\alpha_p + 1)/\alpha_p$ and $\gamma \equiv \sigma\beta^M$, revenue is $R_{it} = Y^\sigma_{it}\exp(D_{it}/|\alpha_p|)$. Materials solve the static first-order condition $\sigma\beta^M R_{it}/M_{it} = p^M_t$ after z^q_{it} is observed, with the not-yet-realised D_{it} replaced by its forecast $\rho_1 D_{i,t-1}$. Labour solves $\sigma\beta^L R_{it}/L_{it} = w_t$ at $t-b$, with both states replaced by their $t-b$ forecasts and a Jensen correction for the productivity innovation still to realise. Optimisation errors $u^m_{it} \sim N\left(0, \sigma^2_{u^m}\right)$ and $u^l_{it} \sim N\left(0, \sigma^2_{u^l}\right)$ are added to the implied

log inputs.

Capital. Capital accumulates as $K_{it} = (1-\delta)K_{i,t-1} + I_{i,t-1}$. Investment follows the closed-form policy of the firm's dynamic problem when productivity is AR(1) and demand does not move — increasing in current productivity, decreasing in current wages — discounted at $\beta_{\mathcal{D}}$ and scaled by a firm-specific adjustment cost b_{it}, whose inverse is log-AR(1), $\ln b_{it}^{-1} = \phi_\rho \ln b_{i,t-1}^{-1} + \phi_\sigma \varsigma_{it}^b$ with $\varsigma_{it}^b \sim N(0,1)$. We use this policy for every process: it forecasts productivity with AR(1) persistence equal to the process's average linear persistence and ignores the demand advantage, so in the simulated economy investment approximates, rather than solves, the dynamic problem.

Calibration. Table D.1 collects the parameter values, which are common to all seven productivity processes.

TABLE D.1. Monte Carlo: parameters of the simulated economy

Technology and demand			*Shocks and errors*		
β^L	Labour elasticity	0.30	σ_z	S.d. of z_{it}^q (stationary)	0.30
β^M	Materials elasticity	0.50	σ_D	S.d. of D_{it} (stationary)	0.30
β^K	Capital elasticity	0.20	σ_ϵ	Output measurement error	0.05
β^0	Intercept	0	σ_p	Price measurement error	0.01
α_p	Demand elasticity	−2.50	σ_{u^l}	Labour optimisation error	0.30
σ	$(\alpha_p+1)/\alpha_p$	0.60	σ_{u^m}	Materials optimisation error	0.25
γ	$\sigma\beta^M$	0.30	σ^{cross}	corr(ξ_{it}^q, ξ_{it}^d)	0
Demand process			*Input prices*		
ρ_0	Constant	0	$\bar{w} = \bar{p}^M$	Long-run mean	0.50
ρ_1	Linear term	0.60	$\rho_w = \rho_m$	Persistence	0.87
ρ_2	Quadratic term	0	$\sigma_{\varsigma^w} = \sigma_{\varsigma^m}$	Innovation s.d.	$\sqrt{0.06}$
Capital and timing			*Panel*		
δ	Depreciation	0.20	N	Firms	2,500
$\beta_{\mathcal{D}}$	Discount factor	0.95	T	Retained periods	10
ϕ_ρ	Adjustment-cost persistence	0.90	—	Burn-in periods	90
ϕ_σ	Adjustment-cost s.d.	0.30	—	Seed	123
b	Labour timing	0.50			

Notes. Parameters of the data-generating process of Section D.1. All seven productivity processes of Section D.2 share these values; process-specific parameters are reported in Section D.2. One replication per process, from a single seed.

D.2 The menu of productivity processes

Each specification below is a first-order Markov process for z_{it}^q satisfying Assumption 2. They differ in whether persistence, $\varrho\left(z_{i,t-1}^q, \xi_{it}^q\right) \equiv \partial z_{it}^q / \partial z_{i,t-1}^q$, is constant, varies with the firm's position in the productivity distribution, varies with the rank of the shock it receives, or both. In DGP1,

DGP2 and DGP4 the innovations are scaled so that the stationary standard deviation of z_{it}^q equals σ_z. DGP3 applies the scaling of DGP1 at its average persistence and DGP6 that of DGP4, so both match σ_z only approximately; in DGP5 and DGP7 the dispersion follows from the coefficients. Realisations of DGP5–DGP7 are clipped to $[-5, 5]$ to prevent explosive paths.

DGP1 — AR(1), Gaussian.

$$z_{it}^q = \rho\, z_{i,t-1}^q + \xi_{it}^q, \qquad \xi_{it}^q \sim N\big(0, \sigma_{\xi^q}^2\big)\,, \qquad \sigma_{\xi^q} = \sqrt{(1-\rho^2)\,\sigma_z^2}, \qquad \varrho = \rho, \tag{D.6}$$

with $\rho = 0.7$. Persistence is flat: this is the linear benchmark that the textbook estimator is designed for.

DGP2 — AR(1), skewed innovations. As Equation D.6, with ξ_{it}^q drawn from a skew-normal with shape $\varkappa = 5$, centred and rescaled to mean zero and standard deviation σ_{ξ^q}:

$$\xi_{it}^q \;\propto\; \varsigma\,|\zeta_{1,it}| + \sqrt{1-\varsigma^2}\,\zeta_{2,it} - \varsigma\sqrt{2/\pi}, \qquad \varsigma = \frac{\varkappa}{\sqrt{1+\varkappa^2}}, \qquad \zeta_{1,it}, \zeta_{2,it} \sim N(0,1). \tag{D.7}$$

Persistence is still flat at ρ; only the shape of the innovation distribution changes.

DGP3 — Decile-dependent persistence. Firms are ranked by $z_{i,t-1}^q$ each period and assigned to a decile $d_{i,t-1} \in \{1, \ldots, 10\}$:

$$z_{it}^q = \rho(d_{i,t-1})\, z_{i,t-1}^q + \xi_{it}^q, \qquad \varrho = \rho(d_{i,t-1})\,, \tag{D.8}$$

with the hump-shaped profile $\{\rho(d)\}_{d=1}^{10} = (0.50,\, 0.55,\, 0.65,\, 0.75,\, 0.85,\, 0.85,\, 0.75,\, 0.65,\, 0.55,\, 0.50)$. Persistence is high in the middle of the distribution and low in both tails, but does not depend on the shock.

DGP4 — Location-scale.

$$z_{it}^q = \gamma_0 z_{i,t-1}^q + \Big(\gamma_1 + \gamma_2 z_{i,t-1}^q\Big)\,\varepsilon_{it}, \qquad \sigma_\varepsilon^2 = \frac{\sigma_z^2\,(1-\gamma_0^2)}{\gamma_1^2 + \gamma_2^2 \sigma_z^2}, \qquad \varrho = \gamma_0 + \gamma_2 \varepsilon_{it}, \tag{D.9}$$

with $(\gamma_0, \gamma_1, \gamma_2) = (0.70,\, 1.00,\, 0.40)$ and the scale factor floored at 0.01. Persistence now varies with the shock but not with the firm's position.

DGP5 — Quantile polynomial. With $u_{it} = \Phi(\xi_{it}^q / \sigma_{\xi^q})$ the rank of the shock,

$$z_{it}^q = a(u_{it}) + b_1(u_{it})\, z_{i,t-1}^q + b_2(u_{it})\,\Big(z_{i,t-1}^q\Big)^2, \qquad c(u) = c^m + c^s\,\Phi^{-1}(u) \quad \text{for } c \in \{a, b_1, b_2\}\,, \tag{D.10}$$

with $(a^m, a^s) = (0,\, 0.10)$, $(b_1^m, b_1^s) = (0.70,\, 0.30)$, $(b_2^m, b_2^s) = (0,\, -0.05)$, so that $\varrho = b_1(u_{it}) + 2b_2(u_{it})\, z^q_{i,t-1}$. Persistence varies in both position and shock rank.

DGP6 — Quadratic-mean location-scale. Equation D.9 augmented with a quadratic drift,

$$z^q_{it} = \gamma_0 z^q_{i,t-1} + \gamma_3 \left(z^q_{i,t-1}\right)^2 + \left(\gamma_1 + \gamma_2 z^q_{i,t-1}\right) \varepsilon_{it}, \qquad \varrho = \gamma_0 + 2\gamma_3 z^q_{i,t-1} + \gamma_2 \varepsilon_{it}, \tag{D.11}$$

with $\gamma_3 = 0.15$ and the remaining parameters as in DGP4.

DGP7 — Hermite quantile process. The estimating equation of Section 4 used as a data-generating process: with $\tilde{z}_{i,t-1} = \left(z^q_{i,t-1} - \bar{z}\right)/s_z$ standardised by the mean and standard deviation of all realisations simulated up to $t-1$,

$$Q_\tau\left(z^q_{it} \mid z^q_{i,t-1}\right) = \sum_{k=0}^{3} a_k(\tau)\, H_k(\tilde{z}_{i,t-1})\,, \qquad \tau_j = \frac{j}{J+1},\;\; j = 1, \ldots, J,\;\; J = 8, \tag{D.12}$$

where H_k are the probabilists' Hermite polynomials, the shock rank is $\tau = \Phi(\xi^q_{it}/\sigma_{\xi^q})$, and the $4 \times J$ coefficient matrix is interpolated linearly between knots and extrapolated linearly from the two nearest knots outside them. The coefficients impose a hump-shaped persistence profile with mild antisymmetric curvature:

$$\begin{pmatrix} a_0(\tau_j) \\ a_1(\tau_j) \\ a_2(\tau_j) \\ a_3(\tau_j) \end{pmatrix} = \begin{pmatrix} -0.40 & -0.25 & -0.15 & -0.05 & 0.05 & 0.15 & 0.25 & 0.40 \\ 0.55 & 0.62 & 0.68 & 0.72 & 0.72 & 0.68 & 0.62 & 0.55 \\ -0.05 & -0.03 & -0.01 & 0.00 & 0.00 & 0.01 & 0.03 & 0.05 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{pmatrix}. \tag{D.13}$$

D.3 Estimation on simulated data

We apply the procedure of Section 3 to each simulated panel, treating it as a single industry. The simulation code shares the second stage of Section 3.1 — the cubic conditional mean, the nine instruments and the multi-start minimisation — but departs from the estimator we apply to the Portuguese data in the following respects.

First stage. At iteration 1 the first stage is a second-order polynomial in (l_{it}, k_{it}, m_{it}) interacted with year dummies, with no demand controls. From iteration 2 onwards it is

$$\begin{aligned} \Phi_t = \Big[&\mathrm{poly}_2(l_{it}, k_{it}, m_{it}) + l_{it}^3 + k_{it}^3 + m_{it}^3 + \widehat{D}_{i,t-1} + \widehat{D}_{i,t-1}^2 \\ &+ p_{it} + ms_{it} + p_{it} ms_{it} + p_{it}^2 + ms_{it}^2 + \mathrm{region}_i \Big] \times \mathrm{year}_t, \end{aligned} \tag{D.14}$$

where ms_{it} is the demeaned market share and region_i are simulated region dummies. Relative to Section 3.1, this first stage adds the own cubes of the inputs and the price, interacts all terms with year dummies, and uses the demand advantage in levels rather than net of its firm mean.

Demand block. Demand is CES ($\alpha_{pm} = 0$), and the quantity sold equals the quantity produced. The parameters $(\alpha_p, \rho_0, \rho_1, \rho_2)$ are estimated jointly by GMM with an identity weighting matrix and instrument vector

$$\Xi_{it} = \{1,\ \widehat{z}^q_{it},\ q_{i,t-2},\ q_{i,t-3},\ p_{i,t-2},\ p_{i,t-3},\ k_{i,t-1},\ k_{i,t-2}\}\,, \tag{D.15}$$

where $\widehat{z}^q_{it} = \widehat{\Phi}_{it} - \hat{\beta}^0 - \hat{\beta}^L l_{it} - \hat{\beta}^K k_{it} - \hat{\beta}^M m_{it}$ is the productivity from the latest production step, neither year-demeaned nor computed from raw output. Relative to Section 3.2, the simulation uses the second and third lags of quantities and prices rather than the first and second, keeps the GMM estimates of (ρ_0, ρ_1, ρ_2), and feeds $\widehat{D}_{it} = q_{it} - \hat{\alpha}_p p_{it}$ back into Equation D.14 through its lag.

Convergence and inference. The demand elasticity is replaced by its new GMM estimate in every iteration. The stopping rule is that of Section 3.3 — maximum and Euclidean norm of the changes in $(\beta^L, \beta^M, \beta^K)$ below $\varepsilon = 0.03$ — with the side conditions $\beta^K > 0.011$ and $\beta^M > 0.02$ and a cap of 100 iterations; each iteration restarts the second-stage optimiser from values around the OLS estimates. All seven processes converge at the third iteration. Standard errors come from a panel block bootstrap: firms are resampled with replacement, preserving each firm's time series, and the production block is re-estimated on each of 30 resamples at the converged demand elasticity and demand advantages. The standard errors therefore omit the sampling variability of the demand block.

D.4 Results

Table D.2 reports the full set of estimates. Two patterns hold across the menu. First, the initialisation is biased in every specification: β^M is over-estimated by 0.12–0.32 and β^L misses in both directions, by up to 0.17. Second, the converged estimator removes most of this bias: across the seven processes $\beta^L \in [0.270, 0.316]$, $\beta^M \in [0.461, 0.497]$, $\beta^K \in [0.199, 0.248]$ and $\alpha_p \in [-2.53, -2.17]$, against true values of 0.30, 0.50, 0.20 and -2.50. Some residual gaps exceed twice the bootstrap standard error, most clearly β^M under DGP5 (0.461, standard error 0.009). The largest gaps arise under DGP7, the Hermite quantile process, where β^L, β^K and α_p are all furthest from the truth — the case in which the cubic conditional mean is only an approximation to the true quantile-dependent dynamics. The estimated Markov coefficients (g_1, g_2, g_3) have no single true counterpart and change across processes with the curvature of the true conditional mean. All estimates come from a single simulated panel per process.

TABLE D.2. Monte Carlo: production and demand estimates across data-generating processes

		Structural parameters				Estimated Markov polynomial			
DGP	Stage	β^L (0.30)	β^M (0.50)	β^K (0.20)	α_p (−2.50)	g_0	g_1	g_2	g_3
DGP1: AR(1), Gaussian	Iter. 1	0.216	0.620	0.215	—	−0.515	0.459	−0.068	−0.024
		(0.012)	(0.025)	(0.007)			(0.286)	(0.283)	(0.092)
	Final	**0.300**	**0.497**	**0.203**	**−2.436**	−0.547	0.222	−0.372	−0.097
		(0.010)	(0.008)	(0.009)			(0.066)	(0.061)	(0.018)
DGP2: AR(1), skewed	Iter. 1	0.465	0.694	0.134	—	0.152	1.196	−0.978	0.246
		(0.069)	(0.061)	(0.021)			(0.275)	(0.530)	(0.284)
	Final	**0.291**	**0.495**	**0.201**	**−2.421**	−0.641	0.231	−0.343	−0.084
		(0.005)	(0.009)	(0.009)			(0.096)	(0.078)	(0.021)
DGP3: Decile-dependent ρ	Iter. 1	0.191	0.699	0.190	—	−0.482	−0.019	−0.662	−0.313
		(0.064)	(0.047)	(0.023)			(0.202)	(1.125)	(2.234)
	Final	**0.304**	**0.493**	**0.199**	**−2.426**	−0.645	−0.125	−0.679	−0.226
		(0.018)	(0.011)	(0.006)			(0.075)	(0.078)	(0.029)
DGP4: Location-scale	Iter. 1	0.287	0.617	0.193	—	−0.212	0.522	0.241	0.243
		(0.005)	(0.023)	(0.010)			(0.059)	(0.108)	(0.071)
	Final	**0.301**	**0.490**	**0.203**	**−2.411**	−0.315	0.552	−0.138	−0.046
		(0.005)	(0.005)	(0.011)			(0.050)	(0.066)	(0.027)
DGP5: Quantile polynomial	Iter. 1	0.254	0.824	0.145	—	0.711	−1.229	2.060	−0.886
		(0.030)	(0.186)	(0.035)			(0.750)	(0.850)	(0.515)
	Final	**0.316**	**0.461**	**0.206**	**−2.451**	−0.211	0.738	−0.127	−0.188
		(0.007)	(0.009)	(0.004)			(0.055)	(0.085)	(0.049)
DGP6: Quadratic-mean loc.-scale	Iter. 1	0.282	0.621	0.204	—	0.101	1.331	0.011	−0.067
		(0.006)	(0.020)	(0.011)			(0.017)	(0.008)	(0.004)
	Final	**0.308**	**0.473**	**0.220**	**−2.526**	0.107	1.272	−0.013	−0.042
		(0.008)	(0.010)	(0.009)			(0.014)	(0.006)	(0.002)
DGP7: Hermite quantile	Iter. 1	0.211	0.694	0.289	—	−0.343	1.039	0.082	0.009
		(0.018)	(0.039)	(0.022)			(0.085)	(0.029)	(0.003)
	Final	**0.270**	**0.497**	**0.248**	**−2.166**	−0.458	0.820	−0.021	−0.002
		(0.014)	(0.016)	(0.037)			(0.057)	(0.020)	(0.002)

Notes. True values in parentheses in the column headers. *Iter. 1* is the control-function estimator without demand controls; *Final* is the converged estimate of Section 3 (see Section D.3). Panel block bootstrap standard errors in parentheses below each estimate, from 30 resamples of firms drawn with replacement; each resample re-estimates the production block at the converged demand elasticity, so no standard error is reported for α_p. α_p is not estimated at iteration 1, which conditions on no demand information; no standard error is reported for the concentrated intercept g_0. g_1, g_2, g_3 are the coefficients of the cubic Markov projection of Section D.3 and have no single true counterpart: they are the cubic approximation to the conditional mean implied by each process in Section D.2. One panel of $N = 2{,}500$ firms and $T = 10$ years (22,500 firm-year observations) per process; all seven converge in three iterations at $\varepsilon = 0.03$.

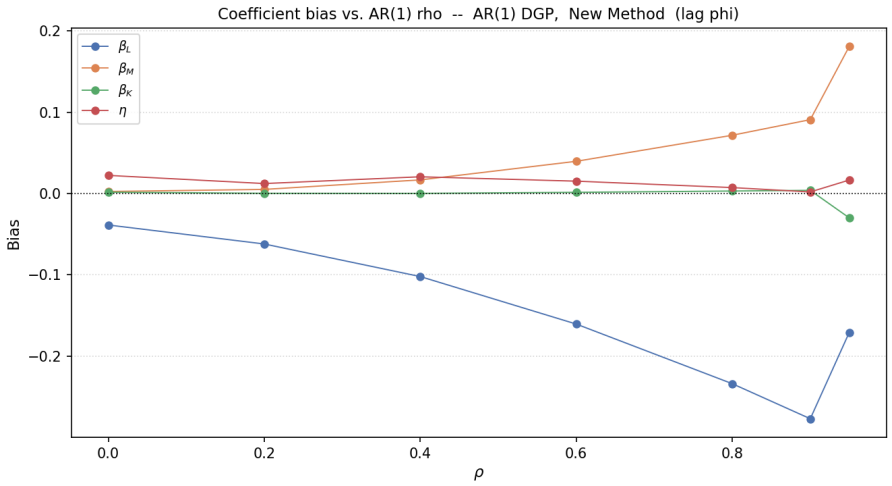
D.5 Sensitivity: correlated productivity and demand innovations

Identification of the production block rests on the current demand innovation being absent from the materials proxy. This holds by construction when $\sigma^{\text{cross}} = 0$: materials load on the demand advantage only through its forecast $\rho_1 D_{i,t-1}$. The empirically relevant violation is therefore not contemporaneous correlation but the case in which *lagged* demand innovations predict current productivity innovations, so that the proxy itself becomes informative about ξ^q_{it}. We re-simulate the AR(1) economy with independent demand innovations and load the productivity innovations on the demand innovation of the previous period. Because labour is set at $t-b$, both latent states are simulated on the sub-period grid $t-1 \to t-b \to t$, and the productivity innovation of sub-period $j \in \{1,2\}$ of year t is

$$\xi^q_{j,it} = \sigma_{\xi^q_j}\left[\sigma^{\text{cross}}\,\frac{\xi^d_{2,i,t-1}}{\sigma_{\xi^d_2}} + \sqrt{1-(\sigma^{\text{cross}})^2}\,\nu_{j,it}\right], \qquad \sigma^{\text{cross}} \in \{0,\,0.2,\,0.4,\,0.6,\,0.8,\,0.9,\,0.95\}, \tag{D.16}$$

where $\xi^d_{2,i,t-1}$ is the demand innovation of the second sub-period of year $t-1$ and $\nu_{j,it} \sim N(0,1)$. The parameter σ^{cross} is therefore a loading, not a correlation. Because both sub-period innovations load on the same lagged demand innovation, the implied correlation between the annual productivity innovation and the annual demand innovation of $t-1$ is 0.41, 0.70 and 0.75 at $\sigma^{\text{cross}} = 0.4$, 0.8 and 0.9, and the standard deviation of the annual productivity innovation is about a third higher at $\sigma^{\text{cross}} = 0.9$ than at zero. At each node we re-run the iterative estimator with the cubic Markov moment condition. The exercise differs from Table D.2 in three respects: it uses the simulator's default optimisation errors ($\sigma_{u^l} = 0.05$, $\sigma_{u^m} = 0.10$), it declares convergence at a tolerance of 0.001 with at most 20 iterations, and it warm-starts the production block at the previous iteration's estimates.

FIGURE D.1. Bias under violations of the innovation exclusion restriction



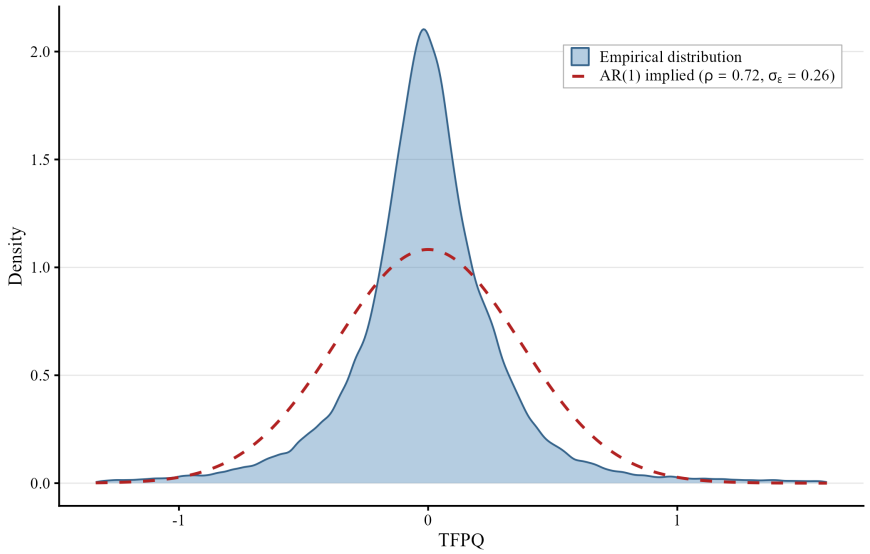
Note: Bias of $\widehat{\beta}^L$, $\widehat{\beta}^M$, $\widehat{\beta}^K$ and $\widehat{\alpha}_p$ from the converged iterative estimator, as a function of the loading σ^{cross} in Equation D.16. The point at $\sigma^{\text{cross}} = 0.95$ reflects a failed estimation (see text). AR(1) productivity process (DGP1), cubic Markov moment condition of Section D.3. The horizontal axis is labelled ρ in the underlying output.

At the identifying assumption the estimator is close to the truth in all four parameters (bias of -0.039, 0.003, 0.002 and 0.022 for β^L, β^M, β^K and α_p). Bias then grows smoothly in σ^{cross} for the two flexible inputs, reaching -0.28 for β^L and 0.09 for β^M at $\sigma^{\text{cross}} = 0.9$, while β^K and α_p stay within 0.025 of the truth up to that point. The last grid point is not informative: at $\sigma^{\text{cross}} = 0.95$ the estimator does not move from its initialisation, and its criterion value is more than two orders of magnitude above that at the other nodes. The reversal in β^L and the jump in β^M at the right end of Figure D.1 reflect this failure, not the violation itself. The quasi-fixed input and the demand elasticity are thus robust to the violation; the flexible-input elasticities are not.

Appendix E: The Unconditional Distribution of TFPQ

Figure E.1 plots the empirical density of TFPQ against the stationary Gaussian implied by the fitted AR(1). The data are more concentrated around the mode and carry more mass in both tails, so the two densities cross twice on each side.

FIGURE E.1. Unconditional TFPQ (z^q) distribution

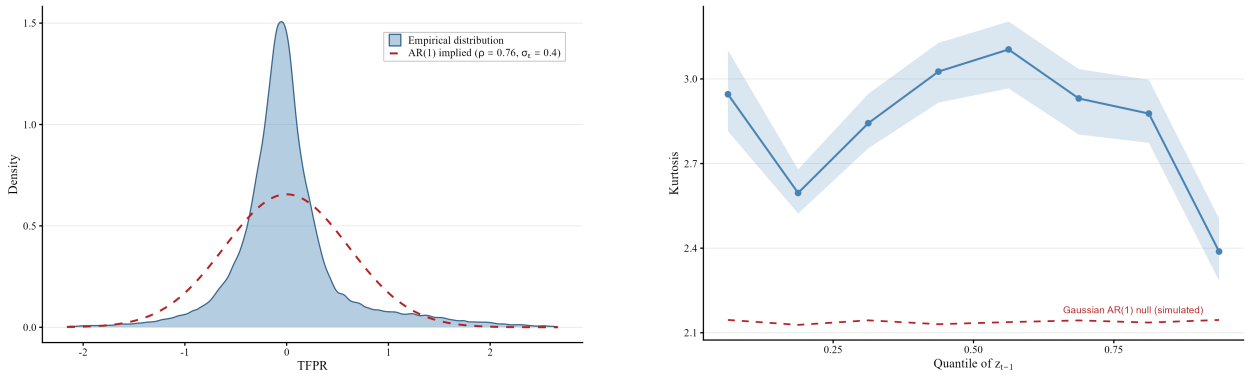


Note: Empirical density of TFPQ in blue. The dashed red line shows the Gaussian density implied by the stationary distribution of a fitted AR(1) process.

Appendix F: The Distribution of TFPR

Figure F.1 reports two further panels on the distribution of TFPR. Panel (a) plots the unconditional density against the Gaussian implied by an AR(1) fitted to it. The departures are the same in kind as for TFPQ but larger in degree. Panel (b) reports conditional kurtosis by quantile of lagged TFPR, which, unlike conditional dispersion, is high everywhere rather than only at the extremes.

FIGURE F.1. The unconditional distribution and conditional kurtosis of TFPR



(a) Unconditional TFPR distribution

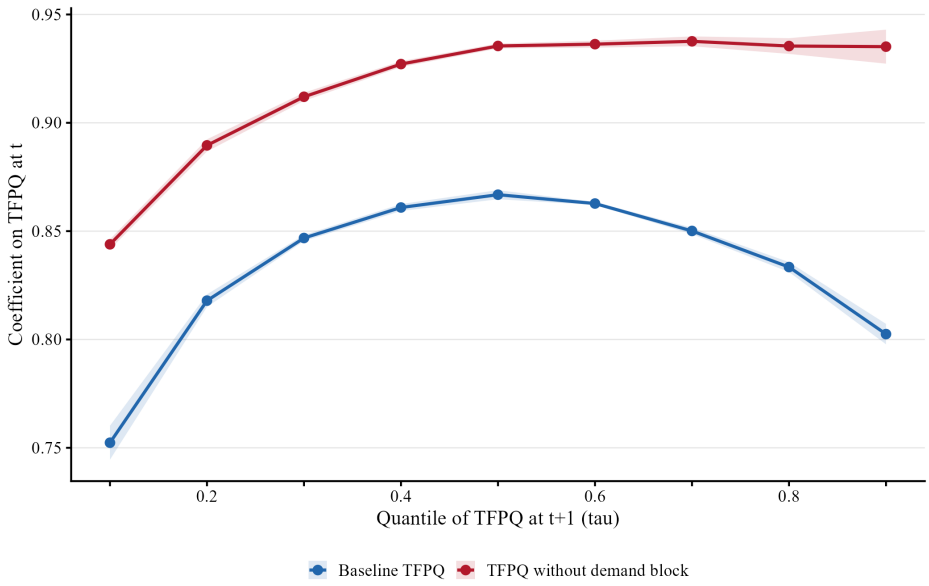
(b) Conditional TFPR kurtosis

Notes: Panel (a) shows the empirical density of TFPR against the Gaussian density implied by the stationary distribution of a fitted AR(1). Panel (b) orders observations by quantiles of $t-1$ residualized TFPR and reports the $(P_{92.5} - P_{7.5})/(P_{75} - P_{25})$ kurtosis; the dashed red line is the AR(1) benchmark.

Figure F.2 separates the role of demand from the nonlinearity of TFPQ itself. It plots the

coefficient from quantile regressions of next year's TFPQ on this year's TFPQ, controlling for the demand advantage D_{ist}, for our baseline TFPQ and for TFPQ estimated without the demand block.

FIGURE F.2. TFPQ persistence with and without the demand block



Notes: Coefficient on z^q_{ist} from quantile regressions of $z^q_{is,t+1}$ on z^q_{ist} and D_{ist} at quantiles $\tau \in \{0.1, \ldots, 0.9\}$, consecutive years only. Blue: baseline TFPQ. Red: TFPQ estimated without the demand block. Shaded areas are 95% confidence intervals.

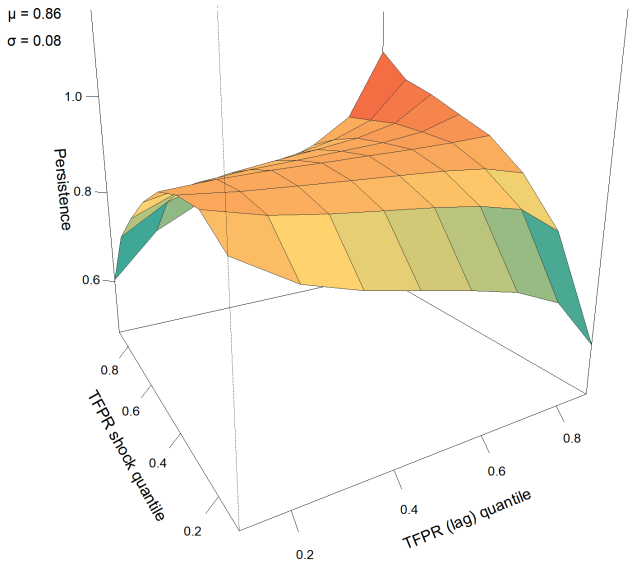
Appendix G: The Persistence of TFPR

This appendix applies the quantile autoregression of Section 4 to TFPR. Figure G.1 displays the estimated persistence surface $\rho(z^r_{i,t-1}, \tau)$, which captures how the effect of lagged revenue productivity varies with both the inherited state and the rank of the current shock.

Revenue dynamics show nonlinearities similar to those documented for TFPQ, including memory-wiping at the extremes, but TFPR is more persistent overall. The mean persistence across states and shocks is 0.86 for TFPR against 0.84 for TFPQ.

The combination of elevated persistence and memory-wiping at extremes implies rich transitional dynamics. While revenue states are generally durable, sufficiently large shocks can still reset a firm's market position—particularly for firms at the tails of the distribution. A low-TFPR firm experiencing a large positive shock (high τ) sees its revenue disadvantage partially erased, while a high-TFPR firm hit by a large negative shock (low τ) can lose its established market position.

FIGURE G.1. Persistence of TFPR



Notes: This figure displays the estimated average derivative of the conditional quantile function for residualized log TFPR. The surface illustrates how the persistence of history varies across both the previous productivity state and the rank of the current shock.

Appendix H: Markup Estimation

To decompose TFPR into its constituent components, we require firm-level estimates of markups. We follow the production-based approach of De Loecker and Warzynski (2012) and De Loecker et al. (2016), which recovers markups from the first-order condition of cost minimization with respect to a variable input.

Under the assumption that firms minimize costs taking input prices as given, the markup μ_{ist} satisfies:

$$\mu_{ist} = \theta_s^V \cdot \frac{P_{ist}Q_{ist}}{P_{ist}^V V_{ist}} \tag{H.1}$$

where θ_s^V is the output elasticity of a variable input V (estimated from the production function), $P_{ist}Q_{ist}$ is firm revenue, and $P_{ist}^V V_{ist}$ is expenditure on the variable input. In our implementation, we use raw materials as the variable input, with the output elasticity $\theta_s^M = \beta_s^M$ recovered from the production function estimation described in Section 3; it is a quantity elasticity, since output is measured in physical units, so the critique of revenue-based markup estimates in Bond et al. (2021) does not apply.

Taking logarithms of Equation H.1, the log markup is:

$$\mu_{ist} = \log \theta_s^M + (p_{ist} + q_{ist}) - (p_{ist}^M + m_{ist}) \tag{H.2}$$

where lowercase letters denote logs. This expression shows that markups are increasing in the revenue-to-materials ratio, scaled by the materials elasticity. Intuitively, firms that generate high revenue relative to their variable input usage—conditional on the production technology—are charging prices above marginal cost.

Given the log markup, we recover the log marginal cost as a residual:

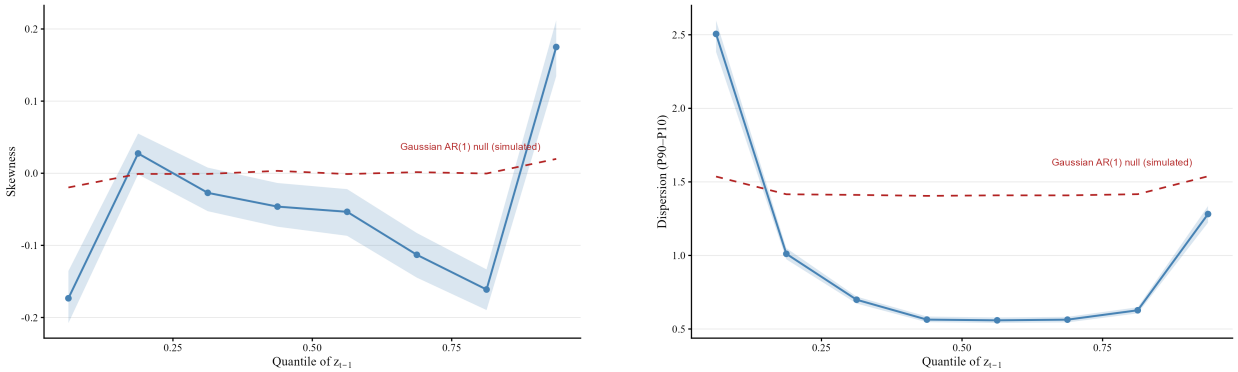
$$mc_{ist} = p_{ist} - \mu_{ist} \tag{H.3}$$

Appendix I: Nonlinearities under Cost-Share Elasticities

Under the cost-share approach, output elasticities are recovered from the first-order conditions of cost minimisation rather than from the estimation procedure of Section 3.1. For a Cobb–Douglas technology, if firms minimise costs taking input prices as given, the output elasticity of each input equals its share in total costs. We measure expenditure on labour and materials directly from the data, and price capital services at their user cost by applying a sector- and year-specific rental rate to the capital stock. Output elasticities are set to the sector-level median cost shares, and log TFPQ is recovered as the corresponding production-function residual.

This procedure replaces the identifying assumptions of Section 3.1 with a different set: (i) firms minimise costs, taking input prices as given; (ii) all inputs are statically optimised, so that observed expenditures reflect frictionless input demand – capital, in particular, is treated as if its services were rented period by period at the user cost; and (iii) technology is Cobb–Douglas with constant returns to scale, so that cost shares, which sum to one by construction, exhaust the output elasticities. In exchange, no assumptions are required on the productivity process, on the timing of input choices, or on the invertibility of input demand: elasticities are measured rather than estimated. Because shares are computed relative to total cost rather than revenue, the approach is furthermore robust to imperfect competition in the output market, as markups leave cost shares unaffected (De Loecker and Warzynski, 2012).

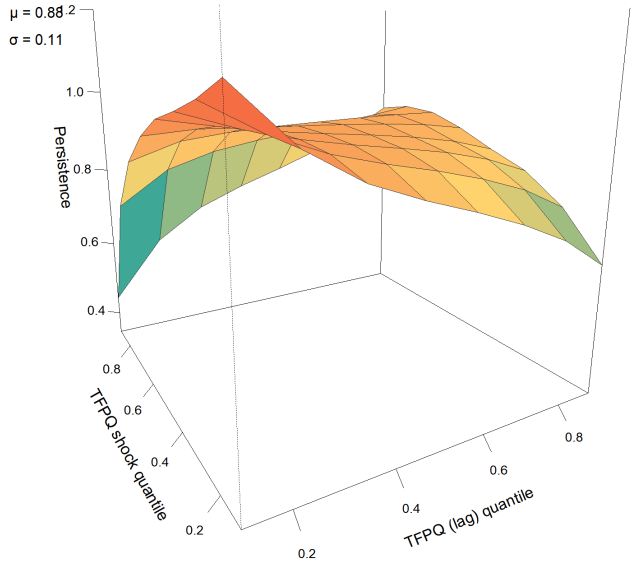
FIGURE I.1. Higher-order moments of TFPQ by productivity group, cost-share elasticities



(a) Conditional skewness

(b) Conditional dispersion

FIGURE I.2. Nonlinear persistence of TFPQ, cost-share elasticities



Appendix J: Discretising the Estimated Process

This appendix describes how the estimated conditional quantile function of Section 4 becomes the transition matrix used in Section 6, and reports the checks that the translation is faithful.

Construction

The estimated object is a conditional quantile function $Q(z, \tau)$: the τ-th quantile of next period's idiosyncratic TFPQ given that this period's value is z. Because Q *is* the inverse conditional CDF, no

Tauchen-style approximation of a linear law of motion is required — the process can be discretised directly, and the rank τ is by construction uniform and independent of the state. Monotonicity of $\tau \mapsto Q(z,\tau)$ is imposed by rearrangement (Chernozhukov et al., 2010).

Fix a grid $\{z_1,\dots,z_n\}$. For each node z_i and each of M equally spaced rank quadrature points $u_m = (m-\frac{1}{2})/M$, compute where the firm lands, $z^\star = Q(z_i,u_m)$, and split the mass $1/M$ between the two adjacent grid nodes with linear-interpolation weights:

$$\omega = \frac{z_{k+1}-z^\star}{z_{k+1}-z_k}, \qquad P_{i,k} \mathrel{+}= \frac{\omega}{M}, \qquad P_{i,k+1} \mathrel{+}= \frac{1-\omega}{M}, \tag{J.1}$$

where $z_k \le z^\star < z_{k+1}$. This is Young's rank lottery. It is locally mean preserving — which matters, because it is the conditional mean that drives the firm's decision — and it delivers non-negative entries and exact unit row sums without renormalisation. We use $n = 150$ nodes and $M = 2{,}000$ ranks.

Support and tails. The grid spans $[z_{p001}, z_{p999}] = [-2.25, 2.76]$, the same range Q itself can return. An earlier version capped the support at ± 3 unconditional standard deviations and clipped to the 0.5th and 99.5th percentiles, which put the grid edge at ± 1.13; the chain then piled genuine tail mass onto its endpoints, and because z is right-skewed it cut the top harder than the bottom. Nodes are placed as sinh of a uniform grid, anchored at the median rather than at the midpoint of the support (the support is asymmetric while the mass sits at zero), so resolution is concentrated where the ergodic density is. The quantile function is estimated on 108 rank knots spanning $\tau \in [0.001, 0.999]$, with additional knots at $\tau = 0.001,\ 0.0025,\ 0.005,\ 0.0075$ and their mirror images; the coefficient functions are interpolated linearly between knots and extended beyond them by a Gaussian (probit) rule, so that the extreme ranks inherit Gaussian rather than exponential tail behaviour.

Making the two chains comparable. The AR(1) chain is built by the same rank lottery on the same grid, from $Q^{\mathrm{AR}}(z,\tau) = \rho z + \sigma_\varepsilon \Phi^{-1}(\tau)$ with $\rho = 0.7508$ and $\sigma_\varepsilon = \sigma_z\sqrt{1-\rho^2}$. To put the nonlinear chain on exactly the same two moments we apply an affine adjustment to the estimated quantile function,

$$\widetilde{Q}(z,\tau) = c\,Q(z/c,\,\tau) + \delta z, \tag{J.2}$$

which shifts the persistence surface rigidly by δ and rescales z by c, and solve for (δ, c) so that the chain's exact ergodic autocorrelation and standard deviation — computed from P and its stationary distribution, not simulated — equal the empirical 0.7508 and 0.3780. The solution is $\delta = -0.0324$ and $c = 1.2190$. Because a rigid shift moves the average local persistence $\mu = \mathbb{E}[\partial Q/\partial z]$ and the linear slope ρ by the same amount, the adjustment leaves the wedge $\mu - \rho$ untouched; we verify this rather than assume it by building the unadjusted chain as well. The conditional quantile function is estimated on 108 rank knots between $\tau = 0.001$ and $\tau = 0.999$, with extra resolution at

$\tau = 0.001$, 0.0025, 0.005, 0.0075 and their mirror images, and the chain integrates over 2,000 equally spaced ranks. A firm drawing a one-in-a-thousand shock is therefore sent where the estimated quantile function says it should go, rather than to the last fitted knot or the edge of the grid.

Checks

The two chains are observationally equivalent except in one number. By construction both have $\rho = 0.751$ and $\sigma_z = 0.378$. The average local persistence is 0.809 for the nonlinear chain against 0.751 for the AR(1), a wedge of $\mu - \rho = 0.058$ against 0.000. That wedge is identically zero for *any* AR(1), whatever its parameters, so it is the sharpest single summary of what a linear regression cannot see. The estimated surface itself gives $\mu = 0.842$; the chain recovers 0.809, so the discretisation is mildly conservative — it understates the nonlinearity rather than manufacturing it.

The chain reproduces the estimated surface. We compare $\partial Q/\partial z$ implied by the chain, cell by cell, with the surface estimated directly in Section 4, on the same state-by-rank grid. The range across ranks is 0.1188 in the chain against 0.1181 in the estimated surface, and the state-by-rank interaction is 0.0461 against 0.0456. The four corners agree up to the calibration shift. The AR(1) block is constant at ρ in every cell, which is what "an AR(1) has no persistence surface" means numerically.

The results do not depend on the grid. We rebuilt both chains under six combinations of node count and spacing (uniform and sinh-curved). The ergodic moments, the wedge, and the conditional dispersion profile agree to roughly three decimals across all of them, and the mass on the two edge nodes is unchanged by refining n — which is the relevant test, since edge mass is what a support that is too narrow would inflate. As a further cross-check the matrices were rebuilt by the alternative Tauchen-style route (numerically inverting $\tau \mapsto Q(z_i, \tau)$ at the grid midpoints and differencing the conditional CDF); the two constructions agree, which they would not if the grid were too coarse.

Conditional dispersion survives discretisation. The width of the central 90 percent of each row of P is flat across rows for the AR(1) chain, by construction, and U-shaped for the nonlinear chain, reproducing the conditional dispersion profile estimated on the data. This is the feature the model responds to and the one a fitted AR(1) has no parameter to express.

Appendix K: The Process and the Frictions: Five Configurations

Section 6 reports two configurations of the model, each estimated under its own productivity process. This appendix reports three more, which together separate what the process does from what the frictions do. The five are:

(i) **AR(1)**. The linear-Gaussian process at its own SMM estimate. This is what a standard paper would have produced, end to end.

(ii) **NL**. Our estimated process at its own SMM estimate.

(iii) **NL@AR(1)**. Our process solved at the AR(1)'s estimated frictions. Identical model, identical frictions, identical state space: only the transition matrix is swapped. Nothing here can be absorbed by re-estimation, which is why it is worth reporting separately.

(iv) **AR(1)**† and **NL**†. Each process with every adjustment cost set to zero — convex, linear and fixed — at the same estimated wage, so that firm scale stays comparable and the only thing removed is the friction. These bound the exercise: whatever the two processes still do differently with no adjustment costs at all is the pure effect of the process on the firm's problem, with nothing for the frictions to absorb.

The frictionless columns. Read AR(1)† against NL† in Table K.1. Neither has an inaction region in the usual sense — there is no fixed cost to create one — so the inaction row measures the share of firm-years whose desired employment change falls below the same 2.5 percent yardstick used on the data. The AR(1) leaves 3.5 percent of firm-years there; our process leaves 10.9 percent. Both sit on the same employment grid and face the same wage, so the difference is the conditional distribution and nothing else. The excess kurtosis row is starker still: with no frictions the AR(1) generates -0.42, a growth distribution *thinner-tailed than a Gaussian*, while our process generates 4.24 against 5.32 in the data. Four-fifths of the empirical fat tail is delivered by the process before any adjustment technology is introduced.

Skewness is the clean counter-example, and it is what makes the decomposition informative rather than a claim that frictions do not matter. Both frictionless columns are essentially symmetric (0.002 and -0.008) while the estimated columns are not, so the asymmetry of the firm growth distribution is produced by the kinked hiring and firing cost and not by the process. Fat tails and state-dependent inaction come from the process; asymmetry comes from the friction.

State-dependence without state-dependent costs. Table K.2 makes the first point sharper. The frictionless AR(1) profile is essentially flat: 0.031 in the bottom decile, 0.036 in the top, a spread of 0.012. This is what homoskedasticity means operationally — every firm faces the same distribution of news whatever its state, so every firm adjusts at the same rate. The frictionless

TABLE K.1. Five configurations: what the process does and what the frictions do

	Data	AR(1)	NL	NL@AR(1)	AR(1)†	NL†
Panel A. Frequency of employment adjustment						
Inaction rate	0.434	0.436	0.440	0.560	0.035	0.109
Hiring rate	0.301	0.287	0.284	0.219	0.483	0.449
Firing rate	0.265	0.277	0.277	0.221	0.482	0.442
Positive spike rate	0.093	0.091	0.095	0.054	0.331	0.190
Negative spike rate	0.088	0.092	0.091	0.055	0.330	0.189
Panel B. Employment growth among adjusters						
Standard deviation	0.268	0.197	0.290	0.227	0.442	0.384
Skewness	-0.123	-0.046	-0.164	0.169	0.002	-0.008
Excess kurtosis	5.316	-0.806	6.258	5.259	-0.420	4.241
$P_{90} - P_{10}$	0.571	0.484	0.567	0.452	1.154	0.792
$(P_{90} - P_{50})/(P_{50} - P_{10})$	0.759	0.599	0.788	1.731	0.870	0.824

Notes. † marks a frictionless configuration: the convex, linear and fixed adjustment costs are all set to zero, at the same wage, so the only difference from the corresponding estimated column is that the firm may reset employment freely every period. NL@AR(1) is the nonlinear process solved at the AR(1)'s estimated frictions – the same model, the same state space, only the transition matrix swapped – and is not re-estimated, so its misses are the object of interest rather than a failure of fit. The two † columns are the ones to read against each other: with no adjustment cost of any kind the AR(1) produces almost no inaction and thinner-than-Gaussian tails, whereas the nonlinear process on its own produces a sizeable inaction region and excess kurtosis of the empirical order of magnitude. Skewness is the mirror image – both frictionless columns are essentially zero, so the asymmetry of firm growth is the kinked cost and not the process.

TABLE K.2. Inaction by lagged-productivity decile, five configurations

Configuration	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	Spread
AR(1)	0.461	0.446	0.434	0.434	0.429	0.422	0.429	0.429	0.434	0.442	0.039
NL	0.446	0.459	0.469	0.465	0.468	0.457	0.460	0.445	0.416	0.318	0.151
NL@AR(1)	0.537	0.600	0.606	0.599	0.599	0.593	0.590	0.575	0.526	0.389	0.217
AR(1)†	0.031	0.031	0.031	0.031	0.031	0.039	0.042	0.041	0.037	0.036	0.012
NL†	0.103	0.093	0.113	0.084	0.145	0.126	0.126	0.117	0.099	0.074	0.071

Notes. Share of firm-years with $|g| \leq 0.025$ within bins of lagged idiosyncratic productivity; 'Spread' is the maximum less the minimum across bins. Columns are labelled as in Table K.1. The empirical counterpart of the spread is 0.080, computed over eight bins rather than ten (Section 4), which makes it the conservative comparison. The frictionless AR(1) profile is flat almost by construction – a homoskedastic, symmetric process gives every firm the same distribution of news whatever its state – whereas the frictionless nonlinear process reproduces most of the empirical state-dependence with no adjustment cost at all.

nonlinear process spreads by 0.071 across deciles, against 0.080 in the data. Almost all of the empirical state-dependence in the adjustment margin is reproduced by a model with no adjustment

cost of any kind. A researcher who observed that profile and attributed it to nonconvex costs — the standard inference — would be reading a property of the productivity process as a property of the adjustment technology.

The swap. The NL@AR(1) column is the pure consequence of feeding a nonstandard process into an otherwise standard, already-calibrated model. It raises the inaction rate from 0.436 to 0.560 and cuts the hiring and firing rates by roughly a quarter, so a model calibrated on an AR(1) and then confronted with the true process misses the adjustment margin by twelve percentage points. It also over-corrects the asymmetry: the ratio $(P_{90} - P_{50})/(P_{50} - P_{10})$ goes to 1.73 against 0.76 in the data, that is, from too little right-tail mass to far too much. Re-estimation, in other words, does absorb a great deal; that is precisely why the targeted-moment fit in Table 7 looks so reassuring, and precisely why it cannot be relied on.