Essays in modern Macroeconomics

Timo Felix Haber

Faculty of Economics
University of Cambridge

This thesis is submitted for the degree of
Doctor of Philosophy

Clare Hall

September 2022

To Berend, Hanne and Maximilian.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution. It does not exceed the prescribed word limit of 60,000 words.

Timo Felix Haber
September 2022

Essays in modern Macroeconomics

Timo Felix Haber

This thesis contains three chapters, each addressing a highly relevant area of interest in modern macroeconomics. The first chapter tests the presence of state-dependent pricing frictions by analysing the effects of monetary policy during high inflation and after large monetary shocks. The second chapter focuses on the interaction of heterogeneity in firm potential with financial frictions and its effect on macroeconomic outcomes. The final chapter presents a global solution method for heterogeneous agent models that can handle interesting and relevent extensions to the standard setting.

The first chapter is co-authored with Professor Guido Ascari and has been published in *The Economic Journal* in January 2022 (Ascari & Haber, 2022). An earlier version of this chapter was submitted as my MPhil thesis at the University of Oxford in 2017. It has undergone several substantial modifications since then. First, the main estimation technique has changed, as we are using smooth local projections developed by Barnichon & Brownlees (2019). Second, there are novel complementary results, such as a sectoral analysis. Finally, we have included numerous additional robustness tests at the request of the referees. The main message of the chapter, however, is still the same. A sticky price theory of the transmission mechanism of monetary policy shocks based on state-dependent pricing yields two testable implications that do not hold in time-dependent models. First, large monetary policy shocks should yield proportionally larger initial responses of the price level. Second, in a high trend inflation regime, the response of the price level to monetary policy shocks should be larger and real effects smaller. Our analysis provides evidence supporting these non-linear effects in the response of the price level in both aggregate and sectoral US data, indicating state-dependent pricing as an important feature of the transmission mechanism of monetary policy.

The second chapter is co-authored with Dr Christian Rörig and Dr Miguel Ferreira. Using a unique dataset covering the universe of Portuguese firms and their credit situation we show that financially constrained firms (1) are found across the entire firm size distribution, (2) account for a sizeable asset share, and (3) exhibit a higher cyclical sensitivity, conditional on size. As these findings are counterfactual to the canonical heterogeneous firms model with financial frictions we identify a factor in the data that reconciles theory and data: permanent heterogeneity among firms. Incorporating ex-ante dispersion of firm potential into the standard model allows us to match the distribution of

constrained firms conditional on size. This, together with the fact that constrained firms have a higher capital elasticity, leads to up to four times larger aggregate fluctuations and capital misallocation.

The third chapter is single-authored. I present a global solution method for the truncated history method developed by Le Grand & Ragot (2022) to solve heterogeneous agent models with aggregate shocks. Their method provides a finite state-space representation of economies by truncating idiosyncratic histories. The global solution to this representation is based on the parameterized expectations algorithm developed by den Haan & Marcet (1990). This solution algorithm is accurate and does not rely on the assumption of a constant set of constrained agents. Further it is applicable to model setups where local methods may struggle or fail. I provide three examples where this is the case; cyclical borrowing constraints, uncertainty shocks and a two asset economy. I show that all three modifications can be solved with the global method and provide interesting insights into the importance of heterogeneity beyond the standard model.

Acknowledgements

I am deeply grateful to my supervisor, Elisa Faraglia, for her continued guidance, wisdom and moral support throughout these past years. I would not be the economist and person I am today without her patience, humor and enthusiasm. I am deeply privileged to have had a supervisor that I can also call my friend.

I would like to thank Florin Bilbiie, Charles Brendon, Vasco Carvalho, Tiago Cavalcanti, Giancarlo Corsetti, Chryssi Giannitsarou, and many others at the Faculty of Economics. Their continous encouragement and brilliant counsel inspired me both professionally and personally, and will do so for a long time to come. This is equally true for Guido Ascari, Sarah Holton, Oliver Landmann, Cezar Santos and Riccardo Trezzi who, by sharing some of their economic thinking, became role models along the way.

I am also indebted to the great members of staff at the Faculty of Economics who always had an open ear, put a smile on my face and found a way out of the regulatory jungle of Cambridge.

Further, I would like to thank the Cambridge Trust, Clare Hall and the Studienstiftung des Deutschen Volkes for their financial support. This PhD would have not started, let alone have finished, without the support of these great institutions.

The friendships that originated at Cambridge will remain for life, no matter where individual paths may lead us. And so I want to thank my brilliant friends Sophie Allcock, Michael Ashby, Jakob Berndt, Tim Coorens, Miguel Ferreira, Zeina Hasna, Salvatore Lattanzio, Emile Marin, Patrick McClanahan, Charles Parry and Christian Rörig. It was an honour and, most of all, a pleasure to share these years with so many gifted, enthusiastic and caring individuals who will make a difference in this world.

I also am indebted to the Clare Hall Boat Club, a home away from home. The early mornings and tough training sessions were paid back a thousand times by the fantastic teammates and kind-hearted friends.

I would also thank my friends outside of Cambridge. There are some that have been there for a very long time. Johanna Albert, Daniel Albrecht, Maximilian Bretschneider,

Karina Graf, Paul Kahle, Marina Kaiser, Sara Koller, Simon Sanktjohanser, Ricarda Schomburgk are just a few of the amazing individuals that inspired me to pursue a PhD even when I didn't believe I could myself. I am eternally grateful for that. There are some friends that had a profund impact on me later in life. I am indebted to Gregor Becker, Raees Chowdhury, Wodzislaw Kicinski, Matthias Mitterbacher, Juan Pradera and Philipp Heuermann for sharing their brilliance, encouragement and inquisiveness with me. I am looking forward to more journeys with these friends, both intellectually and geographically.

I would not be writing these words without my family. My parents have always believed in me and I am infinitely thankful that they have always supported me in whatever I wanted to do. To have the resolve of my father and the empathy of my mother is a great blessing. I am also incredibly lucky to call my brother my best friend and I am looking forward to exploring the world together further.

Finally, I am grateful beyond words to my girlfriend, Simone, who encouraged, supported and motivated me beyond measure throughout these past years. I know how fortunate I am to have a partner who unconditionally accepts and appreciates me for who I am.

Contents

List of figures

List of tables

Chapter 1

Non-linearities, state-dependent prices and the transmission mechanism of monetary policy

Abstract

A sticky price theory of the transmission mechanism of monetary policy shocks based on state-dependent pricing yields two testable implications that do not hold in time-dependent models. First, large monetary policy shocks should yield proportionally larger initial responses of the price level. Second, in a high trend inflation regime, the response of the price level to monetary policy shocks should be larger and real effects smaller. Our analysis provides evidence supporting these non-linear effects in the response of the price level in both aggregate and sectoral US data, indicating state-dependent pricing as an important feature of the transmission mechanism of monetary policy.

Keywords: Sticky prices, state-dependent pricing, monetary policy
JEL Codes: E30, E52, C22

1.1 Introduction

The New Keynesian (NK) paradigm is one of the main frameworks for the analysis of business cycle fluctuations and the effects of monetary and fiscal policy, both among academic researchers and policy institutions. While there can be alternative mechanisms for the transmission of monetary policy shocks, price stickiness remains the major reason why monetary policy has effects on real variables in virtually all NK models. This pivotal assumption has been substantiated by numerous empirical studies that show that individual prices do indeed change infrequently (e.g. Bils & Klenow, 2004; Eichenbaum et al., 2011; Klenow & Kryvtsov, 2008; Nakamura, 2008; Nakamura & Steinsson, 2008; Nakamura & Zerom, 2010). Moreover, this literature has demonstrated that microfounded state-dependent models are able to replicate the empirical distribution of price changes. Unlike time-dependent pricing models, such as Calvo (1983) and Rotemberg (1982), microfounded state-dependent models of nominal rigidities assume that the individual firm can change its price, subject to an adjustment cost. Hence, the frequency of price changes, the number of firms that decide to change their prices following a monetary policy shock, and the size of these changes are endogenous (the so-called "selection effect", see Golosov & Lucas, 2007). In particular, these models predict that individual prices (i) are more flexible in response to large shocks (e.g. Karadi & Reiff, 2019) and (ii) change more frequently when inflation is higher (e.g., Álvarez et al., 2019).

Yet, despite the large literature on sticky prices, to the best of our knowledge, there is surprisingly little direct evidence that *aggregate* prices behave as state-dependent models suggest and thus that predictions (i) and (ii) affect the transmission mechanism of monetary policy in aggregate US data. We aim to fill this gap. If staggered prices are of such paramount importance to the transmission mechanism of monetary policy, and if there is a significant fraction of state-dependent prices, then the two very general predictions (i) and (ii) should also emerge in the data. First, *large absolute value* monetary policy shocks should lead more firms to adjust their prices, and hence yield a proportionally larger response of inflation, whereas the real effects should be subdued. Second, the frequency of price changes should be an increasing function of underlying levels of inflation; that is, prices should be more flexible in a *high trend inflation* regime than otherwise. Hence, the higher trend inflation, the larger the response of inflation and the smallerthe real effects of monetary policy shocks should be.

These theoretical results are quite intuitive and general since they derive from a broad variety of state-dependent sticky price models in the literature. In particular, Álvarez

et al. (2017) show that the impulse response after a monetary shock is size-independent in time-dependent models, whilst it is not in state-dependent models. In this sense, our contribution is really a minimal, first-pass test for state-dependent price theories: if state-dependency is pivotal, aggregate data should exhibit these two features. We take these two theoretical predictions to US data between 1969 and 2007, applying smooth local projections (Barnichon & Brownlees, 2019; Jordà, 2005) and the smooth transition function methodology of Granger & Teräsvirta (1993) together with US monetary policy shocks identified with the narrative method of Romer & Romer (2004). The empirical methodology of this paper is most closely related to the work of Tenreyro & Thwaites (2016), focusing on the differential effects of monetary policy in recessions and expansions. As such, our paper also contributes to the literature about the state-dependent effects of monetary shocks.

Our analysis provides new, statistically significant evidence in favor of state-dependent pricing models in aggregate US data. First, large absolute value shocks have disproportionately larger effects on prices on impact, but are less persistent and have weaker real effects, matching the first theoretical prediction. Second, the impulse response functions (IRFs) in the high and low trend inflation regimes are significantly different for prices and inflation and are also in line with the second theoretical prediction of higher price flexibility in the high trend inflation regime. However, for the second prediction, we do not find statistically significant evidence of muted real effects. Importantly, the non-linearity of the impulse response functions is not due to a different feedback of monetary policy to inflation in response to small vs. large shocks or in periods of high vs. low trend inflation. Moreover, we supplement our analysis of the aggregate price series with a smooth local projection panel analysis of seventeen sectoral price series and find similar results, strengthening the evidence of state-dependent pricing in the data. Hence, this study is also the first to apply the methodology of Barnichon & Brownlees (2019) in a panel setting, to the best of our knowledge. In the appendix we conduct a comprehensive sensitivity analysis to establish the robustness of these results, regarding different empirical specifications, sub-samples, controls and measures of monetary policy surprises.

Literature. The findings of our study are related to the empirical literature on state-dependent pricing, as mentioned before. Some papers find evidence of state-dependent pricing behavior by investigating the response of prices at the micro-economic level to a particular event. Hobijn et al. (2006) document a dramatic increase in restaurant prices in the Euro area after the introduction of the Euro and explain it through the lens of a menu

cost model. Karadi & Reiff (2019) document that micro-level price responses to three large value-added tax (VAT) changes, which occurred in Hungary in 2004 and 2006, were flexible and asymmetrical with respect to positive versus negative tax changes.[1] In order to analyse the pass-through of the Swiss exchange rate shock into import product prices from the Euro area, Bonadio et al. (2019) exploit the Swiss National Bank's (SNB) decision to discontinue the minimum exchange rate policy of one euro against 1.2 Swiss francs on January 15, 2015. The empirical literature on exchange rate pass-through generally concerns measuring the extent of the exchange rate pass-through, and, while somewhat related, it is not concerned explicitly with state-dependent pricing models. Two notable exceptions are Álvarez et al. (2017) and Álvarez & Neumeyer (2020) who compare the observed price dynamics at the micro level after large devaluations to the predictions of an adjustment cost model of price setting. Álvarez et al. (2017) use monthly data on Consumer Price Index (CPI) inflation and the nominal exchange rate from numerous countries in a panel data analysis. Álvarez & Neumeyer (2020) analyse several episodes involving large changes in the nominal price of inputs in Argentina over 2012–2018 by using micro-level price data for the city of Buenos Aires. The work of Álvarez et al. (2017) is the closest to ours because they use panel and aggregate data, and they link the empirical analysis explicitly to the prediction of state-dependent pricing models that the inflation response should depend on the size of the shock. Finally, related to our second test, Veirman (2009) investigates empirically whether the flattening of the Japanese Phillips Curve depends on lower trend inflation, as state-dependent models would imply.

Our results are the first ones (to the best of our knowledge) that point towards a significant presence of state-dependent pricing in the US economy from an aggregate perspective. Our macro evidence is therefore a useful complement to existing micro evidence.

Outlook. The rest of this chapter is structured as follows. Section 1.2 reviews the theory and derived testable implications for our empirical analysis. Section 1.3 discusses the empirical methodology, including the data and the technique of smooth local projections. In Section 1.4 we present our findings and in Section 1.5 we conduct the robustness tests, including our complementary analysis of sectoral price data. Section 1.6 concludes.

[1] Álvarez et al. (2006) and Gagnon et al. (2012) are earlier papers that find significant increases in the frequency of price changes in the months with VAT tax changes, consistent with state-dependent pricing models.

1.2 Theory and testable implications

Contrary to early standard time-dependent pricing models (e.g. Calvo, 1983; Fischer, 1977; Taylor, 1979, 1980), microfounded state-dependent pricing models feature endogenous price adjustments triggered by changes in the economic environment. The firms that decide to change their prices by paying the adjustment costs after a monetary shock are the ones further off from their optimal price ("selection effect"). Since these are not random firms and the sizes of the price changes are relatively large, early studies (Caplin & Spulber, 1987; Golosov & Lucas, 2007) find that the aggregate price level can mimic a flexible price environment. Nonetheless, later studies investigate the robustness of this result to various extensions and show that state-dependent models yield a large degree of aggregate price stickiness and are important for the transmission mechanism of monetary policy.[2] Crucially, all of these state-dependent models also offer two testable implications for aggregate data to verify the empirical validity of sticky price theories.

First Testable Implication: The impulse response functions of inflation and output to a monetary policy shock should depend on the size of the shock. If there is a non-negligible fraction of state-dependent price-setters, the impulse response should be a non-linear function of the size of the shock. The larger the shock, the larger the number of firms that decide to pay the adjustment costs and change their prices immediately, such that the reaction of the aggregate price level at short horizon is increasing in the size of the monetary policy shock (see Álvarez & Lippi, 2014). Moreover, the effect on the price level should be less persistent, because the larger the shock, the higher the number of firms that adjust immediately and therefore the lower the number of firms that eventually would change their price as the shock tapers off. The real effects of monetary policy shocks, are instead hump-shaped with respect to the size of the shock because of two counteracting effects. First, larger shocks, *ceteris paribus*, give rise to stronger real effects, just like in a time-dependent model. Second, larger shocks also increase the number of adjusting firms, strengthening the reaction of the aggregate price level and thus reducing the real effects. For a small shock, the first effect prevails, so that both the impact and the cumulative

[2] These extensions are informational costs (Álvarez et al., 2011; Bonomo et al., 2013; Gorodnichenko, 2008), multi-product firms (Álvarez & Lippi, 2014; Bhattarai & Schoenle, 2014; Midrigan, 2011), multiple sectors and intermediate inputs (Nakamura & Steinsson, 2010), or errors in the timing and the precision of the adjustment (Costain et al., 2019). The same result holds in a variety of state-dependent models that are carefully calibrated to match the main features of retail price microdata (Álvarez et al., 2016; Costain & Nakov, 2011, 2019; Eichenbaum et al., 2011) and, more recently, of both the size and frequency of price *and wage* changes (Costain et al., 2019).

effect on output are increasing in the size of the shock. For large shocks the opposite occurs. It follows that sufficiently big shocks should have lower real effects than smaller shocks.[3]

Second Testable Implication: The impulse response functions of inflation and output to a monetary policy shock should depend on the average level of inflation. As outlined by Dotsey et al. (1999) or Costain & Nakov (2011), average inflation affects the frequency of price adjustments in state-dependent models because it erodes a firm's relative price so that firms adjust prices more frequently. Indeed, the empirical analysis of Álvarez et al. (2019) provides solid evidence of how the frequency of price changes varies with inflation.[4] This evidence implies different impulse responses to a monetary policy shock in high trend inflation regimes compared to low trend inflation regimes. In particular, we should observe a quicker and less persistent reaction of prices in high inflation regimes. Further, Álvarez et al. (2016) demonstrate that in a large class of sticky-price models, the total cumulative output effect of a small unexpected monetary shock is inversely related to the average number of price changes per year.[5] This theoretical prediction provides the second testable implication.

1.3 Empirical methodology

In this section we describe the data and the empirical methodology used to estimate smooth impulse responses and conduct inference. We test the two predictions by analysing the presence of non-linearities in the impulse response functions to a monetary policy shock for large and small shocks, and during high and low trend inflation. Our empirical methodology follows a growing body of literature employing local projections (Jordà, 2005) to account for the response to non-linear terms and state-dependency in empirical

[3] Álvarez & Lippi (2014) show that the monetary shock that maximises the cumulated effect on output (i.e., the area under the impulse response function) is about one-half of the standard deviation of price changes. Costain et al. (2019) show that a similar effect occurs in a model with state-dependent prices and wages.

[4] Figure 5 and 6 therein show that the frequency of price changes do not react much for levels of annual inflation up to 5%, then it starts accelerating, and finally it increases linearly for values of annual inflation above 14% with an elasticity of about two-thirds. This is in line with Sheshinski & Weiss (1977)'s adjustment cost model with no idiosyncratic shocks.

[5] The same holds in a model in which nominal rigidities in both wages and prices are state-dependent (see Figure 11 in Costain et al., 2019), and in a model that allows for temporary price changes, because firms can set a price plan, rather than a fixed price as in the standard adjustment cost model (see Figure 7 in Álvarez & Lippi, 2020).

impulse responses (e.g. Auerbach & Gorodnichenko, 2012a,b; Caggiano et al., 2015, 2014; Furceri et al., 2016; Ramey & Zubairy, 2018; Tenreyro & Thwaites, 2016).

Data. The monthly sample for our response variables runs from 1969m1 up to 2007m12, excluding the most recent financial crisis, where monetary policy has been very different and the zero-lower bound on nominal interest rates has been binding. We analyse three response variables: output, inflation and the nominal interest rate. The series for output is the industrial production index, for inflation we use personal consumption expenditure (PCE) inflation, and for the nominal interest rate we use the effective Federal funds rate, all from the Federal Reserve Bank of St. Louis Database (FRED).[6]

The main shock variable used in this analysis is based on the narrative analysis of Romer & Romer (2004). They identify monetary policy surprises by using a narrative approach to infer the intended Federal funds rate at every Federal Open Market Committee (FOMC) meeting from 1969 onwards. By regressing changes of this intended rate on Greenbook forecasts they derive a measure of monetary policy surprises that is arguably exogenous to the Fed's information set about the future state of the economy. We utilise this methodology and the extended shock sample until December 2003 by Tenreyro & Thwaites (2016).

There are a number of other ways of identifying monetary policy surprises, for example High Frequency Identification (HFI) (Barakchian & Crowe, 2013; Gertler & Karadi, 2015; Gürkaynak et al., 2005; Kuttner, 2001) or identification restrictions in a Vector Autoregression (VAR) (Bernanke et al., 2005; Bernanke & Mihov, 1998; Christiano et al., 1999; Kim & Roubini, 2000; Uhlig, 2005). We conduct robustness tests with respect to the latter and find largely similar conclusions.[7]

Smooth Local Projections. We estimate local projection coefficients using the recently developed methodology by Barnichon & Brownlees (2019) to improve accuracy and infer-

[6]The variables are all in natural log levels and then multiplied by 100 except for the Federal funds rate which remains in percentage points. This standard transformation enables the interpretation of the strength of the coefficients as approximate percentage points.

[7]We do not, however, include HFI shocks for two main reasons. First, the available shock series are too short for our analysis. Even the backward extension up to 1979 by Gertler & Karadi (2015) omits a significant proportion of the Great Inflation period; a major source of variation in our smooth transition function. Second, Ramey (2016) argues that these shocks may not be robust to samples where anticipation effects are important. Moreover, (Ramey, 2016, p.109) shows that the impulse response function to HFI shocks can look very different depending on whether one uses a VAR or local projections. For example, a contractionary Gertler & Karadi (2015) shock in a local projection *increases* output and leaves the price level largely unchanged whereas it produces the literature standard effect in their proxy SVAR.

ence over the standard least squares approach.[8] With this technique, the impulse response coefficients, i.e. all shock-dependent coefficients, are modelled as linear combinations of B-spline basis functions. One can then estimate the coefficients of these linear combinations using generalised ridge estimation, with a penalty parameter that selects the degree of shrinkage. When the shrinkage parameter is close to zero, the estimation yields the standard least squares estimates. Conversely, if the parameter is high, the impulse response is converging to a smooth limit polynomial distributed lag model. We follow Barnichon & Brownlees (2019) and select the shrinkage parameter using k-fold cross validation (Racine, 1997).

We also follow Barnichon & Brownlees (2019) on conducting inference. In particular, in order to take into account potential autocorrelation and heteroscedasticity, we estimate the variance of the coefficients using a modified Newey & West (1987) estimator, corrected for the penalty parameter. We use the resulting variance matrix to construct confidence intervals and t-statistics in the ordinary way. We provide further details on this and the estimation technique as a whole in the Appendix.

1.4 Results

This section presents the main results regarding the two testable implications discussed above.

1.4.1 Implication 1: Size-dependent effects of monetary policy shocks

In order to test the size-dependence of impulse responses we consider the following non-linear local projection:

$$y_{t+h} = \alpha_h + \tau_h t + \beta_h e_t + \zeta_h (e_t \cdot |e_t|) + \sum_{k=1}^{K} \gamma_{h,k} w_{t,k} + v_{t+h}, \tag{1.1}$$

which is estimated for $h = 0, 1, \ldots, \mathrm{H}$. We set $\mathrm{H} = 48$ which corresponds to an impulse response horizon of four years. y_{t+h} denotes the variable of interest, in our case either the industrial production index, PCE inflation or the Federal funds rate. e_t are the narrative Romer & Romer (2004) shocks. $w_{t,k}$ denotes the kth control variable and v_{t+h} the estimation error, possibly heteroskedastic and serially correlated. The set of control variables includes up to two months of lags of industrial production, PCE inflation and the Federal

[8] We provide the results with standard least squares coefficients in the Appendix.

funds rate. Moreover, we follow Ramey (2016) and include contemporaneous values of the industrial production index and PCE inflation. This is equivalent to assuming recursiveness between the three different variables of interest since inflation and industrial production can contemporaneously affect the Federal funds rate but not vice versa.[9]

While the coefficient β_h captures the linear component, the coefficient ζ_h on the absolute value interaction term accounts for non-linearities in the impulse response function due to the size of the shock. The interaction term, $(e_t \cdot |e_t|)$, magnifies the size of the shock, but it keeps the same sign of the shock. Hence, contrary to a simple quadratic term, it isolates the pure effect of a change in the size of the shock. If ζ_h has the same sign of β_h then the non-linear interaction term *amplifies* the linear effects of the impulse response. On the contrary, if ζ_h has the opposite sign to β_h, it *counteracts* the linear impulse response, possibly even tilting the overall effect from one sign to another for large enough shocks. Whenever $\zeta_h = 0$ the impulse response function is linear with respect to the shock size.

Therefore, ζ_h is the main coefficient of interest to test our first theoretical prediction. Let y_t signify prices, and assume a large monetary contraction, i.e., a positive value of e_t. Large monetary policy shocks should induce a more price-flexible impulse response function of the price level and inflation. If firms exhibit state-dependent pricing we should see a negative ζ_h at small horizons as we expect β_h to be close to zero or negative. This would mean that firms decrease prices quicker and so prices decline disproportionately at small horizons. Furthermore, we would then expect to see a positive ζ_h at larger horizons, weakening the price response, as more firms have already changed prices earlier. Consequently, a combination of a negative ζ_h at small horizons, as more firms change prices right away, and a positive ζ_h at larger horizons, as persistence is lower due to earlier price changes, would speak in favor of state-dependent pricing as a valid aggregate propagation mechanism of monetary policy shocks.

Coefficient estimates. The resulting coefficients from estimating local projection (1.1) for PCE inflation, industrial production and the Federal funds rate are reported in the panels of Figure 1.1. The solid line plots the coefficients of the linear term, $\hat{\beta}_h$, and the dashed-dotted line plots the coefficients of the non-linear absolute value interaction term, $\hat{\zeta}_h$, together with their 90% confidence interval bands.

[9] As Ramey (2016) points out, relaxing this assumption would otherwise lead to a number of puzzles. A contractionary monetary policy shock would actually be expansionary for about a year and produce a very pronounced price puzzle.

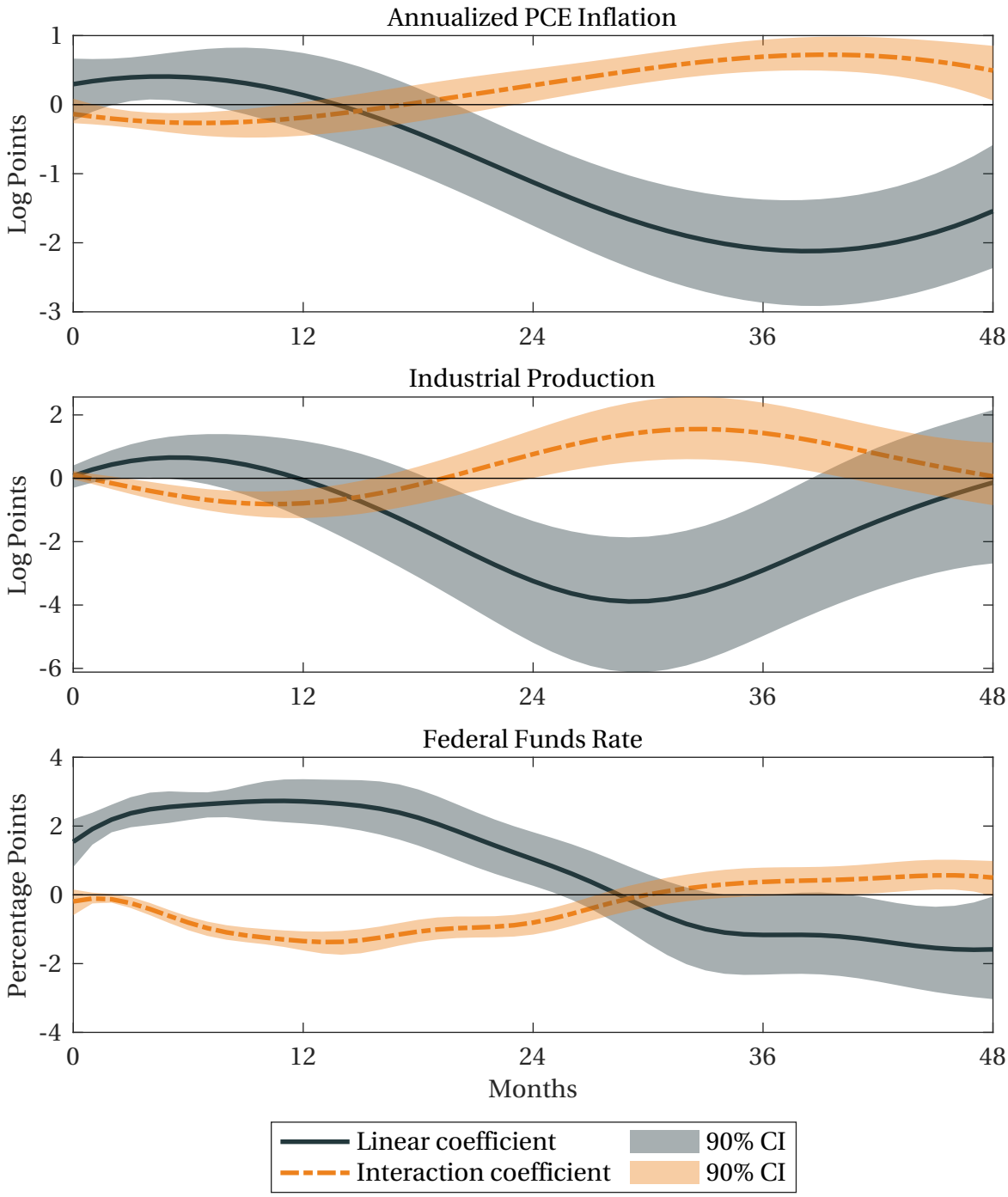


Fig. 1.1: Smooth local projection coefficients for annualised PCE inflation, industrial production and the Federal funds rate. Every panel depicts both the point estimates of the linear coefficient (solid line) and the absolute value interaction coefficient (dashed-dotted), together with their 90% confidence intervals. The coefficients are depicted over a four-year horizon.

The linear terms in the projection deliver a familiar picture in the top panel. After a positive (contractionary) monetary policy shock the linear coefficients yield an initially muted response of inflation followed by significant decline thereafter. There is an initial positive response, however it is not significant on impact and marginally so for just few months in the initial year. As such, the response displays a slight price puzzle if we were to consider only the linear effect.

The non-linear effects implied by the absolute value interaction coefficients are supporting the theoretical predictions. Initially the $\hat{\zeta}_h$ coefficients are negative, counteracting the price puzzle as the shock size increases and indicating a quicker negative response of inflation on impact. Moreover, after about two years, the coefficients on the non-linear term turn positive (and increase), meaning that inflation responds less to a large shock at long horizons. Hence, the evidence suggests that inflation has a stronger reaction at short horizons and a weaker one at long horizons. This is consistent with theoretical models of state-dependent pricing as, for large shocks more firms adjust immediately and hence less adjust later on. Most importantly, the confidence interval shows that the coefficients are statistically significant for most of the horizons.

The other two panels report the results for the linear coefficient in the industrial production and Federal funds rate local projection. These results are in line with the theoretical implications from the literature. The linear output coefficient starts to fall (after a small, positive, but not significant, initial response), reaching its trough two and a half years after the shock, and then recovers. The linear Federal funds rate coefficient exhibits a hump-shaped response and remains positive for more than two years after the shock before turning negative.

Further, the green dashed-dotted line in the middle panel shows that the output response supports the predictions of state-dependent pricing models. The coefficients on the non-linear term in (1.1) are negative at the beginning, counteracting the small positive output response, and then turn positive at longer horizons. As such, the $\hat{\zeta}_h$ estimate counteracts the linear response for both short and long horizons, flattening the overall response of output. While the top panel shows that large monetary policy shocks predict a higher degree of price flexibility, the middle panel shows that larger monetary policy shocks have weaker real effects. Again, the confidence band indicates that the non-linear interaction term is statistically significant.

It is crucial to note that these results are not due to a stronger response of monetary policy, because the coefficients of the Federal funds rate on the non-linear term in the bottom panel are negative for most of the short to medium horizons, suggesting a

proportionally weaker response of monetary policy to a larger shock.

In sum, prices exhibit a non-linear, size-dependent impulse response function, reacting strongly at short and weakly at long horizons for large shocks. Coherently, the output response seems to be smaller, and monetary policy feedback does not seem to drive the above results. As such, we interpret these findings as evidence in favor of state-dependent prices as an important propagation mechanism of monetary policy, as larger shocks induce more firms to change prices early, thus reducing the real effects of a monetary shock.

Impulse Response Functions. In order to further assess and clarify the importance of the non-linear effect, Figure 1.2 compares simulated IRFs to different shock sizes of all three headline variables over a four-year horizon. It depicts the impulse responses for a 25 (dashed line), 100 (solid black line) and 200 (dashed-dotted line) basis point shock, where each impulse is standardised by dividing it by the respective shock size. The standardised IRFs clearly visualise that inflation responds more strongly to a larger shock at short horizons, but then the reaction is less persistent so that the response is weaker at long horizons, as theory would predict. First, the larger the shock, the quicker inflation decreases. Second, for a large enough shock, the initial price puzzle on impact tend to disappear.[10]

The response of output to small and large shocks also supports the theoretical prediction. The standardised IRFs show that the trough in output is smaller relative to the size of the shock, consistent with the behavior of the responses of inflation. Recall that, in state-dependent models, there are two opposite effects on output. First, the larger the shock, *ceteris paribus*, the larger the response of output. This standard effect is the only one also present in time-dependent pricing models. Second, the larger the shock, the greater the number of firms that adjust the price, the larger the response of inflation and the smaller the one of output. This second effect is absent in time-dependent models. Therefore, the output response to a large shock is proportionally flatter in state-dependent models because of this second effect that counteracts the first one. By showing the response relative to the size of the shock, the standardised IRFs isolate the second effect,

[10] Figure A.5, discussed in Appendix A.3.3, depicts the unscaled (i.e., non standardised) impulse responses for a 25 basis point shock and a 200 basis point shock and their 90% confidence interval calculated with the Delta method, respectively (see Appendix for details). The response of inflation to a 200 basis point shock is initially not significantly different from zero and then significantly negative, while it is positive for some months for a 25 basis point shock (even if only marginally significant). Consequently, a sufficiently large shock counteracts the small linear coefficient and switches the sign of the overall impulse response of inflation, removing a potential price puzzle.

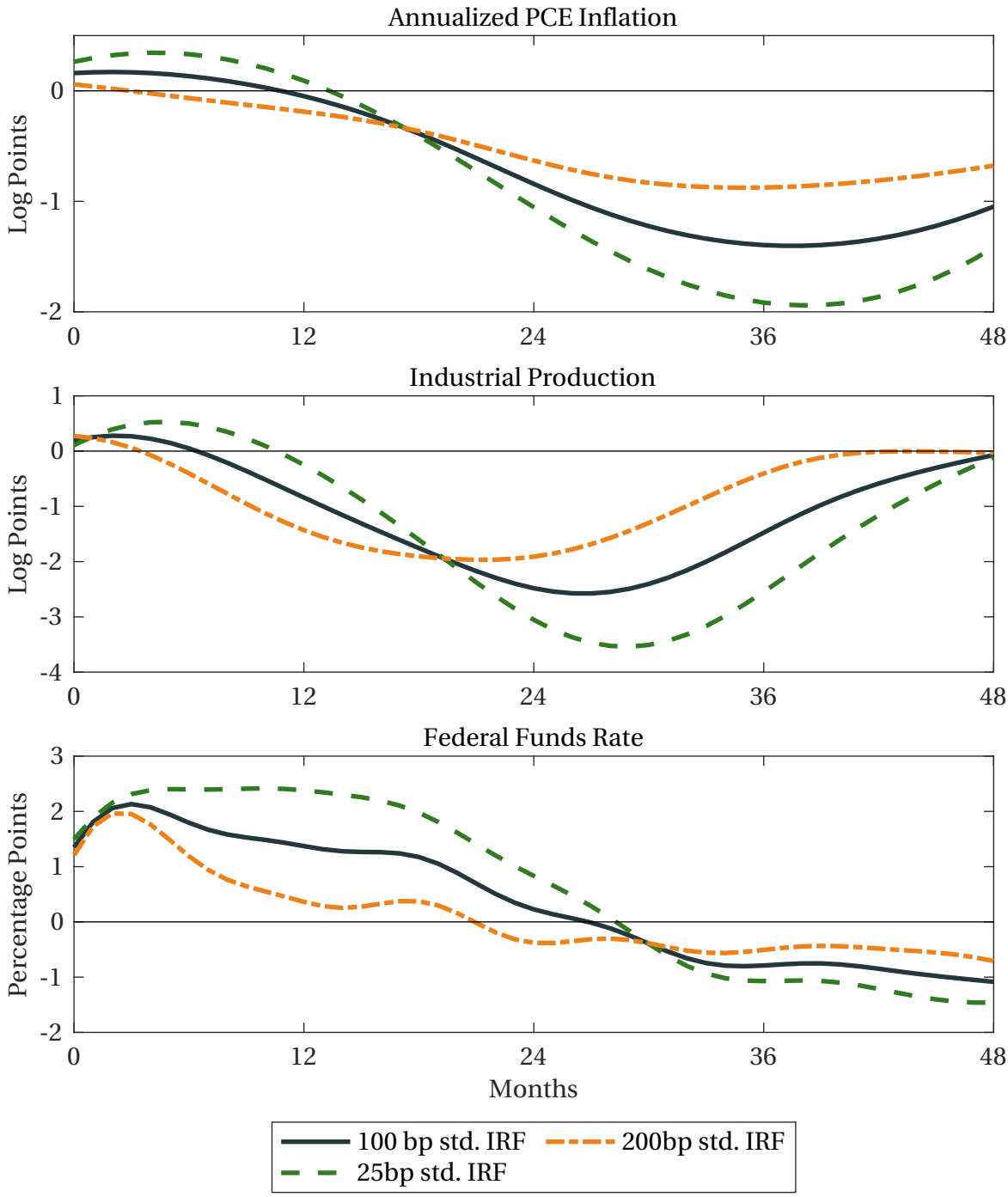


Fig. 1.2: Panel of simulated size-dependent impulse responses for annualised PCE inflation, industrial production and the Federal funds rate over a four-year horizon. The Figure depicts the impulse response for a 25 (dashed line), 100 (solid) and 200 (dashed-dotted) basis point shock, rescaled by dividing by the size of the shock. The impulse responses are depicted over a four-year horizon.

thus revealing whether there is a significant effect coming from state-dependent pricing. Finally, the standardised IRFs also highlight the effect of state-dependent pricing both on the scale, i.e. decreasing real impact as the shock gets larger, and on the timing, i.e. arriving sooner with larger shocks, of the response of output and inflation to the policy shocks.

Standardized cumulative impulse responses for a 25bp and 200bp shock

	PCE Deflator				
Horizon	$h=1$	$h=12$	$h=24$	$h=36$	$h=48$
$\hat{\beta}_h+\hat{\zeta}_h\lvert 0.25\rvert$	0.02	0.23	−0.16	-1.75^{**}	-3.57^{***}
	(0.03)	(0.21)	(0.51)	(0.81)	(1.13)
$\hat{\beta}_h+\hat{\zeta}_h\lvert 2.00\rvert$	-0.03^{***}	-0.17^{*}	-0.43^{**}	-1.20^{***}	-1.83^{***}
	(0.01)	(0.09)	(0.20)	(0.41)	(0.57)
	Cumulative Industrial Production				
Horizon	$h=1$	$h=12$	$h=24$	$h=36$	$h=48$
$\hat{\beta}_h+\hat{\zeta}_h\lvert 0.25\rvert$	0.18	3.29	−16.32	-57.63^{**}	-69.67^{**}
	(0.15)	(4.54)	(14.31)	(25.43)	(34.13)
$\hat{\beta}_h+\hat{\zeta}_h\lvert 2.00\rvert$	0.25^{***}	-7.75^{***}	-30.13^{***}	-40.53^{***}	-38.98^{***}
	(0.07)	(1.64)	(5.37)	(8.52)	(11.54)

Table 1.1: Impulse responses of cumulative PCE inflation, i.e. the PCE deflator, and cumulative industrial production after a 25bp and a 200 bp point shock, standardized by the respective size of the shock. Newey-West standard errors in parentheses.
***: Significant at the 1% level; **: Significant at the 5% level; *: Significant at the 10% level

To strengthen our point, Table 1.1 shows the cumulative effect of a monetary policy shock on inflation (i.e. the PCE deflator) and on output for a small and a large shock, standardised by the size of the shock, and their significance levels, at different horizons: 1, 12, 24, 36 and 48 months.[11] Coherent with the theoretical predictions, prices move significantly downwards on impact for large shocks, while they do not for small shocks. Due to this lagged and inertial behavior in case of small shocks, the cumulative response of the price level after four years is almost doubled, relative to the size of the shock, in comparison with the response to a large shock. The cumulative response of output reflects the response of the price level. The initial response of output is sharper for large shocks, but the cumulative drop in output after four years is about 80% proportionally larger for small shocks.

[11] We simply cumulate the point estimates from the standard least squares estimation and statistics using the Delta method.

Finally, the last row in Figure 1.2 again illustrates that the empirical results are not due to the different behavior of monetary policy after a large shock. The standardised IRFs of the Fed funds rate is milder for larger shocks, relative to shock size, and hence, if anything, it would play against our results.[12]

Figure 1.2 and Table 1.1 reinforce our previous empirical results concerning the significant size-dependent effects of monetary policy shocks. Large shocks induce firms to change prices early on and thus reduce the real effects of such a monetary shock, in accordance with our first theoretical prediction.

1.4.2 Implication 2: Trend inflation-dependent effects of monetary policy shocks

We use smooth transition local projections to test whether the impulse responses after a monetary policy shock are different in high and low inflation regimes. Auerbach & Gorodnichenko (2012a) and Tenreyro & Thwaites (2016) popularised this method, and we follow their approach to a large extent. The impulse response of the variable of interest y_t at horizon h in state $s = \mathrm{HI}, \mathrm{LO}$[13] to a unitary structural shock e_t is the estimated coefficient β_h^s in:

$$\begin{aligned} y_{t+h} = \tau_h t + \mathrm{F}(z_t)(\alpha_h^{\mathrm{HI}} + \beta_h^{\mathrm{HI}} e_t + \sum_{k=1}^{\mathrm{K}} \gamma_{h,k}^{\mathrm{HI}} w_{t,k}) \\ +(1-\mathrm{F}(z_t))(\alpha_h^{\mathrm{LO}} + \beta_h^{\mathrm{LO}} e_t + \sum_{k=1}^{\mathrm{K}} \gamma_{h,k}^{\mathrm{LO}} w_{t,k}) + u_{t+h}, \end{aligned} \tag{1.2}$$

for $h = 0, 1, ..., \mathrm{H}$. Again, our set of controls $\{w_{t,k}\}_{k=1}^{\mathrm{K}}$ includes the contemporaneous values of industrial production and PCE inflation and up to two month lags of industrial production, PCE inflation and the effective Federal funds rate. $\mathrm{F}(z_t)$ is a smooth transition function which indicates the state of the economy (Granger & Teräsvirta, 1993). We use a logistic function with the following form:

$$\mathrm{F}(z_t) = \frac{\exp(\gamma \frac{(z_t - c)}{\sigma_z})}{1 + \exp(\gamma \frac{(z_t - c)}{\sigma_z})} \in [0, 1]. \tag{1.3}$$

[12] Moreover, the response of the Federal funds rate is statistically different (in the first year) between the 25 versus 200 basis points shock (see Figure A.5 in the Appendix).

[13] HI stands for high inflation, LO for low inflation.

If state-dependent prices are an important aggregate propagation mechanism we would expect β_h^{HI} to be statistically significantly more negative than β_h^{LO} for the response of the price level and inflation, especially at short horizons. Prices should be more flexible and so react both more quickly and strongly to monetary policy shocks in a high inflation regime.

In the main specification of our smooth transition function local projection, i.e. equation (1.2), the state variable z_t represents smoothed personal consumption expenditure (PCE) inflation, so we take a 24-month centered moving average (MA) to capture trend inflation.[14] We set $\gamma = 5$ as this gives an intermediate degree of regime switching intensity. This is relatively standard in the literature and also fits our inflation data well. Finally, c corresponds to the 75th percentile of the historical trend inflation distribution. This is equivalent to assuming that about 70% of the time trend inflation is classified as negligible (i.e. $F(z_t) \in [0, 0.1]$) and 30% of the time there is some trend inflation (i.e. $F(z_t) \in (0.1, 1]$).

Figure 1.3 displays the resulting smooth transition function, $F(z_t)$. The solid line, measured on the left vertical axis, shows PCE inflation on an annual basis. The dashed line, measured on the right vertical axis, depicts the smooth transition function based on our MA-filtered measure of PCE inflation. The period of the Great Inflation from around 1974 to 1983 is characterised by two pronounced spikes of inflation of up to 11%. The smooth transition function reaches 1 around these two peaks and stays above 0.4 for the entire period of the Great Inflation, classifying the latter period mostly as a high inflation regime. We take this to be a reasonable approximation for periods of high and low trend inflation in the United States.[15]

Coefficient estimates. The panels of Figure 1.4 display the coefficient from estimating equation (1.2) for each response variable (by row). The first column shows the point estimates of the $\hat{\beta}_h^i$ coefficients for the linear (solid line), the high inflation (dash-dotted line) and low inflation (dashed line) regimes. Column two and three depict the impulse responses conditional on the high inflation and low inflation regime respectively, with their 90% confidence intervals. The last column displays the t-statistic that tests the null of equality of the high and low inflation regime coefficients, i.e., $\hat{\beta}_h^{HI} = \hat{\beta}_h^{LO}$, where the gray area represents the 90% z-values. A positive value means that the high inflation response

[14] The moving average is our benchmark smoothing procedure. However, the results from the local projections are very similar with an HP-Filter ($\lambda = 14400$) smoothing procedure, see the Appendix.

[15] This calibration is also in accordance with Álvarez et al. (2019), who show that the frequency of price changes starts increasing significantly from annual inflation rates of 5%. Our smooth transition function indicates a value of approximately 0.5 with such an annualised inflation rate.

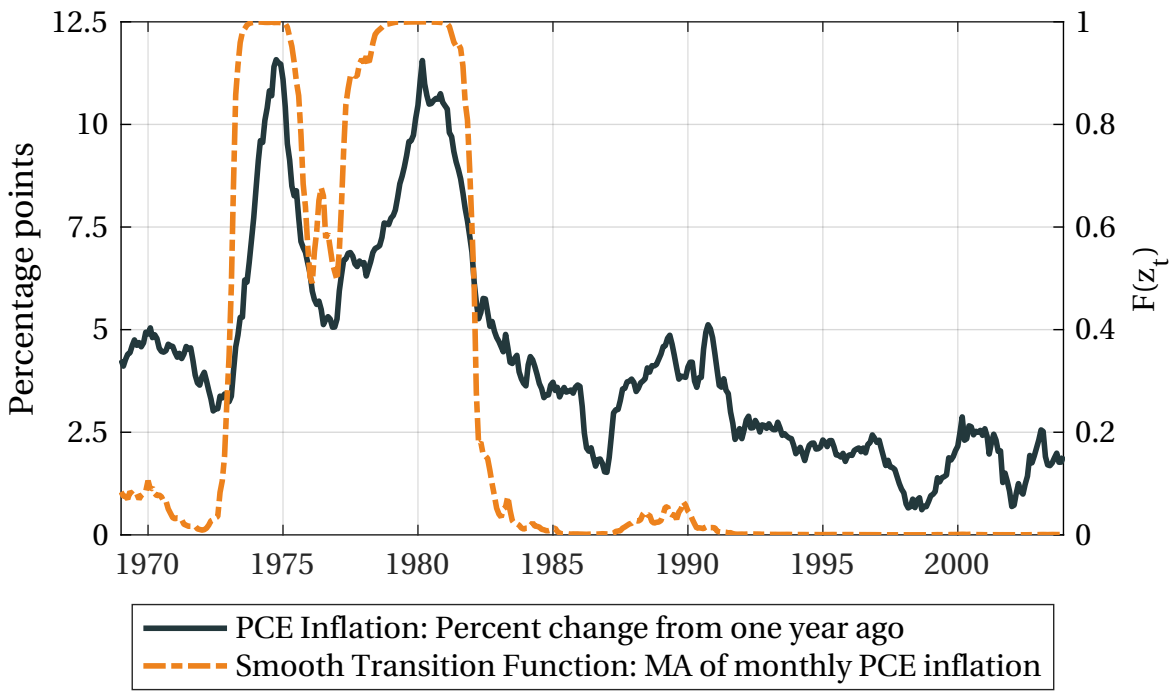


Fig. 1.3: The solid line plots PCE inflation, measured in percentage change from one year ago on the left axis. The dashed line shows the benchmark smooth transition function based on the two-year centered moving average of monthly PCE inflation, measured on the right axis. Both are depicted from January 1969 until December 2003.

is larger whereas a negative value of the t-statistic indicates the opposite. First, the linear terms for PCE inflation in the top left panel show the familiar picture in the literature. PCE inflation declines eventually, after an initial positive response, however it is not statistically significant. Second, inflation in a high inflation regime declines right away on impact after a contractionary monetary policy shock. On the contrary, the impulse response function in a low inflation regime exhibits a price puzzle for about one year, which is marginally statistically significant for only few months. This suggests that in this regime firms are not willing to change price as frequently, so the price level stays persistently around zero for a longer period. Moreover, at long horizons the response is smaller in a high trend inflation regime. Again, this is consistent with the idea that the effect is less persistent in a high inflation regime, because more firms adjust on impact after the shock. Third, the last column shows that the responses of inflation in a high and low inflation regime are statistically significantly different both at short and at long horizons. We interpret these results as evidence in favor of state-dependent prices models as a key propagation mechanism of monetary policy shocks, because they predict a faster, and less persistent, reaction to a monetary disturbance in a high trend inflation regime. This is exactly what the impulse response functions show.

The first panel in the second row shows that the IRFs for output exhibit the usual hump shaped dynamics. Output reacts with a larger delay in a low inflation regime compared to a high inflation regime, but this reaction is stronger and reaches a trough after two years which is roughly twice as deep as the one in a high inflation regime. The difference in the IRFs between the two regimes however is not statistically significant, so we do not find evidence for our second theoretical implication regarding output. Table 1.2 displays similar results for the cumulated IRFs of inflation and output at different horizons. Prices drop from the outset in a high inflation regime, while the reaction is sluggish in a low inflation regime. However, the latter is more persistent so that it eventually catches up and overtakes the cumulative drop in a high inflation regime, so that the cumulative drop after four years is 50% higher. The difference in the initial response, up to two years, is statistically significant providing supporting evidence for state-dependent pricing. This is not the case for the cumulative response of output, that is both not significant for low inflation regimes and not statistically different between low and high inflation regime. Again, the point estimates are in line with the theoretical prediction, but the standard errors get so large (especially for the low inflation regime, as evident also from Figure 1.4) that these differences are not significant.

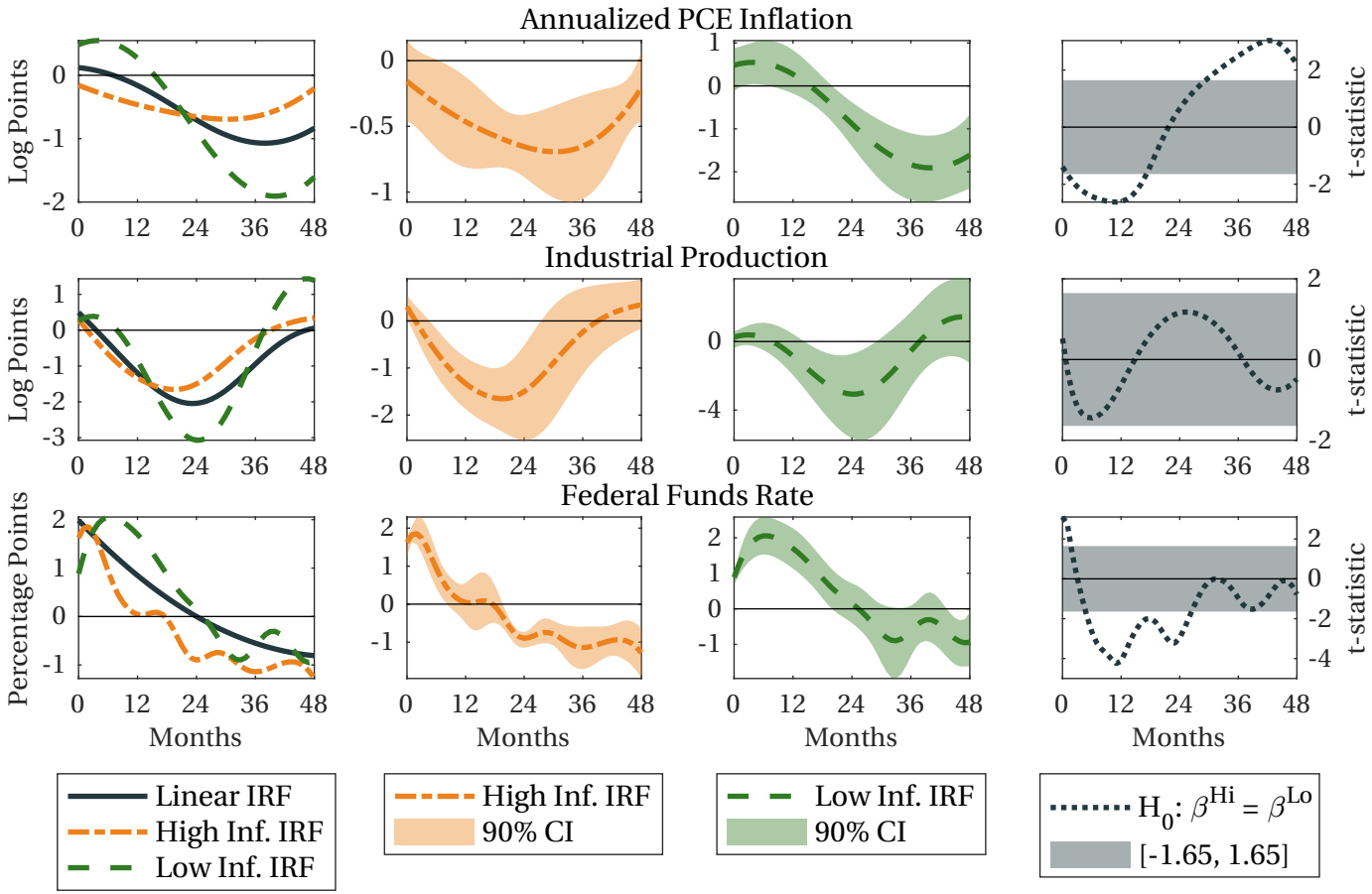


Fig. 1.4: Panel of smooth impulse responses in different inflation states for annualised PCE inflation, industrial production and the Federal funds rate. Column 1 depicts the point estimates for the linear (solid line), high inflation (dash-dotted) and low inflation (dashed) impulse response. Column 2 and 3 depict the high inflation and low inflation impulse responses, together with their 90% confidence bands. Column 4 depicts the t-statistic for the null hypothesis of equality of the high and low inflation responses (dotted), together with 90% z-values (shaded area). The impulse responses are depicted over a four-year horizon.

Cumulative impulse responses in high and low inflation regimes

	PCE Deflator				
Horizon	$h = 1$	$h = 12$	$h = 24$	$h = 36$	$h = 48$
$\hat{\beta}_h^{HI}$	-0.06^{***}	-0.40^{**}	-0.92^{***}	-1.60^{***}	-2.06^{***}
	(0.02)	(0.19)	(0.25)	(0.20)	(0.24)
$\hat{\beta}_h^{LO}$	0.08^{*}	0.49	0.39	-1.31^{**}	-3.10^{***}
	(0.05)	(0.34)	(0.51)	(0.57)	(0.83)
t-stat: $\beta_h^{HI} - \beta_h^{LO}$	-2.60^{**}	-2.21^{**}	-2.61^{**}	-0.47	1.20
	Cumulative Industrial Production				
Horizon	$h = 1$	$h = 12$	$h = 24$	$h = 36$	$h = 48$
$\hat{\beta}_h^{HI}$	0.06	-8.49^{***}	-27.14^{***}	-36.40^{***}	-34.92^{***}
	(0.07)	(2.75)	(5.35)	(9.74)	(13.32)
$\hat{\beta}_h^{LO}$	0.25	1.87	-26.22	-54.52	-41.12
	(0.29)	(5.99)	(18.03)	(35.77)	(41.27)
t-stat: $\beta_h^{HI} - \beta_h^{LO}$	-0.60	-1.39	-0.05	0.51	0.15

Table 1.2: Impulse responses of cumulative PCE inflation, i.e. the PCE deflator, and cumulative industrial production in the high and low inflation regimes. The last rows report the t-statistic with the null hypothesis of the two coefficients being equal. Newey-West standard errors are in parentheses.
***: Significant at the 1% level, **: Significant at the 5% level; *: Significant at the 10% level

Finally, the panels in the third row show that the interest rate increases after a monetary policy shock and stays positive for about two years in the linear case and low inflation regime, but only for one year in the high inflation regime. Thus, monetary policy initially reacts differently to a shock in the two regimes, and these differences are statistically significant for the first two years. In a high inflation regime, the nominal interest rate initially reacts more, but then it decreases much faster than in a low inflation regime. While one might argue that this pattern may explain the initially quicker reaction of the price level in the high inflation regime, prices in the low regime react considerably more sluggishly even though the interest rate is positive for a longer period of time. Indeed, the last column shows that for most of the IRFs the coefficients in the low inflation regime are larger than the ones in a high inflation regime, perhaps signalling a stronger endogenous feedback of monetary policy in response to the shock (as also evident from the IRFs in column one). On the one hand, it would be hard to argue that the different monetary policy behavior is driving the wedge in prices between the high and low inflation regimes. On the other hand, the fact that the path of the Federal funds rate after the initial monetary contraction is different across the two regimes blurs the comparison, possibly explaining

why the evidence for the differences in output responses is not statistically significant.

To conclude, we find evidence in favor of state-dependent prices regarding our second testable implication with respect to the behavior of inflation. Our results show that in a high trend inflation regime, inflation declines right away after a policy shock, and there is no price puzzle, as theory predicts. Moreover, inflation is more persistent in a low inflation regime, despite the interest rate staying positive for a longer amount of time. Regarding the response of output, however, the point estimates are coherent with the theoretical prediction, but the differences between the high and low inflation regime are not statistically significant.

1.5 Robustness

This section reports the results of three particularly important robustness exercises regarding our local projection estimates. First, we present our results using a panel smooth local projection analysis with seventeen sectoral price series. Second, we conduct stability exercises of our main regression coefficients. Finally, we exclude the NBR targeting period from our analysis. Finally, we conduct further comprehensive robustness checks on many other dimensions, that we confine to the Appendix, and we briefly summarise these below.

1.5.1 Sectoral evidence

Our analysis so far focussed on the aggregate evidence of state-dependent pricing. Yet, if there is a significant degree of state-dependent pricing in the aggregate, similar results should also emerge on a more disaggregated level. In order to investigate this, we apply the smooth impulse response function analysis to a panel of the seventeen sectoral price and quantity indices that comprise the aggregate Personal Consumption Index, following Stock & Watson (2016).[16] Our two main variables of interest here are the monthly sectoral Personal Consumption deflator and the monthly sectoral Real Personal Consumption index between 1969 and 2007, as before. We create the latter by deflating the nominal personal consumption with the respective sectoral price index. We then demean all series by their sector specific average and run similar local projections using the Romer & Romer (2004) shocks. In doing so, we are the first, to the best of our knowledge, to estimate fixed effects with the technique of Barnichon & Brownlees (2019).

[16] Table A.4 in the Appendix described the seventeen components used in this exercise

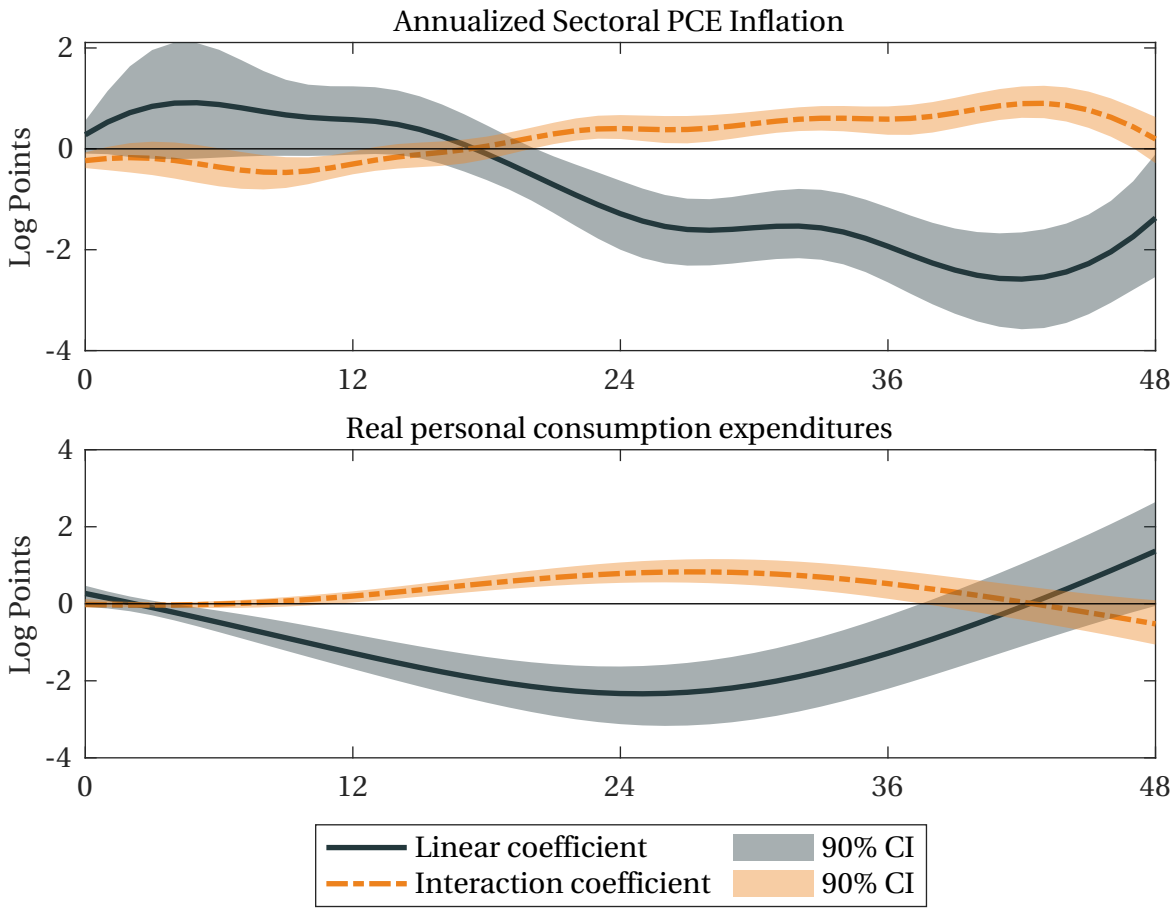


Fig. 1.5: Smooth local projection coefficients for annualised sectoral PCE inflation, and real personal consumption expenditures using sector fixed effects. Every panel depicts both the point estimates of the linear coefficient (solid line) and the absolute value interaction coefficient (dashed-dotted), together with their 90% confidence intervals. The coefficients are depicted over a four-year horizon.

Size dependence on a sectoral level. In order to test our first hypothesis we run a similar local projection to (1.1) with sector fixed effects. We use contemporaneous controls of the both the annualised sectoral PCE inflation, and real personal consumption expenditures. Furthermore, we include a trend and two lags of the sectoral real personal consumption expenditure, the sectoral PCE inflation, the Fed funds rate and aggregate industrial production, respectively. The results are depicted in Figure 1.5. As one can see, the linear coefficient on sectoral PCE inflation again is first marginally positive and then turns negative. The absolute value interaction coefficient (dashed-dotted line) is again first negative, then positive, in line with our aggregate results. The linear coefficient in the real personal consumption expenditure projection is U-shaped, as expected, whilst the absolute value interaction coefficient dampens the consumption response for larger shocks, as it is positive for most of the horizons. So, the sectoral results with respect to our first hypothesis again confirm the aggregate results and point to the presence of state-dependent pricing in US data.

Inflation dependence on a sectoral level. Second, we repeat our test for trend-inflation dependence in the sectoral data. We use the Stock & Watson (2016) univariate unobserved components and stochastic volatility outlier-adjustment model to estimate trend inflation at a sectoral level. We then normalize them between 0 and 1 and use them as sector-specific trend inflation measures $F(z_t)$ in this exercise.[17] The set of controls is the same aforementioned. As above, by demeaning the data we include fixed effects and estimate the sectoral equivalent to (1.2). The results are depicted in Figure 1.6. First, with respect to prices, we see again that sectoral inflation falls quicker in the high inflation regime and recovers more quickly, too, in line with the second prediction. Further, and unlike the aggregate data, here we also find weaker real effects. Sectoral consumption responds by less and recovers quicker in the high inflation regime. This is additional and complementary evidence to support our second test; state-dependent pricing shows up in US data as real effects weaken and inflation responds more price flexible in high inflation regimes, even in sectoral data.

[17] These results are also robust to using the original measure

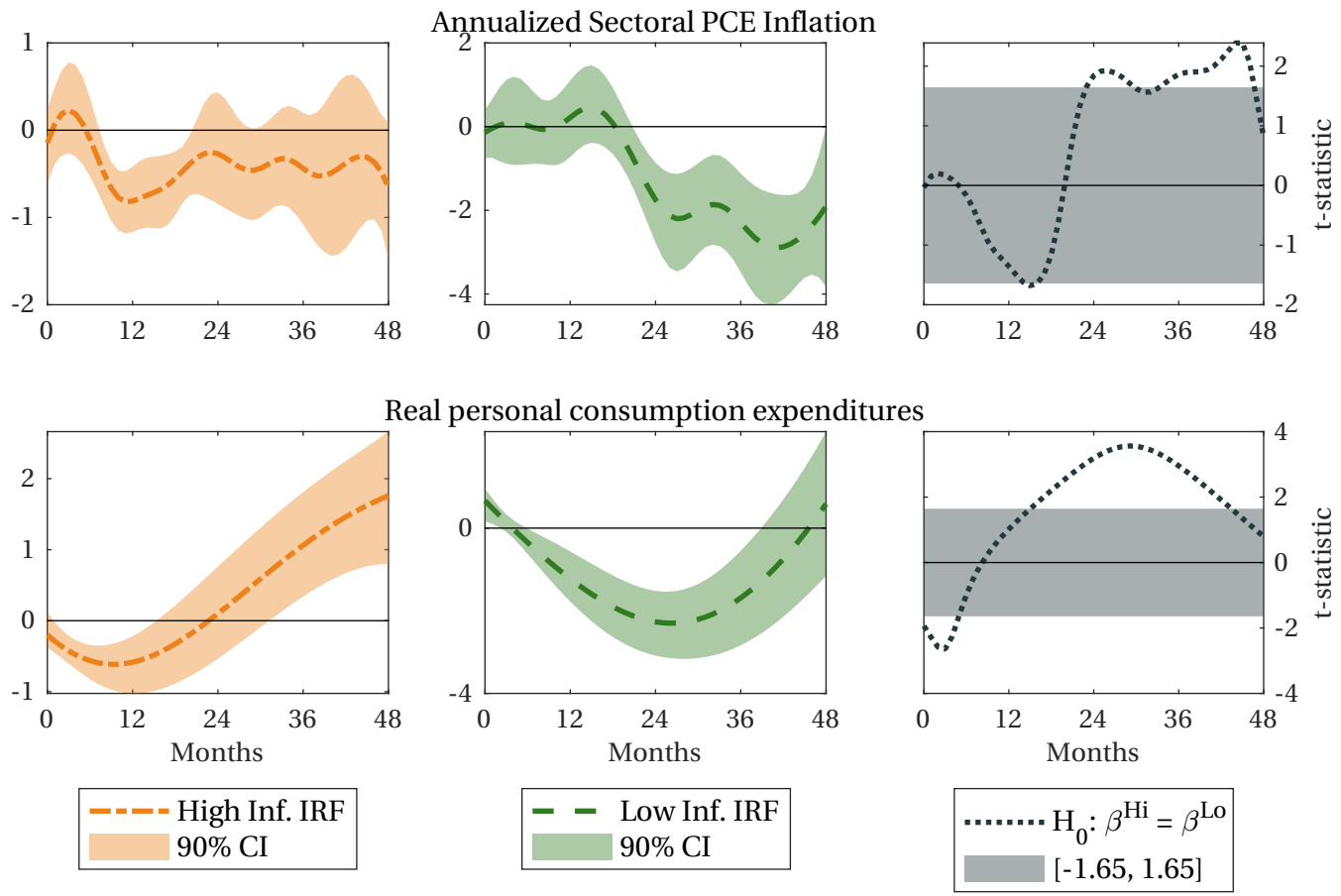


Fig. 1.6: Panel of smooth impulse responses in different inflation states for annualised sectoral PCE inflation and real consumption expenditures. Column 1 and 2 depict the high inflation and low inflation impulse responses, together with their 90% confidence bands. Column 3 depicts the t-statistic for the null hypothesis of equality of the high and low inflation responses (dotted), together with 90% z-values (shaded area). The impulse responses are depicted over a four-year horizon.

1.5.2 Stability

It is crucial for our analysis to be relatively confident that the non-linear and state-dependent dynamic behavior of the impulse response functions of output and inflation is not driven by a feedback effect of monetary policy.[18] To show that monetary policy feedback plays a limited role with respect to large and small shocks, we check if inflation coefficients with respect to the linear and absolute value interaction term do not exhibit structural breaks when monetary policy conduct may have changed. If these estimates stay stable over a variety of monetary policy regimes throughout time, then we can reject the thesis of monetary feedback driving the result of size-dependent effects.

We apply the Hansen (1992) stability test to the coefficients in the local projection (1.1) for PCE inflation. This test has locally optimal power and needs no *a priori* assumption concerning the breakpoint. Furthermore, this test is robust to heteroscedasticity, a potential concern in this analysis. Table 1.3 shows the results for the two coefficients of interest and the joint test for parameter stability at a 1, 3, 6, 12, 24 and 36 month horizon for our PCE inflation local projection.[19] We cannot reject the null of individual parameter constancy for neither the linear nor the absolute value interaction coefficients at these horizons (except for the linear coefficient at a 36-month horizon). However, the joint test statistic does indicate a rejection of the null that all parameters in the local projection are constant. This suggests that, even though the dynamic feedback of monetary policy via lagged control values may have changed throughout time, the shape and non-linearity of the impulse response after a monetary policy shock has stayed relatively constant throughout the sample. The same results hold for the linear and interaction coefficients when the dependent variable is either the industrial production or the Fed funds rate (see Appendix).[20]

[18] As discussed in section 1.4.1, several of our results seem to contradict this possibility. For example, monetary policy seems to react weakly after a large shock whereas prices react by more at the beginning of the horizon. The same applies to a large extent to the test of the second implication in section 1.4.2 as the funds rate stays positive for longer in the regime where prices seem to be more sticky. These patterns seem to indicate that the dynamics of inflation and output are not defined by the reaction of monetary authorities only.

[19] For a test on an individual coefficient, we can reject the null of parameter constancy at the 5% significance level, if the relevant test-statistic is larger than the asymptotic critical value of 0.47. The null hypothesis is that each coefficient in (1.1) is constant, the respective distribution is non-standard and depends on the number of parameters tested for stability. The intuition is that under the null hypothesis the cumulative sums of the first-order conditions from the estimation will have mean zero and wander around zero. However, under the alternative hypothesis of parameter instability, these first-order conditions will not be mean zero for parts of the sample and so the test statistic will be large, leading us to reject the null (see Appendix).

[20] These results substantiate the visual impression from the plots of the recursive estimation of the local

PCE Inflation Local Projection - Hansen (1992) test statistic

Horizon	$h = 1$	$h = 3$	$h = 6$	$h = 12$	$h = 24$	$h = 36$
Linear coeff.	0.10	0.04	0.08	0.04	0.09	0.54^{**}
Interaction coeff.	0.01	0.03	0.02	0.01	0.01	0.06
Joint: all coeffs.	3.13^{**}	3.04^{**}	3.68^{***}	5.88^{***}	9.43^{***}	9.76^{***}

Table 1.3: Estimated Hansen (1992) test statistics for parameter constancy of the PCE inflation local projection with both a linear and an absolute value interaction shock term. The first row reports the individual test statistic for the linear coefficient $\hat{\beta}$ at different horizons, the second row reports those for the absolute value interaction coefficient $\hat{\zeta}$ and the final row reports the test statistic for the joint hypothesis of all parameters (ie. regression coefficients and variance) to be constant.
Individual critical values are $c_{1,1\%} = 0.75$, $c_{1,5\%} = 0.47$ and $c_{1,10\%} = 0.35$. Joint critical values for a model with K = 12 parameters are $c_{12,1\%} = 3.51$, $c_{12,5\%} = 2.96$ and $c_{12,10\%} = 2.69$.
***: Significant at the 1% level; **: Significant at the 5% level; *: Significant at the 10% level

1.5.3 Excluding the NBR targeting period

Coibion (2012) and others have suggested that the exclusion of the NBR targeting period October 1979 and September 1982 can account for a difference in results between the Romer & Romer (2004) and VAR approach. Critically, the largest absolute monetary shock values lie in this period and may thus play a significant role for the conclusion on our first implication. In order to account for this suggestion we exclude this part of the shock sample, modifying the non-linear local projection equation (1.1) by interacting the linear and the non-linear coefficient with a time-dummy that takes a value of 1 for the sample between October 1979 and September 1982 and 0 otherwise. As in Figure 1.2, the three rows in Figure 1.7 display the IRFs for this specification for PCE inflation, industrial production and the Fed funds rate, respectively, for both the coefficients of the linear term, $\hat{\beta}_h$ (green, solid line), and the ones on the non-linear absolute value interaction term, $\hat{\zeta}_h$ (blue, dashed line). The impulse response coefficients are qualitatively similar to the benchmark results. The results, however, change somewhat both in terms of magnitude and, especially, in terms of significance, for the interaction coefficients. This is hardly surprising. The reduced sample does not include the large shocks of the NBR targeting period, and it features low sample variation with values mostly below 1. Hence, the non-linear effects of large shocks are more difficult to identify, inducing wide confidence intervals.

projection coefficients displayed in the Appendix.

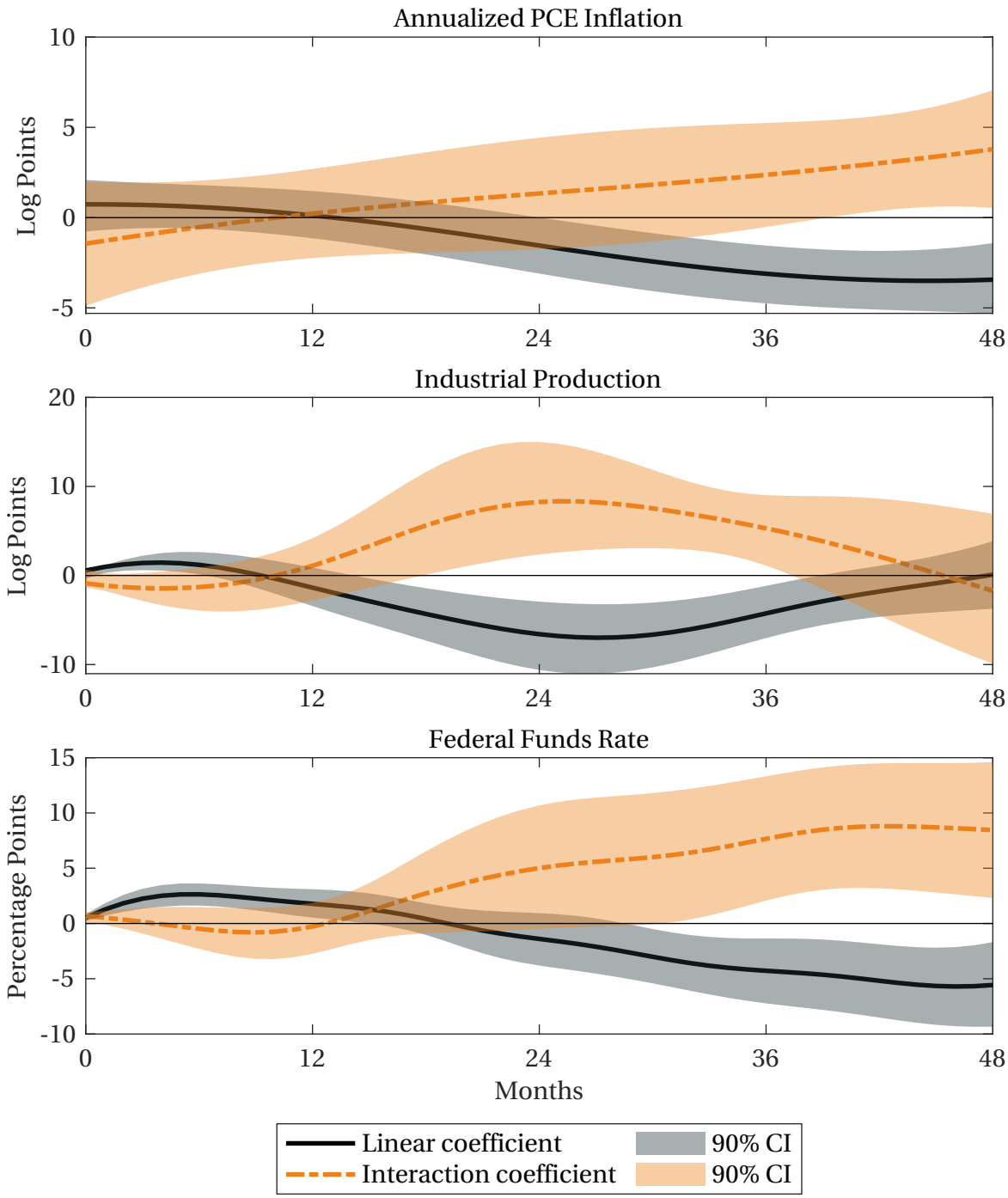


Fig. 1.7: Smooth local projection coefficients for annualised PCE inflation, industrial production and the Federal funds rate, excluding the NBR targeting period between October 1979 and September 1982. Every panel depicts both the point estimates of the linear coefficient (solid line) and the absolute value interaction coefficient (dashed-dotted), together with their 90% confidence intervals. The coefficients are depicted over a four-year horizon.

1.5.4 Other robustness checks

The Appendix presents comprehensive robustness checks. It displays the plots of the recursive estimates of the local projection coefficients in (1.1), to further visually inspect their stability. In addition, results are robust to a different specification of the non-linear terms in (1.1), which includes a squared and a cubed term for the shock value, rather than the absolute value interaction term.

Results for the smooth transition local projections (1.2) are robust to: 1) changes in γ and c; 2) using HP-filtered inflation for z; 3) using a model based trend inflation measure from Ireland (2007).

Finally, the estimates of the coefficients in both (1.1) and (1.2) are also robust to: 1) using the CPI instead of the PCE; 2) using quarterly data with GDP as the output measure; 3) including as controls: (i) the commodity price index, (ii) the corporate bond credit spread by Gilchrist & Zakrajsek (2012) to control for financial frictions, (iii) proxies for fiscal policy using either measure of excess returns on stocks of military contractors from Fisher & Peters (2010) or the exogenous tax changes from Romer & Romer (2010); 3) different measures of shocks obtained from non-linear models: (i) from the Romer & Romer (2004) regression using the smooth transition function; (ii) from a smooth transition VAR; 4) including leads and lags of the shocks to control for potential autocorrelation and misspecification as suggested by Alloza et al. (2019).

1.6 Conclusion

The assumption of sticky prices lies at the very center of the current workhorse model for the analysis of business cycle fluctuations and, in particular, monetary policy effects. The literature features two type of sticky price models: time-dependent and state-dependent prices. A sticky price theory of the transmission mechanism of monetary policy shock based on state-dependent pricing yields two testable implications that do not hold in time-dependent models; the impulse response function of the aggregate price level and inflation should be more flexible both after a large shock and during high trend inflation regimes. Employing the methodology of local projections, we tested these predictions on aggregate US data. We found some evidence in favor of state-dependent models of price stickiness rather than time-dependent ones. This was also supported in sectoral price data. Regarding the response to large shocks, the coefficient of the absolute value interaction shock projections matched our theoretical prior, both in terms of output and inflation.

When the NBR targeting period of US monetary policy, October 1979 and September 1982, is taken out of the sample, our results lose statistical significance, because too little variation is present to identify the non-linear effect and the significance bands becomes very wide. The empirical investigation during large trend inflation regimes also showed that inflation reacts significantly more quickly to a monetary policy shock in times of high trend inflation. In this case, the evidence for output is not significant, however, the point estimates behave according to the theoretical predictions. Furthermore, the Appendix shows that the results are robust to a very large variety of robustness tests.

To the best of our knowledge, our results are the first to point towards a significant presence of state-dependent pricing in the US economy using aggregate data. Our macro evidence is a useful complement to the large empirical literature on individual firm price microdata. Hence, *'Prices are sticky after all'* - just as recent literature has shown (Kehoe & Midrigan, 2015) - but less so when shocks are large or inflation is high. These results are in line with what would be expected if state-dependent pricing played a significant role in the US economy. This supports the theoretical implication that the frequency of changing prices is, at least to some extent, endogenous to the economic environment, as suggested by Álvarez et al. (2017). So, although the Calvo (1983) model may work quite well for "normal" times, when considering situations where high-trend inflation is present or large shocks are likely, a state-dependent sticky price framework that accounts for these phenomena seems more appropriate (as e.g. Álvarez et al., 2016; Álvarez & Lippi, 2014; Costain & Nakov, 2011, 2019; Costain et al., 2019).

References

Alloza, M., Gonzalo, J., & Sanz, C. (2019). *Dynamic effects of persistent shocks.* Working Paper 1944, Banco de España. Working Paper.

Álvarez, F., Beraja, M., Gonzalez-Rozada, M., & Neumeyer, P. A. (2019). From Hyperinflation to Stable Prices: Argentina's Evidence on Menu Cost Models. *The Quarterly Journal of Economics*, 134(1), 451–505.

Álvarez, F., Le Bihan, H., & Lippi, F. (2016). The Real Effects of Monetary Shocks in Sticky Price Models: A Sufficient Statistic Approach. *American Economic Review*, 106(10), 2817–2851.

Álvarez, F. & Lippi, F. (2014). Price Setting With Menu Cost for Multiproduct Firms. *Econometrica*, 82(1), 89–135.

Álvarez, F. & Lippi, F. (2020). Temporary Price Changes, Inflation Regimes, and the Propagation of Monetary Shocks. *American Economic Journal: Macroeconomics*, 12(1), 104–52.

Álvarez, F., Lippi, F., & Paciello, L. (2011). Optimal Price Setting With Observation and Menu Costs. *The Quarterly Journal of Economics*, 126, 1909–1960.

Álvarez, F., Lippi, F., & Passadore, J. (2017). Are State and Time Dependent Models Really Different? In M. Eichenbaum, E. Hurst, & J. A. Parker (Eds.), *NBER Macroeconomics Annual*, volume 31 (pp. 379–457). University of Chicago Press.

Álvarez, F. & Neumeyer, P. A. (2020). The Pass-Through of Large Cost Shocks in an Inflationary Economy. In J. G. Gonzalo Castex & D. Saravia (Eds.), *Changing Inflation Dynamics, Evolving Monetary Policy* chapter 2, (pp. 7–48). Santiago, Chile: Banco Central de Chile.

Álvarez, L. J., Dhyne, E., Hoeberichts, M., Kwapil, C., Le Bihan, H., Lünnemann, P., Martins, F., Sabbatini, R., Stahl, H., Vermeulen, P., & Vilmunen, J. (2006). Sticky Prices in the Euro Area: A Summary of New Micro-Evidence. *Journal of the European Economic Association*, 4(2-3), 575–584.

Auerbach, A. J. & Gorodnichenko, Y. (2012a). Fiscal Multipliers in Recession and Expansion. In A. Alesina & F. Giavazzi (Eds.), *Fiscal Policy after the Financial Crisis* chapter 2, (pp. 63–98). University of Chicago Press.

Auerbach, A. J. & Gorodnichenko, Y. (2012b). Measuring the Output Responses to Fiscal Policy. *American Economic Journal: Economic Policy*, 4(2), 1–27.

Barakchian, S. M. & Crowe, C. (2013). Monetary policy matters: Evidence from new shocks data. *Journal of Monetary Economics*, 60(8), 950–966.

Barnichon, R. & Brownlees, C. (2019). Impulse Response Estimation by Smooth Local Projections. *The Review of Economics and Statistics*, 101(3), 522–530.

Barnichon, R. & Matthes, C. (2018). Functional Approximation of Impulse Responses. *Journal of Monetary Economics*, 99, 41–55.

Bernanke, B. S., Boivin, J., & Eliasz, P. (2005). Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach. *The Quarterly Journal of Economics*, 120(1), 387–422.

Bernanke, B. S., Gertler, M., & Gilchrist, S. (1999). The financial accelerator in a quantitative business cycle framework. In J. B. Taylor & M. Woodford (Eds.), *Handbook of Macroeconomics*, volume 1 of *Handbook of Macroeconomics* chapter 21, (pp. 1341–1393). Elsevier.

Bernanke, B. S. & Mihov, I. (1998). Measuring Monetary Policy. *The Quarterly Journal of Economics*, 113(3), 869–902.

Bhattarai, S. & Schoenle, R. (2014). Multiproduct firms and price-setting: Theory and evidence from U.S. producer prices. *Journal of Monetary Economics*, 66, 178 – 192.

Bils, M. & Klenow, P. (2004). Some Evidence on the Importance of Sticky Prices. *Journal of Political Economy*, 112(5), 947–985.

Bonadio, B., Fischer, A. M., & Sauré, P. (2019). The Speed of Exchange Rate Pass-Through. *Journal of the European Economic Association*, 18(1), 506–538.

Bonomo, M., Carvalho, C., & Garcia, R. (2013). *Time- and State-Dependent Pricing: A Unified Framework*. 2013 Meeting Papers 759, Society for Economic Dynamics. 2013 Meeting Papers 759.

Caggiano, G., Castelnuovo, E., Colombo, V., & Nodari, G. (2015). Estimating fiscal multipliers: News from a non-linear world. *The Economic Journal*, 125(584), 746–776.

Caggiano, G., Castelnuovo, E., & Figueres, J. M. (2017). Economic policy uncertainty and unemployment in the United States: A nonlinear approach. *Economics Letters*, 151(C), 31–34.

Caggiano, G., Castelnuovo, E., & Groshenny, N. (2014). Uncertainty shocks and unemployment dynamics in U.S. recessions. *Journal of Monetary Economics*, 67, 78 – 92.

Calvo, G. A. (1983). Staggered prices in a utility-maximizing framework. *Journal of Monetary Economics*, 12(3), 383–398.

Caplin, A. S. & Spulber, D. (1987). Menu Costs and the Neutrality of Money. *The Quarterly Journal of Economics*, 102(4), 703–725.

References

Christiano, L. J., Eichenbaum, M., & Evans, C. L. (1999). Monetary policy shocks: What have we learned and to what end? In J. B. Taylor & M. Woodford (Eds.), *Handbook of Macroeconomics*, volume 1 chapter 2, (pp. 65–148). Elsevier.

Coibion, O. (2012). Are the Effects of Monetary Policy Shocks Big or Small? *American Economic Journal: Macroeconomics*, 4(2), 1–32.

Costain, J. & Nakov, A. (2011). Price Adjustments in a General Model of State-Dependent Pricing. *Journal of Money, Credit and Banking*, 43(2-3), 385–406.

Costain, J. & Nakov, A. (2019). Logit price dynamics. *Journal of Money, Credit and Banking*, 51(1), 43–78.

Costain, J., Nakov, A., & Petit, B. (2019). *Monetary policy implications of state-dependent prices and wages*. Working papers, Banco de España. Working Paper.

Cover, J. P. (1992). Asymmetric Effects of Positive and Negative Money-Supply Shocks. *The Quarterly Journal of Economics*, 107(4), 1261–1282.

Dotsey, M., King, R. G., & Wolman, A. (1999). State-Dependent Pricing and the General Equilibrium Dynamics of Money and Output. *The Quarterly Journal of Economics*, 114(2), 655–690.

Eichenbaum, M., Jaimovich, N., & Rebelo, S. (2011). Reference Prices, Costs, and Nominal Rigidities. *American Economic Review*, 101(1), 234–62.

Fischer, S. (1977). Long-term Contracts, Rational Expectations, and the Optimal Money Supply Rule. *Journal of Political Economy*, 85(1), 191–205.

Fisher, J. D. & Peters, R. (2010). Using Stock Returns to Identify Government Spending Shocks. *The Economic Journal*, 120(544), 414–436.

Furceri, D., Loungani, P., & Zdzienicka, A. (2016). *The Effects of Monetary Policy Shocks on Inequality*. IMF Working Papers 16/245, International Monetary Fund. Working Paper 16/245.

Gagnon, E., Lòpez-Salido, D., & Vincent, N. (2012). Individual Price Adjustment along the Extensive Margin. In D. Acemoglu, J. Parker, & M. Woodford (Eds.), *NBER Macroeconomics Annual 2012, Volume 27* (pp. 235–281). University of Chicago Press.

Gertler, M. & Karadi, P. (2015). Monetary Policy Surprises, Credit Costs, and Economic Activity. *American Economic Journal: Macroeconomics*, 7(1), 44–76.

Gilchrist, S. & Zakrajsek, E. (2012). Credit Spreads and Business Cycle Fluctuations. *American Economic Review*, 102(4), 1692–1720.

Golosov, M. & Lucas, R. J. (2007). Menu Costs and Phillips Curves. *Journal of Political*

Economy, 115, 171–199.

Gorodnichenko, Y. (2008). *Endogenous information, menu costs and inflation persistence*. Working Paper 14184, National Bureau of Economic Research. Working Paper 14184.

Granger, C. & Teräsvirta, T. (1993). *Modelling Non-Linear Economic Relationships*. Oxford University Press.

Gürkaynak, R. S., Sack, B., & Swanson, E. (2005). Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements. *International Journal of Central Banking*, 1(1), 55–93.

Hansen, B. E. (1992). Testing for parameter instability in linear models. *Journal of Policy Modeling*, 14(4), 517–533.

Hobijn, B., Ravenna, F., & Tambalotti, A. (2006). Menu Costs at Work: Restaurant Prices and the Introduction of the Euro. *The Quarterly Journal of Economics*, 121(3), 1103–1131.

Ireland, P. N. (2007). Changes in the Federal Reserve's Inflation Target: Causes and Consequences. *Journal of Money, Credit and Banking*, 39(8), 1851–1882.

Jordà, O. (2005). Estimation and Inference of Impulse Responses by Local Projections. *American Economic Review*, 95(1), 161–182.

Karadi, P. & Reiff, A. (2019). Menu Costs, Aggregate Fluctuations, and Large Shocks. *American Economic Journal: Macroeconomics*, 11(3), 111–146.

Kehoe, P. & Midrigan, V. (2015). Prices are sticky after all. *Journal of Monetary Economics*, 75, 35–53.

Kim, S. & Roubini, N. (2000). Exchange rate anomalies in the industrial countries: A solution with a structural VAR approach. *Journal of Monetary Economics*, 45(3), 561–586.

Kiyotaki, N. & Moore, J. (1997). Credit Cycles. *Journal of Political Economy*, 105(2), 211–248.

Klenow, P. J. & Kryvtsov, O. (2008). State-Dependent or Time-Dependent Pricing: Does it Matter for Recent U.S. Inflation? *The Quarterly Journal of Economics*, 123(3), 863–904.

Kuttner, K. N. (2001). Monetary policy surprises and interest rates: Evidence from the Fed funds futures market. *Journal of Monetary Economics*, 47(3), 523–544.

Midrigan, V. (2011). Menu Costs, Multiproduct Firms, and Aggregate Fluctuations. *Econometrica*, 79(4), 1139–1180.

Nakamura, E. (2008). Pass-Through in Retail and Wholesale. *American Economic Review*, 98(2), 430–37.

Nakamura, E. & Steinsson, J. (2008). Five Facts about Prices: A Reevaluation of Menu Cost

Models. *The Quarterly Journal of Economics*, 123(4), 1415–1464.

Nakamura, E. & Steinsson, J. (2010). Monetary Non-neutrality in a Multisector Menu Cost Model. *The Quarterly Journal of Economics*, 125(3), 961–1013.

Nakamura, E. & Zerom, D. (2010). Accounting for Incomplete Pass-Through. *The Review of Economic Studies*, 77(3), 1192–1230.

Newey, W. K. & West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3), 703–708.

Racine, J. (1997). Feasible Cross-Validatory Model Selection for General Stationary Processes. *Journal of Applied Econometrics*, 12(2), 169–179.

Ramey, V. (2016). Macroeconomic Shocks and Their Propagation. In J. B. Taylor & H. Uhlig (Eds.), *Handbook of Macroeconomics*, volume 2 chapter 2, (pp. 71 – 162). Elsevier.

Ramey, V. A. & Zubairy, S. (2018). Government Spending Multipliers in Good Times and in Bad: Evidence from US Historical Data. *Journal of Political Economy*, 126(2), 850–901.

Romer, C. D. & Romer, D. H. (2004). A New Measure of Monetary Shocks: Derivation and Implications. *American Economic Review*, 94(4), 1055–1084.

Romer, C. D. & Romer, D. H. (2010). The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks. *American Economic Review*, 100(3), 763–801.

Rotemberg, J. J. (1982). Monopolistic price adjustment and aggregate output. *Review of Economic Studies*, 49, 517–531.

Sheshinski, E. & Weiss, Y. (1977). Inflation and Costs of Price Adjustment. *Review of Economic Studies*, 44(2), 287–303.

Stock, J. H. & Watson, M. W. (2016). Core Inflation and Trend Inflation. *The Review of Economics and Statistics*, 98(4), 770–784.

Taylor, J. B. (1979). Staggered Wage Setting in a Macro Model. *American Economic Review*, 69(2), 108–113.

Taylor, J. B. (1980). Aggregate Dynamics and Staggered Contracts. *Journal of Political Economy*, 88(1), 1–23.

Tenreyro, S. & Thwaites, G. (2016). Pushing on a String: US Monetary Policy Is Less Powerful in Recessions. *American Economic Journal: Macroeconomics*, 8(4), 43–74.

Uhlig, H. (2005). What are the effects of monetary policy on output? Results from an agnostic identification procedure. *Journal of Monetary Economics*, 52(2), 381–419.

Veirman, E. D. (2009). What Makes the Output–Inflation Trade-Off Change? The Absence of Accelerating Deflation in Japan. *Journal of Money, Credit and Banking*, 41(6), 1117–

1140.

Chapter 2

Financial factors, firm size and firm potential

Abstract

Using a unique dataset covering the universe of Portuguese firms and their credit situation, we show that financially constrained firms (1) are found across the entire firm size distribution, (2) account for a sizable asset share, and (3) exhibit a higher sensitivity to shocks, conditional on size. As these findings are counterfactual to the canonical heterogeneous firms model with financial frictions, we identify a factor in the data that reconciles theory and data: permanent heterogeneity among firms. Incorporating ex-ante dispersion of firm potential into the standard model allows us to match the distribution of constrained firms conditional on size. In contrast to previous models, the granularity of large constrained firms, together with the fact that constrained firms present a higher capital elasticity, can generate sizable recessions even for relatively minor shocks to the financial sector.

Keywords: Firm size, business cycle, financial accelerator
JEL Codes: E62, E22, E23

2.1 Introduction

A substantial amount of research in macroeconomics focuses on the propagation of aggregate shocks via financial factors and their relation to individual firm characteristics. In a seminal work on this topic, Gertler & Gilchrist (1994) propose firm size as an effective proxy for financial constraints. Smaller firms are arguably more risky, less liquid and face an elevated external finance premium. Accordingly, small firms are more sensitive to aggregate shocks, as they tend to be in a weaker financial position. Standard heterogeneous firms models reflect this idea, generating a strong correlation between firm size and their financial situation.

This paper provides new empirical evidence that casts doubt on a strong association between size and financial constraints. Using the Bank of Portugal's confidential credit registry database, matched with bank and firm balance sheet data between 2006 and 2017, we construct detailed, firm-specific, and credit-based measures of financial constraints. The credit registry database contains monthly information on actual, potential, short-term and long-term credit above 50 Euros extended to individuals and non-financial corporations by all financial institutions in Portugal. Using this substantial granularity of the data we provide three empirical facts that are counterfactual to the standard heterogeneous firms model with financial frictions.

First, financially constrained firms are found across the entire size distribution. Across all our measures of constraints, there is a non-zero fraction of constrained firms in every size percentile. In fact, going from the bottom 5% of the size distribution to the top 5% only reduces the probability of being constrained by approximately 12% for our preferred measure. Second, we find that the share of assets held by these firms is considerable, and can reach up to 12% of total assets. This is a consequence of the fact that there are large constrained firms that naturally account for a larger asset share. Third, financially constrained firms exhibit a higher elasticity of turnover and employment to both 1) the business cycle, 2) idiosyncratic Total Factor Productivity (TFP) shocks and 3) idiosyncratic financial shocks, conditional on size. Hence, financial constraints explain in part the heterogeneous elasticities across firms in support of the financial accelerator mechanism. Moreover, this channel seems to be independent of potential size channels, such as the one identified by Crouzet & Mehrotra (2020).

In summary, our first contribution is to show that constrained firms (1) exist across the entire size distribution, (2) have a substantial asset share and (3) exhibit higher elasticity to various shocks, conditional on their size. A standard heterogeneous firms model with

a transitory productivity process is unable to match these facts. As the productivity is mean reverting there is a relatively homogeneous optimal size for all firms. Hence, the model produces only small constrained and large unconstrained firms, contrary to our empirical finding (1). It also implies an underestimation of the share of assets held by constrained firms, and size as a sufficient proxy for differential sensitivity to the exogenous innovations, in contrast to empirical findings (2) and (3).

In light of this we explore whether a richer structure of heterogeneity exists in the data, possibly reconciling the model with the empirical results above. In particular, our second contribution is to demonstrate that a considerable degree of ex-ante heterogeneity exists across Portuguese firms. This type of heterogeneity can be thought of as a firm's potential, persists over the firm's life cycle, and affects constrained and unconstrained firms differently. We establish this using a variety of approaches. Firstly, the standard deviation of employment across firms is high and increasing with age, implying large size differences early in the life cycle and a wide range of optimal firm sizes. Secondly, the autocorrelation of employment remains high throughout a firm's life cycle. These two results point towards the importance of permanent firm differences. Thirdly, we also show that these statistics are diminished for constrained firms, in line with our theoretical predictions. Finally, we confirm the importance and differential incidence of ex-ante heterogeneity using the flexible statistical model developed by Pugsley et al. (2021).

The empirical evidence of ex-ante heterogeneity then serves as motivation to refine the canonical theoretical model by including a permanent component of the firms' productivity process. This rather simple addition to the model enables it to match the stylised facts, as it introduces a large and persistent heterogeneity in optimal firm sizes and spells of financial constraints. We then use the augmented model to analyse how the existence of ex-ante heterogeneity shapes aggregate outcomes, our third contribution. Whilst the aggregate responses are relatively similar following a productivity shock, financial shocks are greatly amplified relative to the standard model with only a transitory productivity component. This is due to the granularity of large constrained firms in the top percentiles of the size distribution that need to cut investment when the borrowing constraint is tightened. This mechanism amplifies the output drop in response to a financial shock up to four times and consequently implies that smaller financial shocks are required to elicit sizable recessions. Further, we demonstrate that matching the joint distribution of size and constrained firms gives rise to almost three times stronger output losses due to capital misallocation.

Overall, this paper emphasises the importance of targeting the joint distribution of size

and financial constraints in order to correctly quantify the propagation and amplification of aggregate shocks in existing financial friction models. In other words, models that ignore the existence of large constrained firms may significantly underestimate the pass-through of a tightening of financial conditions on output.

Literature. Our work follows a large literature in macroeconomics that has analysed heterogeneous firms and financial frictions both theoretically and empirically.

Firstly, we relate to the empirical literature that assesses the differences in the cyclicality of constrained firms and the debate on how to identify these firms in the data. Gertler & Gilchrist (1994) find empirical evidence for the financial accelerator mechanism. They analyse the differential cyclical behavior of small and large manufacturing firms and interpret this as evidence for the financial accelerator. Their main assumption is that size is a good proxy for financial constraints. Sharpe (1994) detects a statistically significant relationship between a firm's leverage ratio and the cyclicality of its labor force. Employment growth at highly leveraged firms is more sensitive as they are less likely to hoard labor. This cyclicality also holds for the size dimension, implicitly confirming Gertler & Gilchrist (1994)'s evidence. Related, Gilchrist & Himmelberg (1995) find that investment still responds to cash flow even after controlling for its role for forecasting future investment opportunities, with the effect being stronger for firms without full access to the capital market.

More recently, Crouzet & Mehrotra (2020), using firm level data underlying the Quarterly Financial Reports (QFR) provided by the US Census Bureau, document that differences in size-related cyclicality only arise at the very top of the distribution, with the bottom 99.5% of firms having non-significant differences in cyclicality. Arguably, this evidence, together with the insignificance of standard financial proxies for financial constraints speaks against financial factors driving cyclicality differences.

These results are also related to Farre-Mensa & Ljungqvist (2016) findings, who suggest that typical measures of financial constraints are not associated with firms that behave as if they were constrained. Even indices that combine different firm characteristics such as the ones proposed by Kaplan & Zingales (1997), Whited & Wu (2006) and Hadlock & Pierce (2010) do not correlate well with firms that behave as financially constrained. These findings are also supported by Bodnaruk et al. (2015), who use text analysis of the 10-k financial reports to gauge if firms are constrained or not, and find a weak correlation with common constraint measures. Buehlmaier & Whited (2018) equally contribute to this literature by developing a new financial constraint measure based on text analysis. Finally,

focusing on sensitivity of monetary policy Cloyne et al. (2018) find that age and dividend payments are an empirically relevant proxy for increased sensitivity to the funds rate.

Our paper, by making use of detailed firm level credit data, contributes to this literature by reiterating that size is indeed an insufficient proxy for financial constraints. Moreover, with information on credit lines available to the firm and overdue credit, we also provide evidence that supports a broader financial accelerator mechanism that is only weakly size dependent. Our measures of constraints statistically significantly increase cyclicality even when controlling for size groups.

Secondly, we contribute to the research on heterogeneous firm financial frictions models. One of the early contributions in this literature by Cooley & Quadrini (2001) shares many features with our current model. They augment an otherwise standard Hopenhayn (1992) model of heterogeneous firms with financial frictions and persistent shocks. In doing so, they are able to match the empirical facts that both smaller firms, conditional on age, and younger firms, conditional on size, are more dynamic (i.e. job creation and destruction, growth, volatility of growth and exit are all higher). In similar fashion, Pugsley et al. (2021) highlight the importance of ex-ante heterogeneity in explaining the firm size distribution and the recent decline in firm dynamism.

Another recent instance where permanent productivity differences are important to explain the evidence is Mehrotra & Sergeyev (2020). They argue that financial frictions played a relatively minor role in unemployment increases associated with the Great Recession and that employment was reduced due to shocks that affected unconstrained and constrained firms alike. Conversely, Khan & Thomas (2013) and Ottonello & Winberry (2018) argue for the importance of financial frictions in the propagation of financial and monetary policy shocks, respectively. Our theoretical contribution emphasises the importance of permanent productivity differences for matching the observed distribution of constrained firms, conditional on size. We also highlight the importance of matching this distribution in amplifying both productivity and financial shocks, based on a model very similar to the literature above.

Outlook. The paper is structured as follows; Section 2.2 presents the data we use for the empirical analysis as well as to discipline our theoretical model. We proceed to present the three empirical findings outlined above in section 2.3. Section 2.4 illustrates the importance of ex-ante heterogeneity in our data. In section 2.5 we set out the model to incorporate and account for these facts and in section 2.6 we discuss model predictions of aggregate effects. Finally, section 2.7 concludes.

2.2 Data

We draw on a unique combination of datasets that cover the Portuguese economy between 2006 and 2017, all managed by the Bank of Portugal Microdata Research Laboratory.

The *Informação Empresarial Simplificada* (IES) Central Balance Sheet Database is based on annual accounting data of individual firms. Portuguese firms have to fill out mandatory financial statements in order to comply with their statutory obligation. Consequently, this dataset covers the population of virtually all non-financial corporations in Portugal from 2006 onwards. We combine this dataset with the Central Credit Register (CCR) which contains monthly information on the actual and potential credit above 50 Euros extended to individuals and non-financial corporations, reported by all financial institutions in Portugal.[1] Actual credit includes loans that are truly taken up, such as mortgages, consumer loans, overdrafts and others. Potential credit encompasses all irrevocable commitments to the subject that have not materialized into actual credit, such as available credit on credit cards, credit lines, pledges granted by participants and other credit facilities.[2] We then merge these two databases on the firm level. Moreover, we also add the Monetary Financial Institutions Balance Sheet Database in order to gain information on the balance sheets of banks that extend credit to non-financial institutions. We merge this dataset on a firm level using the bank identifier and the share of loans extended by one firm to arrive at our final dataset.

Similar to Buera & Karmakar (2019), who use the same dataset, we restrict the set of firms in this panel dataset to those with at least five consecutive observations and to firms which are in business at the time of reporting. Furthermore, we only consider privately or publicly held firms and drop micro firms, i.e. those with overall credit amounts of less than 10,000€. Descriptive statistics for the relevant variables can be found in Table B.1 in Appendix B.2.

2.2.1 Measures of financial constraints

Based on the credit information in the data we construct several binary measures indicating whether a firm is financially constrained. Financial constraints are most commonly conceived as a supply side phenomenon. Firms that could potentially obtain credit in perfect credit markets are unable to do so due to asymmetric information considerations on

[1] Given that the firm balance sheet data is of yearly frequency, we consider the month in which the balance sheet data was reported. Results were robust to shifting and averaging the monthly credit data.

[2] Further details on the credit information used are documented in Appendix B.1.

the supply side. For example, a firm that has a profitable investment project that requires external financing cannot realise it as the bank is not satisfied with the creditworthiness of that firm. This may happen either via the price dimension, i.e. a risk premium on the interest rate, or on the quantity dimension i.e. the credit is denied altogether. In this paper, we classify constrained firms along the quantity dimension, using the credit information for each firm.[3] Given that credit allowances are changing over time, this provides us with a time-varying and firm-specific measure for being financially constrained. At the same time, while credit information offers a far more detailed notion of a firm being constrained compared to standard financial ratios such as leverage or liquidity, it is still a proxy. In order to take this into account, we consider a wide range of binary measures to identify whether a firm is constrained.

Measures. Many existing models classify firms as constrained if they have exhausted their maximum borrowing capacity. In our data, the closest counterpart to this metric is potential credit, summarising irrevocable commitments by credit institutions. However, even though this measure enables an understanding of whether firms have drawn down their credit lines and are potentially constrained, it also encompasses a lot of noise. One problem might be that although firms have exhausted their committed credit lines, they could still successfully apply for a short- or long-term loan. To account for this, in our baseline definition, we consider a firm to be credit constrained at time t, if it has no potential credit available at time t and neither its short- nor long-term credit (i.e. effective credit) is growing:

$$\text{Constrained I} := \mathbf{1}_{\text{Potential credit}_t=0 \text{ \& } \Delta\text{Effective credit}_t \leq 0}.$$

The second measure considers firms as constrained if they don't have any potential credit available and overdue credit is positive:

$$\text{Constrained II} := \mathbf{1}_{\text{Potential credit}_t=0 \text{ \& Overdue credit}_t > 0}.$$

The rationale behind this definition is that having overdue credit is likely a signal for a firm in poor financial shape. The third measure is even stricter and considers firms as

[3] See for example Custodio et al. (2021), who use the same dataset to analyse how the price dimension affects firms' investment and employment.

constrained only if overdue credit is increasing:

$$\text{Constrained III} := \mathbf{1}_{\text{Potential credit}_t=0\ \&\ \Delta\text{Overdue credit}_t>0}.$$

While the measures presented so far are conceptually in the spirit of a firm hitting the credit constraint and thus being strictly constrained, it might also be that a firm is in a delicate financial position if it has a large share of their credit to repay within a short period of time. The fourth measure considers this possibility by classifying a firm as constrained if the share of credit to assets that is due within the next year is in the top 10 percent of the distribution:

$$\text{Constrained IV} := \mathbf{1}_{\frac{\text{Credit < 1 Year Maturity}_t}{\text{Total Assets}_t} > \mathbf{P}_{90}}.$$

Our final measure follows the evidence presented by Rampini & Viswanathan (2020) that financially constrained firms use more secured debt, and considers a firm to be constrained if the share of secured debt over total assets is in the top 10 percent of the distribution:

$$\text{Constrained V} := \mathbf{1}_{\frac{\text{Secured Debt}_t}{\text{Total Assets}_t} > \mathbf{P}_{90}}.$$

Appendix B.1 provides a more detailed description of the dataset and the underlying variables for the constrained measures. Table B.3 reports the correlation matrix between the different measures. Finally, Figures B.1 and B.2 report the evolution of the share of constrained firms and credit over time.

2.3 Empirical analysis

Utilizing our measures of financial constraints, we present three stylised empirical facts that are counterfactual to the standard financial frictions model. First, we illustrate that size is only weakly correlated to the firm's financial health. In fact, constrained firms can be found over the entire firm size distribution. Second, we show that across all of our constrained measures, these firms have a higher asset share compared to what a standard calibrated model would predict. Third, we revisit the relationship between firm characteristics and the elasticity of firm outcomes. We find turnover growth of constrained firms to be more cyclical and also more responsive to TFP and financial shocks, and this to remain significant even when controlling for firm size groups. Finally, we take the fact that constrained firms react more to financial shocks also as evidence of the validity of our constrained measures.

2.3.1 Constrained firms are found across the firm distribution

Our first stylised fact states that financially constrained firms populate the entire firm distribution along a number of common proxies for financial constraints. Figure 2.1 plots the share of firms that have zero potential credit and no increase in effective credit (measure I) over percentiles of total assets, age, liquidity ratio and leverage. Evidently, constrained firms can be found in every bin of the firm distribution. In particular, there are constrained firms across the entire firm size distribution, as illustrated by the plot over percentiles of total assets. This finding is robust across all binary identifiers for being constrained, with only the overall fraction of constrained firms changing depending on the strictness of the specific measure, as documented in Figures B.3 - B.6 in Appendix B.3.1. While correlations are in line with the existing literature, they are not as strong as existing models would predict. In fact, when running a linear probability model, the probability of being constrained only reduces by about 12% for two standard deviation increase in total assets, which is equivalent to going from the bottom 5% to the top 5% of the size distribution.[4] Even after accounting for potential attenuation bias, the main conclusion stands: standard firm models typically produce exclusively small constrained firms and large unconstrained firms, yet our data does not support this strong dichotomy. Moreover, even when controlling for a battery of financial variables the explanatory power to predict whether a firm is constrained is relatively low compared to the firms' fixed effects. Hence, existing proxies of financial constraints may be unable to capture this unobserved heterogeneity, which seems to play a substantial role in credit decisions.

2.3.2 Constrained firms account for a larger asset share

The first stylised fact establishes that there are constrained firms across the entire size distribution. This finding contradicts the standard heterogeneous firms model which generates only large unconstrained and small constrained firms. As a consequence, these types of models severely underestimate the fraction of total assets belonging to constrained firms compared to the data, which constitutes our second empirical fact.

This is illustrated in Table 2.1 which compares the empirical values for the fraction of constrained firms and their implied asset share with the values implied by a standard Khan & Thomas (2013) model, calibrated to match 23% of constrained firms, in line with the constrained measure I.[5] Depending on how strict the constrained measure is,

[4] See Table B.2 in Appendix B.2 for the results of the linear probability model.
[5] For more details on the model and calibration see Section 2.5.

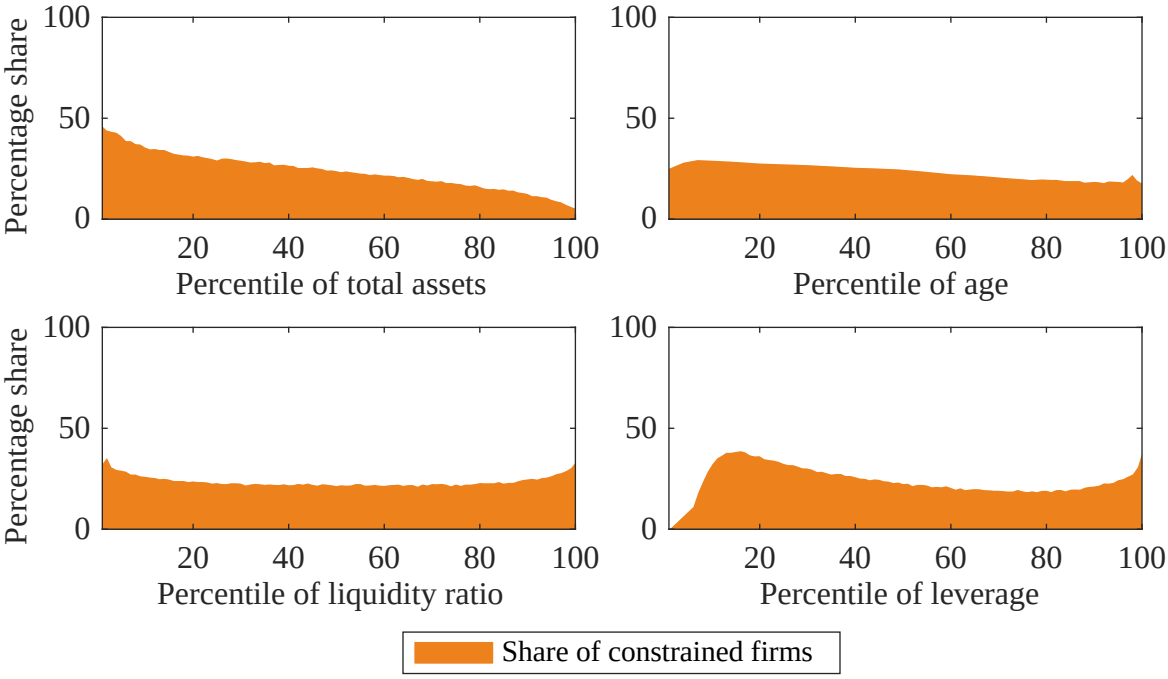


Fig. 2.1: Decomposition of constrained and unconstrained firms across percentiles of firm variables. Constrained firms are identified using measure I which classifies firms as constrained if they have exhausted their potential credit and were not obtaining additional credit in that period.

the percentage of constrained firms in the data varies substantially from 3% to 23% of total firms. Nonetheless, even for the strictest measure that only labels 2% of firms as constrained, the share of total assets in this group of firms is 1%. This value is equal to what is implied by the calibrated model, which is only 1%, despite the model being calibrated to generate 23% of constrained firms. This low share of assets in constrained firms is a consequence of the implied distribution of firms. Whereas in the data, even for the stricter measures there are still constrained firms at the top of the size distribution, in the model the constrained firms are all concentrated at the bottom of the distribution, causing an underestimation of the share of total assets in constrained firms.

2.3.3 Size and financial factors matter for elasticity

A higher asset share held by constrained firms and a presence of constrained firms across the distribution do not necessarily warrant a reassessment of the cyclical properties of financial frictions models. These measures only imply relevance for aggregate cyclicality if these firms also exhibit a differential elasticity to shocks. This section aims to demonstrate that this is the case. In particular, we show that constrained firms are more cyclical than

Table 2.1: Percentage of constrained firms and share of total assets in constrained firms

	Standard Model	Constrained measure				
		I	II	III	IV	V
% of constrained firms	24	23	4	2	10	10
% of assets in constrained firms	1	12	2	1	7	8

their unconstrained counterparts, and are more responsive to both TFP and financial shocks, conditional on size. In order to illustrate this point we use a specification similar to Crouzet & Mehrotra (2020), augmented with the set of firm-specific and time-varying measures of financial factors:

$$g_{i,t} = \kappa u_{i,t} + \sum_{j \in \mathscr{J}} (\alpha_j + \beta_j u_{i,t}) \mathbf{1}_{i \in \mathscr{S}_t^{(j)}} + \sum_{l \in \mathscr{L}} (\gamma_l + \delta_l u_{i,t}) \mathbf{1}_{i \in \mathscr{L}} \tag{2.1}$$
$$+ (\zeta + \eta u_{i,t}) \text{Const.} n_{i,t} + \epsilon_{i,t},$$

where i identifies a firm and t identifies a year. The dependent variable $g_{i,t}$ is the year-on-year log change in turnover. The set $\mathscr{S}_t^{(j)}$ is a jth size group, e.g. all firms above the 90th but below the 99th percentile. We include three size groups, $j \in \{[90,99], [99,99.5], [99.5,100]\}$. Furthermore, $u_{i,t}$ takes the form of three different variables: 1) the year-on-year growth rate of GDP; 2) TFP shocks at the firm level, using the method proposed by Ackerberg et al. (2015); and 3) bank level shocks aggregated at the firm level, following Amiti & Weinstein (2018). $\mathscr{L}$ is a set of industry dummies. Const.$n_{i,t}$ refers to the firm-specific variable measuring financial constraints, indexed by n. Finally, for both TFP shocks and bank level shocks we also include firm fixed effects.

Table 2.2 reports estimates of the coefficient of interest η, the semi-elasticity of firm-level growth in turnover to the different shocks relative to the control group of financially unconstrained firms. In the first line we have the semi-elasticity to the economic cycle, captured by GDP growth. In the second line we report the semi-elasticity relative to firm-level TFP shocks estimated as in Ackerberg et al. (2015). The third line presents the semi-elasticity to a financial shock, identified using the methodology proposed by Amiti & Weinstein (2018). The columns present the elasticity for the different constrained measures.

The first column reports the semi-elasticity of the control group, unconstrained firms, to the different shocks. The results can be interpreted as follows: for a 1% increase in GDP growth, firm-level TFP or credit supply, turnover of unconstrained firms increases by 2.5%,

Table 2.2: Semi-elasticity of turnover conditional on size and measures of financial constraints

	Un-constrained	Constrained measure				
		I	II	III	IV	V
% Δ GDP	2.512***	0.150***	1.975***	1.762***	0.988***	0.394
	(0.023)	(0.054)	(0.427)	(0.529)	(0.272)	(0.268)
TFP shock	0.522***	0.019***	0.096***	0.100***	0.097***	0.072***
	(0.004)	(0.004)	(0.007)	(0.007)	(0.005)	(0.006)
Fin. shock	0.005	0.106***	0.146	0.078	0.340*	0.293*
	(0.019)	(0.039)	(0.146)	(0.172)	(0.148)	(0.124)

Notes. Estimates report the constrained firms semi-elasticity of turnover relative to the control group of unconstrained firms, with respect to GDP, TFP and Financial shocks. Constrained measures are constructed as outlined in Section 2.2.1. All specifications contain a constant term and non-interacted indicators. Standard errors are reported in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

0.5% and 0% respectively. While unconstrained firms turnover responds to both changes in GDP and TFP, it presents a semi-elasticity of 0 to credit supply shocks, which translates into financial shocks not affecting, on average, financially unconstrained firms.

The remaining columns report the semi-elasticity of constrained firms, for the different constrained measures outlined in Section 2.2.1, relative to the unconstrained firms. The results for our baseline measure indicate that constrained firms have, on average, a semi-elasticity to GDP growth, TFP and financial shocks, that is 0.15, 0.02 and 0.1 percentage points higher than unconstrained firms, offering support for the financial accelerator mechanism.

However, as already pointed out when introducing the different measures for being constrained, the baseline measure might capture firms for which potential credit is zero, but are in fact unconstrained. Hence, the baseline measure offers a lower bound of the increased sensitivity of constrained firms. We therefore consider other binary measures trying to overcome these drawbacks, reported in columns II to V. These estimation results are supportive of the notion that the baseline measure acts as a lower bound and that the sensitivity might be up to one order of magnitude higher for constrained firms, as measured by measure II in the GDP regression.

Additionally, we take the evidence that constrained firms are the only ones reacting to credit supply shocks as validation of our measures. This is in line with growing evidence in the literature on the causal effect of financing constraints on firm level outcomes.

Financially constrained firms are found to have a higher elasticity of investment and employment with respect to shocks to the collateral value (Gan (2007) and Chaney et al. (2012)), to financial shocks (Chodorow-Reich (2014)) and to monetary policy shocks (Greenwald et al. (2019) and Ottonello & Winberry (2018)). The results presented in this section are in line with the literature results and suggest that our financial constraint measures are indeed capturing the constrained firms.

Results for the remaining regression coefficients are presented in Tables B.4, B.5 and B.6 in Appendix B.2. It is worth noting that the estimation coefficients with respect to size groups hardly change when including the different constraint measures. This is indicative of the fact that the mechanism going through size is somewhat independent to any financial accelerator mechanism and that size might not be a good proxy for the latter, as already pointed out by Crouzet & Mehrotra (2020).

In Table B.7 in Appendix B.2 we present the results when considering growth in employees instead of turnover. Besides using different measures, we also consider a battery of robustness checks for our GDP regression in particular. First, we include firm fixed effects. Second, we include time fixed effects to account for broader macroeconomic circumstances. Third, we estimate the model excluding those firms for which potential credit is zero throughout. Fourth, we control for supply effects using aggregated bank data. Estimates are robust across all specifications and the results can be found in Tables B.8-B.11 in Appendix B.2.

2.4 Firm potential

In contrast to our stylised facts, the canonical firm financial frictions model à la Khan & Thomas (2013) predicts a very strong correlation between firm size and financial constraints, as firms require a relatively uniform minimum size to become unconstrained. One factor that could potentially break this strong correlation is heterogeneous ex-ante conditions for firms, such as firm potential. Small firms may be unconstrained as they already have reached their potential - i.e. optimal size - while large firms may still be growing and are still constrained. Equally, heterogeneous potential creates a dispersion of unconstrained firms across the entire firm size distribution, similar to our first stylised fact. Further, larger constrained firms may elevate the fraction of assets held by constrained firms closer to what we observe in the data. Finally, this heterogeneity may also explain why financial factors matter for firm cyclicality even when controlling for firm size, as demonstrated in our third stylised fact. Accordingly, this section investigates whether

such ex-ante heterogeneity exists in our dataset.

Looking at the standard deviations of log employment by age a and autocorrelation structure of log employment between age a and h, we find evidence that there is ex-ante heterogeneity, as firms at birth are not all equal, suggesting that ex-ante conditions are persistent and affect firms even in the long run, in line with evidence presented by Pugsley et al. (2021).[6]

Additionally, we find evidence that the ex-ante heterogeneity affects constrained and unconstrained firms differently.[7] The standard deviation is lower throughout the life-cycle and the autocorrelation structure converges to a lower level for constrained firms compared to unconstrained ones. One may have expected the opposite to be true, as constrained firms potentially have less resources to grow and so their employment tomorrow could have a stronger correlation with employment today. Yet, the fact that the autocorrelation tends to be higher across the life-cycle for unconstrained firms may be indicative that they are born closer to their optimal size, when compared to constrained firms. This may then explain why some young firms are constrained and others are not: the ones born closer to their optimal size have lower investments and do not become constrained, while firms that need to grow to reach the optimal size exhaust their credit lines. The results are depicted in Figures B.7 and B.8 in Appendix B.3.

Statistical model. To gain understanding beyond descriptive statistics of the importance of ex-ante and ex-post heterogeneity for the life-cycle of firms, we again follow Pugsley et al. (2021) and adopt their statistical model. This model uses the information provided by the autocovariance structure of log employment to capture the importance of both types of heterogeneity.

Consider the following decomposition for employment n by firm i at age a:

$$\underbrace{\ln n_{i,a}}_{\text{log employment}} = \underbrace{u_{i,a} + v_{i,a}}_{\text{Ex-ante component}} + \underbrace{w_{i,a} + z_{i,a}}_{\text{Ex-post component}} , \tag{2.2}$$

[6]To prevent differences across sectors and business cycle conditions from explaining the majority of the standard deviation and autocorrelation, we first control for sector and year fixed effects and then use the residuals of log employment.

[7]Here we are using the baseline measure Const I, taking into account both potential credit and growth of effective credit. A firm is considered constrained if at age $a - h$ it has potential credit equal to zero and if the effective credit is not growing.

Table 2.3: Calibrated model parameters for the unbalanced panel, including all, constrained and unconstrained firms according to measure I

	ρ_u	ρ_v	ρ_w	σ_θ	σ_u	σ_v	σ_ϵ	σ_z
Total	0.425	0.799	0.904	0.369	0.748	0.708	0.305	0.185
Unconstrained	0.431	0.770	0.884	0.399	0.769	0.744	0.311	0.158
Constrained	0.493	0.874	0.911	0.255	0.655	0.641	0.265	0.176

where

$$
\begin{aligned}
u_{i,a} = \quad & \rho_u u_{i,a-1} + \theta_i, && u_{i,-1} \sim \text{iid}\left(\mu_{\tilde{u}}, \sigma^2_{\tilde{u}}\right), && \theta_i \sim \text{iid}\left(\mu_\theta, \sigma^2_\theta\right), && \left|\rho_u\right| \leq 1 \\
v_{i,a} = \quad & \rho_v v_{i,a-1}, && v_{i,-1} \sim \text{iid}\left(\mu_{\tilde{v}}, \sigma^2_{\tilde{v}}\right), && && \left|\rho_v\right| \leq 1 \\
w_{i,a} = \quad & \rho_w w_{i,a-1} + \varepsilon_{i,a}, && w_{i,-1} = 0, && \varepsilon_{i,a} \sim \text{iid}\left(0, \sigma^2_\varepsilon\right), && \left|\rho_w\right| \leq 1 \\
z_{i,a} \sim \quad & \text{iid}\left(0, \sigma^2_z\right)
\end{aligned}
$$

In this employment process, the terms $u_{i,a}$ and $v_{i,a}$ capture the ex-ante profile while $w_{i,a}$ and $z_{i,a}$ capture the ex-post one. The ex-ante component is determined by three shocks that are drawn just prior to the birth year, at $a = -1$. The shocks $v_{i,-1}$ and $u_{i,-1}$ represent the initial conditions of the firm, which allow for rich heterogeneity even at birth. θ_i is the permanent component, which will accumulate over the life-cycle at speed ρ_u. In particular, with $\rho_u < 1$, the long-run steady state level of employment will be given by $\frac{\theta_i}{1-\rho_u}$. Further, this specification allows for rich heterogeneity not only in terms of optimal size of the firms, depending on the distribution of θ_i, but also in terms of the speed at which firms reach the steady state. As firms start at different points depending on $u_{i,-1}$ and $v_{i,-1}$ and each shock has its own persistence parameter, the path from initial to steady state employment will highly differ across firms.

The ex-post component is formed of two different shocks, one i.i.d. shock with expected value of zero, and a persistent one that follows an AR(1) process with i.i.d. innovations $\epsilon_{i,a}$ and persistence ρ_w. To abstract the ex-post component from affecting the ex-ante one, we set the initial conditions of the persistent shock to $w_{i,-1} = 0$.

We calibrate the model for all, constrained and unconstrained firms separately by minimising the sum of squared differences between the model and empirical autocovariance. Firms are again split into constrained and unconstrained categories according to the measure Constrained I.

Table 2.3 presents the parameters resulting from the calibration strategy.[8] Two key

[8] Figure B.9 in Appendix B.3.2 plots the model fit to the data for both types of firms.

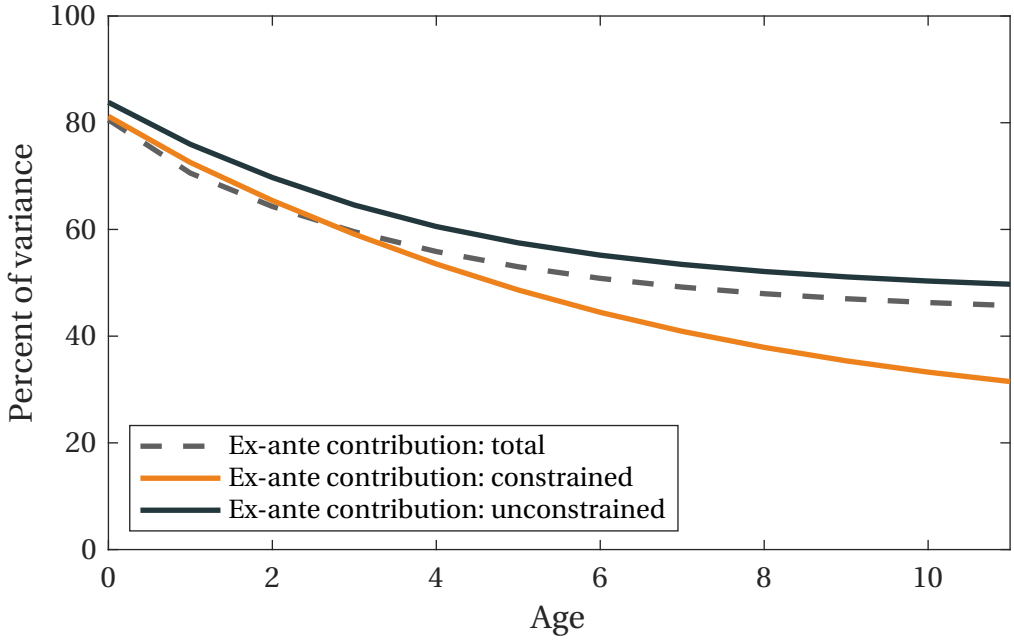


Fig. 2.2: Ex-ante variance contribution. Values for constrained firms presented in orange (light), while blue stands for the unconstrained firms (dark).

parameters of the model are ρ_u and σ_θ, as, together, they imply that permanent heterogeneity exists. First, using the total panel, the point estimates imply that ex-ante conditions matter, as both ρ_u and σ_θ are nonzero. Second, the point estimates imply a standard deviation of steady state employment, σ_θ for unconstrained firms of 0.399 and 0.255 for constrained ones. This again demonstrates that there seem to be differences between both types of firms that originate from ex-ante conditions.

Finally, to more clearly identify the ex-post and ex-ante contributions, one can also derive the formula for the model autocovariance, enabling a clear identification of the contribution of both components. The autocovariance formula is given by

$$\mathrm{Cov}\left[\ln n_{i,a}, \ln n_{i,a-j}\right] = \underbrace{\left(\sum_{k=0}^{a} \rho_u^k\right)\left(\sum_{k=0}^{a-j} \rho_u^k\right)\sigma_\theta^2 + \rho_u^{2(a+1)-j}\sigma_{\hat{u}}^2 + \rho_v^{2(a+1)-j}\sigma_{\hat{v}}^2}_{\text{Ex-ante component}}$$
$$\underbrace{+\sigma_\epsilon^2 \rho_w^j \sum_{k=0}^{a-j} \rho_w^{2k} + \sigma_z^2 \mathbf{1}_{j=0}}_{\text{Ex-post component}}$$

and its derivation can be found in Appendix B.4. The autocovariance is a function of variance and persistence parameters of both ex-ante and ex-post shocks, as described

above. Figure 2.2 illustrates the importance of the ex-ante component for the variance as a function of a firm's age. For all categories of firms, the ex-ante component contribution is above 80% at birth. Differences between the constrained and unconstrained firms start to arise after year 1, with the ex-ante component explaining more than 60% of the variance for unconstrained firms in the long run, while for constrained firms it is below 40%.

The fact that the ex-ante contribution is stronger for unconstrained firms is indicative that these firms are born closer to their optimal size. At the same time, constrained firms have not reached their optimal size yet, and so naturally less contribution to the employment dispersion originates from permanent conditions.

All the empirical evidence in this section suggests that ex-ante heterogeneity: 1) matters both in the short and in the long-run and 2) more strongly affects unconstrained than constrained firms. This may be indicative that unconstrained firms start closer to their steady state level of employment, while firms that still need to grow exhaust their credit lines to reach their optimal size and so become constrained. This mechanism is mirrored in our general equilibrium firm dynamics model in the next section.

2.5 Model

In this section we present a heterogeneous firms model with financial frictions which aims to reconcile the stylised facts of Section 2.3. We build on Khan & Thomas (2013) and introduce ex-ante heterogeneity through a permanent productivity component which can be interpreted as the firm's business potential. This will break up the strong correlation between size and financial constraints. Firms with lower permanent productivity will reach their optimal amount of capital and will be unconstrained from then on, while firms which draw a higher permanent component may be constrained even when very large as they are still growing to reach their high potential.

2.5.1 Households

Households choose consumption, savings and labor supply according to the following maximisation problem:

$$\mathrm{V}(k) = \max_{c,l,k'} \left\{ \mathrm{U}(c,l) + \beta \mathbb{E}\mathrm{V}(k') \right\}$$

subject to:

$$k' + c = (1+r)k + \omega l + \mathrm{D},$$

where c is consumption, l is labor, k is capital and D are dividends. ω is the wage, r is the real interest rate. The first-order conditions for the household problem are standard:

$$\begin{aligned} U_l(c,l) &= \omega U_c(c,l) \\ U_c(c,l) &= \beta \mathbb{E}\left[(1+r')U_c(c',l')\right]. \end{aligned}$$

We use the following Greenwood-Hercowitz-Huffman (GHH) utility formulation:

$$U(C,N) = \log(C) + \psi(1-N)$$

Consequently, in the absence of aggregate risk, the first-order conditions are:

$$\begin{aligned} (1+r) &= \frac{1}{\beta} \\ \omega &= \psi C \end{aligned}$$

2.5.2 Production

The production sector features a continuum of firms, indexed by i. Firms are either classified as entrants or incumbents, as detailed below.

Incumbents. An incumbent firm i produces according to the following production function:

$$y_i = \varphi_i k_i^{\alpha} l_i^{\upsilon}, \quad \alpha + \upsilon < 1,$$

where k and l are capital and labor inputs and φ denotes idiosyncratic productivity. Every firm's productivity comprises two components:

$$\log \varphi_i = w_i + \theta_i,$$

where w_i is an idiosyncratic *transitory* productivity shock, which follows an AR(1) process with persistence ρ_w and variance of innovations σ_ε^2. θ_i is the *permanent* productivity component, drawn at birth from a normal distribution with mean μ_θ and variance σ_θ^2 [9]

$$\begin{aligned} &\theta_i \overset{\text{iid}}{\sim} \mathcal{N}\left(\mu_\theta, \sigma_\theta^2\right) \\ &w_i' = \rho_w w_i + \varepsilon_i \quad \varepsilon_i \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right), \quad \left|\rho_w\right| \leq 1 \end{aligned}$$

[9] Henceforth, when we refer to a model with only a transitory shock we mean that $\ln \varphi_i = w_i$.

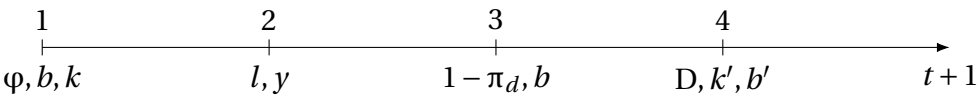


Fig. 2.3: Within period timing of incumbent firm

The firm's total profits before investment are revenue minus labor costs (in what follows we suppress i, the firm indicator, to ease on notation where possible):

$$\pi = y - \omega l,$$

where ω is the wage per unit of labor.

Figure 2.3 summarises the within-period timing of the incumbent. The firm enters the period with predetermined levels of debt b and capital k and immediately observes its idiosyncratic productivity φ composed of a permanent and transitory component. Next, the firm's labor decision is a static choice that can be found through the firm's first order condition:

$$l(k, \varphi; \omega) = \left(\frac{\upsilon\varphi}{\omega} k^{\alpha}\right)^{\frac{1}{1-\upsilon}}.$$

After the production stage, the firm may suffer an exogenous exit shock. The shock happens with probability π_d. Consequently, the value of the firm after the production stage is given by

$$V^1(x, \varphi) = \pi_d x + (1 - \pi_d) V^2(x, \varphi)$$

If the firm survives the exit shock, at the end of the period it chooses debt b' and capital k' to take to the next period and dividends to distribute this period D to maximize its value

$$V^2(x, \varphi) = \max_{k', b', D} \left[D + \mathbb{E}_{\varphi'|\varphi} \Lambda V^1(x', \varphi') \right] \tag{2.3}$$

s.t.:

$$D \equiv x + qb' - k' \geq 0$$

$$b' \leq \xi x$$

$$x' \equiv x(k', b', \varphi') = y(l(k', z'), k', \varphi') - wl(k', \varphi') + (1 - \delta)k' - b'$$

where ξ is the financial parameter that captures the financial frictions in the economy, x is the net worth with which the firm starts the period, given as the sum of profits plus the value of the non-depreciated capital minus the debt the firm has to pay back. q is the

price of the bonds firms issue, with $\frac{1}{q} - 1$ equal to the equilibrium interest rate, r. Λ is the firm discount factor. As the representative household is the owner of the firm, we assume $\Lambda = \beta$ in steady state.

The firm faces two critical constraints according to (2.3). First, the firm cannot issue negative dividends or, equivalently, raise equity. Second, the firm is only able to borrow up to an exogenous fraction ξ of its total cash on hand. We opt for a cash on hand collateral constraint following evidence from Kermani & Ma (2020) or Lian & Ma (2021), which illustrates firms' debt contracts and financial constraints do not depend solely on assets, but also on the firm's value and cash flow. Our measure of cash on hand captures exactly these two sides, as it takes into account both the cash flow and the non-depreciated capital.

Entrants. Entry in this model is exogenous. We assume there is a fixed measure, M_e, of entrants equal to the mass of firms exiting after receiving a death shock. The entrants are assumed to enter with zero debt ($b_0 = 0$) and are log normally distributed over their initial capital k_0 with the mean anchored at a fraction of the mean of optimal capital levels. The choice of a log normal distribution is motivated by the right skewed distribution of entrants in the data. The initial productivity of each entrant, φ_0, follows the same process as the incumbents' productivity. Note that firm entry takes place at the end of a period, and entrants start operating in the next period, given their initial state, (k_0, b_0, φ_0).

2.5.3 Firm level decisions

To characterise the firms' decisions we divide the firms into three groups, following Khan and Thomas (2013). This simplifies the solution of the model significantly.

1. **Unconstrained firms.** Firms that can implement the optimal amount of capital and guarantee that in the future they will never be constrained again.

2. **Constrained firms, type 1.** Firms that can implement the optimal amount of capital but not the minimum savings policy that guarantees they will never be constrained again in the future.

3. **Constrained firms, type 2.** Firms that are constrained and cannot implement the optimal amount of capital nor the minimum savings policy.

For model details on the decisions of the firms in each group see Appendix B.5.

2.5.4 Simplified model predictions

The way in which firms respond to different types of shocks will ultimately depend on whether they have reached their optimal amount of capital or whether they are still growing. Hence, in what follows, we refer to firms which can implement their optimal capital level as being unconstrained and otherwise as constrained. Consequently, type 1 constrained firms are considered unconstrained as they can implement the optimal amount of capital and their investment policy is the same as for unconstrained firms if shocks are relatively small.[10]

To gain more insight into the respective investment elasticities to aggregate shocks and the role of ex-ante heterogeneity, we consider a slightly simplified version of the model as outlined in Appendix B.6. In this model we abstract from labor and assume there is no uncertainty except for a stochastic death shock. The main intuition about differential investment elasticities is captured in Proposition 1.

Proposition 1. Constrained firms are more elastic to an aggregate TFP shock than unconstrained firms, absent any cyclicality in the constraint, if

$$\text{mpk} > \rho \frac{\alpha}{1-\alpha} \frac{1}{1+q_t \xi}.$$

Proof: The proof is provided in Appendix B.6.

Constrained firms will only respond more to an aggregate productivity shock if either their marginal product of capital is large enough, i.e. they are far from their potential, or if the aggregate shock is quickly fading (ρ is close to 0) which gives unconstrained firms little incentive to adjust their capital amount, as productivity is quickly mean reverting. In fact, the elasticity of unconstrained firms is independent of their potential. On the other hand, the marginal product of capital of constrained firms is higher; the higher their potential, the farther they are from reaching their potential.

Hence, the overall aggregate response of output and capital depends on the distribution of constrained firms across the firms size distribution. Furthermore, the financial accelerator mechanism will only be present in the model economy, if Proposition 1 holds on average. In our discussion about aggregate implications below, we separately consider the case of a temporary aggregate shock to total factor productivity (TFP) and a credit shock as a negative shock to borrowing conditions that revert to the steady state value after 1 period.

[10] Large shocks could make the constraint bind again, and they would become strictly constrained.

2.5.5 Solving and calibrating the model

Solution Method. As outlined in Subsection 2.5.3, one can categorize firms into constrained, potentially constrained and unconstrained firms. The two cash-on-hand thresholds that define to which group a firm belongs are derived in Appendix B.5. One can then directly solve for the capital and bond policy function numerically.

To solve for the general equilibrium, we approximate the firm distribution over a fixed grid of net worth using the histogram method proposed by Young (2010).

The steady state solution is then given at the wage which is leading to a clearance of the goods market.[11] Given the steady state wage, we also conduct a Monte Carlo simulation to study the firms' policy responses to aggregate shocks in partial equilibrium.

Calibration. For most of the parameters, which are unrelated to distributions in the model, we follow Khan & Thomas (2013). The set of parameters chosen is documented in the upper part of Table B.12 in Appendix B.2. The discount factor, β, is set to yield an average annual real interest rate of 4%. The production parameters, η and α, imply a labor share of 60% and capital share of 30%, respectively. Leisure preferences imply that households work one third of their available time.

Firm exit rates in the data are heterogeneous and tend to be lower for larger and older firms. In order to account for this without introducing a size based exit rate schedule, we compute a size weighted average exit rate. When not accounting for lower exit rates among performing firms, small firms with high potential are likely to drop out prior to reaching their optimal amount of capital.[12]

The mean productivity levels for the permanent and transitory component, μ_θ and μ_w, are normalised such that when transforming them into a log-normal distribution, the average productivity component equals one.[13] The rest of the parameters are calibrated using the simulated method of moments (SMM).

The values presented in the lower part of Table B.12 in Appendix B.2 minimise the distance between a set of empirical unconditional and conditional moments of the firm

[11] Market clearing interest rates are given by $1/\beta$ due to the household's first-order condition.

[12] The model can still fit the data reasonably well for higher exit rates and far better than a model with just a transitory shock component, yet it gets harder to match the skewness of the firm size distribution as firms with high potential and a long growth path are proportionally more likely to exit before they reach their full size.

[13] Note that the mean of a log-normal distribution is affected not only by the location parameters but also the scale parameter. We adjust it accordingly, such that for any scale parameter, $\mu = 0$ yields an average productivity of 1, when transformed to a log-normal.

Table 2.4: Calibrated model fit

Moment	Data	Model trans. + perm.	Model trans. only
Percentage of const. firms	0.23	0.24	0.24
Share of const. firms in bottom 20%	0.33	0.28	0.94
Size of 90th percentile / median	9.44	9.65	9.75
Size of 90th percentile / bottom 20%	30.24	38.17	51.75
Size of const. firms 90th percentile / median	7.35	7.57	2.43
Size of unconst. firms 90th percentile / median	9.67	9.27	4.90
Asset share of const. firms	0.12	0.13	0.01
Share of const. firms in top 10% vs. bottom 20%	0.36	0.35	0
Percentage of const. firms in top 1%	0.09	0.06	0

Notes. All constrained firms moments are calculated using constrained measure I.

distribution. Table 2.4 compares the fit of a model with just a transitory productivity component to a model including both a transitory and a permanent productivity component.

Both models were separately calibrated to find the best match to the data, except for the collateral constraint parameter ξ. When calibrating the model with just a transitory shock to firms' productivity, we use the same ξ as the one internally calibrated in the two shock model and place extra weight on the unconditional share of constrained firms and size moments. Two reasons motivate us to restrict this calibration: 1) when comparing the aggregate responses across the two models we want the elasticity of constrained firms to be comparable and, as established in Proposition 1, the elasticity of these firms depends directly on ξ. Equally, this allows us to simulate the effect of an identical financial shock in both models. 2) The model with just a transitory shock cannot match the conditional moments such as the relative size of constrained and unconstrained firms, or the share of constrained by size.[14] Given this, we specifically target the fraction of constrained firms and moments of the unconditional size distribution, giving less weight to those moments the model with just a transitory shock could not fit in the first place. This ensures that the underlying size distribution is the same across both models and different predictions are down to differences in the distribution of constrained firms over the firm size distribution.

As documented in the far right column of Table 2.4, the model with just transitory shocks to firms' productivity is not able to match the data well. While it matches the frac-

[14] In Table B.13 in Appendix B.3.3 we report the calibration of the one shock model without any restrictions, with a free ξ and equal weight for all moments. As can be seen, the model is not able to match the conditional moments and the share of constrained firms across the firm distribution cannot match the empirically observed one, as illustrated in Figure B.11.

tion of constrained firms and the unconditional size distribution, it is unable to generate large constrained firms and small unconstrained firms. Hence, with constrained firms concentrated at the bottom of the size distribution, a standard model with just a transitory shock drastically underestimates the asset share of constrained firms. In contrast, when accounting for ex-ante heterogeneity by including a permanent component, and thereby breaking the strong link between size and financial conditions, the model matches the data remarkably well, as documented in the second column of Table 2.4.

2.6 Discussion

In this section we start by discussing how the model fits the joint size constrained distribution presented in Section 2.3 and then explore the impacts of accounting for different permanent productivity components across the distribution of firms.

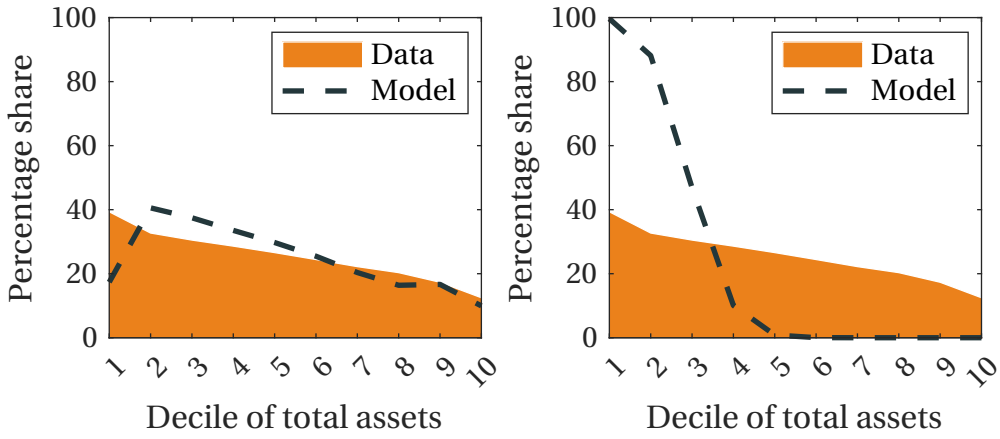
First, we show that while a model that incorporates both transitory and permanent components of the productivity process can generate constrained firms across the entire distribution, similar to what Figure 2.1 suggests, a model with only a transitory component fails to account for this. The implied distributions generated by both models will then cause the model with two productivity components to match the share of total assets in constrained firms, whereas the one shock model will severely underestimate this value.

Second, we assess the implications of accounting for large constrained firms when faced with an aggregate productivity shock and a financial shock, respectively. Furthermore, we compare the degree of misallocation implied by the model including a permanent productivity component to the standard model with just a transitory component.

2.6.1 Replicating the constrained size joint distribution

In section 2.3.1 we highlight that constrained firms are found across the entire distribution of firms. As illustrated in Figure 2.1, even at the top of the distribution in terms of size close to 10% of the firms are financially constrained.

Figure 2.4 compares the model generated share of constrained firms across the size distribution with its empirical equivalent. When disregarding for ex-ante heterogeneity between firms, the model can still produce the same overall share of constrained firms, yet the distribution is completely off. Using only a transitory component, the model can neither generate small unconstrained firms nor large constrained firms as depicted in the right panel of Figure 2.4. On the other hand, the model with a transitory and a permanent



(a) Transitory + permanent shock (b) Transitory shock

Fig. 2.4: Share of constrained firms across the distribution. Constrained firms are identified using measure constrained I which classifies firms as constrained if they have exhausted their potential credit and were not granted additional short- or long-term credit in that period.

productivity component generates small unconstrained firms as well as large constrained firms and is also able to match the untargeted deciles of the empirical distribution quite well, as depicted on the left.

The distribution generated by the model with two shocks is explained by the fact that some larger firms are still growing to reach their steady state capital and are still constrained. At the same time, the model with the two components accounts for a larger share of small firms that are born at or close to their steady state level of capital.[15]

Finally, the distributions generated by both models imply very different values for the share of total assets in constrained firms. Whereas the model with both a permanent and transitory shock generates a value for this moment in line with the data (13% in the model compared to 12% in the data), the model with just a transitory shock only generates 1% of total assets in constrained firms. Again, the permanent component model matches this key empirical moment as outlined in section 2.3.2 whereas the simple, transitory shock model can not.

[15] Figure B.12 in Appendix B.3.3 offers a slightly different perspective, plotting the density distribution of constrained and unconstrained firms. It is possible to observe that while in the two shock model case the distributions overlap, in the one shock case they are completely separated, with the model only generating small constrained firms and large unconstrained ones.

2.6.2 Aggregate effects

We now proceed to assess the aggregate implications of accounting for constrained firms across the entire firm size distribution. We assume a drop in the maximum borrowing capacity of 50%.[16] Given the sudden and transitory nature of the financial shock, we assume wages to be fixed at the general equilibrium level before the shock hits.[17]

Figure 2.5 shows the responses to the credit shock depicted in the upper left panel. Since the firm's capital stock is pre-determined, there is no direct impact in period $t = 2$, when the financial shock hits. However, the lower maximum borrowing capacity affects constrained firms in their investment decision, while unconstrained firms remain unaffected by the shock as they finance their investment completely internally.

The resulting aggregate effect of constrained firms reducing their investment depends heavily on the distribution of these constrained firms along the firm size distribution. In a model with only transitory productivity shocks, all constrained firms will be concentrated at the lower end of the size distribution. When further accounting for the skewness in the firm size distribution, the capital and asset share of these constrained firms becomes marginal. Hence, despite the drastic shock to financing conditions, the aggregate responses in production factors and ultimately output is relatively minor.

However, when accounting for large constrained firms by introducing ex-ante heterogeneity via a permanent productivity component, aggregate effects get massively amplified simply due to the higher capital share of constrained firms. In the case depicted in Figure 2.5 aggregate output drops by about 4 times as much compared to the model without the ex-ante heterogeneity (i.e. peak drop of approximately 10% versus 2.5%). The quantitative magnitude of the effect clearly depends on the fraction of firms identified as being constrained by the different binary measures ranging from 24% (Measure I) to as low as 2% (No potential credit and increasing overdue credit) of all firms. Yet, since all measures are suggestive of the notion that constrained firms exist along the entire firm size distribution, a model with just a transitory productivity component could drastically underestimate the aggregate effects of a credit shock.

When comparing the two models, we can observe that the aggregate investment response is higher in a model with a permanent and a transitory productivity component.

[16] Khan & Thomas (2013) simulate an 88 percentage point drop in ξ. However, in their calibration the initial level of ξ is 1.38. In our calibration ξ is 0.56, hence a 50% drop equals a 28 percentage point drop in maximum borrowing allowances.

[17] General equilibrium results for this exercise lead to the same qualitative conclusions, but we prefer the partial equilibrium analysis to isolate the effect coming from the differences in the distribution of constrained firms.

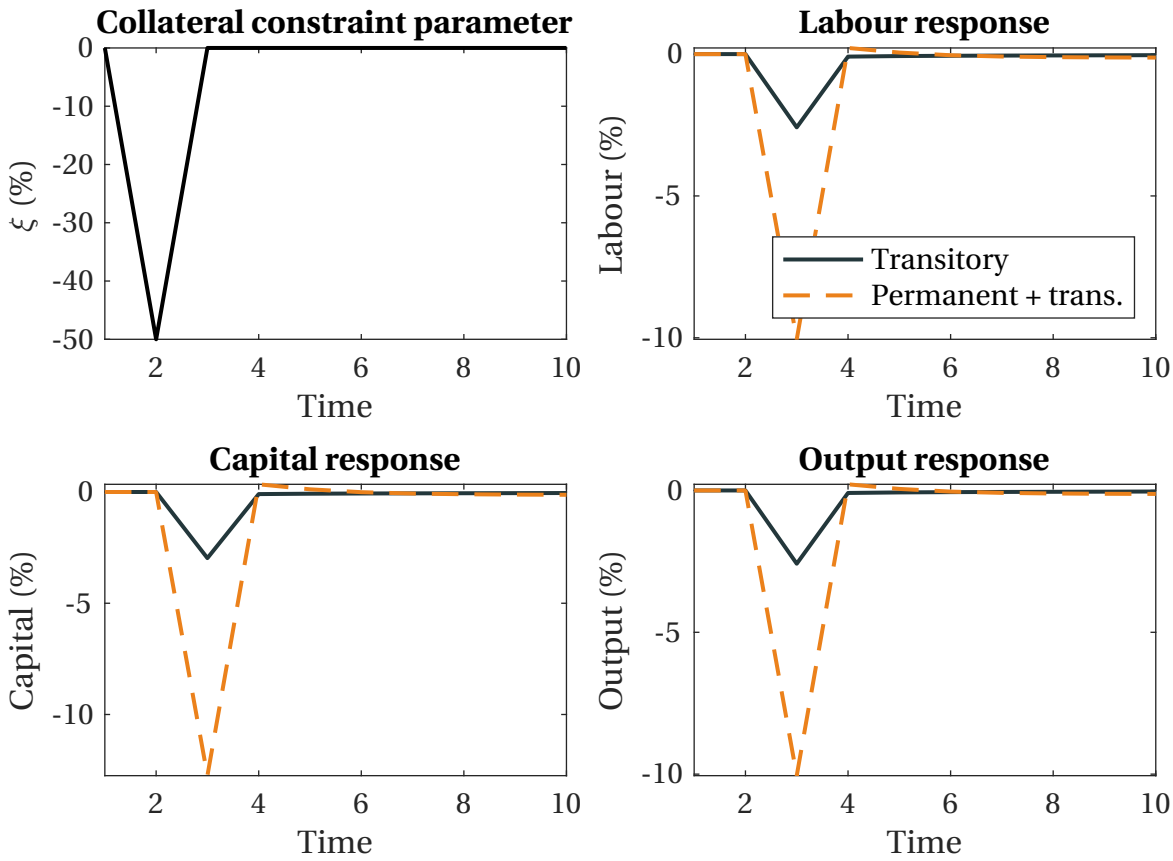


Fig. 2.5: IRFs to a financial shock. Lines indicate the partial equilibrium response to a shock to ξ in the upper left panel, with wages fixed at their steady state level.

This is simply due to the fact that the asset share of constrained firms is substantially larger than in the transitory shock model. This is a direct consequence of the differences in the distribution of constrained firms, as highlighted in Figure 2.4.

In fact, Figure 2.6, which shows the capital elasticity over the size distribution, illustrates the key difference between both models quite well. For unconstrained firms, as already pointed out, the elasticity is zero and for constrained firms, the elasticity is decreasing with size due to decreasing returns to scale. The dashed line is indicating the unconditional average elasticity per decile bin. In a model with just a transitory shock, the overall elasticity is high for small constrained firms but drops to zero at some size cutoff after which all firms become unconstrained. When including a permanent component and thereby generating small unconstrained firms and large constrained firms, the average capital elasticity for small firms is lower than in the one component model but stays above zero for top quantiles of the size distribution. Hence, the capital weighted average elasticity is much higher in the model with a permanent and transitory component, and

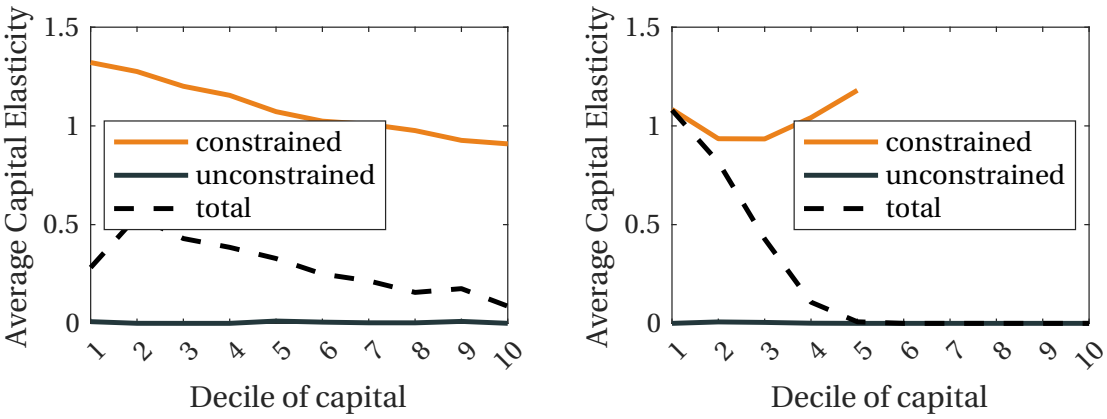


Fig. 2.6: Conditional elasticity over the capital distribution

thus leads to a stronger aggregate capital response.[18]

This result highlights that even a small shock in the financial sector can lead to large aggregate effects due to the granular effects coming from large constrained firms. In fact, to generate a drop in aggregate output similar to the one observed in the U.S. in the Great Financial Crisis, one would only need a 12.5% decrease in the collateral constraint parameter, opposite to shocks larger than 50% as in, for example, Khan & Thomas (2013).

Additionally, in Figure B.15 in Appendix B.3.3 we consider an unexpected and temporary 1% increase in total factor productivity (TFP). In a direct response to the shock, firms employ more labor for any predetermined level of capital. While unconstrained firms do not increase their investment in capital due to the transitory nature of the shock, constrained firms leverage their increased net worth to borrow more. This explains why the lagged response in capital is much smaller than the response in labor, as only constrained firms react to the shown case of $\rho = 0$, which is only 23% of all firms in this calibration.

However, given the small magnitude of the capital response relative to the response in labor, the difference barely shows up in aggregate output. Note, the effect would become stronger if the borrowing constraint was cyclical or the fraction of constrained firms in the economy was higher.[19]

Lastly, we assess the implication of accounting for the firm size distribution in terms of capital misallocation and the effects of policies that target small firms. In terms of capital misallocation, a more realistic distribution of financial constrained firms can amplify

[18]Further, Figure B.13 in Appendix B.3.3 reinforces that the mechanism comes from where the constrained firms are and not from the average elasticity of constrained firms. It depicts the elasticity density distribution across the two models. As these are not very different it follows that the large share of differential aggregate capital elasticity comes from the fact that the model features large constrained firms.

[19]One should also note that the difference between the models would vanish and eventually flip if the TFP shock gets more persistent and unconstrained firms become more cyclical, as shown in proposition 1.

output losses induced by financial frictions up to three times larger. This is due to the higher dispersion of MPKs and the existence of large constrained firms, which create larger MPKs at the top of the distribution. In terms of policy implications, we show that, when a more realistic constrained firm distribution is considered, the effects of policies that aim to help small firms are more limited, with the impact on output being 12% lower than in the model with just one shock, which does not match the constrained size joint distribution. More details regarding these results can be found in Appendix B.7.

2.7 Conclusion

This paper documents three empirical facts that are counterfactual to the standard heterogeneous firm financial frictions model and subsequently analyses the importance of matching these facts in a quantitative financial frictions model with heterogeneous firms. Empirically we show that at any point of the firm distribution there are both constrained and unconstrained firms. These constrained firms also account for an elevated share of assets and are more cyclical, independent of size. We proceed to demonstrate that heterogeneous ex-ante conditions, a potential explanatory factor for these facts, exist and affect constrained and unconstrained firms differently.

Next, we build a standard firm dynamic model, adding a permanent productivity component. We demonstrate that by adding this extra component to the productivity process helps us match the distribution of constrained firms across the size distribution, breaking the typical strong correlation between financial constraints and size, generating a large mass of small unconstrained and large constrained firms. This mechanism has significant implications for aggregate responses to shocks. We find aggregate capital and output respond slightly more to a productivity shock in a model that accounts for ex-ante firm heterogeneity than a model where idiosyncratic productivity is purely driven by a transitory component. More importantly, the effects of a financial shock are strongly affected, with up to four times higher aggregate cyclicality compared to the standard model. This is due to large constrained firms which a model with only a transitory component is unable to generate.

References

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

Amiti, M. & Weinstein, D. E. (2018). How much do idiosyncratic bank shocks affect investment? evidence from matched bank-firm loan data. *Journal of Political Economy*, 126(2), 525–587.

Bodnaruk, A., Loughran, T., & McDonald, B. (2015). Using 10-k text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, 50(4), 623–646.

Buehlmaier, M. M. & Whited, T. M. (2018). Are financial constraints priced? evidence from textual analysis. *The Review of Financial Studies*, 31(7), 2693–2728.

Buera, F. & Karmakar, S. (2019). *Real effects of financial distress: the role of heterogeneity*. Bank of England working papers 814, Bank of England.

Chaney, T., Sraer, D., & Thesmar, D. (2012). The collateral channel: How real estate shocks affect corporate investment. *American Economic Review*, 102(6), 2381–2409.

Chodorow-Reich, G. (2014). The employment effects of credit market disruptions: Firm-level evidence from the 2008–9 financial crisis. *The Quarterly Journal of Economics*, 129(1), 1–59.

Cloyne, J., Ferreira, C., Froemel, M., & Surico, P. (2018). *Monetary Policy, Corporate Finance and Investment*. Working Paper 25366, National Bureau of Economic Research.

Cooley, T. F. & Quadrini, V. (2001). Financial Markets and Firm Dynamics. *American Economic Review*, 91(5), 1286–1310.

Crouzet, N. & Mehrotra, N. (2020). Small and Large Firms over the Business Cycle. *American Economic Review*. Conditionally accepted.

Custodio, C., Bonfim, D., & Raposo, C. C. (2021). The sensitivity of sme's investment and employment to the cost of debt financing. *Available at SSRN 3879737*.

Farre-Mensa, J. & Ljungqvist, A. (2016). Do measures of financial constraints measure financial constraints? *The Review of Financial Studies*, 29(2), 271–308.

Gan, J. (2007). Collateral, debt capacity, and corporate investment: Evidence from a natural experiment. *Journal of Financial Economics*, 85(3), 709–734.

Gertler, M. & Gilchrist, S. (1994). Monetary Policy, Business Cycles, and the Behavior of Small Manufacturing Firms. *The Quarterly Journal of Economics*, 109(2), 309–340.

Gilchrist, S. & Himmelberg, C. P. (1995). Evidence on the role of cash flow for investment. *Journal of Monetary Economics*, 36(3), 541–572.

Greenwald, D. et al. (2019). *Firm debt covenants and the macroeconomy: The interest coverage channel*, volume 1. MIT Sloan School of Management.

Hadlock, C. J. & Pierce, J. R. (2010). New evidence on measuring financial constraints: Moving beyond the kz index. *The Review of Financial Studies*, 23(5), 1909–1940.

Hopenhayn, H. A. (1992). Entry, Exit, and Firm Dynamics in Long Run Equilibrium. *Econometrica*, 60(5), 1127–1150.

Kaplan, S. N. & Zingales, L. (1997). Do investment-cash flow sensitivities provide useful measures of financing constraints? *The quarterly journal of economics*, 112(1), 169–215.

Kermani, A. & Ma, Y. (2020). *Two tales of debt*. Technical report, National Bureau of Economic Research.

Khan, A. & Thomas, J. K. (2013). Credit shocks and aggregate fluctuations in an economy with production heterogeneity. *Journal of Political Economy*, 121(6), 1055–1107.

Lian, C. & Ma, Y. (2021). Anatomy of corporate borrowing constraints. *The Quarterly Journal of Economics*, 136(1), 229–291.

Mehrotra, N. & Sergeyev, D. (2020). Financial shocks, firm credit and the great recession. *Journal of Monetary Economics*.

Ottonello, P. & Winberry, T. (2018). *Financial heterogeneity and the investment channel of monetary policy*. Technical report, National Bureau of Economic Research.

Pugsley, B., Sedláček, P., & Sterk, V. (2021). The nature of firm growth. *American Economic Review*, 111(2), 547–79.

Rampini, A. A. & Viswanathan, S. (2020). *Collateral and secured debt*. Technical report.

Sharpe, S. (1994). Financial market imperfections, firm leverage, and the cyclicality of employment. *American Economic Review*, 84(4), 1060–74.

Whited, T. M. & Wu, G. (2006). Financial constraints risk. *The Review of Financial Studies*, 19(2), 531–559.

Young, E. R. (2010). Solving the incomplete markets model with aggregate uncertainty using the krusell–smith algorithm and non-stochastic simulations. *Journal of Economic Dynamics and Control*, 34(1), 36–41.

Chapter 3

A Global solution to truncated history models of heterogeneous agents

Abstract

I present a global solution method for the truncated history method developed by Le Grand & Ragot (2022a) to solve heterogeneous agent models with aggregate shocks. Their method provides a finite state-space representation of economies by truncating idiosyncratic histories. The global solution to this representation is based on the parameterized expectations algorithm developed by den Haan & Marcet (1990). This solution algorithm is accurate and does not rely on the assumption of a constant set of constrained agents. Further it is applicable to model setups where local methods may struggle or fail. I provide three examples where this is the case; cyclical borrowing constraints, uncertainty shocks and a two asset economy. I show that all three modifications can be solved with the global method and provide interesting insights into the importance of heterogeneity beyond the standard model.

Keywords: Heterogeneous agents, numerical methods, global solution
JEL Codes: D31, D52, E21

3.1 Introduction

Models with incomplete markets and heterogeneous agents are becoming the new workhorse of theoretical and computational macroeconomic analysis, providing crucial insights into the distributional effects on macroeconomic aggregates and *vice versa*. Most of these environments feature infinitely-lived agents that can only borrow up to some exogenous limit and are subject to idiosyncratic productivity shocks (Aiyagari, 1994; Bewley, 1983; Huggett, 1993; İmrohoroğlu, 1998). These assumptions enable a detailed analysis of the relationship between micro- and macroeconomic developments within an economy and provide a more granular understanding of frictions in credit markets.

However, as is well known in the literature, these models feature an infinite-dimensional and time-varying state in the presence of aggregate uncertainty; the wealth distribution. This state makes these environments notoriously costly to solve numerically and almost always requires some approximating assumption. A recent solution method in this research area was proposed by François Le Grand and Xavier Ragot, focussing on a truncation method to cast the environment into a finite-space representation.[1] In particular, their method involves a truncation in the space of idiosyncratic histories, such that all agents with the same shock history for the last N periods are assumed to belong to the same history bin, irrespective of their shock history before that. Using this logic one can aggregate individual choices and constraints so that the resulting environment is represented in terms of a finite number of bins instead of a continuum of individual agents. The projected model can then replicate the aggregate dynamics of the original environment with the introduction of a finite number of time-varying correcting coefficients in the Euler Equation. Fixing these coefficients equal to their steady state values over the dynamic solution of the model is the key approximation assumption that enables accurate numerical solutions to this environment. This method is particularly appealing since it effectively discretizes a distribution without having to rely on very large number of agents (such as the Reiter (2009) method) and, due to the finite-space representation in the history space, can be used to study optimal Ramsey policy with heterogeneous agents.

In this chapter I propose a global solution algorithm to this truncated history environment, based on the parameterized expectations algorithm by den Haan & Marcet (1990). This involves directly parameterizing the expectations operator in the Euler Equations of the different history bins and simulating the economy until convergence of the coefficients on the respective state variables. The method has a number of advantages over the

[1] This methodology builds on earlier work by Bilbiie (2019); Bilbiie & Ragot (2021); Challe & Ragot (2016) among others.

local perturbation methods used to solve these environments up until now. First, unlike local methods, the global solution method does not rely on a constant set of constrained history bins in the presence of aggregate risk, as it can deal with occasionally binding constraints very well (Christiano & Fisher, 2000). Consequently, this solution method provides very similar aggregate results to the local method for small innovations, a successful validation for this solution mechanism. I demonstrate this by comparing the stochastic simulations and impulse response functions of both the local and global solution method. However, the results differ for the consumption response of the credit-constrained bins. The reason is, as mentioned above, that the local method assumes that all agents remain credit-constrained in the face of aggregate risk whilst the global method does not rely on this assumption. It follows that after a positive TFP shock the bottom consumption response is smaller on impact but more persistent in the global solution, as some agents are on their Euler equation in the stochastic simulation. This result forms the first contribution of this chapter. It suggests that if one is interested in solving for aggregate behavior of these environments for small shocks, the solution method does not matter much and one is probably better off with the faster, local method. However, if the focus is on the bottom end of the income distribution the results differ considerably and suggest that the global solution provides a better approximation in that case. This may be quite important if there is a large share of agents that moves between their borrowing constraint and the Euler Equation over the business cycle.

The second contribution is to show the applicability of the global method to solve environments where local perturbation methods may either struggle or fail entirely. I provide three examples; a setting with cyclical borrowing constraints, uncertainty shocks and multiple assets. First, when borrowing constraints depend on the business cycle the standard method of perturbation fails as the steady state equations around which it finds the local approximation are false, even for small shocks. I show that the global method can solve this issue easily and that especially the bottom of distribution exhibits a much higher contemporary elasticity to consume in this setting. Second, I study an environment with aggregate uncertainty shocks. Solving models with uncertainty shocks is technically feasible with local methods, but involves additional tools such as pruning algorithm by Andreasen et al. (2018). On the other hand, I show that the global method can handle uncertainty shocks in heterogeneous agent models quite well. In particular I study a standard Krusell-Smith economy where the second moment of the TFP process is also exogenously time-varying. After an increase in uncertainty in this economy output increases and the capital stock increase, as precautionary savings increase, creating a

supply-side driven boom after an uncertainty shock. As this is counterfactual to most of the empirical evidence, I introduce an aggregate demand externality that endogenously creates recessions after uncertainty shocks. Finally, the third example analyses an economy with a second, safe but non-productive asset in the economy. As documented in the literature, there are investor clienteles. Unprodutive agents save mostly in the safe asset whereas richer, more productive agent exhibit a mixed portfolio. Note that this is again a setting where the local perturbation approach would fail as the portfolio allocation at steady state is indeterminate. PEA, by using stochastic simulations in the algorithm, can avoid this as the portfolio choice is pinned down by the risk characteristics and can potentially provide answers to an optimal policy exercise in this setting.

In general this chapter makes two contributions, a numerical and an economic one. The numeric contribution is to show that there is a feasible, sophisticated and accurate alternative to the local approximation method - the parameterized expectations algorithm adapted to reduced history environments. Moreover, I show that multiple environments that may be highly relevant in the context of heterogeneity can be solved successfully without having to rely on either local approximations for the dynamic response or parameterized laws of motion for the states. The economic contribution is to study these extensions to gain additional economic understanding of the importance of heterogeneity in macroeconomic outcomes.

Nevertheless, that there are some limitations to the methodology proposed in this chapter. Most importantly, the method assumes that the within-bin heterogeneity stays constant and equal to the steady state value over any stochastic simulation. In models where there are large changes in the distribution over the cycle and where these matter considerably for optimality, the presented solution algorithm will yield inferior results. Further, the computation time is also considerably higher compared to the perturbation method, as it solves for policy functions across entire stochastic state space.

Literature. As has been outlined by Krusell & Smith (1998) the main issue with solving heterogeneous agent models with aggregate risk is that there is a state variable that is infinitely dimensional and time-varying; the distribution over wealth and idiosyncratic productivity. Agents experience different histories of idiosyncratic shocks, make different idiosyncratic savings choices and thus end up at different areas of the wealth-productivity distribution. As time increases, the amount of possible histories goes to infinity, making the resultant wealth distribution infinitely dimensional. Aggregate shocks influence households' decisions, too, and so the distribution is time-varying in the presence of

aggregate risk. Finally, the distribution is a state variable as agents need it to forecast prices away from the deterministic steady state.

The literature on numerical methods has dealt with this issue in two major ways. The first technique approximates the distribution with a small number of moments, for example its mean (Krusell & Smith, 1998) or a parameterized distribution (Algan et al., 2008). One then specifies a law of motion for these moments and solves the household's problem conditional on this law of motion. Iterating on a simulation until the conjectured law of motion fits the implied moments fairly well completes the algorithm. This method works well in many applications, but it involves repeatedly simulating the economy and solving the household problem for different prices, which can be very time-consuming. A recent application is provided by Fernández-Villaverde et al. (2019) who use neural networks for approximating a non-linear aggregate law of motion.

The dual step of solving and simulation can be avoided with the second technique, solving for individual decision rules at the steady state and then linearly approximating aggregate behavior in the neighborhood of the steady state. One major advantage is that the wealth distribution (or moments thereof) are not state variable in the individual problem. One can therefore solve the individual problem using value function iteration or projection. Then one approximates the aggregate behavior of the economy around the steady state using perturbation (Reiter (2009)), MIT shocks (Boppart et al. (2018)) or sequence-based Jacobians (Auclert et al. (2021)).

More recently, a number of paper have also used deep neural networks to solve models with a more complete distribution of agents in a simulation approach. Examples of these include Azinovic et al. (2022) who solve a Krusell-Smith economy where agents take into account the distribution encoded with the histogram method by Young (2010) and Maliar et al. (2021) who simulate a panel of 1000 agents. Finally, a number of papers use the advantages of continuous time to both gain additional analytical results and improve the speed of numerical solutions. Examples include Achdou et al. (2022) and Kaplan et al. (2018).

This chapter relates to this literature by proposing an alternative method that can solve discrete time Krusell-Smith models without having to rely on an approximated law of motion or the additional simulation step in the standard algorithm. At the same time, the model can be solved without linearization, enabling the analysis of richer environments, as outlined above.

Outlook. The chapter is structured as follows. In section 3.2 I describe the basic environment for the analysis of heterogeneous agents with aggregate risk. section 3.3 outlines the reduced history method, whilst section 3.4 discusses and compares both the local and the newly proposed global solution method to truncated economies. In section 3.5 I use the global solution method for some settings where the local method either struggles or completely fails. Finally, section 3.6 concludes.

3.2 Environment

In order to keep the exposition simple and comparable, I closely follow the setup of Le Grand & Ragot (2022b). The environment is a simple heterogeneous-agent economy with aggregate productivity risk as popularized by Krusell & Smith (1998). There is a continuum of ex-ante identical agents of measure one, one homogeneous good and time is discrete.

3.2.1 Risk

I assume that there are two types of risk; idiosyncratic and aggregate risk.

Idiosyncratic risk. Idiosyncratic risk derives from a stochastic labor productivity process. Agents cannot insure against labor productivity shocks, as they are not able to borrow against future income shocks. Idiosyncratic productivity is denoted by y and can take values in a finite set $\mathscr{Y}$. The productivity process of a given agent follows a first-order Markov chain with constant transition probabilities $\left(\Pi_{yy'}\right)_{y,y'\in\mathscr{Y}}$. The probability that an agent currently endowed with productivity y will have productivity y' in the following period is equal to $\Pi_{yy'} \in [0,1]$. Furthermore, transition probabilities satisfy the following: $\Pi_{yy'} \geq 0$ and $\sum_{y'\in\mathscr{Y}} \Pi_{yy'} = 1$. An individual history of productivity shocks up to date t is denoted by $y^t = y_0, \ldots, y_t \in \mathscr{Y}^{t+1}$. Finally, agents are assumed to have inelastic labor supply equal to 1 so that their labor productivity equals their effective labor supply.

Aggregate risk. In the standard setup, aggregate risk only affects total labor productivity (TFP) of the representative firm. The current value of TFP is denoted by Z_t, and it follows a continuous Markov process. Note that Section 3.5 will introduce a second source of risk, stochastic volatility of the TFP process.

3.2.2 Production

The supply side of the economy is deliberately simple and standard. There is a representative firm that uses aggregate inputs of Labour $L_t = \bar{L}$ and capital K_t to produce output Y_t with current productivity Z_t. The parameterization used is Cobb-Douglas Production function with capital intensity $\alpha \in (0,1)$ and capital depreciation rate $\delta \in (0,1)$:

$$Y_t = Z_t K_{t-1}^{\alpha} \bar{L}^{1-\alpha} - \delta K_{t-1}, \tag{3.1}$$

where the capital K_{t-1} is requested to be installed one period in advance. The firm rents labor and capital at respective factor prices w_t and r_t. The profit maximization conditions of the firm imply the following expression for factor prices:

$$w_t = (1-\alpha) Z_t K_{t-1}^{\alpha} \bar{L}^{-\alpha} \text{ and } r_t = \alpha Z_t K_{t-1}^{\alpha-1} \bar{L}^{1-\alpha} - \delta. \tag{3.2}$$

3.2.3 Households

Household derive instantaneous utility from consumption, denoted by u. It is assumed that $u : \mathbb{R}_+ \to \mathbb{R}$ is twice continuously differentiable, strictly increasing, and strictly concave, with $u'(0) = \infty$. As usual, households maximize the expected sum of discounted lifetime utility: $\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t)$, where: $0 < \beta < 1$ is the discount factor, $(c_t)_{t\geq 0}$ a consumption path, and $\mathbb{E}_0$ the expectation operator over future aggregate and individual shocks.

Apart from consuming the homogeneous good, households can save in capital shares k_t, but cannot borrow. Capital is risky, as households do not know the interest rate paid out tomorrow, r_{t+1}. The borrowing constraint implies that households cannot insure against productivity shocks and creates a distribution of wealth. Given an initial endowment k_{-1} and an initial productivity level y_0, households choose their consumption path $(c_t)_{t\geq 0}$ and saving plans $(k_t)_{t\geq 0}$ in order to maximize their utility, subject to the intertemporal budget and borrowing constraints. This can be written as:

$$\begin{gathered} \max_{(c_t,k_t)_{t\geq 0}} \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t) \\ c_t + k_t \leq (1+r_t) k_{t-1} + w_t y_t \\ k_t \geq 0, \\ k_{-1}, y_0 \text{ given.} \end{gathered} \tag{3.3}$$

As is well known in the literature, the optimal behaviour of households is chararcterized

by the Karush-Kuhn-Tucker conditions for this problem:

$$u'(c_t) = \beta \mathbb{E}_t \left[u'(c_{t+1})(1 + r_{t+1}) \right] + \lambda_t \tag{3.4}$$

$$\lambda_t k_t = 0 \tag{3.5}$$

$$k_t \geq 0; \lambda \geq 0 \tag{3.6}$$

where the first equation is the Euler Equation, the second equation the complementary slackness condition and third condition denotes the borrowing constraint and positivity constraints.

3.2.4 Aggregation and equilibrium

As agents are ex-ante identical but experience different productivity histories, I distinguish these ex-post heterogeneous agents by subscript i. Their distribution of these agents over financial wealth and productivity is denoted by $\Gamma_t \left(k_{i,t}, y_{i,t}\right)$. Hence the financial market clearing condition is that the sum of individual savings equals total capital supply:

$$\int k_{i,t} d\Gamma_t(k_{i,t}, y_{i,t}) = \mathrm{K}_t, \tag{3.7}$$

whilst the labour market clearing condition is that the sum of all individual productivities equal effective labour supply:

$$\int y_i d\Gamma_t(k_{i,t}, y_{i,t}) = \bar{\mathrm{L}}. \tag{3.8}$$

Again, total labour supply is constant since the individual labour supply decision is exogenous and aggregate risk does not influence the transition probabilities between different productivity states. Finally, the resource constraint in this economy implies that aggregate consumption plus aggregate capital equals total output and non-depreciated capital:

$$\int c_{i,t} d\Gamma_t(k_{i,t}, y_{i,t}) + \mathrm{K}_t = \mathrm{Y}_t + \mathrm{K}_{t-1}. \tag{3.9}$$

Having described the risk, the production sector, the household sector and the aggregation equations, the market equilibrium in this economy follows and is relatively standard.

Definition 3.2.1 (Market Equilibrium)**.** A competitive equilibrium is a collection of individual plans $\left(c_{i,t}, k_{i,t}\right)_{t \geq 0, i \in \mathscr{I}}$, of aggregate quantities $(\mathrm{K}_t, \mathrm{Y}_t)_{t \geq 0}$, and of price processes

$(w_t, r_t)_{t\geq 0}$, such that, for an initial wealth and productivity distribution $\left(k_{i,-1}, y_{i,0}\right)_{i\in\mathscr{I}}$, and for initial values of capital stock verifying $\mathrm{K}_{-1} = \int k_{-1,i} d\Gamma_t(k_{i,-1}, y_{i,0})$, we have:

- given prices, the functions $\left(c_{i,t}, k_{i,t}\right)_{t\geq 0, i\in\mathscr{I}}$ solve the agent's optimization program in equations (3.3)
- financial and goods markets clear at all dates: for any $t \geq 0$, equations (3.7), (3.8) and (3.9) hold
- factor prices $(w_t, r_t)_{t\geq 0}$ are consistent with profit maximization conditions (3.2)

3.3 Reduced history truncation

In this section I outline the method to recast this model into finite-state representation, namely the truncation method offerd by Le Grand & Ragot (2022a). This method groups together agents according to their $\mathrm{N} > 0$ past idiosyncratic shocks, independent of their idiosyncratic history anytime before $t - (\mathrm{N}+1)$. N is called the truncation length and a numerical parameter that ultimately governs the coarseness of the approximation. Take for example $\mathrm{N} = 2$ in a world with where y, the productivity shock, can only take two values, 0 and 1. In that case there would be four histories in the truncated representation, $\{(0,0),(0,1),(1,0),(0,0)\}$, and the infinite-dimensional distribution has been collapsed to only four dimensions.

A general framework. Following Le Grand & Ragot (2022b), a truncated history or history bin is a vector $y^{\mathrm{N}} = \left(y_{-\mathrm{N}+1}, \ldots, y_{-1}, y_0\right)$, where y_0 represents the current productivity status, and $y_{-\mathrm{N}+1}$ the productivity N periods ago. The truncation method assigns agents with infinite history $y^{\infty} = \left(\ldots, y_{-\mathrm{N}-1}, y_{-\mathrm{N}}, y_{-\mathrm{N}+1}, y_{-\mathrm{N}+2}, \ldots, y_{-1}, y_0\right)$ at date t to the truncated history y^{N}, independently of productivity levels more than N periods ago. The infinite history of a given agent is updated in every period, and the composition of truncated histories does not remain fixed through time. For instance, if we assume that a agent with history y^{∞} at t draws the productivity $\tilde{y}_0$ at date $t+1$, her history at $t+1$ will become: $\tilde{y}^{\infty} = \left(\ldots, y_{-\mathrm{N}-1}, y_{-\mathrm{N}}, y_{-\mathrm{N}+1}, \ldots, y_{-1}, y_0, \tilde{y}_0\right)$. At $t+1$ the agent will be assigned to truncated history $\tilde{y}^{\mathrm{N}} = \left(y_{-\mathrm{N}+2}, \ldots, y_1, y_0, \tilde{y}_0\right)$. Consequently, if $n_y = \mathrm{Card}(\mathscr{Y})$ denotes the number of productivity levels, there will be $\mathrm{N}_{\mathrm{tot}} = n_y^{\mathrm{N}}$ truncated histories of length N.

Transition probabilities. Agents move from their current history bin y^{N} to the next history bin $\tilde{y}^{\mathrm{N}}$ according to the probability $\Pi_{y^{\mathrm{N}}, \tilde{y}^{\mathrm{N}}}$ which can be inferred from the exogenous

productivity transition matrix and whether the follow-up history is indeed a continuation. Formally:

$$\Pi_{y^{\mathrm{N}},\tilde{y}^{\mathrm{N}}} = \mathbb{1}_{\tilde{y}^{\mathrm{N}} \succcurlyeq y^{\mathrm{N}}} \Pi_{y_0,\tilde{y}_0} \tag{3.10}$$

where $\Pi_{y_0,\tilde{y}_0}$ is the exogenous productivity transition matrix, and $\mathbb{1}_{\tilde{y}^{\mathrm{N}} \succcurlyeq y^{\mathrm{N}}}$ is an indicator function for whenever $\tilde{y}^{\mathrm{N}}$ is indeed a successor history of y^{N}. In the above example, the successor histories to the current history bin $(0,0)$ would be $(0,0)$ and $(0,1)$. As the quantities in the exogenous transition matrix are positive, the quantitites $\Pi_{y^{\mathrm{N}},\tilde{y}^{\mathrm{N}}}$ are also larger or equal to 0. Together with the fact that any agent always belongs to one and only one history, the transition matrix for all histories, $(\Pi_{y^{\mathrm{N}},\tilde{y}})_{y^{\mathrm{N}},\tilde{y}^{\mathrm{N}} \in \mathscr{Y}^{\mathrm{N}}}$, is well defined and sums to 1 for all possible histories y^{N}.

Sizes. The size of every truncated history $\mathrm{S}_{y^{\mathrm{N}}}$ can be derived from the transition matrix:

$$\mathrm{S}_{y^{\mathrm{N}}} = \sum_{\hat{y}^{\mathrm{N}} \in \mathscr{Y}^{\mathrm{N}}} \mathrm{S}_{\hat{y}^{\mathrm{N}}} \Pi_{\hat{y}^{\mathrm{N}} y^{\mathrm{N}}} \tag{3.11}$$

meaning that the size of history y^{N} is the sum of predecessor history sizes, multiplied by their respective transition probability towards y^{N}.

Savings. Every history bin can choose their own end-of-period savings $k_{t,y^{\mathrm{N}}}$ that translate to beginning-of-period savings $\tilde{k}_{t+1,y^{\mathrm{N}}}$ next period. These are again connected through the transition probabilities as the share of agents that transition to a new history take that fraction of end-of-period savings with them. Formally:

$$\tilde{k}_{t,y^{\mathrm{N}}} = \frac{1}{\mathrm{S}_{y^{\mathrm{N}}}} \sum_{\hat{y}^{\mathrm{N}} \in \mathscr{Y}^{\mathrm{N}}} \mathrm{S}_{\hat{y}^{\mathrm{N}}} \Pi_{\hat{y}^{\mathrm{N}} y^{\mathrm{N}}} k_{t-1,\hat{y}^{\mathrm{N}}} \tag{3.12}$$

Budget constraint. Having defined transitions, sizes and savings, the budget constraint for every history bin follows:

$$c_{t,y^{\mathrm{N}}} + k_{t,y^{\mathrm{N}}} \leq (1 + r_t)\tilde{k}_{t,y^{\mathrm{N}}} + w_t y_{0,y^{\mathrm{N}}} \tag{3.13}$$

where $c_{t,y^{\mathrm{N}}}$ denotes consumption, r_t the interest rate, w_t the wage and $y_{0,y^{\mathrm{N}}}$ the current productivity of history y^{N}.

Optimality conditions. Further, one can also translate the Karush-Kuhn-Tucker conditions (3.4) - (3.6) into the reduced history representation. Importantly, however, as the

marginal utility of aggregated consumption is different from the aggregated marginal utility, one must correct for this within-history heterogeneity. This is done by introducing the (possibly time-varying) parameters $\xi_{t,y^{\mathrm{N}}}$ that measure the within-bin heterogeneity by computing the ratio of the marginal utility of aggregate consumption and aggregated marginal utility. Only with these parameters included the Euler equations holds also for aggreate consumption levels. Formally:

$$\xi_{t,y^{\mathrm{N}}} u'(c_{t,y^{\mathrm{N}}}) = \beta \mathbb{E}_t \left[(1 + r_{t+1}) \sum_{\tilde{y}^{\mathrm{N}} \in \mathscr{Y}^{\mathrm{N}}} \Pi_{y^{\mathrm{N}},\tilde{y}^{\mathrm{N}}} \xi_{t,\tilde{y}^{\mathrm{N}}} u'(c_{t+1,\tilde{y}^{\mathrm{N}}}) \right] + \lambda_{y^{\mathrm{N}}} \tag{3.14}$$

$$k_{t,y^{\mathrm{N}}} \lambda_{t,y^{\mathrm{N}}} = 0 \tag{3.15}$$

$$k_{t,y^{\mathrm{N}}} \geq 0, \lambda_{t,y^{\mathrm{N}}} \geq 0 \tag{3.16}$$

where $\lambda_{t,y^{\mathrm{N}}}$ denotes the multiplier on the borrowing constraint. I will outline in section 3.4 how to compute the $\xi_{t,y^{\mathrm{N}}}$s in more detail.

Aggregate variables. In the truncated representation, aggregate variables are simply the discrete counterpart of the ones above. In particular, the capital stock in the economy is:

$$\mathrm{K}_t = \sum_{y^{\mathrm{N}} \in \mathscr{Y}^{\mathrm{N}}} \mathrm{S}_{y^{\mathrm{N}}} k_{t,y^{\mathrm{N}}} \tag{3.17}$$

and total labour is:

$$\mathrm{L}_t = \sum_{y^{\mathrm{N}} \in \mathscr{Y}^{\mathrm{N}}} \mathrm{S}_{y^{\mathrm{N}}} y_{0,y^{\mathrm{N}}} = \bar{\mathrm{L}} \tag{3.18}$$

and the resource constraint is

$$\sum_{y^{\mathrm{N}} \in \mathscr{Y}^{\mathrm{N}}} \mathrm{S}_{y^{\mathrm{N}}} c_{y^{\mathrm{N}},t} + \mathrm{K}_t = \mathrm{Y}_t + \mathrm{K}_{t-1}. \tag{3.19}$$

The rest of the equations, such as the first-order conditions of the representative firm stay the same as in the fully-fledged model.

3.4 Numerical solutions

With the description of the exact truncated model at hand, this Section will first outline how to obtain the residual heterogeneity parameters at steady state. Then I proceed to describe and compare two numerical solution methods to truncated history models. The

first, used by Le Grand & Ragot (2022a,b), relies on local pertubations of the dynamic equations around the steady state. The second, the solution method of this paper, uses the parameterized expectations algorithm to provide a global solution to the truncation.

3.4.1 Obtaining the residual heterogeneity parameters

As outlined in the previous section, the residual parameters $(\xi_{y^N})_{y^N \in \mathscr{Y}^N}$ are key to make the truncated model conistent with the fully-fledged enviroment outlined in section 3.2. In the presence of aggregate shocks these parameters are time-varying and impossible to compute at every point of the stochastic simulation. However, these parameters can be obtained *at steady state* quite easily. So one critical assumption for both numerical solution methods is as follows:

Assumption 1. The preference parameters $(\xi_{y^N})_{y^N \in \mathscr{Y}^N}$ remain constant and equal to their steady-state values.

One can solve for the fully fledged model at steady state quite easily with methods such as the endogenous Grid method. From the steady state wealth distribution, the set of credit-constrained histories can be identified (i.e. where $\lambda_{y^N} > 0$). Further, at steady state, the aggregated history variables can be calculated. Using linear algebra, the Euler Equation can then be inverted to find the preference parameters $\xi_y^N \in \mathscr{Y}^N$. I refer the reader to the excellent explanation in Le Grand & Ragot (2022a), explaining the details of this methods.

3.4.2 Local method: Perturbation

One easy and relatively quick method to solve the truncated economy is by perturbation of the model equations (3.11) - (3.19). However, there is one critical additional assumption, namely that all histories that were identified as credit-constrained remain so during the presence of aggregate disturbances. The reason for this assumption is that incorporating dozens of possible occasionally binding constraints for all history bins quickly becomes cumbersome in local perturbation methods.[2] So, this assumption enables packages like Dynare to handle the set of equations with ease. Using this assumption the system of

[2]Despite recent advances, e.g. see Guerrieri & Iacoviello (2015), these occassionally binding constraints are mostly only piece wise linear and have difficulty capturing precautionary savings.

equation to be solved with perturbation is:

$$\text{for } y^{\mathrm{N}} \in \mathscr{Y}^{\mathrm{N}}: \quad \tilde{k}_{t,y^{\mathrm{N}}} = \frac{1}{\mathrm{S}_{y^{\mathrm{N}}}} \sum_{\hat{y}^{\mathrm{N}} \in \mathscr{Y}^{\mathrm{N}}} \mathrm{S}_{\hat{y}^{\mathrm{N}}} \Pi_{\hat{y}^{\mathrm{N}} y^{\mathrm{N}}} k_{t-1,\hat{y}^{\mathrm{N}}}, \tag{3.20}$$

$$\text{for } y^{\mathrm{N}} \in \mathscr{Y}^{\mathrm{N}}: \quad c_{t,y^{\mathrm{N}}} + k_{t,y^{\mathrm{N}}} \leq (1+r_t)\tilde{k}_{t,y^{\mathrm{N}}} + w_t y_{0,y^{\mathrm{N}}}, \tag{3.21}$$

$$\text{for } y^{\mathrm{N}} \notin \mathscr{C}: \quad \xi_{t,y^{\mathrm{N}}} u'(c_{t,y^{\mathrm{N}}}) = \beta \mathbb{E}_t \left[(1+r_{t+1}) \sum_{\tilde{y}^{\mathrm{N}} \in \mathscr{Y}^{\mathrm{N}}} \Pi_{y^{\mathrm{N}},\tilde{y}^{\mathrm{N}}} \xi_{t,\tilde{y}^{\mathrm{N}}} u'(c_{t+1,\tilde{y}^{\mathrm{N}}}) \right] \tag{3.22}$$

$$\text{for } y^{\mathrm{N}} \in \mathscr{C}: \quad k_{t,y^{\mathrm{N}}} = 0 \tag{3.23}$$

As one can see, all bins that are part of constrained group $\mathscr{C}$ remains so over the course of the stochastic simulation.

3.4.3 Global method: Parameterized expectations

A global solution using the parameterized expectations algorithm does not rely on assuming that the set of credit constrained histories remains the same as it can easily handle occasionally binding constraints. I use the of parameterized expectations algorithm popularized by den Haan & Marcet (1990) and Christiano & Fisher (2000). The basic idea of PEA is relatively straightforward. In any rational expectations equilibrium, the true conditional expectation of the Euler equation at time t is a function of the information available at time t, specified by the state variables of the system. Hence, it should be possible to approximate the conditional expectations with parameterized functions of the relevant state variables. One then iterates until the time-series generated is such that the assumed function is the best predictor of the conditional expectations in the model. This has several advantages, especially in the setting. First, it handles the course of dimensionality better than standard value function iteration for example. Second, it is much easier to solve for the policy functions of a large number of different bins by Monte Carlo methods than by traditional projection as the evaluation of the Euler Equation only needs to happen at the visited states.

In particular, the algorithm starts by parameterizing the expectations of the bins as follows:

$$\mathbb{E}_t \left[(1+r_{t+1}) \sum_{\tilde{y}^{\mathrm{N}} \in \mathscr{Y}^{\mathrm{N}}} \Pi_{y^{\mathrm{N}},\tilde{y}^{\mathrm{N}}} \xi_{t,\tilde{y}^{\mathrm{N}}} u'(c_{t+1,\tilde{y}^{\mathrm{N}}}) \right] \approx \Psi(\boldsymbol{\gamma}_{y^{\mathrm{N}}}; y^{\mathrm{N}}, \tilde{k}_{y^{\mathrm{N}},t}, \mathrm{K}_t, \mathrm{Z}_t) \tag{3.24}$$

Assuming that the bins are not borrowing constrained, it is easy to retrieve implied con-

sumption by inverting the Euler Equation:

$$c_{t,y^{\mathrm{N}}} = u^{-1}\left(\xi^{-1}_{t,\tilde{y}^{\mathrm{N}}}\beta\Psi(\boldsymbol{\gamma}_{y^{\mathrm{N}}}; y^{\mathrm{N}}, \tilde{k}_{y^{\mathrm{N}},t}, \mathrm{K}_t, \mathrm{Z}_t)\right) \tag{3.25}$$

and using the budget constraint one can retrieve implied savings:

$$k_{t,y^{\mathrm{N}}} = (1+r_t)\tilde{k}_{t,y^{\mathrm{N}}} + w_t y_{0,y^{\mathrm{N}}} - c_{t,y^{\mathrm{N}}} \tag{3.26}$$

If any of the savings are below the borrowing constraint one imposes the borrowing constraint and retrieves consumption as a residual. This way the algorithm can easily handle occasionally binding constraints. Furthermore, the global solution has a number of further advantages in certain settings, discussed in more detail below.

3.4.4 Discussion

As outlined above there are some critical differences between the local and global solution method. One critical advantage of the local method is the computation time; it takes only a few seconds to compute the solution to the equation system (3.21) - (3.23) whilst it may take several minutes or even hours to compute global solutions to an equivalent problem using parameterized expectations. There are however, also several advantages of the global method, especially when one extends the standard heterogeneous agent model with relevant extensions.

Cyclical borrowing constraints. The local method cannot handle cyclical borrowing constraints, that is borrowing constraints that loosen when total factor productivity increases. This again stems from the fact that perturbation methods have difficulties handling occasionally binding constraints, even more so if the constraint moves over the simulation. As seen from equation (3.23), the local method assumes that all borrowing constrained bins save 0, the borrowing constraint *at steady state*. But as the borrowing constraint moves over the cycle so should the savings of these constrained bins. This cannot be handled by standard local methods. The global method, however, is able to solve settings with cyclical borrowing constraints quite easily. The algorithm is almost identical to (3.24) - (3.26), but one imposes different borrowing constraints based on the current level of TFP.

Uncertainty. Another instance where standard first-order perturbation methods may struggle to compute reliable solution are environments with uncertainty shocks, i.e.

shocks to the second moment of the exogenous disturbances. Local methods need to rely on 2nd or 3rd order approximations or additional machinery such as the pruning toolbox developed by Andreasen et al. (2018). So even though it is technically feasible, the success of these methods, especially in a setting with reduced history heterogeneity, is unproven. On the other hand, global methods tend to deliver accurate solutions to settings with uncertainty shocks and are simple to account for, as they are just another state variable to incorporate in the parameterized expectation.

Multiple assets. Further, local perturbations around the steady state are impractical in a portfolio choice setting with multiple assets. This is due to the fact that the portfolios of the individual bins are indeterminate at the steady state. The effective returns of the two assets are equalized without aggregate risk, enabling a continuum of possible steady states. *A priori*, it is impossible to determine which one of these is the right steady state to approximate around. Conversely, a global method approximates policy functions not only around the steady state but across the entire state space. This enables the computation of portfolio choice problems as the effective returns and covariances are determined by the stochastic simulation.

3.4.5 Comparison of the methods

In this section, as a validation for the global method, I analyze the success of the global method in replicating the local method. As the perturbation technique yields a very good approximation around the steady state I would hope that the global method is able to replicate this (see e.g. Le Grand & Ragot (2022b) who compare the local method to other methods such as Reiter (2009) or Boppart et al. (2018)).

Calibration. The calibration of the model is very standard, following the RBC literature. As model is quarterly, I set the discount factor β to 0.98. The per-period utility function is $\log(c)$, the capital intensity α is set to 0.36 and the depreciation rate δ is equal to 0.025. Further, the TFP shock Z_t is assumed to follow an AR(1) process:

$$Z_t = (1-\rho_z) + \rho_z Z_{t-1} + \epsilon_{z,t} \quad \text{with } \epsilon_{z,t} \sim \mathcal{N}(0, \sigma_z^2) \tag{3.27}$$

where the persistence parameters ρ_z is set equal to 0.95 and σ_z is equal to 0.31%. I also parameterize the idiosyncratic risk process as an AR(1) process, with a persistence of $\rho_y = 0.99$ and a quarterly standard deviation σ_y of 10%. I discretize this using the

Rouwenhorst (1995) method with 5 idiosyncratic states. Finally, I choose N = 3 as the standard truncation length.

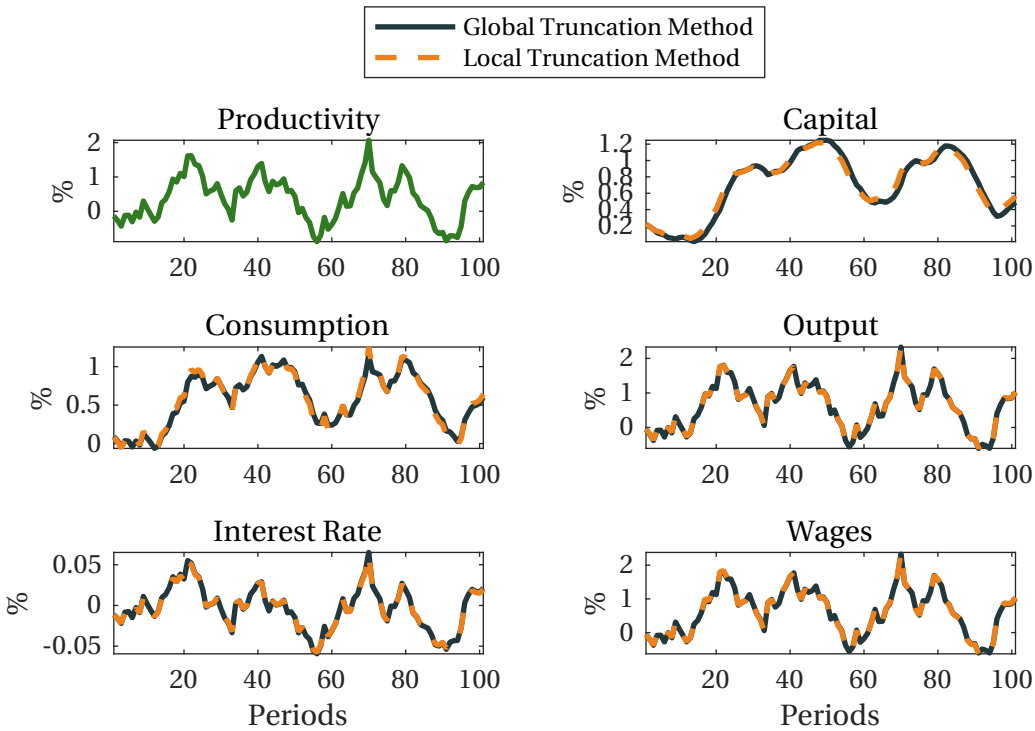


Fig. 3.1: Comparison of a stochastic simulation using the local solution (solid line) and the global solution (dashed line). Except for the interest rate, which is show in percentages changes, the variables are measured as percentage deviations of the deterministic steady state (local) and stochastic steady state (global), respectively.

Simulation. As a first check we compare the solutions of the local and global method along a simulated path of 100 periods of the economy. The results of this simulation are depicted in Figure 3.1. It is quite obvious from the picture that both methods are close to each other, as aggregate quantities are virtually indistinguishable over the simulation. This is reassuring, since the local truncation method has been shown to be close to other local method such as Reiter (2009) or Boppart et al. (2018) (see e.g. Le Grand & Ragot (2022b)).

Impulse Response Functions. This success of the global method in replicating the local solution is also confirmed by Figure 3.2 which depicts the impulse responses to an aggregate productivity shock in both the global and local solution method. All aggregate impulse responses are virtually indistinguishable from each other.

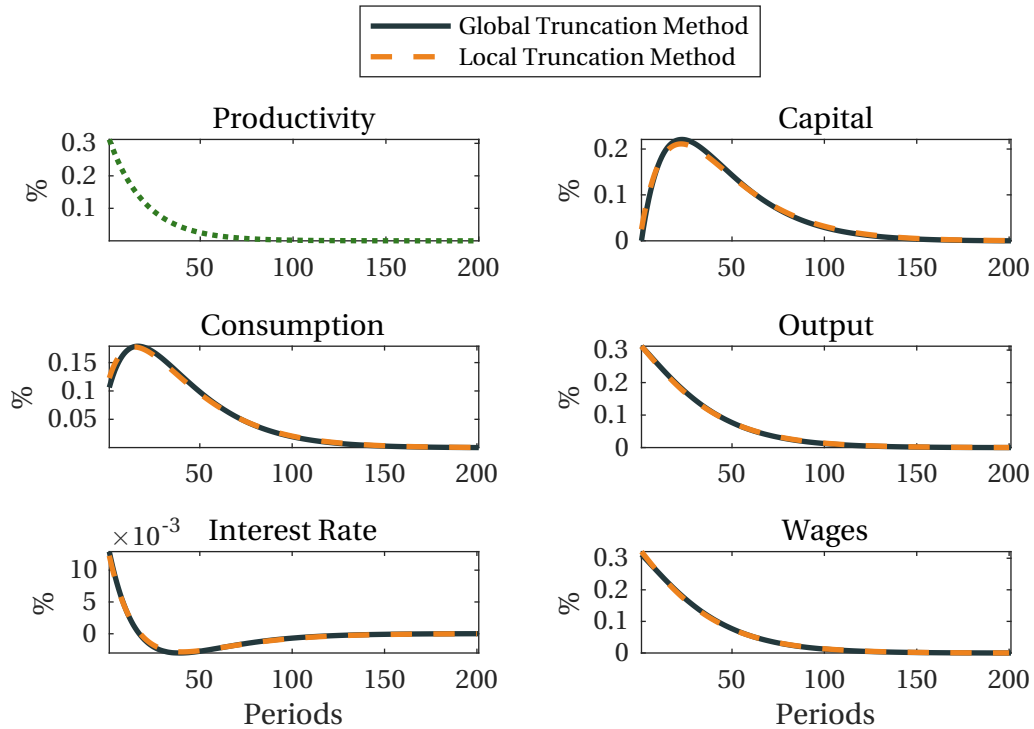


Fig. 3.2: Comparison of impulse response functions of aggregate variables using the global solution (solid line) and the local solution (dashed line) after a one standard deviation productivity shock. The impulse responses are measured as percentage deviations of the deterministic steady state (local) and stochastic steady state (global), respectively.

There are some important differences between the two methods, however, in particular when one disaggregates the results above. In Figure 3.3 I plot the impulse response of aggregate consumption, and of the top 10%, top 50% and bottom 10% of the wealth distribution. It is quite apparent that both solutions again deliver very similar results for the top, but the bottom end of the consumption distribution responds differently. The reason for this is the assumption of constant credit constrained history bins, even in the face of aggregate uncertainty. Unlike in the local method where all of these bins are assumed to be hand-to-mouth by assumption, the global method allows for occasionally binding constraints. This means that some bins will actually move out of the credit constraint and use the additional income for savings. This mediates the consumption response at the bottom quite significantly. It follows that both methods are highly similar in aggregate quantities. But if one is interested in measuring the response of the bottom of the distribution one should use the global method as it can reflect transitions out and into credit constraints. This may matter significantly for policy analysis of the constrained consumption response to say a government spending shock or monetary policy shock,

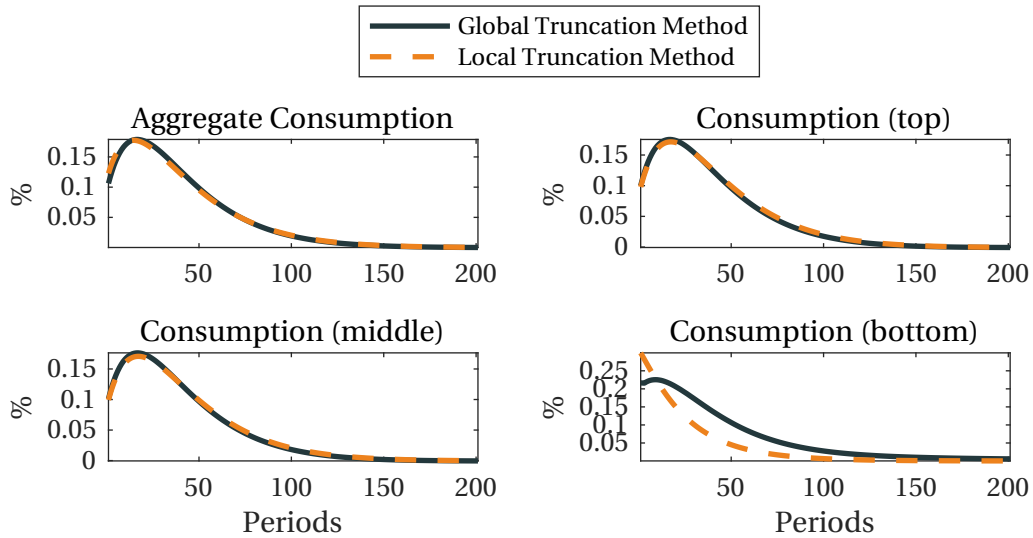


Fig. 3.3: Comparison of impulse response functions of consumption using the global solution (solid line) and the local solution (dashed line) after a one standard deviation productivity shock. Except for the interest rate which is show in percentages changes, the impulse responses are measured as percentage deviations of the deterministic steady state (local) and stochastic steady state (global), respectively.

and may have significant aggregate implications if these constrained agents are relatively wealthy (see e.g. Kaplan et al. (2018); Kaplan & Violante (2014); Kaplan et al. (2014)).

3.5 Three applications

As already discussed in Subsection 3.4.4 there are multiple settings where the global solution algorithm can be used more effectively than the local one. In this section I analyze three examples numerically. First, I consider a procyclical borrowing constraint, second an addition of an uncertainty shock to the standard setting and introducing a second safe asset but non-productive asset in the final exercise.

3.5.1 Cyclical borrowing constraints

The effects of procyclical financial conditions in heterogeneous agent economies have been analyzed by Guerrieri & Lorenzoni (2017), for example, to explain the 2008 credit crunch. In order to introduce cyclical borrowing constraints in the truncated history environment I copy the exact calibration from Subsection 3.4.5 and only modify the borrowing constraint by making it a function of the business cycle. In particular, I assume that the agents cannot borrow in a downturn but that the borrowing constraint loosens as

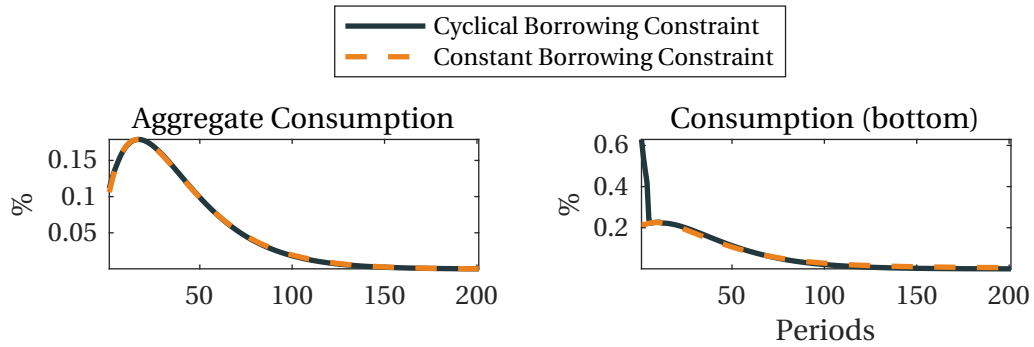


Fig. 3.4: Comparison of impulse response functions of consumption with (dashed) and without (solid) cyclical borrowing constraints after a one standard deviation productivity shock. The impulse responses are measured as percentage deviations from the ergodic mean.

a linear function of the boom. Formally, the borrowing constraint $\underline{k}_t$ evolves as follows:

$$\underline{k}_t = \mathbb{1}_{Z_t>1}\,(100(1-Z_t)) \tag{3.28}$$

This effectively means in our calibration that the borrowing constraint only loosens in booms, has an unconditional mean of -0.4113 and a minimum of around -3.

As the results for the aggregate quantities are virtually the same to Figure 3.2, we focus again on the disaggregation of consumption here. Figure 3.4 depicts the consumption responses after a one standard deviation productivity shocks in the case with cyclical borrowing constraints. As expected, it is the bottom of the distribution that exhibits a much higher elasticity to technology shocks with cyclical borrowing constraints compared to the one without. The bottom 10% raise their consumption by around 0.6% after a shock of 0.3% in our calibration. Finally, note that aggregate consumption does not differ noticeably; the reason is that the bottom 10% only account for a small fraction of total consumption in our calibration so that an increase in their consumption hardly shows up in the aggregate.

3.5.2 Uncertainty shocks

In this subsection I analyze the effect of an uncertainty shock in a heterogeneous agent economy. This means that the process for the exogenous innovations need to be updated. In particular, I follow Fernández-Villaverde & Guerrón-Quintana (2020) and assume that Z_t follows an AR(1) process as before with time varying volatility $\sigma_{Z,t}$ which also follows

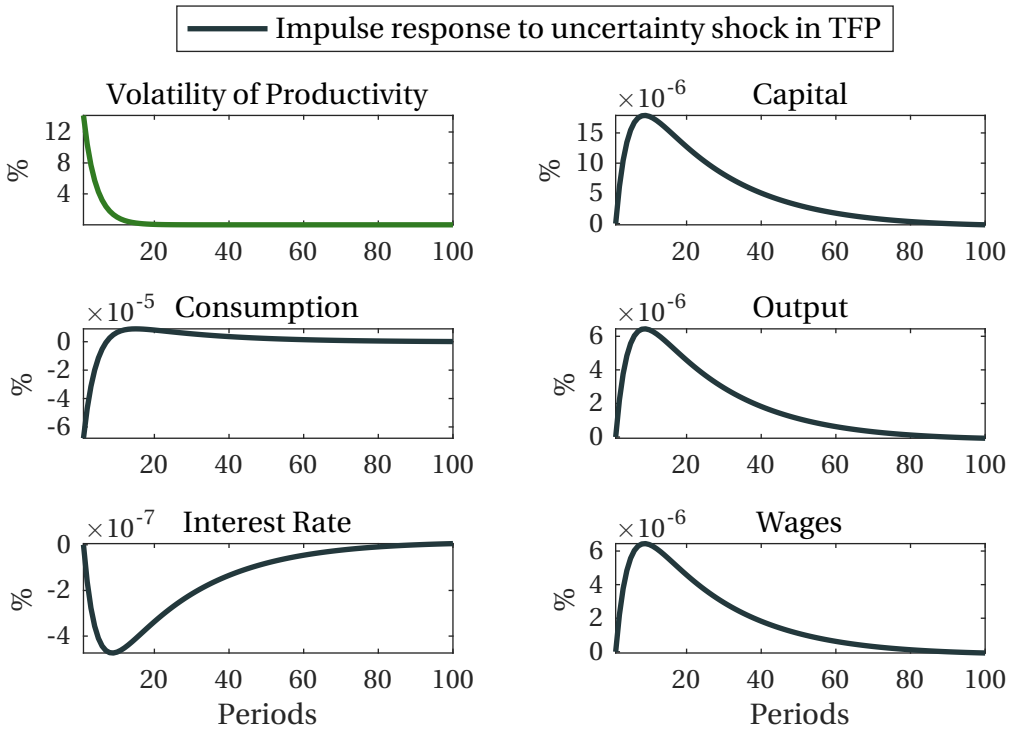


Fig. 3.5: Impulse response functions of aggregate variables after an uncertainty shock at the stochastic steady state. The impulse responses are measured as percentage deviations from the ergodic mean.

an AR(1) process. Formally, we have:

$$Z_t = (1-\rho_Z) + \rho_Z Z_{t-1} + e^{\sigma_{Z,t}} \epsilon_{Z,t} \quad \text{where } \epsilon_{Z,t} \sim \mathcal{N}(0,1) \tag{3.29}$$

and

$$\sigma_{Z,t} = (1-\rho_\sigma)\sigma_Z + \rho_\sigma \sigma_{Z,t-1} + (1-\rho_\sigma^2)^{1/2} \nu_Z u_{Z,t} \quad \text{where } u_{Z,t} \sim \mathcal{N}(0,1) \tag{3.30}$$

where I set the persistence of the technology process to $\rho_Z = 0.95$, the average standard deviation of σ_Z to $\sigma_Z = \log 0.005$, the persistence of the uncertainty shock $\rho_{\sigma_Z} = 0.75$ and ν_Z equal to 0.2. Figure 3.5 plots the impulse response to an uncertainty shock where the standard deviation increases by 12% on impact. As can be discerned from the panel, the responses are small, but expansionary. This is due to the precautionary savings motive, as agents raise savings and thus create a supply side boom. This is counterfactual to evidence by Bloom (2009), for example, who show that uncertainty produces drops in output. I provide one possible remedy to this counterfactual result.

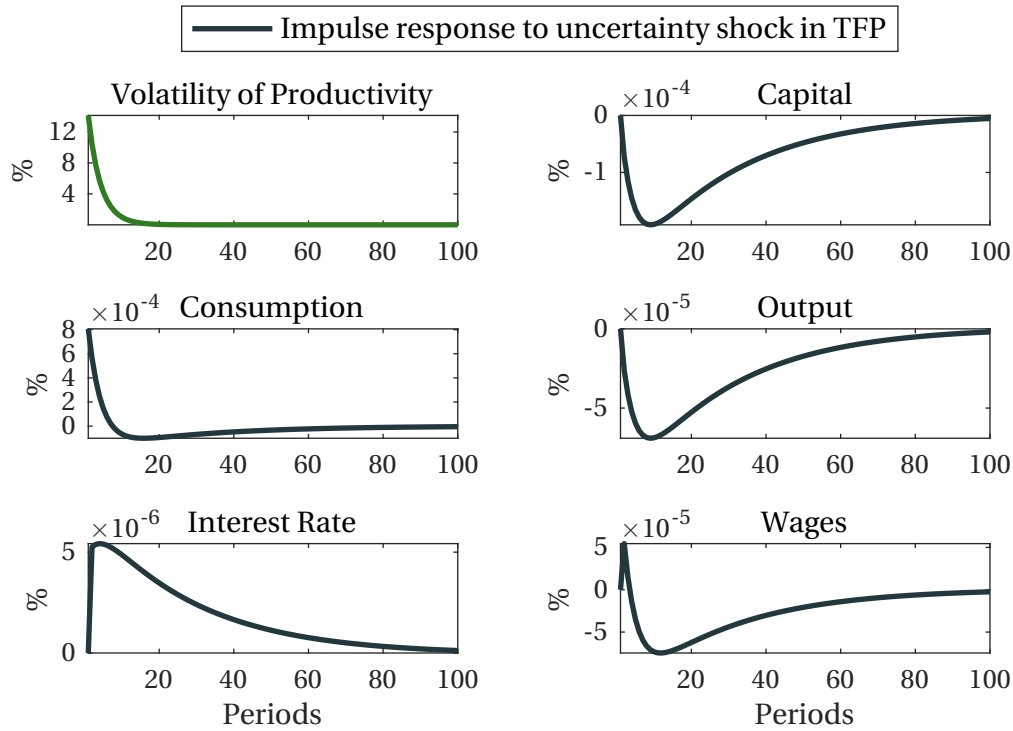


Fig. 3.6: Impulse response functions of aggregate variables after an uncertainty shock with aggregate demand externality at the stochastic steady state. The impulse responses are measured as percentage deviations from the stochastic steady state

An aggregate demand externality. In order to address the counterfactual increase in output after an uncertainty increase, I introduce an aggregate demand externality to TFP. I follow Krueger et al. (2016) an model TFP as a product of the original "pure" TFP shock and a non-linear function of past period's aggregate consumption:

$$\tilde{Z}_t = Z_t C_{t-1}^{\omega} \tag{3.31}$$

This means that, holding the pure TFP process constant, TFP falls today if consumption was lower yesterday and vice versa. Microfoundations for this assumption can be found in Huo & Rios-Rull (2015) and Kaplan & Menzio (2016), among others. I set $\omega = 0.1$, around the mid-value of parameters that Krueger et al. (2016) provide. The impulse responses after an uncertainty shock with an aggregate demand externality are shown in Figure 3.6. All impulse responses have the opposite sign compared to Figure 3.5, and, indeed, the capital and output responses are negative throughout for all horizons. The reasoning is as follows. When the uncertainty shock hits the economy, agents realise that future productivity innovations are more dispersed. Moreover, this dispersion is skewed to the

downside due to the non-linearity of the externality, as a lower exogenous TFP component will reduce composite TFP by more than a higher exogenous TFP component will raise the composite term, due to procyclical aggregate consumption. Hence agents perceive capital as more risky, especially on the downside, and reduce savings on impact. As total income on impact has not changed, consumption must increase as an accounting identity. This explains the immediate rise in consumption and drop in capital. As interest rates rise and wages fall over time it reduces the incentive to consume and so consumption falls until the economy reverts back to the stochastic steady state. Finally, note again, that the responses are rather small, something common in the uncertainty literature. Introducing frictions, such as in Andreasen et al. (2018) or Basu & Bundick (2017), may ameliorate this weak amplifiation of uncertainty shocks.

3.5.3 Two assets

Another interesting extension to the standard model that is only feasible with the global solution technique is a meaningful introduction of multiple assets. Incorporating a portfolio choice mechanism into a heterogeneous agent model is of great interest for monetary policy transmission but can also explain long-run trends in inequality. In order to demonstrate that the global method can handle this, I introduce a second asset in fixed supply $\bar{B}$ to the economy. In particular, I assume that history bins can now save both in capital k_{t,y^N} which has a return of $(1 + r_{t+1})$ and bonds b_{t,y^N} at price q_t today that has a return of 1 tomorrow. Thus, the bond is safe, as the return is known today. Further, I assume that there is a small quadratic adjustment cost associated with buying bonds, $\pi = 0.01$. Thus, the new budget constraint is:

$$c_{t,y^N} + k_{t,y^N} + qb_{t,y^N} + \frac{\pi}{2} b^2_{t,y^N} \leq (1 + r_t)\tilde{k}_{t,y^N} + \tilde{b}_{t,y^N} + w_t y_{0,y^N}, \tag{3.32}$$

and the additional market clearing condition is

$$\sum_{y^N \in \mathscr{Y}^N} b_{t,y^N} = \bar{B} \tag{3.33}$$

I calibrate the total bond supply $\bar{B} = 4$ in order to target the U.S. debt to GDP ratio of 1.2.

The solution to a heterogeneous agent model with bonds is in need of one additional step. Unlike the rental rate of capital r_t, pinned down by the firm's first order condition, the market clearing price of the bond is not uniquely determined by any supply condition. This is no problem in a complete markets model, since the Euler equations hold with

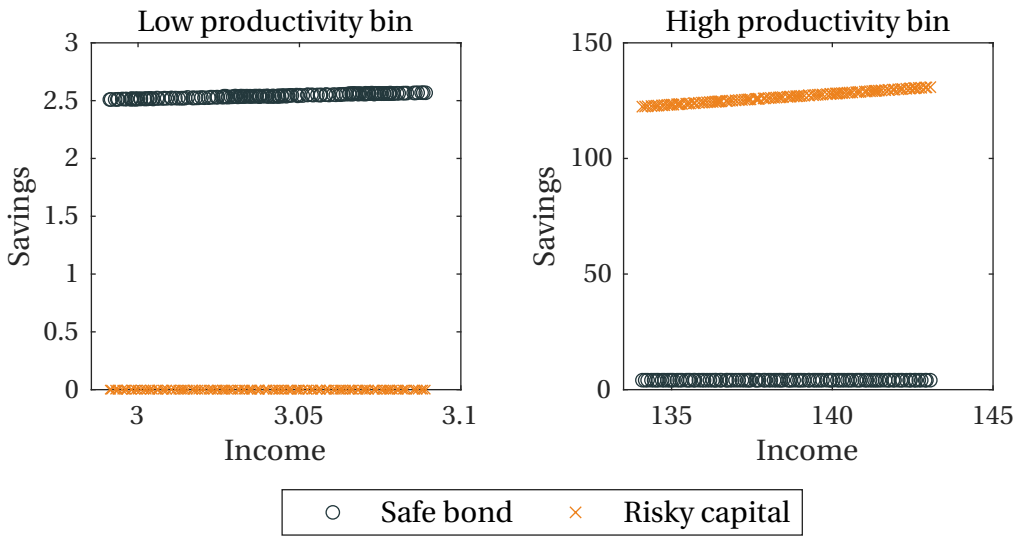


Fig. 3.7: Portfolio decision rules for exemplary low and high productivity history bins

equality at every point in the state space, uniquely pinning down the market clearing price. However, in a heterogeneous agents model every individual has her own idiosyncratic stochastic discount factor. This means that the market clearing price cannot be discerned from the Euler Equation anymore. Thus, I follow Krusell & Smith (1997) and parameterize the price function in order to find the market clearing price.[3] More precisely, I assume the following law of motion for the bond price:

$$q_t(z_t, \mathrm{K}_t; \boldsymbol{\vartheta}) = \vartheta_0 + \vartheta_1 \log(z_t) + \vartheta_2 \log(\mathrm{K}_{t-1}) \tag{3.34}$$

This means that the bond market clearing price is a function of the technology shock and preinstalled capital stock, the two aggregate state variables here. The exact algorithm to solving the entire environment is described in more detail in appendix C.1, but a general outline is as follows. I guess a vector $\boldsymbol{\vartheta}$ of price parameters and find the optimal policy functions of the agents based on this guess. I then simulate the economy for T periods, using a non-linear solver to find the market clearing price based on these policy functions. Next, I regress this new series of market-clearing prices on the series of shocks and capital from that simulation, updating the price parameters using homotopy. I repeat this procedure until the regression error is below some predefined tolerance level. The portfolio allocation results for two history bins are shown in Figure 3.7. These are the

[3]Note that there are other ways of arriving at the market clearing price, too. For example, Den Haan & Rendahl (2010) propose the use of explicit aggregation, parameterizing the sum of the individual policy function and the price. Integrating these across all agents should give the market clearing price since total bond supply equals zero.

bins with a history of three of the lowest productivity shocks and the one with three of the highest productivity shocks. The figure plots the portfolio decision rules as scatter plots as a function of income of the bins. The solution of the model shows that the low productivity bins do not invest in risky capital as it is correlated with wage risk. This is consistent with the empirical evidence by Fagereng et al. (2020), for example. On the other hand, safe asset demand is an increasing function of income. The highly productive bin exhibits a mixed portfolio, as shown in the right panel of 3.7, with a heavy bias towards capital. This is in line with the idea that an investor, who has the means to do so, will diversify his investment but is more tolerant of risk. This is also confirmed by the fact that demand for the risky asset is more varied, backing the theory that the latter is held for excess returns whilst the former is an insurance policy.

The importance of these investor clienteles go beyond simple portfolio predictions and extend into optimal fiscal policy, for example. Consider a social planner who is able to issue safe debt and tax capital and labour returns. That planner faces a critical tradeoff between issuing safe public debt and taxing productive capital. The former eases the self-insurance motive of poorer agents whilst a decrease in capital tax will increase the utility of those agents that save in capital and raise total capital stock. This is even more pronounced over the business cycle, something that has not been studied so far. However, an analysis of this kind is possible with the current solution mechanism, as it can handle optimal policy problems with heteroegneous agents, portfolios and aggregate risk. This is left for future research.

3.6 Conclusion

This chapter presented a global solution method for reduced history representations of heterogeneous agent economies. The algorithm relies on parameterized expectations, a simple and intuitive way of solving for global policy functions. This method is feasible and correct as it can replicate the local results in the aggregate. Further, I have shown that there are numerous settings where this method can deliver interesting and economically meaningful answers in settings where the local method would fail. I provided three examples; a setting with a cyclical borrowing constraint, raising the marginal propensity to consume for borrowing constrained. Second, a setting where uncertainty shocks can be either recessionary or expansionary, depending on whether aggregate demand can influence TFP. And finally, I have applied the method to a two asset economy where different investor clienteles can give rise to unanswered optimal policy questions. The

latter is the most likely direction for future research since the global solution method, together with reduced history representations, is indeed perfectly suited for optimal policy problems with heterogeneous portfolios and aggregate risk.

References

Achdou, Y., Han, J., Lasry, J.-M., Lionse, P.-L., & Moll, B. (2022). Income and Wealth Distribution in Macroeconomics: A Continuous-Time Approach [On the Existence and Uniqueness of Stationary Equilibrium in Bewley Economies with Production]. *Review of Economic Studies*, 89(1), 45–86.

Aiyagari, S. R. (1994). Uninsured Idiosyncratic Risk and Aggregate Saving. *The Quarterly Journal of Economics*, 109(3), 659–684.

Algan, Y., Allais, O., & Den Haan, W. J. (2008). Solving heterogeneous-agent models with parameterized cross-sectional distributions. *Journal of Economic Dynamics and Control*, 32(3), 875–908.

Andreasen, M. M., Fernández-Villaverde, J., & Rubio-Ramírez, J. F. (2018). The Pruned State-Space System for Non-Linear DSGE Models: Theory and Empirical Applications. *Review of Economic Studies*, 85(1), 1–49.

Auclert, A., Bardóczy, B., Rognlie, M., & Straub, L. (2021). Using the Sequence-Space Jacobian to Solve and Estimate Heterogeneous-Agent Models. *Econometrica*, 89(5), 2375–2408.

Azinovic, M., Gaegauf, L., & Scheidegger, S. (2022). Deep equilibrium nets. *International Economic Review*, n/a(n/a).

Basu, S. & Bundick, B. (2017). Uncertainty shocks in a model of effective demand. *Econometrica*, 85(3), 937–958.

Bewley, T. (1983). A difficulty with the optimum quantity of money. *Econometrica*, 51(5), 1485–504.

Bilbiie, F. (2019). *Monetary Policy and Heterogeneity: An Analytical Framework*. 2019 Meeting Papers 178, Society for Economic Dynamics.

Bilbiie, F. O. & Ragot, X. (2021). Optimal monetary policy and liquidity with heterogeneous households. *Review of Economic Dynamics*, 41, 71–95. Special Issue in Memory of Alejandro Justiniano.

Bloom, N. (2009). The impact of uncertainty shocks. *Econometrica*, 77(3), 623–685.

Boppart, T., Krusell, P., & Mitman, K. (2018). Exploiting MIT shocks in heterogeneous-agent economies: the impulse response as a numerical derivative. *Journal of Economic Dynamics and Control*, 89(C), 68–92.

Challe, E. & Ragot, X. (2016). *Precautionary Saving Over the Business Cycle*. Sciences Po publications info:hdl:2441/54k1p7hdq38, Sciences Po.

Christiano, L. J. & Fisher, J. D. M. (2000). Algorithms for solving dynamic models with occasionally binding constraints. *Journal of Economic Dynamics and Control*, 24(8), 1179–1232.

den Haan, W. J. & Marcet, A. (1990). Solving the stochastic growth model by parameterizing expectations. *Journal of Business & Economic Statistics*, 8(1), 31–34.

Den Haan, W. J. & Rendahl, P. (2010). Solving the incomplete markets model with aggregate uncertainty using explicit aggregation. *Journal of Economic Dynamics and Control*, 34(1), 69–78.

Fagereng, A., Guiso, L., Malacrino, D., & Pistaferri, L. (2020). Heterogeneity and persistence in returns to wealth. *Econometrica*, 88(1), 115–170.

Fernández-Villaverde, J. & Guerrón-Quintana, P. A. (2020). Uncertainty shocks and business cycle research. *Review of Economic Dynamics*, 37, S118–S146. The twenty-fifth anniversary of “Frontiers of Business Cycle Research".

Fernández-Villaverde, J., Hurtado, S., & Nuño, G. (2019). *Financial Frictions and the Wealth Distribution*. Working Paper 26302, National Bureau of Economic Research.

Guerrieri, L. & Iacoviello, M. (2015). OccBin: A toolkit for solving dynamic models with occasionally binding constraints easily. *Journal of Monetary Economics*, 70(C), 22–38.

Guerrieri, V. & Lorenzoni, G. (2017). Credit Crises, Precautionary Savings, and the Liquidity Trap*. *The Quarterly Journal of Economics*, 132(3), 1427–1467.

Huggett, M. (1993). The risk-free rate in heterogeneous-agent incomplete-insurance economies. *Journal of Economic Dynamics and Control*, 17(5-6), 953–969.

Huo, Z. & Rios-Rull, J.-V. (2015). Tightening Financial Frictions on Households, Recessions, and Price Reallocations. *Review of Economic Dynamics*, 18(1), 118–139.

İmrohoroğlu, S. (1998). A Quantitative Analysis of Capital Income Taxation. *International Economic Review*, 39(2), 307–328.

Kaplan, G. & Menzio, G. (2016). Shopping externalities and self-fulfilling unemployment fluctuations. *Journal of Political Economy*, 124(3), 771 – 825.

Kaplan, G., Moll, B., & Violante, G. L. (2018). Monetary Policy According to HANK. *American Economic Review*, 108(3), 697–743.

Kaplan, G. & Violante, G. L. (2014). A model of the consumption response to fiscal stimulus payments. *Econometrica*, 82(4), 1199–1239.

Kaplan, G., Violante, G. L., & Weidner, J. (2014). The Wealthy Hand-to-Mouth. *Brookings Papers on Economic Activity*, 45(1 (Spring), 77–153.

Krueger, D., Mitman, K., & Perri, F. (2016). Macroeconomics and Household Heterogeneity. In J. B. Taylor & H. Uhlig (Eds.), *Handbook of Macroeconomics*, volume 2 of *Handbook of Macroeconomics* chapter 0, (pp. 843–921). Elsevier.

Krusell, P. & Smith, A. A. (1997). Income And Wealth Heterogeneity, Portfolio Choice, And Equilibrium Asset Returns. *Macroeconomic Dynamics*, 1(02), 387–422.

Krusell, P. & Smith, Jr., A. A. (1998). Income and Wealth Heterogeneity in the Macroeconomy. *Journal of Political Economy*, 106(5), 867–896.

Le Grand, F. & Ragot, X. (2022a). Managing inequality over business cycles: Optimal policies with heterogeneous agents and aggregate shocks. *International Economic Review*, 63(1), 511–540.

Le Grand, F. & Ragot, X. (2022b). Refining the truncation method to solve heterogeneous-agent models. *Annals of Economics and Statistics*, (146), 65–92.

Maliar, L., Maliar, S., & Winant, P. (2021). Deep learning for solving dynamic economic models. *Journal of Monetary Economics*, 122, 76–101.

Reiter, M. (2009). Solving heterogeneous-agent models by projection and perturbation. *Journal of Economic Dynamics and Control*, 33(3), 649–665.

Rouwenhorst, K. G. (1995). Asset pricing implications of equilibrium business cycle models. In T. F. Cooley (Ed.), *Frontiers of Business Cycle Research* (pp. 294–330). Princeton: Princeton University Press.

Young, E. R. (2010). Solving the incomplete markets model with aggregate uncertainty using the krusell–smith algorithm and non-stochastic simulations. *Journal of Economic Dynamics and Control*, 34(1), 36–41. Computational Suite of Models with Heterogeneous Agents: Incomplete Markets and Aggregate Uncertainty.

A

Appendix for Chapter 1

A.1 Introduction

In this appendix we describe the estimation technique, perform a thorough sensitivity analysis for our results in the main text of the paper and describe further details of the Hansen (1992) test. First, we describe the estimation technique of smooth local projections by Barnichon & Brownlees (2019) in section A.2. Second, we present further robustness checks specific to our test of the size effect of monetary shocks (equation (1.1) in the main text) in section A.3. Third, we present robustness checks specific to our test of regime dependency of monetary shocks (equation (1.2) in the main text) in section A.4. Fourth, we present a number of robustness checks on the results regarding both tests in section A.5. We also provide some additional figures displaying the shock distribution and results regarding their regime dependency in section A.6. Finally, we discuss our application of the Hansen (1992) coefficient constancy test in more detail in section A.7. All the figures are collected in section A.8 and all the tables are collected in section A.9. The numbers of equations and sections purely in Arabic, with no letters, refer to the equations and the sections in the main text of the paper.

A.2 Estimation technique

In this section we describe the methodology used for our estimation. We follow Barnichon & Brownlees (2019) very closely and advise the reader to consult their work in case some aspects remain unclear.

Smooth local projections. Below we will set out the estimation technique, using the local projection (1.1) as our reference point. Note, however, that the same technique applies to all other local projections too, including (1.2) and the two coefficients of interest β_h^{HI} and β_h^{LO}.
Recall our main equation for the first test (1.1), repeated below for convenience:

$$y_{t+h} = \alpha_h + \tau_h t + \beta_h e_t + \zeta_h (e_t \cdot |e_t|) + \sum_{k=1}^{\mathrm{K}} \gamma_{h,k} w_{t,k} + v_{t+h}, \tag{A.1}$$

We can approximate both the β_h and the ζ_h coefficient using B-spline basis function expansions as follows:

$$\beta_h \approx \sum_{j=1}^{\mathrm{J}} b_j \mathrm{B}_j(h) \tag{A.2}$$

$$\zeta_h \approx \sum_{j=1}^{\mathrm{J}} z_j \mathrm{B}_j(h) \tag{A.3}$$

where $\mathrm{B}_j : \mathbb{R} \to \mathbb{R}$ for $j = 1, \ldots, \mathrm{J}$ is a set of B-spline basis function and b_j and z_j for $j = 1, \ldots, \mathrm{J}$ are a set of scalar parameters. With this we can write (A.1) as:

$$y_{t+h} = \alpha_h + \tau_h t + \sum_{j=1}^{\mathrm{J}} b_j \mathrm{B}_j(h) e_t + \sum_{j=1}^{\mathrm{J}} z_j \mathrm{B}_j(h)(e_t \cdot |e_t|) + \sum_{k=1}^{\mathrm{K}} \gamma_{h,k} w_{t,k} + v_{t+h}, \tag{A.4}$$

One can represent this in linear regression form. Let $\mathscr{Y}_t$ for $t = 1, \ldots, \mathrm{T}$ be defined as the vector $(y_{\min(t,\mathrm{T}-\mathrm{H})}, \ldots, y_{\min(\mathrm{T},t+\mathrm{H})})'$ with size H. Let $\mathscr{X}_{\beta,t}$ and $\mathscr{X}_{\zeta,t}$ be defined as $\mathrm{H} \times \mathrm{J}$ matrices where the (h, j)th element is $\mathrm{B}_j(h) e_t$ and $\mathrm{B}_j(h)(e_t \cdot |e_t|)$, respectively. Define $\mathscr{X}_{\alpha,t}$ as a diagonal $\mathrm{H} \times \mathrm{H}$ matrix with 1s on its main diagonal. Define $\mathscr{X}_{\tau,t}$ also as a diagonal $\mathrm{H} \times \mathrm{H}$ matrix with t on its main diagonal. Finally, define $\mathscr{X}_{\gamma,t}$ also as a diagonal $\mathrm{H} \times \mathrm{H}$ matrix with $w_{t,k}$ on its main diagonal for $k = 1, \ldots, \mathrm{K}$. Stacking these horizontally we obtain $\mathscr{X}_t = (\mathscr{X}_{\alpha,t}, \mathscr{X}_{\tau,t}, \mathscr{X}_{\beta,t}, \mathscr{X}_{\zeta,t}, \mathscr{X}_{\gamma,t})$. We can now write equation (A.4) as

$$\mathscr{Y}_t = \mathscr{X}_t \theta + \mathscr{U}_t \tag{A.5}$$

where $\mathscr{U}_t$ denotes the $\mathrm{H} \times 1$ prediction error vector term. Finally, one can stack these for $t = 1, \ldots, \mathrm{T}$ to obtain $\mathscr{Y}, \mathscr{X}$ and $\mathscr{U}$. The smooth local projections can then be estimated by

generalized ridge estimation:

$$\hat{\theta} = \arg\min_{\theta}\{||\mathscr{Y} - \mathscr{X}\theta||^2 + \lambda\theta'\mathbf{P}\theta\} \tag{A.6}$$

$$= (\mathscr{X}'\mathscr{X} + \lambda\mathbf{P})^{-1}\mathscr{X}'\mathscr{Y}, \tag{A.7}$$

where λ is a positive shrinkage parameter and $\mathbf{P}$ is a symmetric positive semidefinite penalty matrix. The value of λ determines the bias/variance trade-off in the estimation. The estimation collapses to the least squares estimator when $\lambda = 0$ whilst the estimator is biased with a large value of λ but may have a smaller variance.

Penalty matrix. We use a penalty matrix $\mathbf{P} = \mathbf{D}_r'\mathbf{D}_r$ where $\mathbf{D}_r'$ is the matrix representation of Δ^r, the r-th difference operator. This lets the estimated impulse response - with a very large λ - shrink to the $r-1$ polynomial. We set $r = 3$ so that the limit polynomial is quadratic.

Shrinkage parameter. We select the optimal λ using k-fold cross validation (Racine, 1997). We set $k = 0$ so that the cross validation function becomes:

$$\mathrm{CV} = \frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}} \frac{\hat{\mathscr{U}}_t^2}{(1-h_{tt})^2} \tag{A.8}$$

where h_{tt} denotes the tth diagonal element of the projection matrix $\mathscr{X}_t(\mathscr{X}_t'\mathscr{X}_t + \lambda\mathbf{P})^{-1}\mathscr{X}_t'$.

Inference. Further, we also follow Barnichon & Brownlees (2019) in conducting inference. The Newey & West (1987) variance of $\hat{\theta}$ is estimated as follows:

$$\begin{aligned}\widehat{\mathbf{V}}(\hat{\theta}) =& \mathrm{T}\left[\sum_{t=1}^{\mathrm{T}} \mathscr{X}_t'\mathscr{X}_t + \lambda\mathbf{P}\right]^{-1}\left[\hat{\Gamma}_0 + \sum_{l=1}^{\mathrm{L}} w_l\left(\hat{\Gamma}_l + \hat{\Gamma}_l'\right)\right] \\ &\times\left[\sum_{t=1}^{\mathrm{T}} \mathscr{X}_t'\mathscr{X}_t + \lambda\mathbf{P}\right]^{-1}\end{aligned} \tag{A.9}$$

where $w_l = 1 - l/(1+\mathrm{L})$ and $\hat{\Gamma}_l = \frac{1}{\mathrm{T}}\sum_{t=l+1}^{\mathrm{T}} \mathscr{X}_t'\hat{\mathscr{U}}_t\hat{\mathscr{U}}_{t-l}'\,\mathscr{X}_{t-l}$ with $\hat{\mathscr{U}}_t$ denoting regression residuals. We set L to H and λ to 0.5 times the degree of shrinkage determined by k -fold cross-validation. We construct the standard errors for coefficient β_h as $\sqrt{\mathbf{B}(h)'\widehat{\mathbf{V}}(\hat{b})\mathbf{B}(h)}$ where $\mathbf{B}(h) = \left(\mathrm{B}_1(h), \ldots, \mathrm{B}_{\mathrm{J}}(h)\right)'$, and $\hat{b}$ and $\widehat{\mathbf{V}}(\hat{b})$ denote the subvector and submatrix of, respectively, $\hat{\theta}$ and $\widehat{\mathbf{V}}(\hat{\theta})$ relative to the b parameter.

Consequently, the $1-p$ confidence interval for β_h is $\mathbf{B}(h)'\hat{b} \pm z_{1-p/2}\sqrt{\mathbf{B}(h)'\widehat{\mathbf{V}}(\hat{b})\mathbf{B}(h)}$, where $z_{1-p/2}$ denotes the $1-p/2$ quantile of a standard normal. And the t-statistic can be computed as $\frac{\beta_h}{\sqrt{\mathbf{B}(h)'\widehat{\mathbf{V}}(\hat{b})\mathbf{B}(h)}}$. An identical procedure applies to the coefficient ζ_h.

Delta method. When constructing functions $g(\hat{\theta})$ that depend on our estimated coefficients we use the Delta method to conduct inference. The variance of these functions can be approximated as:

$$\hat{\mathrm{V}}(g(\hat{\theta})) \approx \nabla g(\hat{\theta})'\widehat{\mathbf{V}}(\hat{\theta})\nabla g(\hat{\theta}) \tag{A.10}$$

where $\nabla g(\hat{\theta})$ is the Jacobian matrix of $g(\hat{\theta})$ with respect to $\hat{\theta}$.

Non-smoothed and cumulative responses. When estimating non-smoothed responses we set $\lambda = 0$ to obtain the least squares estimates. We also construct cumulative responses from this, by cumulating the non-smoothed response over its projection horizon. The variance matrix for the cumulative estimates can then be obtained with the Delta method, as set out above.

A.3 Robustness: Size-dependent impulse response

In this section we perform a number of robustness checks with respect to equation (1.1), the test for a size-dependent impulse response. In A.3.1 we conduct a visual stability assessment of recursive estimates of the absolute value interaction term. In A.3.2 we include the squared and cubed value of the shock instead of the absolute value interaction term. In A.3.3 we show the impulse responses to a 25bp and a 200bp shock.

A.3.1 Recursive estimates

We check the stability of the absolute value interaction coefficients by recursively estimating equation (1.1), adding one month for every new estimation. We start with a third of our sample, corresponding to the 140 months between January 1969 to August 1980. Figures A.1, A.2 and A.3 plot the results for the absolute value interaction term coefficient with respect to PCE inflation, industrial output and the federal funds rate, respectively. The x-axis indicates the horizon of the impulse response, the y-axis indicates the last period in the sample for the recursive estimates and the z-axis indicates the coefficient

size. The recursive coefficient sequences are relatively stable. Take, for example, the inflation coefficients with respect to the absolute value interaction of the monetary shock. Qualitatively, the negative and then positive dynamics of this coefficient are certainly constant over the recursive estimation. Furthermore, its size is also relatively similar across different samples, apart from some fluctuations for the long horizons at the beginning of the sample until the mid 1980s.

A.3.2 Alternative specification with quadratic and cubic terms

Here we consider a different specification for the non-linear local projection to investigate non-linearities in the impulse response function with respect to the size of the shock:

$$y_{t+h} = \alpha_h + \tau_h t + \beta_h e_t + \vartheta_h e_t^2 + \psi_h e_t^3 + \sum_{k=1}^{K} \gamma_{h,k} w_{t,k} + v_{t+h}. \tag{A.11}$$

The inclusion of a squared and a cubed shock value accounts for non-linearities in the impulse response function in a different way. The coefficient with respect to squared shocks, ϑ_h, captures possible asymmetries of the impulse response functions with respect to positive and negative shocks. In general, a ϑ_h with a sign equal to that of β_h amplifies the linear coefficient of the impulse response with respect to positive shocks. Conversely, a ϑ_h with an opposite sign to β_h counteracts the linear coefficient of the impulse response with respect to positive shocks.

The coefficient with respect to cubed shocks, ψ_h, captures possible non-linearities with respect to the size of the shock, and is the main coefficient of interest in this specification. In general, a ψ_h with a sign equal to the one of β_h amplifies the latter coefficient after larger shocks. On the contrary, a ψ_h with an opposite sign to the one of β_h counteracts the latter coefficient after larger shocks. We prefer our benchmark specification (1.1) to (A.11) because a larger shock affects the shape of the impulse response also via the squared term and may thus obfuscate interpretation for our test. Hence, the interaction term in (1.1) should capture and illustrate the size effect more neatly than the cubed shock term in (A.11).

The estimated coefficients are reported in Figure A.4. As previously, the three rows correspond to the three response variables, annualized PCE inflation (row 1), industrial production (row 2) and the federal funds rate (row 3). The first column combines all three coefficients of the non-linear impulse response, $\hat{\beta}_h$ (black, solid line), $\hat{\vartheta}_h$ (red, dashed line) and $\hat{\psi}_h$ (green, dashed-dotted line) into one graph. The second column shows the coeffi-

cient with respect to the squared shock value ($\hat{\vartheta}_h$) again, together with its 90% confidence interval. The third column depicts the coefficient with respect to the cubed shock value ($\hat{\psi}_h$) again, together with its 90% confidence interval. A similar pattern compared to the main results of Section 1.4.1 emerges. The literature standard conclusions with respect to the linear coefficients still hold, and, more importantly, also this specification provides a consistent picture with respect to the non-linear interaction terms. The cubed shock coefficient counteracts the linear term for both inflation and output. Larger monetary policy shocks prompt a weaker price and output puzzle at early horizons, and thus a negative overall effect for sufficiently large shocks. Furthermore, there seems to be less persistence as the cubed shock coefficients become significantly positive whilst the linear effect is negative for both inflation and output in the second part of the IRFs. Consequently, this again seems to point towards state-dependency as, for large shocks, inflation reacts stronger at short horizons and weaker at long horizons. To conclude, results are in line with the main specification of Section 1.4.1 and confirm our first theoretical predictions once again.

Regarding the squared shock coefficient, there is some weak, statistically insignificant, evidence that prices react more negatively at short horizons for positive (contractionary) monetary policy shocks, reducing somewhat the linear price puzzle. Conversely, negative (expansionary) shocks would prompt an amplified price puzzle. There is stronger and statistically significant evidence that positive shocks reduce the persistence and depth of the inflation impulse response at longer horizons. The squared shock coefficient in the output impulse response is largely insignificant, except for the early horizons where we can also observe a weakening of the output puzzle for positive shocks. This result is consistent with the findings of Cover (1992) or Barnichon & Matthes (2018) that positive money-supply (expansionary) shocks have a weaker effect on output than negative money-supply (contractionary) shocks. However, the coefficient with respect to the federal funds rate is positively significant at the very early horizons, suggesting that monetary policy reacts stronger after positive shocks. Yet, this effect is marginal and dies away quite quickly. Overall, the results suggest asymmetric responses to monetary policy shocks, in accordance with the previous literature on this issue, as the ones in Tenreyro & Thwaites (2016) or Barnichon & Matthes (2018).

A.3.3 Unscaled impulse responses to large and small shocks

Figure A.5 depicts the unscaled (i.e., non-standardised) impulse responses for a 25 basis point shock and a 200 basis point shock and their 90% confidence interval calculated with the Delta method, respectively. This complements Figure 1.2 in the main text. First, the larger the shock, the quicker inflation decreases. Second, for a large enough shock, the initial price puzzle on impact disappears: the IRF to a large shock is firstly not significantly different from zero and then significantly negative, while it is positive for some months for small shocks (even if only marginally significantly). Consequently, a sufficiently large shock counteracts the small linear coefficient and switches the sign of the overall impulse response of inflation, removing a potential price puzzle.

A.4 Robustness: Smooth transition local projection

We now turn to the sensitivity of the impulse response estimates during high and low inflation regimes with respect to the specification of the smooth transition function, equation (1.3) in the main text.

A.4.1 Varying the regime switching parameter

A change of the regime switching parameter from to $\gamma = 3$ or $\gamma = 10$ does not have any significant impact on any of the impulse responses (see Figure A.6 and A.7).

A.4.2 Varying the percentile of inflation parameter

Figures A.8 and A.9 show that the results for both lowering c to the 70th percentile of trend inflation and increasing it to the 80th percentile are very similar to the baseline findings.

A.4.3 Using HP-filtered PCE inflation

Figure A.10 displays the results of the smooth transition local projection with HP-filtered PCE inflation as a state variable (z_t in equation (1.3)). We set the smoothing parameter $\lambda = 14440$, which is the standard value for monthly data. The results are again very close to the results in the main text..

A.4.4 Using the trend inflation measure from Ireland (2007)

Figure A.11 shows the results of the smooth transition impulse response estimation, using an alternative measure of trend inflation from the estimated DSGE model with a time-varying inflation target by Ireland (2007). We center the model-based estimate between 0 and 1 and use it as an alternative measure for the smooth transition function $F(z_t)$ in (1.3). The main results of the main specification still goes through, as prices react faster in a high inflation regime compared to a low regime. However, there is a significant price puzzle and output reacts significantly weaker in the low inflation regime.

A.5 Robustness: Both tests

This section presents the results for robustness checks performed with respect to both local projection of the main text, (1.1) and (1.2).

A.5.1 Alternative price measure: CPI

This robustness test uses Consumer Price Index (CPI) inflation instead of PCE inflation as the inflation measure. Figure A.12 and A.13 provide the results for the two main local projections using the CPI measure. It is clear that the results are quite robust to this change in the inflation response variable. The only change compared to the main results is that both regime-dependent impulse response functions now exhibit a significant price puzzle.

A.5.2 Controlling for commodity prices

For this robustness exercise, we add two lags and the contemporaneous value of the commodity price index in the controls. This is done in order to take into account of the fact that inflation is sensitive to movements in commodities and/or oil prices. Results, depicted in Figures A.14 and A.15 are largely unchanged in this case.

A.5.3 Controlling for financial frictions

Financial frictions (Bernanke et al. (1999); Kiyotaki & Moore (1997)) are another prominent propagation mechanism of monetary policy shocks in the literature. Variations in financial frictions over time could then affect our estimate, making them spurious. We control for financial frictions by including the contemporaneous value and two lags of the highly

informative corporate bond credit spread, introduced by Gilchrist & Zakrajsek (2012), as a proxy for financial frictions (hereafter GZ-spread). This series is available from January 1973, so we are estimating our local projection on a truncated sample.

The results, reported in Figures A.16 and A.17 are mostly unchanged. With respect to our first implication, we still can discern a weaker effect of large shocks on output and a more pronounced, yet less persistence response of prices. The figure for the second hypothesis shows that inflation behaves as in the baseline specification, with a statistically significantly different and quicker response in the high inflation regime compared to the low inflation regime. Yet, we do observe a stronger reaction of output in the high inflation regime. This is mostly driven by a significant output expansion and a non-significant contraction in the low inflation regime. However, if we calculate the cumulative output response as in Table 1.2 the cumulative output difference is statistically insignificant. Taking this into account, we conclude that our evidence in favor of a sticky price theory holds up to controlling for the presence of financial frictions.

A.5.4 Non-linear Romer and Romer (2004) regression

The shocks used in the main local projections are residuals of the estimated reaction function of the central bank. More specifically, Romer & Romer (2004) assume a concrete form of reaction function by regressing the change in the intended federal funds rate ($\widehat{\Delta \mathrm{FFR}_t}$) on a measure of forecast variables primarily obtained from the Greenbook. This method assumes that the reaction function is linear. Following Tenreyro & Thwaites (2016), we re-estimate the Romer & Romer (2004) regression using the smooth transition function of the main specification, as:

$$\widehat{\Delta \mathrm{FFR}_t} = \mathrm{F}(z_t)(\mathbf{X}_t\beta^{\mathrm{H}}) + (1 - \mathrm{F}(z_t))(\mathbf{X}_t\beta^{\mathrm{L}}) + e_t^{\mathrm{NL}}. \tag{A.12}$$

In doing so, the identified shocks now account for the possibility that monetary authorities reacted differently to forecasts in a high and low trend inflation regime. The new series of non-linear shocks (e_t^{NL}) has a correlation of 0.92 with the linear shocks sample (see Figure A.32). This suggests that, whilst the new reaction function is picking up some non-linearities, the original shocks are a good instrument. Figures A.18 and A.19 shows that the results are again similar to the original ones. Note that in the smooth transition local projection, output reacts significantly stronger at short horizons in the high inflation regime, but effect disappears very quickly having no significant impact on the cumulative response. Further, still see a quicker decrease of prices in that regime, reinforcing our

original results.

A.5.5 Shocks from a smooth transition VAR

This subsection investigates the sensitivity of the results with respect to an alternative shock measure. More specifically, we use a smooth transition version of the classical three equation recursive VAR including industrial production, PCE inflation and the federal funds rate to recover the structural VAR monetary policy shocks and then use these in our local projections instead of the Romer & Romer (2004) ones.[1] The correlation of these monetary policy shocks with the Romer & Romer (2004) ones is 0.32. Moreover, they are generally much larger with a standard deviation of 1.07 compared to our standard shocks that have a standard deviation of 0.31 (see Figure A.33). They also exhibit much more variation across subsamples. Figure A.20 and A.21 are generally supporting our baseline results but there are some differences that are worth pointing out. The local projection estimates of the linear and absolute value interaction exhibit the same pattern, but they are generally weaker in magnitude. Regarding the smooth transition local projection for the second theoretical implication, the reaction of inflation in the two regimes is in line with the prior of state dependence, yet the significance of this result is weaker compared to our baseline result. The differential reaction of output is mostly insignificant, except for early horizons where output reacts stronger in a high inflation regime. The reaction of the Fed Funds rate is much stronger in the high inflation regime in the first six months, which could explain the stronger initial output response. We conclude based on this that the main conclusion of higher price flexibility in a high inflation regime still applies, especially when one considers the estimates of the inflation response.

A.5.6 Including leads and lags of the shocks

As shown in Alloza et al. (2019) it is important to account for potential persistence of narrative shocks by including leads of the shock. We follow this suggestion by re-estimating our main specifications with one lag and one lead of the shock included. The results in Figures A.22, A.23 are largely unchanged and consequently suggest that our results are not sensitive to the inclusion of a lag or lead structure of shocks.

[1] We use a lag polynomial of 3, as suggested by various information criteria. This lag length is also very similar to the one used in Caggiano et al. (2017) where they use a non-linear VAR with a lag length of 2 months.

A.5.7 Quarterly estimation using GDP as output measure

In order to check the robustness of our results with respect to the output measure, we switch to a quarterly specification of our local projections and use real GDP instead of industrial production as real GDP is measured at a quarterly frequency. The results are depicted in Figure A.24 and A.25. The results are very similar to our monthly baseline specification with industrial production as our output measure.

A.5.8 Quarterly estimation including fiscal policy controls

In order to control for the effects of fiscal policy on our estimates we also use a quarterly estimation of our local projection as most of these shocks are measured at a quarterly frequency. First, we control for government spending shocks by including two lags and the contemporaneous value of the excess returns of military contractors, popularised by Fisher & Peters (2010). These results are reported in A.26 and A.27. The results with respect to the absolute value interaction are basically unchanged. The results with respect to our smooth transition local projection are a bit different compared to the baseline. There is no significant difference between the two inflation impulse responses at early horizons, yet the point estimates and the significance at later horizons do support our main results. Further, we now see a stronger output response in the low inflation regime, in line with our state-dependent pricing. Second, we control for tax shocks by including two lags and the contemporaneous value of the exogenous tax series by Romer & Romer (2010) in A.28 and A.29. The results with respect to both tests are in line with our main results, and even somewhat better, as there is a significantly stronger reaction of industrial production in the low inflation regime.

A.5.9 Unsmoothed results

Figures A.30 and A.31 report the unsmoothed results of the benchmark non-linear and smooth transition local projections, respectively.

A.6 Shock distribution: Asymmetries and business-cycle dependencies

Figures A.32 and A.33 provide the time series of the shocks used in our local projections and their respective correlations.

Recent research has documented that monetary policy shocks have different effects in booms versus recessions as well as asymmetric effects in terms of positive versus negative shocks (see Barnichon & Matthes, 2018; Tenreyro & Thwaites, 2016). Hence, it is important to analyse whether the distribution of the size of the shock and the high vs. trend inflation regime are independent of the state of the economy and the sign of the shock. Figures A.34 and A.35 show the state-dependent distributions of the linear Romer and Romer shocks for high- and low inflation regimes and recessions and booms, respectively.[2] First, regarding our results for high vs. low trend inflation, the high inflation regime exhibits slightly fatter tails than the low trend inflation regime, but both for positive and negative shocks. Hence, it does not seem that the results about trend inflation are due to having an asymmetric distribution of positive and negative shocks in low vs. high trend inflation regimes. Moreover, we included a dummy term for recessions in order to control for differential effects of high and low inflation responses, depending on the business cycle. The results are shown in Figure A.36. It is evident that inflation reacts quicker in high inflation regimes, both during booms and recessions. This validates the robustness of our results to the state of the business cycle. Second, regarding our results for large vs. small shocks, A.35 shows that the distribution of the shocks is different between recessions and booms, as already highlighted by Tenreyro & Thwaites (2016). Recessions feature large negative shocks. This complicates the identification of all these effects in US data, because recessions, negative shocks and large shocks tend to appear at the same time. As the previous cited papers that wanted to distinguish the effects of the state of the economy (recession vs. boom) from the effects of the sign of the shock (asymmetric effect), our analysis is subject to this caveat. Our specification (1.1), however, does control for the sign of the shock, given the presence of a quadratic term in it, as explained above. While we do find some evidence of asymmetric effects, our results are unchanged when controlling for it. Moreover, when we interact our non-linear and absolute value interaction with a dummy term for recessions we find that the shape of the point estimate is constant across regimes and the difference between these two IRF is largely insignificant (see Figure A.37).

[2]The average shock distribution is computed with the linear Romer and Romer shocks, smoothed with a normally distributed kernel. The high inflation and low inflation estimates are generated by weighting the kernel function with the headline smooth transition function. The boom and recession estimates are generated by weighting the kernel function with the NBER recession dates.

A.7 Hansen (1992) test procedure

This description closely follows the exposition in Hansen (1992). We use the following specification in order to test for possible structural breaks in the individual impulse response coefficients:

$$y_{t+h} = \alpha_h + \beta_h \epsilon_t + \zeta_h (\epsilon_t \cdot |\epsilon_t|) + \sum_{k=1}^{K} \gamma_{h,k} w_{t,k} + v_{t+h} \equiv b_h' x_t + v_{t+h} \tag{A.13}$$

and assume $E(v_{t+h}|x_t) = 0$ and $E(v_{t+h}^2) = \sigma_{t,h}$ and $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \sigma_{t,h}^2 = \sigma_h^2$. Furthermore, the variables cannot contain any deterministic or stochastic trends. Accordingly, we modify our original specification by excluding the deterministic time trend and taking first differences of the industrial production and federal funds rate control variables to ensure the fulfillment of this assumption.
Estimating equation (A.13) with ordinary least squares yields $(\hat{b}_h, \hat{\sigma}_h)$ and the following system of first-order conditions:

$$0 = \sum_{t=1}^{T} f_{k,t,h} \quad k = 1, \ldots, K+1 \tag{A.14}$$

where the variables $\{f_{k,t,h}\}$ are defined as:

$$f_{k,t,h} = \begin{cases} x_{k,t}\hat{v}_{t+h} & \text{for} \quad k = 1, \ldots, K \\ (\hat{v}_{t+h}^2 - \hat{\sigma}_h) & \text{for} \quad k = K+1 \end{cases} \tag{A.15}$$

where K is the number of coefficients to be estimated in the local projection. The Hansen (1992) individual test statistic is then based on the cumulative of these first order conditions. Defining the cumulative first-order condition at time t for horizon h and estimate k as $S_{k,t,h} \equiv \sum_{j=1}^{t} f_{k,j,h}$, the individual test statistic can be written as:

$$L_{k,h} = \frac{1}{T}\frac{1}{V_k}\sum_{t=1}^{T} S_{k,t,h}^2 \tag{A.16}$$

where $V_k = \sum_{t=1}^{T} f_{k,t,h}^2$. This is essentially an average of squared cumulative sums of first-order conditions related to parameter k. The null hypothesis is that $b_{k,h}$ is constant and so the first-order conditions are mean zero and thus the cumulative sums wander around zero. Note that the respective distribution is non-standard and depends on the number of parameters tested for stability. The alternative is that the parameter k is not stable and will

develop a non-zero mean and thus $S_{k,h}$ will be large and thus increasing the test statistic.

A.8 Figures

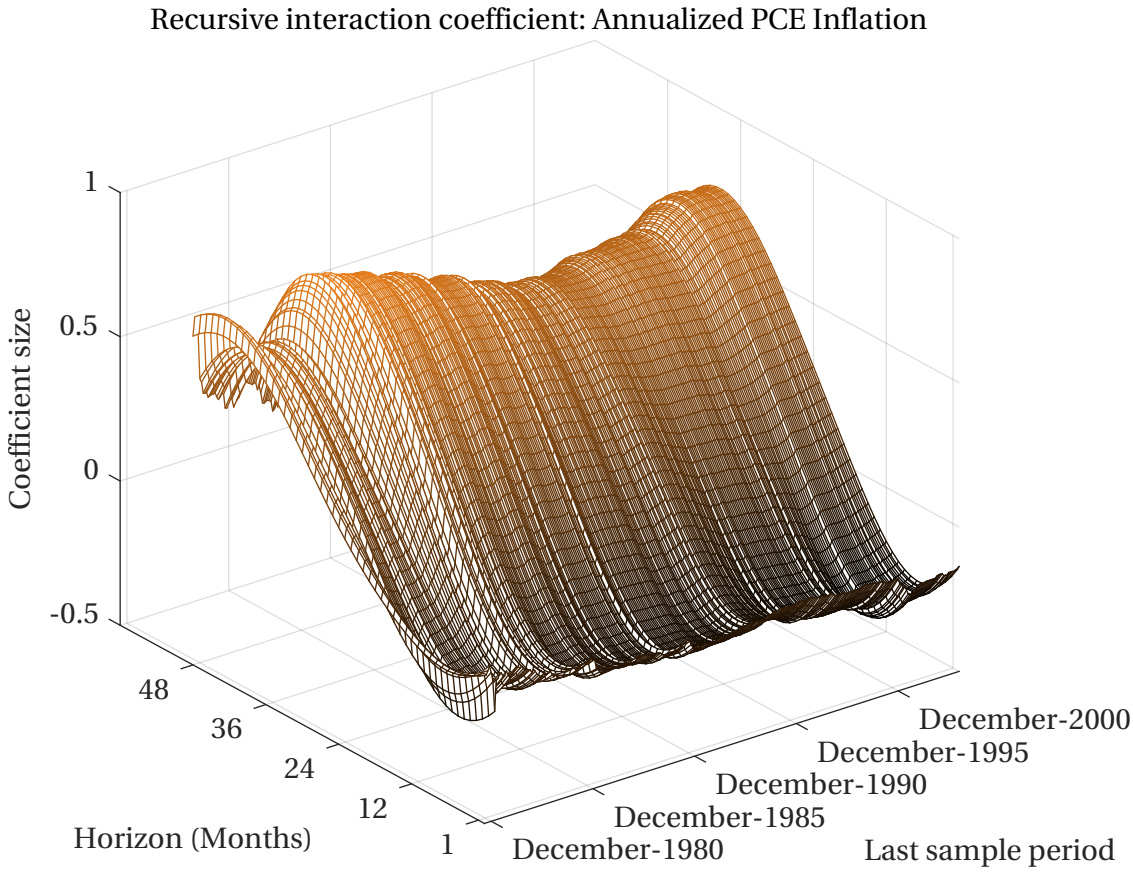


Fig. A.1: Recursive smooth PCE inflation local projection: absolute value interaction coefficient. The x-axis indicates the horizon of the impulse response, the y-axis indicates the last period of that estimate and the z-axis indicates the coefficient. The first sample contains the data between January 1969 and August 1980 whilst the last sample contains data between January 1969 and December 2003.

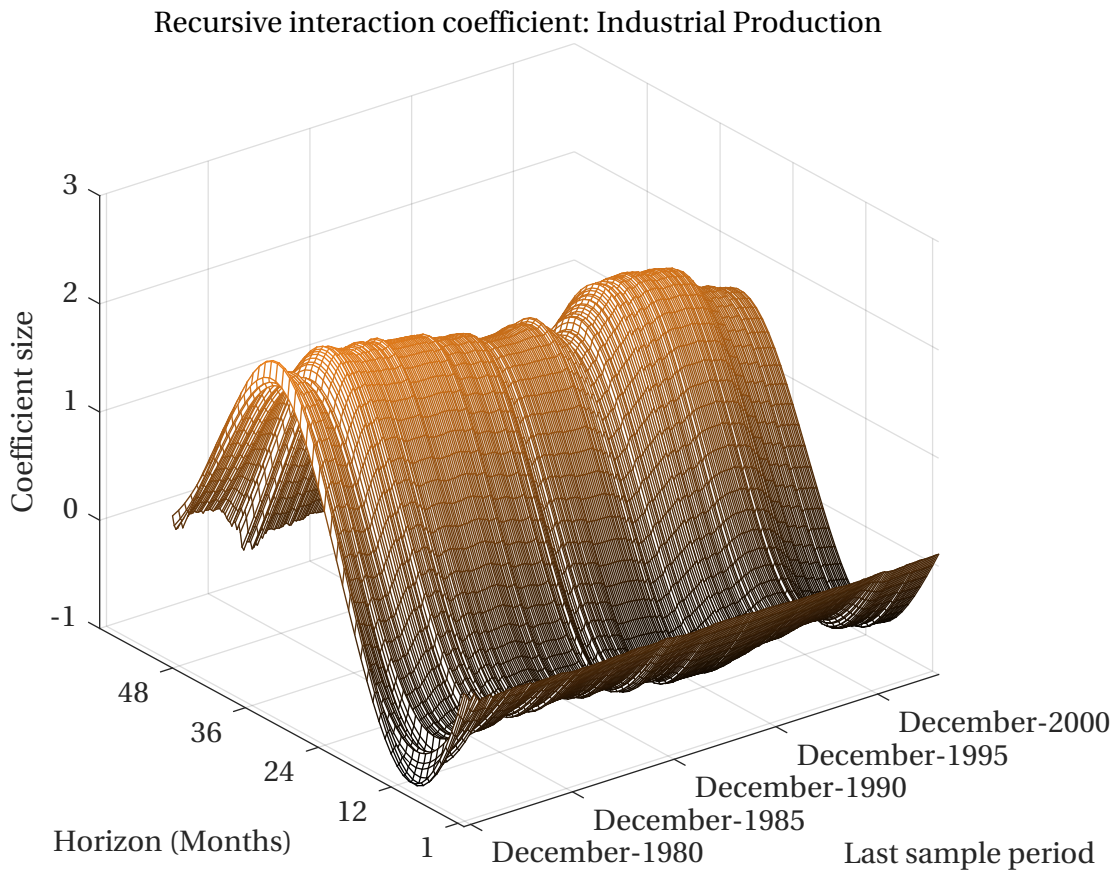


Fig. A.2: Recursive smooth industrial output local projection: absolute value interaction coefficient. The x-axis indicates the horizon of the impulse response, the y-axis indicates the last period of that estimate and the z-axis indicates the coefficient. The first sample contains the data between January 1969 and August 1980 whilst the last sample contains data between January 1969 and December 2003.

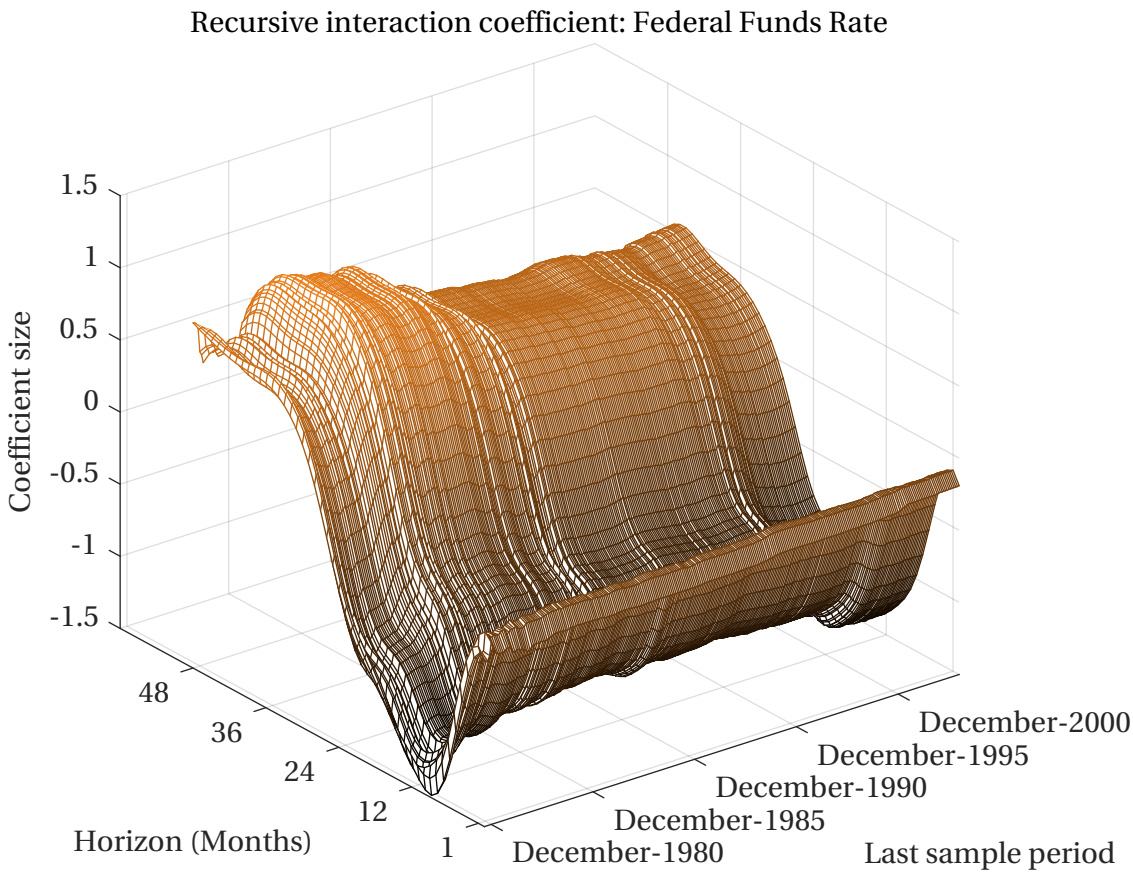


Fig. A.3: Recursive smooth federal funds rate local projection: absolute value interaction coefficient. The x-axis indicates the horizon of the impulse response, the y-axis indicates the last period of that estimate and the z-axis indicates the coefficient. The first sample contains the data between January 1969 and August 1980 whilst the last sample contains data between January 1969 and December 2003.

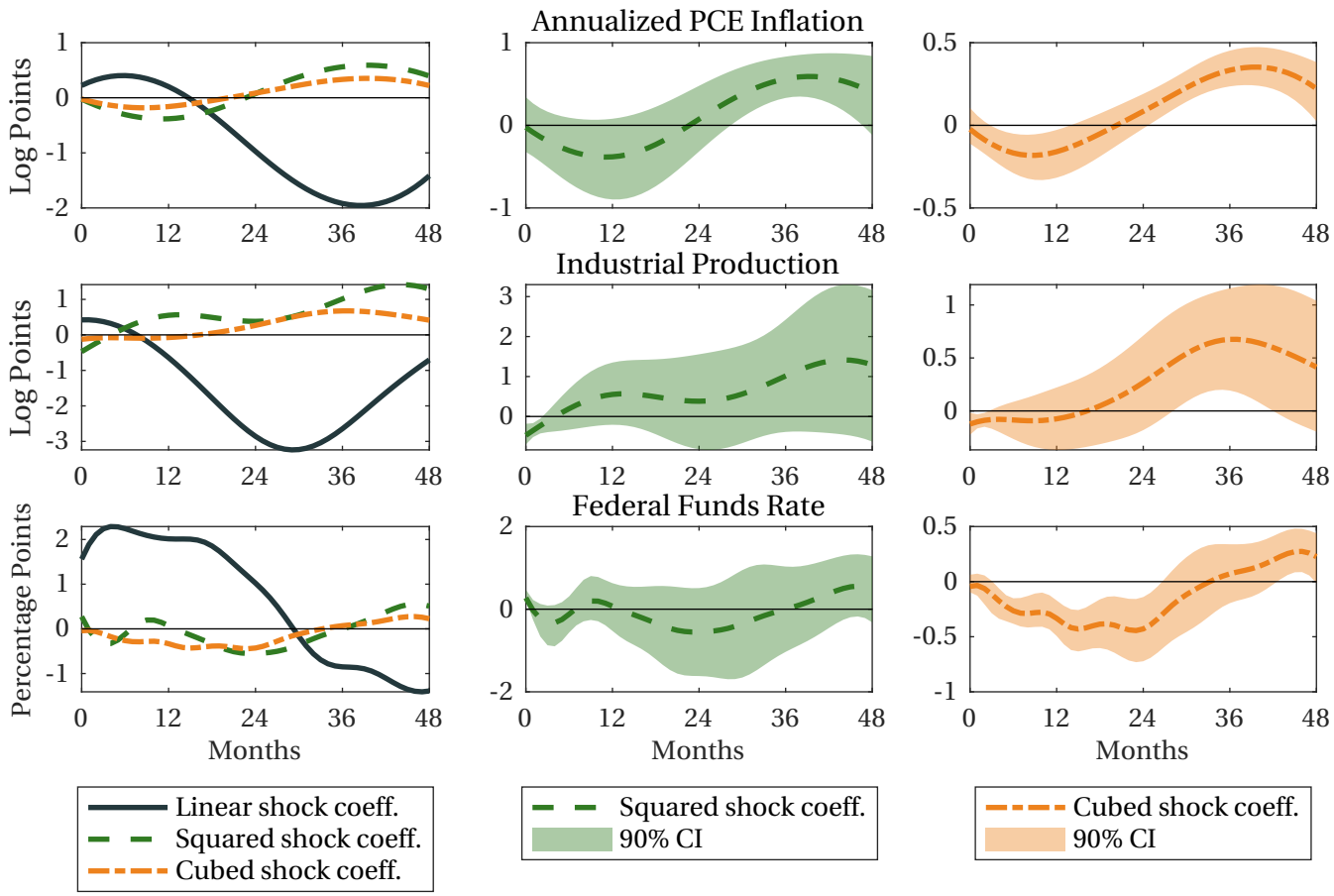


Fig. A.4: Smooth local projection coefficients for annualized PCE inflation (first row), industrial production (second row) and the federal funds rate (third row). The first column depicts the point estimate with respect to the shock (solid line), its squared value (dashed line) and its cubed value (dashed-dotted line). The second column depicts the point estimate of the squared value again, together with its 90% confidence interval. The third column depicts the point estimate of the cubed value again, together with its 90% confidence interval. All of the coefficients are depicted over a four year horizon.

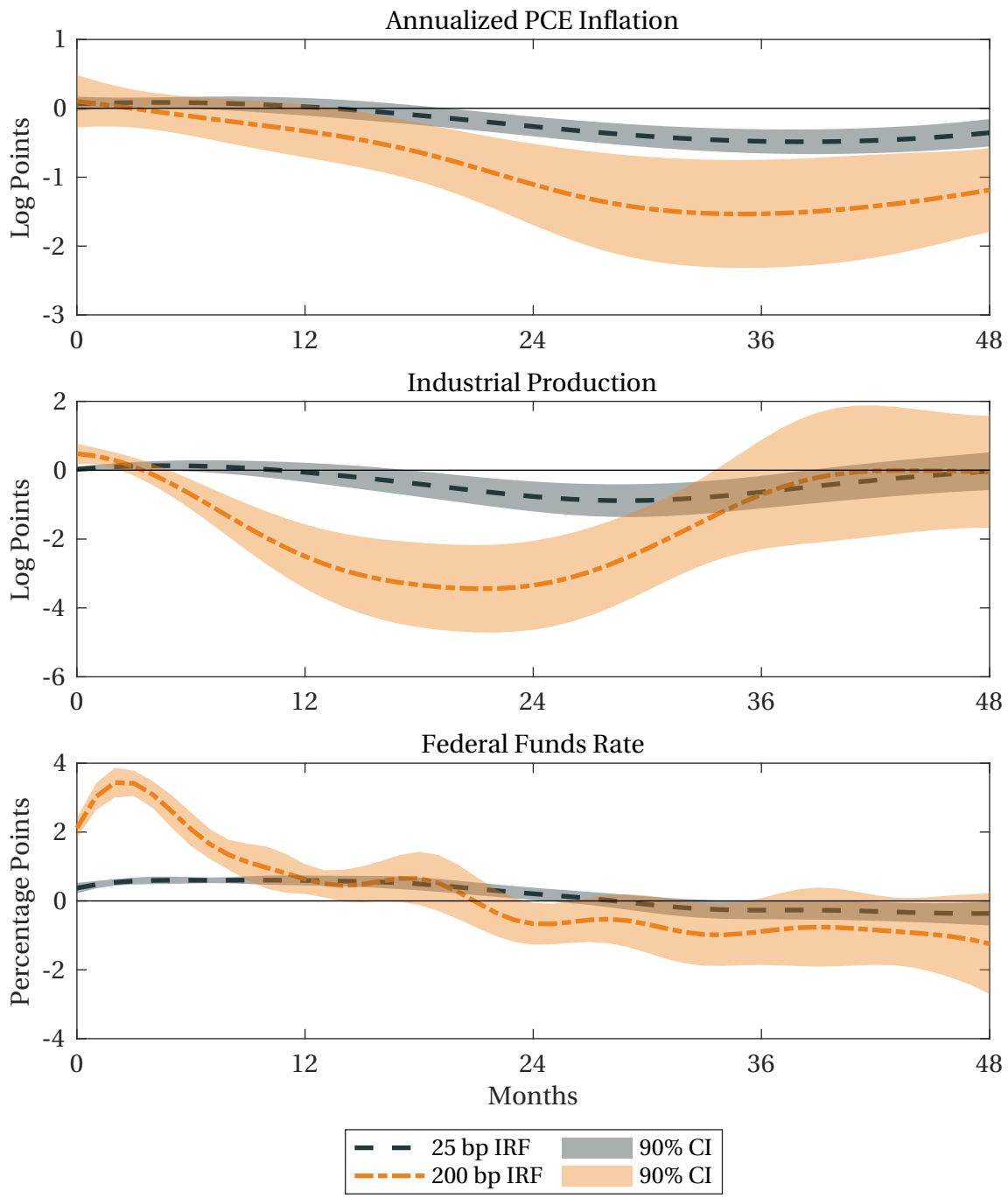


Fig. A.5: Simulated size-dependent impulse responses for annualised PCE Inflation, Industrial Production and the Federal Funds Rate over a four year horizon. The Figure depicts the impulse response for a 25 (dashed line) and 200 (dashed-dotted) basis point shock. The impulse responses are depicted over a four year horizon.

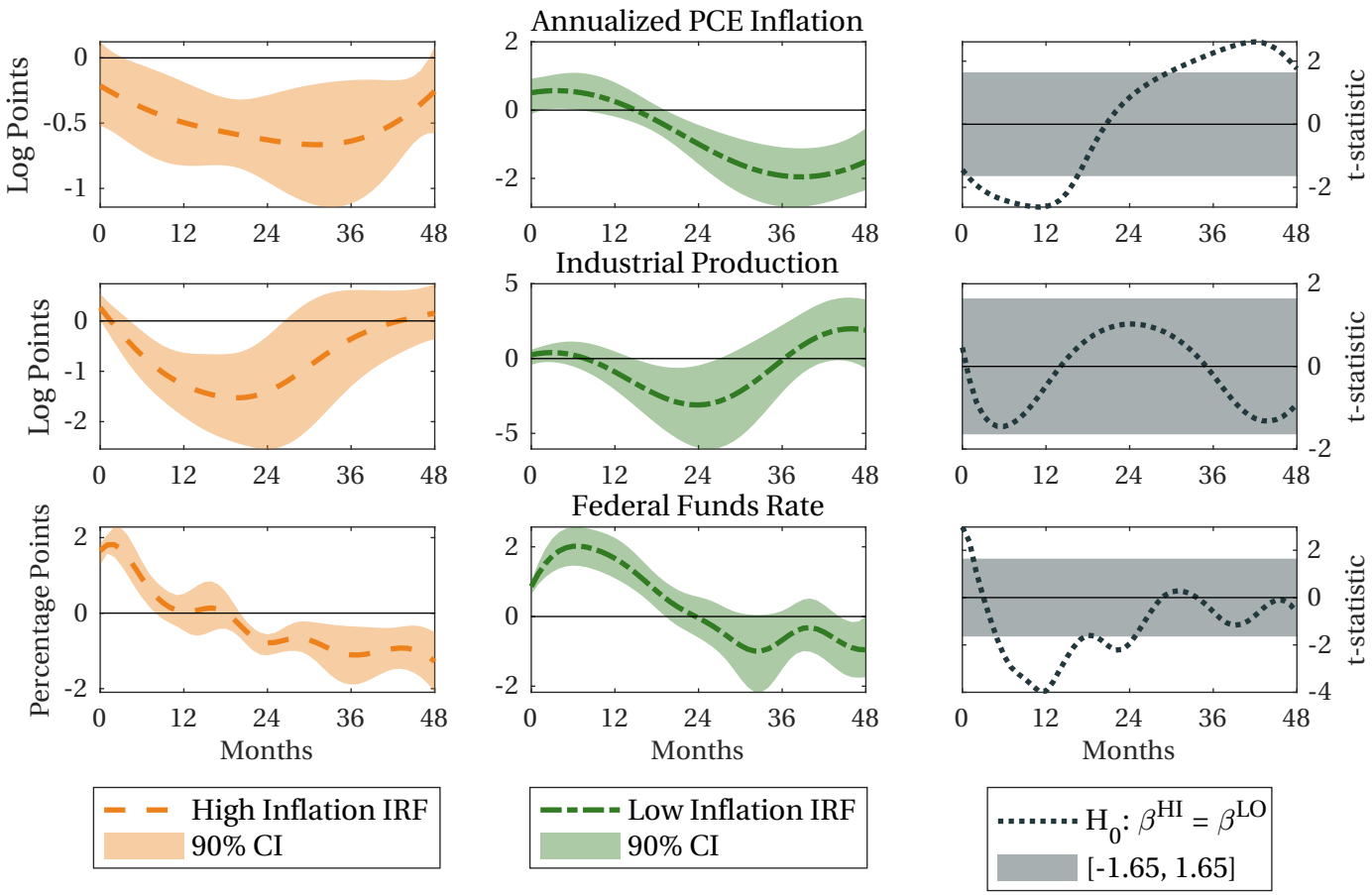


Fig. A.6: Smooth impulse response functions in different inflation states with a lower speed of transition ($\gamma = 3$). The three response variables are annualized PCE Inflation (first row), industrial output (second row) and the federal funds rate (third row). The first column depicts the point estimates of the high inflation impulse response (dashed line) together with its 90% confidence interval. The second column depicts the point estimates of the low inflation impulse response (dashed-dotted line) together with its 90% confidence interval. The third column depicts the t-statistic for the null hypothesis of equality of the high and low inflation impulse responses (dotted line), together with the 90% z-values (grey area). All of the coefficients are depicted over a four year horizon.

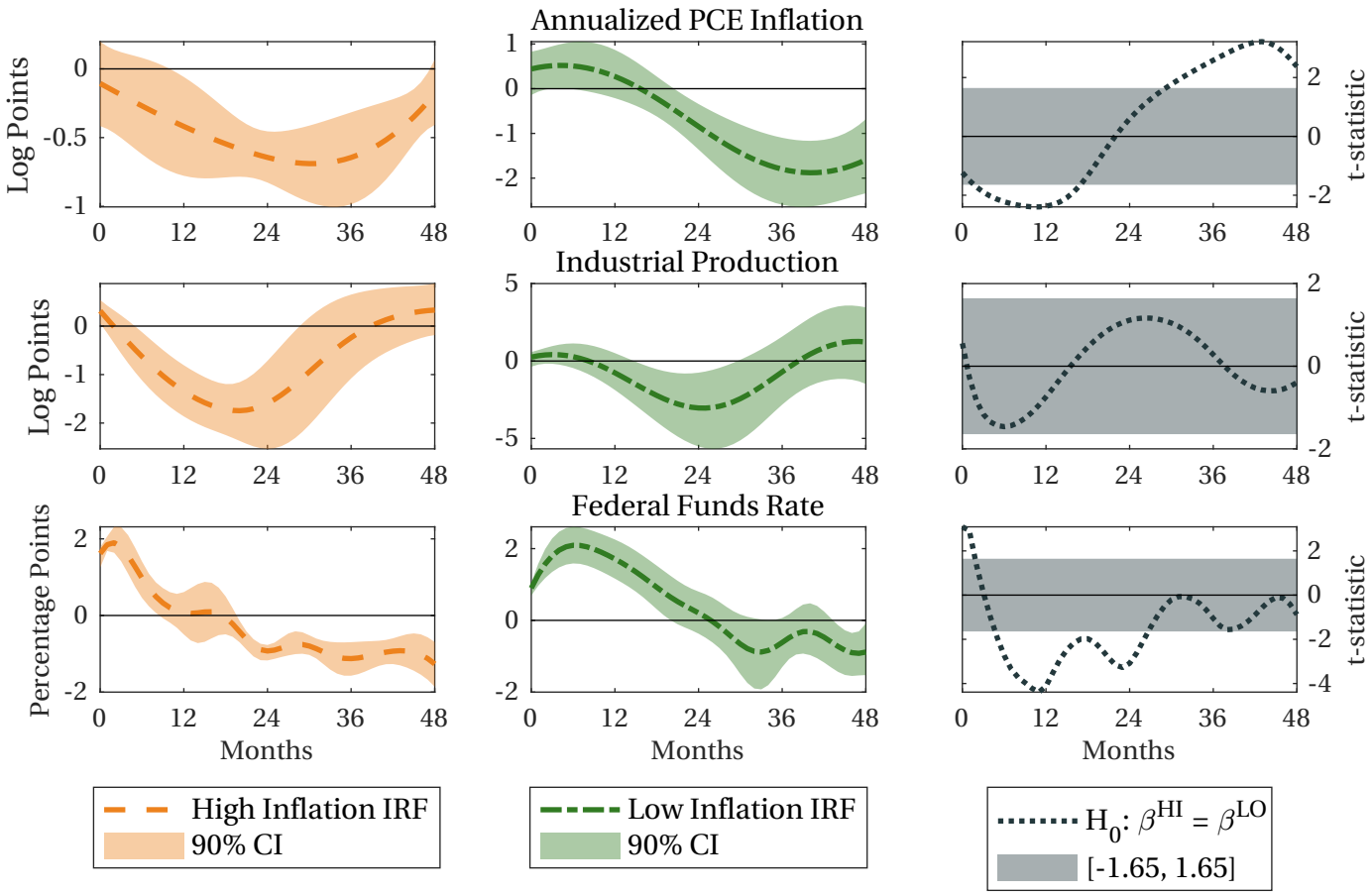


Fig. A.7: Smooth impulse response functions in different inflation states with a higher speed of transition ($\gamma = 10$). The three response variables are annualized PCE Inflation (first row), industrial output (second row) and the federal funds rate (third row). The first column depicts the point estimates of the high inflation impulse response (dashed line) together with its 90% confidence interval. The second column depicts the point estimates of the low inflation impulse response (dashed-dotted line) together with its 90% confidence interval. The third column depicts the t-statistic for the null hypothesis of equality of the high and low inflation impulse responses (dotted line), together with the 90% z-values (grey area). All of the coefficients are depicted over a four year horizon.

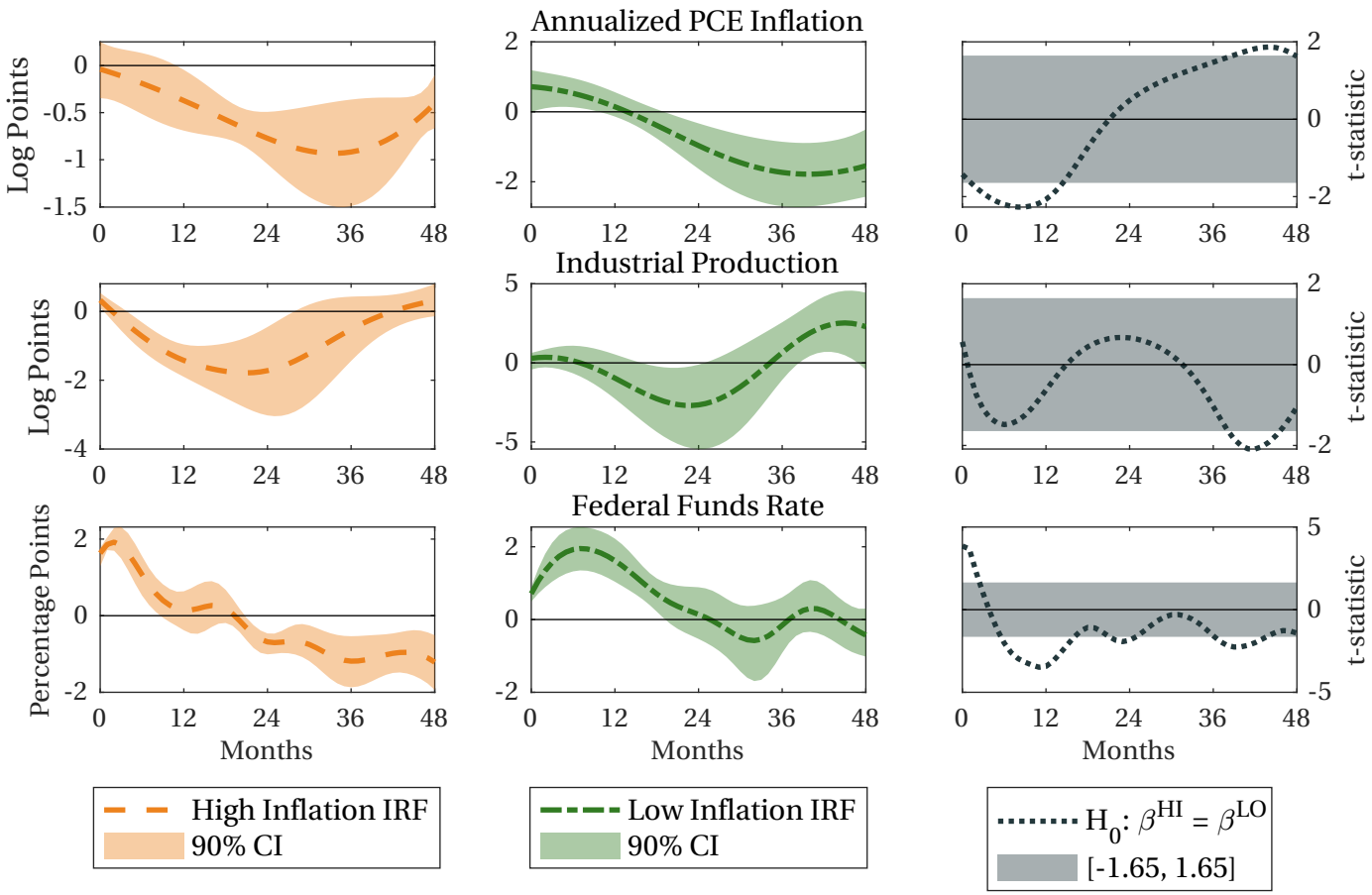


Fig. A.8: Smooth impulse response functions in different inflation states with a lower inflation threshold ($c = 0.7$). The three response variables are annualized PCE Inflation (first row), industrial output (second row) and the federal funds rate (third row). The first column depicts the point estimates of the high inflation impulse response (dashed line) together with its 90% confidence interval. The second column depicts the point estimates of the low inflation impulse response (dashed-dotted line) together with its 90% confidence interval. The third column depicts the t-statistic for the null hypothesis of equality of the high and low inflation impulse responses (dotted line), together with the 90% z-values (grey area). All of the coefficients are depicted over a four year horizon.

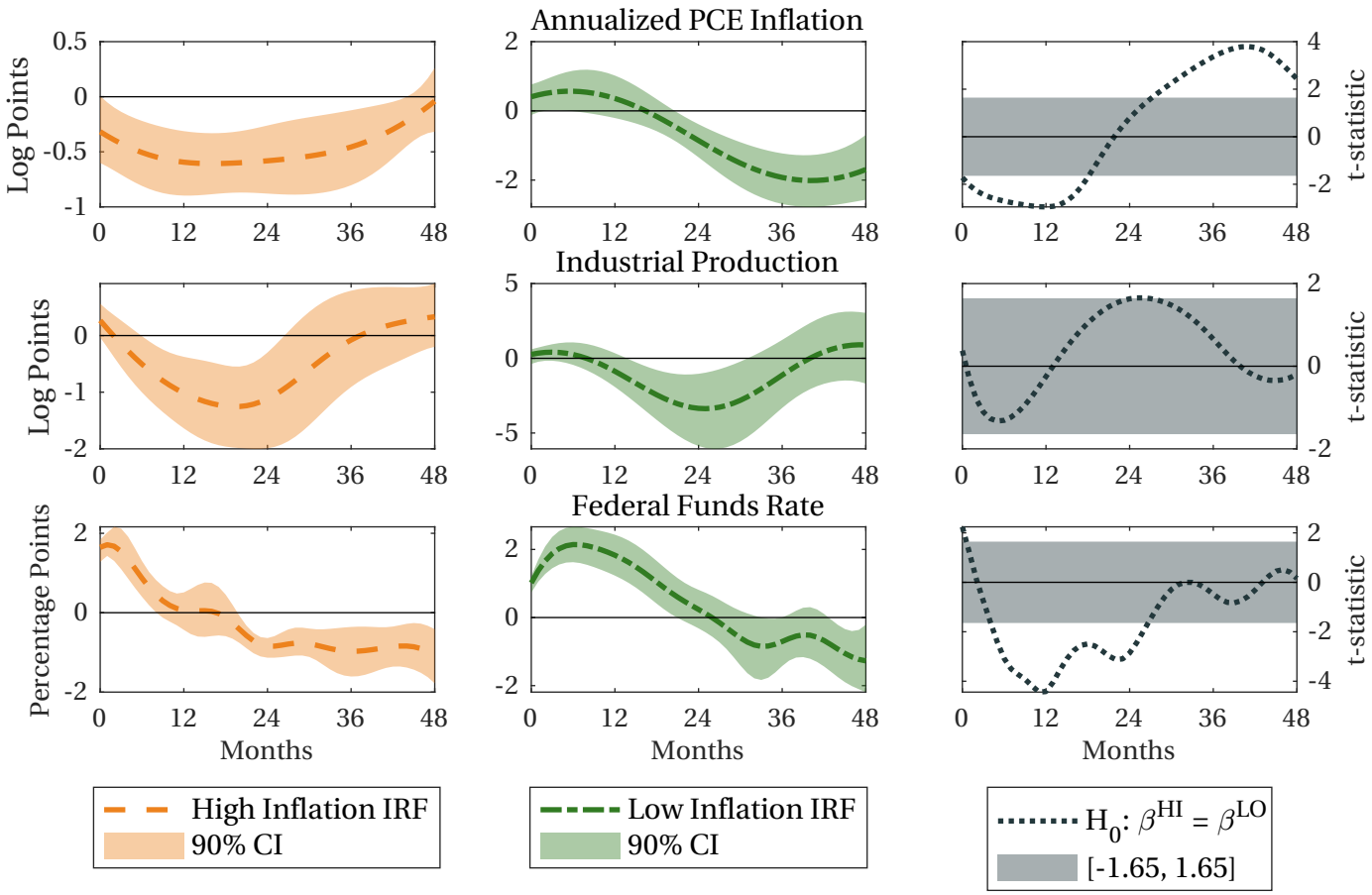


Fig. A.9: Smooth impulse response functions in different inflation states with a lower inflation threshold ($c = 0.8$). The three response variables are annualized PCE Inflation (first row), industrial output (second row) and the federal funds rate (third row). The first column depicts the point estimates of the high inflation impulse response (dashed line) together with its 90% confidence interval. The second column depicts the point estimates of the low inflation impulse response (dashed-dotted line) together with its 90% confidence interval. The third column depicts the t-statistic for the null hypothesis of equality of the high and low inflation impulse responses (dotted line), together with the 90% z-values (grey area). All of the coefficients are depicted over a four year horizon.

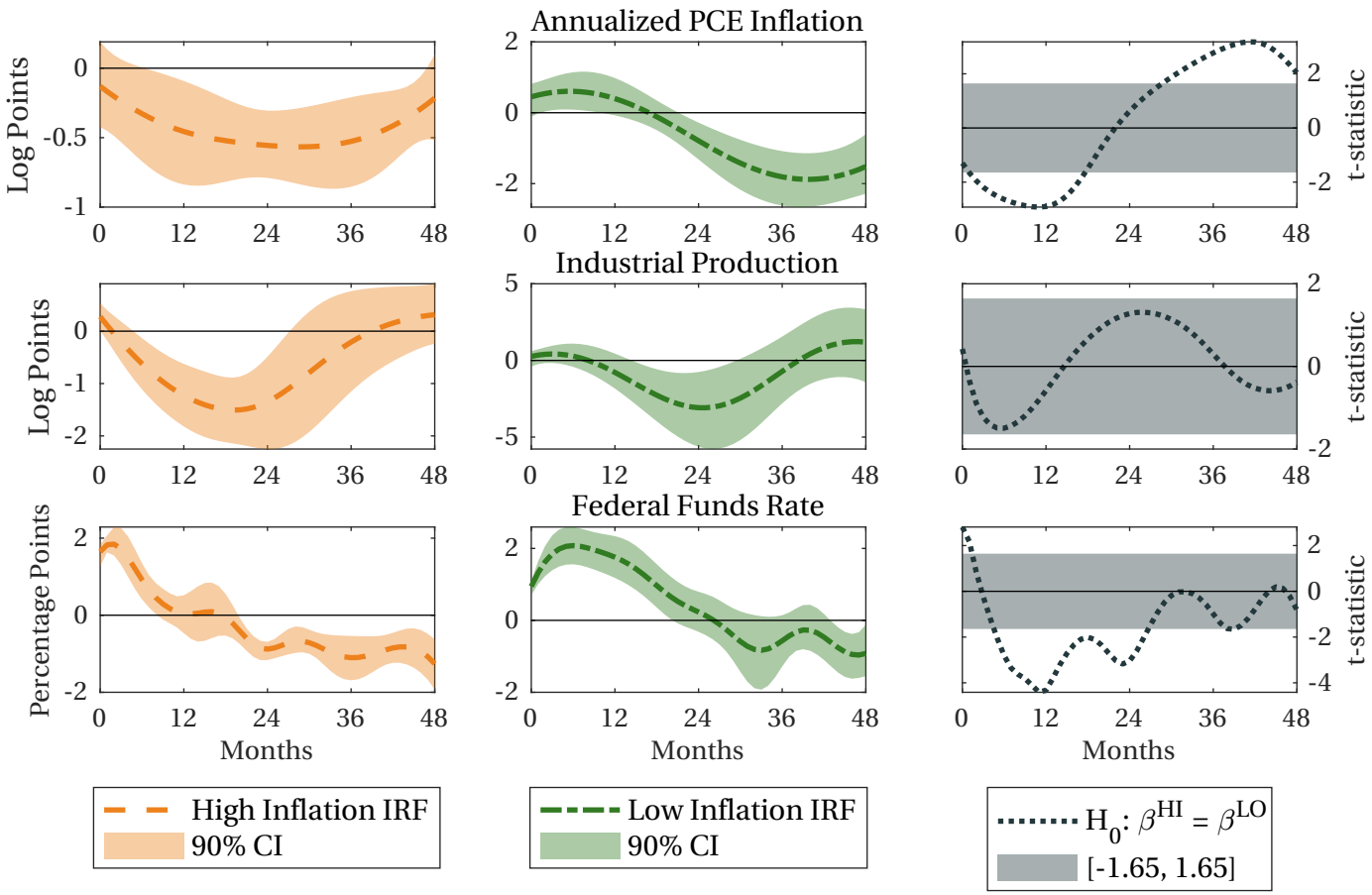


Fig. A.10: Smooth impulse response functions in different inflation states with HP-filtered PCE inflation as the state ($\lambda = 14400$). The three response variables are annualized PCE Inflation (first row), industrial output (second row) and the federal funds rate (third row). The first column depicts the point estimates of the high inflation impulse response (dashed line) together with its 90% confidence interval. The second column depicts the point estimates of the low inflation impulse response (dashed-dotted line) together with its 90% confidence interval. The third column depicts the t-statistic for the null hypothesis of equality of the high and low inflation impulse responses (dotted line), together with the 90% z-values (grey area). All of the coefficients are depicted over a four year horizon.

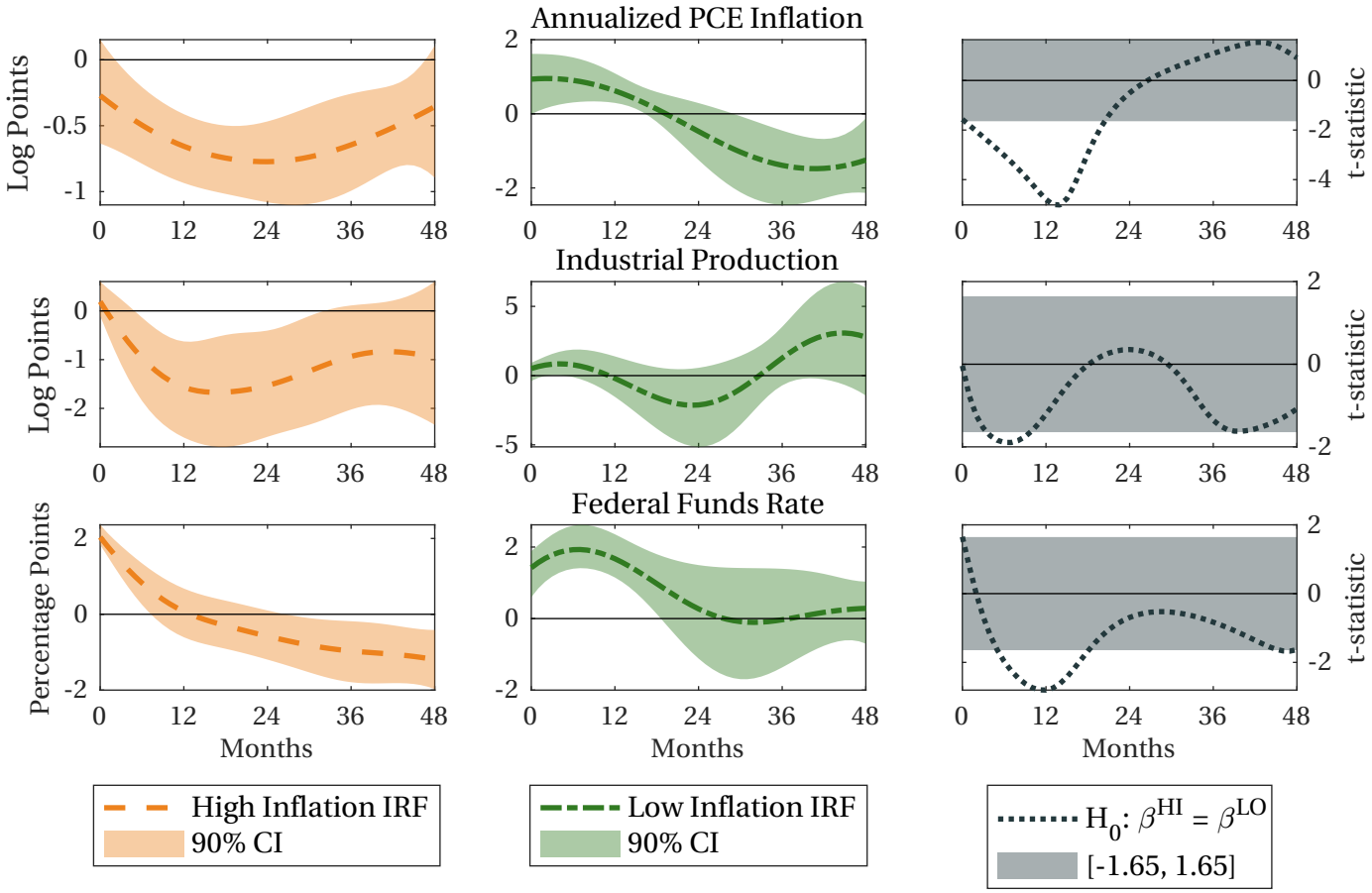


Fig. A.11: Smooth impulse response functions in different inflation states with Peter Ireland's (2007) measure of trend inflation as the state. The three response variables are annualized PCE Inflation (first row), industrial output (second row) and the federal funds rate (third row). The first column depicts the point estimates of the high inflation impulse response (dashed line) together with its 90% confidence interval. The second column depicts the point estimates of the low inflation impulse response (dashed-dotted line) together with its 90% confidence interval. The third column depicts the t-statistic for the null hypothesis of equality of the high and low inflation impulse responses (dotted line), together with the 90% z-values (grey area). All of the coefficients are depicted over a four year horizon.

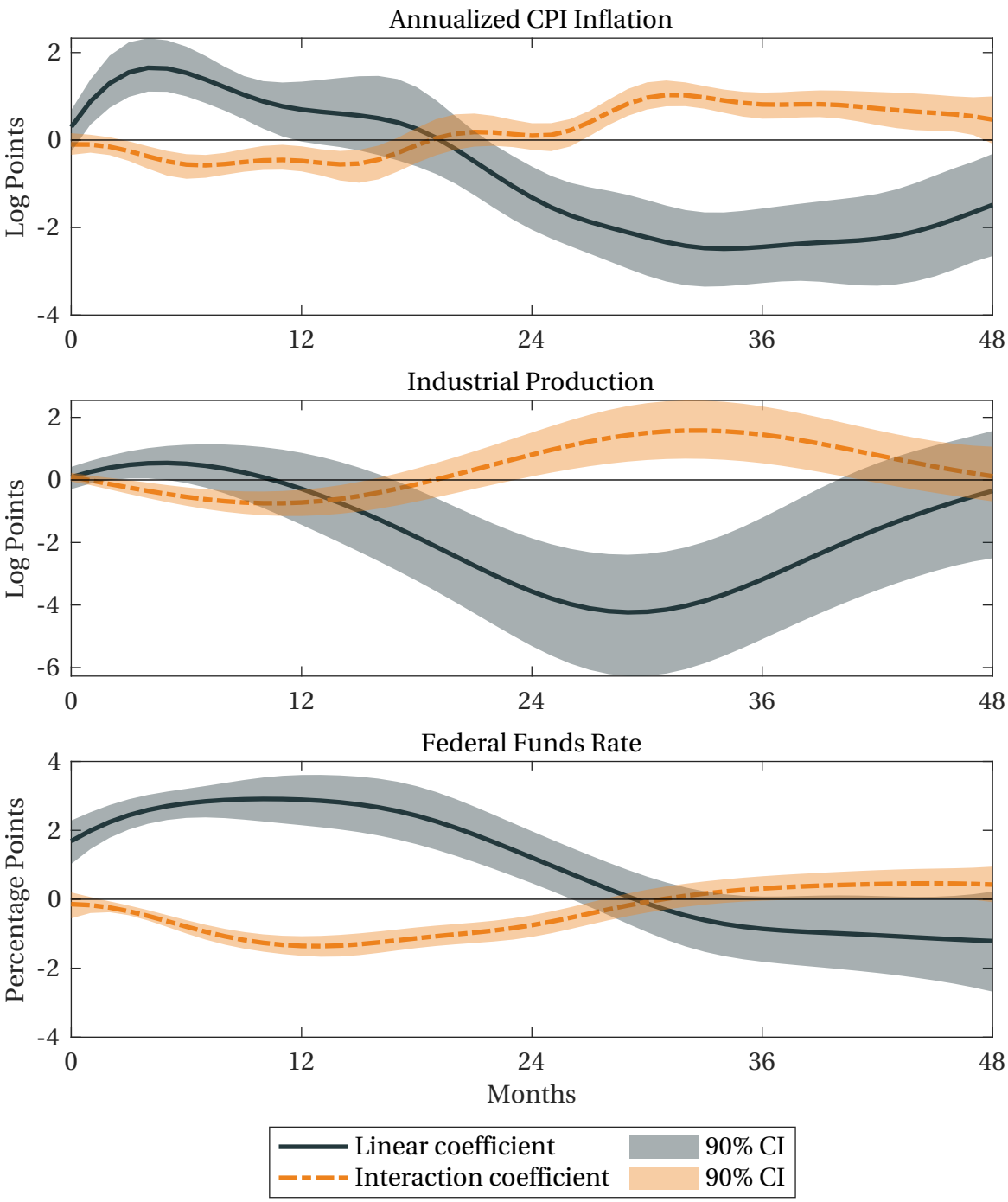


Fig. A.12: Smooth local projection coefficients with annualized CPI inflation as the price measure. The three response variables are annualized CPI Inflation, industrial output and the federal funds rate. Every panel depicts both the point estimates of the linear coefficient (solid line) and the absolute value interaction coefficient (dashed-dotted), together with their 90% confidence intervals. All of the coefficients are depicted over a four year horizon.

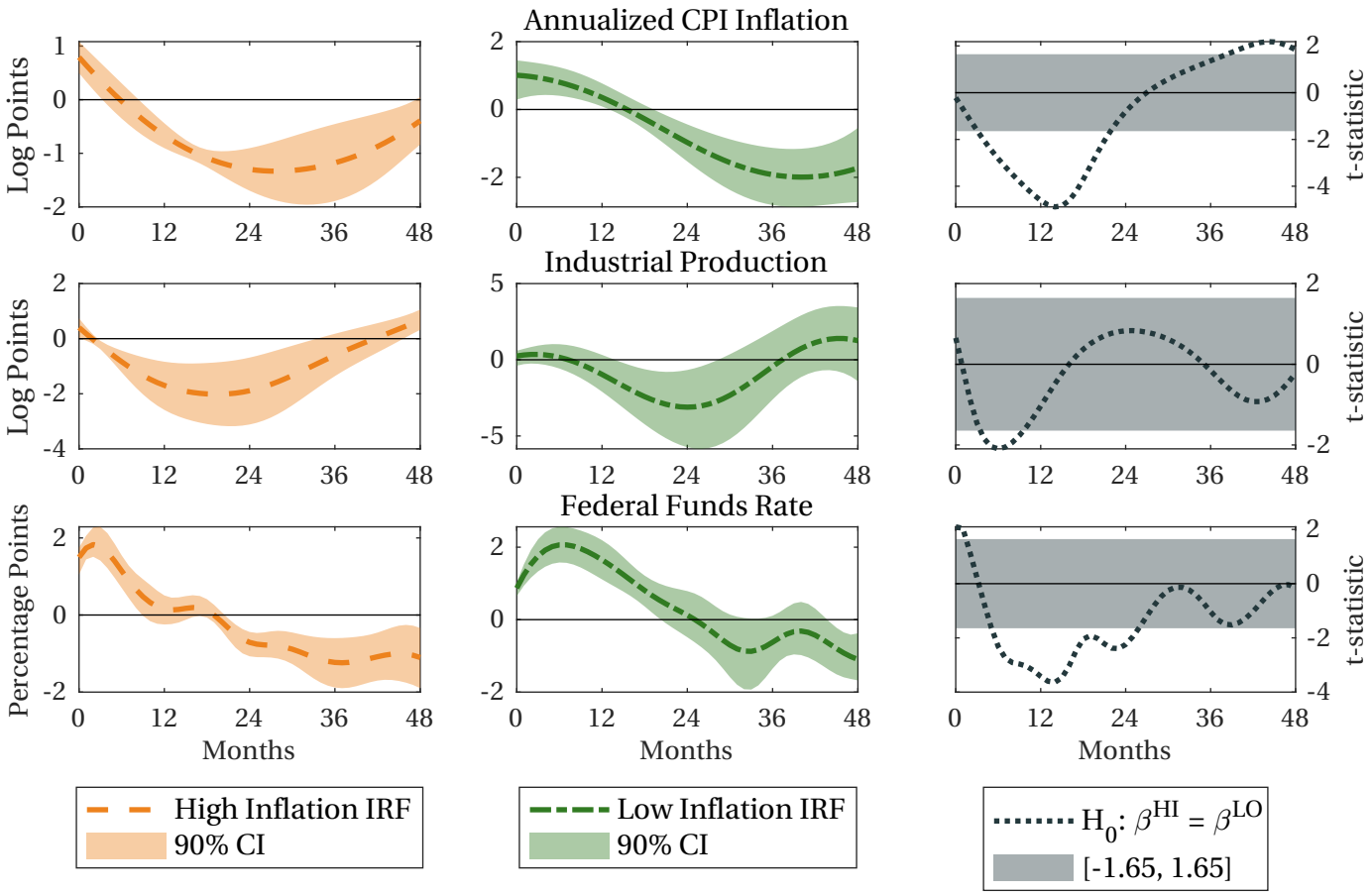


Fig. A.13: Smooth impulse response functions in different inflation states with annualised CPI inflation as its price measure. The three response variables are annualized CPI Inflation (first row), industrial output (second row) and the federal funds rate (third row). The first column depicts the point estimates of the high inflation impulse response (dashed line) together with its 90% confidence interval. The second column depicts the point estimates of the low inflation impulse response (dashed-dotted line) together with its 90% confidence interval. The third column depicts the t-statistic for the null hypothesis of equality of the high and low inflation impulse responses (dotted line), together with the 90% z-values (grey area). All of the coefficients are depicted over a four year horizon.

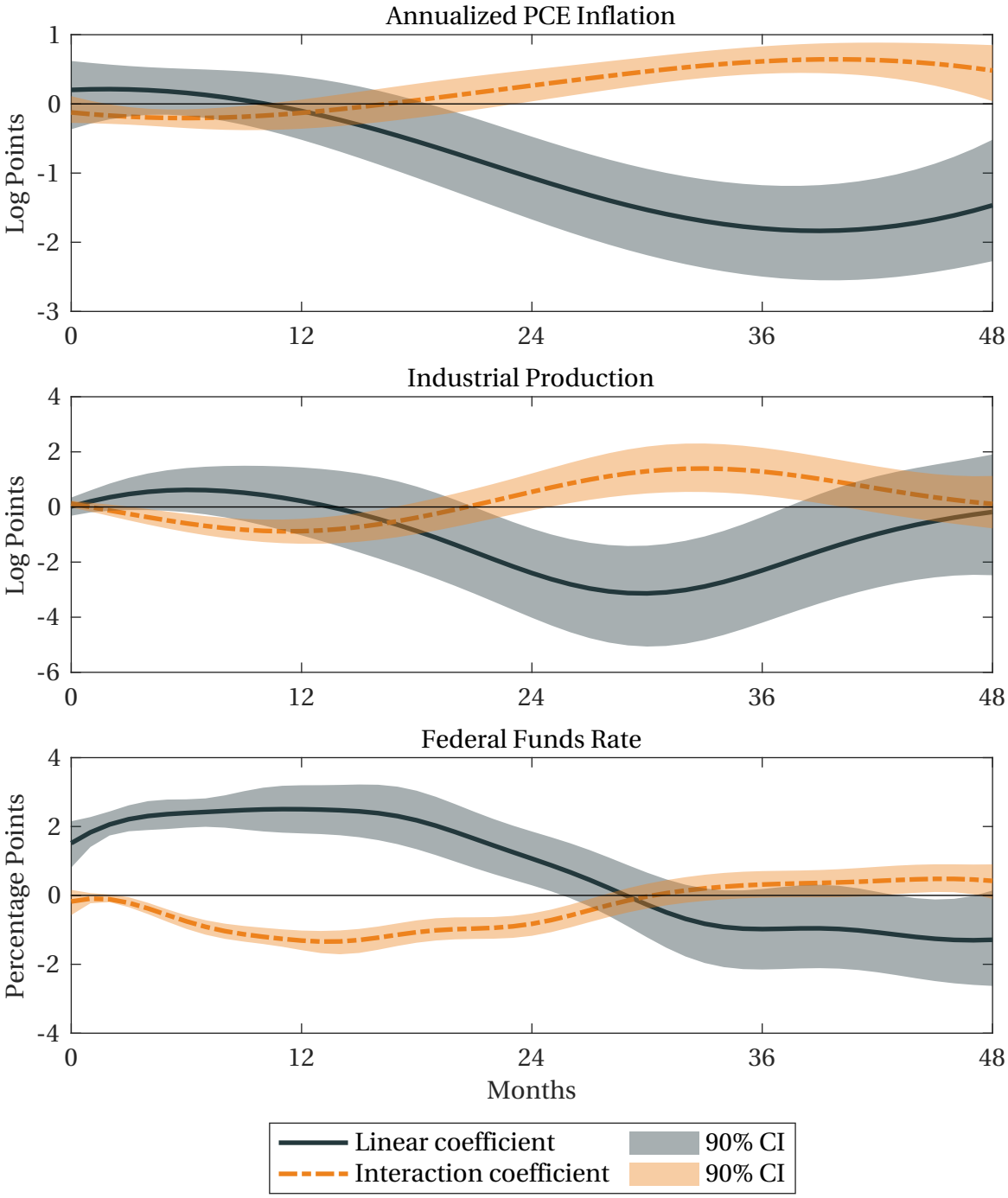


Fig. A.14: Smooth local projection coefficients, controlling for the commercial price index (PCOM). The three response variables are annualized PCE Inflation, industrial output and the federal funds rate. Every panel depicts both the point estimates of the linear coefficient (solid line) and the absolute value interaction coefficient (dashed-dotted), together with their 90% confidence intervals. All of the coefficients are depicted over a four year horizon.

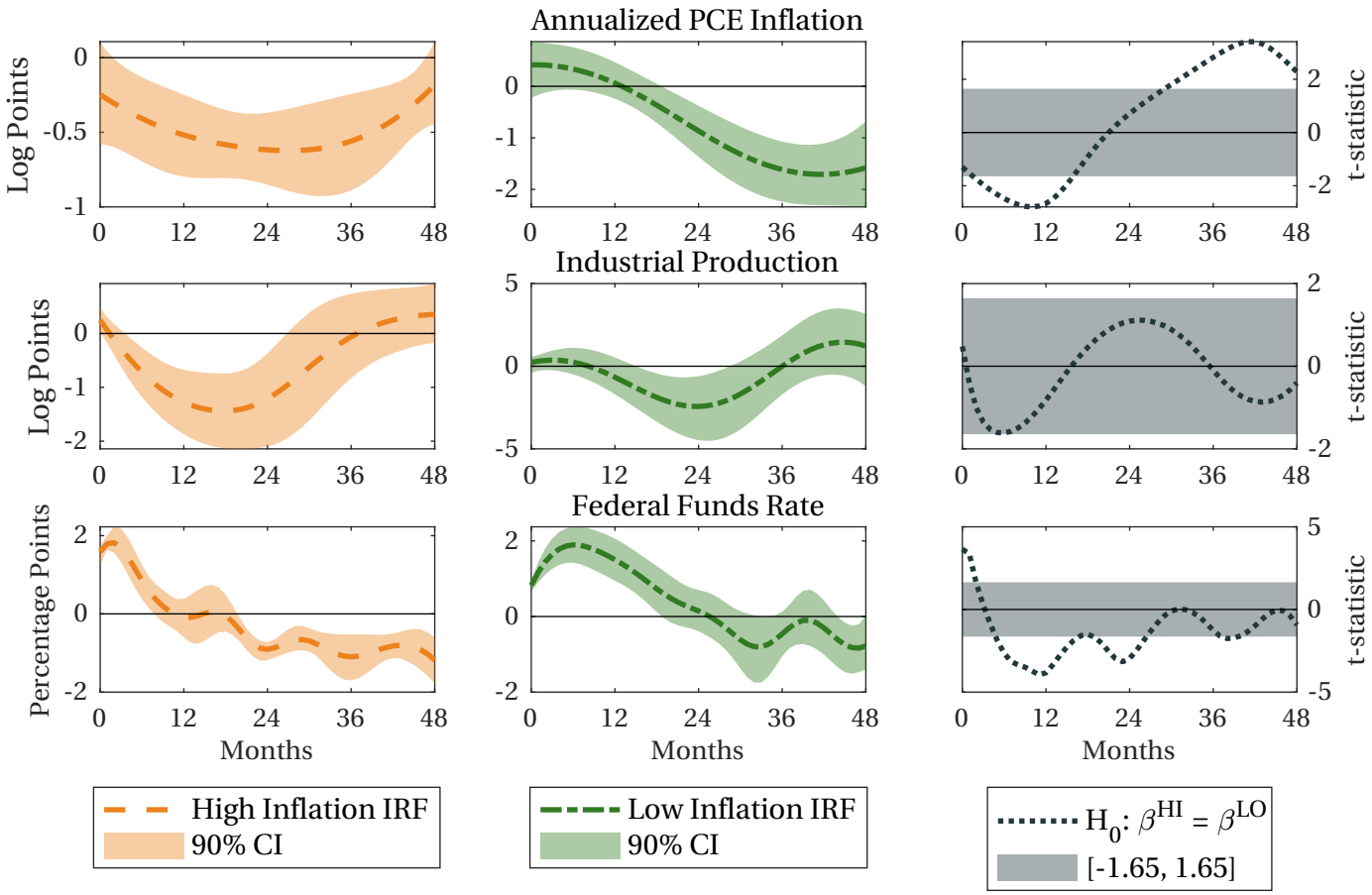


Fig. A.15: Smooth impulse response functions in different inflation states, controlling for the commercial price index (PCOM). The three response variables are annualized PCE Inflation (first row), industrial output (second row) and the federal funds rate (third row). The first column depicts the point estimates of the high inflation impulse response (dashed line) together with its 90% confidence interval. The second column depicts the point estimates of the low inflation impulse response (dashed-dotted line) together with its 90% confidence interval. The third column depicts the t-statistic for the null hypothesis of equality of the high and low inflation impulse responses (dotted line), together with the 90% z-values (grey area). All of the coefficients are depicted over a four year horizon.

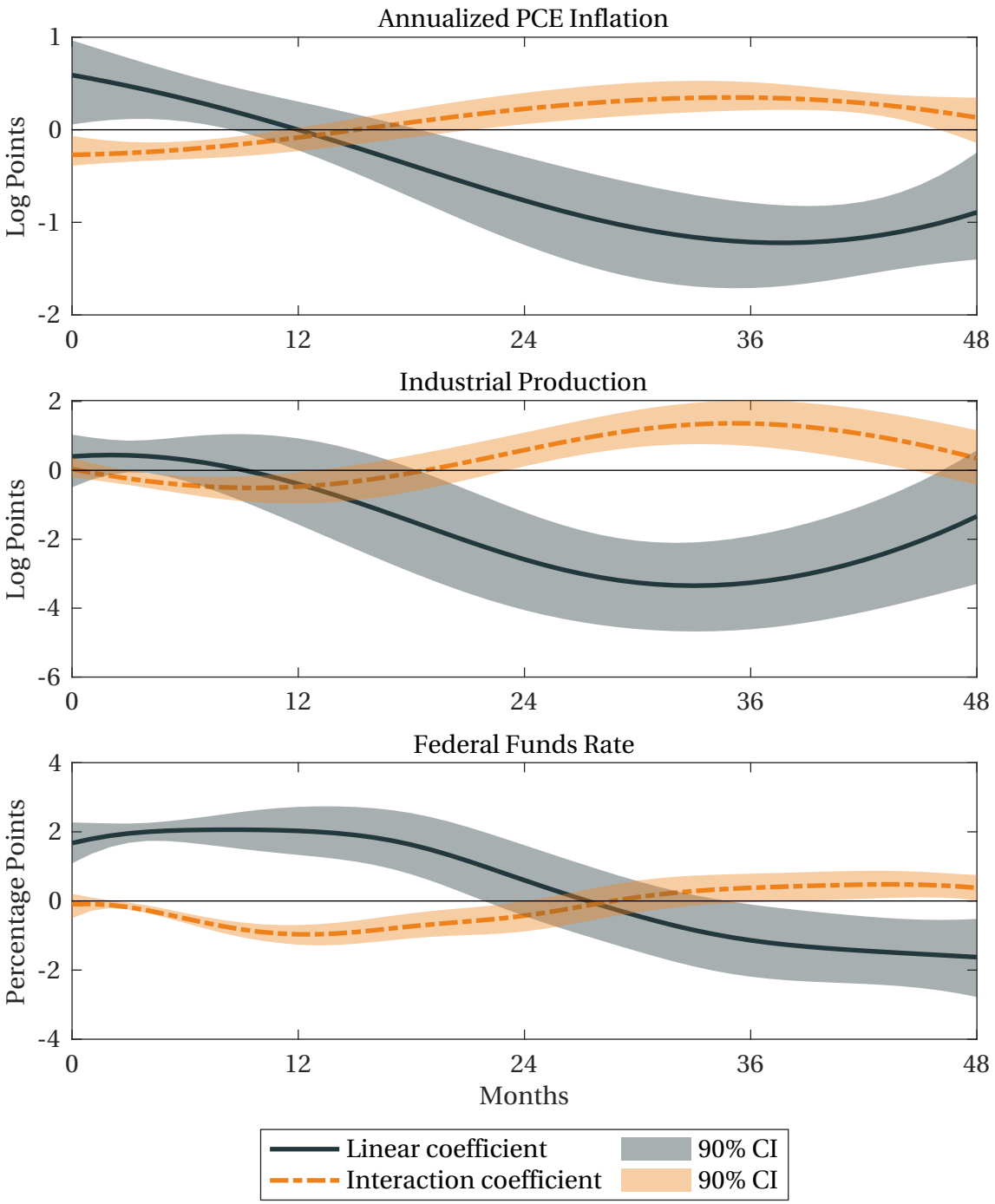


Fig. A.16: Smooth local projection coefficients, controlling for the Gilchrist & Zakrajsek (2012) index. The three response variables are annualized PCE Inflation, industrial output and the federal funds rate. Every panel depicts both the point estimates of the linear coefficient (solid line) and the absolute value interaction coefficient (dashed-dotted), together with their 90% confidence intervals. All of the coefficients are depicted over a four year horizon.

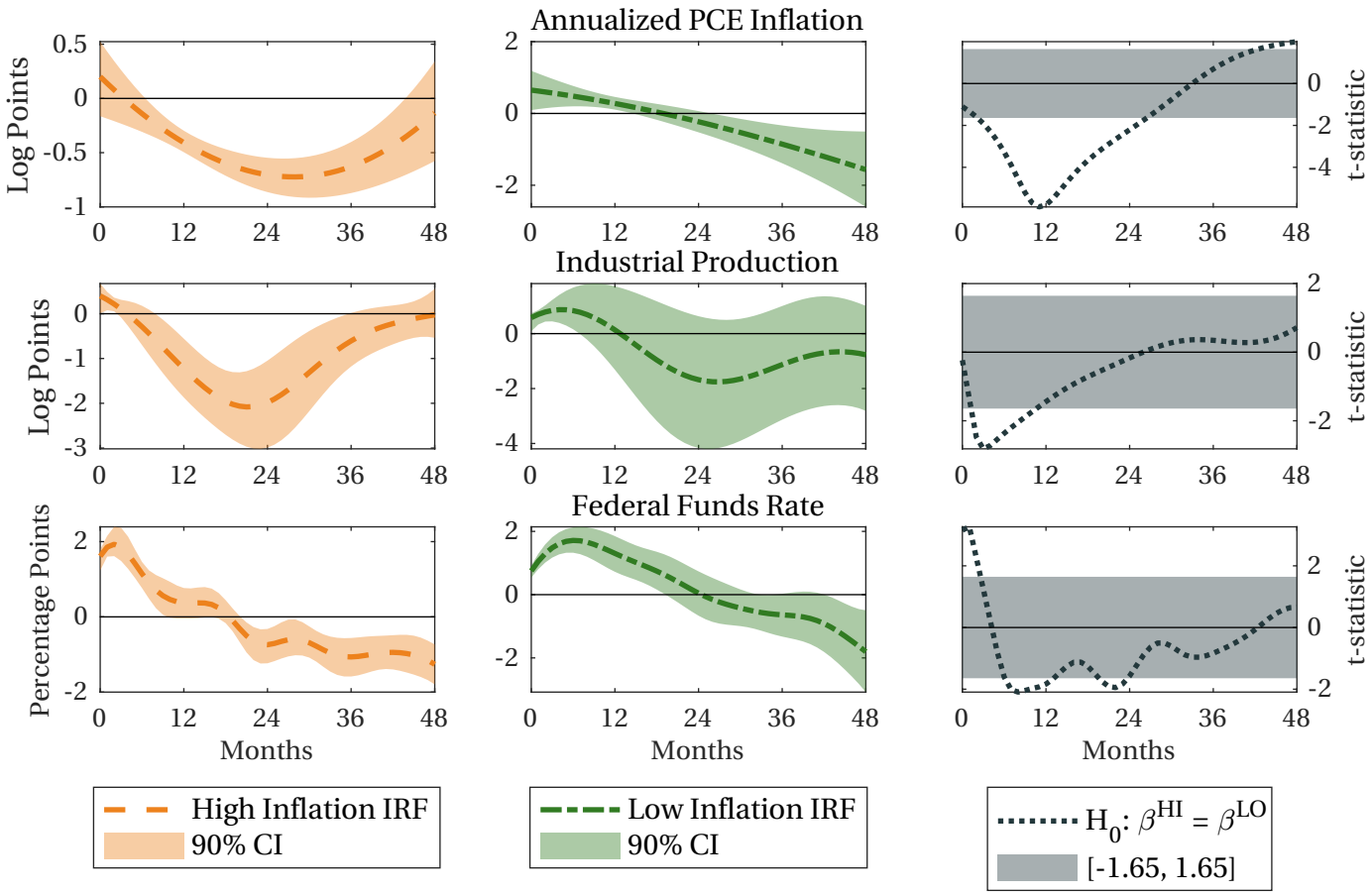


Fig. A.17: Smooth impulse response functions in different inflation states, controlling for the Gilchrist & Zakrajsek (2012) index. The three response variables are annualized PCE Inflation (first row), industrial output (second row) and the federal funds rate (third row). The first column depicts the point estimates of the high inflation impulse response (dashed line) together with its 90% confidence interval. The second column depicts the point estimates of the low inflation impulse response (dashed-dotted line) together with its 90% confidence interval. The third column depicts the t-statistic for the null hypothesis of equality of the high and low inflation impulse responses (dotted line), together with the 90% z-values (grey area). All of the coefficients are depicted over a four year horizon.

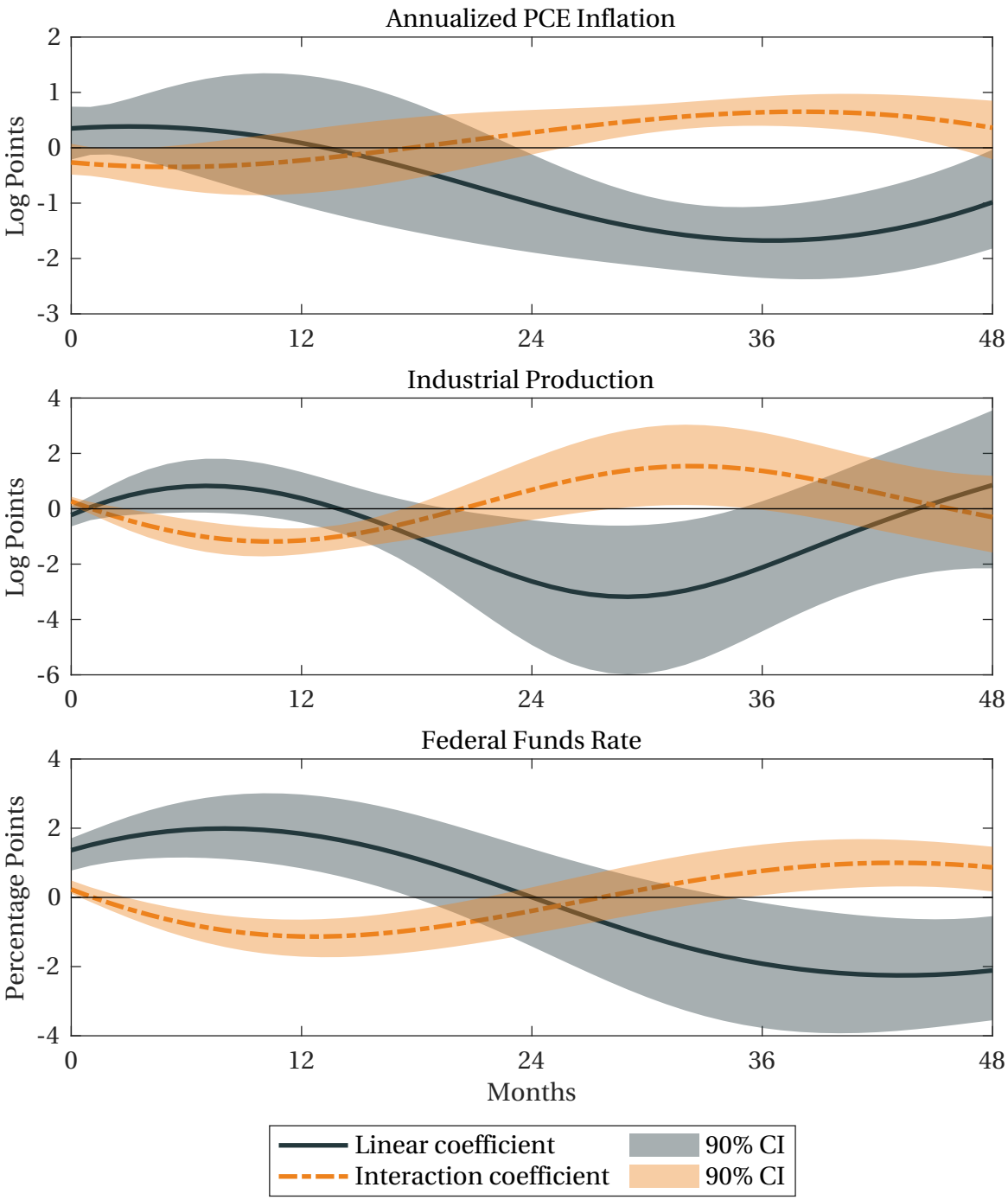


Fig. A.18: Smooth local projection coefficients, using non-linearly identified Romer and Romer shocks. The three response variables are annualized PCE Inflation, industrial output and the federal funds rate. Every panel depicts both the point estimates of the linear coefficient (solid line) and the absolute value interaction coefficient (dashed-dotted), together with their 90% confidence intervals. All of the coefficients are depicted over a four year horizon.

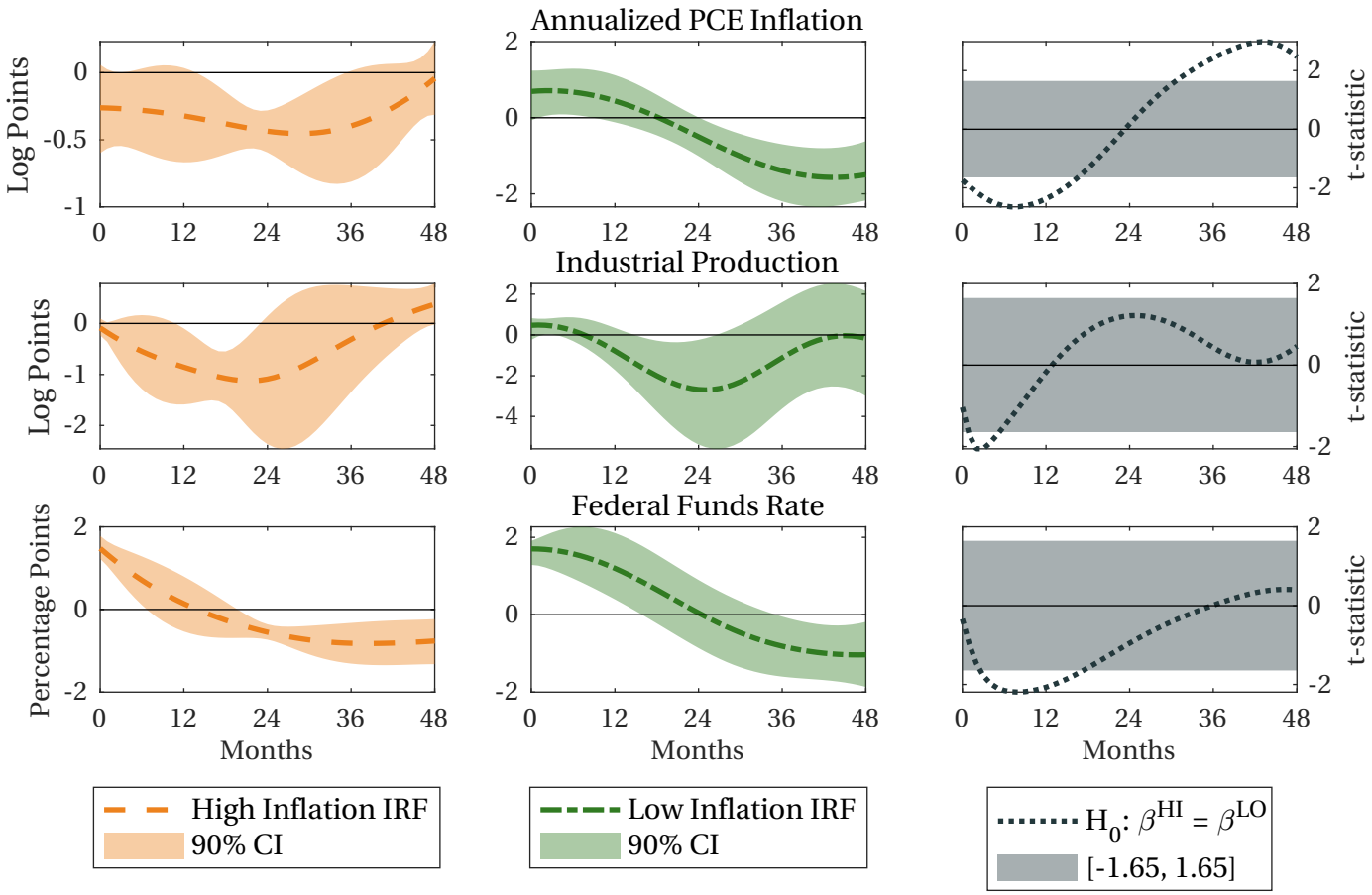


Fig. A.19: Smooth impulse response functions in different inflation states, using non-linearly identified Romer and Romer shocks. The three response variables are annualized PCE Inflation (first row), industrial output (second row) and the federal funds rate (third row). The first column depicts the point estimates of the high inflation impulse response (dashed line) together with its 90% confidence interval. The second column depicts the point estimates of the low inflation impulse response (dashed-dotted line) together with its 90% confidence interval. The third column depicts the t-statistic for the null hypothesis of equality of the high and low inflation impulse responses (dotted line), together with the 90% z-values (grey area). All of the coefficients are depicted over a four year horizon.

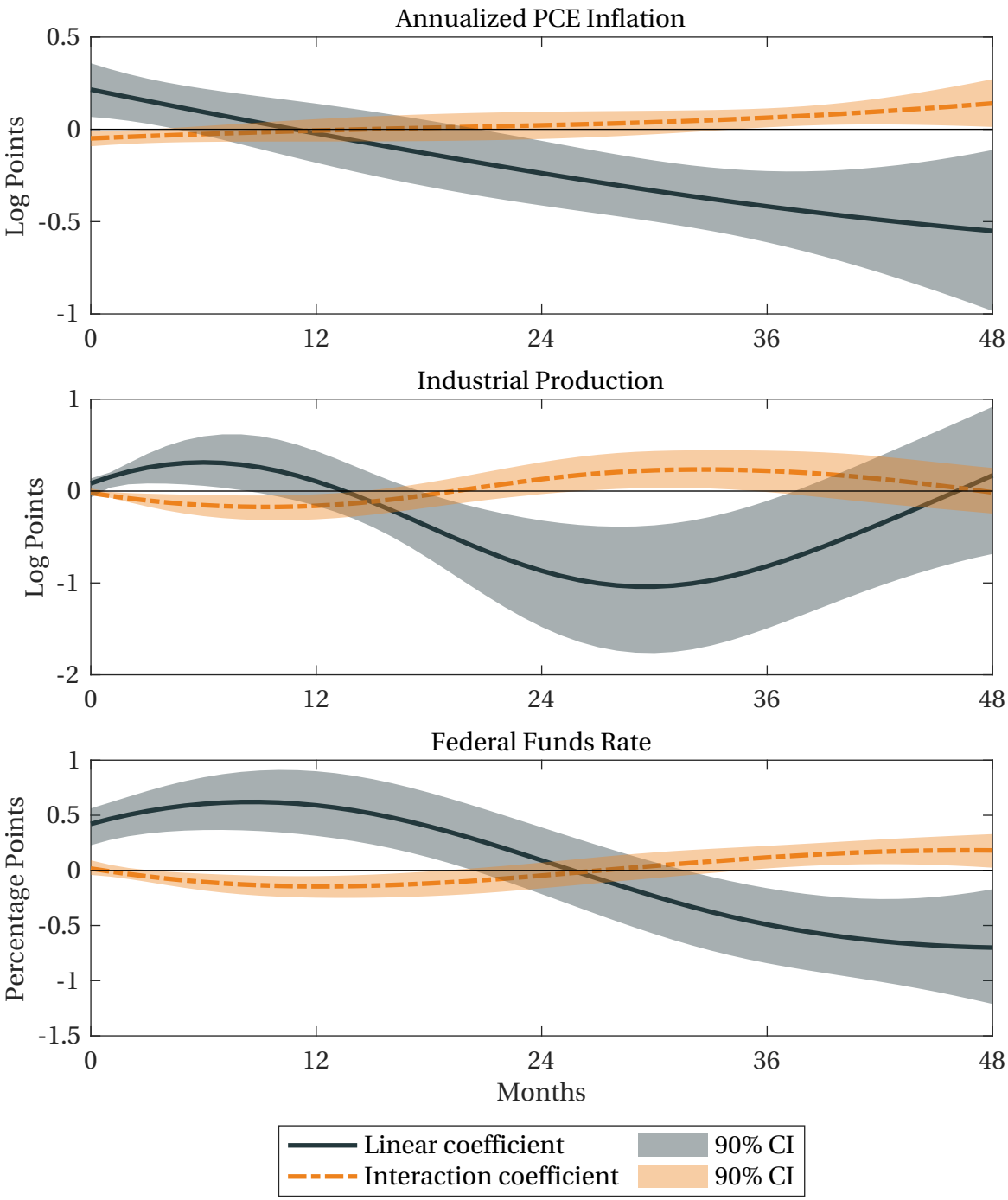


Fig. A.20: Smooth local projection coefficients, using shocks identified from a Smooth Transition VAR. The three response variables are annualized PCE Inflation, industrial output and the federal funds rate. Every panel depicts both the point estimates of the linear coefficient (solid line) and the absolute value interaction coefficient (dashed-dotted), together with their 90% confidence intervals. All of the coefficients are depicted over a four year horizon.

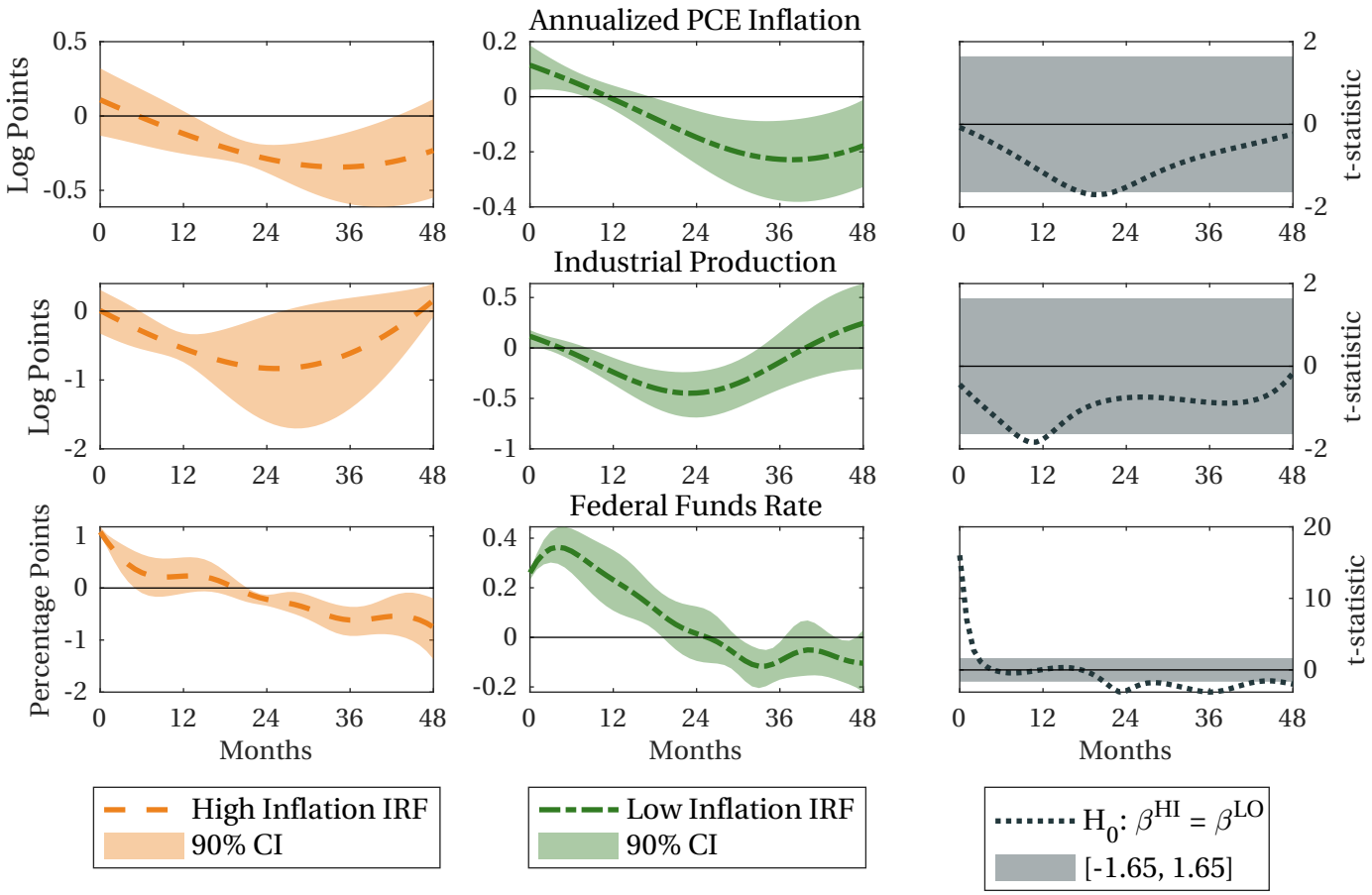


Fig. A.21: Smooth impulse response functions in different inflation states, using shocks identified from a Smooth Transition VAR. The three response variables are annualized PCE Inflation (first row), industrial output (second row) and the federal funds rate (third row). The first column depicts the point estimates of the high inflation impulse response (dashed line) together with its 90% confidence interval. The second column depicts the point estimates of the low inflation impulse response (dashed-dotted line) together with its 90% confidence interval. The third column depicts the t-statistic for the null hypothesis of equality of the high and low inflation impulse responses (dotted line), together with the 90% z-values (grey area). All of the coefficients are depicted over a four year horizon.

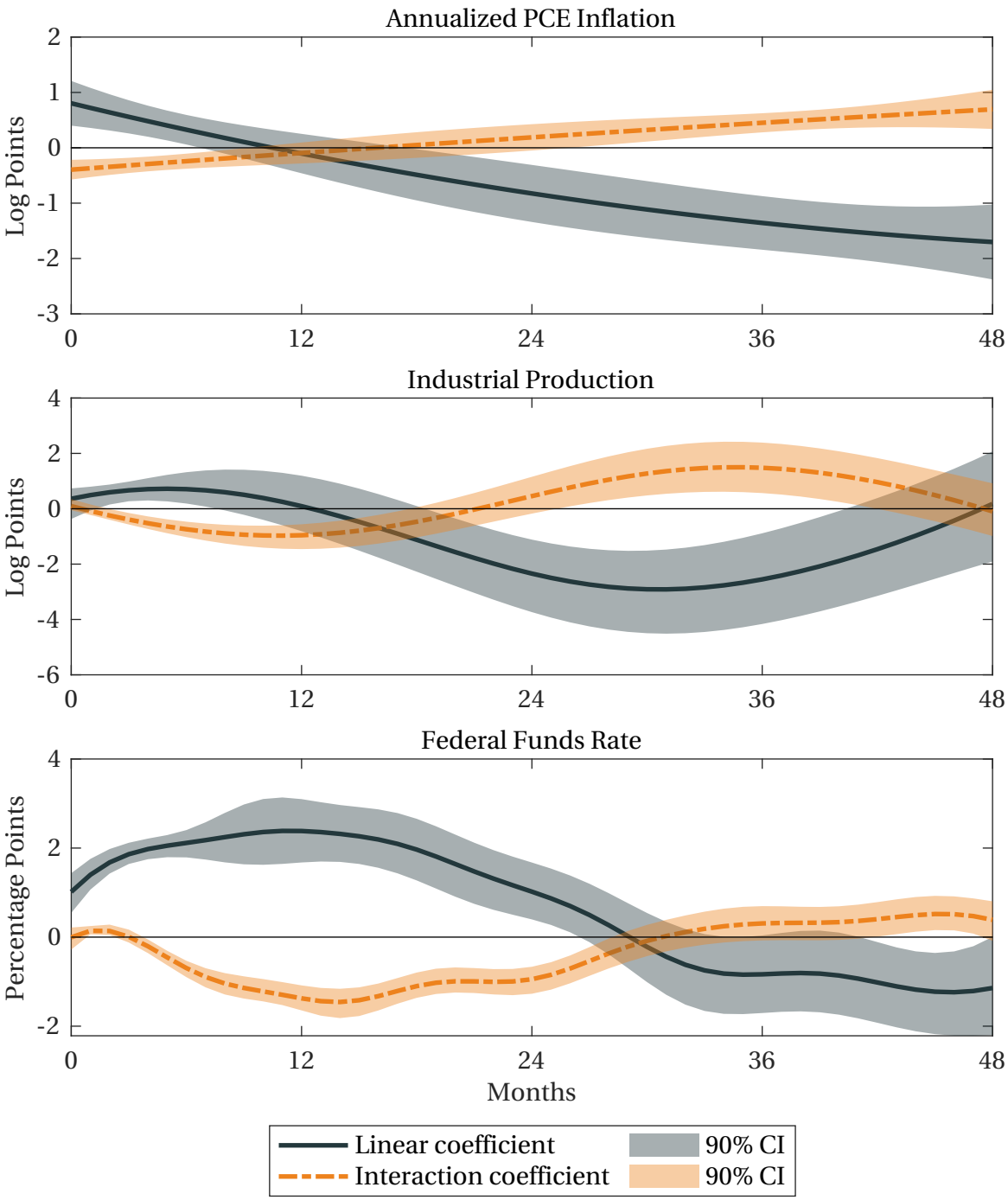


Fig. A.22: Smooth local projection coefficients, controlling for one lead and one lag of the shock itself. The three response variables are annualized PCE Inflation, industrial output and the federal funds rate. Every panel depicts both the point estimates of the linear coefficient (solid line) and the absolute value interaction coefficient (dashed-dotted), together with their 90% confidence intervals. All of the coefficients are depicted over a four year horizon.

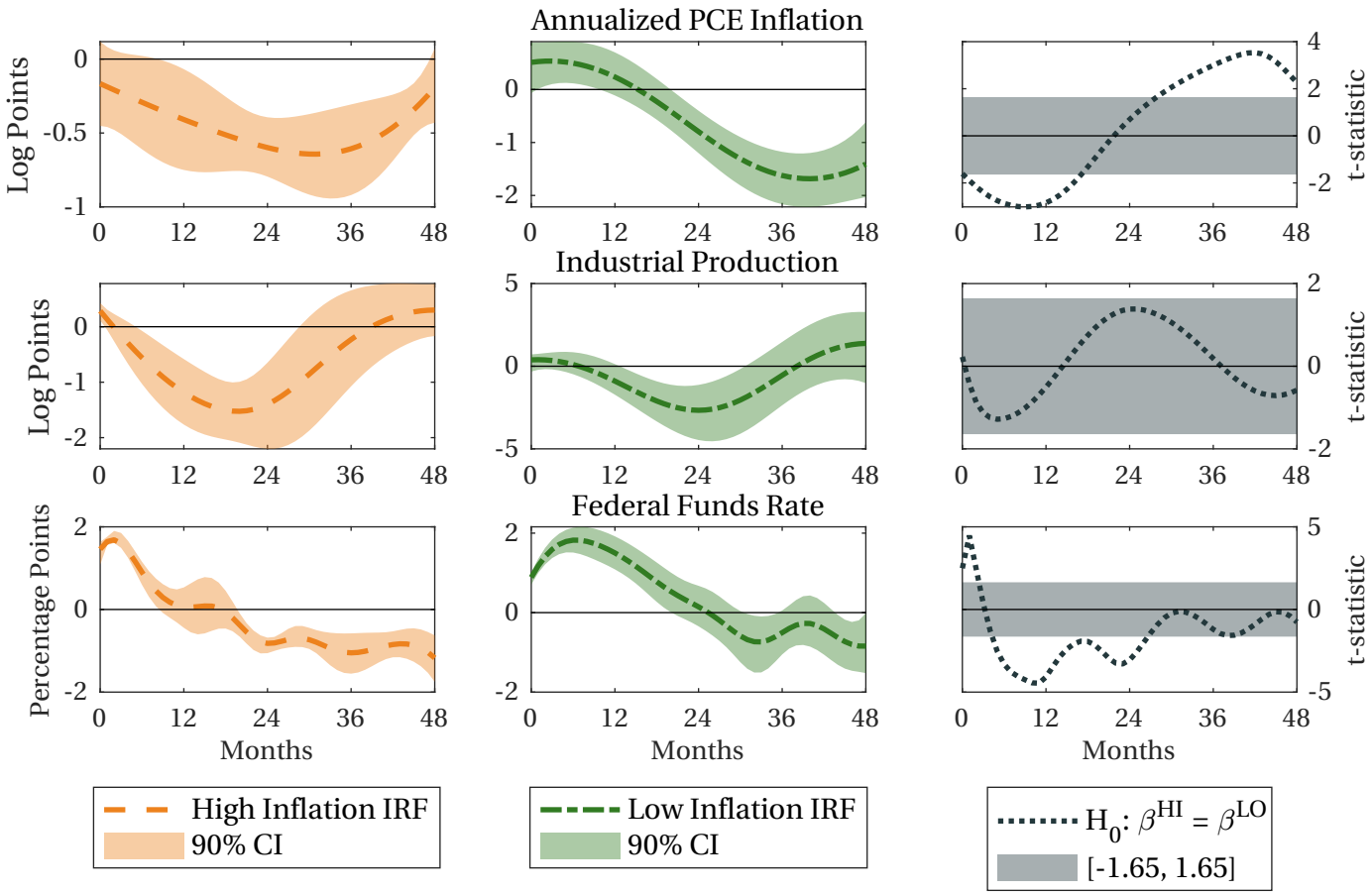


Fig. A.23: Smooth impulse response functions in different inflation states, controlling for for one lead and one lag of the shock itself. The three response variables are annualized PCE Inflation (first row), industrial output (second row) and the federal funds rate (third row). The first column depicts the point estimates of the high inflation impulse response (dashed line) together with its 90% confidence interval. The second column depicts the point estimates of the low inflation impulse response (dashed-dotted line) together with its 90% confidence interval. The third column depicts the t-statistic for the null hypothesis of equality of the high and low inflation impulse responses (dotted line), together with the 90% z-values (grey area). All of the coefficients are depicted over a four year horizon.

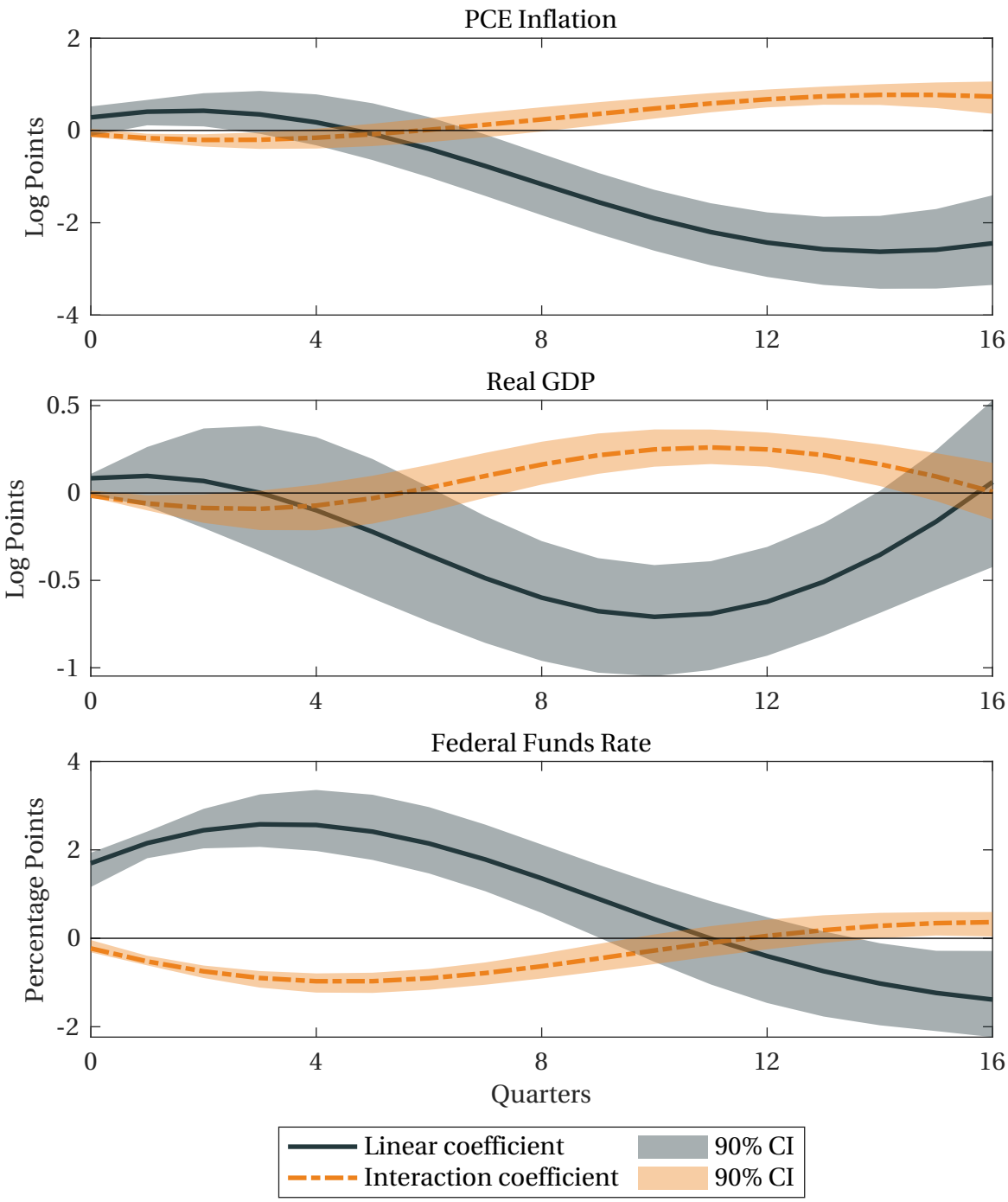


Fig. A.24: Smooth local projection coefficients, quarterly estimation, using real GDP as the output measure. The three response variables are annualized PCE Inflation, real GDP and the federal funds rate. Every panel depicts both the point estimates of the linear coefficient (solid line) and the absolute value interaction coefficient (dashed-dotted), together with their 90% confidence intervals. All of the coefficients are depicted over a four year horizon.

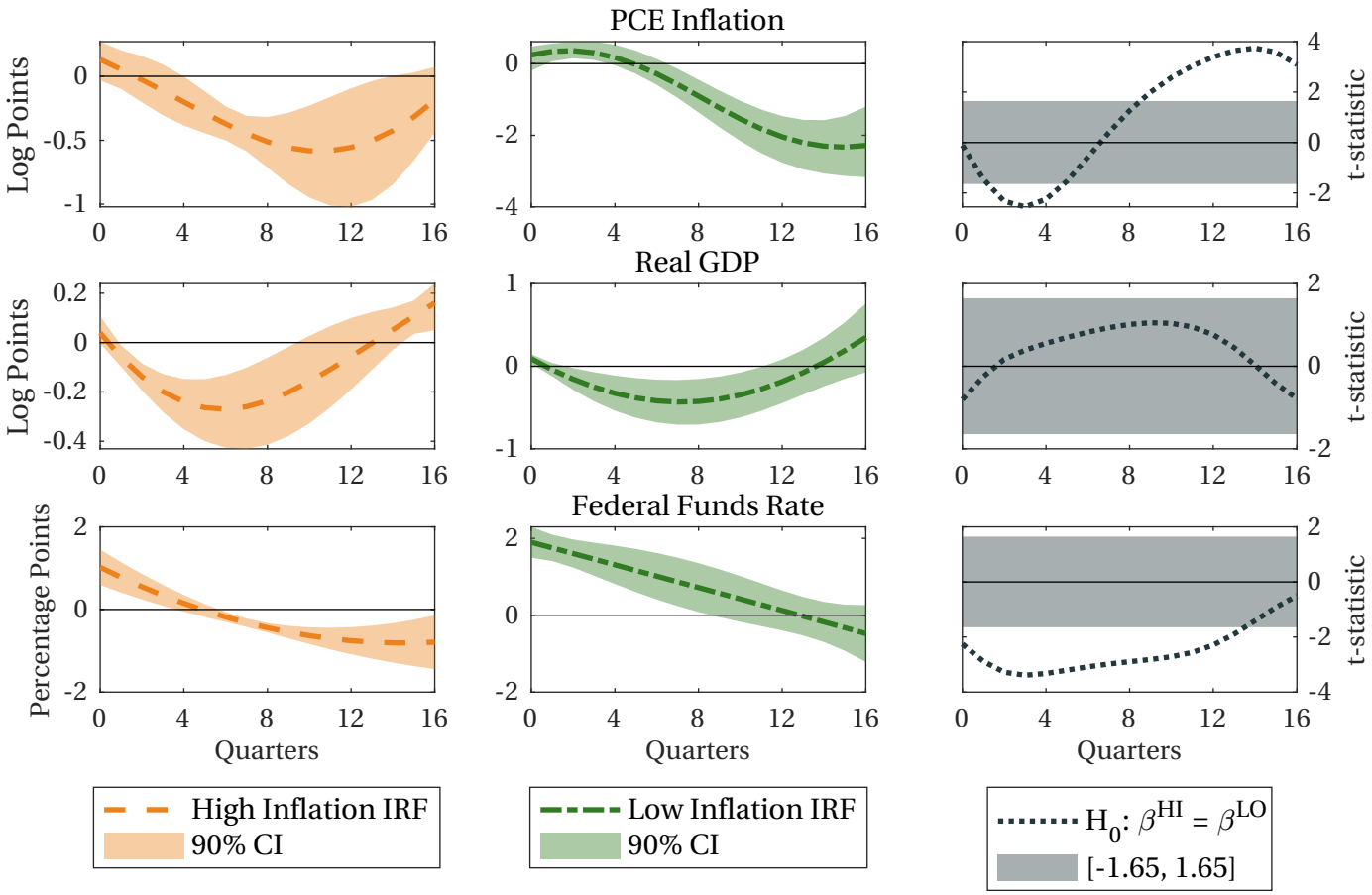


Fig. A.25: Smooth impulse response functions in different inflation states, quarterly estimation, using real GDP as the output measure. The three response variables are annualized PCE Inflation (first row), real GDP (second row) and the federal funds rate (third row). The first column depicts the point estimates of the high inflation impulse response (dashed line) together with its 90% confidence interval. The second column depicts the point estimates of the low inflation impulse response (dashed-dotted line) together with its 90% confidence interval. The third column depicts the t-statistic for the null hypothesis of equality of the high and low inflation impulse responses (dotted line), together with the 90% z-values (grey area). All of the coefficients are depicted over a four year horizon.

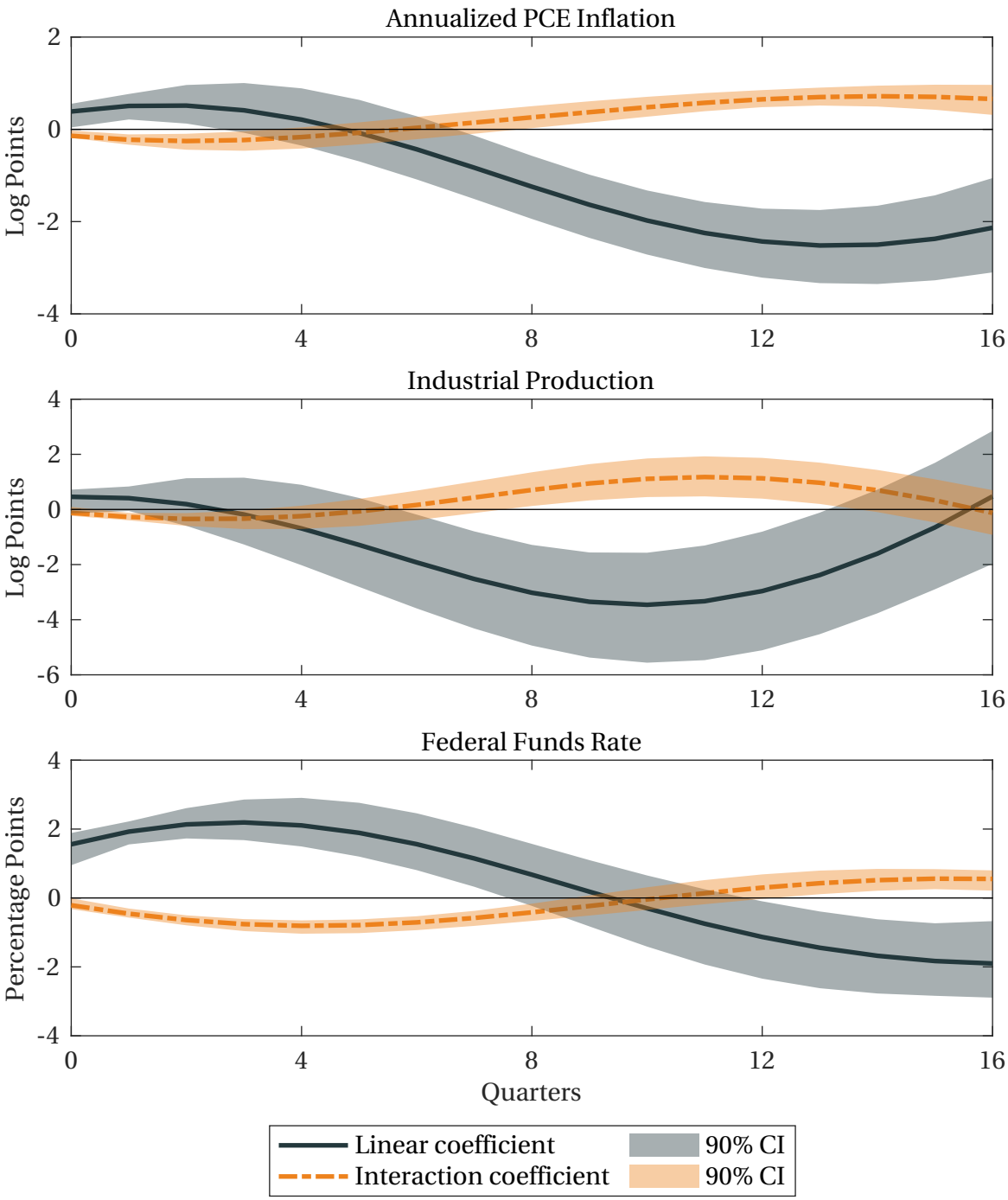


Fig. A.26: Smooth local projection coefficients, quarterly estimation, controlling for fiscal policy with the Fisher & Peters (2010) shocks. The three response variables are annualized PCE Inflation, industrial output and the federal funds rate. Every panel depicts both the point estimates of the linear coefficient (solid line) and the absolute value interaction coefficient (dashed-dotted), together with their 90% confidence intervals. All of the coefficients are depicted over a four year horizon.

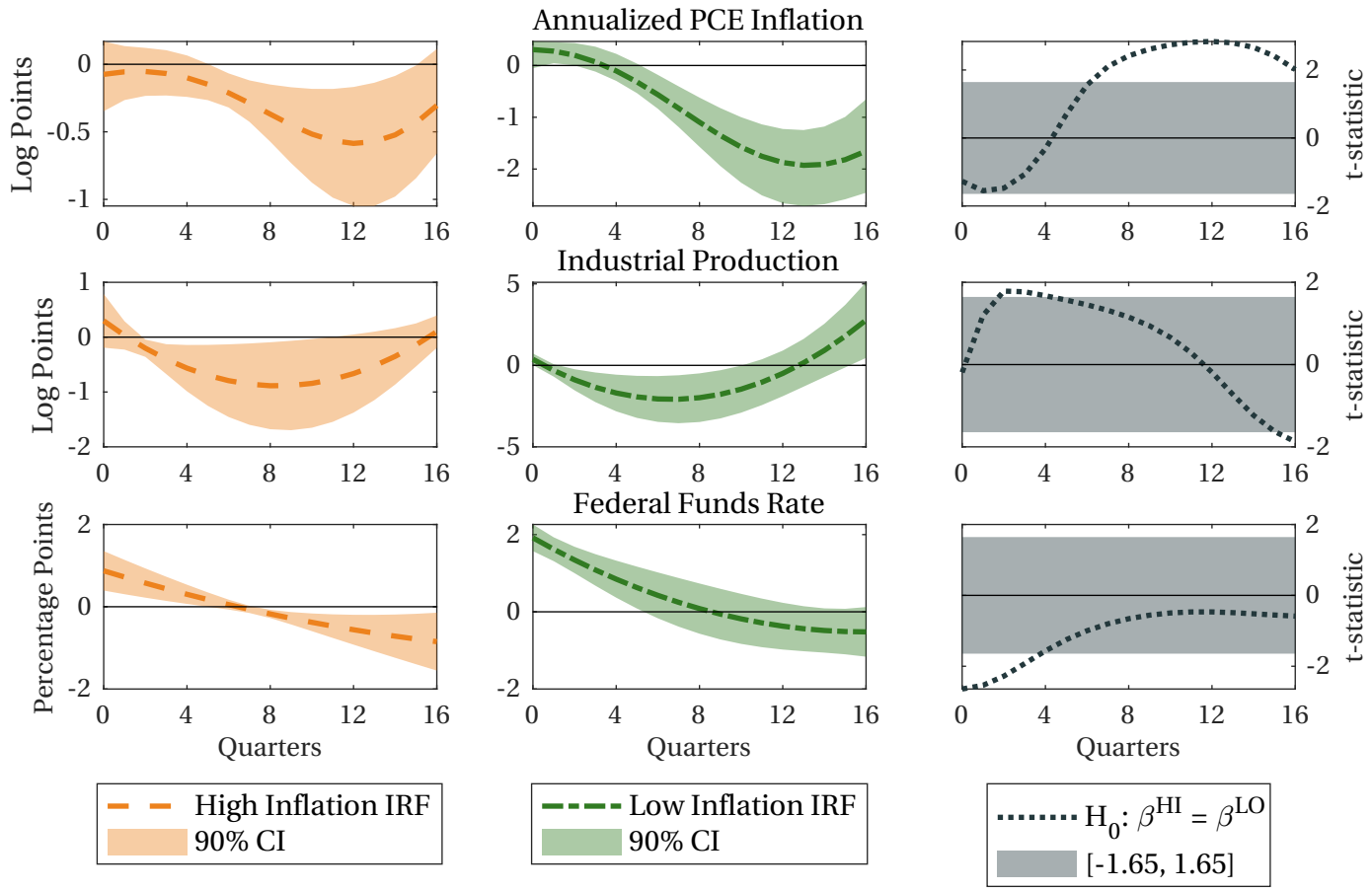


Fig. A.27: Smooth impulse response functions in different inflation states, controlling for quarterly fiscal shocks (Fisher & Peters, 2010). The three response variables are annualized PCE Inflation (first row), industrial output (second row) and the federal funds rate (third row). The first column depicts the point estimates of the high inflation impulse response (dashed line) together with its 90% confidence interval. The second column depicts the point estimates of the low inflation impulse response (dashed-dotted line) together with its 90% confidence interval. The third column depicts the t-statistic for the null hypothesis of equality of the high and low inflation impulse responses (dotted line), together with the 90% z-values (grey area). All of the coefficients are depicted over a four year horizon.

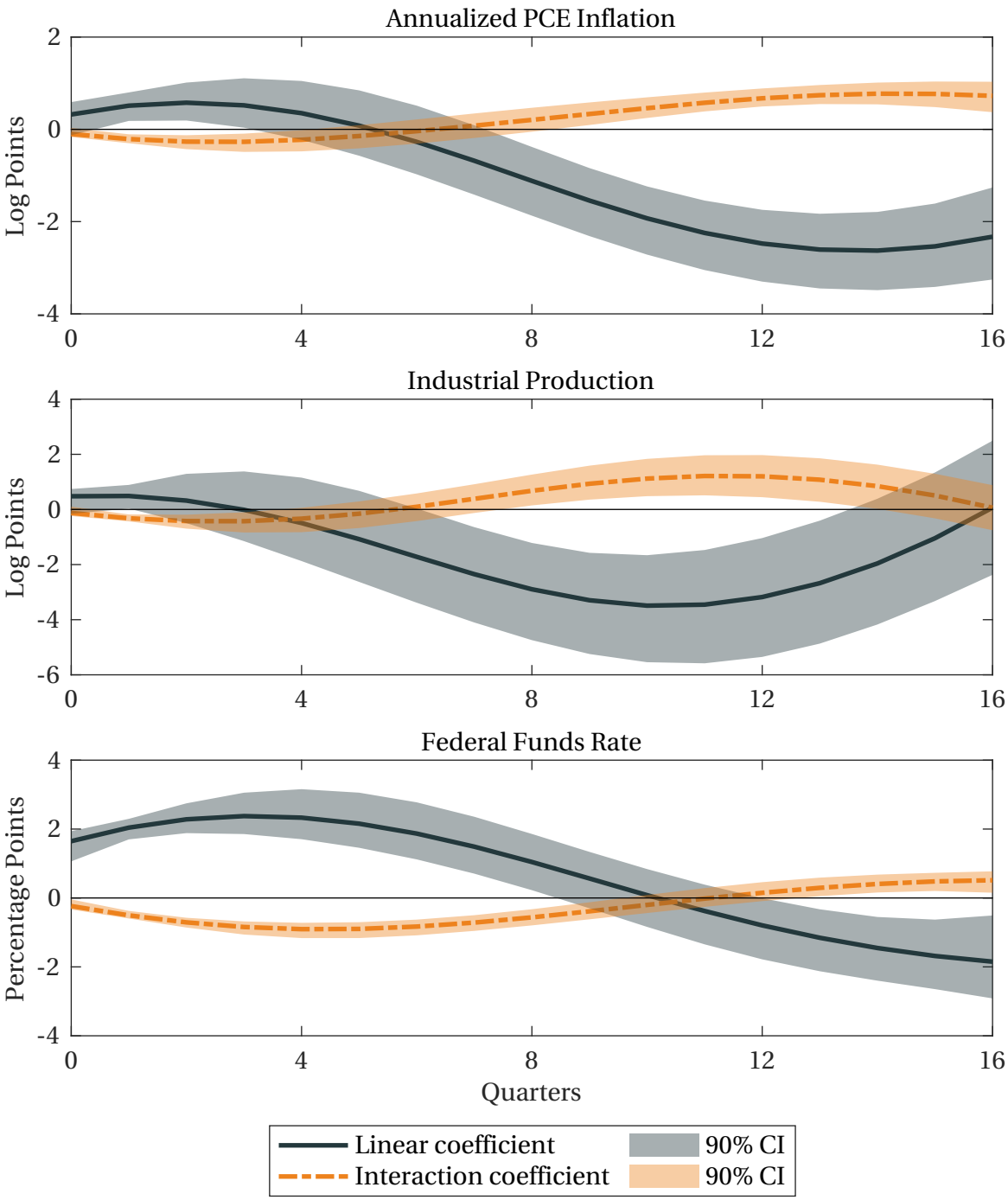


Fig. A.28: Smooth local projection coefficients, quarterly estimation, controlling for fiscal policy with the Romer & Romer (2010) exogenous tax series. The three response variables are annualized PCE Inflation, industrial output and the federal funds rate. Every panel depicts both the point estimates of the linear coefficient (solid line) and the absolute value interaction coefficient (dashed-dotted), together with their 90% confidence intervals. All of the coefficients are depicted over a four year horizon.

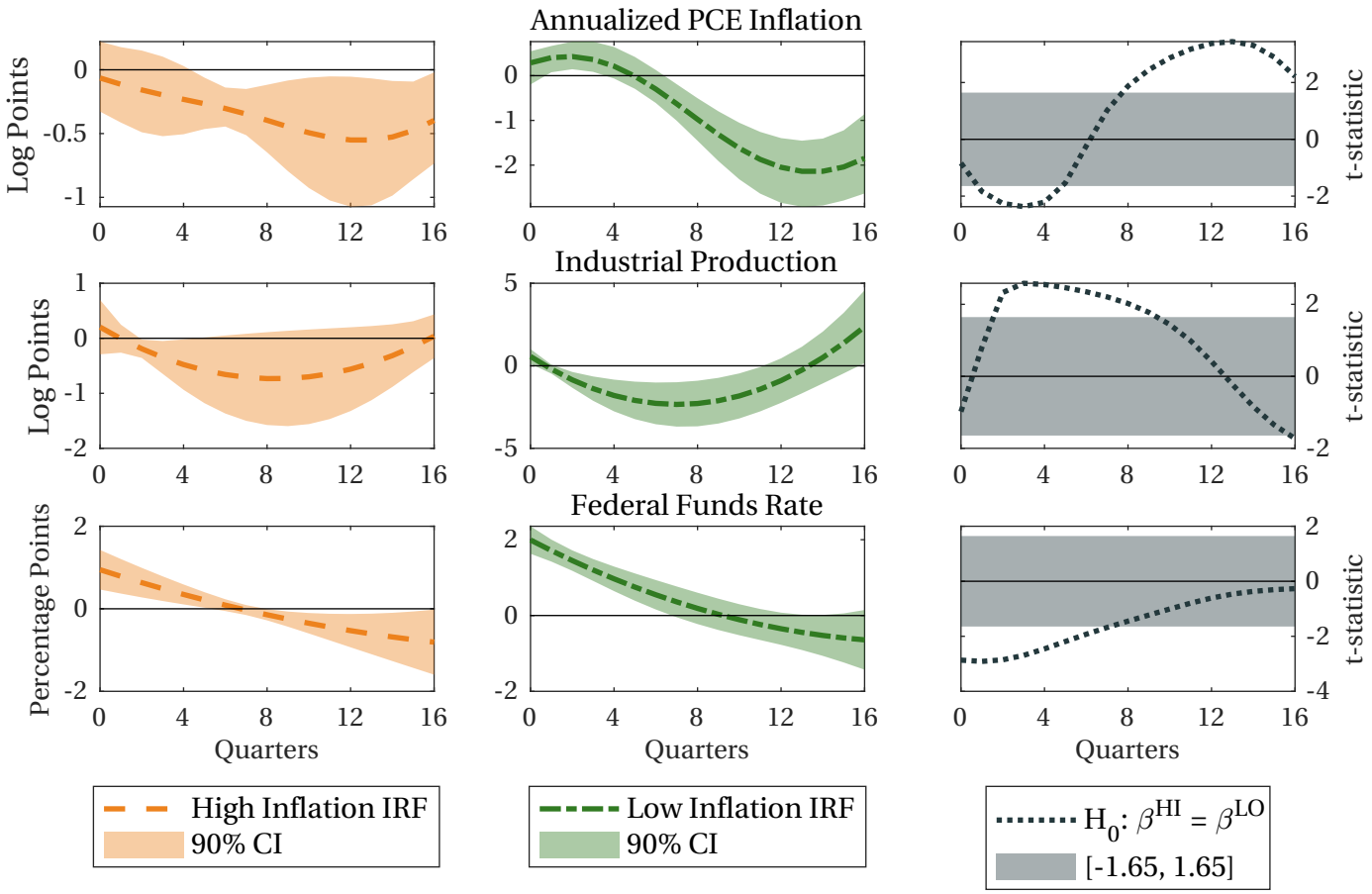


Fig. A.29: Smooth impulse response functions in different inflation states, controlling for quarterly tax shocks (Romer & Romer, 2010). The three response variables are annualized PCE Inflation (first row), industrial output (second row) and the federal funds rate (third row). The first column depicts the point estimates of the high inflation impulse response (dashed line) together with its 90% confidence interval. The second column depicts the point estimates of the low inflation impulse response (dashed-dotted line) together with its 90% confidence interval. The third column depicts the t-statistic for the null hypothesis of equality of the high and low inflation impulse responses (dotted line), together with the 90% z-values (grey area). All of the coefficients are depicted over a four year horizon.

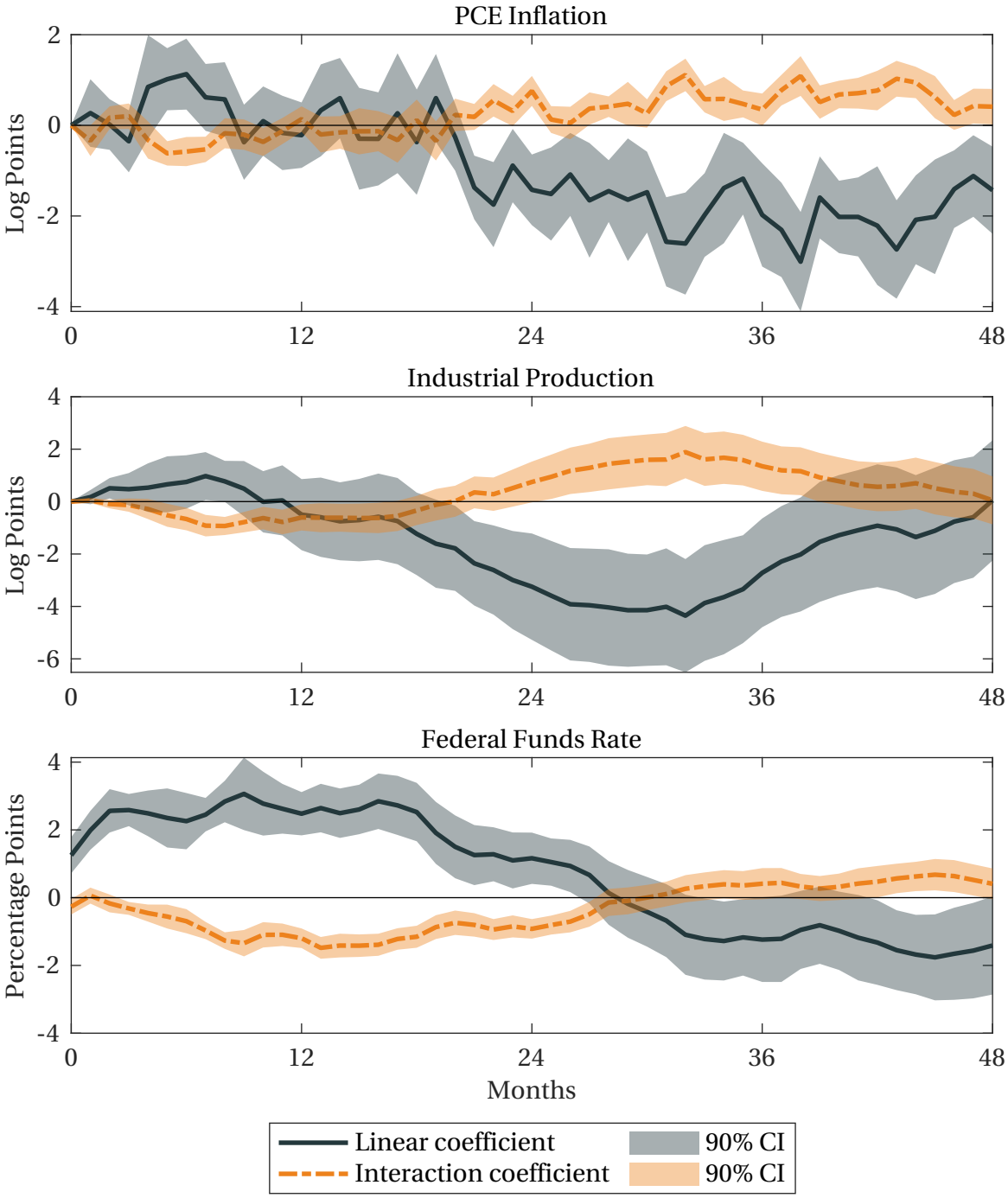


Fig. A.30: Local projection coefficients, unsmoothed. The three response variables are annualized PCE Inflation, industrial output and the federal funds rate. Every row depicts both the point estimates of the linear coefficient (solid line) and the absolute value interaction coefficient (dashed-dotted), together with their 90% confidence intervals. All of the coefficients are depicted over a four year horizon.

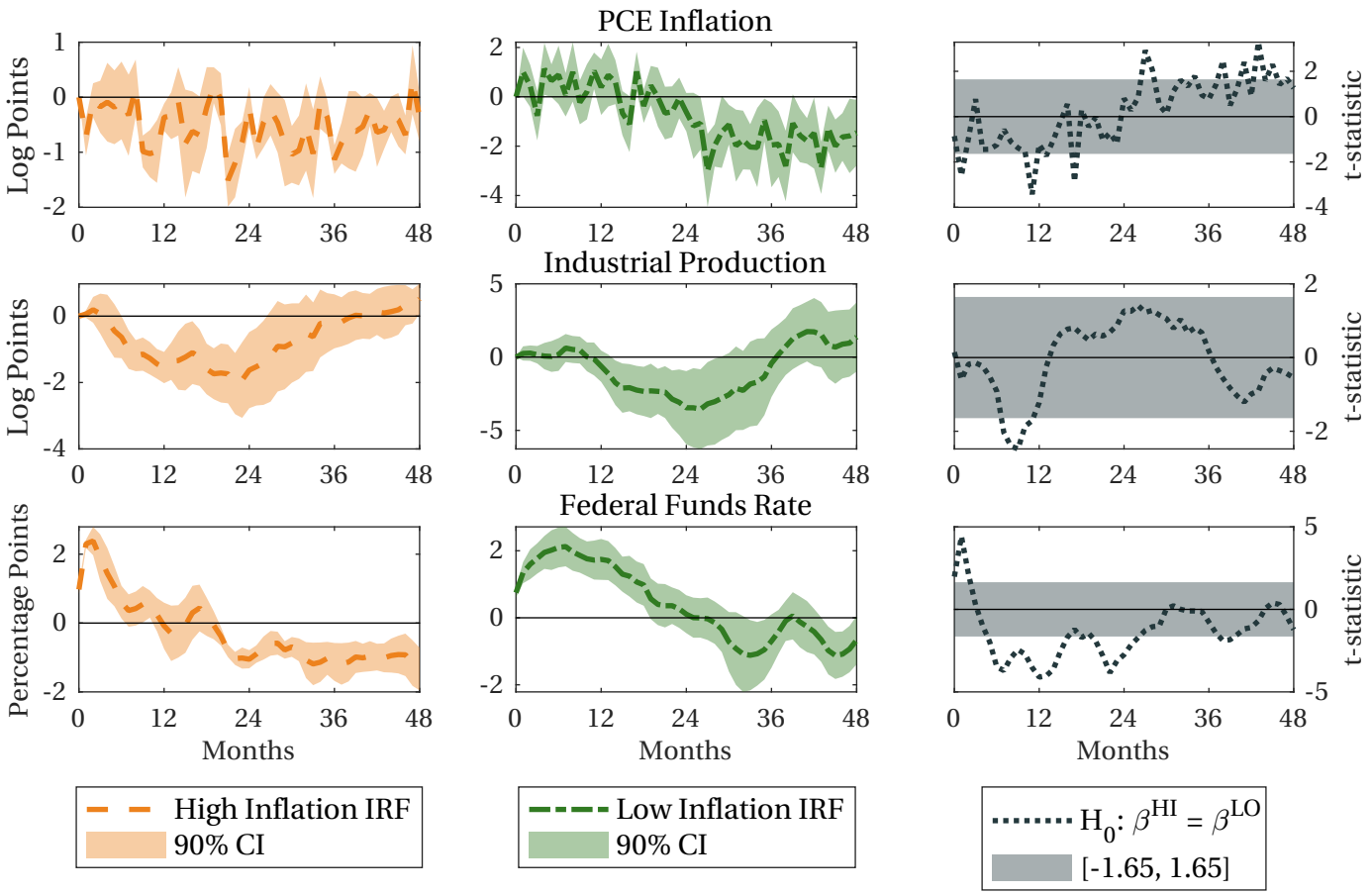


Fig. A.31: Impulse response functions in different inflation states, unsmoothed. The three response variables are annualized PCE Inflation (first row), industrial output (second row) and the federal funds rate (third row). The first column depicts the point estimates of the high inflation impulse response (dashed line) together with its 90% confidence interval. The second column depicts the point estimates of the low inflation impulse response (dashed-dotted line) together with its 90% confidence interval. The third column depicts the t-statistic for the null hypothesis of equality of the high and low inflation impulse responses (dotted line), together with the 90% z-values (grey area). All of the coefficients are depicted over a four year horizon.

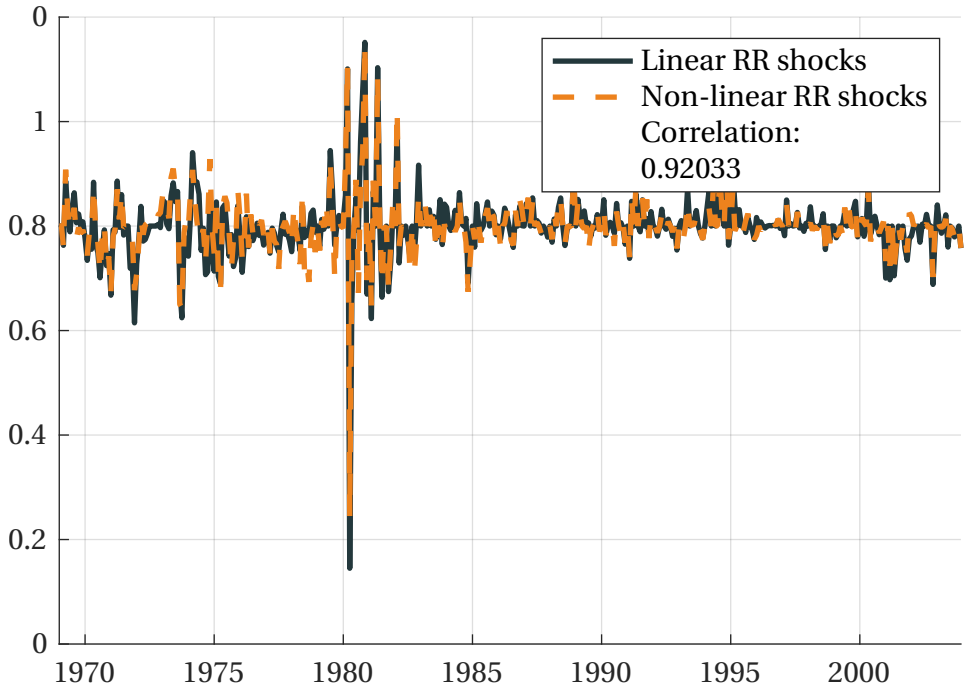


Fig. A.32: Time series of linear and non-linear Romer and Romer shocks. The linear shocks are indicated by the solid line, whereas the non-linear shocks are indicated by the dashed line. The correlation coefficient between two shock series is equal to 0.92, as indicated in the top right hand corner.

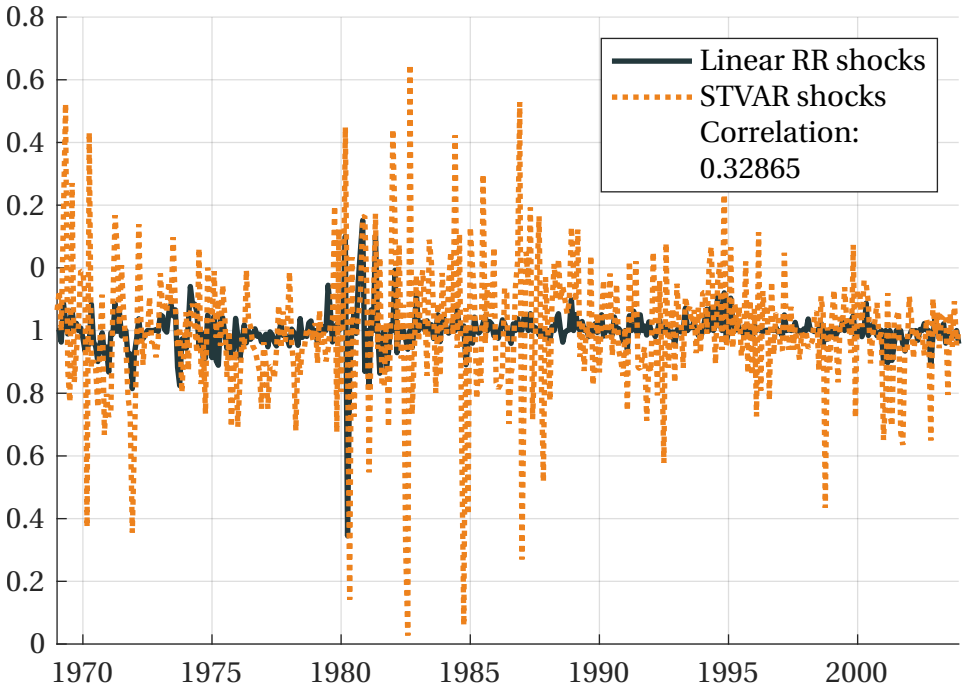


Fig. A.33: Time series of linear Romer and Romer and STVAR shocks. The linear shocks are indicated by the solid line, whereas the STVAR shocks are indicated by the dashed-dotted line. The correlation coefficient between the two shock series is equal to 0.32, as indicated in the top right hand corner.

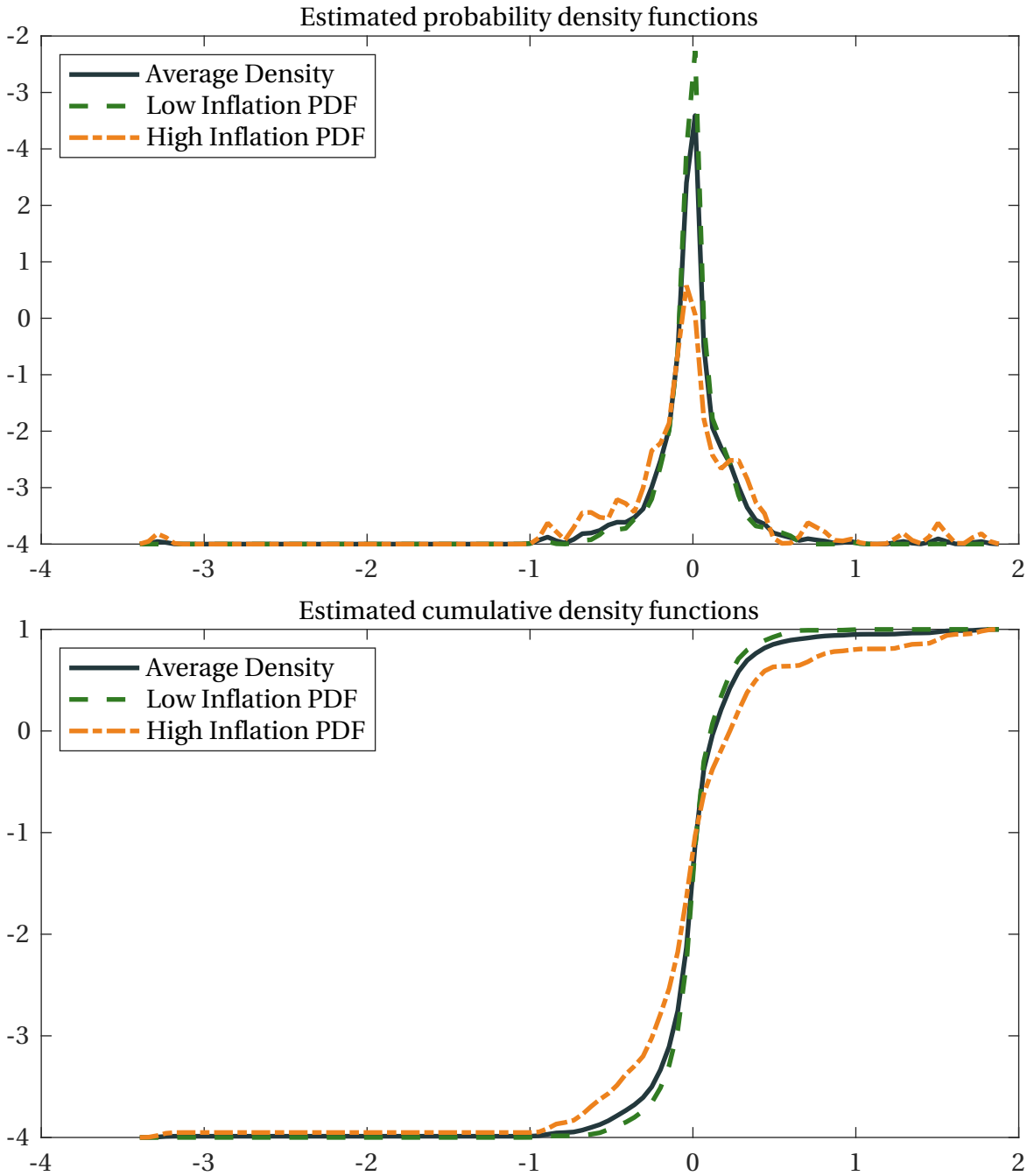


Fig. A.34: Distribution of shocks over high and low trend inflation regimes: Panel A shows the estimated probability density functions whilst Panel B shows the estimated cumulative density function. The average density is indicated by the solid line, the low inflation density with the dotted line and the high inflation density with the dashed-dotted line.

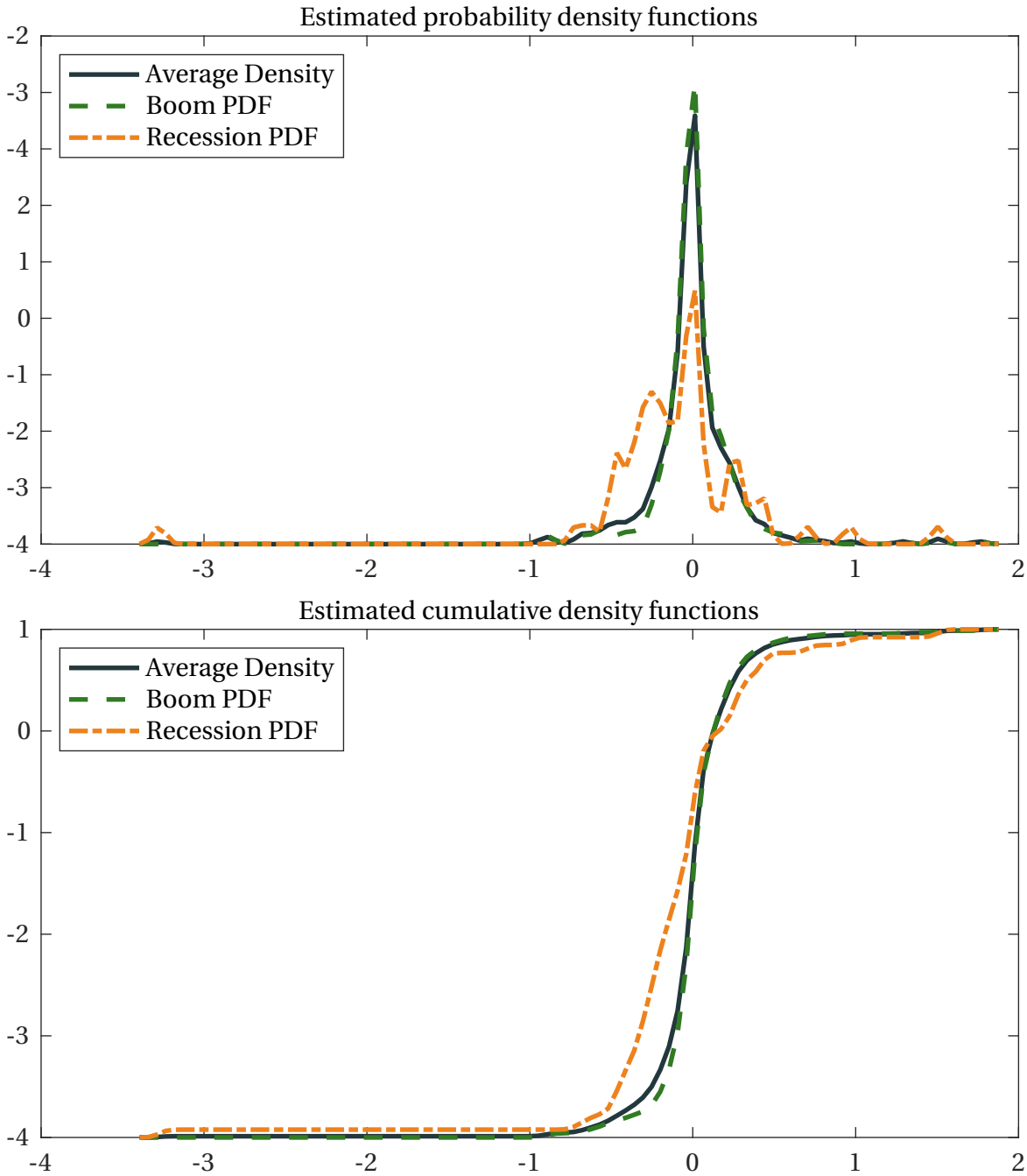


Fig. A.35: Distribution of shocks over booms and recessions: Panel A shows the estimated probability density functions whilst Panel B shows the estimated cumulative density functions. The average density is indicated by the solid line, the expansion density with the dotted line and the recession density with the dashed-dotted line.

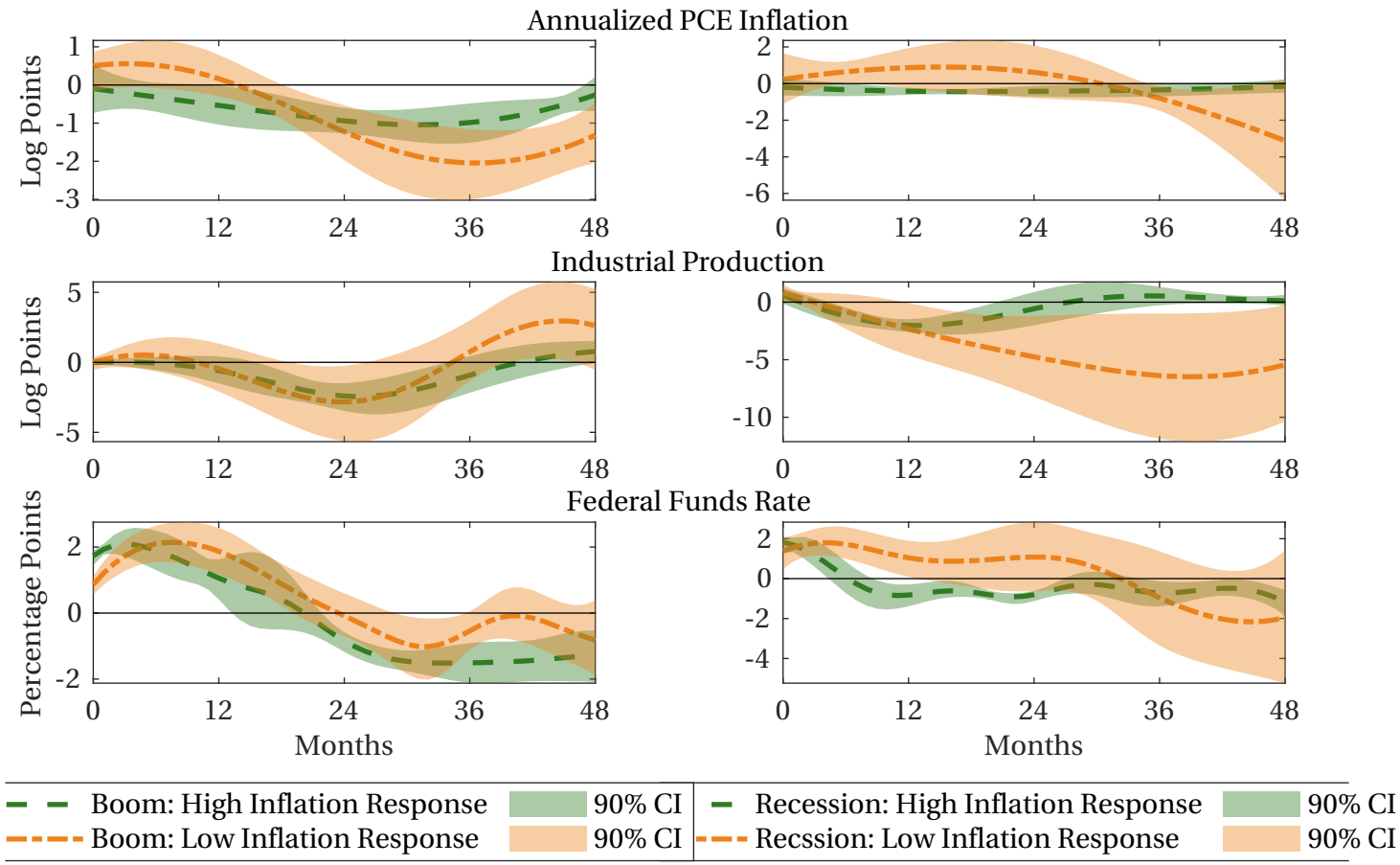


Fig. A.36: Smooth impulse response functions in different inflation states and in different business cycles states, ie. recessions and expansions, as defined by the NBER recession dates. The three response variables are annualized PCE Inflation (first row), industrial output (second row) and the federal funds rate (third row). The first column depicts the point estimates of the high (dashed line) and low (dashed-dotted line) inflation impulse response together with its 90% confidence interval in economic expansions. The second column depicts the point estimates of the high (dashed line) and low (dashed-dotted line) inflation impulse response together with its 90% confidence interval in economic recessions. All of the coefficients are depicted over a four year horizon.

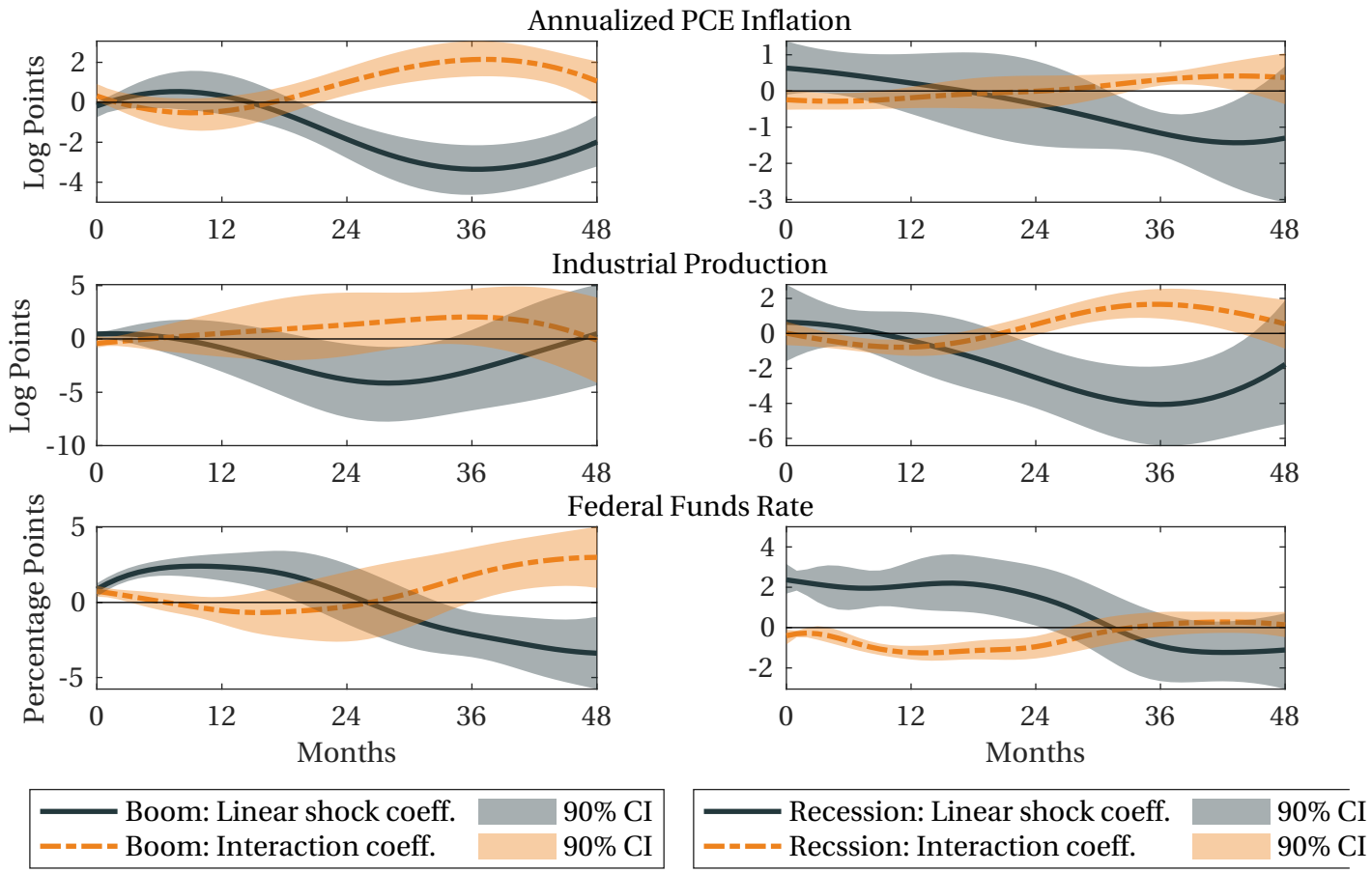


Fig. A.37: Smooth local projection coefficients, in both expansions and recessions, as defined by the NBER recession dates. The three response variables are annualized PCE Inflation, industrial output and the federal funds rate. The first column depicts both the point estimates of the linear coefficient (solid line) and the absolute value interaction coefficient (dashed-dotted), together with their 90% confidence intervals in economic expansions. The second column depicts both the point estimates of the linear coefficient (solid line) and the absolute value interaction coefficient (dashed-dotted), together with their 90% confidence intervals in economic recessions. All of the coefficients are depicted over a four year horizon.

A.9 Tables

Shock summary statistics

Statistic	Linear R&R Shock	Non-linear R&R Shock	STVAR Shock
Mean	-0.004	-0.002	0.006
Median	0	0	0.029
Std. dev.	0.312	0.288	1.073
Min.	-3.275	-2.776	-4.863
Max.	1.758	1.735	4.246
AR(1) Coefficient	0.084	0.009	-0.034

Table A.1: Summary statistics for the monetary policy shocks used in the analysis

Δ IP Local Projection - Hansen (1992) test statistic

Horizon	$h = 1$	$h = 3$	$h = 6$	$h = 12$	$h = 24$	$h = 36$
Linear coeff.	0.11	0.09	0.17	0.14	0.22	0.20
Interaction coeff.	0.02	0.03	0.01	0.03	0.05	0.19
Joint: all coeffs.	3.02^{**}	4.04^{***}	4.38^{***}	4.26^{***}	3.64^{***}	3.68^{***}

Table A.2: Estimated Hansen (1992) test statistics for parameter constancy of the change of industrial production local projection with both a linear and an absolute value interaction shock term. The first row reports the individual test statistic for the for the linear coefficient $\hat{\beta}$ at different horizons, the second row reports those for the absolute value interaction coefficient $\hat{\zeta}$ and the final row reports the test statistic for the joint hypothesis of all parameters (ie. regression coefficients and variance) to be constant.
Individual critical values are $c_{1,1\%} = 0.75$, $c_{1,5\%} = 0.47$ and $c_{1,10\%} = 0.35$. Joint critical values for a model with K = 12 parameters are $c_{12,1\%} = 3.51$, $c_{12,5\%} = 2.96$ and $c_{12,10\%} = 2.69$.
***: Significant at the 1% level; **: Significant at the 5% level; *: Significant at the 10% level;

Δ FFR Local Projection - Hansen (1992) test statistic

Horizon	$h = 1$	$h = 3$	$h = 6$	$h = 12$	$h = 24$	$h = 36$
Linear coeff.	0.13	0.05	0.03	0.11	0.11	0.18
Interaction coeff.	0.02	0.01	0.01	0.03	0.05	0.14
Joint: all coeffs.	3.36**	2.28	2.45	4.19***	3.66***	2.78*

Table A.3: Estimated Hansen (1992) test statistics for parameter constancy of the change of the federal funds rate local projection with both a linear and an absolute value interaction shock term. The first row reports the individual test statistic for the linear coefficient $\hat{\beta}$ at different horizons, the second row reports those for the absolute value interaction coefficient $\hat{\zeta}$ and the final row reports the test statistic for the joint hypothesis of all parameters (ie. regression coefficients and variance) to be constant.
Individual critical values are $c_{1,1\%} = 0.75$, $c_{1,5\%} = 0.47$ and $c_{1,10\%} = 0.35$. Joint critical values for a model with K = 12 parameters are $c_{12,1\%} = 3.51$, $c_{12,5\%} = 2.96$ and $c_{12,10\%} = 2.69$.
***: Significant at the 1% level; **: Significant at the 5% level; *: Significant at the 10% level

Sector	Expenditure share (1960-2015)
Durable goods	
Motor vehicles and parts	0.053
Furnishings and durable household equipment	0.036
Recreational goods and vehicles	0.029
Other durable goods	0.016
Nondurable goods	
Food and beverages purchased for off-premises consumption	0.117
Clothing and footwear	0.054
Gasoline and other energy goods	0.037
Other nondurable goods	0.078
Services	
Housing excluding gas and electric utilities	0.153
Gas and electric utilities	0.025
Health care	0.114
Transportation services	0.032
Recreation services	0.029
Food services and accommodations	0.064
Financial services and insurance	0.063
Other services	0.081
Final consumption expenditures of nonprofit institutions serving households (NPISHs)	0.020

Table A.4: The seventeen components of the PCE Price Index used in this study and their expenditure shares, from Stock & Watson (2016)

B

Appendix for Chapter 2

B.1 Variable definitions

Central Credit Responsibility Database (Central de Responsabilidades de Crédito)

Identifier (tina): Anonymized tax identification number.

Global Credit (valor_global): is the sum of effective credit and potential credit, representing the total available credit that a firm accesses.

Effective Credit (valor_efectivo): is credit effectively used in a regular situation, i.e., without payment delays as defined in the respective contract. Examples of effective responsibilities are:

- Loans for the acquisition of financial instruments (shares, bonds, etc.);
- Discount and other credits secured by effects;
- Overdrafts on bank accounts;
- Leasing and factoring;
- Used amounts of credit cards.

Potential Credit (valor_potencial): represents irrevocable commitments of the participating entities. Banco de Portugal requires all credit-granting institutions to report to the CCR their outstanding loan exposure by instrument of all irrevocable credit obligations. Examples of potential responsibilities are:

- Unused amounts of credit cards;
- Lines of credit;
- Guarantees provided by participating entities;
- Guarantees and guarantees given in favor of the participating entities;
- Any other credit facilities likely to be converted into effective debts.

Overdue Credit (valor_vencido): All outstanding credit exposures recorded as non-performing (including overdue, written off, renegotiated credit, overdue credit in litigation, and written off credit in litigation) are aggregated to calculate overdue credits. It includes principal, interest and related fees.

Short-term Credit (valor_curto): Short-term credit is calculated using two different definitions. In the first place, short-term credit is defined based on the term-to-maturity as agreed in the credit contract, denoted by valor_curto_o. Specifically, short-term credit has original maturity of equal to or less than one year. Before 2009, the CCR dataset did not streamline credit exposure based on the maturity structure. Therefore, for the data before 2009, the short-term credit is defined as the aggregation of commercial credit, discount funding, and other short-term funding, which are short-term funding by their nature. In the second place, short-term credit is defined based on residual maturity – the remaining time until the expiration or the repayment of the instrument, denoted by valor_curto_r. Specifically, it is credit with residual maturity of equal to or less than one year. This variable is only available from 2009 onwards. Potential credit is excluded for both calculations.

Long-term Credit (valor_longo): Similar to short-term credit, long-term credit is defined based on original and residual maturities. More precisely, long-term credit is credit with an original or residual maturity of more than one year, denoted by valor_longo_o and valor_longo_r, respectively. Long-term credit defined on an original maturity basis

(valor_longo_o) for the data before 2009 is the aggregation of total credit excluding commercial credit (type 1), discount funding (type 2), and other short-term funding (type 3). Potential credit is excluded for both calculations.

B.2 Additional tables

Table B.1: Descriptive statistics of Portuguese firms between 2006 and 2017

Variable	Mean	Median	Std. Dev.	Size group median			
				<90th	90th-99th	99-99.5th	>99.5th
Total Assets	3.15	0.28	85.10	0.25	5.06	42.71	135.70
Turnover	1.86	0.23	33.59	0.21	3.25	19.93	27.94
Potential credit	0.19	0.03	4.56	0.03	0.14	0.95	2.95
Effective credit	0.53	0.04	5.96	0.04	1.15	6.93	126.73
Leverage	0.28	0.20	0.38	0.20	0.24	0.17	0.08
Liquidity ratio	0.14	0.06	0.19	0.06	0.02	0.01	0.01
Age	15.01	12.00	12.26	12.00	21.00	23.00	21.50
Employees	14.47	4.00	130.58	4.00	25.00	95.00	98.00
# Banks	2.45	2.00	1.89	2.00	4.00	4.00	5.00

Notes. Total assets, turnover, potential credit and effective credit are measured in 2010 Euro Millions.

Table B.2: Linear probability regression: How age, total assets, leverage and liquidity ratio affect the probability of being constrained according to measure I

	Constrained binary			
	(1)	(2)	(3)	(4)
Age	-0.03***			
	(0.000)			
Total assets		-0.06***		
		(0.000)		
Leverage			-0.02***	
			(0.000)	
Liquidity ratio				0.01***
				(0.000)
Constant	0.24***	0.24***	0.24***	0.24***
Observations	1,365,913	1,365,913	1,365,913	1,365,913
R-squared	0.006	0.024	0.015	0.000

Notes. Here we use winsorized response variables at the 99.5th and 0.5th percentile.
Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table B.3: Correlation between different measures of financial constraints

Measure	I	II	III	IV	V
Constrained I	1				
Constrained II	0.298***	1			
Constrained III	0.248***	0.812***	1		
Constrained IV	-0.032***	0.062***	0.052***	1	
Constrained V	0.013***	0.053***	0.050***	0.237***	1

Table B.4: Elasticity of turnover to GDP changes conditional on size bins and financial constraints

	(1)	(2)	(3)	(4)	(5)	(6)
[90, 99] × GDP Growth	-0.00238	-0.00232	-0.00138	-0.00127	-0.00283	-0.00242
	(0.00275)	(0.00276)	(0.00275)	(0.00275)	(0.00283)	(0.00283)
[99, 99.5] × GDP Growth	0.00144	0.00146	0.00248	0.00276	-0.00330	-0.00273
	(0.00957)	(0.00957)	(0.00956)	(0.00957)	(0.00948)	(0.00949)
>99.5 × GDP Growth	0.00335	0.00333	0.00483	0.00531	-0.00676	-0.00630
	(0.0105)	(0.0105)	(0.0105)	(0.0105)	(0.0104)	(0.0104)
Const. Adj. Eff. × GDP Growth		0.150**				
		(0.054)				
Const. Overdue × GDP Growth			1.975***			
			(0.427)			
Const. Overdue Inc. × GDP Growth				1.762***		
				(0.529)		
Const. Maturing × GDP Growth					0.988***	
					(0.272)	
Const. Secured × GDP Growth						0.394
						(0.268)
Industry × GDP Growth FE	Yes	Yes	Yes	Yes	Yes	Yes
Clustering	Firm	Firm	Firm	Firm	Firm	Firm
N	1365463	1365463	1365463	1365463	1122067	1122067

Notes. Estimates report the semi-elasticity of turnover with respect to GDP changes. Constrained measures are constructed as documented in Section 2.2.1. All specifications contain a constant term and non-interacted indicators. Standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table B.5: Elasticity of turnover to idiosyncratic TFP shocks conditional on size bins and financial constraints

	(1)	(2)	(3)	(4)	(5)	(6)
TFP	0.522***	0.520***	0.511***	0.515***	0.529***	0.534***
	(0.00416)	(0.00429)	(0.00426)	(0.00420)	(0.00460)	(0.00463)
[90, 99] × TFP	0.0211**	0.0220**	0.0254**	0.0235**	0.0273**	0.0279**
	(0.00818)	(0.00819)	(0.00818)	(0.00818)	(0.00920)	(0.00924)
[99, 99.5] × TFP	-0.0297	-0.0286	-0.0255	-0.0268	-0.0178	-0.0202
	(0.0205)	(0.0205)	(0.0205)	(0.0205)	(0.0233)	(0.0235)
>99.5 × TFP	-0.0896**	-0.0880**	-0.0821**	-0.0852**	-0.0606	-0.0647*
	(0.0288)	(0.0288)	(0.0287)	(0.0288)	(0.0320)	(0.0322)
Const. Adj. Eff. × TFP		0.0185***				
		(0.00350)				
Const. Overdue × TFP			0.0957***			
			(0.00662)			
Const. Overdue Inc. × TFP				0.1000***		
				(0.00692)		
Const. Maturing × TFP					0.0976***	
					(0.00543)	
Const. Secured × TFP						0.0720***
						(0.00551)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Industry × Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Clustering	Firm	Firm	Firm	Firm	Firm	Firm
N	1361912	1361912	1361912	1361912	1116635	1116635

Notes. Estimates report the semi-elasticity of turnover with respect to idiosyncratic TFP shocks. Constrained measures are constructed as documented in Section 2.2.1. All specifications contain a constant term and non-interacted indicators. Standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table B.6: Elasticity of turnover to idiosyncratic financial shocks conditional on size bins and measures of financial constraints

	(1)	(2)	(3)	(4)	(5)	(6)
Bank shock	0.00547	-0.0407	-0.0125	-0.00243	-0.0146	-0.0213
	(0.0193)	(0.0229)	(0.0190)	(0.0190)	(0.0258)	(0.0264)
[90, 99] × Bank shock	0.194	0.224	0.199	0.199	0.173	0.179
	(0.118)	(0.119)	(0.118)	(0.118)	(0.156)	(0.155)
[99, 99.5] × Bank shock	0.579	0.618	0.599	0.596	0.0783	0.0813
	(0.722)	(0.721)	(0.723)	(0.722)	(0.601)	(0.601)
>99.5 × Bank shock	-0.281*	-0.237	-0.264*	-0.274*	-0.206	-0.202
	(0.133)	(0.133)	(0.133)	(0.132)	(0.162)	(0.160)
Const. Adj. Eff. × Bank shock		0.106**				
		(0.0393)				
Const. Overdue × Bank shock			0.146			
			(0.146)			
Const. Overdue Inc. × Bank shock				0.0777		
				(0.172)		
Const. Maturing × Bank shock					0.340*	
					(0.148)	
Const. Secured × Bank shock						0.293*
						(0.124)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Industry × Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Clustering	Firm	Firm	Firm	Firm	Firm	Firm
N	1143962	1143962	1143962	1143962	936084	936084

Notes. Estimates report the semi-elasticity of turnover with respect to financial shocks estimated with the Amiti & Weinstein (2018) methodology. Constrained measures are constructed as outlined in Section 2.2.1. All specifications contain a constant term and non-interacted indicators. Standard errors are reported in parentheses.
*** p<0.01, ** p<0.05, * p<0.1

Table B.7: Cyclicality in employees conditional on size bins and measures of financial constraints

	(1)	(2)	(3)	(4)	(5)	(6)
[90, 99] × Δ GDP	0.001***	0.001***	0.001***	0.001***	0.001***	0.002***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
[99, 99.5] × Δ GDP	0.001	0.001	0.002	0.002	0.002	0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
>99.5 × Δ GDP	0.000	0.000	0.000	0.001	-0.000	-0.000
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Const. Adj. Eff. × Δ GDP		-0.852**				
		(0.0301)				
Const. Overdue × Δ GDP			0.361***			
			(0.0587)			
Const. Overdue Inc. × Δ GDP				0.231**		
				(0.0731)		
Const. Maturing × Δ GDP					0.278***	
					(0.0361)	
Const. Secured × Δ GDP						0.122***
						(0.0365)
Industry × Δ GDP FE	Yes	Yes	Yes	Yes	Yes	Yes
Clustering	Firm	Firm	Firm	Firm	Firm	Firm
N	1363763	1363763	1363763	1363763	1121965	1121965

Notes. Estimates report the semi-elasticity of employees with respect to GDP. Constrained measures are constructed as documented in Section 2.2.1. All specifications contain a constant term and non-interacted indicators. Standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table B.8: Cyclicality in turnover conditional on size bins and measures of financial constraints including firm fixed effects

	(1)	(2)	(3)	(4)	(5)	(6)
[90, 99] × Δ GDP	-0.001	-0.001	-0.001	-0.001	-0.000	-0.000
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
[99, 99.5] × Δ GDP	0.005	0.005	0.005	0.005	-0.000	-0.000
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
>99.5 × Δ GDP	0.006	0.006	0.006	0.006	-0.001	-0.001
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Const. Adj. Eff. × Δ GDP		0.245***				
		(0.0571)				
Const. Overdue × Δ GDP			1.694***			
			(0.456)			
Const. Overdue Inc. × Δ GDP				0.915		
				(0.570)		
Const. Maturing × Δ GDP					0.163	
					(0.303)	
Const. Secured × Δ GDP						-0.271
						(0.301)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Industry × Δ GDP FE	Yes	Yes	Yes	Yes	Yes	Yes
Clustering	Firm	Firm	Firm	Firm	Firm	Firm
N	1361996	1361996	1361996	1361996	1116719	1116719

Notes. Estimates report the semi-elasticity of turnover with respect to GDP. Constrained measures are constructed as documented in Section 2.2.1. All specifications contain a constant term and non-interacted indicators. Standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table B.9: Cyclicality in turnover conditional on size bins and measures of financial constraints including time fixed effects

	(1)	(2)	(3)	(4)	(5)	(6)
[90, 99] × Δ GDP	-0.002	-0.002	-0.001	-0.001	-0.003	-0.002
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
[99, 99.5] × Δ GDP	0.001	0.001	0.002	0.003	-0.003	-0.003
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.0010)
>99.5 × Δ GDP	0.003	0.003	0.005	0.005	-0.007	-0.006
	(0.011)	(0.011)	(0.011)	(0.011)	(0.010)	(0.010)
Const. Adj. Eff. × Δ GDP		0.150***				
		(0.0541)				
Const. Overdue × Δ GDP			1.975***			
			(0.427)			
Const. Overdue Inc. × Δ GDP				1.762***		
				(0.529)		
Const. Maturing × Δ GDP					0.988***	
					(0.272)	
Const. Secured Alt. × Δ GDP						0.394
						(0.268)
Industry × Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Clustering	Firm	Firm	Firm	Firm	Firm	Firm
N	1365463	1365463	1365463	1365463	1122067	1122067

Notes. Estimates report the semi-elasticity of turnover with respect to GDP.Constrained measures are constructed as documented in Section 2.2.1. All specifications contain a constant term and non-interacted indicators. Standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table B.10: Cyclicality in turnover conditional on size bins and measures of financial constraints excluding firms that have 0 potential credit in all periods

	(1)	(2)	(3)	(4)	(5)	(6)
[90, 99] × Δ GDP	-0.002	-0.002	-0.002	-0.001	-0.001	-0.001
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
[99, 99.5] × Δ GDP	-0.001	-0.001	-0.001	-0.000	-0.005	-0.005
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
>99.5 × Δ GDP	0.002	0.002	0.003	0.004	-0.007	-0.007
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Const. Adj. Eff. × Δ GDP		-0.00548				
		(0.0696)				
Const. Overdue × Δ GDP			2.109***			
			(0.527)			
Const. Overdue Inc. × Δ GDP				1.446*		
				(0.658)		
Const. Maturing × Δ GDP					1.093***	
					(0.276)	
Const. Secured × Δ GDP						0.524
						(0.281)
Industry × Δ GDP FE	Yes	Yes	Yes	Yes	Yes	Yes
Clustering	Firm	Firm	Firm	Firm	Firm	Firm
N	1190494	1190494	1190494	1190494	980907	980907

Notes. Estimates report the semi-elasticity of turnover with respect to GDP. Constrained measures are constructed as documented in Section 2.2.1. All specifications contain a constant term and non-interacted indicators. Standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table B.11: Cyclicality in turnover conditional on size bins and measures of financial constraints including bank controls

	(1)	(2)	(3)	(4)	(5)	(6)
[90, 99] × Δ GDP	-0.003	-0.003	-0.002	-0.002	-0.004	-0.004
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
[99, 99.5] × Δ GDP	-0.002	-0.002	-0.001	-0.001	-0.007	-0.007
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
>99.5 × Δ GDP	0.002	0.002	0.004	0.005	-0.008	-0.008
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Const. Adj. Eff. × Δ GDP		0.205**				
		(0.0641)				
Const. Overdue × Δ GDP			1.955***			
			(0.483)			
Const. Overdue Inc. × Δ GDP				1.706**		
				(0.594)		
Const. Maturing × Δ GDP					1.036***	
					(0.281)	
Const. Secured × Δ GDP						0.502
						(0.288)
Industry × Δ GDP FE	Yes	Yes	Yes	Yes	Yes	Yes
Clustering	Firm	Firm	Firm	Firm	Firm	Firm
Bank Controls	Yes	Yes	Yes	Yes	Yes	Yes
N	1187112	1187112	1187112	1187112	976408	976408

Notes. Estimates report the semi-elasticity of turnover with respect to GDP. Constrained measures are constructed as documented in Section 2.2.1. All specifications contain a constant term and non-interacted indicators. Standard errors are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table B.12: Parameter values

Parameter	Description	Value		Source
β	Discount factor	0.96		K&T (2013)
α	Returns on capital	0.30		K&T (2013)
η	Returns on labor	0.60		K&T (2013)
δ	Depreciation rate	0.065		K&T (2013)
ψ	Labour preference	2.15		K&T (2013)
π_d	Exogenous probability of exit	0.02		Data
μ_θ	Average: permanent productivity	0		Normalized
μ_w	Average: transitory shock	0		Normalized
Model		Transitory & permanent	Transitory only	
ξ	Collateral constraint	0.57	0.57	Calibrated
σ_θ	Std. dev.: permanent productivity	0.16	-	Calibrated
ρ_w	Persistence of transitory shock	0.07	0.47	Calibrated
σ_w	Std. dev: transitory shock	0.09	0.09	Calibrated
μ_{ke}	Relative size of entrants	0.01	0.09	Calibrated
σ_{ke}	Standard deviation of entrants	0.11	0.12	Calibrated

Notes. K&T (2013) is short for Khan & Thomas (2013). Both models were separately calibrated to find the best match to the data, except for the collateral constraint ξ. See text for explanation.

Table B.13: Calibration fit for the 1 shock model with no restrictions

Moment	Data	Model transitory shock
Percentage of constrained firms	0.23	0.20
Share of constrained firms in bottom 20%	0.33	0.85
Size of 90th-percentile vs. median	9.44	9.08
Size of 90th percentile vs. bottom 20%	30.24	42.65
Size of constrained firms 90th-percentile vs. median	7.35	2.20
Size of unconstrained firms 90th-percentile vs. median	9.67	5.19
Asset share of constrained firms	0.12	0.01
Share of constrained firms in top 10% vs. bottom 20%	0.36	0
Percentage of constrained firms in top 1%	0.09	0

Notes. All moment conditions were equally weighted when minimizing the percentage deviation from the empirical target values.

B.3 Additional figures

B.3.1 Descriptive figures

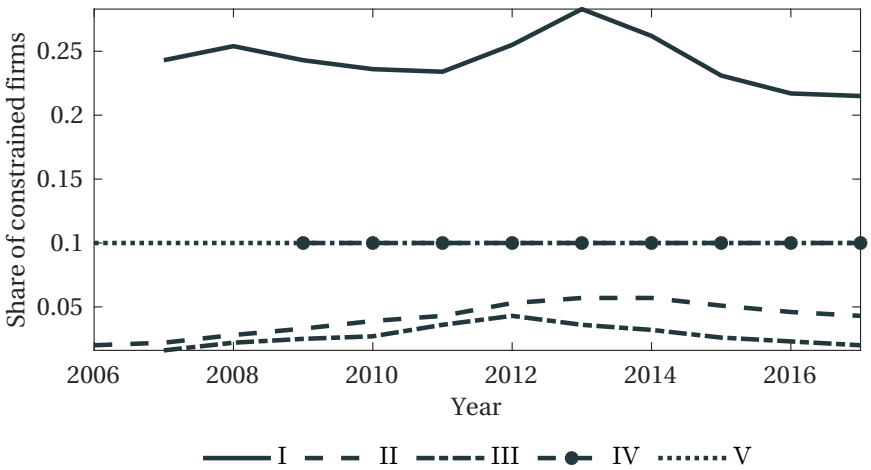


Fig. B.1: Share of constrained firms over time. Measures 1 to 5 as defined in Section 2.2.1

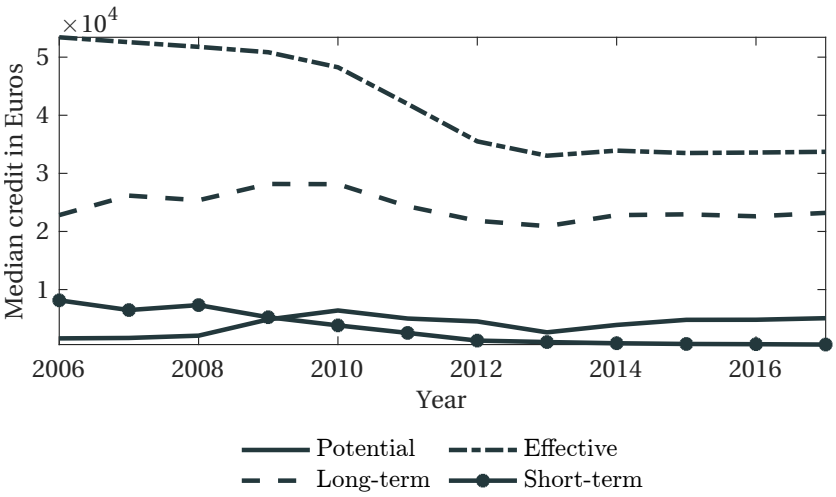


Fig. B.2: Median values for potential, effective, long-term and short-term credit over time.

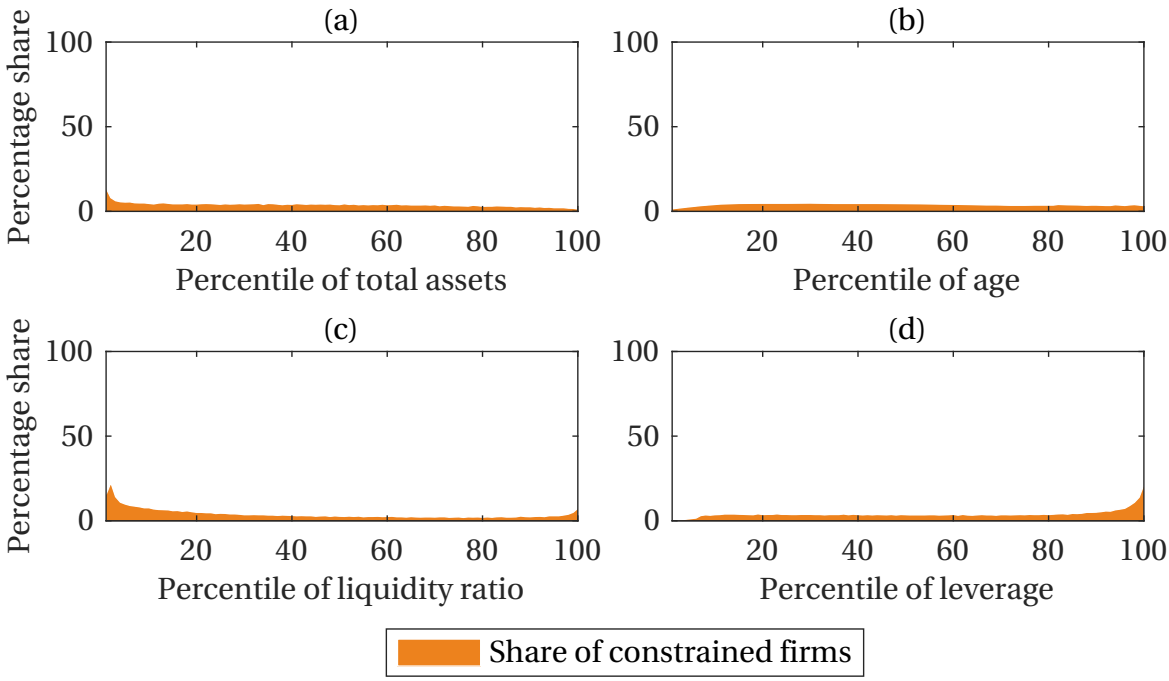


Fig. B.3: Decomposition of constrained and unconstrained firms across percentiles of firm variables using constraint measure II

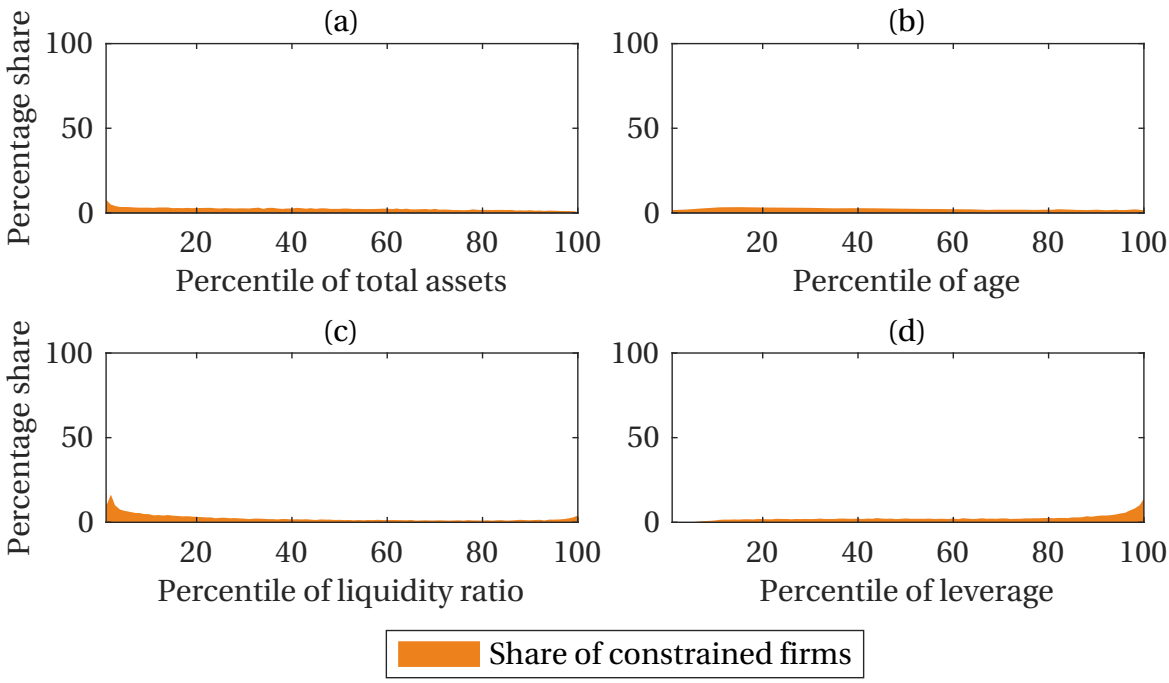


Fig. B.4: Decomposition of constrained and unconstrained firms across percentiles of firm variables using constraint measure III

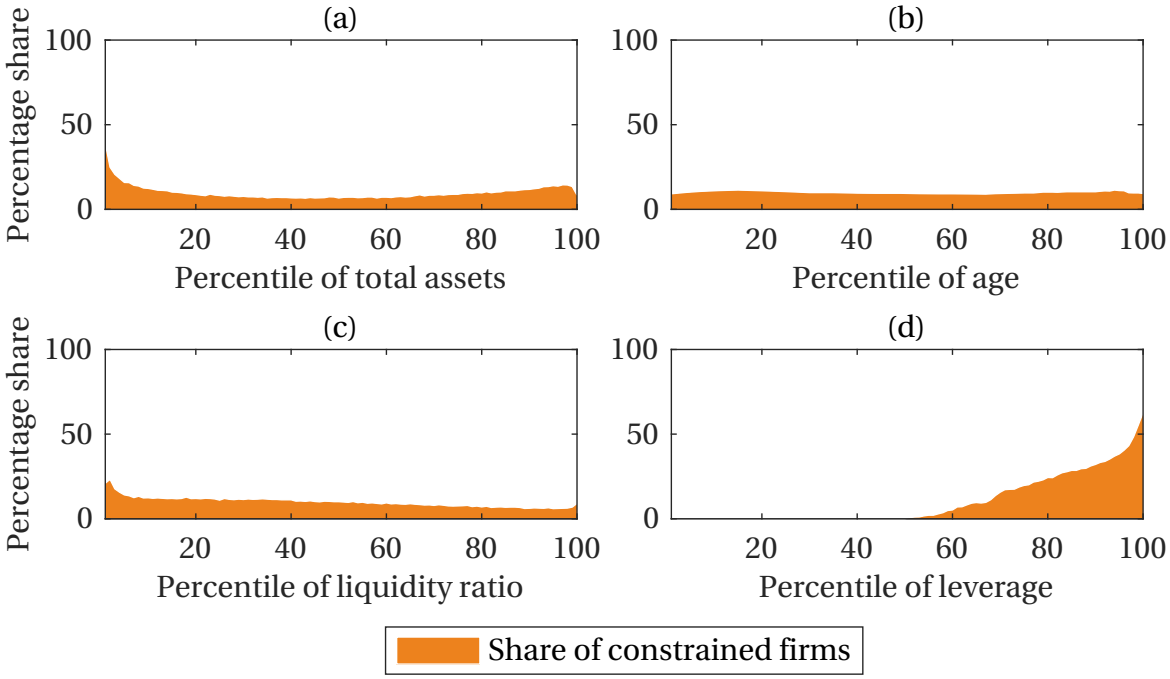


Fig. B.5: Decomposition of constrained and unconstrained firms across percentiles of firm variables using constraint measure IV

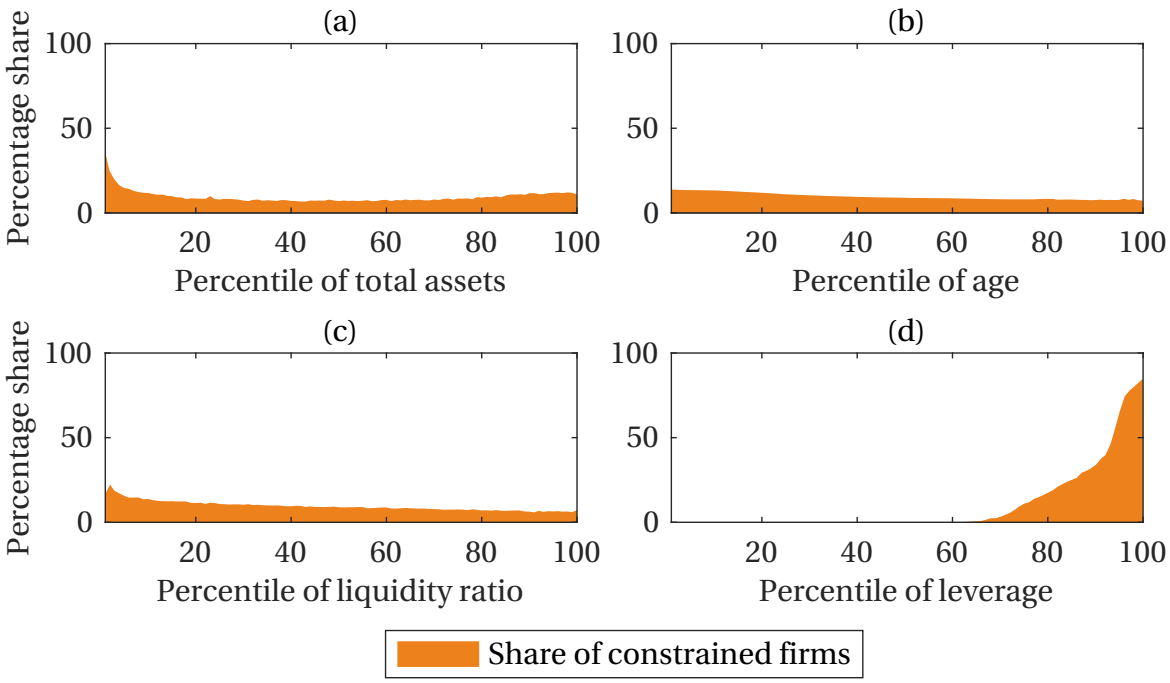


Fig. B.6: Decomposition of constrained and unconstrained firms across percentiles of firm variables using constraint measure V

B.3.2 Statistical model

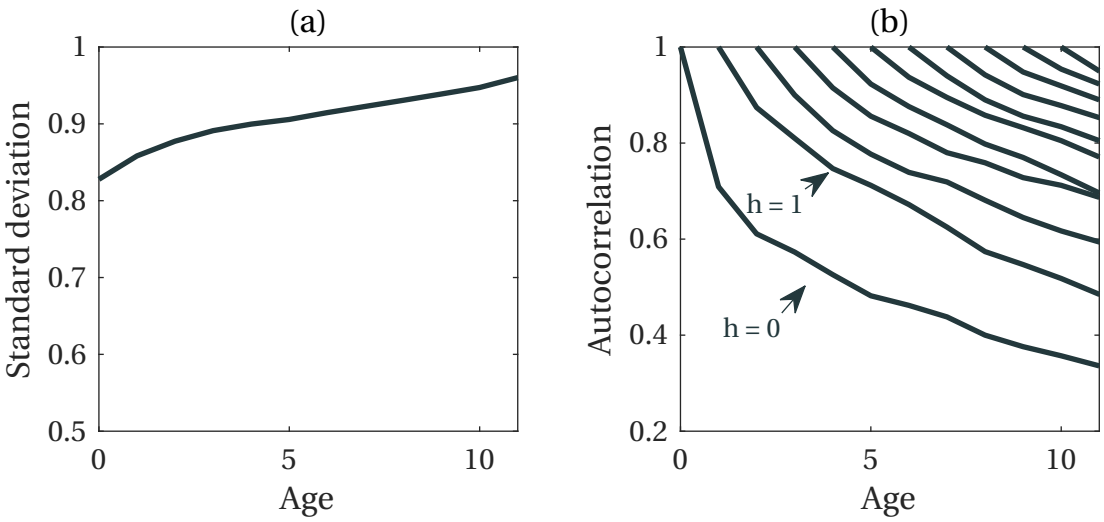


Fig. B.7: Standard deviation and autocorrelation of log employment by age. The left panel presents the standard deviation of log employment by age, after controlling for sector and year fixed effects. The right panel presents the autocorrelation of log employment between ages a and $h \leq a$. Across lines h changes, while a changes along the lines

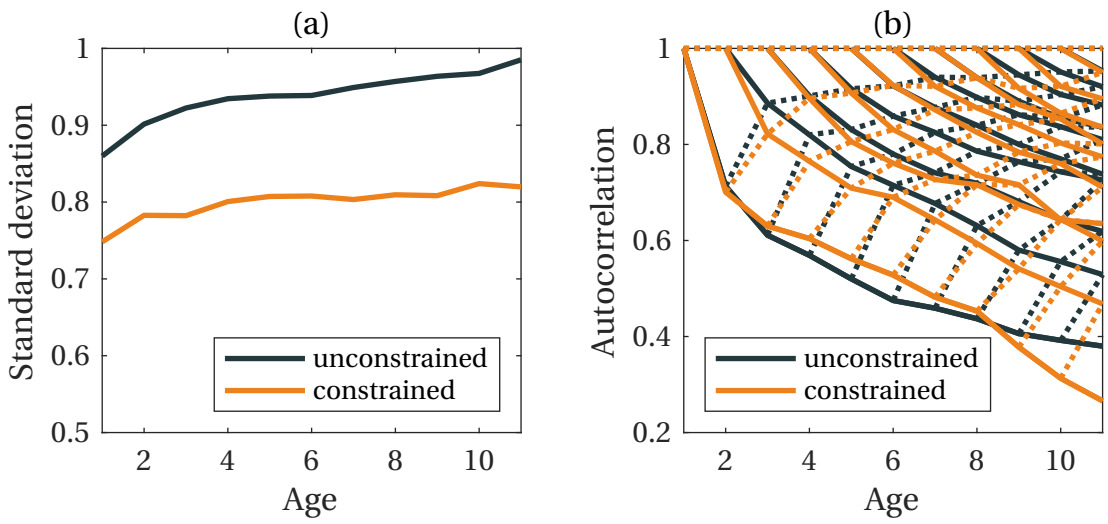


Fig. B.8: Standard deviation and autocorrelation of log employment by age and separated by constraint measure I. The left panel presents the standard deviation of log employment by age, after controlling for sector and year fixed effects. The right panel presents the autocorrelation of log employment between ages a and $h \leq a$. Across lines h changes, while a changes along the lines

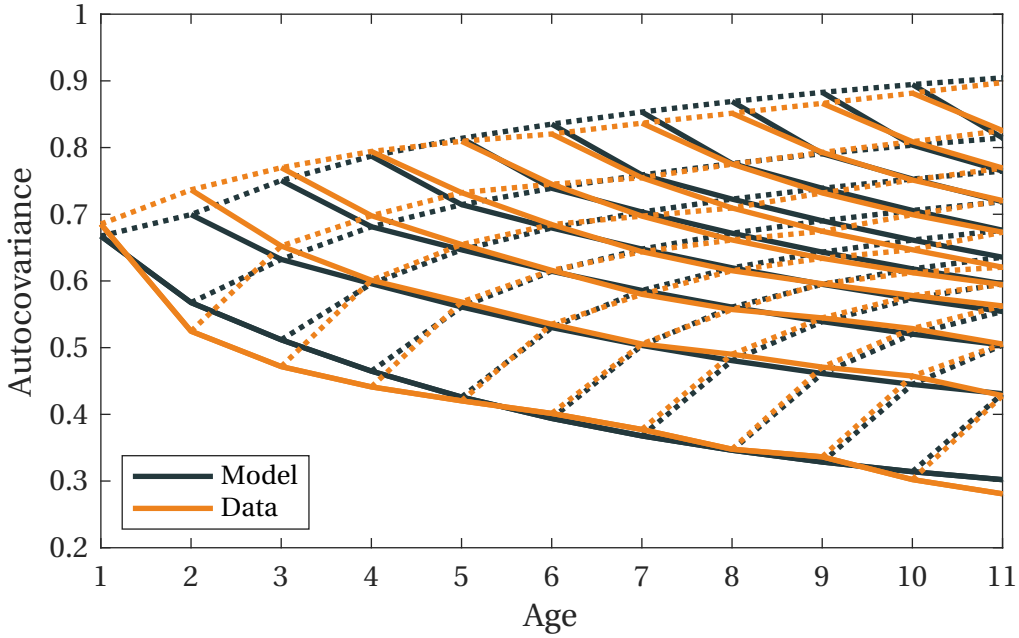


Fig. B.9: Model fit of statistical model for employment process

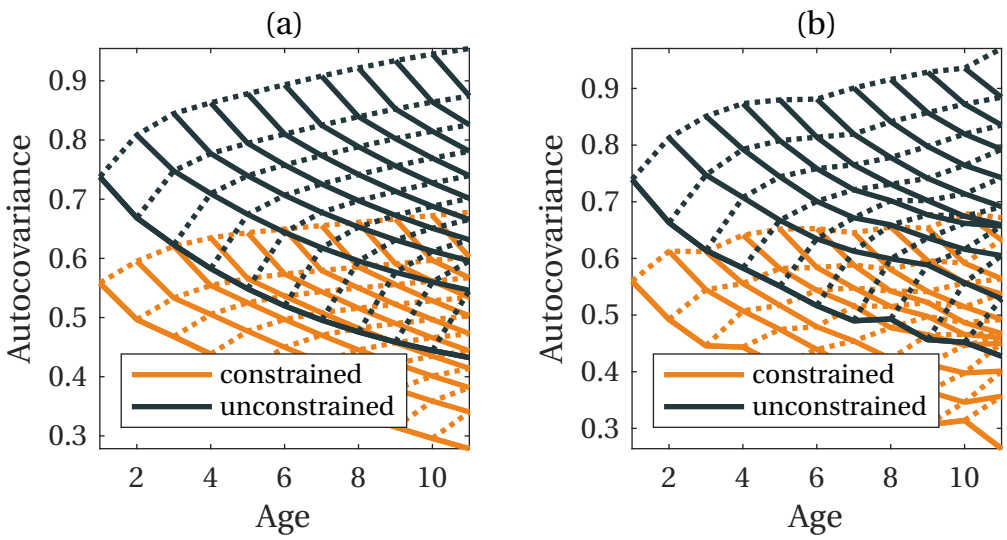


Fig. B.10: Model (a) and empirical (b) autocovariance for constrained firms (orange) and unconstrained firms (blue) using the measure Constrained I

B.3.3 Theoretical model

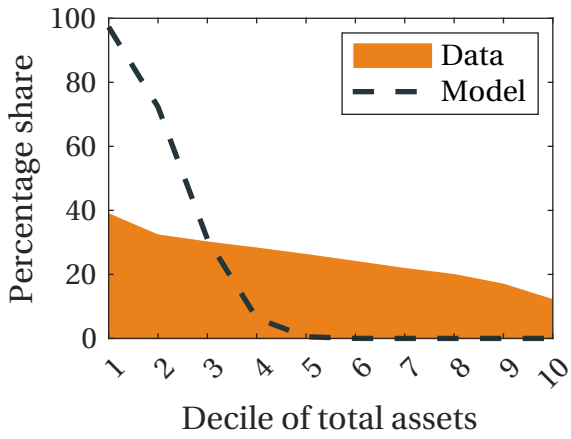
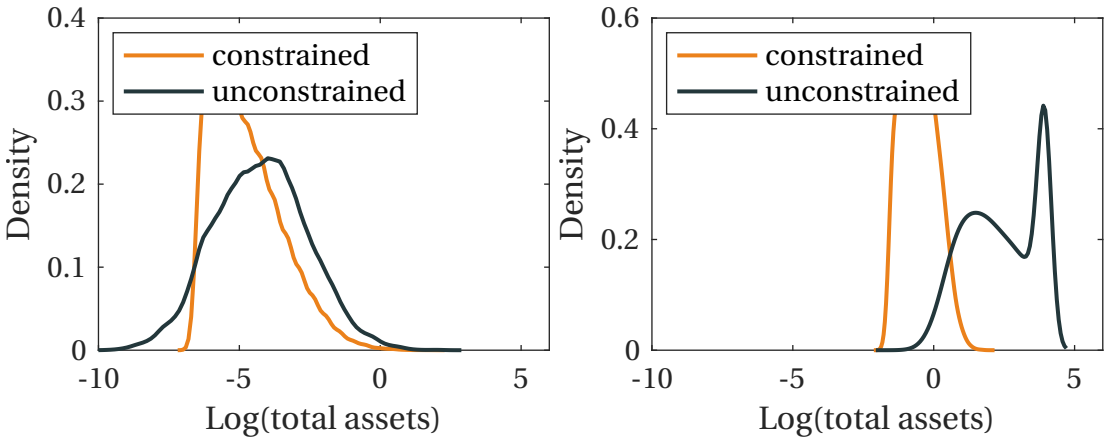


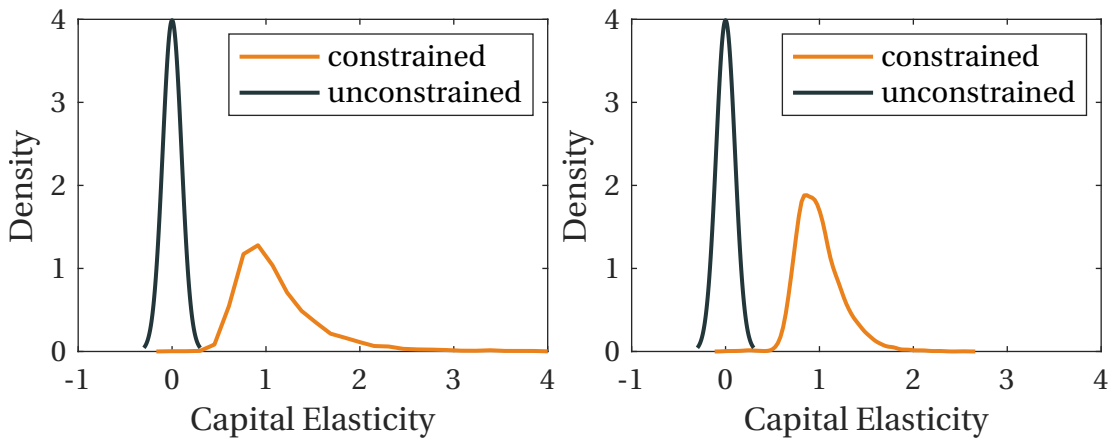
Fig. B.11: Share of constrained firms across the distribution in the transitory shock only model with calibration in Table B.13



(a) Transitory + permanent shock

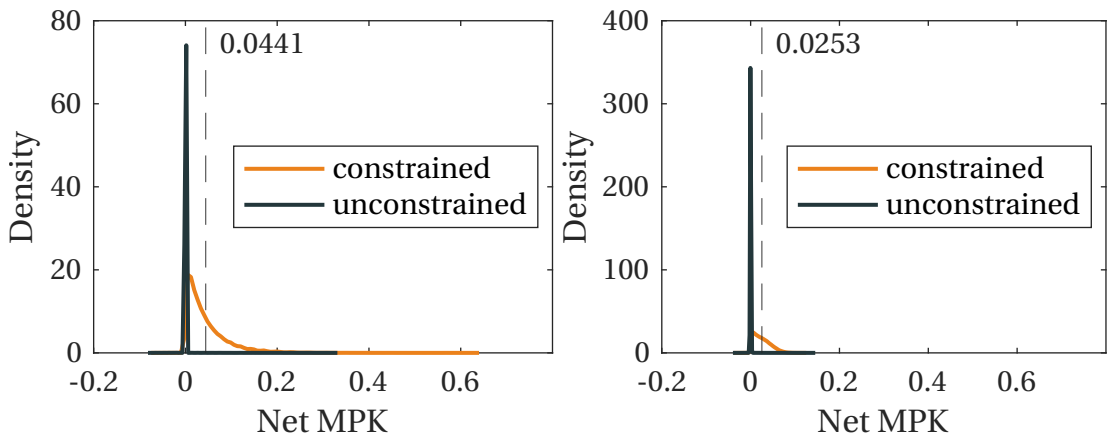
(b) Transitory shock

Fig. B.12: Conditional distributions of log of total assets implied by the model



(a) Transitory + permanent shock (b) Transitory shock

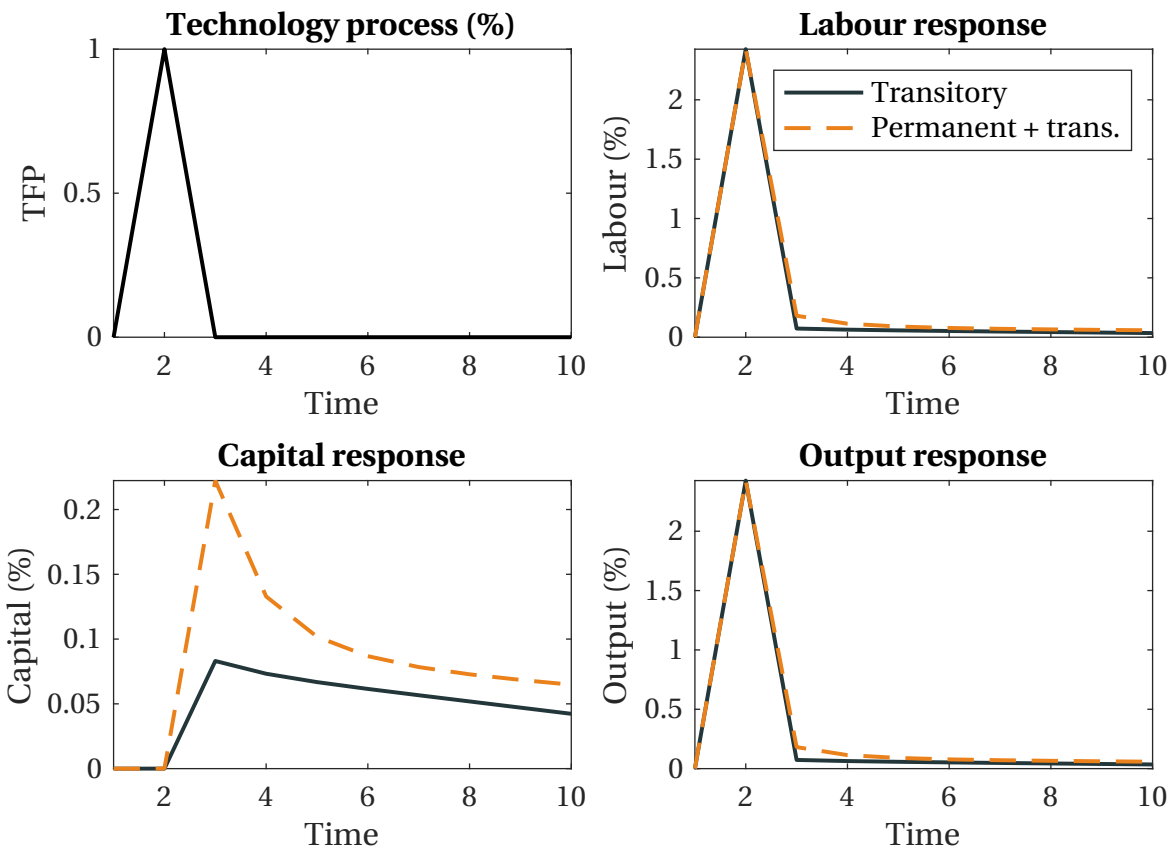
Fig. B.13: Conditional distribution of capital elasticity



(a) Transitory + permanent shock (b) Transitory shock

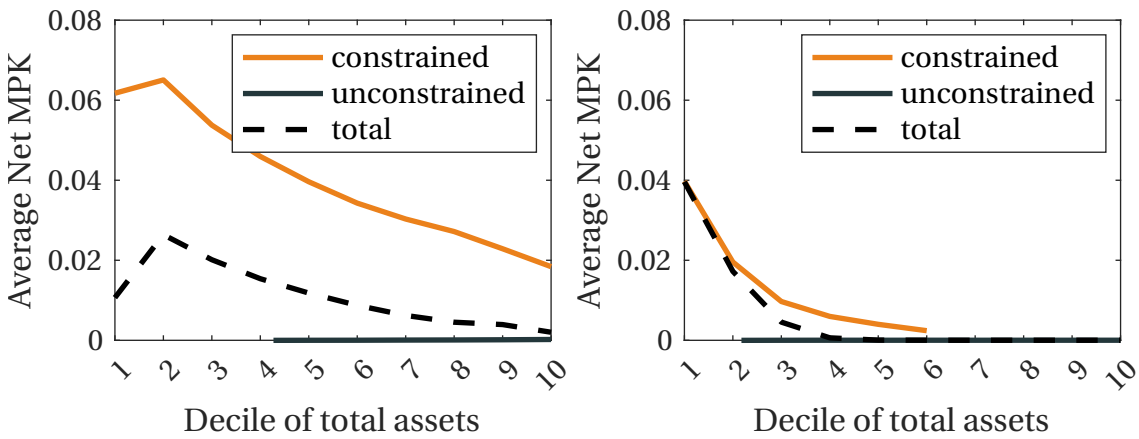
Notes. The dashed line depicts the conditional mean of the marginal product of capital for constrained firms.

Fig. B.14: Conditional distributions of MPKs



Notes. Lines indicate the partial equilibrium response to a shock to overall TFP in the upper left panel, with wages fixed at their steady state level.

Fig. B.15: IRFs after an aggregate productivity shock



(a) Transitory + permanent shock (b) Transitory shock

Notes. The line depicts the average MPK of constrained firms per decile bin of total assets along the entire size distribution.

Fig. B.16: Average MPK along total assets distribution

B.4 Statistical model derivation

This is reproduced from Pugsley et al. (2021) for reference. Write stochastic processes in MA representation:

$$
\begin{aligned}
u_{i,t} &= \rho_u^{t+1} u_{i,-1} + \sum_{k=0}^{a} \rho_u^k \theta_i \\
v_{i,a} &= \rho_v^{a+1} v_{i,-1} \\
w_{i,a} &= \sum_{k=0}^{a} \rho_w^k \varepsilon_{i,a-k} = \sum_{k=0}^{j-1} \rho^k \varepsilon_{i,a-k} + \rho_v^j \sum_{k=0}^{a-j} \rho_v^k \varepsilon_{i,a-j-k} \quad 0 \le j \le a
\end{aligned}
$$

So the level of log employment of firm i at age a is:

$$
\ln n_{i,a} = \rho_u^{a+1} u_{i,-1} + \sum^{a} \rho_u^k \theta_i + \rho_v^{a+1} v_{i,-1} + \sum_{i=1}^{j-1} \rho^k \varepsilon_{i,a-k} + \rho_v^j \sum_{i=1}^{a-j} \rho_v^k \varepsilon_{i,a-j-k} + z_{i,a}
$$

Then the autocovariance of log employment at age a and $a-j$ for $j \ge 0$ is:

$$
\begin{aligned}
\mathrm{Cov}\left[\log n_{i,a}, \log n_{i,a-j}\right] &= \left(\sum_{k=0}^{a} \rho_u^k\right) \sigma_\theta^2 \left(\sum_{k=0}^{a-j} \rho_u^k\right) + \rho_u^{a+1} \sigma_{\tilde{u}}^2 \rho_u^{a-j+1} + \rho_v^{a+1} \sigma_{\tilde{v}}^2 \rho_v^{a-j+1} \\
&\quad + \mathrm{Cov}\left[\rho_v^j \sum_{k=0}^{a-j} \rho_v^k \varepsilon_{i,a-j-k}, \sum_{k=0}^{a-j} \rho_v^k \varepsilon_{i,a-j-k}\right] + \mathbf{1}_{\{j=0\}} \sigma_z^2 \\
&= \sigma_\theta^2 \left(\sum_{k=0}^{a} \rho_u^k\right) \left(\sum_{k=0}^{a-j} \rho_u^k\right) + \sigma_{\tilde{u}}^2 \rho_u^{2(a+1)-j} + \sigma_{\tilde{v}}^2 \rho_v^{2(a+1)-j} + \sigma_\varepsilon^2 \rho_w^j \sum_{k=0}^{a-j} \rho_w^{2k} + \mathbf{1}_{\{j=0\}} \sigma_z^2
\end{aligned}
$$

B.5 Model: Firm level decisions

Unconstrained Firms This group of firms can implement both the optimal amount of capital and the minimum savings policy that guarantees these firms will never be constrained in the future again. Given the absence of adjustment costs and the stochastic process for φ the optimal amount of capital is the solution to:

$$\max_{k'} \quad -k' + \beta \mathbb{E}_{\varphi'|\varphi}\left[(\pi(k',\varphi') + (1-\delta)k'\right]$$

So the optimal amount of capital solves the following equation

$$\beta \mathbb{E}_{\varphi'|\varphi}\left[\frac{\partial \pi}{\partial k'}(k',\varphi')\right] = 1 + \beta\delta - \beta$$

which is when the expected marginal productivity of capital is equal to the marginal cost of an extra unit. The minimum savings policy these firms implement guarantees they will never be constrained again. It is given by

$$\mathrm{B}^*(\varphi_i) = \min_{\varphi_j} \tilde{\mathrm{B}}(k^*(\varphi_i),\varphi_j)$$

where $\tilde{\mathrm{B}}(k^*(\varphi_i),\varphi_j)$ is the minimum savings that guarantees that going from state φ_i to φ_j the firm is still able to implement the optimal amount of capital. It is given by

$$\begin{aligned}\tilde{\mathrm{B}}(k^*(\varphi_i),\varphi_j) =& \pi(k^*(\varphi_i),\varphi_j) + (1-\delta)k^*(\varphi_i) - k'^*(\varphi_j) + \\ & q\min\left\{\mathrm{B}^*(\varphi_j), \xi\left(\pi(k^*(\varphi_i),\varphi_j) + (1-\delta)k^*(\varphi_i) - \tilde{\mathrm{B}}(k^*(\varphi_i),\varphi_j)\right)\right\}\end{aligned}$$

Given the optimal amount of capital and the minimum savings policy, the dividends distributed by the unconstrained firms are given by

$$\mathrm{D} = x - k^* + q\mathrm{B}^*$$

From the dividend constraint $\mathrm{D} \geq 0$ we can extract the minimum threshold for cash-on-hand that guarantees the firm is not constrained

$$\tilde{x} = k^* - q\mathrm{B}^*$$

and the firms is constrained if $x \leq \tilde{x}$.

Constrained Firms: Type 1 These firms can implement the optimal amount of capital, k^*, but not the optimal savings policy and are therefore partially constrained. As they may still be constrained in future states, they value internal financing more than households value dividends. As a result, for this type of firms, D = 0. The amount of debt is given by

$$b' = \frac{(k^* - x)}{q}$$

A firm is type 1 if it can adopt the above amount of debt and capital and at the same time guaranteeing that it does not default in the next period.

Constrained Firms: Type 2 Strictly constrained firms can not implement the optimal amount of capital. Those firms utilize all their borrowing capacity as their marginal value of net worth is greater than unity. Hence, their savings policy is simply

$$b' = \xi x,$$

and their maximum possible investment is consequently

$$k' = x + q\xi x < k^*,$$

which is strictly smaller than their optimal level of capital k^*.

B.6 Simple model: Results

Take a very simple model to analyze the impact of heterogeneous productivities on cyclicality, following Crouzet & Mehrotra (2020). Firms can only invest in physical capital, have permanent productivity and face no uncertainty, except for a stochastic death shock. The problem can be written as:

$$\mathrm{V}(k_{t,i}, b_{t,i}, \theta_i) = \pi_d x_{t,i} + (1-\pi_d)\left(x_{t,i} - k_{t+1,i} + q_t b_{t+1,i} + \beta \mathrm{V}(k_{t+1,i}, b_{t+1,i}, \theta_i)\right)$$

subject to

$$x_{t,i} = z_t \theta_i k_{t,i}^{\alpha} + (1-\delta) k_{t,i} - b_{t,i}$$

$$\xi x_{t,i} \geq b_{t+1,i}$$

$$k_{t+1,i} \leq x_{t,i} + q_t b_{t+1,i}$$

B.6.1 Unconstrained firms

Steady state growth. Unconstrained firms optimal capital $k^*_{t+1,i}$ is the solution to:

$$\beta^{-1} = (1-\delta) + \alpha z_t \theta_i k_{t+1,i}^{*\alpha-1}$$

Hence optimal capital $k^*_{t+1,i}$ is

$$k^*_{t+1,i} = \theta_i^{\frac{1}{1-\alpha}} \left(\frac{\alpha z_{t+1}}{\beta^{-1} - (1-\delta)}\right)^{\frac{1}{1-\alpha}}$$

where we can choose $z := \left(\frac{\beta^{-1}-(1-\delta)}{\alpha}\right)$ such that, at steady state and for $\theta = 1$, we have that $k^*_{t+1,i} = 1$. In the absence of idiosyncratic shocks and constant total factor productivity z, unconstrained firms are not growing at steady state as they reached their optimal level of capital.

$$g_{uncons} = \frac{k^*_{t+1,i}(\theta_i)}{k^*_{t,i}(\theta_i)} = 1$$

Cyclicality. Now consider the following setup; at time $t = -1$, $z_t = z$. At time $t = 0$, firms learn the future path of z_t, for $t \geq 0$ will be

$$z_t = z \exp(\rho^t \epsilon)$$

The growth rate then becomes

$$g_{uncons} = \frac{k^*_{t+1,i}(\theta_i)}{k^*_{t,i}(\theta_i)} = \frac{\exp(\frac{\rho}{1-\alpha}\epsilon)\theta_i^{1/(1-\alpha)}}{\theta_i^{1/(1-\alpha)}} = \exp\left(\frac{\rho}{1-\alpha}\epsilon\right)$$

Hence, the elasticity of capital is the same across all unconstrained firms, independent of firm size and firm-specific productivity.

$$\frac{\Delta g_{uncons}}{\Delta\epsilon}|_{\epsilon\approx 0} = \frac{\rho}{1-\alpha}$$

B.6.2 Constrained firms

Steady state growth. Constrained firms invest according to their maximum investment capacity which is capped by the net worth constraint.

$$\begin{aligned} k_{t+1,i} &= n_{t,i} + q_t b_{t+1,i} \\ &= n_{t,i} + q_t \xi n_{t,i} \\ &= (1+q_t\xi)(z_t\theta_i k_{t,i}^{\alpha} + (1-\delta)k_{t,i} - b_{t,i}) \end{aligned}$$

Hence,

$$\begin{aligned} g_{cons} &= (1+q_t\xi)(z_t\theta_i k_{t,1}^{\alpha-1} + (1-\delta) - b_{t,i}/k_{t,i}) \\ &= (1+q_t\xi)(z_t\theta_i k_{t,1}^{\alpha-1} + (1-\delta) - \frac{\xi}{1+q_{t-1}\xi}) \end{aligned}$$

Due to decreasing returns to scale, the growth rate is affected by the size of the firm, with larger firms growing slower

$$\frac{\Delta g_{cons}}{\Delta k_{t,i}} = (1+q_t\xi)(\alpha-1)z_t\theta_i k_{t,1}^{\alpha-2} < 0$$

For firms of the same size, those with a higher permanent productivity component grow quicker

$$\frac{\Delta g_{cons}}{\Delta\theta_i} = (1+q_t\xi)z_t k_{t,1}^{\alpha-1} > 0$$

Cyclicality Now consider the same setup as for unconstrained firms; at time $t = -1$, $z_t = z$. At time $t = 0$, firms learn the future path of z_t, for $t \geq 0$ will be

$$z_t = z \exp(\rho^t \epsilon)$$

The growth rate on impact then becomes

$$g_{cons} = (1 + q_t \xi)\left(z \exp(\rho^0 \epsilon)\theta_i k_{t,1}^{\alpha-1} + (1-\delta) - \frac{\xi}{1 + q_{t-1}\xi}\right)$$

So, the elasticity of capital with respect to the shock ϵ is decreasing on capital and increasing on the productivity of the firm

$$\frac{\Delta g_{cons}}{\Delta \epsilon}|_{\epsilon \approx 0} = (1 + q_t \xi)(z\theta_i k_{t,1}^{\alpha-1}) = \frac{(1 + q_t \xi)}{\alpha} \text{mpk}_i$$

With the derivative of the elasticity with respect to the size and productivity of the firm being negative and positive respectively

$$\frac{\Delta^2 g_{cons}}{\Delta \epsilon \Delta \theta_i}|_{\epsilon \approx 0} = (1 + q_t \xi)(z k_{t,1}^{\alpha-1}) > 0$$

$$\frac{\Delta^2 g_{cons}}{\Delta \epsilon \Delta k_{t,1}}|_{\epsilon \approx 0} = (\alpha - 1)(1 + q_t \xi)(z\theta_i k_{t,1}^{\alpha-2}) < 0$$

When is the elasticity of constrained larger than unconstrained?

$$\frac{\Delta g_{cons}}{\Delta \epsilon}|_{\epsilon \approx 0} > \frac{\Delta g_{uncons}}{\Delta \epsilon}|_{\epsilon \approx 0}$$

This happens when the marginal product of capital of constrained firms is above a given threshold

$$\text{mpk} > \rho \frac{\alpha}{1-\alpha} \frac{1}{1 + q_t \xi}$$

So, two factors will determine which elasticity is larger: (i) the marginal product of capital of constrained firms, which depends on the distribution in terms of both size and productivity. The smaller and the more productive constrained firms are, the higher their elasticity; (ii) the persistence of the aggregate shock. As ρ approaches zero, unconstrained firms will not react to the shock, while the elasticity of constrained firms on impact does not depend on the persistence of the shock.

B.7 Model: Additional results

Capital misallocation Besides the amplification of financial shocks, what does a more realistic distribution of financially constrained firms imply for the degree of misallocation in the economy? Pugsley et al. (2021) show that when accounting for ex-ante heterogeneity, the economy exhibits a stronger degree of misallocation. In Table B.14 we compare a number of model aggregate statistics in with their first-best counterparts, i.e. the case where financial frictions eliminated. While our results in Table B.14 are qualitatively in line with Pugsley et al. (2021), they suggest that the degree of misallocation can be substantially larger when accounting for the skewness in the firms' size and capital distribution, as well as large constrained firms. Consider, for example the degree of misallocation for output, which is almost three times as large in with both transitory and permanent productivity components (0.939 vs 0.976).[1]

Table B.14: Deviations from frictionless economy

	Deviations from 1st best	
	trans. + perm.	transitory only
Consumption	0.949	0.982
Capital	0.894	0.952
Output	0.939	0.976
Employment	0.989	0.994
MPK stdev	0.048	0.034

Notes. Reported values are relative to the models without any financial frictions, i.e. when setting the collateral constraint parameter ξ to a sufficiently large value that firms can directly implement their optimal amount of capital.

Policy implications Lastly, we discuss the impacts of policies designed to aid small businesses. For example, the U.S. Federal government has the requirement to allocate 23% of total procurement contracts to small businesses. di Giovanni et al. (2022) argue that allocating procurement contracts to small firms, can help these firms overcome financial constraints and grow faster. In light of the new stylized facts presented in this paper, of

[1] Figures B.14 and B.16 in Appendix B.3.3 illustrate the mechanism explaining the larger capital misallocation. Figure B.14 showcases the density distribution of the MPKs in both models. It is possible to observe that in the two shock model there is a much higher dispersion of MPKs and average. Figure B.16 illustrates that the existence of large constrained firms also contributes to having larger MPK at the top of the distribution.

Table B.15: Deviations from steady state

	Deviations from steady state	
	trans. + perm.	transitory only
Consumption	1.014	1.015
Capital	1.018	1.023
Output	1.015	1.017
Employment	1.015	1.017
MPK stdev	0.890	0.958

Notes. Reported values are relative to the benchmark models without any subsidy.

the weak correlation between size and constraint, we assess if the impacts of this type of policies has been overestimated.

The policy here implemented takes the form of a government subsidy granted to the 20% smallest firms in the economy. In total, the subsidy amounts to 1% of the GDP and is paid by households, with a lump sum tax. Results in Table B.15 show that, despite the effects of this policy being positive, with consumption, capital, output and employment all going up, the effects are diminished when accounting for the joint constrained size distribution, as is the case in the transitory and permanent shocks model. When that joint distribution is matched, consumption grows less 7%, capital less 22% and output and employment less 12% than in the model with just a transitory productivity shock that cannot match the joint constrained size distribution. [2]

[2] Cloyne et al. (2018) argue that the firm's age is a better proxy for financial constraints than size. Although we do not account for the joint age constraint distribution, empirically we find that the correlation between age and constraint is positive but equally weak. In case we would account for the age constraint distribution, our aggregate results would not change, as the main mechanism of the share of productive capital in constrained firms is not affected by the age distribution.

C

Appendix for Chapter 3

C.1 Algorithm to solve two asset economy

he precise algorithm for the solution of the benchmark model is as follows:

1. Parameterise bond price q_t, using equation (3.34). Guess a set of coefficients $\boldsymbol{\vartheta}$.

2. Parameterise the conditional expectations of the Euler equations of all history bins $y^{\mathrm{N}} \in \mathscr{Y}^{\mathrm{N}}$ as follows:

$$\mathbb{E}_t \left[(1 + r_{t+1}) \sum_{\tilde{y}^{\mathrm{N}} \in \mathscr{Y}^{\mathrm{N}}} \Pi_{y^{\mathrm{N}}, \tilde{y}^{\mathrm{N}}} \xi_{t, \tilde{y}^{\mathrm{N}}} u'(c_{t+1, \tilde{y}^{\mathrm{N}}}) \right] \approx \Phi_{y^{\mathrm{N}}}(\mathbf{X}_{t, y^{\mathrm{N}}}, \boldsymbol{\gamma}_{y^{\mathrm{N}}}; \boldsymbol{\vartheta}), \tag{C.1}$$

$$\mathbb{E}_t \left[\sum_{\tilde{y}^{\mathrm{N}} \in \mathscr{Y}^{\mathrm{N}}} \Pi_{y^{\mathrm{N}}, \tilde{y}^{\mathrm{N}}} \xi_{t, \tilde{y}^{\mathrm{N}}} u'(c_{t+1, \tilde{y}^{\mathrm{N}}}) \right] \approx \Psi_{y^{\mathrm{N}}}(\mathbf{X}_{t, y^{\mathrm{N}}}, \boldsymbol{\kappa}_{y^{\mathrm{N}}}; \boldsymbol{\vartheta}), \tag{C.2}$$

where $\Phi_{y^{\mathrm{N}}}, \Psi_{y^{\mathrm{N}}}$ are the parameterised expectations functions, $\mathbf{X}_{t, y^{\mathrm{N}}}$ the state variables of island y^{N} at time t and $\boldsymbol{\gamma}_{y^{\mathrm{N}}}$, $\boldsymbol{\kappa}_{y^{\mathrm{N}}}$ are the corresponding coefficients of the assumed parameterised expectations function for island y^{N}. In particular, assume an exponential-log form for the parameterised expectations function, following Marcet & Lorenzoni (1998), so that

$$\Phi_{y^{\mathrm{N}}}(\mathbf{X}_{t, y^{\mathrm{N}}}, \boldsymbol{\gamma}_{y^{\mathrm{N}}}; \boldsymbol{\vartheta}) = \exp(\boldsymbol{\gamma}'_{y^{\mathrm{N}}} \cdot [1, \log(\mathbf{X}_{t, y^{\mathrm{N}}})]), \tag{C.3}$$

$$\Psi_{y^{\mathrm{N}}}(\mathbf{X}_{t, y^{\mathrm{N}}}, \boldsymbol{\kappa}_{y^{\mathrm{N}}}; \boldsymbol{\vartheta}) = \exp(\boldsymbol{\kappa}'_{y^{\mathrm{N}}} \cdot [1, \log(\mathbf{X}_{t, y^{\mathrm{N}}})]). \tag{C.4}$$

Guess a set of parameter vectors $(\boldsymbol{\gamma}_{y^{\mathrm{N}}}, \boldsymbol{\kappa}_{y^{\mathrm{N}}})_{y^{\mathrm{N}} \in \mathscr{Y}^{\mathrm{N}}}$.

3. Initialise a distribution $(\mathrm{S}_{-1,y^{\mathrm{N}}}, k_{-1,y^{\mathrm{N}}}, b_{-1,y^{\mathrm{N}}})_{y^{\mathrm{N}}\in\mathscr{E}^{\mathrm{N}}}$. Draw a long shock series for the aggregate productivity process $\{z_t\}_{t\geq 0}^{\mathrm{T}}$.

4. Simulate the economy for T periods. In particular, calculate

$$(k_{t,y^{\mathrm{N}}}(\boldsymbol{\gamma}_{y^{\mathrm{N}}}, \boldsymbol{\kappa}_{y^{\mathrm{N}}}; \boldsymbol{\vartheta}), b_{t,y^{\mathrm{N}}}(\boldsymbol{\gamma}_{y^{\mathrm{N}}}, \boldsymbol{\kappa}_{y^{\mathrm{N}}}; \boldsymbol{\vartheta}), c_{t,y^{\mathrm{N}}}(\boldsymbol{\gamma}_{y^{\mathrm{N}}}, \boldsymbol{\kappa}_{y^{\mathrm{N}}}; \boldsymbol{\vartheta}))_{t\geq 0, y^{\mathrm{N}}\in\mathscr{Y}^{\mathrm{N}}}^{\mathrm{T}}$$

   in any period t as follows. I omit the explicit dependency on the various coefficients for easier notation wherever possible.

   (a) **Prices and labour supply**: Solve for prices r_t and w_t using the firm first order conditions.

   (b) **Unconstrained individual problem**: Solve the system of budget constraint and complementary slackness conditions by first assuming that none of the history bin specific inequality constraints is binding. This means that the Euler equation for capital gives bin specific consumption $c_{t,y^{\mathrm{N}}}$. Using this in the Euler equation for bonds enables us to recover $b_{t,y^{\mathrm{N}}}$. Finally, $k_{t,y^{\mathrm{N}}}$ follows from the budget constraint.

   (c) **Constrained individual problem**: Check whether any of the implied allocations above violate the borrowing constraints. If that is the case solve by imposing the borrowing constraint on this asset. This gives three possible cases for every island y^{N}:

      i. **Constrained in bonds**. Impose $b_{t,y^{\mathrm{N}}} = \bar{b}$. Solve the budget constrain and the capital Euler equation for $c_{t,y^{\mathrm{N}}}$ and $k_{t,y^{\mathrm{N}}}$ with $b_{t,y^{\mathrm{N}}} = \bar{b}$.

      ii. **Constrained in capital**. Impose $k_{t,y^{\mathrm{N}}} = \bar{k}$. Solve the budget constraint and the bond Euler equation for $c_{t,y^{\mathrm{N}}}$ and $b_{t,y^{\mathrm{N}}}$ with $k_{t,y^{\mathrm{N}}} = \bar{k}$.

      iii. **Constrained in both assets** Impose $k_{t,y^{\mathrm{N}}} = \bar{k}$ and $b_{t,y^{\mathrm{N}}} = \bar{b}$. Solve the budget constraint for $c_{t,y^{\mathrm{N}}}$.

   (d) Use the generated series to produce vectors with the realisation of the expecta-

tions in Euler equations:

$$\Theta_{t,y^{\mathrm{N}}}(s_t(\boldsymbol{\gamma}_{y^{\mathrm{N}}},\boldsymbol{\kappa}_{y^{\mathrm{N}}};\boldsymbol{\vartheta}),s_{t+1}(\boldsymbol{\gamma}_{y^{\mathrm{N}}},\boldsymbol{\kappa}_{y^{\mathrm{N}}};\boldsymbol{\vartheta})) =$$
$$\beta\sum_{\tilde{y}^{\mathrm{N}}\succ y^{\mathrm{N}}}\Pi_{t,y^{\mathrm{N}},\tilde{y}^{\mathrm{N}}}\xi_{t,\tilde{y}^{\mathrm{N}}}\mathrm{U}_c(c_{t+1,y^{\mathrm{N}}}(\boldsymbol{\gamma}_{y^{\mathrm{N}}},\boldsymbol{\kappa}_{y^{\mathrm{N}}};\boldsymbol{\vartheta}))(1+r_{t+1,y^{\mathrm{N}}}(\boldsymbol{\gamma}_{y^{\mathrm{N}}},\boldsymbol{\kappa}_{y^{\mathrm{N}}};\boldsymbol{\vartheta})),$$

$$\Omega_{t,y^{\mathrm{N}}}(s_t(\boldsymbol{\gamma}_{y^{\mathrm{N}}},\boldsymbol{\kappa}_{y^{\mathrm{N}}};\boldsymbol{\vartheta}),s_{t+1}(\boldsymbol{\gamma}_{y^{\mathrm{N}}},\boldsymbol{\kappa}_{y^{\mathrm{N}}};\boldsymbol{\vartheta})) =$$
$$\beta\sum_{\tilde{y}^{\mathrm{N}}\succ y^{\mathrm{N}}}\Pi_{t,y^{\mathrm{N}},\tilde{y}^{\mathrm{N}}}\xi_{t,\tilde{y}^{\mathrm{N}}}\mathrm{U}_c(c_{t+1,y^{\mathrm{N}}}(\boldsymbol{\gamma}_{y^{\mathrm{N}}},\boldsymbol{\kappa}_{y^{\mathrm{N}}};\boldsymbol{\vartheta})).$$

Calculate aggregate quantities and move to period $t+1$.

5. For every island y^{N} regress the realised expectation on the assumed parameterised expectation function, using the newly created series of state variables as explanatory variables.

6. If $|\boldsymbol{\gamma}^{\boldsymbol{new}}_{y^{\mathrm{N}}}-\boldsymbol{\gamma}_{y^{\mathrm{N}}}|>tol_{pea}$ or $|\boldsymbol{\kappa}^{\boldsymbol{new}}_{y^{\mathrm{N}}}-\boldsymbol{\kappa}_{y^{\mathrm{N}}}|>tol_{pea}$ update the vectors $(\boldsymbol{\gamma}_{y^{\mathrm{N}}},\boldsymbol{\kappa}_{y^{\mathrm{N}}})$ using homotopy:

$$\boldsymbol{\gamma}'_{y^{\mathrm{N}}}=\mu\boldsymbol{\gamma}^{\boldsymbol{new}}_{y^{\mathrm{N}}}+(1-\mu)\boldsymbol{\gamma}_{y^{\mathrm{N}}},$$
$$\boldsymbol{\kappa}'_{y^{\mathrm{N}}}=\mu\boldsymbol{\kappa}^{\boldsymbol{new}}_{y^{\mathrm{N}}}+(1-\mu)\boldsymbol{\kappa}_{y^{\mathrm{N}}},$$

where μ is an updating parameter and move to item 4 again with the new set of coefficients. If not, move to 7.

7. Simulate the economy with the newly found parameterised expectations. At every t solve for the market clearing price $q_{t,sim}$ using a non-linear solver.

8. Run the following regression:

$$q_{t,sim}=\vartheta_0^{new}+\vartheta_1^{new}\log(z_{t,sim})+\vartheta_2^{new}\log(\mathrm{K}_{t-1,sim})+\vartheta_3^{new}\log(\sigma_{\mathrm{Z},t})$$

If $|\boldsymbol{\vartheta}^{\boldsymbol{new}}-\boldsymbol{\vartheta}|>tol_{price}$, update the vector using homotopy:

$$\boldsymbol{\vartheta}'=\mu\boldsymbol{\vartheta}^{\boldsymbol{new}}+(1-\mu)\boldsymbol{\vartheta}$$

and move to item 4 again. If not, the algorithm is done.