The London School of Economics and Political Science

Essays in Macroeconometrics

Miguel Bandeira

A thesis submitted to the Department of Economics of the London School of Economics and Political Science for the degree of Doctor of Philosophy, London, July 2020.

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work.

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of approximately 24,000 words, excluding graphs, tables and appendices.

Abstract

This thesis is composed of three chapters.

Chapter 1 introduces a statistical framework to study dynamic (S,s) economies. The proposed framework enables researchers to estimate unit level cumulated changes in frictionless variables based on a panel of variables for which changes are intermittent and lumpy. Formally, the estimates are grounded on an exact closed-form solution for the smoothing problem associated with a nonlinear and non-Gaussian state-space representation of an economy composed of microeconomic unit pursuing two-sided (S,s) policies subject to costless adjustment opportunities. This state-space representation is semi-structural and can accommodate some classic problems that have been analysed through the lens of (S,s) models such as pricing subject to menu costs, cash withdrawals and plant-level investment and hiring and firing decisions. The resulting unit level estimates can be used to construct estimates of frictionless aggregate variables and of the cross-sectional distribution of state gaps for any time period.

Chapter 2 applies the theoretical results developed in chapter 1 to a large micro price dataset underlying the United Kingdom Consumer Price Index to produce a novel measure of inflation which I label *frictionless inflation*. This measure is theoretically grounded on a random menu cost model and it should be interpreted as the inflation that would have been observed in a counterfactual world where menu costs of price adjustment did not exist. I use this measure to answer four questions. First, what is the importance of menu costs for the aggregate inflation dynamics? Second, what is the importance of menu costs for the transmission of monetary policy shocks? Third, what is the relationship between frictionless inflation and the movements in the output gap? Fourth, can frictionless inflation be used as a leading indicator for headline inflation?

Chapter 3 studies the estimation of impulse responses functions (IRFs) of different individuals to an aggregate shock. The commonplace approach to this

problem involves first grouping individuals according to some external classification or observable explanatory variables and subsequently estimating the associated group-specific IRFs. This chapter starts by showing that the IRF estimates based on this approach are subject a *misclassification bias* that arises whenever the grouping of individuals imposed by the researcher groups together individuals that do not react in the same way to aggregate shocks. Motivated by this results, this chapter introduces an alternative methodology to estimate disaggregated IRFs using the C-Lasso framework which asymptotically eliminates the misclassification bias without the need for the researcher to take a stance on individual group membership. The proposed estimator is used to revisit the dynamic responses of firm-level debt to an aggregate investment specific technology shock.

Acknowledgements

I owe a debt of gratitude to my supervisor, Ricardo Reis, for patiently guiding me through the process of writing this thesis and learning the craft of economic research. He has been a true role model. His passion for research, dedication to students and colleagues along with an extensive knowledge of the field, are some of the attributes I hope to take with me in my future endeavours. I also thank Silvana Tenreyro for her supervision during my MRes year. Her natural optimism and focus on the big picture were very inspiring to me at a time when I was taking my first steps in research.

This thesis benefited from the generous help of various faculty members and students at the LSE. I specially thank Charlie Bean, Francesco Caselli, Andreas Ek, Wouter Den Haan, Chao He, Ethan Ilzetzki, Xavier Jaravel, Nobuhiro Kiyotaki, Sevim Kösem, Per Krusell, William Matcham, Benjamin Moll, John Moore, Tsogsag Nyamdavaa, Łukasz Rachel, Kevin Sheedy and Shengxing Zhang. Teaching was an integral part of my life at the LSE and for the opportunity of teaching in their courses I thank Marcia Schafgans, Taisuke Otsu, Tatiana Komarova, Steve Pischke and Greg Fischer – I hope my future students will benefit from many of the lessons I learned from you.

Completing this PhD would have been all the more difficult were it not for the friendship provided by Andrea Alati, Patrick Coen, Dita Eckardt, Martina Fazio, Friedrich Geiecke, Jay Euijung Lee, Wolfgang Ridinger, Claudia Robles-Garcia and Celine Zipfel. I thank Thomas Drechsel for playing the role of an older brother in the PhD programme, Laura Castillo-Martinez for our many coffees and laps around Lincoln's Inn Fields and Adrien Bussy for being an inseparable companion and supporting me through the many ups and downs of the PhD life. A special thanks to my friends Rafael Cervelli and Daniel de Lima for reminding me there was a life outside the PhD and for making me feel closer to home.

The support of my family has been a key input in this process. I thank my

parents, Ana and Luis, for encouraging me to pursue whatever made me happy and to inspire me to always aim to give the best version of myself. A heartfelt thanks to my parents-in-law, Sonia and José, for coming to London whenever they could and for supporting me like a true son. My most profound gratitude goes to my grandparents, Graça and Miguel. To my grandmother for always being there for me since the moment I was born. To my grandfather for instilling in me the love for economics and for teaching me so many lessons that formed my character – I wish you could be here to celebrate this moment with me.

Finally, reaching this milestone would not have been possible without the unconditional love and support of my wife, Silvia. Words fall short in describing how grateful I am of having her by my side. This thesis is dedicated to her.

Contents

List of Figures

List of Tables

Chapter 1

State-Space Modeling of Dynamic (S,s) Economies

1.1. Introduction

At the microeconomic level many changes can be described as intermittent and lumpy. Retail prices usually remain unchanged for several months before they change by large amounts. The same pattern has been documented for several variables including inventory orders, individual cash-withdrawals, plant-level investment and firm's hiring and firing decisions. This empirical pattern led to the widespread application of (S,s) policies, formally justified by the existence of fixed adjustment costs, to a variety of economic problems. The fundamental variable driving individual decisions under those policies is the so-called state gap, understood as the difference between a state variable that determines individual payoffs and its *frictionless* counterpart, that is, the value that would maximise individual payoffs for that time period. The behaviour implied by those policies consists in adjusting the state variable to bring the state gap to a pre-specified level when the current state gap is sufficiently large and leaving the state variable unchanged otherwise.

Even though the intermittent and lumpy changes observed in disaggregated data can be rationalised by microeconomic units pursuing (S,s) policies it is *a priori* unclear what are implications of those policies for economy-wide outcomes. Does the adoption of (S,s) policies at the microeconomic level affect the dynamics of aggregate variables? Does it affect their response to shocks? What are the welfare costs induced by presence of those policies? From an empirical perspective, the key information to answer these questions is the values of each unit's *frictionless* state variables. These are, by construction, the values that would have been observed for each unit's state variable in a counterfactual world where (S,s) policies were not adopted. Unfortunately these hypothetical values are not directly observed in the data and, on top of that, the adoption of (S,s) policies makes inference about these values particularly difficult for outside observers because periods of inaction act as a store of private information about the unit's frictionless state variable values (Caplin and Leahy, 2010).

This paper introduces a statistical framework to study economies that are composed of microeconomic units adopting (S,s) policies. This framework is designed to enable researchers to estimate cumulated changes in each unit's frictionless state variable for any time period based *only* on data on each unit's actual state variable values. The problem researchers face is illustrated in figure 1.A.1. For a given microeconomic unit, the blue-starred line represents a sequence of cumulated changes in its frictionless state variable generated from an exogenous process. Assuming this unit adopts (S,s) policies, the implied path for cumulated changes in actual state variable is given by the black solid line, which changes when its distance to the blue starred-line crosses certain thresholds (black dashed lines) and remains constant otherwise. In actual data researchers only observe a panel black solid lines. The statistical framework proposed in this paper is designed to enable researchers to go from a panel of observed black lines to a panel of estimates of the blue-starred lines, which are the ultimate object of interest.

To solve this statistical inversion problem, this paper proceeds in three logical steps. The first step is to introduce a process that is assumed to generate actual state variable values in an economy where each microeconomic unit adopts a two-sided (S,s) policy subject to costless adjustments opportunities. Building from the continuous-time literature on (S,s) models, the cumulated changes in each unit's frictionless state variable are assumed to evolve according to a random walk with drift whilst costless adjustment opportunities are assumed to arrive according to Bernoulli random draws. Combining these two processes specifying the evolution of the unobserved variables and (S,s) policies to map these unobserved variables to actual state variable values yields a nonlinear and non-Gaussian state-space representation of the data generating process. This representation is semi-structural and can accommodate some classic problems that have been analysed through the lens of (S,s) models such as pricing subject to menu costs, cash-management, plant-level investment and labor hiring and firing decisions.

The second step consists in devising a procedure that makes possible estimation of unknown parameters that enter the data generating process by using information on actual state variable values observed by the researcher. In the literature on (S,s) models it has become commonplace to estimate unknown parameters based on moment conditions involving changes in observed state variables. The estimation procedure here proposed follows this tradition but divides the estimation in two stages. The first-stage uses moment conditions that do *not* depend on the initial cross-sectional distribution of state gaps to estimate all parameters bar unit-specific initial state gaps. In the second-stage, parameters estimated in the first-stage are kept fixed and a different set of moment conditions is then used to estimate the initial cross-sectional distribution of state gaps. Given the proposed state-space representation does not admit closed-form solution for moment conditions involving state variable changes, estimation in both stages is conducted via Simulated Method of Moments (SMM).

The third step involves solving a smoothing problem, that is, solving for the probability density function of the latent variables of interest (*i.e.* cumulated changes in each unit's frictionless state variable) *conditional* on all the state variable values observed by the researcher and a set of values for the unknown parameters. Despite the marked nonlinearities and non-Gaussianity of the data generating process, this paper explores a local version of the forward-filtering and backward-smoothing recursion techniques in Kitagawa (1987, 1994, 1996) to obtain an exact closed-form expression for the smoothed probability density function. Finally, by setting parameter values equal to their estimates from the second step, the *smoothed estimates* of cumulated changes in each unit's frictionless state variable are obtained by computing expectation of the smoothed probability density function. These smoothed estimates can then be combined with the estimated initial cross-sectional distribution of state gaps to produce estimates of the cross-sectional distribution of state gaps in the economy at any point in time.

Lastly, the properties of the proposed estimator are illustrated through a Monte Carlo experiment. Samples of artificial data are generated from the assumed data generating process and, in each of these samples, smoothed estimates for each unit's cumulated changes in frictionless state variable are obtained and compared against the estimates obtained from two alternative estimators. The first alternative estimator is based on an incomplete expression for the smoothed probability density function that holds only in a special case. The second alternative estimator does not make use of the information contained in the state variable values observed in the data. Across a variety of data generating process specifications and sample sizes, the calculated mean squared error (MSE) for smoothed estimates is between 5 and 55% smaller than the MSE of the first alternative estimator and more than a full order of magnitude smaller than the MSE of the second alternative estimator.

Relation to the literature. This paper relates and contributes to two strands of literature. First, it relates broadly to an extensive literature that has employed (S,s) rules to model individual decisions underlying the intermittent and lumpy adjustments observed in dissagregated data. In particular, it relates to other papers that have explicitly attempted to estimate frictionless state variables and state gaps. In the context of plant-level labor adjustments, Caballero, Engel and

Haltiwanger (1997) use information on hours worked to recover labor gaps at the plant level.[1] In the context of pricing, Campbell and Eden (2014) and Carvalho and Kryvstov (2018) use information contained in prices of similar quote-lines to estimate frictionless prices. In a similar environment to the one considered in the present paper, Baley and Blanco (2020) provide an analytical mapping between moments of the cross-sectional distribution of observed state variable adjustments and moments of the cross-sectional distribution of gaps in steady state. There are three aspects of the statistical framework here proposed that set it apart from the existing literature. First, the estimates of cumulated changes in frictionless states are formally grounded on the theory of filtering and smoothing which contrasts with alternative estimators that lack any statistical underpinning. Second, the framework here proposed is semi-structural and can accommodate many problems that have been analysed through the lens of (S,s) models. Third, the estimator here introduced produces estimates of cumulated changes in frictionless inflation for each unit using only information contained in the observed state variable values.

Second, this paper is relates to a vast and long-standing literature on nonlinear and non-Gaussian filtering and smoothing. The intersection of this literature with economics is mostly confined to the estimation of nonlinear dynamic stochastic general equilibrium models (Herbst and Schorfheide, 2015), however, it has been fruitfully applied to tackle a wide range of problems spanning different fields including GPS tracking, brain imaging, audio signal processing and autonomous navigation (Chen, 2003; Särkkä, 2013). In linear and Gaussian state-space representations exact closed form solutions to the filtering and smoothing problems are given respectively by the celebrated Kalman filter (Kalman, 1960) and the Rauch-Tung-Striebel smoother (Rauch, Tung and Striebel, 1965). However, for nonlinear and/or non-Gaussian state-space representations filtering and smoothing problems in general do not admit closed form solutions and approximate

[1] For further discussion of this approach refer to Cooper and Willis (2004), Caballero and Engel (2004), Bayer (2009) and Cooper and Willis (2009).

solutions have to be obtained numerically.[2] This paper contributes to the literature on nonlinear and non-Gaussian filtering and smoothing by providing exact closed-form solutions for the filtering and smoothing problems given a state-space representation of a dynamic (S,s) economy. The fact that closed-form solutions are derived is particularly important, not just because they are exact, but also because they allow smoothed estimates to be easily calculated in a setting where most of the existing smoothing algorithms either do not apply due to the discontinuities in the measurement equation (*e.g.* smoothers of the "Kalman family") or are computationally too expensive to be implemented on a large scale (*e.g.* Markov chain Monte Carlo or particle filters).

Structure of the paper. Section 1.2 presents the state-space representation that is assumed to generate the data that is observed by the researcher. Section 1.3 introduces a two-stage estimation procedure to estimate the unknown parameters that enter the data generating process. Section 1.4 presents the exact closed-form expression for the smoothed probability density which can be used to compute smoothed estimates of the cumulated changes in each unit's frictionless state variable. Section 1.5 uses a Monte Carlo experiment to illustrate the properties of the proposed estimator. Section 1.6 concludes and discusses some promising avenues for future research.

Notation. In all that follows, bold letters are used to denote vectors or matrices and non-bold fonts denote scalars. For a function $f : \mathbb{R} \to \mathbb{R}$ and a matrix $\mathbf{A} \in \mathbf{R}^{m\times n}$, the expression $\mathbf{B} = f \circ (\mathbf{A})$ is equivalent to $b_{i,j} = f(a_{i,j})$, $\forall i, j$. The symbol $\odot$ denotes the Hadamard product, $\mathbf{1}_{\boldsymbol{m\times n}}$ and $\mathbf{0}_{\boldsymbol{m\times n}}$ denote $m \times n$ matrices of ones and zeros, $\mathbb{1}\{\cdot\}$ denotes the indicator function and $\|\cdot\|$ denotes the Euclidean norm. When working with indexed sequences of variables or vectors, $\mathbf{X}_{\boldsymbol{t}}$ is used to denote the t-th element in the sequence and $\mathbf{X}^{\boldsymbol{t}}$ refers to the subsequence $\{\mathbf{X}_{\boldsymbol{i}}\}_{i=0}^{t}$. For any two real numbers $a < b$, denote by $\mathbb{Z}_{[a,b]}$ the set of

[2] Several algorithms have been proposed in the literature, including: filters and smoothers of the "Kalman family" (such as extended and unscented Kalman filters and smoother), grid based approximation methods, sequential Monte Carlo methods or particle filters and smoothers. All of these alternative algorithms are covered in Särkkä (2013). An excellent survey of particle filters in particular is Doucet and Johansen (2011).

all integers in $[a, b]$. Random variables or vectors are denoted by upper case letters whilst their realisations are denoted by their lower case counterparts. For a continuous (discrete) random vector $\mathbf{X}$, the notation $f_{\mathbf{X}|\mathbf{Y}}(\boldsymbol{x}|\boldsymbol{y})$ is used to denote its probability density (mass) function evaluated at a specific value $\boldsymbol{x}$ conditional on the random vector $\mathbf{Y}$ taking the value $\boldsymbol{y}$. The function $\delta(\cdot)$ is used to denote the Dirac delta function whereas the function $\phi(\cdot)$ denotes the standard normal probability density function.

1.2. A state-space representation of dynamic (S,s) economies

A foundational assumption for the framework developed in this paper is that the researcher observes a panel of data on state variable values that is generated by units adopting a two-sided (S,s) policy subject to costless adjustment opportunities. This section formally introduces that assumption, discusses some of its main aspects and how it can accommodate some well known problems in the (S,s) literature.

1.2.1. Two-sided (S,s) policies subject to costless adjustment opportunities

Consider an economy composed of n units in which each unit chooses every period the value of a single state variable that determines its payoffs. Let $z_{i,t}$ denote the value of unit i's state variable at time t. It is assumed that $z_{i,t}$ is chosen according to the following policy function,

$$z_{i,t}\left(z_{i,t-1}, z^{\star}_{i,t}, \ell_{i,t}\right) = \begin{cases} z_{i,t-1}, \text{ if } d_{i,t}\left(z_{i,t-1}, z^{\star}_{i,t}, \ell_{i,t}\right) = 1 \\ z^{\star}_{i,t} + c_i, \text{ if } d_{i,t}\left(z_{i,t-1}, z^{\star}_{i,t}, \ell_{i,t}\right) = 0 \end{cases} \tag{1.1}$$

where,

$$
\begin{aligned}
d_{i,t}\left(z_{i,t-1}, z^{\star}_{i,t}, \ell_{i,t}\right) = {} & \underbrace{\mathbb{1}\left\{z_{i,t-1} - z^{\star}_{i,t} \in (L_i, U_i)\right\}(1 - \ell_{i,t})}_{\substack{\text{=1 if state gap at the previous state value} \\ \text{is inside the inaction region and state} \\ \text{variable cannot be changed costlessly}}} \\
& + \underbrace{\mathbb{1}\left\{z_{i,t-1} - z^{\star}_{i,t} = c_i\right\}\ell_{i,t}}_{\substack{\text{=1 if state variable can be changed} \\ \text{costlessly but state gap at the current} \\ \text{value is already at the reset value}}}
\end{aligned}
\tag{1.2}
$$

and $z^{\star}_{i,t}$ denotes the *frictionless* value of unit i's state variable at time t, that is, the value of $z_{i,t}$ that would be chosen in the absence of (S,s) rules; the open interval (L_i, U_i) denotes the *inaction region*; the parameter $c_i \in (L_i, U_i)$ denotes the *reset value* and $\ell_{i,t}$ is a dummy variable equal to one if unit i at time t can make a *costless adjustment*.

The *state gap*, defined as the difference between a state variable and its frictionless counterpart, is the fundamental variable driving each unit's state variable choices. Equation (1.1) implies that, at every time period, a unit either leaves its state variable unchanged or chooses a new value such that its state gap equals the reset value. In addition, the two terms in the right-hand-side of equation (1.2) imply that there are two instances where a unit choses *not* to change its state variable. The first instance is when it cannot adjust costlessly and by leaving its state variable unchanged the resulting state gap lies within the inaction region. The second instance is when it can adjust costlessly but by leaving its state variable unchanged its state gap is equal to the reset value.

Timing. The implicit timing assumption in (1.1) and (1.2) is that each period starts with the realisation of any exogenous shocks that determine $z^{\star}_{i,t}$ and $\ell_{i,t}$ and ends with each unit's state variable choices. This timing convention is commonplace in (S,s) models cast in discrete-time (*e.g.* Caballero, Engel and Haltiwanger, 1997, p.118).

Microfoundation. In this paper two-sided (S,s) policies subject to random costless adjustment opportunities are taken exogenously and all results presented henceforth are independent of the specific reason why individual units would choose to adopt such rules. Starting with Arrow, Harris and Marschak (1951) and Scarf (1959) there is a long-standing literature showing that (S,s) policies tend to arise optimally in situations with three defining features: a state variable that affects flow payoffs, fixed costs of exerting control over this state variable and a driving force that causes the state to drift absent control.[3,4] Notice however that, in the spirit of Caballero and Engel (1991), the results here presented also apply to a broader class of problems where (S,s) rules are not optimal but can be justified as either simple rules that approximate more complex first best rules or as arising from nearly rational behavior. Finally, the two-sided (S,s) policies presented in (1.1) and (1.2) also include costless adjustment opportunities. In the literature these opportunities have been incorporated into (S,s) models to give them additional flexibility and, in particular, to allow the size of state variable adjustments to vary both across units and for the same unit over time as it is typically observed in the data.[5]

1.2.2. The data generating process

This subsection defines important variables and notation that will be used throughout the rest of the paper, then it presents the state-space representation that is assumed to be generating the data observed by the researcher and discusses some

[3] In continuous time, Stokey (2009, chapter 7) and Plehn-Dujowich (2005) present environments in which two-sided (S,s) policies with a single reset point are optimal. In discrete time, it is in general difficult to analytically characterise the policy function for problems involving fixed adjustment adjustment costs but there are several precedents in the literature in which two-sided (S,s) rules with a single reset point are either assumed or derived as an approximate solution for the original problem. Some examples of such precedents include Caballero and Engel (1999) in the context of plant-level investment, Gertler and Leahy (2008) for price setting subject to menu costs and Elsby and Michaels (2019) for firm's hiring and firing decisions.

[4] It is important to notice that the type of two-sided (S,s) policies with a single reset point as assumed in (1.1) and (1.2) tend to arise optimally when the fixed cost is the *only* cost of exerting control. Alternatively, in situations where exerting control over the state involves both fixed and a proportional components the optimal policy takes the form of a two-sided (S,s) policy with *two* reset points. For a formal proof of these type of rule when adjustment involves both fixed and proportional costs, refer to Harrison, Sellke and Taylor (1983) or Stokey (2009, chapter 8).

[5] The costless adjustment opportunities here considered are a special case of a broader class of (S,s) models, commonly referred to as generalised (S,s) models or second-generation state-dependent models, in which the fixed costs are allowed to vary both across units and over time. This type of formulation was originally introduced in the context of plant-level investment by Caballero and Engel (1999) and in the context of pricing by Dotsey, King and Wolman (1999).

its most important aspects.

Terminology and notation. First, let the *cumulated change in unit i's state variable at time t* be denoted by $\mathrm{Z}_{i,t} \equiv z_{i,t} - z_{i,0}$ and, likewise, the *cumulated change in unit i's frictionless state variable at time t* be denoted by $\mathrm{Z}^{\star}_{i,t} \equiv z^{\star}_{i,t} - z^{\star}_{i,0}$. Second, let the *re-centered inaction region* be denoted by $(\underline{x}_i, \overline{x}_i)$ where $\underline{x}_i \equiv L_i - c_i < 0$ and $\overline{x}_i \equiv U_i - c_i > 0$. Lastly, let the *re-centered state gap* be denoted by $x_{i,t} \equiv z_{i,t} - z^{\star}_{i,t} - c_i$.

***Assumption* 1.1** *The researcher observes a sequence of vectors* $\mathbf{Z^T} = \{\mathbf{Z_1}, \ldots, \mathbf{Z_T}\}$ *generated by the following state-space representation,*

$$\mathbf{Z_t} = \mathbf{Z_{t-1}} \odot \mathbf{d_t} + \left(\mathbf{Z^{\star}_t} - \boldsymbol{x_0}\right) \odot \left(\mathbf{1}_{\boldsymbol{n}\times\mathbf{1}} - \mathbf{d_t}\right) \tag{1.3}$$

$$\mathbf{d_t} = \mathbb{1} \circ \left\{\mathbf{Z_{t-1}} - \mathbf{Z^{\star}_t} + \boldsymbol{x_0} \in (\underline{\boldsymbol{x}}, \overline{\boldsymbol{x}})\right\} \odot \left(\mathbf{1}_{\boldsymbol{n}\times\mathbf{1}} - \boldsymbol{\ell_t}\right) \;+\; \mathbb{1} \circ \left\{\mathbf{Z^{\star}_t} = \mathbf{Z_{t-1}} + \boldsymbol{x_0}\right\} \odot \boldsymbol{\ell_t} \tag{1.4}$$

$$\mathbf{Z^{\star}_t} = \boldsymbol{\mu} + \mathbf{Z^{\star}_{t-1}} + \boldsymbol{\varepsilon_t} \tag{1.5}$$

$$\boldsymbol{\ell_t} = \mathbb{1} \circ \{\boldsymbol{\nu_t} \leqslant \boldsymbol{\lambda}\} \tag{1.6}$$

where $\boldsymbol{\varepsilon_t} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ *with* $\boldsymbol{\Sigma} = \mathit{diag}\left(\sigma^2_{\varepsilon,1}, \ldots, \sigma^2_{\varepsilon,n}\right)$ *and iid across time,* $\boldsymbol{\nu_t}$ *is such that* $\nu_{i,t} \sim$ *Uniform*$(0,1)$ *and iid across units and time and* $\mathbf{Z_0} = \mathbf{Z^{\star}_0} = \mathbf{0}_{\boldsymbol{n}\times\mathbf{1}}$*. The vectors* $\boldsymbol{x_0}$*,* $\underline{\boldsymbol{x}}$*,* $\overline{\boldsymbol{x}}$*,* $\boldsymbol{\mu}$*,* $\boldsymbol{\lambda} \in \mathbb{R}^n$ *and the matrix* $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ *contain parameter values that are unknown to the researcher.*

A semi-structural state-space representation. The state-space representation in assumption 1.1 is *semi-structural.* Although its form could be rationalised from an environment where individual units face fixed adjustment costs, the mapping between parameters in assumption 1.1 and other deep structural parameters is left *unspecified.* There are four reasons why a semi-structural representation is chosen over a fully structural one. First and foremost, for the main objective of this paper, which is to introduce an estimator of $\mathbf{Z}^{\star}_t$'s based on observations of $\mathbf{Z}_t$'s, the representation in assumption 1.1 is sufficient. Second, since there exists more than one environment that can give rise to this representation, a semi-structural representation can more easily accommodate different problems that have been considered in the (S,s) literature. Third, a semi-structural representation naturally tends to provide a better fit of the data since it does not impose cross-restrictions on parameters. Fourth, a semi-structural representation greatly reduces the computational burden of the simulation-based parameter estimation procedure that will be proposed in section 1.3.[6]

Representation in cumulated changes instead of levels. Although the data generating process could be equivalently represented in terms of state variable levels and their frictionless counterparts, assumption 1.1 expresses it in terms of *cumulated changes* in those variables. This representation is chosen over the one in levels because, in the former, individual reset points do not enter as parameters. This is an advantage due to the difficulties in identifying reset points from data on state variable adjustments.[7] Nonetheless, considering a representation in terms of cumulated changes implies that, without further assumptions on reset point values or initial frictionless state variable, this framework only allows the estimation of the cross-sectional distributions of the *re-centered* state gaps and the construction of frictionless aggregates for which the aggregator function could be written as a function of cumulated changes.[8]

[6] The reduction in computational burden comes from the fact that for a given combination of parameter values, state-space representation in assumption 1.1 can be directly used simulate data on $\mathbf{Z}_t$'s without solving an inter-temporal optimisation problem first. Given the non-convexities that arise in the presence of fixed adjustment costs, solving this inter-temporal optimisation problem typically requires the use of global methods such as value function interaction which can be costly in terms of computational time.

[7] In a continuous-time environment, similar to the single product random menu cost model in Stokey (2009) and Alvarez, Le Bihan and Lippi (2016), it is possible to derive a closed form expression for the steady state distribution of state variable changes where the inaction region boundaries enter *only* in deviation from the reset point, that is, the steady state distribution of state variable changes is a function of $L_i - c_i$ and $U_i - c_i$. This implies that, in that environment, there are no moments of that distribution that could be used to separately identify L_i, c_i and U_i. In a version of state-space representation in assumption 1.1 cast in levels it is in general not possible to derive the implied distribution of state variable changes, but some early attempts to estimate parameters in that representation where also indicative that the boundaries of the inaction region and the reset point cannot be separately identified. The impossibility to separately identify the boundaries of the inaction region and the reset point also appears in the two-sided (S,s) model of Bonomo, Correa and Medeiros (2013).

[8] More precisely, consider an aggregate variable that is a function of a sequence of observed state variables, that is, $A_t = a\left(\mathbf{z}^{\boldsymbol{t}}\right)$ where $a(\cdot)$ is the aggregator function. Working with a representation in terms of cumulated changes as in assumption 1.1, implies that this framework would only allow the construction of frictionless aggregates for which the aggregator functions that could be equivalently represented as a function of observed sequence of cumulated changes in state variables. For example, in the context of pricing A_t could be a price index and the

Frictionless *versus* reset states. A maintained assumption is that each unit's state variable choices are driven by the *state gap* which is *defined* as the difference between each unit's state variable and its *frictionless* value (*i.e* the value that would be observed if (S,s) policies where *permanently* removed). In some specifications of (S,s) models, it is assumed instead that adjustment decisions are driven by the *reset state gap* which is defined as the difference between each unit's state variable and it's *reset* or *mandated* value (*i.e.* the value that would be observed if (S,s) policies where *momentarily* removed).[9] It is important to notice that, although frictionless and reset values are conceptually different, in the presence of two-sided (S,s) policies with a single reset point their difference is *constant* over time. This implies that for the state-space representation in assumption 1.1 this difference is immaterial, since $\mathbf{Z}^{\star}_{t}$ is equal to the cumulated change in reset state variable and re-centered state gaps are equal to reset state gaps.[10]

Transition equation for frictionless state. In equation (1.5) it is assumed that $\mathbf{Z}^{\star}_{t}$ evolves according to a random walk with drift. There is two reasons underlying that assumption. First, it is the discrete time analogue of a Brownian motion with drift which is the most commonly assumed process for the frictionless states in (S,s) models cast in continuous time. Second, assuming a linear transition equation with normally distributed disturbances makes possible to obtain exact closed-form solutions for the filtered and smoothed densities of $\mathbf{Z}^{\star}_{t}$.[11] Finally, it is important to notice that, besides the random walk with drift assumption, it is also assumed that the frictionless states of different units evolve

aggregator function represents the aggregation used by statistical authorities to produce the index from micro prices. To produce a *frictionless* price index based on estimates of $\mathbf{Z}^{\star}_{t}$'s, one needs to be able to write the aggregator as a function of the sequence of cumulated changes.

[9] See, for instance, Caballero, Engel and Haltiwanger (1995).

[10] The reset state variable value is given by $z^{r}_{i,t} = z^{\star}_{i,t} + c_i$, that is, the value that would make state gap in current period equal to its reset value. From that it follows that $\mathrm{Z}^{\star}_{i,t} = \mathrm{Z}^{r}_{i,t}$ and $x_{i,t} = z_{i,t} - z^{r}_{i,t}$. This point was made in the context of pricing by Bonomo *et al.* (2013).

[11] It is useful to think of this assumption as a *weaker* version of the linearity and Gaussianity requirements necessary to obtain the Kalman filter (Kalman, 1960) and the Rauch-Tung-Striebel smoother (Rauch, Tung and Striebel, 1965) as exact solutions for the filtering and smoothing problems, respectively. In this paper, closed-form expressions for the filtered and smoothed probability densities of $\mathbf{Z}^{\star}_{t}$ are derived despite the marked non-linearities in (1.3), (1.4) and (1.6) and the non-Gaussianity of the shocks $\boldsymbol{\nu}_{t}$ determining the arrival of costless adjustment opportunities.

independently of each other.[12] The motivation behind this assumption is to keep the filtering and smoothing problems tractable. In particular, as we shall see in section 1.4, this assumption allows me the joint filtered and smoothed probability densities of $\mathbf{Z}_t^\star$ to be obtained by solving the filtering and smoothing problems for each unit *separately* which greatly reduces the dimensionality of the problem at hand.

Transition equation for costless adjustment opportunities. The dummy vector $\boldsymbol{\ell}_t$ takes the value one for units that can adjust their state variables costlessly in period t and zero for units that cannot. Equation (1.6) determines that in any given time period, a unit i receives a costless adjustment opportunity with probability λ_i. Moreover, as it is commonly done in the literature, it is assumed that the arrival of such opportunities is independent across individuals and time and also independent of the state gap.[13]

The measurement equation. The mapping between the two latent unobserved variables, cumulated changes in frictionless states and costless adjustment opportunities, and cumulated changes in states observed by the econometrician is provided by equations (1.3) and (1.4). Notice that the i-th element of $\mathbf{Z}_t$ and $\mathbf{d}_t$ is given by (1.1) and (1.2) re-expressed in terms of cumulated changes and re-centered gaps and inaction regions.[14]

[12] This is a strong assumption as it rules out the presence of any shocks that are common across units. To accommodate the presence of common shocks one would need to consider a variance-covariance matrix $\boldsymbol{\Sigma}$ that is symmetric and positive definite but not necessarily diagonal. I am actively working on the generalization of all the results in this paper under that weaker assumption.

[13] In models where the fixed costs are allowed to stochastically vary over time, the common assumption is that each period units draw a fixed adjustment cost and those draws are independent across units and time. See Caballero and Engel (1999) in the context of investment and Dotsey, King and Wolman (1999) and Costain and Nakov (2011a,b) in the context of pricing decisions.

[14] In order to go from (1.1) and (1.2) to (1.3) and (1.4) simply subtract $z_{i,0}$ on both sides of (1.1) and note that using the definitions of $\mathrm{Z}_{i,t}$ and $\mathrm{Z}_{i,t}^\star$ the re-centered state gap can be equivalently written as $x_{i,t} = \mathrm{Z}_{i,t} - \mathrm{Z}_{i,t}^\star + x_{i,0}$.

1.2.3. Applications

This subsection concludes with a discussion of some well known problems in the (S,s) literature that can be accommodated by the data generating process in assumption 1.1.

Pricing subject to menu costs. By letting $z_{i,t}$ denote the (log of) nominal price charged by firm i and $z^{\star}_{i,t}$ denote the (log of) nominal frictionless prices (*i.e.* the price that would maximize firm i's flow of profits in period t), the state-space representation in assumption 1.1 is capable of accommodating some well known models of pricing under menu costs. First, it can accommodate a single product random menu cost model of pricing in which firms can draw either a positive or a zero cost of adjustment as in Alvarez, Le Bihan and Lippi (2016).[15] Second, by imposing $\underline{\boldsymbol{x}} \to -\infty$ and $\overline{\boldsymbol{x}} \to \infty$ each firm adjusts its prices at random time intervals as in Calvo (1983) model of staggered price setting. Third, by imposing $\boldsymbol{\lambda} = \mathbf{0}_{\boldsymbol{n}\times\mathbf{1}}$ each firm adopts a two-sided (S,s) pricing policy like those that arise in the canonical menu cost model of Golosov and Lucas (2007).[16]

Cash withdrawals problem. If $z_{i,t}$ denotes the accumulated cash withdrawals by individual/firm i and $z^{\star}_{i,t}$ denotes the accumulated cash expenditures by individual/firm i, then the state gap corresponds for individual/firm i current cash balance and the state-space representation in assumption 1.1 can accommodate some cash-balance problems. For example, it can represent an economy in which individuals adopt optimal cash withdrawal policies that arise in Alvarez and Lippi (2009) model of demand for cash in which individuals receive random opportunities of withdrawing cash for free. Moreover, by imposing $\boldsymbol{\lambda} = \mathbf{0}_{\boldsymbol{n}\times\mathbf{1}}$ each unit would choose its cash holdings according to a two-sided (S,s) rule like the one

[15] This type of model is also considered by Stokey (2009), Blanco (2017), Luo and Villar (2017) and Gautier and Le Bihan (2018) and it can be seen as a special case of the *CalvoPlus* model in Nakamura and Steinsson (2010) where firms draw either a low or a high menu cost at the beginning of each period.

[16] If in addition to $\boldsymbol{\lambda} = \mathbf{0}_{\boldsymbol{n}\times\mathbf{1}}$, one imposes $\overline{\boldsymbol{x}} \to \infty$ then each firm in this economy follows a one-sided (S,s) pricing policy as the ones derived in earlier models of pricing subject to menu costs such as Sheshinski and Weiss (1977, 1983).

adopted in Miller and Orr (1966) model of demand for money by firms.[17]

Plant-level investment and labor adjustments. Other settings where (S,s) policies have been extensively used is plant level investment and labor hiring and firing decisions. If $z_{i,t}$ denotes the capital/labor stock of plant i and $z^{\star}_{i,t}$ its frictionless counterpart (*i.e.* the capital/labor stock that would maximize plant i's flow of profit in period t), the representation in assumption 1.1 could also accommodate some well known models of lumpy plant-level investment and hiring and firing decisions. In the context of plant-level investment, assumption 1.1 could be interpreted as the data generating process for the *Bernoulli fixed-cost* environment in Baley and Blanco (2020). Moreover, in the special case without random costless adjustment opportunities ($\boldsymbol{\lambda} = \mathbf{0_{n\times 1}}$) that representation can accommodate environments with one or two-sided (S,s) rules environments considered by Bertola and Caballero (1994) and Caballero, Engel and Haltiwanger (1995), respectively. In the context of hiring and firing decisions two-sided (S,s) rules without random adjustment opportunities are also adopted by Elsby and Michaels (2019) and Elsby, Michaels and Ratner (2019).

1.3. Parameter Estimation via Simulated Method of Moments

For each unit there is a total of six parameters that are unknown to the researcher, namely, the initial re-centered state-gap ($x_{i,0}$), the boundaries of the re-centered inaction region ($\underline{x}_i$ and $\bar{x}_i$), the drift (μ_i) and the variance of shocks ($\sigma^2_{\varepsilon,i}$) in the transition equation for cumulated changes in frictionless state variable and the probability of arrival of a costless adjustment opportunity (λ_i). This section presents a two-stage procedure that allows the estimation of these parameters

[17] If in addition to $\boldsymbol{\lambda} = \mathbf{0_{n\times 1}}$, one imposes $\overline{\boldsymbol{x}} \to \infty$ then individuals/firms adopt one-sided (S,s) policies like those used in Frenkel and Jovanovic (1980) model of precautionary demand for money. In the limiting case with one-sided (S,s) policies and with non-stochastic cash withdrawals ($\sigma_{\varepsilon,i} \to 0$, $\forall i$), the state-space representation in assumption 1.1 could also be interpreted as an economy-wide version of the early-environments considered Baumol (1952) and Tobin (1956). Obviously, in a non-stochastic environment the problem of estimating $\mathbf{Z}^{\star}_t$ boils down to the problem of estimating the vector $\boldsymbol{\mu}$ from a set of observations $\mathbf{Z}^{\boldsymbol{T}}$.

based on actual state variable values observed by the researcher. A necessary requirement to be able to estimate those parameters at the *unit level* is that a sufficiently large number of state variable adjustments is observed for that unit in the data. This section starts by describing the estimation procedure for an arbitrary unit that is assumed to satisfy that requirement and then discusses how it can be adapted to situations where that requirement is not satisfied.

Terminology and notation. The variable $\Delta \mathrm{Z}_{i,t}$ is referred to as a state variable *change* regardless of the value it takes. The terminology *state variable adjustment* refers specifically to a *non-zero* state variable change. The vector of unit-specific parameters is denoted by $\Theta_i \equiv \{x_{i,0}, \Gamma_i\}$ where $\Gamma_i \equiv \left\{\underline{x}_i, \bar{x}_i, \mu_i, \sigma^2_{\varepsilon,i}, \lambda_i\right\}$. Denote by $\left\{\mathrm{Z}^s_{i,t}(x_{i,0}, \Gamma_i)\right\}_{t=0}^{T}$ a simulated time-series for unit i generated by the state-space representation in assumption 1.1 using unit specific parameters equal to $x_{i,0}$ and Γ_i.

1.3.1. A two-stage estimation procedure

The fundamental idea underlying the two-stage procedure is to use different sets of moment conditions to estimate Γ_i *separately* from $x_{i,0}$. The advantages of splitting parameter estimation in two stages are twofold. First, it reduces the dimensionality of the estimation problem at hand. Second, as it will later become clear, for a given unit this procedure only requires one state variable adjustment to be observed in the data to the estimate $x_{i,0}$, even in contexts where the total number of state variable adjustments observed is not sufficient to estimate all the parameters at the unit level. Formally, the choice of moment conditions for each of the stages is grounded on the following result,

***Proposition* 1.1** *Consider a sequence $\{\mathrm{Z}_{i,t}\}_{t=0}^{\mathrm{T}}$ generated by the data generating process in assumption 1.1 and assume it contains at least one state variable adjustment. Let τ_i^1 denote the period at which the first adjustment occurs. Then, the subsequence $\{\Delta \mathrm{Z}_{i,t}\}_{t=\tau_i^1+1}^{\mathrm{T}}$ does not depend on $x_{i,0}$.*

Proof. See appendix 1.D. □

The intuition for this result is illustrated in figure 1.A.2. It formalises the fact that, under two-sided (S,s) policies, an adjustment period is, by construction, a period where the re-centered state variable gap is set to zero. From there onwards the decisions of whether to adjust or not and the respective size of adjustment are independent of any re-centered state variable gap values before that occurred before the last adjustment.

Given proposition 1.1, parameter estimation procedure proceeds as follows. The first-stage uses moment conditions computed from the subsequence $\{\Delta \mathrm{Z}_{i,t}\}_{t=\tau_i^1+1}^{\mathrm{T}}$ to estimate all the parameters except for $x_{i,0}$. The second-stage uses moment conditions computed from the subsequence $\{\Delta \mathrm{Z}_{i,t}\}_{t=1}^{\tau_i^1}$ to estimate $x_{i,0}$ whilst keeping the remaining parameters equal their first-stage estimates. Since there is no closed form solution for the moment conditions implied by the state-space representation in assumption 1.1, the estimation in both stages is done via Simulated Method of Moments (SMM).[18]

First stage. The vector of parameters Γ_i is estimated from,

$$\widehat{\Gamma}_i = \arg\min_{\Gamma_i} \left\| \mathbf{\Omega}^{\frac{1}{2}} \left[g\left(\{\Delta \mathrm{Z}_{i,t}\}_{t=\tau_i^1+1}^{\mathrm{T}}\right) - \frac{1}{\mathcal{S}} \sum_{s=1}^{\mathcal{S}} g\left(\{\Delta \mathrm{Z}_{i,t}^{s}(\Gamma_i)\}_{t=\tau_i^{1,s}+1}^{\mathrm{T}}\right) \right] \right\|^2 \tag{1.7}$$

where $\mathbf{\Omega}$ is a positive definite weight matrix and $g(\cdot)$ is a vector-valued function containing the *frequency of state variable adjustments* and *four or more percentiles of the distribution of state-variable adjustments.* The summation term in the above expression is called a *simulator* and it approximates the unknown mapping between parameters and the moment conditions chosen for estimation by averaging moments over $\mathcal{S}$ individual time-series simulated from the state-space representation in assumption 1.1. A distinguishing aspect of this simulator

[18] See Gouriéroux and Monfort (1996), Adda and Cooper (2003) and Davidson and MacKinnon (2004).

is that it does not depend on some parameters that enter the data generating process, namely, the initial re-centered state variable gap. This follows directly from proposition 1.1 since the moments are computed based on the subsample of state variable changes that occur *after* the first state variable adjustment. For a given number of simulations, the estimator in (1.7) is consistent as T → ∞ (see Gouriéroux and Monfort, 1996, proposition 2.3).[19]

Second stage. Initial re-centered state variable gap estimates can be obtained from,

$$\hat{x}_{i,0} = \underset{x_{i,0}}{\arg\min} \left\| \tilde{\mathbf{\Omega}}^{\frac{1}{2}} \left[h\left(\{\Delta \mathrm{Z}_{i,t}\}_{t=1}^{\tau_i^1} \right) - \frac{1}{S} \sum_{s=1}^{S} h\left(\{\Delta \mathrm{Z}_{i,t}^{s}(x_{i,0}, \widehat{\Gamma}_i)\}_{t=1}^{\tau_i^{1,s}} \right) \right] \right\|^2 \tag{1.8}$$

where $\tilde{\mathbf{\Omega}}$ is a positive definite weight matrix and $h(\cdot)$ is a function containing the *time elapsed until the first state-variable adjustment* and *the value of that first adjustment*. In contrast with the first-stage, the simulator term in (1.8) *does* depend on all the parameters that enter the data generating process because the moments are computed based on the subsample of state-variable changes that occur *up* to the first non-zero state variable change. However, when generating simulated data to calculate this simulator the values of Γ_i are kept fixed at their estimated values from the first-stage. It is important to acknowledge that, although it is expected that the two moments contained in $h(\cdot)$ are informative about the initial re-centered state variable gap, there is no theoretical guarantee that the estimator in (1.8) is consistent. This steams from the fact that, regardless of sample size, the estimates of $x_{i,0}$ in the second-stage only use information from the subsample up to the first non-zero state variable change.[20] However, as

[19] It is important to notice that the estimator is consistent for any positive definite weight matrix $\mathbf{\Omega}$. Nonetheless, efficiency gains could be achieved by choosing the weight matrix optimally (see Gouriéroux and Monfort, 1996, pp.31-34).

[20] As an analogy, consider the problem of estimating the mean of a random variable $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ by taking the first observation from a sample $\{x_1, x_2, \ldots, x_N\}$ of *i.i.d* observations, in other words, consider $\hat{\mu}_x = x_1$ as an estimator for μ_x. Despite being unbiased (in this particular case), the estimator $\hat{\mu}_x$ will not converge in probability to μ_x as there are no laws of large numbers that apply for an estimator based on a single observation.

shown in section 1.5, in artificially generated data the estimates of $x_{i,0}$ obtained from (1.8) are *on average* equal to their true values.

1.3.2. Choice of moment conditions and parameter identification

Global point identification of the parameters of interest requires that, for a given choice of moment conditions, the true parameter values are the unique minimizers of the population counterparts of the SMM objective functions. Although a formal discussion of parameter identification is beyond the scope of this paper, this subsection provides an informal discussion under what circumstances parameters are likely to be point identified for the choices of moment condition in (1.7) and (1.8).[21]

Parameter identification in the first stage. Given the data generating process in assumption 1.1, state variable adjustments can be classified in two classes. The first class, known as *time-dependent* adjustments, comprises adjustments that are triggered by the arrival of costless adjustment opportunities. The second class, known as *state-dependent* adjustments, comprises adjustments triggered by state variable gap lying outside the inaction region. For a given process governing the evolution of the frictionless state variable, the number of time-dependent adjustments is determined by the value of λ_i whereas the number of upward (downward) state-dependent adjustments is determined by the values of $\underline{x}_i$ ($\bar{x}_i$). In order for those parameters to be point identified from moments of the distribution of state-variable adjustments it is crucial that this distribution is composed by a *mixture* of time and state-dependent adjustments. Visually this requirement is equivalent to having a distribution of state variable adjustments that is either bimodal or trimodal and in which the smallest mode is negative

[21] In some instances, local parameter identification can be formally shown by checking invertibility of the Hessian of the population counterparts of (1.7) and (1.8) across the parameter space. This is particularly difficult to do in contexts like the present one where there is no closed form mapping between the true parameters and the implied moment conditions that enter the objective function.

and the largest positive. Some distributions satisfying this requirement are depicted in figure 1.A.3. Intuitively, the smallest and largest modes are important to identify the boundaries of the inaction region whilst the mass of state variable changes *within* these two modes is important to identify the probability of arrival of a costless adjustment opportunity.[22] The fact that the exact location of those peaks matters for parameter identification echoes the fact that, in the present setting, identification is achieved through the *shape* of the whole distribution of state variable adjustments and not from some specific summary statistics. With this mind, it is recommended to include in $g(\cdot)$ a large number of wide-ranging percentiles of this distribution.

Parameter identification in the second stage. In terms of parameter identification in (1.8) the inclusion of the *value* of the first state variable adjustment is very important. To understand why notice that, for a given value of other parameters, the time elapsed until the first state variable adjustment is typically concave in $x_{i,0}$ with a maximum achieved at an interior point of the inaction region.[23] In contrast, the value of the first state variable adjustment is typically strictly decreasing in $x_{i,0}$. Therefore, the value of the first state variable adjustment is useful to discriminate between the two values of $x_{i,0}$ that are consistent with a given time elapsed until the first state variable adjustment.

Diagnosis tests. Since parameter identification cannot be formally tested, following Canova (2007, pp. 207-211) two general diagnosis tests are suggested. The first one is to check invertibility of the objective function's Hessian evaluated at the estimated parameters. The second is to plot the objective functions in (1.7) and (1.8) around the estimated parameters varying one parameter at the time and check whether that function is "flat" or not. In case any of these diagnosis

[22] Other way to understand this requirement is to consider a case where true parameter values are such that all the state variable adjustments generated from the data generating process in assumption 1.1 are time-dependent ones. In that case the distribution of state-variable changes is unimodal and neither the frequency of state-variable adjustments nor the distribution of those adjustments are affected by the values of $\underline{x}_i$ and $\bar{x}_i$. In that case, the population counterpart of (1.7) will be "flat" at the true parameter values and parameters are not point identified.

[23] For the typical relationship between the initial state-variable gap and the time elapsed until the first adjustment refer to Stokey (2009, figure 5.3)

is indicative of lack of parameter identification it is important to investigate the sensitivity of the final estimates of $\mathbf{Z}_t^\star$ to different parameter values that minimize the first and second stage objective functions.

1.3.3. Number of state variable adjustments and parameter heterogeneity

The estimation procedure was described thus far under the assumption that enough state variable adjustments are observed for the arbitrary unit under consideration. In practice, and specially given the intermittent nature adjustments in the presence of (S,s) policies, it is likely that in any dataset at least some units don't satisfy this requirement. This section briefly discusses how each of the estimation stages can be adapted to handle this type of situation.

Modified first stage. This stage is the most demanding in terms of number of state variable adjustments since it requires enough adjustments after the first one such that one can meaningfully compute four or more percentiles of their distribution.[24] In cases where this is not possible, a potential solution lies in imposing the vector of parameters Γ_i is *common* across a group of units and compute moment conditions using the distribution of state variable changes *pooled* across that group of units.[25] The fundamental trade-off faced by the researcher in these situations is between obtaining a sufficiently large number of state variable adjustments whilst not missing important dimensions of heterogeneity. In the most extreme case where the researcher does not observe any other unit characteristics, parameters could be assumed to be common across all units.

[24] There is no explicit cutoff or rule of thumb to determine how many state variable changes are sufficient to compute those percentiles. In practice it is suggested not to use percentiles that are based on a distribution with less than 20 observations of state variable adjustments. In simulated data attempts to estimate parameters based on distributions with few observations led to considerably imprecise estimates.

[25] The criteria that could be used to classify units in groups for which the parameters are assumed to be common depends on what additional information is available to the researcher. For example, in the case of plants it could be assumed that parameters are common across plants producing a specific type of product or located in a certain region. In the case of prices it could be assumed parameters are common across quote-lines of narrow (or broadly) defined product category.

Modified second stage. Once the vector of first stage parameter is obtained, the second stage only requires *one* state variable adjustment in order to be able to estimate $x_{i,0}$. Although this requirement is much less demanding than the first stage one, it is possible that for some units no state variable adjustment is observed. A potential solution for these cases is to *impute* values for $x_{i,0}$ based on their estimated values for other units. For example, the researcher could either impose $x_{i,0} = \hat{x}_{j,0}$ for some other unit j for which at least one state variable adjustment is observed or, alternatively, draw a value from the distribution of estimated initial re-centered state variable gaps for other units.

1.3.4. Model fit

A maintained assumption throughout this paper is that the state variable values observed by the researcher are generated by the state-space representation in assumption 1.1. Nonetheless, once parameter estimates are obtained, one could check whether this is a sensible assumption by comparing how close targeted and non-targeted moments in simulated data are from their counterparts in actual data. Some non-targeted moments that could be used for this purpose include the mean, the variance, skewness and kurtosis of distribution state variable adjustments and the adjustment hazard functions. Finally, global specification tests could also be used (Adda and Cooper, 2003, p. 97).

1.4. Latent variable estimation

This section presents an exact closed-form expression for the probability density function of the vector of cumulated changes in each unit's frictionless state variable values conditional on all the cumulated changes observed by the researcher and a set of values for the unknown parameters. A convenient aspect of the state-space representation in assumption 1.1 is that the evolution of both observed and latent variables are *independent* across units. This implies that the desired smoothed probability density function can be obtained from the product of its marginals.[26] More formally, under assumption 1.1 it holds that,

$$f_{\mathbf{Z}_t^\star \mid \mathbf{Z}^\mathbf{T};\mathbf{\Theta}}\left(\mathbf{z}_t^\star \mid \mathbf{z}^\mathbf{T};\boldsymbol{\theta}\right) = \prod_{i=1}^{n} f_{\mathrm{Z}_{i,t}^\star \mid \mathrm{Z}_i^\mathrm{T};\mathbf{\Theta}_i}\left(\mathrm{z}_{i,t}^\star \mid \mathrm{z}_i^\mathrm{T};\boldsymbol{\theta}\right) \tag{1.9}$$

The remainder of this section presents expressions for the filtered and smoothed probability densities, that is, $f_{\mathrm{Z}_{i,t}^\star \mid \mathrm{Z}_i^t;\mathbf{\Theta}_i}\left(\mathrm{z}_{i,t}^\star \mid \mathrm{z}_i^t;\boldsymbol{\theta}_i\right)$ and $f_{\mathrm{Z}_{i,t}^\star \mid \mathrm{Z}_i^\mathrm{T};\mathbf{\Theta}_i}\left(\mathrm{z}_{i,t}^\star \mid \mathrm{z}_i^\mathrm{T};\boldsymbol{\theta}_i\right)$. Even though the filtered probability density is not directly used to compute smoothed estimates of the latent variables of interest, its expression is helpful in grasping some of the intuition behind the smoothed probability function and, moreover, its expression can be used to evaluate the likelihood of the model or to perform forecasting which makes it an object of independent interest for any researcher working with the state-space representation in assumption 1.1. For expositional purposes separate expressions are presented depending on whether the time period in consideration is an adjustment or an inaction period.

Terminology and notation. In all that follows, an arbitrary realisation of sequence of cumulated state variable changes for a given unit is considered. Without loss of generality, let $K \in \mathbb{Z}_{[0,\mathrm{T}]}$ denote the number of state variable adjustments observed in that sequence, τ^k denotes the time period at which the k-th adjustment is observed and $\tau^0 = 0$ denotes the initial time period. Moreover, the vector of unit specific parameters $\mathbf{\Theta}_i$ is kept fixed at some arbitrary value $\boldsymbol{\theta}$.[27]

1.4.1. Filtered and Smoothed probability density for adjustment periods

The following proposition characterises the filtered and smoothed probability density functions for adjustment periods,

[26] From (1.5) with a *diagonal* matrix $\mathbf{\Sigma}$ it follows that $\mathrm{Z}_{i,t}^\star$'s are independent across units. From (1.6) with $\nu_{i,t}$'s *iid* across i (and t) it follows that $\ell_{i,t}$'s are also independent across units. Finally, from (1.3) and (1.4) it follows that $\mathrm{Z}_{i,t}$'s depend only on $\mathrm{Z}_{i,t-1}$, $\mathrm{Z}_{i,t}^\star$ and $\ell_{i,t}$ and, therefore, $\mathrm{Z}_{i,t}$'s are also independent across units. Under this same independence argument it also holds that $f_{\mathbf{Z}_t^\star \mid \mathbf{Z}^t;\mathbf{\Theta}}\left(\mathbf{z}_t^\star \mid \mathbf{z}^t;\boldsymbol{\theta}\right) = \prod_{i=1}^{n} f_{\mathrm{Z}_{i,t}^\star \mid \mathrm{Z}_i^t;\mathbf{\Theta}_i}\left(\mathrm{z}_{i,t}^\star \mid \mathrm{z}_i^t;\boldsymbol{\theta}_i\right)$.

[27] This vector could be either the parameter estimates obtained from the two-step procedure proposed in section 1.3 or any other set of parameter values.

***Proposition* 1.2** *Consider a sequence* $\{\mathrm{z}_{i,t}\}_{t=0}^{\mathrm{T}}$ *generated by the data generating process in Assumption 1. For* $k \in \mathbb{Z}_{[0,K]}$ *define,*

$$c_i^k \equiv \left(\mathrm{z}_{i,\tau^k} + x_{i,0}\right) \mathbb{1}\left\{k \in \mathbb{Z}_{[1,K]}\right\} \tag{1.10}$$

Suppose $t = \tau^k$ *for some* $k \in \mathbb{Z}_{[0,K]}$, *then,*

$$f_{\mathrm{Z}_{i,t}^\star | \mathrm{Z}_i^{\mathrm{T}}; \mathbf{\Theta}_i}\left(\mathrm{z}_{i,t}^\star \mid \mathrm{z}_i^{\mathrm{T}}; \boldsymbol{\theta}\right) = f_{\mathrm{Z}_{i,t}^\star | \mathrm{Z}_i^t; \mathbf{\Theta}_i}\left(\mathrm{z}_{i,t}^\star \mid \mathrm{z}_i^t\,; \boldsymbol{\theta}\right) = \delta\left(\mathrm{z}_{i,t}^\star - c_i^k\right) \tag{1.11}$$

Proof. See appendix 1.D. □

Intuition. This result states that, at the initial period or at any period where a state variable adjustment is observed, both filtered and smoothed probability densities are degenerate or, in other words, the value of $\mathrm{Z}_{i,t}^\star$ is *known* to the researcher. First, for the initial period this follows by construction since $\mathrm{Z}_{i,t}^\star \equiv z_{i,t}^\star - z_{i,0}^\star$. Otherwise, in any period where a state variable adjustment is observed, it must be the case that the re-centered state gap is set to zero which occurs if and only if $\mathrm{Z}_{i,t}^\star = \mathrm{z}_{i,\tau^k} + x_{i,0}$.

1.4.2. Filtered probability density for inaction periods

The following proposition completes the characterisation of the filtered probability density function by providing an expression that holds for any inaction period,

***Proposition* 1.3** *Consider a sequence* $\{\mathrm{z}_{i,t}\}_{t=0}^{\mathrm{T}}$ *generated by the data generating process in Assumption 1. Suppose* t *is an inaction period. Let* τ^k *denote the largest time period before* t *such that* (1.11) *holds and let* $b \equiv t - \tau^k$ *denote the number of periods elapsed since* τ^k. *Define the boundaries of the inaction region as,*

$$\underline{Z}_i^k \equiv (x_{i,0} - \bar{x}_i)\,\mathbb{1}\left\{k = 0\right\} + \left(c_i^k - \bar{x}_i\right)\mathbb{1}\left\{k > 0\right\} \tag{1.12}$$

$$\bar{Z}_i^k \equiv (x_{i,0} - \underline{x}_i)\,\mathbb{1}\left\{k = 0\right\} + \left(c_i^k - \underline{x}_i\right)\mathbb{1}\left\{k > 0\right\} \tag{1.13}$$

The function $\beta_{i,b}^k(\cdot)$ is defined recursively as,

$$\beta_{i,b}^k\,(x) \equiv \mathbb{1}\left\{b = 1\right\} + \mathbb{1}\left\{b > 1\right\}\left[\int_{\underline{Z}_i^k}^{\bar{Z}_i^k} \frac{1}{\tilde{\sigma}_{i,b-1}}\,\phi\left(\frac{y - \tilde{\mu}_{i,b-1}^k(x)}{\tilde{\sigma}_{i,b-1}}\right)\beta_{i,b-1}^k\,(y)\,dy\right] \tag{1.14}$$

Moreover, define the recursions,

$$\mu_{i,b}^k \equiv b\mu_i + c_i^k \tag{1.15}$$

$$\sigma_{i,b} \equiv b^{\frac{1}{2}}\sigma_{\varepsilon,i} \tag{1.16}$$

$$\tilde{\mu}_{i,b}^k(x) \equiv \left(c_i^k + b\,x\right)/\,(b+1) \tag{1.17}$$

$$\tilde{\sigma}_{i,b} \equiv \left[b/\,(b+1)\right]^{\frac{1}{2}}\sigma_{\varepsilon,i} \tag{1.18}$$

Then, ignoring terms that are zero almost everywhere,

$$f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}^t_i;\boldsymbol{\Theta}_i}\left(\mathrm{z}^\star_{i,t} \mid \mathrm{z}^t_i\,;\boldsymbol{\theta}\right) \propto \frac{1}{\sigma_{i,b}}\phi\left(\frac{\mathrm{z}^\star_{i,t} - \mu_{i,b}^k}{\sigma_{i,b}}\right)\beta_{i,b}^k(\mathrm{z}^\star_{i,t})\,\mathbb{1}\left\{\mathrm{z}^\star_{i,t} \in \left(\underline{Z}_i^k, \bar{Z}_i^k\right)\right\} \tag{1.19}$$

Proof. See appendix 1.D. □

Intuition. The first term in (1.19) is a Gaussian probability density function with mean equal the last *known* value of $\mathrm{Z}^\star_{i,t}$ (denoted by c_i^k) plus a linear increment of μ_i per period elapsed since τ^k and variance equal to the number of periods elapsed since τ^k times $\sigma^2_{\varepsilon,i}$. Alone, this first term resembles the solution to the well-known linear Gaussian filtering problem where the filtered distribution remains Gaussian with parameters obtained from the Kalman filtering recursions (Kalman, 1960). However, the nonlinearities in the measurement equation induced by two-sided $(S{,}s)$ rules give rise to two extra terms in (1.19) which *deform* the Gaussian density from the first term.

To grasp the intuition behind these two extra terms, notice that all periods in between τ^k and t were *inaction* periods. In the presence of two-sided $(S{,}s)$ rules it must be the case that, for those periods, the re-centered state gap was always *within* the re-centered inaction region which occurs, if and only if, $\mathrm{z}^\star_{i,j}$ belongs to the interval $\left(\underline{Z}_i^k, \bar{Z}_i^k\right)$ $\forall j \in \mathbb{Z}_{(\tau^k,t]}$. The last term in (1.19) simply ensures that zero probability is assigned to the event $\mathrm{Z}^\star_{i,t} \notin \left(\underline{Z}_i^k, \bar{Z}_i^k\right)$. The term $\beta^k_{i,b}(\mathrm{z}^\star_{i,t})$ reflects the fact that $\mathrm{z}^\star_{i,t}$ is the last value of a sequence $\left\{\mathrm{z}^\star_{i,j}\right\}_{j=\tau^k}^{t}$ that must satisfy three conditions: (i) starts at value c_i^k, (ii) it is generated by the frictionless state transition equation (1.5) and (iii) it does not contain any value that does not belong to the interval $\left(\underline{Z}_i^k, \bar{Z}_i^k\right)$. Intuitively, the function $\beta^k_{i,b}(\mathrm{z}^\star_{i,t})$ yields *smaller* values for argument values that are *less likely* to be the last value of a sequence satisfying these three conditions, which ultimately results in smaller values of the filtered probability density function being assigned to those $\mathrm{z}^\star_{i,t}$ values.

1.4.3. Smoothed probability density for inaction periods

The following expression characterises the smoothed probability density for any inaction period for which *some* adjustment is observed afterwards,[28]

***Proposition* 1.4** *Consider a sequence* $\{\mathrm{z}_{i,t}\}_{t=0}^{\mathrm{T}}$ *generated by the data generating process in Assumption 1. Suppose* t *is an inaction period. Let* τ^k *denote the largest time period before* t *such that* (1.11) *holds and let* $b \equiv t - \tau^k$ *denote the number of periods elapsed since* τ^k. *Suppose further that there exists a period after* t *such that* (1.11) *holds and let* τ^{k+1} *denote the smallest of those periods. Define* $\Delta^k \equiv \tau^{k+1} - \tau^k$ *and the function* $\chi^k_{i,b}(\cdot)$ *is defined recursively as,*

[28] For the hypothetical unit depicted in figure 1.A.1 the expression in proposition 1.4 applies for $t \in (0, \tau^1) \cup (\tau^1, \tau^2)$.

$$\chi^k_{i,b}\left(x\right) \equiv \mathbb{1}\left\{b = \Delta^k - 1\right\} + \mathbb{1}\left\{b < \Delta^k - 1\right\}\left[\int_{\underline{Z}^k_i}^{\bar{Z}^k_i} \frac{1}{\ddot{\sigma}^k_{i,b+1}}\,\phi\left(\frac{y - \ddot{\mu}^k_{i,b+1}(x)}{\ddot{\sigma}^k_{i,b+1}}\right)\chi^k_{i,b+1}\left(y\right)dy\right] \tag{1.20}$$

Moreover, define the recursions,

$$\breve{\mu}^k_{i,b} \equiv \left[bc^{k+1}_i + (\Delta^k - b)c^k_i\right]/\Delta^k \tag{1.21}$$

$$\breve{\sigma}^k_{i,b} \equiv \left[b(\Delta^k - b)/\Delta^k\right]^{\frac{1}{2}}\sigma_{\varepsilon,i} \tag{1.22}$$

$$\ddot{\mu}^k_{i,b}(x) \equiv \left[c^{k+1}_i + (\Delta^k - b)x\right]/\left(\Delta^k - b + 1\right) \tag{1.23}$$

$$\ddot{\sigma}^k_{i,b} \equiv \left[(\Delta^k - b)/(\Delta^k - b + 1)\right]^{\frac{1}{2}}\sigma_{\varepsilon,i} \tag{1.24}$$

Then, ignoring terms that are zero almost everywhere,

$$f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}^{\mathrm{T}}_i;\boldsymbol{\Theta}}\left(\mathrm{z}^\star_{i,t} \mid \mathrm{z}^{\mathrm{T}}_i\,;\boldsymbol{\theta}\right) \propto \frac{1}{\breve{\sigma}^k_{i,b}}\phi\left(\frac{\mathrm{z}^\star_{i,t} - \breve{\mu}^k_{i,b}}{\breve{\sigma}^k_{i,b}}\right)\beta^k_{i,b}(\mathrm{z}^\star_{i,t})\,\chi^k_{i,b}\left(\mathrm{z}^\star_{i,t}\right)\,\mathbb{1}\left\{\mathrm{z}^\star_{i,t} \in \left(\underline{Z}^k_i, \bar{Z}^k_i\right)\right\} \tag{1.25}$$

where $\underline{Z}^k_i$*,* $\bar{Z}^k_i$ *and* $\beta^k_{i,b}(\cdot)$ *are defined in* (1.12), (1.13) *and* (1.14)*, respectively.*

Proof. See appendix 1.D. □

Intuition. By definition, the smoothed probability density is conditional on measurements observed over the *whole* sample. Conditioning on more information makes the smoothed probability density in (1.25) differ from the filtered probability density in (1.19) in two ways. First, the first term is still a Gaussian probability density function but with different parameters. The mean is now given by the linear interpolation between the last and the next *known* values of $\mathrm{Z}^\star_{i,t}$ (denoted by c^k_i and c^{k+1}_i, respectively), whilst the variance is still proportional to $\sigma^2_{\varepsilon,i}$ but it is maximised at the mid point between the τ^k and τ^{k+1}. Second, in (1.25) there is an additional term given by $\chi^k_{i,b}\left(\mathrm{z}^\star_{i,t}\right)$ for which the intuition is similar to that of $\beta^k_{i,b}\left(\mathrm{z}^\star_{i,t}\right)$ but looking at periods that occur after t. More precisely, given that all periods in between t and τ^{k+1} are inaction periods, $\chi^k_{i,b}\left(\mathrm{z}^\star_{i,t}\right)$ reflects the fact that $\mathrm{z}^\star_{i,t}$ is the initial value of a sequence $\left\{\mathrm{z}^\star_{i,j}\right\}_{j=t}^{\tau^{k+1}}$ that must satisfy three conditions: (i) the last value is c_i^{k+1}; (ii) it is generated by the frictionless state transition equation (1.5) and (iii) it does not contain any value that does not belong to the interval $\left(\underline{Z}_i^k, \bar{Z}_i^k\right)$. The function $\chi_{i,b}^k\left(\mathrm{z}_{i,t}^{\star}\right)$ yields smaller values for $\mathrm{z}_{i,t}^{\star}$ values that are less likely to be the initial value of a sequence satisfying these three conditions.

The following proposition completes the characterisation of the smoothed probability density by considering an inaction period for which no adjustment is observed afterwards,[29]

***Proposition* 1.5** *Consider a sequence* $\{\mathrm{z}_{i,t}\}_{t=0}^{\mathrm{T}}$ *generated by the data generating process in Assumption 1. Suppose* t *is an inaction period. Let* τ^k *denote the largest time period before* t *such that* (1.11) *holds and let* $b \equiv t - \tau^k$ *denote the number of periods elapsed since* τ^k*. Suppose further that there does not exist any period after* t *such that* (1.11) *holds. Define* $\Delta^K \equiv \mathrm{T} - \tau^k$ *and the function* $\iota_{i,b}^k(\cdot)$ *is defined recursively as,*

$$\iota_{i,b}^k(x) \equiv \mathbb{1}\left\{b = \Delta^K\right\} + \mathbb{1}\left\{b < \Delta^K\right\}\left[\int_{\underline{Z}_i^k}^{\bar{Z}_i^k} \frac{1}{\sigma_{\varepsilon,i}} \phi\left(\frac{y - (\mu + x)}{\sigma_{\varepsilon,i}}\right) \iota_{i,b+1}^k(y)\, dy\right] \tag{1.26}$$

Then, ignoring terms that are zero almost everywhere,

$$f_{\mathrm{Z}_{i,t}^{\star}|\mathrm{Z}_i^{\mathrm{T}};\boldsymbol{\Theta}}\left(\mathrm{z}_{i,t}^{\star} \mid \mathrm{z}_i^{\mathrm{T}}; \boldsymbol{\theta}\right) \propto \frac{1}{\sigma_{i,b}} \phi\left(\frac{\mathrm{z}_{i,t}^{\star} - \mu_{i,b}^k}{\sigma_{i,b}}\right) \beta_{i,b}^k(\mathrm{z}_{i,t}^{\star})\, \iota_{i,b}^k\left(\mathrm{z}_{i,t}^{\star}\right)\, \mathbb{1}\left\{\mathrm{z}_{i,t}^{\star} \in \left(\underline{Z}_i^k, \bar{Z}_i^k\right)\right\} \tag{1.27}$$

where $\underline{Z}_i^k$*,* $\bar{Z}_i^k$*,* $\beta_{i,b}^k(\cdot)$*,* $\mu_{i,b}$ *and* $\sigma_{i,b}$ *are defined in* (1.12), (1.13), (1.14), (1.15) *and* (1.16)*, respectively.*

Proof. See appendix 1.D. □

Intuition. The fact that no adjustment is observed before the end of the sample implies that there is no *known* value for $\mathrm{Z}_{i,t}^{\star}$ for any period after t. This affects

[29] For the hypothetical unit depicted in figure 1.A.1 the expression in proposition 1.5 applies for $t \in (\tau^2, \mathrm{T})$.

the expression for the smoothed probability density changes with respect to the one in (1.25) in two ways. First, and similarly to the filtered probability density in (1.19), the mean of the Gaussian probability density term is given by a linear extrapolation starting from c_i^k whereas the variance increases linearly with the number of periods elapsed since τ^k. Second, the $\chi_{i,b}^k\left(\mathrm{z}_{i,t}^\star\right)$ term is replaced by $\iota_{i,b}^k\left(\mathrm{z}_{i,t}^\star\right)$ which, in turn reflects the fact that $\mathrm{z}_{i,t}^\star$ is the initial value of a sequence $\left\{\mathrm{z}_{i,j}^\star\right\}_{j=t}^T$ that instead must satisfy only *two* conditions: (i) it is generated by the frictionless state transition equation (1.5) and (ii) it does not contain any value that does not belong to the interval $\left(\underline{Z}_i^k, \bar{Z}_i^k\right)$. Similarly to $\chi_{i,b}^k\left(\mathrm{z}_{i,t}^\star\right)$, the function $\iota_{i,b}^k\left(\mathrm{z}_{i,t}^\star\right)$ yields smaller values for $\mathrm{z}_{i,t}^\star$ values that are less likely to be the initial value of a sequence satisfying these two conditions.

1.4.4. From probability densities to latent variable point estimates

The expressions in (1.11), (1.25) and (1.27) and a set of values for $\boldsymbol{\Theta}_i$ can be used to produce *smoothed estimates* of $\mathrm{Z}_{i,t}^\star$ by computing the expectations of the smoothed probability density distributions. However, in order to do that it is necessary to keep track of all the terms that enter those expressions. The means and variances of the Gaussian probability density terms are trivial to keep track of but the integral recursions that give rise to the functions $\beta_{i,b}^k\left(\mathrm{z}_{i,t}^\star\right)$, $\chi_{i,b}^k\left(\mathrm{z}_{i,t}^\star\right)$ and $\iota_{i,b}^k\left(\mathrm{z}_{i,t}^\star\right)$ can be computationally challenging to deal with. Appendix 1.E provides an algorithm that uses Gauss-Legendre integration to evaluate the smoothed probability densities in a fast and efficient way.

1.5. Monte Carlo Experiment

This section uses a Monte Carlo experiment to illustrate how the results presented this far can be used to go from a panel of cumulated state variable changes, $\{\mathbf{Z}_t\}_{t=0}^{\mathrm{T}}$, to a panel of estimates of cumulated changes in frictionless state variables, $\{\hat{\mathbf{Z}}_t\}_{t=0}^{\mathrm{T}}$.

1.5.1. Alternative data generating processes

Samples of artificial data are generated from the state-space representation in assumption 1.1. Three different sets of parameter values and panel dimensions are considered. The data generating processes are summarized in table 1.B.1. In all parameter combinations considered the vector of parameters Γ_i is assumed to be *common* across all units whereas $x_{i,0}$'s are assumed to be fully heterogeneous. The key parameter that differs across parameter combinations in table 1.B.1 is the probability of arrival of costless adjustment opportunities. With everything else constant, higher probabilities of costless adjustment opportunities are associated with a larger fraction of time-dependent adjustments which, in turn, affects the shape of the distribution of state variable adjustments by increasing the proportion of small state variable adjustments as illustrated in figure 1.A.3. For each combination of parameter values and panel dimensions, a total of 1,000 samples are generated and, in each of them, parameter and latent variable estimates are obtained and compared to their true values.

1.5.2. Parameter estimates

For each artificially generated dataset first stage parameters are obtained from (1.7) with moment conditions calculated from the vector of state variable changes *pooled* across all units. The moments used for first stage estimation are the frequency of state variable adjustments and the 1^{st}, 5^{th}, 10^{th}, …, 95^{th} and 99^{th} percentiles of the distribution of state variable adjustments. Once first stage parameters are obtained, the $x_{0,i}$'s are estimated from (1.8) for each unit.[30]

First stage estimates. Kernel densities of parameter distribution across Monte Carlo replications are depicted in figure 1.A.4. First and most importantly, as expected from a consistent estimator all densities becomes more concentrated around the true parameter values with the increase in total sample size. Second, and in line with the independence result in proposition 1.1, all the results in fig-

[30] Computational details for the estimation procedure are available in appendix 1.F.

ure 1.A.4 were obtained even though the simulated data in (1.7) was generated assuming $x_{i,0} = 0$ for all units whereas the Monte Carlo samples were generated using linearly spaced $x_{i,0}$'s. Third, for larger values of the probability of arrival of costless adjustment opportunities the boundaries of the inaction region are estimated less precisely. This can be rationalised from the fact that, the larger is that parameter the smaller is the mass of state variable adjustments concentrated around those boundaries as can be seen by comparing the left and right panels in figure 1.A.3. Most importantly, in spite of that decrease in precision, the boundary parameters were estimated without signs of lack of parameter identification. This is suggestive that the inclusion of extreme percentiles of the distribution of state variable changes in first stage moment conditions is helpful for parameter identification, specially in cases where relatively few price changes are triggered by crossing the boundaries of the inaction region.[31]

Second stage estimates. For each combination of data generating process and panel dimensions, figure 1.A.5 presents the kernel density estimates for the distribution of $\hat{x}_{i,0} - x_{i,0}$ pooled across units and Monte Carlo replications. First, it is important to notice that all estimated densities are centered at zero which is indicative that the two moments used for second stage estimation are informative about the initial re-centered state variable gap. Nonetheless, and in contrast with common parameter estimates in figure 1.A.4, the estimates of $x_{i,0}$ are less precise and do not become more concentrated around their true parameter values with the increase in total sample size. As anticipated in section 1.3, the underlying reason for that is that regardless of the total sample size there can only be *one* first state variable adjustment for each individual. In other words, the estimates of $\hat{x}_{i,0}$ in (1.8) are based on moments that involve a single realisation of a random variable and laws of large numbers will not apply. In practice, the imprecisions in estimated $x_{i,0}$'s will translate into more imprecise latent variable estimates even in large samples. This will be illustrated in the next section.

[31] Some versions of this experiment that did not include extreme percentiles of the distribution of non-zero price changes yielded a non-negligible fraction of extreme estimates for the boundary parameters, suggesting the lack of parameter identification.

1.5.3. Latent variable estimates

Lastly, three alternative estimators for $\mathrm{Z}^{\star}_{i,t}$'s are considered and their relative performance is assessed by comparing their respective MSEs.

Three alternative latent variable estimators. The first estimator considered is the estimator based on the smoothed probability density function derived in section 1.4. This estimator is given by,

$$\hat{\mathrm{Z}}_{i,t} = \mathbb{E}\left[\mathrm{Z}^{\star}_{i,t} \mid \mathrm{z}^{\mathrm{T}}_{i}, \boldsymbol{\theta}\right] = \int x\, f_{\mathrm{Z}^{\star}_{i,t} \mid \mathrm{Z}^{\mathrm{T}}_{i};\boldsymbol{\Theta}_i}\left(x \mid \mathrm{z}^{\mathrm{T}}_{i}\,; \boldsymbol{\theta}\right) dx \tag{1.28}$$

where $f_{\mathrm{Z}^{\star}_{i,t} \mid \mathrm{Z}^{\mathrm{T}}_{i};\boldsymbol{\Theta}}\left(x \mid \mathrm{z}^{\mathrm{T}}_{i}\,; \boldsymbol{\theta}\right)$ is given by the expressions in (1.11), (1.25) and (1.27). The second estimator considered is given by,

$$\check{\mathrm{Z}}_{i,t} = \int x\, \check{f}_{\mathrm{Z}^{\star}_{i,t} \mid \mathrm{Z}^{\mathrm{T}}_{i};\boldsymbol{\Theta}_i}\left(x \mid \mathrm{z}^{\mathrm{T}}_{i}\,; \boldsymbol{\theta}\right) dx \tag{1.29}$$

where $\check{f}_{\mathrm{Z}^{\star}_{i,t} \mid \mathrm{Z}^{\mathrm{T}}_{i};\boldsymbol{\Theta}_i}\left(x \mid \mathrm{z}^{\mathrm{T}}_{i}\,; \boldsymbol{\theta}\right)$ is given by (1.11) and the *first term* on the right-hand side of (1.25) and (1.27). In other words, for a time period that is in between two state variable adjustments $\check{\mathrm{Z}}^{\star}_{i,t}$ is equal to the linear interpolation between the last and the next known values of $\mathrm{Z}^{\star}_{i,t}$. Otherwise, for a time period that is after the last observed adjustment, $\check{\mathrm{Z}}^{\star}_{i,t}$ is simply the linear extrapolation starting from the last known value of $\mathrm{Z}^{\star}_{i,t}$. The last estimator considered is given by,

$$\tilde{\mathrm{Z}}^{\star}_{i,t} = \mathbb{E}\left[\mathrm{Z}^{\star}_{i,t}\right] = t \times \mu_i \tag{1.30}$$

which does not condition on the information contained in cumulated state variable changes observed in the data.[32] Figure 1.A.6 plots estimates obtained from each of these estimators against the true values of $\mathrm{Z}^{\star}_{i,t}$ for a given unit in an artificially generated sample. To investigate the impact of parameter uncertainty on latent variable estimates, for any given sample estimates based on (1.28), (1.29) and

[32] Notice that iterating (1.5) backwards and using $\mathbf{Z}^{\star}_{0} = \mathbf{0}_{n\times 1}$ yields $\mathbf{Z}^{\star}_{t} = \boldsymbol{\mu}\, t + \sum_{k=1}^{t} \boldsymbol{\varepsilon}_k$ and, therefore, $\mathbb{E}\left[\mathbf{Z}^{\star}_{t}\right] = \boldsymbol{\mu}\, t$ given that $\mathbb{E}\left[\boldsymbol{\varepsilon}_k\right] = \mathbf{0}_{n\times 1}$ from assumption 1.1.

(1.30) are computed twice: one using the true parameter values and other using the estimated parameters from the two-stage procedure.

MSE comparisons. The MSEs for different estimators and across different DGPs are reported in table 1.B.2. There are four conclusions to be drawn from this table. First, across the different DGP specifications and sample size considered, $\hat{\mathrm{Z}}^{\star}_{i,t}$ always achieves a smaller MSE than $\check{\mathrm{Z}}^{\star}_{i,t}$ which, in turn, always achieves a smaller MSE than $\tilde{\mathrm{Z}}^{\star}_{i,t}$. More precisely, across the different specifications considered, the MSEs based on $\hat{\mathrm{Z}}^{\star}_{i,t}$ are between 5 and 55% smaller than their counterparts based $\check{\mathrm{Z}}^{\star}_{i,t}$ and more than a full order of magnitude smaller than their counterparts based on $\tilde{\mathrm{Z}}^{\star}_{i,t}$. This first conclusion is not surprising since $\hat{\mathrm{Z}}^{\star}_{i,t}$ not only uses all the information available to the researcher it is also based on the correct expression for the smoothed probability density function. The relative performance of alternative estimators could also be anticipated from the illustration in figure 1.A.6. Second, the MSE difference between $\hat{\mathrm{Z}}^{\star}_{i,t}$ and $\check{\mathrm{Z}}^{\star}_{i,t}$ is *smaller* as we move from the first to the third DGP. This is explained by the associated increase in the relative fraction of time-dependent adjustments. In the limiting case where *all* the adjustments are time-dependent ones (that is, if $\underline{x}_i \to -\infty$, $\bar{x}_i \to \infty$, $\forall i$) the estimators $\hat{\mathrm{Z}}^{\star}_{i,t}$ and $\check{\mathrm{Z}}^{\star}_{i,t}$ are *equivalent*.[33] Third, for all estimators considered there is an increase in MSE when using estimated parameters instead of true parameter values and this increase is bigger for $\hat{\mathrm{Z}}^{\star}_{i,t}$ and $\check{\mathrm{Z}}^{\star}_{i,t}$ than for $\tilde{\mathrm{Z}}^{\star}_{i,t}$. This bigger increase can be rationalised by the fact that $\tilde{\mathrm{Z}}^{\star}_{i,t}$ only depends on *one* estimated parameter whilst $\hat{\mathrm{Z}}^{\star}_{i,t}$ and $\check{\mathrm{Z}}^{\star}_{i,t}$ depend on *all* the estimated parameters. Fourth, the difference between the MSEs when using estimated versus true parameter values is roughly constant across the different panel dimensions considered. If all parameters were estimated consistently one would expect this difference to vanish as $\mathrm{T} \to \infty$. However, as illustrated in the previous section, estimates of $x_{i,0}$'s are not consistent which rationalises the constant MSE difference across panel dimensions. Overall, the figures reported in

[33] This result follows from the fact that if $\underline{x}_i \to -\infty$, $\bar{x}_i \to \infty$, $\forall i$ then the smoothed probability densities in (1.25) and (1.27) are equal to the first term on the right-hand-side *only*. This is formally stated and proved in corollary 1.1 in appendix 1.D.

table 1.B.2 indicate that $\hat{Z}^{\star}_{i,t}$ is superior – in MSE terms – to the two alternative latent variable estimators considered.

1.6. Concluding remarks and future research

This paper introduced a statistical framework designed to enable a researcher to transform a panel containing variables for which adjustments are intermittent and lumpy, as implied in presence of (S,s) policies, into a panel of estimated cumulated changes in their *frictionless* counterparts, that is, the cumulated changes that would have been observed in a hypothetical world where (S,s) policies were not in place. This framework is formally grounded on a nonlinear a non-Gaussian state-space representation of the data generating process of an economy composed of microeconomic units pursuing two-sided (S,s) policies subject to costless adjustment opportunities. Given this representation, this paper introduces a two-stage simulation-based procedure for parameter estimation and provides closed-form expressions for the smoothed probability density function. These two ingredients are then combined to obtain smoothed estimates of cumulated changes in frictionless state variables for each microeconomic unit at any point in time.

I conclude by highlighting two promissing avenues for future research motivated by the present paper. The first of these avenues consists in extending the results here presented to more general versions of assumption 1.1 and, more precisely, to more general forms for the transition equation for cumulated changes in frictionless state variables in (1.6). One of these generalisations would be to accommodate the possibility of common shocks in the evolution of frictionless states by allowing $\mathbf{\Sigma}$ to be a non-diagonal matrix. Other generalisations would be to include nonlinearities and/or non-Gaussian disturbances in (1.6). On the one hand, introducing such nonlinearities or non-Gaussian disturbances would most likely make impossible to obtain closed form expressions for the smoothed probability density function and, therefore, smoothed estimates would need to be obtained from algorithm-based approximate solutions for the smoothing problem. On the other hand, by relaxing those assumptions the transition equation for frictionless state would allow the proposed state-space representation to accommodate a wider range of microfoundations for the evolution of frictionless states.

Lastly, given the increasing availability of micro datasets in which variable changes can be characterised as intermittent and lumpy, the second promising avenue for future research consists in the applications of the current framework. By making possible to obtain unit level estimates of cumulated changes in frictionless state variables, the present framework enables researchers to construct empirical estimates of *frictionless aggregate variables* and of the *cross-sectional distribution of re-centered state gaps* for any time period. These estimates can be used to shed a new light on some of the most important questions in the (S,s) literature such as the importance of microeconomic lumpiness for the dynamics of aggregate variables, the responses of aggregate variables to shocks as well as the welfare costs of microeconomic adjustment frictions.

1.7. References

Adda, J. and Cooper, R. (2003). *Dynamic Economics: Quantitative Methods and Applications*. The MIT Press, 1st edn.

Alvarez, F., Le Bihan, H. and Lippi, F. (2016). The real effects of monetary shocks in sticky price models: A sufficient statistic approach. *American Economic Review*, **106** (10), 2817 – 2851.

— and Lippi, F. (2009). Financial innovation and the transactions demand for cash. *Econometrica*, **77** (2), 363–402.

Arrow, K. J., Harris, T. and Marschak, J. (1951). Optimal inventory policy. *Econometrica*, **19** (3), 250–272.

Baley, I. and Blanco, A. (2020). Aggregate dynamics in lumpy economies, working Paper.

Baumol, W. J. (1952). The transactions demand for cash: An inventory theoretic approach. *The Quarterly Journal of Economics*, **66** (4), 545–556.

Bayer, C. (2009). A comment on the economics of labor adjustment: Mind the gap: Evidence from a monte carlo experiment. *American Economic Review*, **99** (5), 2258–2266.

Berger, D., Caballero, R. J. and Engel, E. (2018). Missing aggregate dynamics and var approximations of lumpy adjustment models, working Paper.

Bertola, G. and Caballero, R. J. (1994). Irreversibility and aggregate investment. *Review of Economic Studies*, **61**, 223–246.

Blanco, A. (2017). Optimal inflation target in an economy with menu costs and a zero lower bound.

Bonomo, M., Correa, A. and Medeiros, M. C. (2013). Estimating strategic complementarity in a state-dependent pricing model.

Caballero, R., Engel, E. and Haltiwanger, J. (1995). Plant-level adjustment and aggregate investment dynamics. *Brookings Papers on Economic Activity*, **2**.

—, — and — (1997). Aggregate employment dynamics: Building from microeconomic evidence. *American Economic Review*, **87** (1), 115–137.

Caballero, R. J. and Engel, E. (2004). Three strikes and you're out: Reply to cooper and willis. *American Economic Review*, **94** (4), 1238–1244.

— and Engel, E. M. R. A. (1991). Dynamic (s,s) economies. *Econometrica*, **59** (6), 1659 – 1686.

— and — (1999). Explaining investment dynamics in u.s. manufacturing: A generalized (s,s) approach. *Econometrica*, **67** (4), 783–826.

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

Campbell, J. R. and Eden, B. (2014). Rigid prices: Evidence from U.S. scanner data. *International Economic Review*, **55** (2), 423–442.

Canova, F. (2007). *Methods for Applied Macroeconomic Research*. Princeton University Press, 1st edn.

Caplin, A. and Leahy, J. (2010). Economic theory and the world of practice: A celebration of the (S,s) model. *Journal of Economic Perspectives*, **24** (1), 183–202.

Carvalho, C. and Kryvstov, O. (2018). Price selection, working Paper.

Chen, Z. (2003). Bayesian filtering: From kalman filters to particle filters, and beyond. *Statistics*, **182** (1), 1–69.

Cooper, R. and Willis, J. L. (2004). A comment on the economies of labor adjustment: Mind the gap. *American Economic Review*, **94** (4), 1223–1237.

— and — (2009). A comment on the economics of labor adjustment: Mind the gap: Evidence from a monte carlo experiment: Reply. *American Economic Review*, **99** (5), 2267–76.

Costain, J. and Nakov, A. (2011a). Distributional dynamics under smoothly state-dependent pricing. *Journal of Monetary Economics*, **58** (6-8), 646 – 665.

— and — (2011b). Price adjustments in a general model of state-dependent pricing. *Journal of Money, Credit and Banking*, **43** (2-3), 385–406.

Davidson, R. and MacKinnon, J. G. (2004). *Econometric Theory and Methods*. Cambridge University Press.

Dotsey, M., King, R. G. and Wolman, A. L. (1999). State-dependent pricing and the general equilibrium dynamics of money and output. *The Quarterly Journal of Economics*, **114** (2), 655–690.

Doucet, A. and Johansen, A. M. (2011). *A Tutorial on particle filtering and smoothing: fifteen years later*, Oxford University Press, pp. 656 – 704.

Elsby, M. W. L. and Michaels, R. (2019). Fixed adjustment costs and aggregate fluctuations. *Journal of Monetary Economics*, **101**, 128 – 147.

—, — and Ratner, D. (2019). The aggregate effects of labor market frictions. *Quantitative Economics*, **10**, 803–852.

Frenkel, J. A. and Jovanovic, B. (1980). On transaction and precautionary demand for money. *The Quarterly Journal of Economics*, **95** (1), 25–43.

Gautier, E. and Le Bihan, H. (2018). Shocks vs menu costs: Patterns of price rigidity in an estimated multi-sector menu-cost model, banque de France Working Paper No. 682.

Gertler, M. and Leahy, J. (2008). A phillips curve with an ss foundation. *Journal of Political Economy*, **116** (3), 533 – 572.

Golosov, M. and Lucas, R. E. (2007). Menu costs and phillips curves. *Journal of Political Economy*, **115** (2), 171–199.

Gouriéroux, C. and Monfort, A. (1996). *Simulation-Based Econometric Methods*. CORE Lectures, Oxford University Press.

Harrison, J. M., Sellke, T. M. and Taylor, A. J. (1983). Impulse control of brownian motion. *Mathematics of Operations Research*, **8** (3), 454 – 466.

Herbst, E. P. and Schorfheide, F. (2015). *Bayesian Estimation of DSGE Models*. Princeton University Press, 1st edn.

Judd, K. L. (1998). *Numerical Methods in Economics*. The MIT Press, 1st edn.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME – Journal of Basic Engineering*, **82** (Seies D), 35–45.

Kitagawa, G. (1987). Non-gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association*, **82** (400), 1032–1041.

— (1994). The two-filter formula for smoothing and an implementation of the gaussian-sum smoother. *Annals of the Institute of Statistical Mathematics*, **46** (4), 605–623.

— (1996). Monte carlo filter and smoother for non-gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, **5** (1), 1–25.

Luo, S. and Villar, D. (2017). The skewness of the price change distribution: A new touchstone for sticky price models. *Finance and Economics Discussion Series 2017-028. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2017.028.*

Miller, M. H. and Orr, D. (1966). A model of the demand for money by firms. *The Quarterly Journal of Economics*, **80** (3), 413–435.

Nakamura, E. and Steinsson, J. (2010). Monetary non-neutrality in a mutisector menu cost model. *The Quarterly Journal of Economics*, **125** (3), 961–1013.

Plehn-Dujowich, J. M. (2005). The optimality of a control band policy. *Review of Economic Dynamics*, **8**, 877–901.

Rauch, H. E., Tung, F. and Striebel, C. T. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, **3** (8), 1445–1450.

Särkkä, S. (2013). *Bayesian Filtering and Smoothing*. Institute of Mathematical Statistics Textbooks, Cambridge University Press, 1st edn.

Scarf, H. (1959). The optimality of (S,s) policies in the dynamic inventory problem. In K. J. Arrow, Karlin and P. Suppes (eds.), *Mathematical Methods in Social Sciences*, *13*, Stanford University Press.

Sheshinski, E. and Weiss, Y. (1977). Inflation and the costs of price adjustment. *Review of Economic Studies*, **44** (2), 287–303.

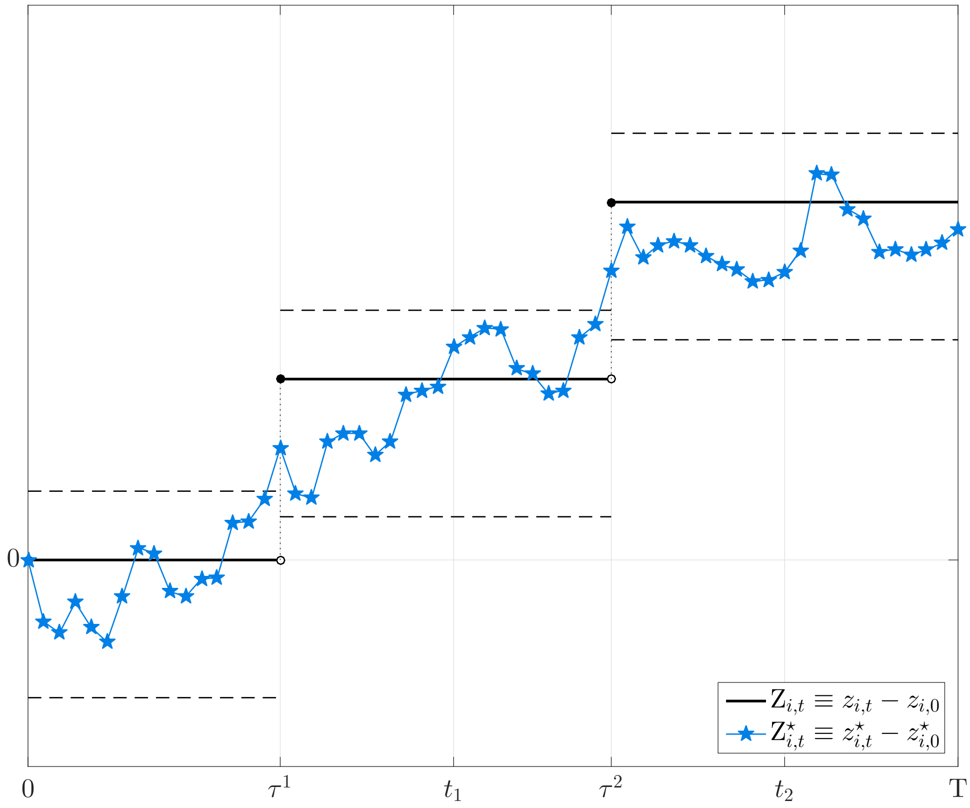
— and — (1983). Optimum pricing policy under stochastic inflation. *Review of Economic Studies*, **50** (3), 513–529.

Stokey, N. L. (2009). *The Economics of Inaction: Stochastic Control Models with Fixed Costs*. Princeton University Press, 1st edn.

Tobin, J. (1956). The interest-elasticity of transactions demand for cash. *The Review of Economics and Statistics*, **38** (3), 241–247.

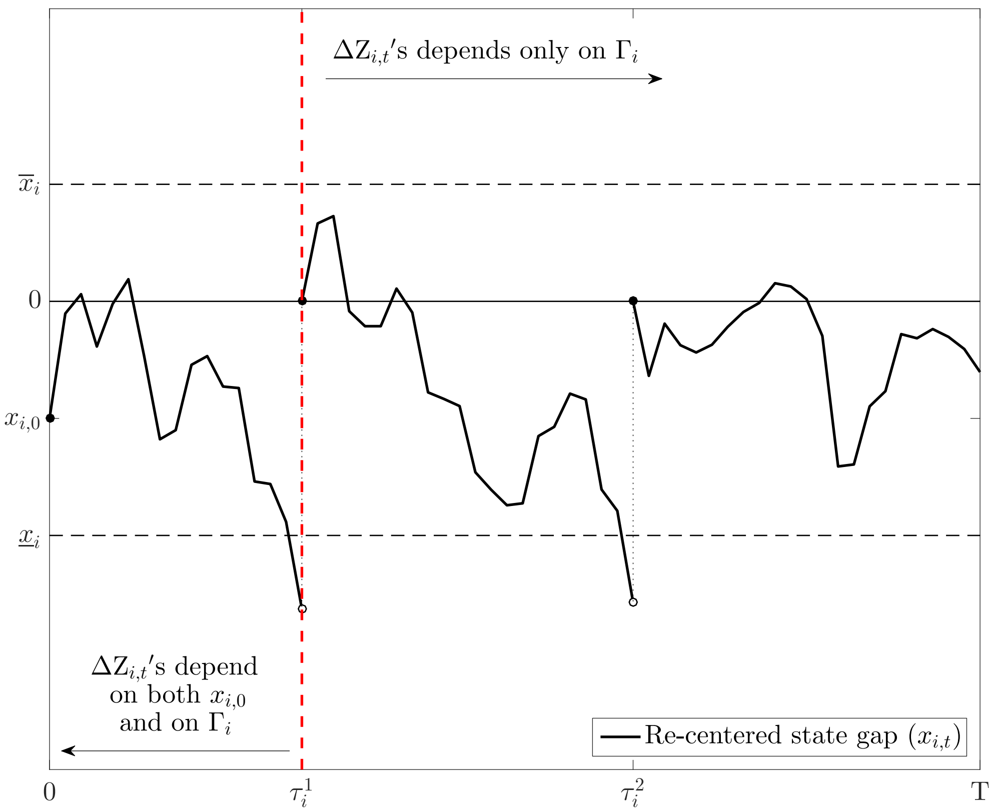
1.A. Figures

Figure 1.A.1: Cumulated changes for observed and frictionless state variables



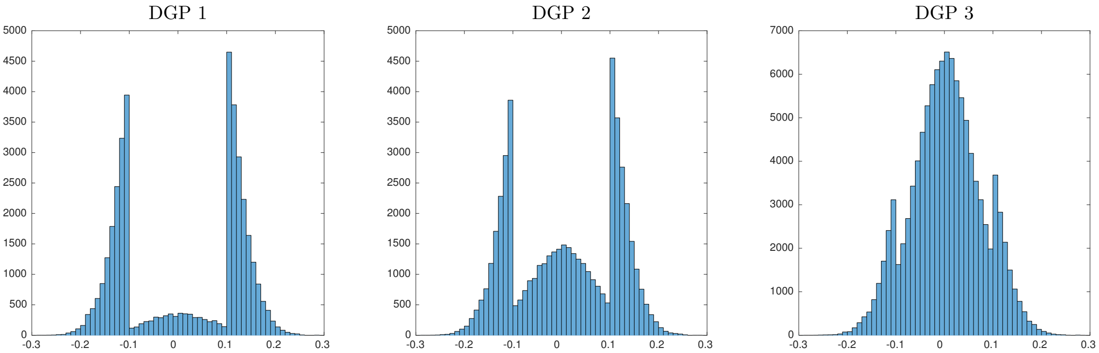
The data underlying this graph is generated from the data generating process specified in assumption 1.1. As defined in section 1.2, $Z_{i,t}$ denotes the cumulated change in unit i's actual state variable and $Z^\star_{i,t}$ denotes the cumulated change in unit i's frictionless state variable. The dashed black lines denote the boundaries of the inaction region for any point in time.

Figure 1.A.2: Re-centered state variable gaps



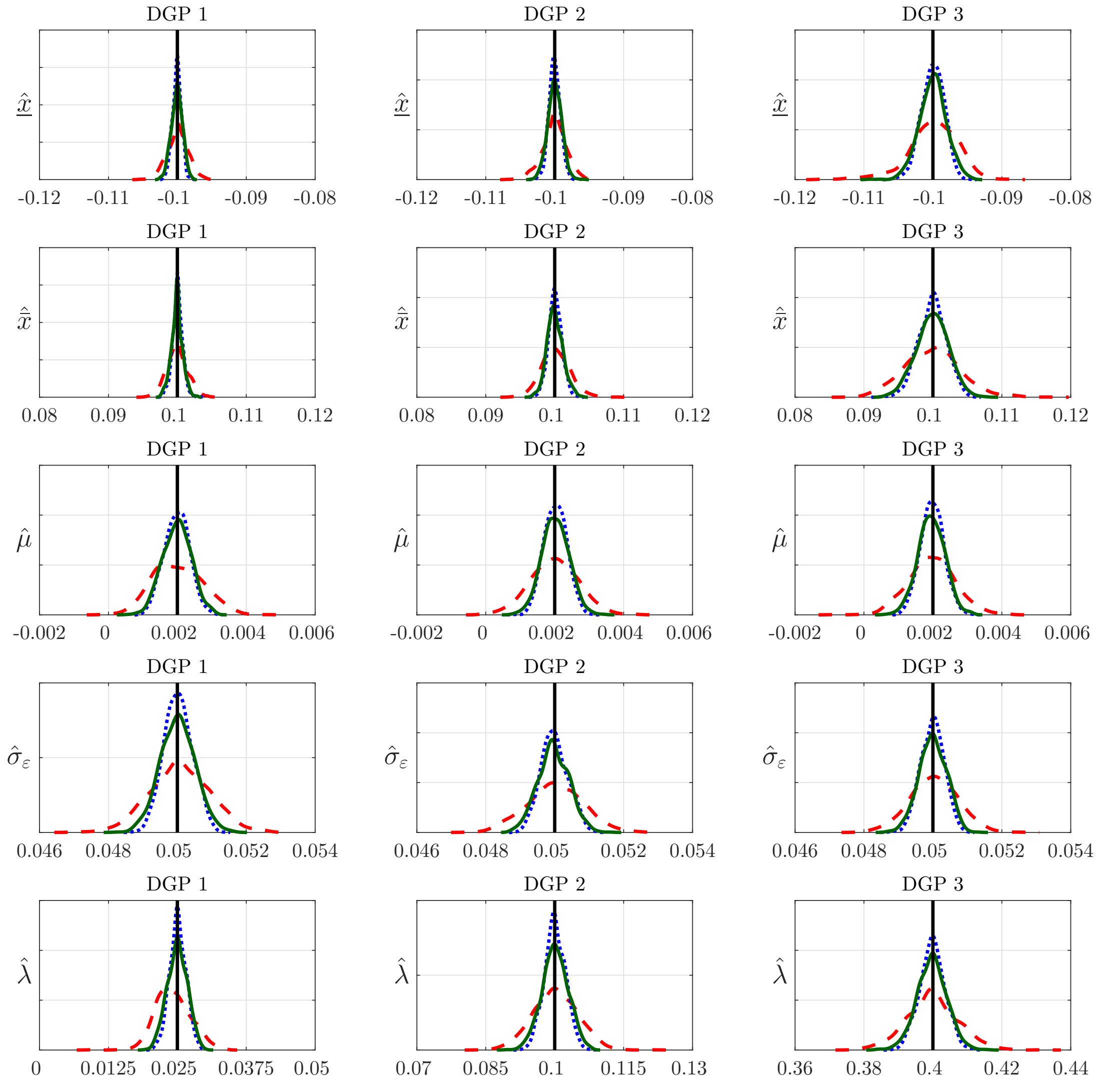
This figure plots the paths for the re-centered state gap implied by figure 1.A.1. The annotations follow from the result in proposition 1.1.

Figure 1.A.3: Distribution of state variable adjustments for alternative DGPs



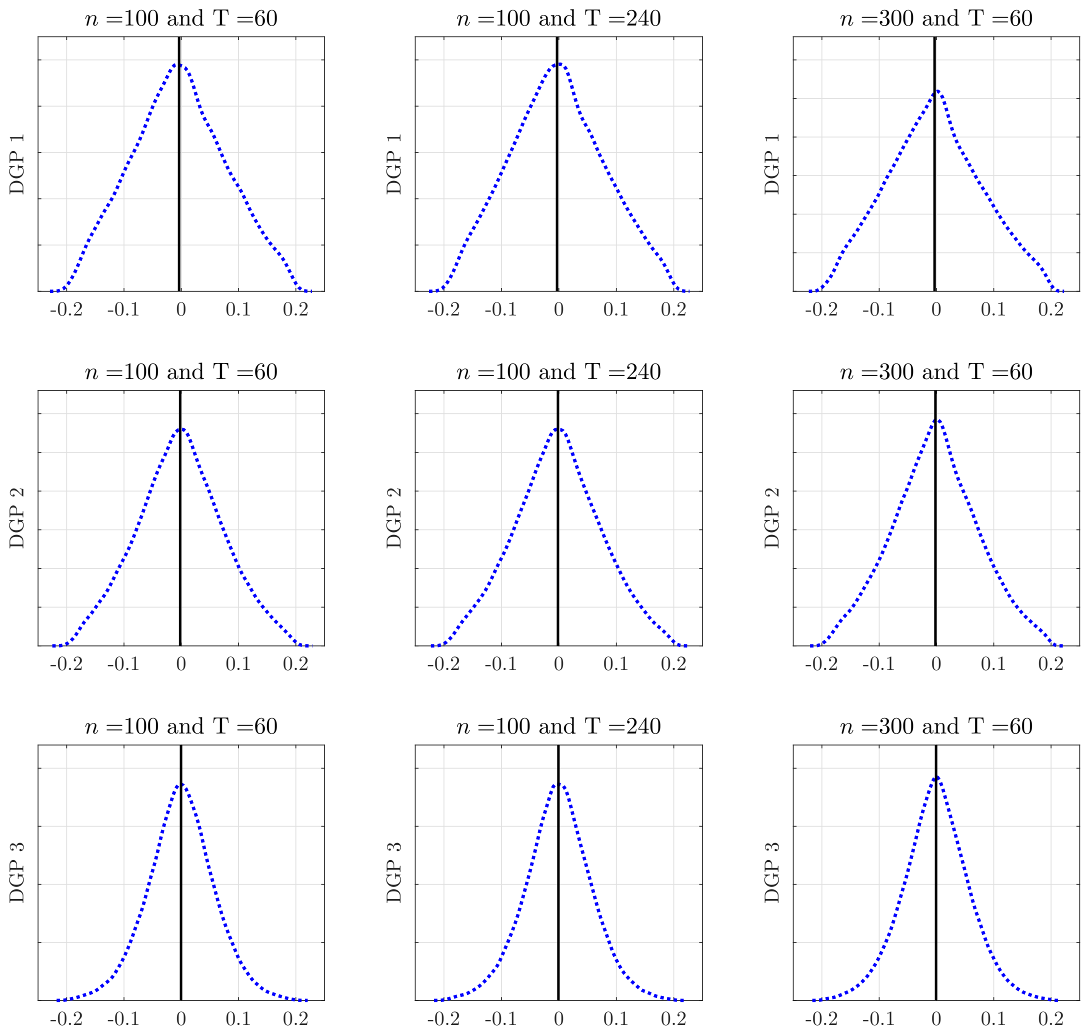
Distributions of state variable adjustments are obtained from a sample with a balanced panel with $n = 1,000$ and T $= 240$. The initial re-centred state-variable gaps are equally spaced points within the inaction region following the same rule as in table 1.B.1.

Figure 1.A.4: First-stage parameters across Monte Carlo replications



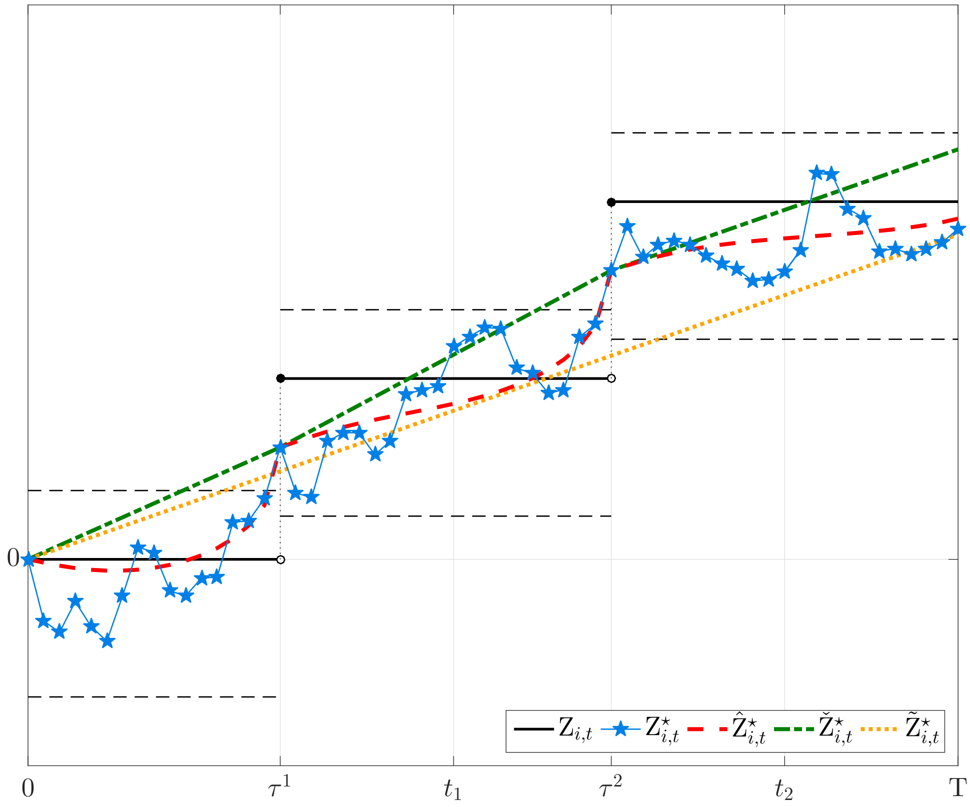
Each line is a normal kernel density estimate of parameter estimates across 1,000 Monte Carlo replications. The red dashed line correspond to panels with dimensions $n = 100$ and $\mathrm{T} = 60$, the blue dotted line to panels with dimensions $n = 100$ and $\mathrm{T} = 240$ and the green solid line to panels with dimensions $n = 300$ and $\mathrm{T} = 60$. The vertical black solid lines are at the true parameter values. The alternative data generating processes are described in table 1.B.1.

Figure 1.A.5: Second-stage parameters across Monte Carlo replications



For each combination of dgp and panel dimensions, the dashed blue lines are the kernel density estimates of the distribution of $(\hat{x}_{i,0} - x_{i,0})$ *pooled* across units and Monte Carlo replications. The vertical solid black lines are the means of the data underlying each kernel density estimate. The rules used to generate the initial conditions, $x_{i,0}$, are described in table 1.B.1.

Figure 1.A.6: Three alternative latent variable estimators



The data underlying this graph is generated from the data generating process specified in assumption 1.1. The dashed black lines denote the boundaries of the inaction region for any point in time. All the latent variable estimators are computed fixing the parameters equal to their true values. For this particular sample the MSEs of $\hat{Z}^{\star}_{i,t}$, $\check{Z}^{\star}_{i,t}$, $\tilde{Z}^{\star}_{i,t}$ are 0.0024, 0.0072, 0.0070, respectively.

1.B. Tables

Table 1.B.1: Data generating processes in the Monte Carlo experiment

DGP #	$\underline{\boldsymbol{x}}$	$\overline{\boldsymbol{x}}$	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$	$\boldsymbol{\lambda}$
1	$-0.1\times\mathbf{1}_{\boldsymbol{n}\times\mathbf{1}}$	$0.1\times\mathbf{1}_{\boldsymbol{n}\times\mathbf{1}}$	$0.002\times\mathbf{1}_{\boldsymbol{n}\times\mathbf{1}}$	$(0.05)^2\times\mathbf{I}_{n\times n}$	$0.025\times\mathbf{1}_{\boldsymbol{n}\times\mathbf{1}}$
2	$-0.1\times\mathbf{1}_{\boldsymbol{n}\times\mathbf{1}}$	$0.1\times\mathbf{1}_{\boldsymbol{n}\times\mathbf{1}}$	$0.002\times\mathbf{1}_{\boldsymbol{n}\times\mathbf{1}}$	$(0.05)^2\times\mathbf{I}_{n\times n}$	$0.1\times\mathbf{1}_{\boldsymbol{n}\times\mathbf{1}}$
3	$-0.1\times\mathbf{1}_{\boldsymbol{n}\times\mathbf{1}}$	$0.1\times\mathbf{1}_{\boldsymbol{n}\times\mathbf{1}}$	$0.002\times\mathbf{1}_{\boldsymbol{n}\times\mathbf{1}}$	$(0.05)^2\times\mathbf{I}_{n\times n}$	$0.4\times\mathbf{1}_{\boldsymbol{n}\times\mathbf{1}}$

The parameter vector notation is the same as in the state-space representation in Assumption 1.1. For each of the three parameter combinations above, three sample sizes are considered: (a) $n = 100$ and $\mathrm{T} = 60$, (b) $n = 100$ and $\mathrm{T} = 240$ and (c) $n = 300$ and $\mathrm{T} = 60$. For each combination of parameter values and sample size, the initial re-centered state variable gaps equally spaced points within the inaction region, more precisely, each element of the vector $\boldsymbol{x}_0$ os generated according to $x_{0,i} = 0.1 + [0.2/(n+2)] \times i$ where $i = 1, \ldots, n$.

Table 1.B.2: Mean Squared Errors for alternative latent variable estimators

DGP		True Parameters			Estimated Parameters		
		$\hat{\mathrm{Z}}^{\star}_{i,t}$	$\check{\mathrm{Z}}^{\star}_{i,t}$	$\tilde{\mathrm{Z}}^{\star}_{i,t}$	$\hat{\mathrm{Z}}^{\star}_{i,t}$	$\check{\mathrm{Z}}^{\star}_{i,t}$	$\tilde{\mathrm{Z}}^{\star}_{i,t}$
1	(a)	0.0014	0.0031	0.0753	0.0079	0.0094	0.0758
	(b)	0.0014	0.0033	0.2979	0.0081	0.0099	0.3004
	(c)	0.0014	0.0031	0.0753	0.0078	0.0093	0.0755
2	(a)	0.0012	0.0022	0.0748	0.0066	0.0076	0.0754
	(b)	0.0012	0.0023	0.3002	0.0068	0.0079	0.3023
	(c)	0.0012	0.0022	0.075	0.0066	0.0075	0.0751
3	(a)	0.0007	0.0009	0.0752	0.004	0.0042	0.0757
	(b)	0.0007	0.0009	0.2979	0.0041	0.0043	0.2999
	(c)	0.0007	0.0009	0.0753	0.004	0.0042	0.0755

The DGP numbers refer to one of the three parameter combinations summarized in table 1.B.1. The letters refer to the panel dimensions where (a) $n = 100$ and $\mathrm{T} = 60$, (b) $n = 100$ and $\mathrm{T} = 240$ and (c) $n = 300$ and $\mathrm{T} = 60$. The latent variable estimators $\hat{\mathrm{Z}}^{\star}_{i,t}$, $\check{\mathrm{Z}}^{\star}_{i,t}$ and $\tilde{\mathrm{Z}}^{\star}_{i,t}$ are described in section 1.5.3. The columns under “True Parameters” / “Estimated Parameters” are computed fixing the parameter values equal to their true parameter values/parameter values estimated from the two-stage procedure. For a given latent variable estimator, the Mean Squared Errors reported are computed as $\mathrm{MSE}(\hat{x}) = (1,000 \times n \times (\mathrm{T}+1))^{-1} \sum_{k=1}^{1,000} \sum_{i=1}^{n} \sum_{t=0}^{\mathrm{T}} (\hat{x}_{i,t,k} - \mathrm{Z}^{\star}_{i,t,k})^2$ where k indexes Monte Carlo replications, $\mathrm{Z}^{\star}_{i,t,k}$ is the true value of cumulated changes in frictionless state for unit i at time t in Monte Carlo replication k.

1.C. Proofs of auxiliary results

Notation. The function $o(\cdot):\mathbb{R}\to\mathbb{R}$ is used to denote any function that is equal to zero almost everywhere in the real line. For a matrix $\mathbf{A}\in\mathbb{R}^{m\times n}$ let $\mathbf{A'}$ denote the matrix transpose, $\boldsymbol{a}_{*,j}$ denote the j-th column of the matrix $\mathbf{A}$; $\boldsymbol{a}_{i,*}$ to denote the i-th row of $\mathbf{A}$. For matrix operations let $\otimes$ denote the Kronecker product, $\oslash$ denote the Hadamard division and the exponent ${}^{\circ-1}$ the Hadamard inverse. The remaining notation is as defined in the main text.

***Lemma* 1.1** *Let $x, y, \mu_x, a, c \in \mathbb{R}$, $\sigma_x, \sigma_y \in \mathbb{R}_{>0}$ and $\mu_y = ax + c$. Then,*

$$\frac{1}{\sigma_y}\phi\left(\frac{y-\mu_y}{\sigma_y}\right)\times\frac{1}{\sigma_x}\phi\left(\frac{x-\mu_x}{\sigma_x}\right)=\frac{1}{\tilde{\sigma}_y}\phi\left(\frac{y-\tilde{\mu}_y}{\tilde{\sigma}_y}\right)\times\frac{1}{\tilde{\sigma}_x}\phi\left(\frac{x-\tilde{\mu}_x}{\tilde{\sigma}_x}\right)\tag{L1.1}$$

where,

$$\tilde{\mu}_x\equiv\frac{\sigma_x^2a(y-c)+\sigma_y^2\mu_x}{\sigma_y^2+a^2\sigma_x^2}\tag{L1.2}$$

$$\tilde{\sigma}_x\equiv\frac{\sigma_y\sigma_x}{\sqrt{\sigma_y^2+a^2\sigma_x^2}}\tag{L1.3}$$

$$\tilde{\mu}_y\equiv a\mu_x+c\tag{L1.4}$$

$$\tilde{\sigma}_y\equiv\sqrt{\sigma_y^2+a^2\sigma_x^2}\tag{L1.5}$$

Proof of Lemma 1.1. Using the definition of the standard normal probability density function:

$$\frac{1}{\sigma_y}\phi\left(\frac{y-\mu_y}{\sigma_y}\right)\times\frac{1}{\sigma_x}\phi\left(\frac{x-\mu_x}{\sigma_x}\right)=\frac{1}{2\pi\sigma_y\sigma_x}\exp\left\{-\frac{(y-\mu_y)^2}{2\sigma_y^2}-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right\}$$

$$=\frac{1}{2\pi\sigma_y\sigma_x}\exp\left\{-\frac{1}{2\sigma_y^2\sigma_x^2}\underbrace{\left[\sigma_x^2(y-\mu_y)^2+\sigma_y^2(x-\mu_x)^2\right]}_{(*)}\right\} \tag{L1.6}$$

Given that $\mu_y = ax + c$, rearrange terms and define $\tilde{\mu}_x \equiv (\sigma_y^2 + a^2\sigma_x^2)^{-1}(\sigma_x^2 a(y - c) + \sigma_y^2\mu_x)$ to obtain,

$$(*)=\sigma_x^2(y-c)^2+(\sigma_x^2a^2+\sigma_y^2)\left(x^2-2x\tilde{\mu}_x\right)+\sigma_y^2\mu_x^2$$

$$=\sigma_x^2(y-c)^2+(\sigma_x^2a^2+\sigma_y^2)\left(x^2-2x\tilde{\mu}_x+\tilde{\mu}_x^2\right)+\sigma_y^2\mu_x^2-(\sigma_x^2a^2+\sigma_y^2)\tilde{\mu}_x^2$$

$$=(\sigma_x^2a^2+\sigma_y^2)\left(x-\tilde{\mu}_x\right)^2+\underbrace{\sigma_x^2(y-c)^2+\sigma_y^2\mu_x^2-(\sigma_x^2a^2+\sigma_y^2)\tilde{\mu}_x^2}_{(**)} \tag{L1.7}$$

Using the definition of $\tilde{\mu}_x$ and rearranging terms yields,

$$(**)=\sigma_x^2\sigma_y^2(\sigma_x^2a^2+\sigma_y^2)^{-1}\left[(y-c)^2-2a\mu_x(y-c)+a^2\mu_x^2\right]$$

$$=\sigma_x^2\sigma_y^2(\sigma_x^2a^2+\sigma_y^2)^{-1}(y-\underbrace{(a\mu_x+c)}_{\equiv\tilde{\mu}_y})^2 \tag{L1.8}$$

Combine (L1.8) and (L1.7) and plug back in (L1.6),

$$\frac{1}{\sigma_y}\phi\left(\frac{y-\mu_y}{\sigma_y}\right)\times\frac{1}{\sigma_x}\phi\left(\frac{x-\mu_x}{\sigma_x}\right)=\frac{1}{2\pi\sigma_y\sigma_x}\exp\left\{-\frac{1}{2}\left(\frac{(x-\tilde{\mu}_x)^2}{(\sigma_x^2a^2+\sigma_y^2)^{-1}\sigma_x^2\sigma_y^2}+\frac{(y-\tilde{\mu}_y)^2}{(\sigma_x^2a^2+\sigma_y^2)}\right)\right\}$$

Finally, define $\tilde{\sigma}_x \equiv (\sigma_y^2 + a^2\sigma_x^2)^{-\frac{1}{2}}(\sigma_y\sigma_x)$ and $\tilde{\sigma}_y \equiv (\sigma_y^2 + a^2\sigma_x^2)^{\frac{1}{2}}$ and rearrange to obtain,

$$\frac{1}{\sigma_y}\phi\left(\frac{y-\mu_y}{\sigma_y}\right)\times\frac{1}{\sigma_x}\phi\left(\frac{x-\mu_x}{\sigma_x}\right) = \underbrace{\frac{1}{(2\pi)^{\frac{1}{2}}\tilde{\sigma}_y}\exp\left\{-\frac{(y-\tilde{\mu}_y)^2}{2\tilde{\sigma}_y}\right\}}_{\frac{1}{\tilde{\sigma}_y}\phi\left(\frac{y-\tilde{\mu}_y}{\tilde{\sigma}_y}\right)}\ \underbrace{\frac{1}{(2\pi)^{\frac{1}{2}}\tilde{\sigma}_x}\exp\left\{-\frac{(x-\tilde{\mu}_x)^2}{2\tilde{\sigma}_x^2}\right\}}_{\frac{1}{\tilde{\sigma}_x}\phi\left(\frac{x-\tilde{\mu}_x}{\tilde{\sigma}_x}\right)}$$

This completes the proof. □

***Lemma* 1.2** *Consider an arbitrary time period $t \in \mathbb{Z}_{[0,\mathrm{T}]}$. If $t = 0$ then,*

$$f_{\mathrm{Z}_{i,t}^{\star}|\mathbf{\Theta}}\left(\mathrm{z}_{i,t}^{\star} \mid \boldsymbol{\theta}\right) \;=\; \delta\left(\mathrm{z}_{i,t}^{\star}\right) \tag{L1.2.1}$$

If $t > 0$, let τ^k denote the largest time period before t such be such that (1.11) *holds and let $b \equiv t - \tau^k$ denote the number of periods elapsed since τ^k, then,*

$$f_{\mathrm{Z}_{i,t}^{\star}|\mathrm{Z}_i^{t-1};\mathbf{\Theta}}\left(\mathrm{z}_{i,t}^{\star} \mid \mathrm{z}_i^{t-1}; \boldsymbol{\theta}\right) \;\propto\; \frac{1}{\sigma_{i,b}}\phi\left(\frac{\mathrm{z}_{i,t}^{\star} - \mu_{i,b}^k}{\sigma_{i,b}}\right)\ \beta_{i,b}^k(\mathrm{z}_{i,t}^{\star}) \tag{L1.2.2}$$

where $\beta_{i,b}^k(\cdot)$ is given by (1.14) *whereas $\mu_{i,b}^k$ and $\sigma_{i,b}$ are given by* (1.15) *and* (1.16).

Proof of Lemma 1.2. For the initial time period (L1.2.1) follows trivially since, by definition, $\mathrm{Z}_{i,t}^{\star} = z_{i,t}^{\star} - z_{i,0}^{\star}$. For any other time period, use the Chapman-Kolmogorov equation and the pdf for $\mathrm{Z}_{i,t}^{\star}$ conditional on $\mathrm{Z}_{i,t}^{\star}$ to obtain,

$$f_{\mathrm{Z}_{i,t}^{\star}|\mathrm{Z}_i^{t-1};\mathbf{\Theta}}\left(\mathrm{z}_{i,t}^{\star} \mid \mathrm{z}_i^{t-1}; \boldsymbol{\theta}\right) = \int \frac{1}{\sigma_\varepsilon}\phi\left(\frac{z^{\star} - (\mu + \tilde{z}^{\star})}{\sigma_\varepsilon}\right) f_{\mathrm{Z}_{i,t-1}^{\star}|\mathrm{Z}_i^{t-1};\mathbf{\Theta}}\left(\mathrm{z}_{i,t-1}^{\star} \mid \mathrm{z}_i^{t-1}; \boldsymbol{\theta}\right) d\tilde{z}^{\star} \tag{L1.3}$$

It remains to be shown that for any t, the expression for the pdf in (L1.2.2) satisfies (L1.3). Given the expressions for the filtered pdf in propositions 1.2 and 1.3 there are two cases to be verified depending on whether $t-1$ is an adjustment or an inaction period. First, suppose that $t-1$ is an *adjustment* period and, hence, the filtered pdf is given by (1.11). Combining the filtered density in (1.11) with (L1.3) yields,

$$
\begin{aligned}
f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}_i^{t-1};\boldsymbol{\Theta}}\left(\mathrm{z}^\star_{i,t} \mid \mathrm{z}_i^{t-1}\,;\boldsymbol{\theta}\right) &= \int \frac{1}{\sigma_{\varepsilon,i}}\phi\left(\frac{\mathrm{z}^\star_{i,t}-(\mu_i+\tilde{z}^\star)}{\sigma_{\varepsilon,i}}\right)\delta(\tilde{z}^\star - c_i^k)\,d\tilde{z}^\star \\
&= \frac{1}{\sigma_{\varepsilon,i}}\phi\left(\frac{\mathrm{z}^\star_{i,t}-(\mu_i+c_i^k)}{\sigma_{\varepsilon,i}}\right) \\
&= \frac{1}{\sigma_{i,1}}\phi\left(\frac{\mathrm{z}^\star_{i,t}-\mu^k_{i,1}}{\sigma_{i,1}}\right)\beta^k_{i,1}(\mathrm{z}^\star_{i,t}) \qquad \text{(L1.4)}
\end{aligned}
$$

where: (i) the second equality uses the properties of the Dirac delta function; (ii) the third equality used the definitions of $\mu^k_{i,b}$ and $\sigma_{i,b}$ in (1.15) and (1.16) for $b=1$ and that $\beta^k_{i,b}(x)=1\ \forall x$ if $b=1$ from (1.14). Second, suppose now $t-1$ is an *inaction* period and, hence, the filtered pdf is given by (1.19). Combining the filtered density in (1.19) with (L1.3) yields,

$$
\begin{aligned}
f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}_i^{t-1};\boldsymbol{\Theta}}\left(\mathrm{z}^\star_{i,t} \mid \mathrm{z}_i^{t-1}\,;\boldsymbol{\theta}\right) &\propto \int_{\underline{Z}_i^k}^{\bar{Z}_i^k} \frac{1}{\sigma_{\varepsilon,i}}\phi\left(\frac{\mathrm{z}^\star_{i,t}-(\mu_i+\tilde{z}^\star)}{\sigma_{\varepsilon,i}}\right)\frac{1}{\sigma_{i,b-1}}\phi\left(\frac{\tilde{z}^\star-\mu^k_{i,b-1}}{\sigma_{i,b-1}}\right)\beta^k_{i,b-1}(\tilde{z}^\star)\,d\tilde{z}^\star \\
&\propto \frac{1}{\sigma_{i,b}}\phi\left(\frac{\mathrm{z}^\star_{i,t}-\mu^k_{i,b}}{\sigma_{i,b}}\right)\int_{\underline{Z}_i^k}^{\bar{Z}_i^k}\frac{1}{\tilde{\sigma}_{i,b-1}}\phi\left(\frac{\tilde{z}^\star-\tilde{\mu}^k_{i,b-1}(\mathrm{z}^\star_{i,t})}{\tilde{\sigma}_{i,b-1}}\right)\beta^k_{i,b-1}(\tilde{z}^\star)d\tilde{z}^\star \\
&\propto \frac{1}{\sigma_{i,b}}\phi\left(\frac{\mathrm{z}^\star_{i,t}-\mu^k_{i,b}}{\sigma_{i,b}}\right)\beta^k_{i,b}(\mathrm{z}^\star_{i,t}) \qquad \text{(L1.5)}
\end{aligned}
$$

where: (i) the first line follows from substituting (1.19) in (L1.3) and re-arranging; (ii) the second line follows from applying lemma 1.1 to combine the two normal densities; (iii) the third line uses the definition of $\beta^k_{i,b}(\mathrm{z}^\star_{i,t})$ in (1.14). □

1.D. Proofs of results in the main text

Proof of Proposition 1.1. For any t it follows from (1.6) that $\ell_{i,j,t}$ depends only on parameter λ_i and realizations of the shock $\nu_{i,t}$ which is drawn from a Uniform (0,1), hence, $\ell_{i,t}$ does not depend on $x_{i,0}$. Iterating (1.5) backwards and using $\mathrm{Z}^\star_{i,0} = 0$ yields,

$$\mathrm{Z}^\star_{i,t} = t\mu_i + \sum_{k=1}^{t} \varepsilon_{i,k} \tag{P1.1.1}$$

Therefore, for any t this depends only on the parameters $\mu_{i,j}$ and $\sigma_{\varepsilon,i,j}$. For any $t > \tau_i^1$, equation (1.3) implies that $\mathrm{Z}_{i,t-1} = \mathrm{Z}^\star_{i,\tau_i^k} - x_{i,0}$ where τ_i^k denotes the last time period where a price change occurred. Subtract $\mathrm{Z}_{i,t-1}$ on both sides of (1.3) and use $\mathrm{Z}_{i,t-1} = \mathrm{Z}^\star_{i,\tau_i^k} - x_{i,0}$ to obtain,

$$\Delta \mathrm{Z}_{i,t} = (\mathrm{Z}^\star_{i,t} - \mathrm{Z}^\star_{i,\tau_i^k})\,(1 - \mathrm{d}_{i,t}) \tag{P1.1.2}$$

Finally, given that $\mathrm{Z}^\star_{i,t}$ is independent of $x_{i,0}$ for any t, it remains to be shown that for $t > \tau_i^1$ also $\mathrm{d}_{i,t}$ is independent of $x_{i,0}$. To see this substitute $\mathrm{Z}_{i,t-1} = \mathrm{Z}^\star_{i,\tau_i^k} - x_{i,0}$ in (1.4) to obtain:

$$\mathrm{d}_{i,t} = \mathbb{1}\{\mathrm{Z}^\star_{i,\tau_i^k} - \mathrm{Z}^\star_{i,t} \in (\underline{x}_i, \bar{x}_i)\}\,(1 - \ell_{i,t}) + \mathbb{1}\{\mathrm{Z}^\star_{i,t} = \mathrm{Z}^\star_{i,\tau_i^k}\}\,\ell_{i,t} \tag{P1.1.3}$$

Given that $\mathrm{Z}^\star_{i,t}$ and $\ell_{i,t}$ do not depend on $x_{i,0}$, this completes the proof. □

Proof of Proposition 1.2. For the initial period, $\mathrm{Z}^\star_{i,t}$ is degenerate at zero since, by definition, $\mathrm{Z}^\star_{i,t} = z^\star_{i,t} - z^\star_{i,0}$. For any other adjustment period, $t = \tau^k$, the distribution must be degenerate at a value that makes the re-centered state gap equal to zero. Using the definition of the re-centered state gap,

$$
\begin{aligned}
x_{i,t} &\equiv z_{i,t} - z^\star_{i,t} - c_i \\
&= (z_{i,t} - z_{i,0}) - (z^\star_{i,t} - z^\star_{i,0}) + (z_{i,0} - z^\star_{i,0} + c_i) \\
&= \mathrm{Z}_{i,t} - \mathrm{Z}^\star_{i,t} + x_{i,0}
\end{aligned}
$$

Therefore, the distribution of $\mathrm{Z}^\star_{i,\tau^k}$ must be degenerate at $\mathrm{z}_{i,\tau^k} + x_{i,0} \equiv c^k_i$. □

Proof of Proposition 1.3. To establish (1.19) start from Bayes' rule,

$$
\begin{aligned}
f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}^t_i;\boldsymbol{\Theta}}\left(\mathrm{z}^\star_{i,t} \mid \mathrm{z}^t_i\,; \boldsymbol{\theta}\right) &= f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}_{i,t};\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}\left(\mathrm{z}^\star_{i,t} \mid \mathrm{z}_{i,t}; \mathrm{z}^{t-1}_i\,; \boldsymbol{\theta}\right) \\
&= \frac{f_{\mathrm{Z}_{i,t}|\mathrm{Z}^\star_{i,t};\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}\left(\mathrm{z}_{i,t} \mid \mathrm{z}^\star_{i,t}; \mathrm{z}^{t-1}_i\,; \boldsymbol{\theta}\right)\, f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}\left(\mathrm{z}^\star_{i,t} \mid \mathrm{z}^{t-1}_i\,; \boldsymbol{\theta}\right)}{\displaystyle\int f_{\mathrm{Z}_{i,t}|\mathrm{Z}^\star_{i,t};\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}\left(\mathrm{z}_{i,t} \mid a; \mathrm{z}^{t-1}_i\,; \boldsymbol{\theta}\right)\, f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}\left(a \mid \mathrm{z}^{t-1}_i\,; \boldsymbol{\theta}\right) da} \\
&\propto f_{\mathrm{Z}_{i,t}|\mathrm{Z}^\star_{i,t};\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}\left(\mathrm{z}_{i,t} \mid \mathrm{z}^\star_{i,t}; \mathrm{z}^{t-1}_i\,; \boldsymbol{\theta}\right)\, f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}\left(\mathrm{z}^\star_{i,t} \mid \mathrm{z}^{t-1}_i\,; \boldsymbol{\theta}\right)
\end{aligned}
\tag{P1.3.1}
$$

Look only at the first term in (P1.3.1) and use the law of total probability,

$$
\begin{aligned}
f_{\mathrm{Z}_{i,t}|\mathrm{Z}^\star_{i,t};\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}\left(\mathrm{z}_{i,t} \mid \mathrm{z}^\star_{i,t}; \mathrm{z}^{t-1}_i; \boldsymbol{\theta}\right) &= f_{\mathrm{Z}_{i,t}|\ell_{i,t};\mathrm{Z}^\star_{i,t};\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}\left(\mathrm{z}_{i,t}|0; \mathrm{z}^\star_{i,t}; \mathrm{z}^{t-1}_i; \boldsymbol{\theta}\right) f_{\ell_{i,t}|\mathrm{Z}^\star_{i,t};\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}\left(0|\mathrm{z}^\star_{i,t}; \mathrm{z}^{t-1}_i; \boldsymbol{\theta}\right) \\
&\quad + f_{\mathrm{Z}_{i,t}|\ell_{i,t};\mathrm{Z}^\star_{i,t};\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}\left(\mathrm{z}_{i,t}|1; \mathrm{z}^\star_{i,t}; \mathrm{z}^{t-1}_i; \boldsymbol{\theta}\right) f_{\ell_{i,t}|\mathrm{Z}^\star_{i,t};\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}\left(1|\mathrm{z}^\star_{i,t}; \mathrm{z}^{t-1}_i; \boldsymbol{\theta}\right)
\end{aligned}
\tag{P1.3.2}
$$

Since t is an inaction period and τ^k is the last adjustment period before t, this implies that $\mathrm{z}_{i,t} = \mathrm{z}_{i,t-1} = \mathrm{z}_{i,\tau^k}$. Using the definition of $d_{i,t}\left(z_{i,t-1}, z^\star_{i,t}, \ell_{i,t}\right)$ in (1.2) and the transition equation for the arrival of costless adjustment opportunities in (1.6), expression (P1.3.2) can be written as,

$$
f_{\mathrm{Z}_{i,t}|\mathrm{Z}^\star_{i,t};\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}\left(\mathrm{z}_{i,t} \mid \mathrm{z}^\star_{i,t}; \mathrm{z}^{t-1}_i; \boldsymbol{\theta}\right) = \mathbb{1}\{\mathrm{z}^\star_{i,t} \in (\underline{Z}^k_i, \bar{Z}^k_i)\}\,(1-\lambda_i) + \mathbb{1}\{\mathrm{z}^\star_{i,t} = z_{i,\tau^k} + x_{i,0}\}\,\lambda_i
\tag{P1.3.3}
$$

where $\underline{Z}_i^k$ and $\bar{Z}_i^k$ are given by (1.12) and (1.13). Substituting (P1.3.3) in (P1.3.1) and re-arranging yields,

$$
\begin{aligned}
f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}^t_i;\boldsymbol{\Theta}}\left(\mathrm{z}^\star_{i,t} \mid \mathrm{z}^t_i\,;\boldsymbol{\theta}\right) \;\propto\; & f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}\left(\mathrm{z}^\star_{i,t} \mid \mathrm{z}^{t-1}_i\,;\boldsymbol{\theta}\right) \mathbb{1}\{\mathrm{z}^\star_{i,t} \in (\underline{Z}^k_i, \bar{Z}^k_i)\} \\
& + \underbrace{\frac{\lambda_i}{(1-\lambda_i)} f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}\left(\mathrm{z}^\star_{i,t} \mid \mathrm{z}^{t-1}_i\,;\boldsymbol{\theta}\right) \mathbb{1}\{\mathrm{z}^\star_{i,t} = z_{i,\tau^k} + x_{i,0}\}}_{=o(z^\star)}
\end{aligned}
\tag{P1.3.4}
$$

Consider now the second term in (P1.3.1) and use the Chapman-Kolmogorov equation to obtain,

$$
f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}\left(\mathrm{z}^\star_{i,t}|\mathrm{z}^{t-1}_i\,;\boldsymbol{\theta}\right) = \int f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}^\star_{i,t-1};\boldsymbol{\Theta}}(\mathrm{z}^\star_{i,t}|\tilde{z}^\star;\boldsymbol{\theta}) f_{\mathrm{Z}^\star_{i,t-1}|\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}(\tilde{z}^\star|\mathrm{z}^{t-1}_i;\boldsymbol{\theta})\, d\tilde{z}^\star
\tag{P1.3.5}
$$

It follows from Assumption 1.1 that,

$$
\mathrm{Z}^\star_{i,t} \mid \mathrm{Z}^\star_{i,t-1};\boldsymbol{\Theta} \sim \mathcal{N}\left(\mu_i + \mathrm{Z}^\star_{i,t-1}, \sigma^2_{\varepsilon,i}\right)
\tag{P1.3.6}
$$

Finally, combining (P1.3.5) and (P1.3.6) and substitute in (P1.3.4) to obtain,

$$
\begin{aligned}
f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}^t_i;\boldsymbol{\Theta}}\left(\mathrm{z}^\star_{i,t} \mid \mathrm{z}^t_i\,;\boldsymbol{\theta}\right) \propto & \left[\int \frac{1}{\sigma_{\varepsilon,i}} \phi\left(\frac{\mathrm{z}^\star_{i,t} - (\mu_i + \tilde{z}^\star)}{\sigma_{\varepsilon,i}}\right) f_{\mathrm{Z}^\star_{i,t-1}|\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}(\tilde{z}^\star|\mathrm{z}^{t-1}_i;\boldsymbol{\theta})\, d\tilde{z}^\star\right] \mathbb{1}\{\mathrm{z}^\star_{i,t} \in (\underline{Z}^k_i, \bar{Z}^k_i)\} \\
& + o(\mathrm{z}^\star_{i,t})
\end{aligned}
\tag{P1.3.7}
$$

This equation is a *filtering forward recursion* as it expresses the filtered pdf a given time period as a function of the filtered pdf in previous time period. Filtering forward recursions are commonplace in the nonlinear non-Gaussian filtering literature.[34] It is important to notice that (P1.3.7) is a "local forward filtering

[34] See, for instance, Kitagawa (1987, equation 2.3) or Särkkä (2013, theorem 4.1).

recursion", since it holds only for inaction periods such that τ^k is the largest time period before t such that (1.11) holds. To complete the proof it remains to be shown that (1.19) satisfies the forward recursion regardless of whether $t-1$ is an *adjustment* or an *inaction* period. Consider first the case where $t-1$ is an *adjustment* period and, in that case, $f_{\mathrm{Z}^\star_{i,t-1}|\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}(\tilde{z}^\star|\mathrm{z}^{t-1}_i;\boldsymbol{\theta})$ is given by (1.11) and (P1.3.7) becomes,

$$
\begin{aligned}
f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}^t_i;\boldsymbol{\Theta}}\left(\mathrm{z}^\star_{i,t} \mid \mathrm{z}^t_i;\boldsymbol{\theta}\right) &\propto \left[\int \frac{1}{\sigma_{\varepsilon,i}}\phi\left(\frac{\mathrm{z}^\star_{i,t}-(\mu_i+\tilde{z}^\star)}{\sigma_{\varepsilon,i}}\right)\delta(\tilde{z}^\star - c^k_i)\,d\tilde{z}^\star\right]\mathbb{1}\{\mathrm{z}^\star_{i,t}\in(\underline{Z}^k_i,\bar{Z}^k_i)\} + o(\mathrm{z}^\star_{i,t})\\
&= \frac{1}{\sigma_{\varepsilon,i}}\phi\left(\frac{\mathrm{z}^\star_{i,t}-(\mu_i+c^k_i)}{\sigma_{\varepsilon,i}}\right)\mathbb{1}\{\mathrm{z}^\star_{i,t}\in(\underline{Z}^k_i,\bar{Z}^k_i)\} + o(\mathrm{z}^\star_{i,t})\\
&= \frac{1}{\sigma_{i,1}}\phi\left(\frac{\mathrm{z}^\star_{i,t}-\mu^k_{i,1}}{\sigma_{i,1}}\right)\beta^k_{i,1}\left(\mathrm{z}^\star_{i,t}\right)\mathbb{1}\{\mathrm{z}^\star_{i,t}\in(\underline{Z}^k_i,\bar{Z}^k_i)\} + o(\mathrm{z}^\star_{i,t})
\end{aligned}
\tag{P1.3.8}
$$

where the first equality uses the properties of the Dirac delta function whereas the second equality uses the definitions of $\mu^k_{i,1}$ and $\sigma_{i,1}$ from (1.15) and (1.16) and $\beta^k_{i,1}(x)\,\forall x$ from (1.14). Therefore, (1.19) satisfies (P1.3.7) if $t-1$ is a period where (1.11) holds. Lastly, consider the case where $t-1$ is an inaction period and $f_{\mathrm{Z}^\star_{i,t-1}|\mathrm{Z}^{t-1}_i;\boldsymbol{\Theta}}(\tilde{z}^\star|\mathrm{z}^{t-1}_i;\boldsymbol{\theta})$ is given by (1.19) lagged by one period. Substituting that into (P1.3.7) yields,

$$
\begin{aligned}
f_{\mathrm{z}^{\star}_{i,t}|\mathrm{z}^{t}_{i};\boldsymbol{\Theta}}\left(\mathrm{z}^{\star}_{i,t}\mid \mathrm{z}^{t}_{i};\boldsymbol{\theta}\right) \propto & \left[\int_{\underline{Z}^{k}_{i}}^{\bar{Z}^{k}_{i}} \frac{1}{\sigma_{\varepsilon,i}}\phi\left(\frac{\mathrm{z}^{\star}_{i,t}-(\mu_i+\tilde{z}^{\star})}{\sigma_{\varepsilon,i}}\right)\frac{1}{\sigma_{i,b-1}}\phi\left(\frac{\tilde{z}^{\star}-\mu^{k}_{i,b-1}}{\sigma_{i,b-1}}\right)\beta^{k}_{i,b-1}\left(\tilde{z}^{\star}\right)\,d\tilde{z}^{\star}\right]\\
& \times \mathbb{1}\{\mathrm{z}^{\star}_{i,t}\in(\underline{Z}^{k}_{i},\bar{Z}^{k}_{i})\}+o(\mathrm{z}^{\star}_{i,t})\\
& = \frac{1}{\sigma_{i,b}}\phi\left(\frac{\mathrm{z}^{\star}_{i,t}-\mu^{k}_{i,b}}{\sigma_{i,b}}\right)\left[\int_{\underline{Z}^{k}_{i}}^{\bar{Z}^{k}_{i}}\frac{1}{\tilde{\sigma}_{i,b-1}}\phi\left(\frac{\tilde{z}^{\star}-\tilde{\mu}^{k}_{i,b-1}(\mathrm{z}^{\star}_{i,t})}{\tilde{\sigma}_{i,b-1}}\right)\beta^{k}_{i,b-1}\left(\tilde{z}^{\star}\right)\,d\tilde{z}^{\star}\right]\\
& \times \mathbb{1}\{\mathrm{z}^{\star}_{i,t}\in(\underline{Z}^{k}_{i},\bar{Z}^{k}_{i})\}+o(\mathrm{z}^{\star}_{i,t})\\
& = \frac{1}{\sigma_{i,b}}\phi\left(\frac{\mathrm{z}^{\star}_{i,t}-\mu^{k}_{i,b}}{\sigma_{i,b}}\right)\beta^{k}_{i,b}\left(\mathrm{z}^{\star}_{i,t}\right)\mathbb{1}\{\mathrm{z}^{\star}_{i,t}\in(\underline{Z}^{k}_{i},\bar{Z}^{k}_{i})\}+o(\mathrm{z}^{\star}_{i,t})
\end{aligned}
\tag{P1.3.9}
$$

where the first line follows from substituting lagged (1.19) in (P1.3.7) and rearranging, the second line follows from using Lemma 1.1 to combine the two normal pdfs inside the integral and the definitions of $\tilde{\mu}^{k}_{i,b-1}(x)$ and $\tilde{\sigma}_{i,b-1}$ from (1.17) and (1.18) and the last line follows from the definition of $\beta^{k}_{i,b}(x)$ in (1.14). Therefore, (1.19) also satisfies (P1.3.7) when $t-1$ is an *inaction* period. □

Proof of Proposition 1.4. To establish (1.25) start from,

$$\begin{aligned} f_{\mathrm{Z}^{\star}_{i,t}|\mathrm{Z}^{\mathrm{T}}_{i};\mathbf{\Theta}}\left(\mathrm{z}^{\star}_{i,t}|\mathrm{z}^{\mathrm{T}}_{i};\boldsymbol{\theta}\right) &= \int f_{\mathrm{Z}^{\star}_{i,t},\mathrm{Z}^{\star}_{i,t+1}|\mathrm{Z}^{\mathrm{T}}_{i};\mathbf{\Theta}}\left(\mathrm{z}^{\star}_{i,t},\tilde{z}^{\star}\mid \mathrm{z}^{\mathrm{T}}_{i};\boldsymbol{\theta}\right)d\tilde{z}^{\star} \\ &= \int f_{\mathrm{Z}^{\star}_{i,t+1}|\mathrm{Z}^{\mathrm{T}}_{i};\mathbf{\Theta}}\left(\tilde{z}^{\star}\mid \mathrm{z}^{\mathrm{T}}_{i};\boldsymbol{\theta}\right) f_{\mathrm{Z}^{\star}_{i,t}|\mathrm{Z}^{\star}_{i,t+1};\mathrm{Z}^{\mathrm{T}}_{i};\mathbf{\Theta}}\left(\mathrm{z}^{\star}_{i,t}\mid \tilde{z}^{\star};\mathrm{z}^{\mathrm{T}}_{i};\boldsymbol{\theta}\right)d\tilde{z}^{\star} \\ &= \int f_{\mathrm{Z}^{\star}_{i,t+1}|\mathrm{Z}^{\mathrm{T}}_{i};\mathbf{\Theta}}\left(\tilde{z}^{\star}\mid \mathrm{z}^{\mathrm{T}}_{i};\boldsymbol{\theta}\right) f_{\mathrm{Z}^{\star}_{i,t}|\mathrm{Z}^{\star}_{i,t+1};\mathrm{Z}^{t}_{i};\mathbf{\Theta}}\left(\mathrm{z}^{\star}_{i,t}\mid \tilde{z}^{\star};\mathrm{z}^{t}_{i};\boldsymbol{\theta}\right)d\tilde{z}^{\star} \\ &= \int f_{\mathrm{Z}^{\star}_{i,t+1}|\mathrm{Z}^{\mathrm{T}}_{i};\mathbf{\Theta}}\left(\tilde{z}^{\star}\mid \mathrm{z}^{\mathrm{T}}_{i};\boldsymbol{\theta}\right) \frac{f_{\mathrm{Z}^{\star}_{i,t+1}|\mathrm{Z}^{\star}_{i,t};\mathbf{\Theta}}\left(\tilde{z}^{\star}\mid \mathrm{z}^{\star}_{i,t};\boldsymbol{\theta}\right) f_{\mathrm{Z}^{\star}_{i,t}|\mathrm{Z}^{t}_{i};\mathbf{\Theta}}\left(\mathrm{z}^{\star}_{i,t}\mid \mathrm{z}^{t}_{i};\boldsymbol{\theta}\right)}{f_{\mathrm{Z}^{\star}_{i,t+1}|\mathrm{Z}^{t}_{i};\mathbf{\Theta}}\left(\tilde{z}^{\star}\mid \mathrm{z}^{t}_{i};\boldsymbol{\theta}\right)}d\tilde{z}^{\star} \\ &= f_{\mathrm{Z}^{\star}_{i,t}|\mathrm{Z}^{t}_{i};\mathbf{\Theta}}\left(\mathrm{z}^{\star}_{i,t}\mid \mathrm{z}^{t}_{i};\boldsymbol{\theta}\right)\int \frac{f_{\mathrm{Z}^{\star}_{i,t+1}|\mathrm{Z}^{\star}_{i,t};\mathbf{\Theta}}\left(\tilde{z}^{\star}\mid \mathrm{z}^{\star}_{i,t};\boldsymbol{\theta}\right) f_{\mathrm{Z}^{\star}_{i,t+1}|\mathrm{Z}^{\mathrm{T}}_{i};\mathbf{\Theta}}\left(\tilde{z}^{\star}\mid \mathrm{z}^{\mathrm{T}}_{i};\boldsymbol{\theta}\right)}{f_{\mathrm{Z}^{\star}_{i,t+1}|\mathrm{Z}^{t}_{i};\mathbf{\Theta}}\left(\tilde{z}^{\star}\mid \mathrm{z}^{t}_{i};\boldsymbol{\theta}\right)}d\tilde{z}^{\star} \end{aligned} \tag{P1.4.1}$$

This equation is a *smoothing backward recursion* as it expresses the smoothed probability density at a given time period as a function of the smoothed probability density in the next period. Just like the forward filtering recursion in (P1.3.7), this type of recursion is commonplace in the nonlinear non-Gaussian filtering and smoothing literature.[35] Suppose t is an inaction period and let τ^k denote the the largest time period before t such that (1.11) holds. In that case, the filtered probability density function $f_{\mathrm{Z}^{\star}_{i,t}|\mathrm{Z}^{t}_{i};\mathbf{\Theta}}\left(\mathrm{z}^{\star}_{i,t}\mid \mathrm{z}^{t}_{i};\boldsymbol{\theta}\right)$ is given by (1.19), the denominator term $f_{\mathrm{Z}^{\star}_{i,t+1}|\mathrm{Z}^{t}_{i};\mathbf{\Theta}}\left(\tilde{z}^{\star}\mid \mathrm{z}^{t}_{i};\boldsymbol{\theta}\right)$ is given by (L1.2.2) and $\mathrm{Z}^{\star}_{i,t+1}\mid \mathrm{Z}^{\star}_{i,t};\mathbf{\Theta}\sim\mathcal{N}\left(\mu_i+\mathrm{Z}^{\star}_{i,t},\sigma^2_{\varepsilon,i}\right)$ from Assumption 1.1. Substituting these three terms into (P1.4.1) and re-arranging terms yields,

$$\begin{aligned} f_{\mathrm{Z}^{\star}_{i,t}|\mathrm{Z}^{\mathrm{T}}_{i};\mathbf{\Theta}}\left(\mathrm{z}^{\star}_{i,t}|\mathrm{z}^{\mathrm{T}}_{i};\boldsymbol{\theta}\right) &\propto \left[\int \frac{\frac{1}{\sigma_{i,b}}\phi\left(\frac{\mathrm{z}^{\star}_{i,t}-\mu^{k}_{i,b}}{\sigma_{i,b}}\right)\frac{1}{\sigma_{\varepsilon,i}}\phi\left(\frac{\tilde{z}^{\star}-(\mu_i+\mathrm{z}^{\star}_{i,t})}{\sigma_{\varepsilon,i}}\right) f_{\mathrm{Z}^{\star}_{i,t+1}|\mathrm{Z}^{\mathrm{T}}_{i};\mathbf{\Theta}}\left(\tilde{z}^{\star}\mid \mathrm{z}^{\mathrm{T}}_{i};\boldsymbol{\theta}\right)}{\frac{1}{\sigma_{i,b+1}}\phi\left(\frac{\tilde{z}^{\star}-\mu^{k}_{i,b+1}}{\sigma_{i,b+1}}\right)\ \beta^{k}_{i,b+1}(\tilde{z}^{\star})}d\tilde{z}^{\star}\right] \\ &\times\beta^{k}_{i,b}(\mathrm{z}^{\star}_{i,t})\ \mathbb{1}\left\{\mathrm{z}^{\star}_{i,t}\in\left(\underline{Z}^{k}_{i},\bar{Z}^{k}_{i}\right)\right\}+o(\mathrm{z}^{\star}_{i,t}) \end{aligned} \tag{P1.4.2}$$

[35] See, for instance, Kitagawa (1987, equation 2.4) or Särkkä (2013, theorem 8.1).

Equation (P1.4.2) is a "local smoothing backward recursion" since it holds only for inaction periods such that τ^k is the largest time period before t such that (1.11) holds. Since I am assuming there is exists a period after t such that (1.11) holds, to complete the proof it remains to be shown that (1.25) satisfies (P1.4.2) regardless of whether $t+1$ is an *adjustment* or an *inaction* period. Consider first the case where $t+1$ is an adjustment period and, in that case, $f_{\mathrm{Z}^\star_{i,t+1}|\mathrm{Z}^{\mathrm{T}}_i;\boldsymbol{\Theta}}\left(\tilde{z}^\star \mid \mathrm{z}^{\mathrm{T}}_i\,;\boldsymbol{\theta}\right)$ is given by (1.11) and $b = \Delta^k - 1$ so (P1.4.2) becomes,

$$f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}^{\mathrm{T}}_i;\boldsymbol{\Theta}}\left(\mathrm{z}^\star_{i,t}|\mathrm{z}^{\mathrm{T}}_i\,;\boldsymbol{\theta}\right) \propto \left[\int \frac{\frac{1}{\sigma_{i,\Delta^k-1}}\phi\left(\frac{\mathrm{z}^\star_{i,t}-\mu^k_{i,\Delta^k-1}}{\sigma_{i,\Delta^k-1}}\right)\frac{1}{\sigma_{\varepsilon,i}}\phi\left(\frac{\tilde{z}^\star-(\mu_i+\mathrm{z}^\star_{i,t})}{\sigma_{\varepsilon,i}}\right)\delta\left(\tilde{z}^\star-c^{k+1}_i\right)}{\frac{1}{\sigma_{i,\Delta^k}}\phi\left(\frac{\tilde{z}^\star-\mu^k_{i,\Delta^k}}{\sigma_{i,\Delta^k}}\right)\ \beta^k_{i,\Delta^k}(\tilde{z}^\star)} d\tilde{z}^\star\right]$$

$$\times\, \beta^k_{i,\Delta^k-1}(\mathrm{z}^\star_{i,t})\ \mathbb{1}\left\{\mathrm{z}^\star_{i,t}\in\left(\underline{Z}^k_i,\bar{Z}^k_i\right)\right\} + o(\mathrm{z}^\star_{i,t})$$

$$= \left[\frac{\frac{1}{\sigma_{i,\Delta^k-1}}\phi\left(\frac{\mathrm{z}^\star_{i,t}-\mu^k_{i,\Delta^k-1}}{\sigma_{i,\Delta^k-1}}\right)\frac{1}{\sigma_{\varepsilon,i}}\phi\left(\frac{c^{k+1}_i-(\mu_i+\mathrm{z}^\star_{i,t})}{\sigma_{\varepsilon,i}}\right)}{\frac{1}{\sigma_{i,\Delta^k}}\phi\left(\frac{c^{k+1}_i-\mu^k_{i,\Delta^k}}{\sigma_{i,\Delta^k}}\right)\beta^k_{i,\Delta^k}(c^{k+1}_i)}\right]\beta^k_{i,\Delta^k-1}(\mathrm{z}^\star_{i,t})\mathbb{1}\left\{\mathrm{z}^\star_{i,t}\in\left(\underline{Z}^k_i,\bar{Z}^k_i\right)\right\} + o(\mathrm{z}^\star_{i,t})$$

$$= \left[\frac{\frac{1}{\breve{\sigma}^k_{i,\Delta^k-1}}\phi\left(\frac{\mathrm{z}^\star_{i,t}-\breve{\mu}^k_{i,\Delta^k-1}}{\breve{\sigma}^k_{i,\Delta^k-1}}\right)\frac{1}{\sigma_{i,\Delta^k}}\phi\left(\frac{c^{k+1}_i-\mu^k_{i,\Delta^k}}{\sigma_{i,\Delta^k}}\right)}{\frac{1}{\sigma_{i,\Delta^k}}\phi\left(\frac{c^{k+1}_i-\mu^k_{i,\Delta^k}}{\sigma_{i,\Delta^k}}\right)\ \beta^k_{i,\Delta^k}(c^{k+1}_i)}\right]\beta^k_{i,\Delta^k-1}(\mathrm{z}^\star_{i,t})\mathbb{1}\left\{\mathrm{z}^\star_{i,t}\in\left(\underline{Z}^k_i,\bar{Z}^k_i\right)\right\} + o(\mathrm{z}^\star_{i,t})$$

$$\propto \frac{1}{\breve{\sigma}^k_{i,\Delta^k-1}}\phi\left(\frac{\mathrm{z}^\star_{i,t}-\breve{\mu}^k_{i,\Delta^k-1}}{\breve{\sigma}^k_{i,\Delta^k-1}}\right)\ \beta^k_{i,\Delta^k-1}(\mathrm{z}^\star_{i,t})\ \chi^k_{i,\Delta^k-1}(\mathrm{z}^\star_{i,t})\mathbb{1}\left\{\mathrm{z}^\star_{i,t}\in\left(\underline{Z}^k_i,\bar{Z}^k_i\right)\right\} + o(\mathrm{z}^\star_{i,t}) \tag{P1.4.3}$$

where the first equality follows from the properties of the Dirac delta function, the second equality uses the result in Lemma 1.1 and the last line follows from noticing that $\beta^k_{i,\Delta^k}(c^{k+1}_i)$ is a constant and $\chi^k_{i,\Delta^k-1}(x) = 1, \forall x$ from the definition in (1.20).[36] Consider now the case where $t+1$ is an *inaction* period, so $b < \Delta^k - 1$ and $f_{\mathrm{Z}^\star_{i,t+1}|\mathrm{Z}^{\mathrm{T}}_i;\boldsymbol{\Theta}}\left(\tilde{z}^\star|\mathrm{z}^{\mathrm{T}}_i\,;\boldsymbol{\theta}\right)$ is given by (1.25) forwarded by one period. In that case, (P1.4.2) becomes,

[36] In the second equality lemma 1.1 is invoked with: $y = c^{k+1}_i$, $x = \mathrm{z}^\star_{i,t}$, $a = 1$, $c = \mu_i$, $\sigma_y = \sigma_{\varepsilon,i}$, $\sigma_x = \sigma_{i,\Delta^k-1}$ and $\mu_x = \mu^k_{i,\Delta^k-1}$.

$$f_{\mathrm{Z}_{i,t}^{\star}|\mathrm{Z}_i^{\mathrm{T}};\boldsymbol{\Theta}}\left(\mathrm{z}_{i,t}^{\star}|\mathrm{z}_i^{\mathrm{T}};\boldsymbol{\theta}\right) \propto \underbrace{\left[\int_{\underline{Z}_i^k}^{\bar{Z}_i^k} \frac{\frac{1}{\sigma_{i,b}}\phi\left(\frac{\mathrm{z}_{i,t}^{\star}-\mu_{i,b}^k}{\sigma_{i,b}}\right)\frac{1}{\sigma_{\varepsilon,i}}\phi\left(\frac{\tilde{z}^{\star}-(\mu_i+\mathrm{z}_{i,t}^{\star})}{\sigma_{\varepsilon,i}}\right)\frac{1}{\breve{\sigma}_{i,b+1}^k}\phi\left(\frac{\tilde{z}^{\star}-\breve{\mu}_{i,b+1}^k}{\breve{\sigma}_{i,b+1}^k}\right)\ \chi_{i,b+1}^k\left(\tilde{z}^{\star}\right)}{\frac{1}{\sigma_{i,b+1}}\phi\left(\frac{\tilde{z}^{\star}-\mu_{i,b+1}^k}{\sigma_{i,b+1}}\right)} d\tilde{z}^{\star}\right]}_{[*]}$$

$$\times\ \beta_{i,b}^k(\mathrm{z}_{i,t}^{\star})\ \mathbb{1}\left\{\mathrm{z}_{i,t}^{\star}\in\left(\underline{Z}_i^k,\bar{Z}_i^k\right)\right\}+o(\mathrm{z}_{i,t}^{\star}) \qquad \text{(P1.4.4)}$$

Working with $[*]$ term only,

$$\begin{aligned}
[*] &= \int_{\underline{Z}_i^k}^{\bar{Z}_i^k} \frac{\frac{1}{\sigma_{i,b+1}}\phi\left(\frac{\tilde{z}^{\star}-\mu_{i,b+1}^k}{\sigma_{i,b+1}}\right)\frac{1}{\tilde{\sigma}_{i,b}}\phi\left(\frac{\mathrm{z}_{i,t}^{\star}-\tilde{\mu}_{i,b}^k(\tilde{z}^{\star})}{\tilde{\sigma}_{i,b}}\right)\frac{1}{\breve{\sigma}_{i,b+1}^k}\phi\left(\frac{\tilde{z}^{\star}-\breve{\mu}_{i,b+1}^k}{\breve{\sigma}_{i,b+1}^k}\right)\ \chi_{i,b+1}^k\left(\tilde{z}^{\star}\right)}{\frac{1}{\sigma_{i,b+1}}\phi\left(\frac{\tilde{z}^{\star}-\mu_{i,b+1}^k}{\sigma_{i,b+1}}\right)} d\tilde{z}^{\star} \\
&= \int_{\underline{Z}_i^k}^{\bar{Z}_i^k} \frac{1}{\breve{\sigma}_{i,b}^k}\phi\left(\frac{\mathrm{z}_{i,t}^{\star}-\breve{\mu}_{i,b}^k}{\breve{\sigma}_{i,b}^k}\right)\frac{1}{\ddot{\sigma}_{i,b+1}^k}\phi\left(\frac{\tilde{z}^{\star}-\ddot{\mu}_{i,b+1}^k(\mathrm{z}_{\mathrm{i,t}}^{\star})}{\ddot{\sigma}_{i,b+1}^k}\right)\chi_{i,b+1}^k\left(\tilde{z}^{\star}\right)d\tilde{z}^{\star} \\
&= \frac{1}{\breve{\sigma}_{i,b}^k}\phi\left(\frac{\mathrm{z}_{i,t}^{\star}-\breve{\mu}_{i,b}^k}{\breve{\sigma}_{i,b}^k}\right)\int_{\underline{Z}_i^k}^{\bar{Z}_i^k}\frac{1}{\ddot{\sigma}_{i,b+1}^k}\phi\left(\frac{\tilde{z}^{\star}-\ddot{\mu}_{i,b+1}^k(\mathrm{z}_{\mathrm{i,t}}^{\star})}{\ddot{\sigma}_{i,b+1}^k}\right)\chi_{i,b+1}^k\left(\tilde{z}^{\star}\right)d\tilde{z}^{\star} \\
&= \frac{1}{\breve{\sigma}_{i,b}^k}\phi\left(\frac{\mathrm{z}_{i,t}^{\star}-\breve{\mu}_{i,b}^k}{\breve{\sigma}_{i,b}^k}\right)\chi_{i,b}^k\left(\mathrm{z}_{i,t}^{\star}\right) \qquad \text{(P1.4.5)}
\end{aligned}$$

where the first equality follows using Lemma 1.1 to combine the first two normal densities in the numerator along with the definitions of $\tilde{\mu}_{i,b}^k(x)$ and $\tilde{\sigma}_{i,b}$ in (1.17) and (1.18), the second equality follows from cancelling out the first term in the numerator with the denominator and using Lemma 1.1 to combine the remaining two normal densities, the last equality follows from the definition of $\chi_{i,b}^k(x)$ in (1.20) for $b < \Delta^k - 1$.[37] Finally, plugging (P1.4.5) back into (P1.4.4) yields,

$$f_{\mathrm{Z}_{i,t}^{\star}|\mathrm{Z}_i^{\mathrm{T}};\boldsymbol{\Theta}}\left(\mathrm{z}_{i,t}^{\star}|\mathrm{z}_i^{\mathrm{T}};\boldsymbol{\theta}\right) \propto \frac{1}{\breve{\sigma}_{i,b}^k}\phi\left(\frac{\mathrm{z}_{i,t}^{\star}-\breve{\mu}_{i,b}^k}{\breve{\sigma}_{i,b}^k}\right)\beta_{i,b}^k\left(\mathrm{z}_{i,t}^{\star}\right)\chi_{i,b}^k\left(\mathrm{z}_{i,t}^{\star}\right)\mathbb{1}\left\{\mathrm{z}_{i,t}^{\star}\in\left(\underline{Z}_i^k,\bar{Z}_i^k\right)\right\}+o(\mathrm{z}_{i,t}^{\star}) \qquad \text{(P1.4.6)}$$

[37] In the first equality, Lemma 1.1 is used with: $y = \tilde{z}^{\star}$, $x = \mathrm{z}_{i,t}^{\star}$, $a = 1$, $c = \mu_i$, $\sigma_x = \sigma_{i,b}$, $\sigma_y = \sigma_{\varepsilon,i}$ and $\mu_x = \mu_b^k$. In the second equality Lemma 1 is again used but now with: $y = \mathrm{z}_{i,t}^{\star}$, $x = \tilde{z}^{\star}$, $a = b/(b+1)$, $c = c_i^k/(b+1)$, $\sigma_y = \tilde{\sigma}_{i,b}$, $\sigma_x = \breve{\sigma}_{i,b+1}^k$ and $\mu_x = \breve{\mu}_{i,b+1}^k$.

Therefore, (1.25) again satisfies the local smoothing backward recursion in (P1.4.2).

□

Proof of Proposition 1.5. This proposition refers to the case where there does *not* exist any period after t such that (1.11) holds and, hence, t can only be either the *end period* (*i.e.* $t = \mathrm{T}$) or an *inaction* period. Consider first the case where t corresponds to the end period. In that case, the smoothed probability density must coincide with the filtered probability density in (1.19). Start from (1.19) with $b = \Delta^K \equiv \mathrm{T} - \tau^k$,

$$f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}^t_i;\boldsymbol{\Theta}}\left(\mathrm{z}^\star_{i,t} \mid \mathrm{z}^t_i\,; \boldsymbol{\theta}\right) \propto \frac{1}{\sigma_{i,\Delta^K}}\phi\left(\frac{\mathrm{z}^\star_{i,t} - \mu^k_{i,\Delta^K}}{\sigma_{i,\Delta^K}}\right)\beta^k_{i,\Delta^K}(\mathrm{z}^\star_{i,t})\mathbb{1}\left\{\mathrm{z}^\star_{i,t} \in \left(\underline{Z}^k_i, \bar{Z}^k_i\right)\right\} + o(\mathrm{z}^\star_{i,t})$$

$$= \frac{1}{\sigma_{i,\Delta^K}}\phi\left(\frac{\mathrm{z}^\star_{i,t} - \mu^k_{i,\Delta^K}}{\sigma_{i,\Delta^K}}\right)\beta^k_{i,\Delta^K}(\mathrm{z}^\star_{i,t})\iota^k_{i,\Delta^K}(\mathrm{z}^\star_{i,t})\mathbb{1}\left\{\mathrm{z}^\star_{i,t} \in \left(\underline{Z}^k_i, \bar{Z}^k_i\right)\right\} + o(\mathrm{z}^\star_{i,t}) \qquad \text{(P1.5.1)}$$

where the equality follows from the fact that $\iota^k_{i,\Delta^K}(x) = 1, \forall x$ from the definition in (1.26). Therefore, (1.27) coincides with the filtered density in (1.19) for the case where the end period is an inaction period. For the case where t is an inaction period but *not* the end period (*i.e* $b < \Delta^K$, the proof is similar to that of proposition 1.4 since it requires showing that the smoothed density in (1.27) satisfies the local backward smoothing backward recursion in (P1.4.2). Suppose $f_{\mathrm{Z}^\star_{i,t+1}|\mathrm{Z}^\mathrm{T}_i;\boldsymbol{\Theta}}\left(\tilde{z}^\star|\mathrm{z}^\mathrm{T}_i\,;\boldsymbol{\theta}\right)$ is given by (1.27) forwarded by one period and substitute in (P1.4.2) to obtain,

$$f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}^\mathrm{T}_i;\boldsymbol{\Theta}}\left(\mathrm{z}^\star_{i,t}|\mathrm{z}^\mathrm{T}_i\,;\boldsymbol{\theta}\right) \propto \underbrace{\left[\int_{\underline{Z}^k_i}^{\bar{Z}^k_i} \frac{\frac{1}{\sigma_{i,b}}\phi\left(\frac{\mathrm{z}^\star_{i,t}-\mu^k_{i,b}}{\sigma_{i,b}}\right)\frac{1}{\sigma_{\varepsilon,i}}\phi\left(\frac{\tilde{z}^\star-(\mu_i+\mathrm{z}^\star_{i,t})}{\sigma_{\varepsilon,i}}\right)\frac{1}{\sigma_{i,b+1}}\phi\left(\frac{\tilde{z}^\star-\mu^k_{i,b+1}}{\sigma_{i,b+1}}\right)\ \iota^k_{i,b+1}\left(\tilde{z}^\star\right)}{\frac{1}{\sigma_{i,b+1}}\phi\left(\frac{\tilde{z}^\star-\mu^k_{i,b+1}}{\sigma_{i,b+1}}\right)} d\tilde{z}^\star\right]}_{[**]}$$

$$\times\, \beta^k_{i,b}(\mathrm{z}^\star_{i,t})\ \mathbb{1}\left\{\mathrm{z}^\star_{i,t} \in \left(\underline{Z}^k_i, \bar{Z}^k_i\right)\right\} + o(\mathrm{z}^\star_{i,t}) \qquad \text{(P1.5.2)}$$

Notice that,

$$\begin{aligned}
[**] &= \frac{1}{\sigma_{i,b}}\phi\left(\frac{\mathrm{z}^\star_{i,t}-\mu^k_{i,b}}{\sigma_{i,b}}\right)\int_{\underline{Z}^k_i}^{\bar{Z}^k_i}\frac{1}{\sigma_{\varepsilon,i}}\phi\left(\frac{\tilde{z}^\star-(\mu_i+\mathrm{z}^\star_{i,t})}{\sigma_{\varepsilon,i}}\right)d\tilde{z}^\star \\
&= \frac{1}{\sigma_{i,b}}\phi\left(\frac{\mathrm{z}^\star_{i,t}-\mu^k_{i,b}}{\sigma_{i,b}}\right)\iota^k_{i,b}(\mathrm{z}^\star_{i,t}) \qquad \text{(P1.5.3)}
\end{aligned}$$

where the first equality follows from cancelling out terms in the numerator in denominator and taking outside the integral terms that do not depend on $\tilde{z}^\star$ whereas the second equality follows from the definition of $\iota^k_{i,b}(x)$ in (1.26) with $b < \Delta^K$. Finally, plugging (P1.5.3) back in (P1.5.2) yields,

$$f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}^{\mathrm{T}}_i;\boldsymbol{\Theta}}\left(\mathrm{z}^\star_{i,t}|\mathrm{z}^{\mathrm{T}}_i;\boldsymbol{\theta}\right) \propto \frac{1}{\sigma_{i,b}}\phi\left(\frac{\mathrm{z}^\star_{i,t}-\mu^k_{i,b}}{\sigma_{i,b}}\right)\beta^k_{i,b}(\mathrm{z}^\star_{i,t})\,\iota^k_{i,b}(\mathrm{z}^\star_{i,t})\,\mathbb{1}\left\{\mathrm{z}^\star_{i,t}\in\left(\underline{Z}^k_i,\bar{Z}^k_i\right)\right\}+o(\mathrm{z}^\star_{i,t}) \tag{P1.5.4}$$

Therefore, (1.27) satisfies the local smoothing backward recursion in (P1.4.2). □

***Corollary* 1.1** *For a given unit i, suppose t is an inaction period. Let τ^k denote the largest time period before t such that* (1.11) *holds and let $b \equiv t-\tau^k$ denote the number of periods elapsed since τ^k. If $\underline{x}_i \to -\infty$ and $\bar{x}_i \to \infty$ then $\beta^k_{i,b}(x) = \chi^k_{i,b}(x) = \iota^k_{i,b} = 1$, $\forall b, k, x$.*

Proof of Corollary 1.1. From the definitions of $\underline{Z}^k_i$ and $\bar{Z}^k_i$ in (1.12) and (1.13) it follows that, for any k, if $\underline{x}_i \to -\infty$ then $\bar{Z}^k_i \to \infty$ and if $\bar{x}_i \to \infty$ then $\underline{Z}^k_i \to -\infty$. Using this fact, I now verify by induction that $\beta^k_{i,b}(x) = 1$ satisfies the recursive definition in (1.14). First, if $b = 1$ it follows by definition that $\beta^k_{i,b}(x) = 1$. Second, suppose $b > 1$ and $\beta^k_{i,b-1}(x) = 1$ then using (1.14) yields,

$$\beta^k_{i,b}(x) = \int_{-\infty}^{\infty}\frac{1}{\tilde{\sigma}_{i,b-1}}\phi\left(\frac{y-\tilde{\mu}^k_{i,b-1}(x)}{\tilde{\sigma}_{i,b-1}}\right)dy = 1 \tag{C1.1.1}$$

This completes the proof for $\beta^k_{i,b}(x)$. Consider now the case of $\chi^k_{i,b}(x)$. If $b =$

$\Delta^k - 1$ then by definition of (1.20) it follows that $\chi^k_{i,b}(x) = 1$. Otherwise, suppose $b < \Delta^k - 1$ and that $\chi^k_{i,b+1}(x) = 1$, then using (1.20) yields,

$$\chi^k_{i,b}(x) = \int_{-\infty}^{\infty} \frac{1}{\ddot{\sigma}^k_{i,b+1}} \phi\left(\frac{y - \ddot{\mu}^k_{i,b+1}(x)}{\ddot{\sigma}^k_{i,b+1}}\right) dy = 1 \tag{C1.1.2}$$

This completes the proof for $\chi^k_{i,b}(x)$. Finally, consider the case of $\iota^k_{i,b}(x)$. If $b = \Delta^K$ then by definition of (1.26) it follows that $\iota^k_{i,b}(x) = 1$. Otherwise, suppose $b < \Delta^K$ and that $\iota^k_{i,b+1}(x) = 1$, then using (1.26) yields,

$$\iota^k_{i,b}(x) = \int_{-\infty}^{\infty} \frac{1}{\sigma_{\varepsilon,i}} \phi\left(\frac{y - (\mu + x)}{\sigma_{\varepsilon,i}}\right) dy = 1 \tag{C1.1.3}$$

This completes the proof. □

1.E. Computational details for smoothed probability densities

This appendix describes how to numerically evaluate the smoothed probability density functions in propositions (1.25) and (1.27) and how to use them to compute smoothed estimates. The challenging terms to keep track of in those expressions are the integral recursions in (1.14), (1.20) and (1.26). In order to compute those terms, the integrals are approximated using Gauss-Legendre quadrature methods.[38] In all that follows, let n^{GL} denote the number of Gauss-Legendre nodes used in those approximations, let $\mathbf{z}^{GL}$ be a $n^{GL} \times 1$ vector of Gauss-Legendre nodes in the interval $[\underline{Z}^k_i, \overline{Z}^k_i]$ sorted in ascending order and $\boldsymbol{\omega}^{GL}$ be the associated $n^{GL} \times 1$ vector of Gauss-Legendre weights.[39]

[38] See, for example, Judd (1998, section 7.2).

[39] To compute the Gauss-Legendre nodes and associated weights on an arbitrary interval $[a, b]$, the *lgwt* function provided by Greg von Winckel on File Exchange and available for download here is used. The description here presented is for a general number of Gauss-Legendre notes n^{GL}. For the Monte Carlo experiment in section 1.5 all the computations are based on 100 Gauss-Legendre nodes.

1.E.1. Computational details for proposition 1.4

Suppose the probability density function is given by expression (1.25) in proposition 1.4. In that case, define $\check{\mu}^k_* \equiv [\check{\mu}^k_{i,1}, \ldots, \check{\mu}^k_{i,\Delta^k-1}]$ and $\check{\sigma}^k_* \equiv [\check{\sigma}^k_{i,1}, \ldots, \check{\sigma}^k_{i,\Delta^k-1}]$. Moreover, let $\mathbf{A}$ be an $n^{GL} \times \Delta^k - 1$ matrix given by,

$$\mathbf{A} = \left[\mathbf{1}_{n^{GL}\times 1} \otimes (\check{\sigma}^k_*)^{\circ -1}\right] \odot \left[\phi \circ \left((\mathbf{1}_{1\times(\Delta^k-1)} \otimes \mathbf{z}^{GL} - \mathbf{1}_{n^{GL}\times 1} \otimes \check{\mu}^k_*) \oslash (\mathbf{1}_{n\times 1} \otimes \check{\sigma}^k_*) \right)\right] \tag{C.1}$$

Moreover, let $\mathbf{B}$ and $\mathbf{C}$ also be $n^{GL} \times \Delta^k - 1$ matrices such that $b_{*,1} = \mathbb{1}_{n\times 1}$ and $c_{*,\Delta^k-1} = \mathbb{1}_{n\times 1}$ and the remaining columns are defined recursively according to,

$$b_{*,j} = \left[\mathbf{I}_n \otimes \boldsymbol{\omega}^{GL'}\right] \left[\frac{1}{\tilde{\sigma}_{i,j-1}} \times \phi \circ \left(\frac{1}{\tilde{\sigma}_{i,j-1}} \left(\mathbf{1}_{n\times 1} \otimes \mathbf{z}^{GL} - \tilde{\mu}^k_{i,j} \circ (\mathbf{z}^{GL} \otimes \mathbf{1}_{n\times 1})\right) \right) \odot (\mathbf{1}_{n\times 1} \otimes b_{*,j-1})\right] \tag{C.2}$$

$$c_{*,j} = \left[\mathbf{I}_n \otimes \boldsymbol{\omega}^{GL'}\right] \left[\frac{1}{\ddot{\sigma}^k_{i,j+1}} \times \phi \circ \left(\frac{1}{\ddot{\sigma}^k_{i,j+1}} \left(\mathbf{1}_{n\times 1} \otimes \mathbf{z}^{GL} - \ddot{\mu}^k_{i,j} \circ (\mathbf{z}^{GL} \otimes \mathbf{1}_{n\times 1})\right) \right) \odot (\mathbf{1}_{n\times 1} \otimes c_{*,j+1})\right] \tag{C.3}$$

Using these three matrices define,

$$\mathbf{F}^c \equiv [\mathbf{A} \odot \mathbf{B} \odot \mathbf{C}] \oslash [\boldsymbol{\omega}^{GL'} (\mathbf{A} \odot \mathbf{B} \odot \mathbf{C})] \tag{C.4}$$

The vector containing the probability density function from proposition (1.25) evaluated at each Gauss-Legendre node is given by:

$$f_{\mathrm{Z}^\star_{i,t}|\mathrm{Z}^{\mathrm{T}}_i;\boldsymbol{\Theta}} \circ \left(\mathbf{z}^{GL} \mid \mathrm{z}^{\mathrm{T}}_i ; \boldsymbol{\theta}\right) \approx \mathbf{F}^c_{*,b} \tag{C.5}$$

Moreover, for a given function $g : \mathbb{R} \to \mathbb{R}$ one can compute its smoothed expectations as,

$$\mathbb{E}\left[\mathrm{Z}^{\star}_{i,t} \mid \mathrm{z}^{\mathrm{T}}_{i} \, ; \boldsymbol{\theta}\right] \approx \boldsymbol{\omega}^{GL'} \left[\, g \circ (\mathbf{z}^{GL}) \odot \mathbf{F}^{c}_{*,b} \,\right] \tag{C.6}$$

Finally, the smoothed estimates of $\mathrm{Z}^{\star}_{i,t}$ at time t are obtained from (C.6) using $g(\cdot)$ equal to the identity function.

1.E.2. Computational details for proposition 1.5

Suppose now the probability density function is given by expression (1.27) in proposition 1.5. In that case, define $\mu_* = [\mu_{i,1}, \ldots, \mu_{i,\Delta^K}]$ and $\sigma_* = [\sigma_{i,1}, \ldots, \sigma_{i,\Delta^K}]$. Let $\mathbf{D}$ be a $n^{GL} \times \Delta^K$ matrix given by,

$$\mathbf{D} = \left[\mathbf{1}_{n\times 1} \otimes (\sigma_*)^{\circ -1}\right] \odot \left[\phi \circ \big((\mathbf{1}_{1\times\Delta^K} \otimes \mathbf{z}^{GL} - \mathbf{1}_{n\times 1} \otimes \mu_*) \oslash (\mathbf{1}_{n\times 1} \otimes \sigma_*) \big)\right] \tag{C.7}$$

Let $\tilde{\mathbf{B}}$ and $\mathbf{E}$ be $n^{GL} \times \Delta^K$ matrices such that $\tilde{b}_{*,1} = \mathbb{1}_{n\times 1}$ and $e_{*,\Delta^K} = \mathbb{1}_{n\times 1}$. The remaining columns of $\tilde{\mathbf{B}}$ are defined recursively according to (C.2) and the remaining columns of $\mathbf{E}$ according to:

$$e_{*,j} = \left[\, \mathbf{I}_n \otimes \boldsymbol{\omega}^{GL'} \right] \left[\frac{1}{\sigma_{\varepsilon,i}} \times \phi \circ \left(\frac{1}{\ddot{\sigma}^{k}_{i,j+1}} \left(\mathbf{1}_{n\times 1} \otimes \mathbf{z}^{GL} - (\mu \mathbf{1}_{n\times 1} + \mathbf{z}^{GL}) \otimes \mathbf{1}_{n\times 1}\right) \right) \odot (\mathbf{1}_{n\times 1} \otimes e_{*,j+1}) \right] \tag{C.8}$$

Using these three matrices define,

$$\mathbf{F}^{i} \equiv [\mathbf{D} \odot \tilde{\mathbf{B}} \odot \mathbf{E}] \oslash [\boldsymbol{\omega}^{GL'} \Big(\mathbf{D} \odot \tilde{\mathbf{B}} \odot \mathbf{E}\Big)] \tag{C.9}$$

The vector containing the probability density function from proposition (1.27) evaluated at each Gauss-Legendre node is given by:

$$f_{\mathrm{Z}^{\star}_{i,t}|\mathrm{Z}^{\mathrm{T}}_{i};\boldsymbol{\Theta}} \circ \left(\mathbf{z}^{GL} \mid \mathrm{z}^{\mathrm{T}}_{i}\,; \boldsymbol{\theta}\right) \approx \mathbf{F}^{i}_{*,b} \tag{C.10}$$

Moreover, for a given function $g : \mathbb{R} \to \mathbb{R}$ one can compute its smoothed expectations as,

$$\mathbb{E}\left[\mathrm{Z}^{\star}_{i,t} \mid \mathrm{z}^{\mathrm{T}}_{i}\,; \boldsymbol{\theta}\right] \approx \boldsymbol{\omega}^{GL'}\left[g \circ \left(\mathbf{z}^{GL}\right) \odot \mathbf{F}^{i}_{*,b}\right] \tag{C.11}$$

Finally, the smoothed estimates of $\mathrm{Z}^{\star}_{i,t}$ at time t are obtained from (C.11) using $g(\cdot)$ equal to the identity function.

1.F. Computational details for parameter estimation

This appendix describes the algorithms used to obtain parameter estimates for the Monte Carlo experiment in section 1.5. As describes in section 1.3, parameter estimation is done in two stages.[40]

1.F.1. Algorithm for first-stage parameter estimation

Since in all the DGPs considered in table 1.B.1, the vector of parameters Γ_i is assumed to be common across all units so the simulator is computed from the distribution of state variable changes after the first adjustment *pooled* across all units. Let $\Delta\mathbf{Z}^{\mathrm{T}}_{t=\tau^1+1} \equiv \{\{\Delta\mathrm{Z}_{i,t}\}^{\mathrm{T}}_{\tau^1_i+1}\}^{n}_{i=1}$ denote that vector observed in actual data. The counterpart for this vector in a panel of simulated data is denoted by $\Delta\mathbf{Z}^{\mathrm{T,s}}_{t=\tau^1+1}\left(\boldsymbol{\Gamma}, \boldsymbol{\xi^s}\right) \equiv \{\{\Delta\mathrm{Z}_{i,t}\left(\boldsymbol{\Gamma}, \boldsymbol{\xi^s_i}\right)\}^{\mathrm{T}}_{\tau^{1,s}_i+1}\}^{n}_{i=1}$ where $\boldsymbol{\xi}^s = \{\boldsymbol{\varepsilon}^s_t, \boldsymbol{\nu^s_t}\}^{\mathrm{T}}_{t=1}$ is a particular simulation for the two vectors of shocks drawn from their respective distribu-

[40] All the algorithms described in this appendix are implemented in MATLAB R2015b. Codes used for parameter estimation and computation of smoothed estimates are available from the author upon request.

tions in assumption 1.1.[41] Using this notation and given that moment conditions are equally weighted in the Monte Carlo experiment, estimates for common parameters are obtained by minimising the following objective function,

$$\mathrm{G}\left(\Delta\mathbf{Z}^{\mathrm{T}}_{t=\tau^1+1},\boldsymbol{\Gamma},\{\boldsymbol{\xi}^s\}_{s=1}^{\mathcal{S}}\right)=\left\|g\left(\Delta\mathbf{Z}^{\mathrm{T}}_{t=\tau^1+1}\right)-\frac{1}{\mathcal{S}}\sum_{s=1}^{\mathcal{S}}g\left(\Delta\mathbf{Z}^{\mathrm{T,s}}_{t=\tau^1+1}\left(\boldsymbol{\Gamma},\boldsymbol{\xi}^s\right)\right)\right\|^2 \tag{D.1}$$

Algorithm 1. To minimise (D.1) the following algorithm is used,

1. Draw 50 vectors of primitive shocks and denote them by $\{\tilde{\boldsymbol{\xi}}^s\}_{s=1}^{50}$.
2. Choose an initial value for the vector of parameters, say $\boldsymbol{\Gamma}^0$.
3. Use a global search algorithm to search for $\boldsymbol{\Gamma}$ that minimises $\mathrm{G}\left(\Delta\mathbf{Z}^{\mathrm{T}}_{t=\tau^1+1},\boldsymbol{\Gamma},\{\tilde{\boldsymbol{\xi}}^s\}_{s=1}^{50}\right)$.[42]
4. The search is subject to the restrictions: $\underline{\boldsymbol{x}}\leqslant 0$, $\overline{\boldsymbol{x}}\geqslant 0$, $\sigma_\varepsilon\geqslant 0$ and $\boldsymbol{\lambda}\in[0,1]$.

Some points about algorithm 1 above are worth emphasising. First, in step 3 global search methods are preferred over gradient based ones since in preliminary simulations the later failed to converge in many instances. Second, the vectors of primitive shocks in $\{\tilde{\boldsymbol{\xi}}^s\}_{s=1}^{50}$ are drawn only *once* at the beginning of the algorithm and kept fixed when searching for a minimum in step 3. This is important as otherwise the algorithm would not numerically converge and the asymptotic statistical properties would no longer be valid.[43] Third, in the Monte Carlo experiment steps 2, 3 and 4 are performed twice starting different initial conditions.[44] In case the results differ, the value of parameters that yields the

[41] It is important to notice that the drawings of primitive shocks are done such that in simulated data, the number of units and their respective starting and end dates exactly match the structure that is observed in actual data. The general principle in simulated based estimation is treating real and simulated data as similarly as possible. For an example of the consequences of not respecting this principle refer to Berger, Caballero and Engel (2018) who show that using a number of units in simulations that is larger than the number of units in actual data can lead to underestimate the shock persistence in a Calvo model (see table 1, p. 12).

[42] In MATLAB R2015b the algorithm *patternsearch* is used with the default options.

[43] See, for instance, p. 29 in Gouriéroux and Monfort (1996).

[44] The first set of initial conditions is designed to be an educated guess for DGP in which most state variable adjustments are triggered by state gaps leaving the inaction region. In that case, the initial values for $-\underline{x}$ and $-\bar{x}$ are set to the average values of positive and negative state variable adjustments changes across all units, respectively, whereas initial value for λ is equal to 25% of the frequency of state variable adjustments in the data. The second initial condition

smallest objective function is chosen.

1.F.2. Algorithm for second-stage parameter estimation

For the estimation of unit-specific initial conditions $x_{i,0}$'s, the minimisation problem in (1.8) has to be solved once for each unit that has at least one state variable adjustment in the data. Given the large number of such unit in some of the proposed DGPs performing separate minimisations using global search methods would be computationally demanding. Instead, given the vector of parameters $\mathbf{\Gamma}$ is common across units, the minimisation for the second stage is performed on a grid of possible values of $x_{i,0}$.

Algorithm 2. To estimate unit-specific initial conditions the following algorithm is used:

1. Determine the number of panels to be simulated as $S = \lceil 10^4/n \rceil$.[45]

2. Generate S vectors of primitive shocks conform with the data template, $\mathbf{\Xi} = \{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^S\}$.

3. Create a grid $\mathcal{G} = [x_{i,0}^{(1)}, \ldots, x_{i,0}^{(50)}]$, where $x_{i,0}^{(k)} = \hat{\underline{x}} + \frac{k}{(50+1)}(\hat{\bar{x}} - \hat{\underline{x}})$.

4. Set $\mathbf{\Gamma} = \hat{\mathbf{\Gamma}}$ and $x_{i,0} = x_{i,0}^{(k)}, \forall i$ and use assumption 1.1 and Ξ to generate S panels of data.

5. For a given collection of panels compute $f(x_{i,0}^{(k)}) = (S\,n)^{-1} \sum_{s=1}^{S} \sum_{i=1}^{n} h(\{\Delta \mathrm{Z}_{i,j}^{s}(x_{i,0}^{(k)}, \hat{\mathbf{\Gamma}})\}_{t=1}^{\tau_i^{1,s}})$

6. Repeat steps 4 and 5 for each $x_{i,0}^{(k)} \in \mathcal{G}$ and store $\mathcal{F} = \{f(x_{i,0}^{(k)})\}_{k=1}^{50}$

7. Create a new grid $\tilde{\mathcal{G}} = [\tilde{x}_{i,0}^{(1)}, \ldots, \tilde{x}_{i,0}^{(50,000)}]$ where $\tilde{x}_{i,0}^{(i)} = x_{i,0}^{(1)} + \frac{(i-1)}{(49,999)}(x_{i,0}^{(50)} - x_{i,0}^{(1)})$

8. For each $\tilde{x}_{i,0}^{(k)} \in \tilde{\mathcal{G}}$ use a cubic spline on the values in $\mathcal{F}$ to approximate $f(\tilde{x}_{i,0}^{(i)})$.[46]

is designed to be an educated guess for a DGP in which most state variable adjustments are triggered by the arrival of costless adjustment opportunities. In that case, $-\underline{x}$ and $-\bar{x}$ are set to the 95th and 5th percentiles of the distribution of state variable changes changes pooled across all units, respectively, and λ is set to 75% of the frequency of state variable changes observed in the data.

[45] This ensures the simulator is based on at least 10 thousand individual price trajectories.

[46] In MATLAB R2015b this cubic spline is constructed using the function *interp1* in with the option 'spline'.

9. For each $\tilde{x}_{i,0}^{(k)} \in \tilde{\mathcal{G}}$ compute $\tilde{H}(\Delta \mathrm{Z}_i, \tilde{x}_{i,0}^{(k)}, \hat{\mathbf{\Gamma}}) = \left\| (h(\{\Delta \mathrm{Z}_{i,t}\}_{t=1}^{\tau_i^1}) - f(\tilde{x}_{i,0}^{(k)})) \oslash h(\{\Delta \mathrm{Z}_{i,t}\}_{t=1}^{\tau_i^1}) \right\|^2$

10. For a given unit with at least one state variable adjustment $\hat{x}_{i,0} = \arg\min_{a \in \tilde{\mathcal{G}}} \tilde{H}(\{\Delta \mathrm{Z}_{i,t}\}_{t=1}^{\tau_i^1}, a, \hat{\mathbf{\Gamma}})$

11. Repeat steps 9 and 10 for each unit with at least one state variable adjustment.

12. For the remaining units set $\hat{x}_{i,0}$ equal to the average of values in step 11.

Some points about algorithm 2 are worth emphasising. First, in step 9 the deviations of data moments from their simulated counterparts are expressed as percentage deviations of the data moments. That normalisation is necessary with equally weighted moments to ensure that one moment condition does not receive a disproportional weight simply due to differences in scale. Second, assuming $\mathbf{\Gamma}$ is common across all units once the grid for the approximation of the moment conditions is constructed (steps 1 to 8) it can be used as the moment simulator for other units, in practice, this implies that only steps 9 and 10 need to be repeated at the quote-line each speeds up the calculations. Third, the initial grid $\mathcal{F}$ could be equivalently generated by simply generating 10,000 separate individual price trajectories all starting from a given initial condition and computing average over those. The construction in steps 1 to 6 simply takes advantage of some functions used to generate simulated data for the estimation of common parameters.

Chapter 2

Frictionless Inflation

2.1. Introduction

The idea that some form of pricing friction hampers the responses of nominal prices to changes in economic conditions lies at the heart of the new Keynesian framework that has emerged as the workhorse for the analysis of monetary policy and its implications for inflation, economic fluctuations and welfare. One pricing friction that is commonly used to rationalise the delayed response of prices to shocks is the existence of menu costs, that is, the existence of costs that have to be incurred whenever prices are changed, independently of the size of the change.[1] In the presence of menu costs, price changes will only occur when the resulting increase in profits is sufficiently large to outweigh the associated costs. In addition to their intuitive appeal, so-called menu cost models are also consistent with some of the stylised facts about price setting observed in micro price data (Klenow and Malin, 2010; Nakamura and Steinsson, 2008, 2013).

This paper introduces a measure of *frictionless inflation* designed to estimate the counterfactual inflation that would have been observed in a hypothetical world where each price-setter in the economy was exposed to the same environment but

[1] The classic example of menu costs of price adjustment is the problem of a restaurant owner that has to print new menus whenever the price of an item is changed. In a broader sense, menu costs can be thought of as resulting from costs of information, decision and implementation of a pricing strategy.

could have changed its prices without incurring in menu costs.[2,3] Formally, the construction of this measure is grounded on a state-space representation of microeconomic pricing behavior implied by a random menu cost model.[4] Given the form of this representation, the smoother for dynamic (S,s) economies developed in Bandeira (2020) is applied to over 2.2 million individual quote-lines underlying the construction of the UK Consumer Price Index (CPI) to produce a time-series of year-over-year inflation at a monthly frequency spanning the period from 1997 to 2018. This novel time-series of frictionless inflation is then used to address four questions.

First, what is the quantitative importance of menu costs at the microeconomic level for aggregate inflation dynamics? In the data, over the last two decades inflation and frictionless inflation co-move positively but not perfectly so and their difference can be up to 2.29 percentage points in year-over-year terms. However, in the second half of the sample the difference between the two-series decreases and their correlation increases. Altogether, the evidence presented is indicative that menu costs at the microeconomic level matter for aggregate inflation dynamics but their importance seems to have decreased over time. Two explanations that are consistent with this decrease in importance are the increase

[2] In a very stylised way, suppose it was possible to go back in time and not only collect the nominal prices of products across different shopping outlets but to also ask the person in charge of setting that price "what is the price of this product that would maximise your profits over the next month?". The answer to this last question is what is referred in this paper as the *frictionless price* and the measure of *frictionless inflation* here presented is designed to estimate the inflation that would have been observed if price indices were constructed based on frictionless prices instead of the actual shelf prices.

[3] The price that maximises firm's profits in the absence of price adjustment frictions has received different labels in the literature, such as, *frictionless profit-maximising price* (Alvarez, Le Bihan and Lippi, 2016), *frictionless optimum* (Midrigan, 2011, figure 3) or *static desired price* (Nakamura and Steinsson, 2010, figure 4).

[4] It is important to notice that the interpretation given to a measure frictionless inflation is necessarily dependent on which frictions are present in the model chosen to characterise microeconomic pricing decisions. This paper studies price setting decisions as implied by a random menu cost model in which the *only* mechanism preventing price adjustment is the presence of menu costs. Therefore, the notion of frictionless inflation here adopted is with respect to *one* particular friction, menu costs, and not to other frictions that have been proposed in the literature such as sticky-information (Mankiw and Reis, 2002) or information capacity constraints (Woodford, 2009). Nonetheless, as we shall later see, the random menu cost model used to construct the frictionless inflation nests as special cases the Calvo (1983) model of staggered price setting and the canonical menu cost model of Barro (1972) which are two of the most popular models of nominal rigidities (section 2.2). Moreover, the proposed model can match most of the key moments observed in micro price data (section 2.3).

in product competition from Chinese import penetration in early 2000s and a decrease in inflation uncertainty generated by the change in inflation target by the Bank of England in late 2003.

Second, what is the importance of menu costs for the transmission of monetary shocks? More precisely, this paper investigates whether in the data the responses of inflation and its frictionless counterpart to a monetary policy shock are line with their model implied behavior from the new Keynesian framework. First, it is shown that in the basic new Keynesian model from Galí (2008) frictionless inflation should respond *more* than inflation upon impact of a monetary policy shock but should respond *less* in all the subsequent periods. In the data, however, the constructed time-series of frictionless inflation does not react significantly to the series of high-frequency identified monetary surprises from Cesa-Bianchi, Thwaites and Vicondoa (2020). In summary, the empirical evidence presented is at odds with the monetary policy transmission mechanism in the basic new Keynesian model.

Third, what is the the relationship between frictionless inflation and the new Keynesian notion of output gap? Based on the version of basic new Keynesian model from Galí (2008), this paper shows that the wedge between inflation and its frictionless counterpart is negatively proportional to the *changes* in the aggregate output gap, which is defined as the log deviation between output and its flexible price counterpart. This relationship is qualitatively supported in the data using several alternative proxies for the output gap. This result is suggestive that, when combined with a specification for product demand, the measures of frictionless inflation could be also used to quantify changes in the output gap at disaggregated levels where typically measures of output are not available.

Fourth, can the constructed time-series of frictionless inflation be used to improve headline inflation forecasts? In order to investigate this question, in the spirit of Blinder and Reis (2005) this paper uses a series of horse-race type regressions in which headline inflation forecasts based on the cumulated headline inflation over the previous 12 months are compared with the forecasts based on the cumulated frictionless inflation over the previous 12 months. At all forecasting horizons considered, the forecasts based on past frictionless inflation outperform those based on past headline inflation both in-sample and out-of-sample. The results here presented are suggestive that the constructed time-series of frictionless inflation does contain information that can be used to improve headline inflation forecasts.

Relation to the literature. This paper relates to two strands of literature. First, it relates with papers that have estimated related measures of inflation, most notably, the *reset price inflation* from Bils, Klenow and Malin (2012) and the *frictionless optimal price inflation* from Bonomo, Correa and Medeiros (2013). In a context where price-setters follow two-sided (S, s) pricing rules, as it will be assumed in section 2.2, all these three measures are theoretically equivalent.[5] Having said that, this paper differs and complements Bils, Klenow and Malin (2012) and Bonomo, Correa and Medeiros (2013) in three dimensions. First, the methodology used to construct the time-series of frictionless inflation here presented is diametrically different from the existing ones. Second, this paper uses micro price data underlying the UK CPI, whilst Bils, Klenow and Malin (2012) use US CPI micro data and Bonomo, Correa and Medeiros (2013) use Brazilian CPI micro data. Third, the measures of inflation in Bils, Klenow and Malin (2012) and Bonomo, Correa and Medeiros (2013) are used to test implications from different models of price rigidities whereas in in this paper the time-series frictionless inflation is used not only to test some implications from the basic new Keynesian model in Galí (2008) but it is also used to evaluate the impact of menu costs on aggregate inflation dynamics and to improve headline inflation

[5] My measure of frictionless inflation is conceptually equivalent to the frictionless optimal price inflation from Bonomo, Correa and Medeiros (2013), that is, the inflation that would be observed in the counterfactual scenario where all price-setters chose the prices that maximise their static profits period by period. That measure is conceptually different from the reset price inflation. In the words of Bils, Klenow and Malin (2012, p. 2803), "[...] new prices need not be viewed as frictionless spot prices. If future spot prices are expected to differ from the current spot price, then a newly set price may be influenced by future expected spot prices. Thus, reset price inflation can deviate from spot price inflation.". However, if price-setters follow two-sided (S, s) pricing rules, then frictionless prices and reset prices will differ by a *constant* and, hence, frictionless inflation and reset price inflation coincide. This point is also made in Bonomo, Correa and Medeiros (2013, pp. 19-20).

forecasts.[6]

Second, it relates to a literature that aims to quantify the size of price adjustment costs and their importance for individual pricing decisions (Levy, Bergen, Dutta and Venable, 1997; Blinder, Canetti, Lebow and Rudd, 1998; Dutta, Bergen, Levy and Venable, 1999; Zbaracki, Ritson, Levy, Dutta and Bergen, 2004; Anderson, Jaimovich and Simester, 2015). It contributes to this literature in two fundamental ways. First, in contrast with previous papers that have focused on quantifying the costs of price adjustments, this paper focuses not on the size of menu costs *per se* but on their influence on aggregate inflation dynamics. In spirit, this approach is similar to Gorodnicheko and Weber (2016) who focus on the implications of menu costs for the responses of the stock market returns of different firms to a monetary policy surprise. The second contribution is in terms of scope, since the measure of frictionless inflation here presented is based on prices of hundreds of different products that are representative of the typical basket of consumer goods and were collected in thousands of outlets across the UK. This contrasts with the existing literature that relies on very detailed micro data that covers a limited set of products or store chains.

Structure of the paper. The remainder of the paper is structured as follows. Section 2.2 describes the methodology used to construct the measure of frictionless inflation. Section 2.3 describes the micro price data used to construct that measure and presents parameter estimates. Section 2.4 investigates the importance of menu costs of price adjustment at the micro level for aggregate inflation dynamics. Section 2.5 present an empirical test of the monetary transmission mechanism in the basic new Keynesian model in Galí (2008). Section 2.6 explores the relationship between frictionless inflation and the movements in the output gap. Section 2.7 inspects whether the constructed time-series of frictionless inflation can be used to improve inflation forecasts. Section 2.8 concludes and discusses some promising avenues for future research.

[6] The key focus of the methodology introduced in Bonomo, Correa and Medeiros (2013) is the estimation of strategic complementarities and the measure of frictionless optimal price inflation and it's subsequent analysis are produced as a by-product of their main exercise.

2.2. Constructing a measure of frictionless inflation

In order to produce estimates of frictionless inflation based on micro price data, this paper proceeds in a sequence of three steps. First, a single product menu cost model with random costless adjustment opportunities is used to provide a connection between observed individual prices and their unobserved frictionless counterparts, which are the ultimate object of interest. Second, the model implied pricing dynamics are cast in a state-space representation that is assumed to be the process generating the observed micro price data. Last, given the form of that state-space representation, the smoother for dynamic (S,s) economies proposed in Bandeira (2020) is used to obtain frictionless inflation estimates from micro price data. The following subsections explain each of these steps in detail.

2.2.1. A menu cost model with random costless adjustment opportunities

Consider the problem of a firm that sells a single product and chooses a pricing policy to maximise her discounted stream profits net of adjustment costs. Let $p(t)$ denote the log of the firm's nominal price with initial value p_0 given. Let $p^\star(t)$ denote the log of the *frictionless* price, that is, the price that maximises firm's instantaneous flow of profits, and assume it evolves according to a Brownian motion with drift,

$$dp^\star(t) = \mu dt + \sigma dW(t) \tag{2.1}$$

where $W(t)$ is a standard Brownian motion and the initial value $p_0^\star$ given. To adjust her nominal prices the firm must pay a menu cost. Following Stokey (2009) and Alvarez, Le Bihan and Lippi (2016) assume that in a period of length dt this menu cost amounts to κ with probability $1 - \lambda dt$ or *zero* with probability λdt.

Let $\tilde{x}(t)$ denote the *price gap* defined as the log difference between the firm's nominal price and its frictionless counterpart, that is, $\tilde{x}(t) \equiv p(t) - p^{\star}(t)$. Finally, assume the firm's instantaneous flow of profits can be written as $f(\tilde{x}(t))$ with $f(\cdot)$ is continuous, strictly increasing in $(-\infty, 0)$ and strictly decreasing in $(0, +\infty)$.[7] Formally, the firm chooses a sequence of adjustment times and corresponding price adjustments to solve the following stochastic impulse control problem,

$$\mathbb{V}(\tilde{x}_0) = \max_{\{\tau_j, \Delta p_{\tau_j}\}_{j=0}^{\infty}} \mathbb{E}\left[\int_0^{\infty} e^{-\rho t} f(\tilde{x}(t))\, dt - \sum_{j=0}^{\infty} e^{-\rho \tau_j} \kappa (1 - \ell_{\tau_j}) \;\Big|\; \tilde{x}(0) = \tilde{x}_0\right] \tag{2.2}$$

where $\tilde{x}(t) = \tilde{x}(0) - \mu t - \sigma W(t) + \sum_{j=0}^{\infty} \mathbb{1}\{\tau_j < t\} \Delta p_{\tau_j}$, the variable ℓ_{τ_j} is an indicator that takes the value one if the adjustment occurs upon a costless adjustment opportunity and ρ denotes the firm's discount factor.

Special cases. On the one hand, in the limiting case without costless adjustment opportunities (*i.e.* $\lambda = 0$) the problem in (2.2) is a simplified version of the problem in Golosov and Lucas (2007) and almost identical to those in Barro (1972) and Alvarez, Beraja, Gonzalez-Rozada and Neumeyer (2019). On the other hand, if changing prices is infinitely costly (*i.e.* $\kappa \to \infty$) price adjustments occur at random time intervals, determined by the arrival of costless adjustment opportunities, as in the Calvo (1983) model of staggered price setting.

Optimal pricing policy. Given the simple form of the firm's instantaneous flow of profits and the assumed law of motion for the log of frictionless prices in (2.1), it is well-known that the solution to (2.2) takes the form of a two-sided (S,s) policy.[8] Such policy is fully characterised by a triplet of parameters $L < c < U$, where the interval (L, U) represents an *inaction region* and c a reset point. The implied optimal pricing behavior consists in adjusting nominal prices when either a costless adjustment opportunity arrives or when the current price

[7] For a microfounded firm problem that gives rise to instantaneous flow of profits satisfying these assumptions refer, for example, to Alvarez, Lippi and Paciello (2018).

[8] For a formal proof refer, for example, to Stokey (2009, chapter 7). For a discussion of other sufficient conditions for the solution to (2.2) to take the form of a two-sided (S,s) policy refer to Plehn-Dujowich (2005).

gap lies outside the inaction region. The size of price adjustments is such that it makes the price gap equal to the reset point at the adjustment periods. The optimal pricing behavior implied by two-sided (S,s) policies is illustrated in figure 2.A.1.

2.2.2. A state-space representation of microeconomic pricing dynamics

In the micro price data used to produce the United Kingdom Consumer Price Index, an individual price quote corresponds to the nominal price of an *item* collected in a unique shopping venue at a given month.[9] A sequence of monthly price quotes collected for the same item in the same shopping venue is referred to as a *quote-line*. A cornerstone assumption underlying the construction of a measure of frictionless inflation in this paper is that each quote-line observed in micro price data is generated by a firm pursuing a discrete-time approximation of the two-sided (S,s) policies that solve the pricing problem in (2.2). This subsection formally introduces that assumption and shows how it maps to the data generating process in Bandeira (2020).

Terminology and notation. Items are indexed by $j = 1, \ldots, J$, shopping venues at which prices of a given item are collected are indexed by $i = 1, \ldots, n_j$ and months indexed by t. The log of an individual price quote is denoted by $p_{i,j,t}$ and the *cumulated inflation* for that price quote is defined as $\mathrm{Z}_{i,j,t} \equiv p_{i,j,t} - p_{i,j,\underline{t}_{i,j}}$ where $\underline{t}_{i,j}$ denotes the month the respective quote-line starts being observed. Similarly, the log of an individual *frictionless* price quote is denoted by $p^{\star}_{i,j,t}$ and the *cumulated frictionless inflation* for that price quote is denoted by $\mathrm{Z}^{\star}_{i,j,t} \equiv p^{\star}_{i,j,t} - p^{\star}_{i,j,\underline{t}_{i,j}}$. The inaction region for that quote-line is denoted by ($L_{i,j}$,$U_{i,j}$) and the reset point by $c_{i,j}$. Define the *re-centered initial price gap* as

[9] An *item* is the most disaggregated level that the Office for National Statistics (ONS) uses for products differentiation for the purposes of price collection. It can be understood as a narrowly defined product category, but less specific than the barcode classification used in some scanner micro price datasets. An ONS *item* description typically does not specify a specific brand or, in many cases, a weight for the product to be collected. Examples of ONS items include "Light Bulb" (item id 430524) or "Bottle still water 500ml" (item id 212012).

$x_{i,j,0} \equiv p_{i,j,\underline{t}_{i,j}} - p^{\star}_{i,j,\underline{t}_{i,j}} - c_{i,j}$, the lower bound of the re-centered inaction region is defined as $\underline{x}_{i,j} \equiv L_{i.j} - c_{i,j} < 0$ and, analogously, the upper bound of the re-centered inaction region is defined as $\bar{x}_{i,j} \equiv U_{i,j} - c_{i,j} > 0$.

Transition equation for cumulated frictionless inflation. Using a discrete time approximation to the Brownian motion assumption in (2.1), the log of frictionless prices is assumed to evolve according to a random walk with drift. From that assumption it follows that cumulated frictionless inflation also evolves according to,

$$\mathrm{Z}^{\star}_{i,j,t} = \mu_{i,j} + \mathrm{Z}^{\star}_{i,j,t-1} + \varepsilon_{i,j,t} \tag{2.3}$$

where $\varepsilon_{i,j,t} \sim \mathcal{N}(0, \sigma_{\varepsilon,i,j})$ and independent and identically distributed across i, j and t.

Transition equation for the arrival of costless adjustment opportunities. Let $\ell_{i,j,t}$ be an indicator variable equal to one if prices for that respective quote-line in month t can be changed for free. The discrete time counterpart of the random menu cost assumption in the previous section is given by,

$$\ell_{i,j,t} = \mathbb{1}\{\nu_{i,j,t} \leqslant \lambda_{i,j}\} \tag{2.4}$$

where $\nu_{i,j,t} \sim \text{Uniform}(0,1)$ and independent and identically distributed across i, j and t.

Measurement equation for cumulated inflation. Assuming each quote-line is generated by a firm pursuing a discrete time approximation of a two-sided (S,s) policy then cumulated inflation evolves according to,

$$\mathrm{Z}_{i,j,t} = \mathrm{Z}_{i,j,t-1}\, d_{i,j,t} + (\mathrm{Z}^{\star}_{i,j,t} - x_{i,j,0})(1 - d_{i,j,t}) \tag{2.5}$$

where $\mathrm{Z}^{\star}_{i,j,t} - x_{i,j,0}$ is the value of cumulated inflation that closes the re-centered price gap at time t and $d_{i,j,t}$ is an indicator variable when its optimal *not* to adjust prices and evolves according to,

$$d_{i,j,t} = \mathbb{1}\{\mathrm{Z}_{i,j,t-1} - \mathrm{Z}^{\star}_{i,j,t} + x_{i,j,0} \in (\underline{x}_{i,j}, \bar{x}_{i,j})\}(1 - \ell_{i,j,t}) + \mathbb{1}\{\mathrm{Z}^{\star}_{i,j,t} = \mathrm{Z}_{i,j,t-1} + x_{i,j,0}\}\ell_{i,j,t} \tag{2.6}$$

where the first term on the right-hand-side is equal to one for periods where adjustment is costly and the price-gap is in the inaction region whilst the second term in the right-hand-side is equal to one for time periods where adjustment is costless but at the previous period price the current price gap is already at the reset point.

2.2.3. From a state-space representation to a measure of frictionless inflation

Assuming that each quote-line in micro price data is generated by equations (2.3), (2.4), (2.5) and (2.6) yields the state-space representation of dynamic (S,s) economies considered in Bandeira (2020). The statistical framework there proposed is used to: (*i*) estimate unknown parameters that enter in (2.3), (2.4), (2.5) and (2.6) and (*ii*) produce smoothed estimates of frictionless inflation for each individual quote-line.

2.2.3.1. Parameter estimation

For each individual quote-line the vector of unknown parameters is given by $\mathbf{\Theta}_{i,j} = \{\mathbf{\Gamma}_{i,j}, x_{i,j,0}\}$ where $\mathbf{\Gamma}_{i,j} \equiv \{\underline{x}_{i,j}, \bar{x}_{i,j}, \mu_{i,j}, \sigma_{\varepsilon,i,j}, \lambda_{i,j}\}$. To estimate these parameters the simulation-based two-stage procedure proposed in Bandeira (2020) is used.

Parameter heterogeneity. Over the last decade with increasing availability of disaggregated micro price data it has been documented that there is substantial heterogeneity in pricing practices across different sectors and even across narrowly defined products within a given sector. To account for this heterogeneity it has become common practice in the micro price literature to allow for rich parameter heterogeneity.[10] Following this tradition, in this paper the vector of parameters $\mathbf{\Gamma}_{i,j}$ is assumed to vary at the item level, that is, $\mathbf{\Gamma}_{i,j} = \mathbf{\Gamma}_j$, $\forall i = 1, \ldots, n_j$.[11] The re-centered initial price gaps, $x_{i,j,0}$'s, are assumed to vary at the quote-line level.

First-stage. As discussed in Bandeira (2020), the first-stage estimates of $\mathbf{\Gamma}_j$ are obtained using moments of the distribution of price changes excluding any observations before the first price adjustment for each quote-line. The moments used for estimation in the first-stage are the frequency of price adjustments as well as the 1st, 5th, 10th, ..., 90th, 95th and 99th percentiles of the price adjustment distribution. In the estimation these moment conditions are equally weighted and, given that $\mathbf{\Gamma}_j$ is assumed to be common across quote-lines of a given item, the moment conditions are computed from the distribution of price changes *pooled* across all quote-lines of a given item.

Second-stage. In the second-stage, the vector of parameters $\mathbf{\Gamma}_j$ is kept fixed at its first-stage estimated value whilst $x_{i,j,0}$'s are estimated using the time elapsed until the first price adjustment and the size of the first price adjustment. For any quote-line for which no price adjustment is observed, the initial re-centered price gap is imputed by taking the average over all $x_{i,j,0}$'s for that item.[12]

2.2.3.2. Smoothed estimates of frictionless inflation

Once the unknown parameters are estimated, the closed-form expression for the smoothed probability density function presented in Bandeira (2020) is used to obtain the *smoothed estimates* of $\mathrm{Z}^{\star}_{i,j,t}$ for each quote-line and, given those, estimates

[10] Heterogeneity in pricing moments had been documented by several authors. See, for example, Bunn and Ellis (2012) for the United Kingdom, Nakamura and Steinsson (2008, 2010) and Klenow and Malin (2010) for the United States and Álvarez *et al.* (2006). Several papers have also analysed the implications of this cross-sectional heterogeneity in pricing behaviors for the transmission of monetary shocks. See, for instance, Carvalho (2006), Nakamura and Steinsson (2010) and Gautier and Le Bihan (2018).

[11] As discussed in Bandeira (2020, section 3.3), this assumption is motivated by data constraints. At more disaggregated levels, there are typically not enough price adjustments in the ONS micro price data to meaningfully compute the pricing moments used for parameter estimation.

[12] Detailed algorithms used for first and second stage parameter estimation can be found in appendix D of Bandeira (2020).

of frictionless inflation for each quote-line are obtained from,

$$\hat{\pi}^{\star}_{i,j,t} = \mathbb{E}\left[\mathrm{Z}^{\star}_{i,j,t} \mid \{\mathrm{Z}_{i,j,t}\}^{\bar{t}_{i,j}}_{\underline{t}_{i,j}}, \widehat{\boldsymbol{\Theta}}_{i,j}\right] - \mathbb{E}\left[\mathrm{Z}^{\star}_{i,j,t-1} \mid \{\mathrm{Z}_{i,j,t}\}^{\bar{t}_{i,j}}_{\underline{t}_{i,j}}, \widehat{\boldsymbol{\Theta}}_{i,j}\right] \tag{2.7}$$

where $\widehat{\boldsymbol{\Theta}}_{i,j}$ denotes an estimated parameter value from the two-stage procedure described in the previous section. Quote-line level estimates from (2.7) can then be used to construct measures of frictionless inflation at any aggregation level.

2.3. Micro Price Data

This section is divided in three parts. The first part describes the primary micro price quotes and cleaning procedures used to obtain the final dataset used to produce measures of frictionless inflation. The second part analyses the item level reduced form parameter estimates obtained from the two-stage procedure described in the previous section. The third and last part assesses whether the proposed state-space representation of microeconomic pricing dynamics can match some of the key pricing moments observed in the data.

2.3.1. Price quotes underlying the UK Consumer Price Index

The estimates of frictionless inflation constructed in this paper are based on publicly available data on locally collected price quotes underlying the construction of the UK Consumer Price Index (CPI). In order to produce the CPI, the Office for National Statistics (ONS) collects on a monthly basis prices of different goods and services that are selected to be representative of general consumer expenditure across the whole of the UK.[13] There are two price collection methods: central and local. Central collection is used for goods and services for which the price

[13] Two major sources of information used to determine the selection of representative items are the Household Final Monetary Consumption Expenditure (HFMCE) and the ONS Living Cost and Food Survey (LCF). The CPI coverage excludes housing costs such as council tax, mortgage interest payments, house depreciation, buildings insurance, ground rent and other house purchase costs such as estate agents' and conveyancing fees.

is the same for all UK residents or the regional price variation can be collected with no field work, for example, via internet, telephone or e-mail enquiries. Since some of these price quotes could reveal the identity of the price setter, the ONS excludes those quotes from the publicly available dataset. For the remaining goods and services, which account for about 60% of the aggregate CPI by weight, price collectors on behalf on the ONS visit every month thousands of shops in over 140 locations spread over the UK to record over 100,000 prices on hand-held computers.[14] The following steps describe how the primary price quotes provided by the ONS are manipulated to obtain the final dataset that is used to produce estimates of frictionless inflation.

Non-uniquely identified price quotes. In order to combine different price quotes over time it is necessary that these quotes possess a unique identifier. In the ONS internal systems price quotes are uniquely identified by a concatenation of month, shop code, location and 6-digit item identifier. For confidentiality reasons, the ONS does not publish the *location* variable and the variable region is used instead as a proxy. However, in some instances it happens that price-quotes are collected in two different locations within the same region and have the same shop code and, in those cases, the two price trajectories could not be separated. Any price quotes that cannot be uniquely identified by a concatenation of month, shop code, region and 6-digit item identifier are excluded from the final sample.

Invalid price quotes. As described in ONS (2014, chapter 6), locally collected price quotes must pass a series of internal validation procedures to ensure that prices have been accurately recorded and indicator codes have been used sensibly and correctly. Any price quote that fails to pass these internal validation procedures is excluded from the final sample and it does not enter the final CPI

[14] Bunn and Ellis (2012) were the first to have access and document stylised facts about consumer prices in the UK. Due to its public availability there is a growing number of papers that use the micro price data provided by the ONS. A non-exhaustive list includes: Chu *et al.* (2018), Petrella, Santoro and Simonsen (2018), Carvalho and Kryvstov (2018), Blanco and Cravino (2019), Kryvstov and Vincent (2019) and Hobijn, Nechio and Shapiro (2019). Since this is not the first paper to use this dataset, this section focuses on the most important aspects of the data cleaning used to obtain the final dataset used to produce a measure of frictionless inflation.

calculation.

Product substitutions. For each outlet in which prices are collected, price collectors start by choosing among all products matching the specification of each item to be priced one product that is representative of what people buy in the area. Once a product is chosen, the price collector returns to the outlet every month to collect prices of that *same* product. However, in some cases the price collector might be forced to change the product being priced either because it becomes unavailable or because the producer changed its physical characteristics such as weight or size. Whenever those forced substitutions occur the price collector must flag the respective price quote accordingly. To deal with product substitutions, whenever a substitution flag is present a break in the original quote-line is generated.

Product sales. The menu cost model with random costless adjustment opportunities underlying the state-space representation in (2.3), (2.4), (2.5) and (2.6) does not include temporary sales, hence, that state-space representation should be interpreted as a data generating process for *regular* prices. However, the prices quotes collected by the price collectors are the product shelf-prices which implies they reflect any temporary sales at the time of collection. In order to have a good match between the underlying theory and the data those sales should be removed before constructing a measure of frictionless inflation. If the collected price is a sale price, the price collectors must flag it with the respective sales indicator and those sales flags are used to obtain quote lines of regular prices. First, collected prices that are not on sale are immediately considered as the regular prices for the period. Second, if the collected price is on sale the last regular price observed is used as the regular price for that period. If the last regular price change is not available, the first regular price observed *after* the current observation is used instead.[15]

[15] If no regular prices are observed before or after the sales observation it is because the respective quote-line is composed exclusively of sales prices and in those cases the entire quote-line is excluded from the sample.

Quote-line gaps. After following the above described steps it is possible that the resulting quote-line contains gaps, for example, because some price quote in the middle of the quote-line was not validated internally by the ONS. Since to produce estimates of the frictionless inflation requires a quote-line of contiguous observations of regular prices, whenever the resulting quote-line contains gaps it is split into separate quote-lines of contiguous observations. Moreover, any resulting singleton quote-line is excluded from the final sample.

An example. To illustrate how to go from the primary price quotes provided by the ONS to the quote-lines of regular prices used to produces estimates of frictionless inflation, fictitious price quotes of an item collected in a particular outlet are considered in figure 2.A.2. Although price quotes are available for a 4 year period, some of them did not pass the ONS internal checks and, because of that, did not contribute to the CPI in the respective periods. Those price quotes are excluded from the final sample. Moreover, in one of the periods the price collector was forced to substitute the product for which prices were being recorded for another brand and/or variety that also matches the specification of the item to be priced. Since from that period onwards it is effectively a different product that is being priced a quote break is created. Finally, during the whole period the price collector also indicated two episodes of temporary sales and in those periods regular prices are imputed for sales prices following the procedure above described. More precisely, for the first sales spell the last regular price observed is imputed in place of the sales prices whereas for the second sales spell the next regular price is imputed since the last regular price is not observed. The outcome of this process are four quote lines of contiguous regular price observations which include a total of six regular price changes and three regular price changes when excluding the first. Estimates for the frictionless cumulated inflation are obtained for each quote line separately.

Items excluded from the final sample. Some research items available in the primary price quote files but did not enter actual CPI calculations are excluded from the final sample. Moreover, since the estimation of common parameters is conducted at the item level and it is based on the moments of the distribution of regular price changes excluding the first price adjustment in any quote-line, any item that has less than 100 such price adjustments is excluded from the sample.

Final Sample. The final sample used to produce frictionless inflation measures spans the period from 1996m1 to 2018m1 and comprises over 23 million price quote observations, over 2.2 million quote-lines and 979 unique items.

2.3.2. Reduced form parameter estimates

As expected there is substantial heterogeneity in the estimated reduced form parameters at the item level (figure 2.A.3 and table 2.B.1). This section investigates whether this heterogeneity is in line with what one would expect from their theoretically implied relationship and key pricing moments of the distribution of price changes.

Inaction region asymmetry. A dimension of interest is the extent to which price setters pricing policies are characterised by symmetric inaction region boundaries. In terms of estimated parameters, a simple way of measuring the asymmetry of the inaction region is by taking the sum of the two boundaries, that is, $\mathcal{A}_j = \hat{\bar{x}}_j + \hat{\underline{x}}_j$. In the data, that measure ranges from -171% to 206% and 58% of the items display negative asymmetry.[16] A common explanation for asymmetric pricing policies is the presence of non-zero trend inflation, since in a pure menu cost model with positive trend inflation the optimal policy is characterised by price increases that are larger than price decreases (Ball and Mankiw, 1994). This explanation is qualitatively in line with the heterogeneity observed in the cross section of items since the correlation between $\mathcal{A}_j$ and the estimated drift of the frictionless inflation process of -0.41 (figure 2.A.4, top-left panel). Nonetheless, that explanation alone is quantitatively insufficient to rationalise all the heterogeneity in the asymmetry measure across different items. Other poten-

[16] The measure of asymmetry is computed only for items that have at least one price changed triggered by crossing the upper bound of the inaction region and one price change triggered by crossing the lower bound of the inaction region (total of 424 items).

tial explanations for asymmetric adjustment policies include differences in size of the menu cost to increase or decrease prices and asymmetries in the profit loss function.[17]

Trends in frictionless inflation. The estimated frictionless inflation trend ($\hat{\mu}_j$) across items is typically in the range of -1.4% to 1.2% per month. The median estimated frictionless inflation trend across products is 0.2% per month which is consistent with the average monthly CPI inflation for the period analysed of 0.17%. In the data, the correlation across items between estimated frictionless inflation trend and the average size of price changes is 0.63 (figure 2.A.4, top right-panel). Moreover, approximately one-fourth of the items have negative estimated trends. These negative estimated trends can be explained by the fact that, despite the positive aggregate inflation, some individual items have become less expensive over time.[18]

Standard deviation of idiosyncratic shocks. Differences in dispersion of price changes are mostly accounted by differences in the volatility of idiosyncratic shocks. More precisely, the correlation between the standard deviation of price changes and the estimated standard deviation of idiosyncratic shocks ($\hat{\sigma}_{\varepsilon,j}$) across items is 0.86 (figure 2.A.4, bottom-left panel). The median value of the estimated standard deviation of the idiosyncratic shocks for the UK micro price data is 7.2% which is in line with the figures reported in Gautier and Le Bihan (2018). Estimating a random menu cost model across 227 products underlying the French CPI, Gautier and Le Bihan (2018) find a median across products for the unconditional productivity standard deviation to be either 5% (in a model without strategic complementarities) or 9% (in a model with strategic complementarities).

[17] It is relatively common to assume a zero trend inflation and profit flow function that is quadratic in the price gap. These two assumptions give rise to an optimal symmetric Ss policy that can be solved in closed form (Dixit, 1991). Despite their analytical convenience, it is noticeable that these assumptions would be at odds with the pricing behaviour observed for most items in the UK micro price data.

[18] Most examples are from items classified as: recreation and culture (e.g personal CD player, CD radio cassette, computer diskettes etc); furniture, household equipment and maintenance (e.g automatic washing machine, vacuum cleaners etc) and clothing and footwear (e.g. boy's casual short sleeve shirt).

Arrival of costless adjustment opportunities. The heterogeneity in frequency of price changes across items in the micro price data is mostly accounted by the estimated probability of a costless adjustment opportunity ($\hat{\lambda}_j$). Across all the items in the sample they correlate at 0.91 (figure 2.A.4, bottom-right panel). This high correlation resounds the fact that, in the UK data, most of the price changes are estimated to be due to the arrival of costless adjustment opportunities.

State-dependent versus time-dependent pricing. An interesting feature of the random menu cost model is that it allows for the occurrence of both time-dependent and state-dependent price adjustments. A way of measuring which of these two extremes can better account for the pricing dynamics in the micro data is to compute the ratio between the number of costless price changes over the total number of price changes.[19] For a given item, a value of this measure closer to a hundred percent indicates that most of the price changes are triggered by the arrival of costless adjustment opportunities and, hence, the pricing dynamics is closer to that implied by the Calvo model. Computing this ratio using all the price changes in the final sample yields a figure of 88%. For individual items this measure ranges from 7.5% to 100% with an average value of 87%. Across different COICOP divisions, this ratio ranges from 66% for restaurants and hotels to 94% in food and non-alcoholic beverages (figure 2.A.5). The large number of time-dependent price changes is broadly in line with previous findings in the literature. For instance, Nakamura and Steinsson (2010, footnote 25) using data underlying the US CPI find that roughly 75% of price changes occur in the low menu cost state. Gautier and Le Bihan (2018, tables 5 and 6) find that costless adjustment opportunities account for 80% of the price changes in micro price data underlying the French CPI. Moreover, Blanco (2017) using UK CPI micro price data finds that 93% of all the price changes are due to either free adjustment opportunities or fat-tailed idiosyncratic shocks.[20]

[19] More formally, for each item in the data this statistic is computed as $\mathcal{C}_j = \sum_i \sum_t \mathbb{1}\{\Delta p_{i,j,t} \in (-\hat{\bar{x}}_j, 0) \cup (0, -\hat{\underline{x}}_j)\} / \sum_i \sum_t \mathbb{1}\{\Delta p_{i,j,t} \neq 0\}$.

[20] More precisely, the ratio of free to total price adjustments is only 48% but adding the price changes after fat-tailed idiosyncratic shocks the figure increases to 93% (p.17). The combination

2.3.3. Model fit

As described in section 2.2, a cornerstone assumption underlying the construction the construction of a frictionless inflation measure is that the observed quote-lines are generated by the state-space representation in (2.3) to (2.6). Figures 2.A.6 to 2.A.11 investigate the extent to which the proposed representation of pricing dynamics can match some of key moments observed in micro price data. First, in terms of the moments directly targeted in the first-stage of parameter estimation, the proposed representation of pricing dynamics can in general match the percentiles of the distribution of non-zero price changes and the model fit tends to be better at the tails of the distribution (figures 2.A.6 and 2.A.7). Second, in terms of the non-targeted moments, the worse fit of the middle percentiles of the distribution of non-zero price changes does not affect the model's ability to match some of the moments (e.g. mean or variance of price changes) but it does imply that for a non-negligible proportion of items the model cannot account for the values of robust skewness and robust kurtosis observed in the micro price data (figure 2.A.8). The inability of the single product version of the random menu cost model to match the kurtosis observed in the micro price data has been pointed before by Alvarez, Le Bihan and Lippi (2016). It is interesting to notice that for the micro price data underlying the UK CPI that is also the case for some individual items but the mismatch almost vanishes when considering the distribution for the price changes across all items (figure 2.A.9). Third, across all individual items considered the actual distribution of regular price changes can take a wide variety of shapes and inevitably the proposed state-space representation will fit some items better than others. Figures 2.A.10 and 2.A.11 illustrate

of the two cases is a measure of how close the slope of the Phillips curve in the model in Blanco (2017) is from a Calvo model and that measure is more comparable to the ratio of free price changes to total price changes reported above. Intuitively, in the state-space representation of the random menu cost model in (2.3) to (2.6), the boundaries of the inaction region are allowed to vary freely to account for the extremes in the distribution whereas the arrival of the costless adjustment opportunities accounts for the "middle" of the distribution. In Blanco (2017) also the middle of the distribution is accounted by the arrival of costless adjustment opportunities but the tails are accounted by fat-tailed idiosyncratic shocks. Despite these methodological differences, it is interesting that we find similar figures in terms of how close the model is from a pure Calvo model.

the variety of shapes that the distribution of price changes can take across items and the model fit for the items with worst and best model fits, respectively.

2.4. Menu costs at the microeconomic level and aggregate inflation dynamics

This section presents a measure of aggregate frictionless inflation for the UK and compares it with its regular price counterpart to investigate the importance of menu costs at the micro level for the aggregate inflation dynamics.

2.4.1. A measure of aggregate frictionless inflation

Estimates of aggregate frictionless inflation are produced by aggregating the quote-line level estimates of frictionless inflation from (2.7). The aggregation procedure is described below and it is based on the methodology used by the ONS to produce the official CPI.

From price quotes to elementary aggregates. The lowest level of aggregation at which price quotes are aggregated into a price index is at the *stratum* level. Each item in the data can be stratified in four different ways, namely, by region, by shop type, by region and shop type or not stratified.[21] For each stratum, individual price quotes are aggregated to produce an elementary aggregate index. As described in ONS (2014, chapter 2), the method primarily used to produce elementary aggregates in the CPI is the geometric mean (also known as the *Jevons* formula). More precisely, the elementary aggregate for stratum s of item j in month t with base period $\underline{t}$ is given by,

$$\mathcal{I}_{s,j,t|\underline{t}} = \left(\prod_{i=1}^{\mathrm{N}_{s,j}} \frac{\mathrm{P}_{i,j,t}}{\mathrm{P}_{i,j,\underline{t}}} \right)^{\frac{1}{\mathrm{N}_{s,j}}} \tag{2.8}$$

[21] There are 12 regions in total (London, Southeast, Southwest, East Anglia, East Midlands, Yorks and Humber, Northwest, North, Wales, Scotland and Northern Ireland) and two shop types (multiple if the shop has 10 or more outlets or independent if it has less than 10 outlets).

where $\mathrm{P}_{i,j,t}$ and $\mathrm{P}_{i,j,\underline{t}}$ are the nominal prices of the same product collected in a particular outlet in period t and in the base period $\underline{t}$ and $\mathrm{N}_{s,j}$ is the number of price quotes in stratum s of item j after using the central shop weights as a replication factors. Using the previous month as the base period, the above expression can be equivalently written as,

$$\mathcal{I}_{s,j,t|t-1} = \exp\left\{\frac{1}{\mathrm{N}_{s,j}}\sum_{i=1}^{\mathrm{N}_{s,j}} \pi_{i,j,t}\right\} \tag{2.9}$$

where $\pi_{i,j,t} \equiv \log(\mathrm{P}_{i,j,t}) - \log(\mathrm{P}_{i,j,t-1})$. To obtain *regular price elementary aggregates* expression (2.9) is used with $\pi_{i,j,t}$ replaced by $\pi^r_{i,j,t} \equiv \log(\mathrm{P}^r_{i,j,t}) - \log(\mathrm{P}^r_{i,j,t-1})$ where $\mathrm{P}^r_{i,j,t}$ denotes the regular nominal prices.[22] Similarly, to obtain *frictionless elementary aggregates* expression (2.9) is used with $\pi_{i,j,t}$ replaced by $\hat{\pi}_{i,j,t}$ obtained from (2.7).[23]

From elementary aggregates to higher level indices. Indices for higher levels of aggregation are weighted averages of the elementary aggregate indices. First, a weighted average of elementary aggregates indices with stratum weights yields item level indices. Second, item level indices and the respective item weights are combined to produces Classification of Individual Consumption by Purpose (COICOP) indices at the class, group and division levels and the aggregate CPI. The exact same aggregation procedure is done twice, one starting from the regular price elementary aggregates and other starting from the frictionless elementary aggregates.[24]

[22] That is, the prices obtained after the sales filter described in section 2.3.1 is applied to each quote-line.

[23] A key difference between the elementary aggregates produced following (2.9) and the elementary aggregates underlying the published CPI is that the later uses the previous month of January as the base period. Because of the breaks in quote-lines created to deal with product substitutions and quote-line gaps, using the previous January as the base period would require to either: (*i*) drop any observations that do not have information for $\mathrm{P}_{i,j,t}$ in the previous month of January which account for almost half of the total sample or (*ii*) impute values for inflation for all the periods between the previous month of January and the period where the actual quote-line starts being observed in the data. Changing base period and chain-link the elementary aggregate indices every month is preferred over either of these two options.

[24] The weights attributed to each item to produce indices at the class level and above are always relative to the total weight of that category in the final sample used in this paper. For example, suppose a given class accounts for 10% of the overall CPI and it is composed by items A, B and C with weights 4%, 4% and 2%, respectively. If item C does not appear in the final

2.4.2. Aggregate frictionless inflation in the UK: 1997 - 2018

The measure of frictionless inflation for the UK against its regular price counterpart is presented in figure 2.A.12. In a nutshell, the conclusion to be drawn from that figure is that menu costs matter for aggregate inflation dynamics but their importance is decreasing over time period analysed.

Menu costs matter for aggregate inflation dynamics... which comes from the fact that the two lines do not perfectly overlap over the period analysed. The correlation between the two-series from 1997 to 2018 is equal to 0.83 and the difference between regular price inflation and its frictionless counterpart can range from -0.79 and 2.29 percentage points. On average over the whole period analysed, frictionless inflation is 0.55 percentage points lower than regular price inflation and the difference is mostly driven by observations prior to 2004. Moreover, from figure 2.A.12 it is also the case that the time-series for frictionless inflation is smoother that its regular price counterpart. More precisely, the historical standard deviation of the time-series for frictionless inflation is 20% lower than the standard deviation for the time-series of regular price inflation (0.91 against 1.13), also, the time-series of frictionless inflation exhibits higher persistence as it is more autocorrelated than regular price inflation at any horizon up to three years.

...but their importance is decreasing over time. From figure 2.A.12 it is also clear that the time-series of frictionless inflation is in general closer to its regular price counterpart after mid-2000s. In particular, considering only the periods prior to 2004 the correlation between the two time-series is 0.52 against 0.94 in the post-2004 period. Moreover, the average difference between regular price inflation and frictionless inflation decreases from 1.14 percentage points prior to 2004 to 0.27 percentage points post 2004.

sample because it is centrally collected, then in the price index computation for that class items A and B will enter with a weight of 50% each. This “composition effect” affects the comparison of inflation figures here reported with the official all items CPI inflation, but does not affect the comparison between regular price inflation and frictionless inflation discussed in this section.

Potential explanations. Interpreted through the lens of the random menu cost model that underlies the construction of the frictionless inflation measure, this change in the correlation of the two-time series can be explained by a narrowing of the boundaries of the inaction region or by an increase in the rate of costless adjustment opportunities or by a combination of the two.[25] There are two events that are consistent with a narrowing of the inaction boundaries, namely, the increase in product competition from import penetration from China around the early 2000s and the change in inflation target in December 2003 from 2.5% of the Retail Price Index excluding mortgage payments (RPIX) to 2% of the CPI. The increase in product competition from imports would make deviations from the frictionless prices more costly in terms of profit flows and, hence, a smaller deviation would be enough to trigger the payment of the menu cost resulting in narrower bands of the inaction region. The decrease in the inflation target could also rationalise a narrowing of the boundaries of the inaction region if interpreted as decreasing the uncertainty regarding aggregate inflation rate. It is more difficult to rationalise an increase in the rate of arrival of costless adjustment opportunities in terms of specific events, but such increase is consistent with an increase in the frequency of price changes that is observed for part of the period between 2005 and 2012 (Petrella, Santoro and Simonsen, 2018, figure 1).

2.5. Frictionless Inflation and the transmission of monetary policy shocks

At the heart of the new Keynesian paradigm lies the idea that prices react slowly in response to changes in economic conditions due to the existence of pricing frictions. More precisely, the basic new Keynesian model presented in Galí (2008, chapter 3) predicts that inflation should react *differently* than its frictionless counterpart in response to a monetary policy shock. This section uses the constructed

[25] On the one hand, the narrowing of the inaction region decreases the range of values that $\hat{\pi}^{\star}_{i,j,t}$ can take since, by construction, $\hat{\pi}^{\star}_{i,j,t} \in (-\bar{x}_{i,j}, -\underline{x}_{i,j})$. On the other hand, an increase in the rate of arrival of costless adjustment opportunities would decrease the number of periods where $\pi^{r}_{i,j,t} = 0$ and $\hat{\pi}^{\star}_{i,j,t} \neq 0$. In the limit, if $\lambda_{i,j} = 1$, then $\pi^{r}_{i,j,t} = \hat{\pi}^{\star}_{i,j,t}, \forall t > \underline{t}_{i,j}$.

time-series of frictionless inflation to empirically test this prediction.

2.5.1. Inflation and frictionless inflation in the basic new Keynesian model

The following proposition presents the implied relationship between the responses of inflation and frictionless inflation to a monetary policy shock in the basic new Keynesian model,

***Proposition* 2.1** *Consider the basic new Keynesian model under an interest rate rule as presented in Galí (2008, chapter 3). Let $\pi_t^\star$ denote frictionless inflation rate obtained under flexible prices, π_t denote the inflation rate under Calvo pricing and ε_t^v denotes a monetary policy shock. Then it holds that,*

$$\frac{\partial \pi_t^\star}{\partial \varepsilon_t^v} < \frac{\partial \pi_t}{\partial \varepsilon_t^v} < 0 \tag{2.10}$$

and for any $h > 0$ it holds that,

$$\frac{\partial \pi_{t+h}}{\partial \varepsilon_t^v} < \frac{\pi_{t+h}^\star}{\partial \varepsilon_t^v} \tag{2.11}$$

Proof. See appendix 2.C. □

Intuition. Equations (2.10) and (2.11) state that, in response to a monetary policy shock, frictionless inflation reacts *more* on impact than actual inflation and reacts *less* in subsequent periods. Figure 2.A.13 illustrates this result by plotting the model implied impulse responses of inflation and frictionless inflation to a monetary policy shock.[26] To understand the intuition behind the result in proposition 2.1 consider the case of a monetary tightening that reduces aggregate demand. On the one hand, in a world without pricing frictions *all* the firms in the economy reduce their prices simultaneously on impact and slowly revert as

[26] Notice that, although the impulse responses in figure 2.A.13 are obtained under the baseline calibration described in Galí (2008, p.52), the result in proposition 2.1 holds irrespective of parameter values.

the initial shock vanishes. On the other hand, in a world with pricing frictions only a *fraction* of the firms reduces their prices on impact whilst the remaining firms keep their prices unchanged, which partially mutes response of aggregate inflation on impact. In the subsequent periods, despite the vanishing effect of the monetary shock, some firms that are allowed to re-price their goods choose to *cut* their prices since they were stuck with prices that were last set prior to the monetary shock. These firms that choose to cut their prices in the subsequent periods makes the response of inflation smaller than frictionless inflation in the periods after the shock.

2.5.2. An empirical test

To empirically test the predictions in (2.10) and (2.11) the impulse responses of regular price inflation and frictionless inflation are estimated via local projections (Jordà, 2005). More precisely, for every horizon $h \in \{0, 1, \dots, 36\}$ the following regressions are estimated:

$$\pi_{t+h} = \alpha_h + \beta_h \widehat{\Delta i_t} + \gamma_h \mathbf{X}_t + \epsilon_t \tag{2.12}$$

$$\pi^\star_{t+h} = \alpha^\star_h + \beta^\star_h \widehat{\Delta i_t} + \gamma^\star_h \mathbf{X}^\star_t + \epsilon^\star_t \tag{2.13}$$

where π_t and $\pi^\star_t$ are year-over-year regular price inflation and frictionless inflation (*i.e.* the solid black and blue-starred lines in figure 2.A.12, respectively), $\widehat{\Delta i_t}$ denotes changes in the Bank of England official rate instrumented by the high-frequency measure monetary policy surprises from Cesa-Bianchi, Thwaites and Vicondoa (2020), $\mathbf{X}_t$ and $\mathbf{X}^\star_t$ are vectors of control variables that include four lags of regular price inflation and four lags frictionless inflation.[27] In terms of the

[27] For analytical convenience the basic new Keynesian model in Galí (2008) used to derive the results in proposition 2.1 uses Calvo pricing whereas the empirical measure of $\pi^\star_t$ used on the left-hand-side of (2.13) is theoretically grounded on the menu cost model subject to random costless adjustment probabilities discussed in section 2.2 which corresponds to the Calvo only in a special case where all the price adjustments are triggered by the arrival of costless adjustment opportunities. This fact can induces a mismatch between the $\pi^\star_t$ that the results in proposition 2.1 refer to and the $\pi^\star_t$ used to estimate the impulse responses in (2.13). Notice, however, that as discussed in section 2.3 about 90% of the price adjustments observed in the UK micro price

coefficients in (2.12) and (2.13), the hypotheses to be tested from proposition 2.1 are: (i) $\beta_0^\star < \beta_0 < 0$ and (ii) $\beta_h^\star > \beta_h$, $\forall h > 0$.

Results. The impulse responses for regular price inflation and frictionless inflation to a 1 percent unexpected increase in nominal interest rates are depicted in figure 2.A.14. The first hypothesis is not supported given that frictionless inflation reacts positively on impact whereas regular inflation is virtually unchanged and neither of the two responses is statistically significant (see figure 2.A.15). The second hypothesis is only *weakly* supported at horizons greater than 21 months when regular price inflation starts declining.[28]

To conclude it is important to notice that this is *not* a test of whether monetary policy has effects inflation. In fast, from the estimated impulse response of regular price inflation in figure 2.A.14 it is indeed the case that regular price inflation declines in response to a monetary tightening albeit with a long delay (*i.e* there is a substantial prize puzzle). Instead, the test of proposition 2.1 should be interpreted instead as a test of the monetary transmission mechanism as implied by the basic new Keynesian model presented in Galí (2008, chapter 3). In that model, frictionless inflation should react quickly to a monetary policy shock and this is at odds with the responses in figure 2.A.14 where frictionless inflation does not significantly react to a monetary policy shock at any horizon considered.

2.6. Frictionless inflation and movements in the output gap

In the new Keynesian framework a key variable of interest is the *output gap* formally defined as "the log deviation of output from its flexible part counterpart" (Galí, 2008, p.48). This section uncovers a relationship between frictionless infla-

data are trigerred by the arrival of costless adjustment opportunities making this mismatch less of a concern.

[28] Local projections in (2.12) and (2.13) were also estimated using: (i) the monetary policy shocks from Cesa-Bianchi, Thwaites and Vicondoa (2020) directly instead of as an instrument for Δi_t and (ii) under a variety of specifications including more or less lags of inflation and frictionless inflation in the vector of controls. In all those specifications, the same qualitative results were obtained.

tion and the output gap implied by the basic new Keynesian model presented in Galí (2008, chapter 3) and uses the constructed time-series of frictionless inflation to empirically test this relationship.

2.6.1. The inflation wedge and output gap in the basic new Keynesian model

The following proposition presents a relationship between the inflation wedge, defined as the difference between inflation and its frictionless counterpart, and the movements in the output gap,

***Proposition* 2.2** *Consider the basic new Keynesian model as presented in Galí (2008, chapter 3). Let $\pi_t^\star$ denote frictionless inflation rate obtained under flexible prices, π_t denote the inflation rate under Calvo pricing and $\tilde{y}_t$ denotes the output gap. Then it holds that,*

$$\pi_t - \pi_t^\star = -\left(\sigma + \frac{\varphi + \alpha}{1 - \alpha}\right) \Delta \tilde{y}_t \tag{2.14}$$

where $\alpha \in (0, 1)$ denotes the returns to scale in the firm's production function, φ is the inverse Frisch elasticity of labor supply and σ is the elasticity of intertemporal substitution.

Proof. See appendix 2.C. □

Intuition. The result in (2.14) simply states that whenever prices increase *more* than their frictionless counterparts the output falls by *less* than it frictionless counterpart, which can be rationalised from the aggregation production decisions of firms that face a negatively slopped demand curve.

2.6.2. An empirical test

Using the constructed time-series of frictionless inflation, the prediction in (2.14) is empirically assessed by a bivariate regression of the form,

$$\pi_t - \pi_t^\star = \alpha + \beta \Delta \tilde{y}_t + w_t \tag{2.15}$$

where π_t and $\pi_t^\star$ are the year-over-year regular price inflation and frictionless inflation (*i.e.* the solid black and blue-starred lines in figure 2.A.12, respectively) and $\Delta \tilde{y}_t$ is a measure of changes in the output gap. A total of six different measures of the output gap are considered. First, to measure *output* at the monthly frequency it is used either the monthly GDP index or industrial production index both produced by the ONS. Second, the *gap* is measured by taking the deviations of the log of these two indices from three different types of trends: a trend extracted from the Hodrick and Prescott (1997) filter, a cubic trend and a trend extracted from the Hamilton (2018) filter. In terms of coefficients in (2.15) the hypothesis to be tested, implied by the relationship in (2.14), is $\beta = -[\sigma + (\varphi + \alpha)/(1 - \alpha)] < 0$.[29]

Results. The slope estimates from (2.15) for different measures of the output gap are presented in table 2.B.2. First, notice that all the estimated coefficients are negative which is qualitatively in line with the relationship in (2.14). Moreover, in three out of the six specifications considered the slope is also statistically significant at a 10% significance level. In quantitative terms, the assessment of (2.14) depends on the parameter values. Using $\alpha = 1/3$, $\varphi = 1$ and $\sigma = 1$, which corresponds to the the baseline calibration in Galí (2008, p.52), yields an approximate value for β of -3. This value is close to the point estimates obtained by measuring the output gap as a deviation of monthly GDP from an Hamilton (2018) trend and as a deviation of industrial production from an Hodrick and Prescott (1997) trend.

[29] From (2.14) it follows that $\frac{Cov(\pi_t - \pi_t^\star, \Delta \tilde{y}_t)}{Var(\Delta \tilde{y}_t)} = -[\sigma + (\varphi + \alpha)/(1 - \alpha)]$.

Using frictionless inflation to *measure* movements in the output gap. Despite the qualitative support (2.14) evidenced by the negative coefficients of all the coefficients in table 2.B.2, quantitatively these coefficients take on a wide range of values depending on how the output gap is measured. Indeed, the $\tilde{y}_t$ in (2.14) corresponds to the new Keynesian output gap whereas all measures of $\tilde{y}_t$ used to estimate (2.15) are *proxies* that do not necessarily correspond to the new Keynesian notion of output gap. This suggests an alternative use for the measures of frictionless inflation computed in this paper: to use them to quantify movements in the output gap. More precisely, given a specification for product demand that relates changes in quantities sold to changes in prices, one could use the estimates of frictionless inflation from (2.7) to quantify changes in the output gap.[30] Such a measure would have two advantages over existing measures of the output gap. First, it would be, by construction, consistent with the new Keynesian notion of the output gap. Second, because the approach described in section 2.2 makes possible the estimation of frictionless inflation at the quote-line level, this approach could be used to produce measures of changes in output gap at virtually any aggregation level including levels for which data on output is typically not available (*e.g.* item level).

2.7. Frictionless inflation and headline inflation forecasts

Over the last three decades inflation targeting regimes have become increasingly popular among Central Banks worlwide. In a nutshell, Central Banks operating under an inflation targeting regime set short term nominal interest rate with the aim of keeping inflation as close as possible to a pre-specified inflation target. In this context, inflation forecasts are of crucial importance for the conduct of monetary policy. This section investigates whether the constructed time-series of

[30] More concretely, given a relationship of the form $\Delta y_{i,j,t} = f(\pi_{i,j,t}, \theta_{i,j})$ where $\theta_{i,j,t}$ is a vector of time-invariant quote-line specific parameters, the changes in the output gap can be estimated from $\Delta\tilde{y}_{i,j,t} = \Delta y_{i,j,t} - \Delta y^{\star}_{i,j,t} = f(\pi_{i,j,t}, \theta_{i,j}) - f(\hat{\pi}^{\star}_{i,j,t}, \theta_{i,j})$ where $\hat{\pi}^{\star}_{i,j,t}$ is obtained from (2.7).

frictionless inflation can be used to improve inflation forecasts.

2.7.1. Can frictionless inflation improve headline inflation forecasts?

To understand the rationale underlying the forecasting exercise proposed in this section, consider an hypothetical quote-line for which prices remains unchanged for 9 months and in the tenth month they change by 30%. The frictionless counterfactual for that quote-line is a quote-line for which the price changes every period by small positive amounts to reflect the changes in economic conditions, say these changes are on average 3% every month. Starting from this disaggregated perspective, the question is whether an aggregate measure based on those small changes contains useful information to forecast the headline inflation rate that is calculated based on infrequent and lumpy price adjustments.

A forecasting exercise. In order to assess the relevance of frictionless inflation to forecast published inflation, the forecasting exercise of Blinder and Reis (2005) is revisited. In particular, linear regressions of the form are estimated,

$$\pi_{t,t+h} = \alpha + \mathbf{X}\beta + \varepsilon_t \tag{2.16}$$

where $\pi_{t,t+h}$ is the inflation between months t and $t+h$ calculated based on the published headline CPI index, $\mathbf{X}$ is a vector containing published inflation over the previous 12 months ($\pi_{t-12,t}$) or frictionless inflation over the previous 12 months ($\pi^{\star}_{t-12,t}$) or both. Three different measures are used to assess forecasting ability of past inflation against past frictionless inflation. First, in-sample forecasting ability is evaluated by estimating the three different specifications of the above regression and computing the standard error of the regressions. Second, the out-of-sample forecasting ability is assessed by estimating the above regressions using data until January 2008 and computing the root mean squared errors from forecasting inflation from then until the end of the sample. Third, the version of the above regression with both $\pi_{t-12,t}$ and $\pi^{\star}_{t-12,t}$ and the significance of the coefficients is compared.

Results. The results for the forecasting exercise are presented in table 2.B.3. First, in terms of in-sample forecasting ability, for all forecasting horizons considered the specification with frictionless inflation only has a better in-sample fit than the specification only with headline inflation. Moreover, the gains of including both frictionless and headline inflation are small at 6 and 12 month horizons and inexistent for 24 and 36 month horizons. Second, the message in terms of out-of-sample forecasting is similar. Again the specification only with frictionless inflation has better out-of-sample forecasting ability when compared with the specification with the exception that there are gains from including both the frictionless and the headline inflation. These gains are larger for longer horizons specially 24 and 36 months. Finally, in terms of in-sample coefficient significance only the coefficient on frictionless inflation is significant at any of the horizons considered and more so for longer horizons.

To conclude the evidence from the forecasting exercise presented in this section is *suggestive* that the constructed time-series of frictionless inflation contains useful information to forecast the headline inflation. It remains to be investigated whether frictionless inflation can also improve inflation forecasts in more sophisticated models featuring nonlinearities and richer set of predictor variables.

2.8. Concluding remarks and future research

The existence of menu costs of price adjustment is one of the leading explanations for the delayed response of prices to changes in economic conditions. This paper introduced a measure of frictionless inflation that estimates the inflation in a counterfactual world where menu costs did not exist. Exploring this measure this paper arrived at four main conclusions. First, menu costs at the microeconomic level matter for aggregate inflation dynamics but their importance has decreased over time. Second, the responses of frictionless inflation to a monetary policy shock are at odds with the monetary transmission mechanism implied by the basic new Keynesian model. Third, the inflation wedge is negatively correlated with changes in the output gap in line with the basic new Keynesian model. Fourth, the constructed time-series frictionless inflation contains useful information to forecast headline inflation.

I conclude by highlighting some promising avenues for future research motivated by the present paper. First, the disaggregated measures of frictionless inflation can be combined with a specification for individual product to measure changes in the output gap at disaggregation levels where for which data on output is typically not available. Second, as discussed in section 2.7, it remains to be investigated whether frictionless inflation can improve forecasting performance in more sophisticated models including nonlinearities and richer set of predictors. Third, the framework here used to construct frictionless inflation could also be used to produce estimates of the cross-sectional distribution of price gaps at time period or any aggregation level. This distribution could in turn be used to quantify the welfare costs of pricing frictions or as a key state variable to understand sign and state-dependent effects of monetary policy.

2.9. References

ALVAREZ, F., BERAJA, M., GONZALEZ-ROZADA, M. and 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.

—, LE BIHAN, H. and LIPPI, F. (2016). The real effects of monetary shocks in sticky price models: A sufficient statistic approach. *American Economic Review*, **106** (10), 2817 – 2851.

—, LIPPI, F. and PACIELLO, L. (2018). Monetary shocks in models with observation and menu costs. *Journal of the European Economic Association*, **16** (2), 353–382.

ÁLVAREZ, L. J., DHYNE, E., HOEBERICHTS, M., KWAPIL, C., BIHAN, H. L., L´UNNEMANN, P., MARTINS, F., SABBATINI, R., STAHL, H., VERMEULEN, P. and 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.

ANDERSON, E., JAIMOVICH, N. and SIMESTER, D. (2015). Price stickiness: Empirical evidence of the menu cost channel. *The Review of Economics and Statistics*, **97** (4), 813–826.

BALL, L. and MANKIW, N. G. (1994). Asymmetric price adjustment and economic fluctuations. *The Economic Journal*, **104**, 247–261.

BANDEIRA, M. (2020). State-Space Modeling of Dynamic (S,s) Economies.

BARRO, R. J. (1972). A theory of monopolistic price adjustment. *Review of Economic Studies*, **39** (1), 17–26.

BERGER, D. and VAVRA, J. (2018). Dynamics of the u.s. price distribution. *European Economic Review*, **103**, 60 – 82.

BILS, M., KLENOW, P. J. and MALIN, B. A. (2012). Reset price inflation and the impact of monetary policy shocks. *American Economic Review*, **102** (6), 2798–2825.

BLANCO, A. (2017). Optimal inflation target in an economy with menu costs and a zero lower bound.

— and CRAVINO, J. (2019). Price rigidities and the relative ppp, working Paper.

BLINDER, A. S., CANETTI, E. D., LEBOW, D. E. and RUDD, J. D. (1998). *Asking about prices: A new approach to understanding price stickiness*. Russell Sage Foundation, 1st edn.

— and Reis, R. (2005). Understanding the greenspan standard. In *Proceedings - Economic Policy Symposium - Jackson Hole*, pp. 11 – 96.

Bonomo, M., Correa, A. and Medeiros, M. C. (2013). Estimating strategic complementarity in a state-dependent pricing model.

Bunn, P. and Ellis, C. (2012). Examining the behaviour of individual uk consumer prices. *The Economic Journal*, **122** (558), F35 – F55.

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

Carvalho, C. (2006). Heterogeneity in price stickiness and the real effects of monetary shocks. *Frontiers of Macreconomics*, **2** (1).

— and Kryvstov, O. (2018). Price selection, working Paper.

Cesa-Bianchi, A., Thwaites, G. and Vicondoa, A. (2020). Monetary policy transmission in the united kingdom: A high frequency identification approach. *European Economic Review*, **123**, 103–375, working Paper.

Chu, B. B., Huynh, K., Jacho-Chávez, D. and Kryvstov, O. (2018). On the evolution of the united kingdom price distributions. *The Annals of Applied Statistics*, **12** (4), 2618 – 2646.

Dixit, A. K. (1991). Analytical approximations in models of hysteresis. *Review of Economic Studies*, **58** (1), 141–151.

Dutta, S., Bergen, M., Levy, D. and Venable, R. (1999). Menu costs, posted prices, and multiproduct retailers. *Journal of Money, Credit and Banking*, **31** (4), 683 – 703.

Galí, J. (2008). *Monetary Policy, Inflation and the Business Cycle: An Introduction to the New Keynesian Framework*. Princeton University Press, 1st edn.

Gautier, E. and Le Bihan, H. (2018). Shocks vs menu costs: Patterns of price rigidity in an estimated multi-sector menu-cost model, banque de France Working Paper No. 682.

Golosov, M. and Lucas, R. E. (2007). Menu costs and phillips curves. *Journal of Political Economy*, **115** (2), 171–199.

Gorodnicheko, Y. and Weber, M. (2016). Are sticky prices costly? evidence from the stock market. *American Economic Review*, **106** (1), 165 – 199.

Hamilton, J. D. (2018). Why you should never use the hodrick-prescott filter. *The Review of Economics and Statistics*, **100** (5), 831–843.

Hobijn, B., Nechio, F. and Shapiro, A. H. (2019). Using brexit to identify the nature of price rigidities, federal Reserve Bank of San Francisco Working Paper 2019-13.

Hodrick, R. J. and Prescott, E. C. (1997). Postwar u.s. business cycles: An empirical investigation. *Journal of Money, Credit and Banking*, **29** (1), 1–16.

Jordà, O. (2005). Estimation and inference of impulse responses by local projections. *American Economic Review*, **95** (1), 161 – 182.

Klenow, P. J. and Malin, B. A. (2010). Microeconomic evidence on price-setting. *Handbook of Monetary Economics*, **3A**, 231–284.

Kryvstov, O. and Vincent, N. (2019). The cyclicality of sales and aggregate price flexibility, working Paper.

Levy, D., Bergen, M., Dutta, S. and Venable, R. (1997). The magnitude of menu costs: Direct evidence from large u. s. supermarket chains. *The Quarterly Journal of Economics*, **112** (3), 791–825.

Mankiw, N. G. and Reis, R. (2002). Sticky information versus sticky prices: A proposal to replace the new keynesian phillips curve. *The Quarterly Journal of Economics*, **117** (4), 1295–1328.

Midrigan, V. (2011). Menu costs, multiproduct firms, and aggregate fluctuations. *Econometrica*, **79** (4), 1139–1180.

Nakamura, E. and Steinsson, J. (2008). Five facts about prices: A reevaluation of menu cost models. *The Quarterly Journal of Economics*, **123** (4), 1415–1464.

— and — (2010). Monetary non-neutrality in a mutisector menu cost model. *The Quarterly Journal of Economics*, **125** (3), 961–1013.

— and — (2013). Price rigidity: Microeconomic evidence and macroeconomic implications. *Annual Review of Economics*, **5**, 133–163.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, **55** (3), 703 – 708.

ONS (2014). *Consumer Price Indices: Technical Manual.* Ofiice for National Statistics, 2014th edn.

Petrella, I., Santoro, E. and Simonsen, L. (2018). Time-varying price flexibility and inflation dynamics, working Paper.

Plehn-Dujowich, J. M. (2005). The optimality of a control band policy. *Review of Economic Dynamics*, **8**, 877–901.

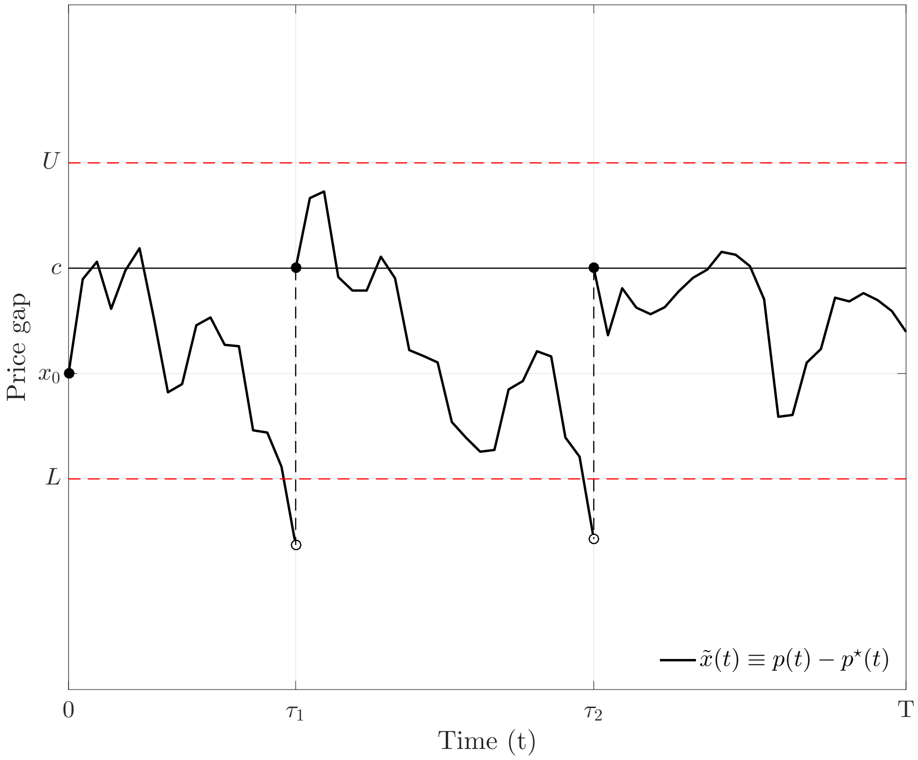
Stokey, N. L. (2009). *The Economics of Inaction: Stochastic Control Models with Fixed Costs*. Princeton University Press, 1st edn.

Woodford, M. (2009). Information-constrained state-dependent pricing. *Journal of Monetary Economics*, **56** (Supplement), S100–S124.

Zbaracki, M. J., Ritson, M., Levy, D., Dutta, S. and Bergen, M. (2004). Managerial and customer costs of price adjustment: Direct evidence from industrial markets. *The Review of Economics and Statistics*, **86** (2), 513–533.

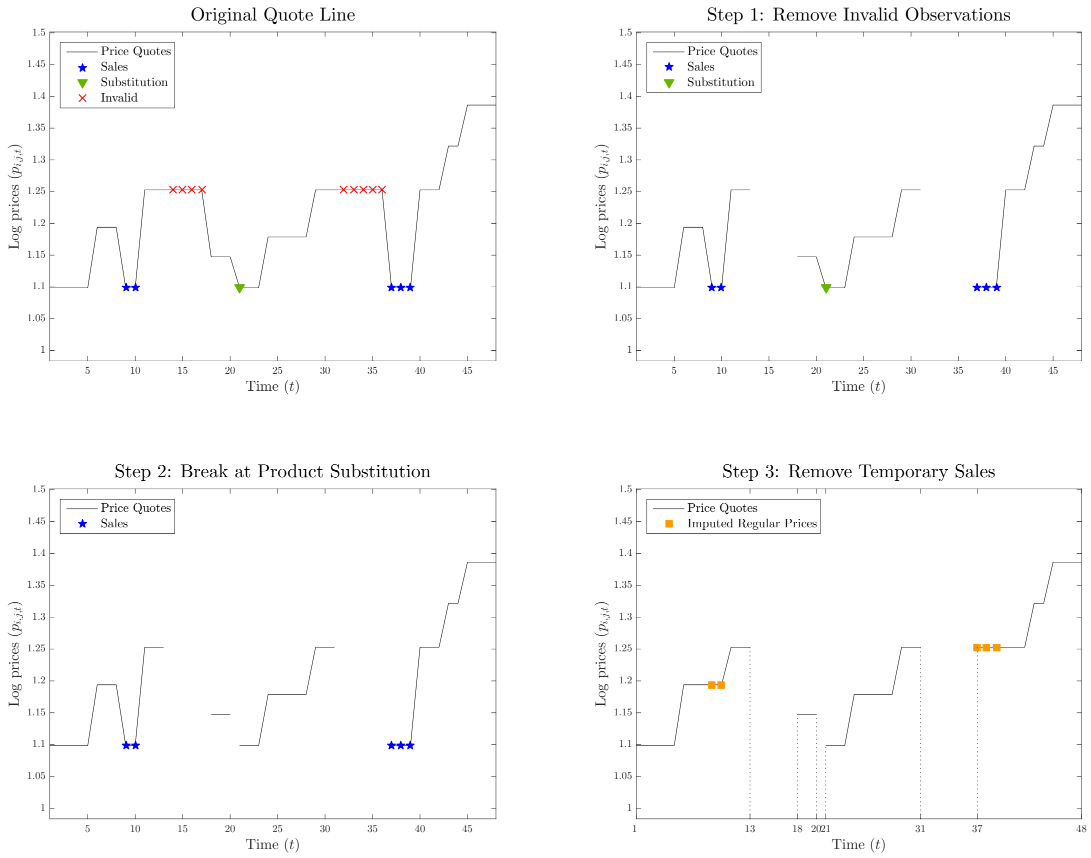
2.A. Figures

Figure 2.A.1: Optimal pricing behavior implied by two-sided (S,s) policies



This figure illustrates the optimal pricing behaviour implied by the random menu cost model described in section 2.2. At periods τ^1 and τ^2 the firm adjusts its nominal prices because at the previous period prices the price gap is outside the inaction region. The size of the adjustments is such that the price gaps at adjustment periods is equal to the reset point (c).

Figure 2.A.2: Step-by-step cleaning of an hypothetical quote-line


Original Quote Line

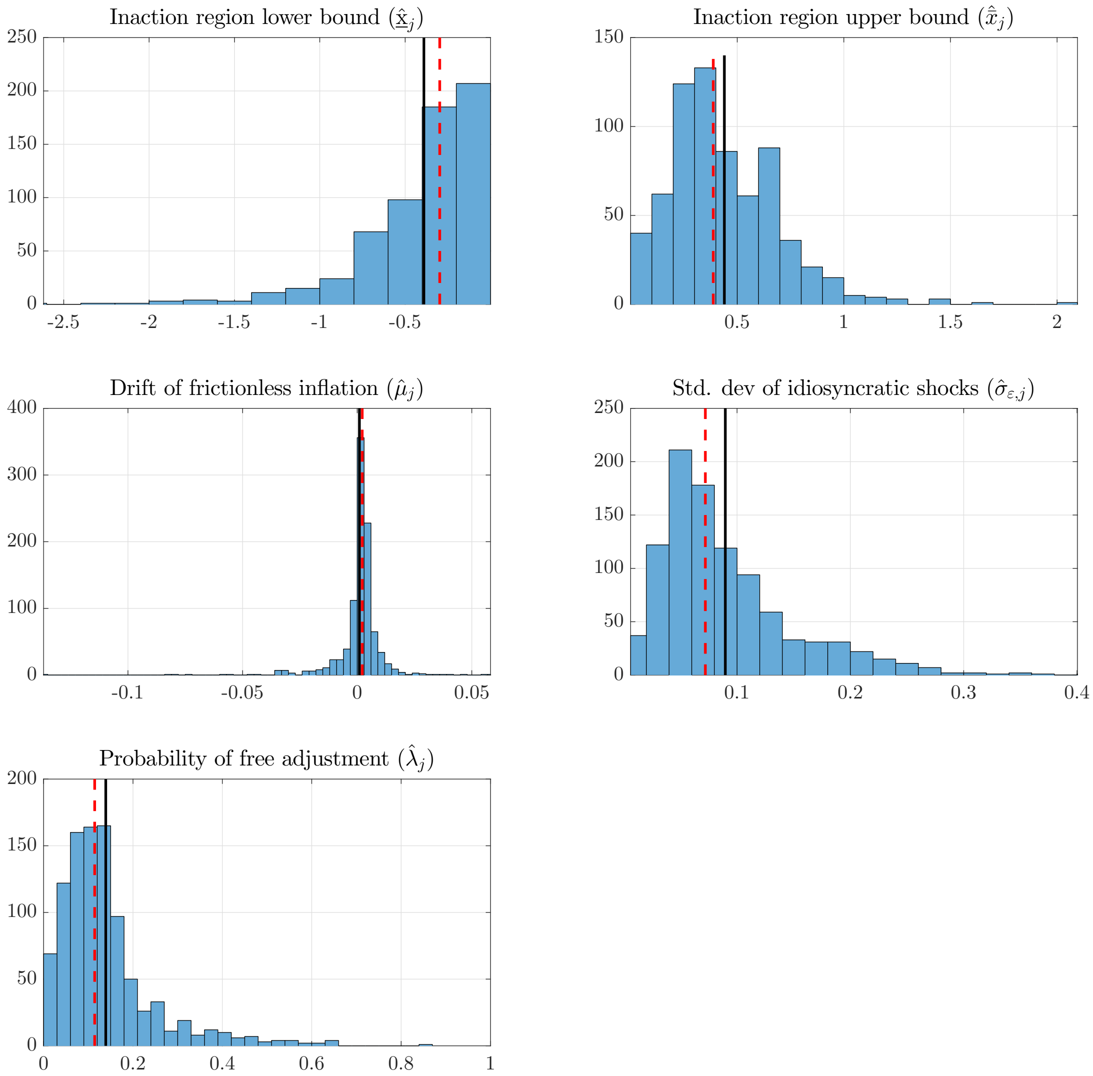
Step 1: Remove Invalid Observations

Step 2: Break at Product Substitution

Step 3: Remove Temporary Sales

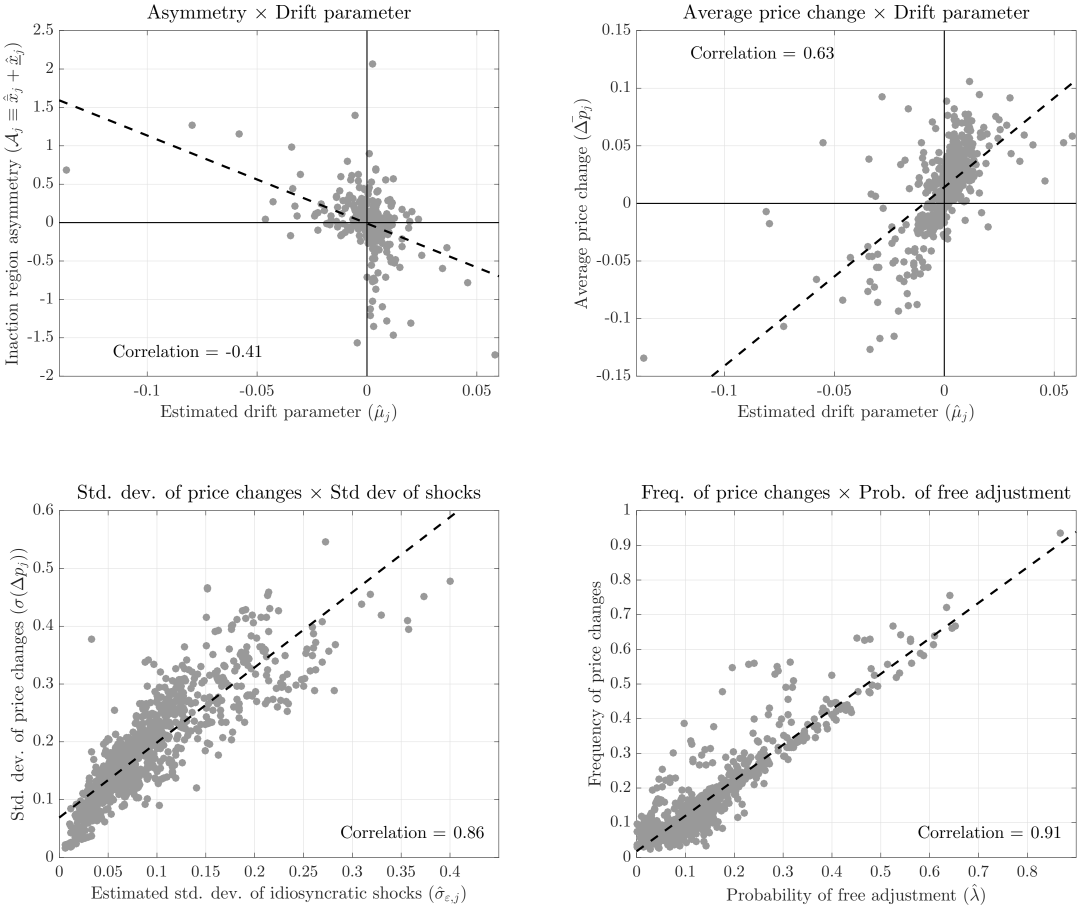
Black solid lines represent raw log price quotes. Blue stars indicate observations flagged as sales. Downward pointing green triangle indicates an observation flagged as a product substitution. Red crosses indicates observations that did not pass the validity checks. Orange squares indicate regular prices imputed for sales prices following the procedure described in the main text.

Figure 2.A.3: Distribution of estimated common parameters at the item level



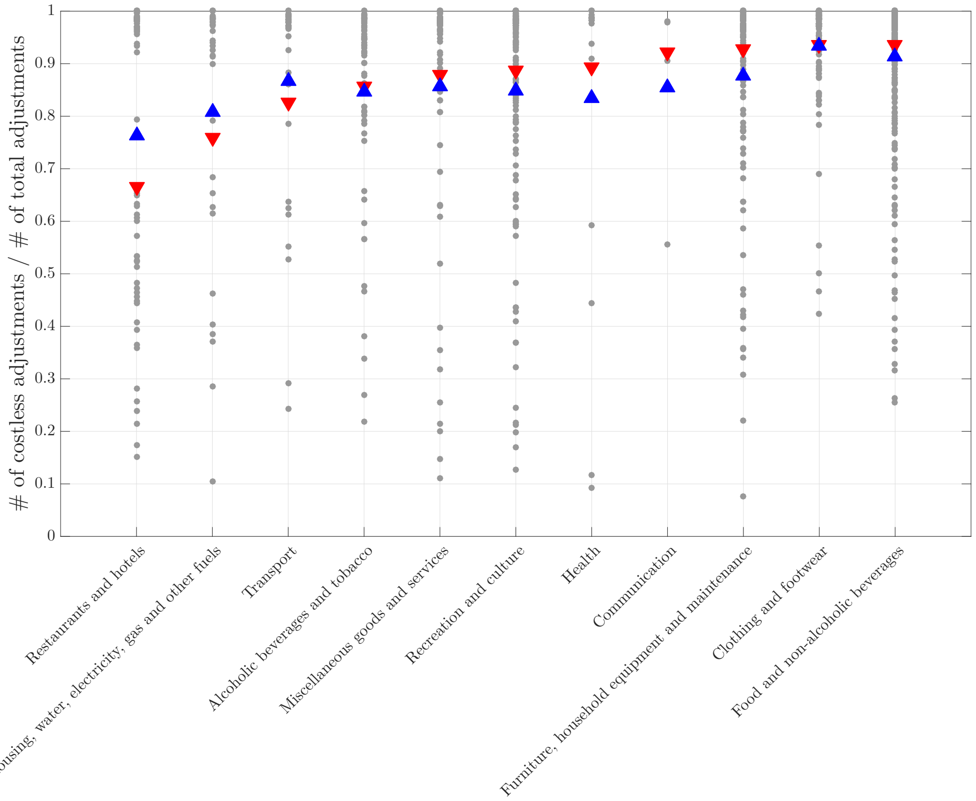
Histograms of estimated common parameters over different items. Histogram for the inaction region lower bound excludes items for which none of the observed price changes is triggered by crossing the lower bound (358 out of 979 items). Histogram for the inaction region upper bound excludes items for which none of the observed price changes is triggered by crossing the upper bound. Solid black and red dashed lines are, respectively, the mean and the median of estimated parameters calculated from the sample that is used to plot the histogram. Descriptive statistics across all items available in table 2.B.1.

Figure 2.A.4: Pricing moments and reduced form parameters



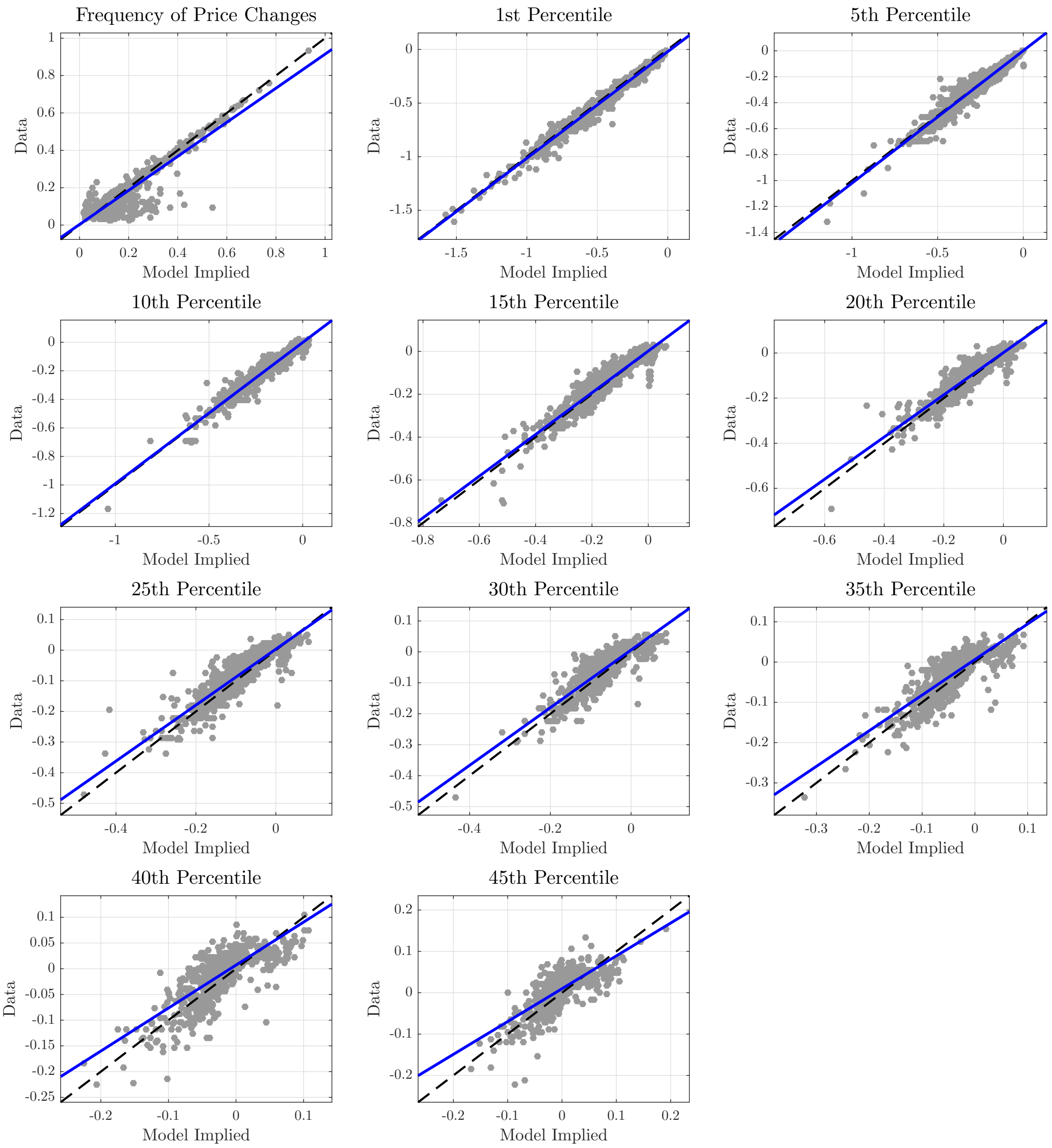
In all the scatter plots, each grey dot represents a given 6-digit item in the sample. In the first scatter plot only items for which there is at least one price change triggered by crossing the upper and the lower boundaries of the inaction region are considered (total of 424 items). In all the scatter plots, the dashed black line is the best fit line obtained from a bivariate regression of the variable on the y-axis on the variable on the x-axis.

Figure 2.A.5: Proportion of costless adjustments by COICOP division



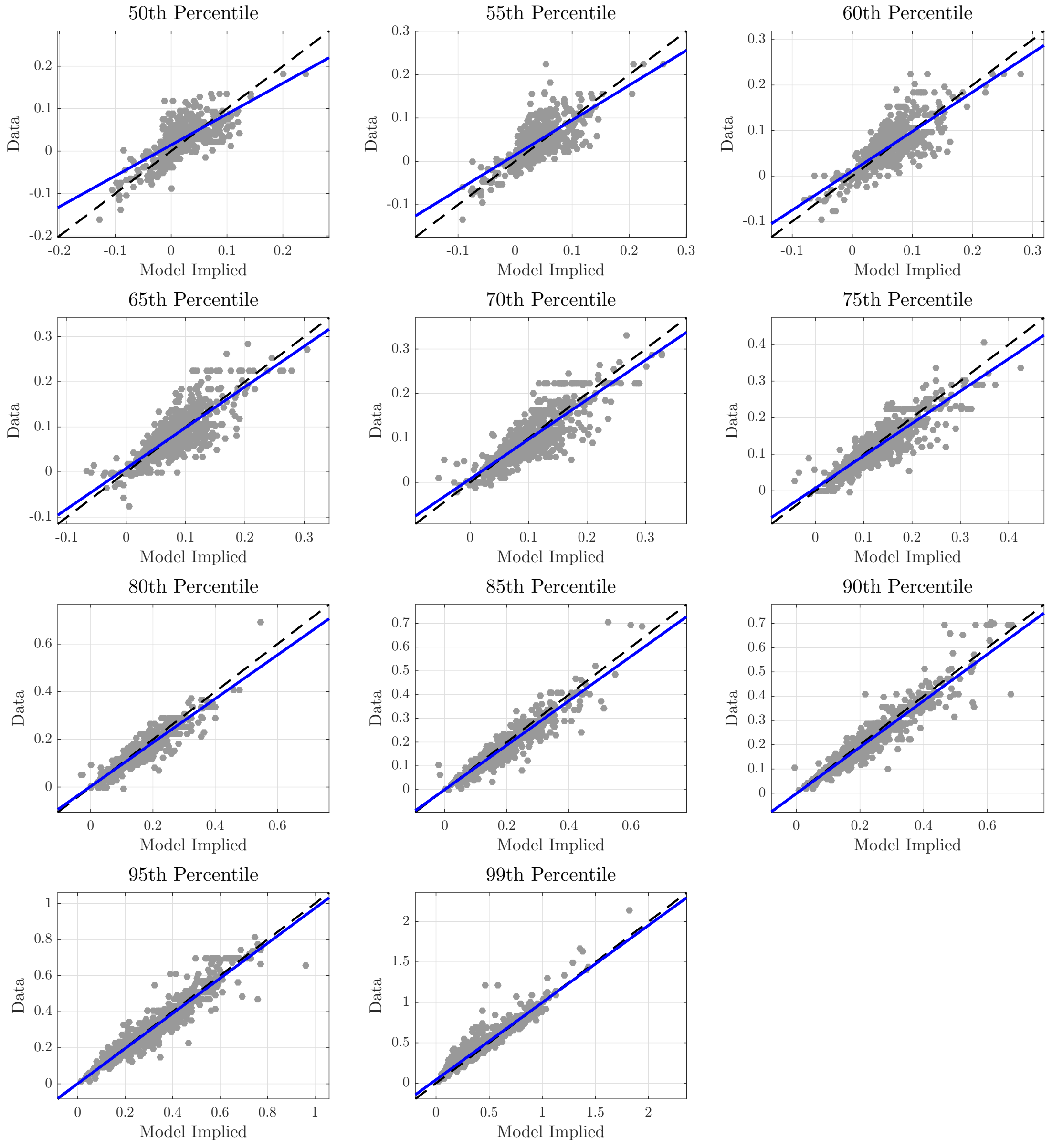
Each grey dot represents the ratio of costless adjustments over the total number of adjustments for each of the 979 unique items in the sample. The red downward pointing triangle contains this ratio computed using all the price changes for that division. The blue upward pointing triangle contains the (unweighted) average ratio across all items in the division.

Figure 2.A.6: Data versus model implied targeted moments (1/2)



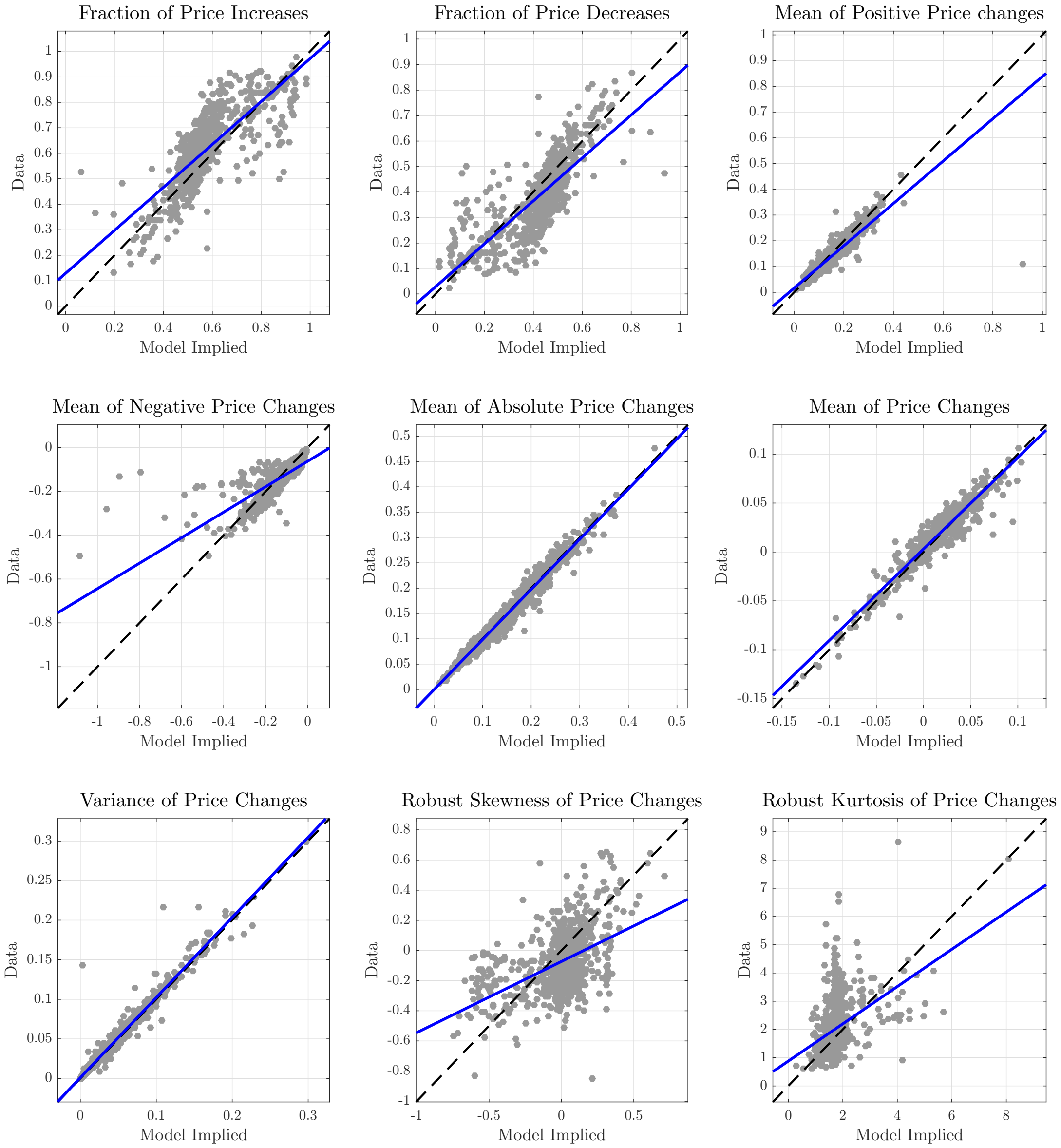
Each grey dot represents, for a given 6-digit item in the sample, the value of the moment in the data (y-axis) against its model implied counterpart (x-axis). The black dashed line is a 45° line. The solid blue line is the best fit line obtained from a bivariate regression of data moments on their model implied counterparts. Both in the data and in simulated data the moments are computed after excluding the first price change of each quote line. To compute the model implied moments a set of 50 panels is generated using the common parameter estimates from the first stage and the same primitive shocks used for the estimation. The model implied moment is the average of the respective moment across simulated panels.

Figure 2.A.7: Data versus model implied targeted moments (2/2)



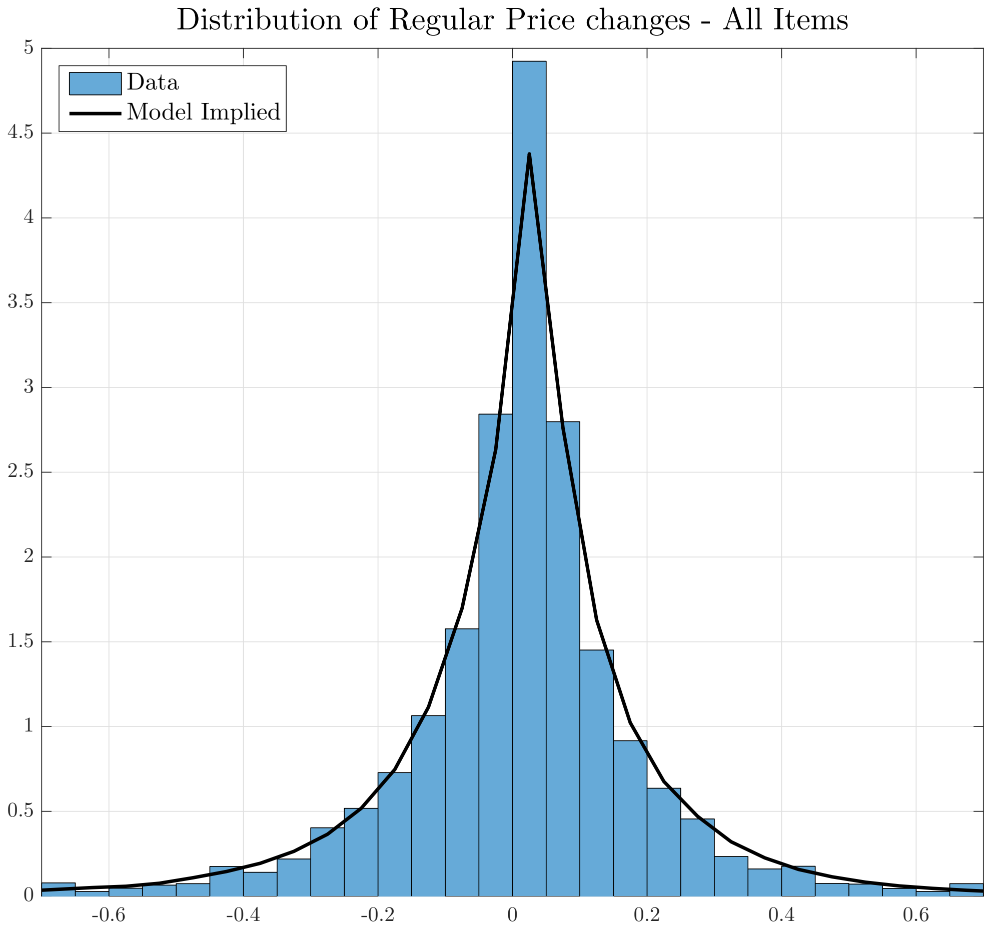
Each grey dot represents, for a given 6-digit item in the sample, the value of the moment in the data (y-axis) against its model implied counterpart (x-axis). The black dashed line is a 45° line. The solid blue line is the best fit line obtained from a bivariate regression of data moments on their model implied counterparts. Both in the data and in simulated data the moments are computed after excluding the first price change of each quote line. To compute the model implied moments a set of 50 panels is generated using the common parameter estimates from the first stage and the same primitive shocks used for the estimation. The model implied moment is the average of the respective moment across simulated panels.

Figure 2.A.8: Data versus model implied non targeted moments



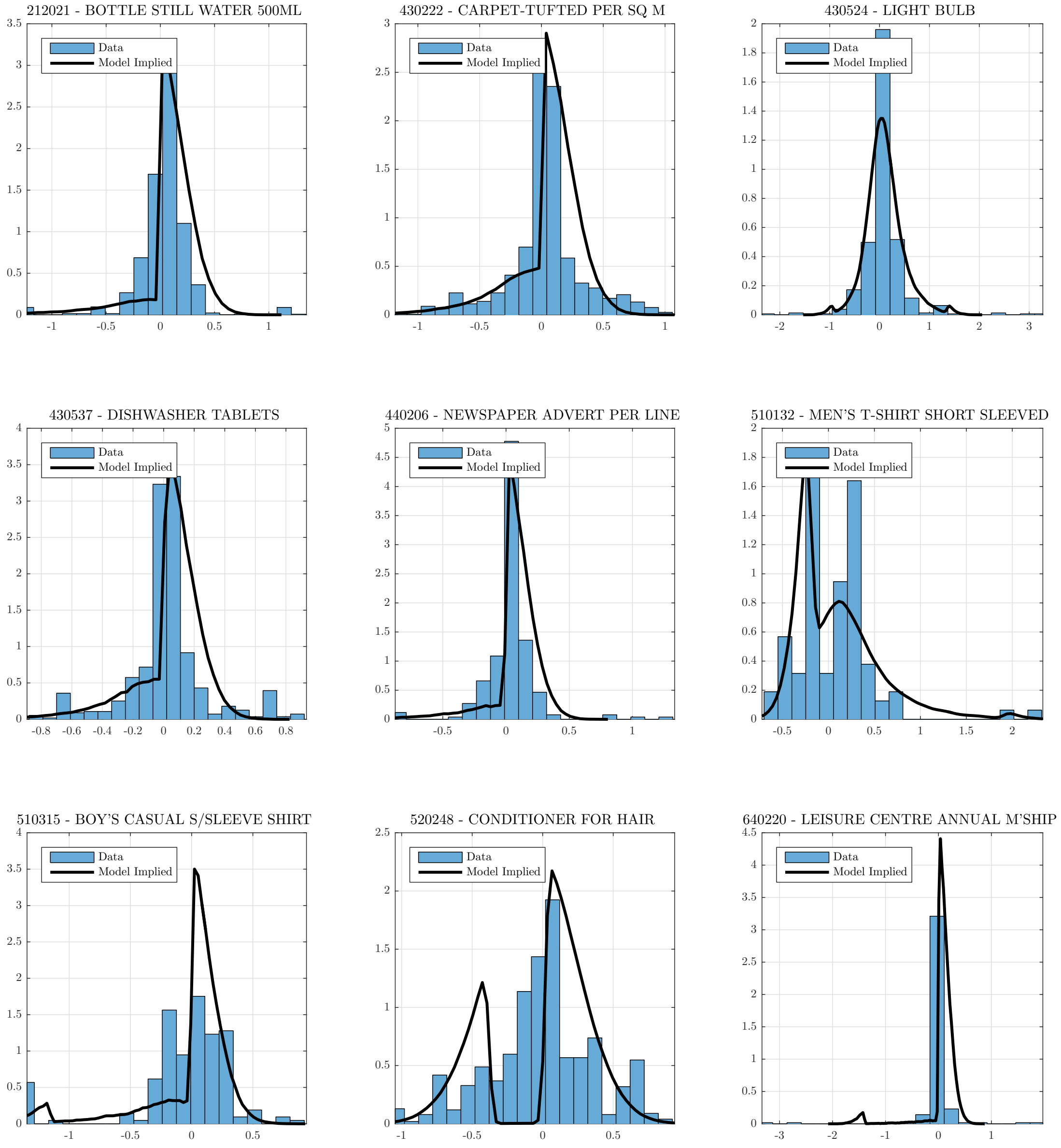
Each grey dot represents, for a given 6-digit item in the sample, the value of the moment in the data (y-axis) against its model implied counterpart (x-axis). The black dashed line is a 45° line. The solid blue line is the best fit line obtained from a bivariate regression of data moments on their model implied counterparts. Both in the data and in simulated data the moments are computed after excluding the first price change of each quote line. To compute the model implied moments a set of 50 panels is generated using the common parameter estimates from the first stage and the same primitive shocks used for the estimation. The robust skewness and robust kurtosis are computed as in Berger and Vavra (2018, table 1), in particular, Robust-Skew = $(P_{90} + P_{10} - 2P_{50})/(P_{90} - P_{10})$ and Robust-Kurt = $(P_{90} - P_{62.5} + P_{37.5} - P_{10})/(P_{75} - P_{25})$.

Figure 2.A.9: Model fit for all items



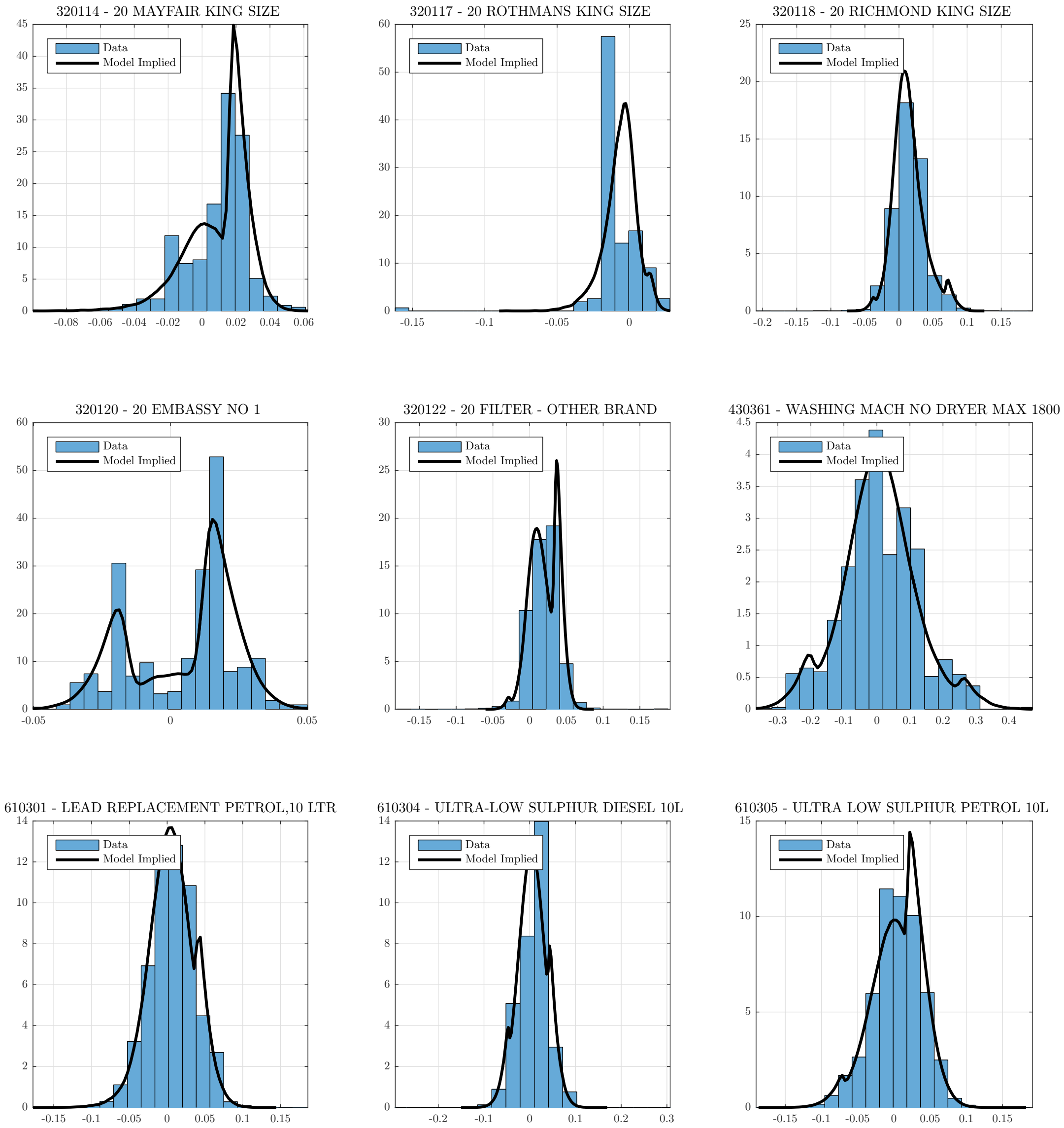
Blue bars represent the histograms for the distributions of regular price changes observed for the whole dataset. The black solid line is kernel density estimate of the distribution of price changes for all items over 50 panels of simulated data. Simulated data is generated by combining data of separate items and simulating using the the estimated common parameters. When considering the price changes, the first price change in each quote line is excluded. The histogram excludes log price changes that are not in the range $[-0.7, 0.7]$. The excluded observations account for 0.78% of the total number of non-zero price changes.

Figure 2.A.10: Model fit for worst fitting 9 items



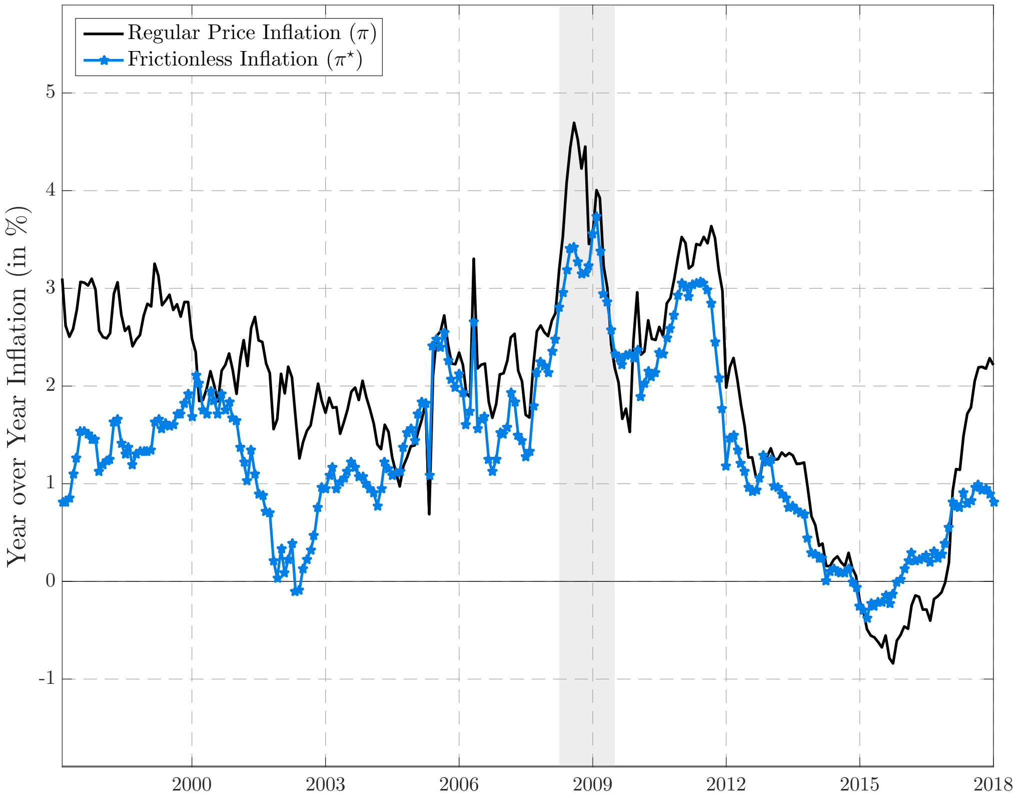
Blue bars represent the histograms for the distributions of regular price changes observed in the data for a given item. The black solid lines the kernel density estimates of the distribution of price changes over 50 panels of simulated data. Each panel was simulated using the estimated parameter values for the first stage for each item and the same primitive shocks used for estimation. When considering the price changes, the first price change in each quote line is excluded. The items chosen are those for which the first-stage SMM objective function evaluated at the estimated parameters displayed the *largest* values.

Figure 2.A.11: Model fit for best fitting items



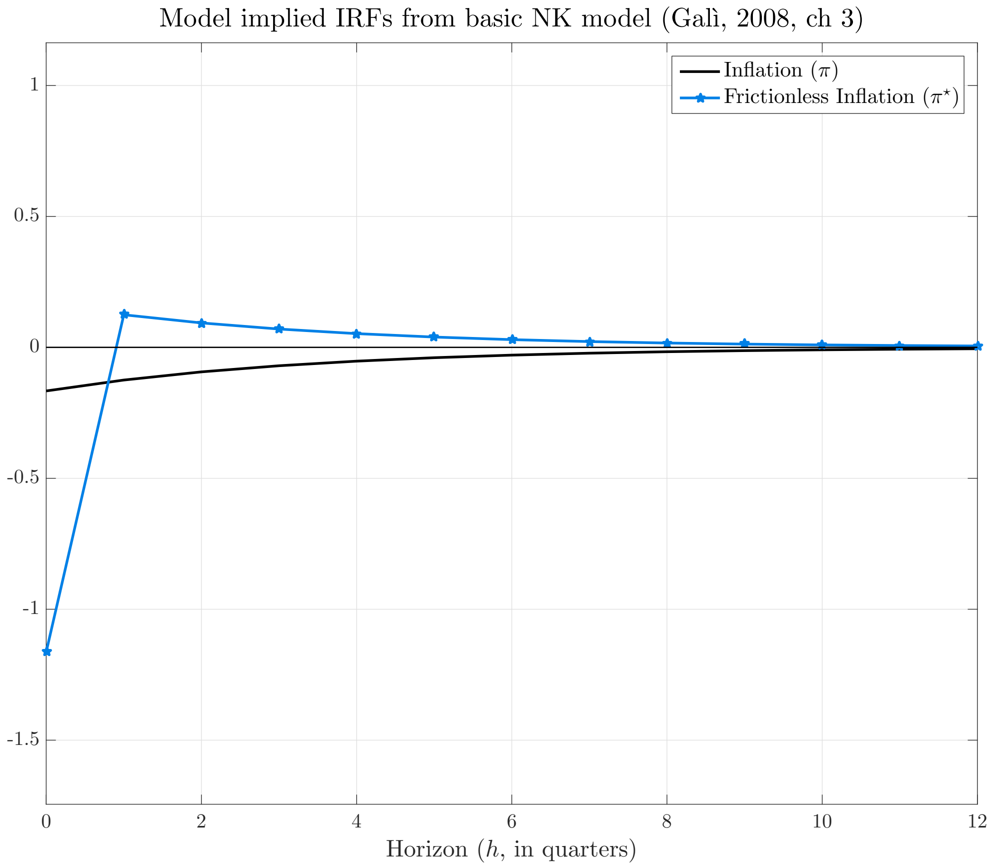
Blue bars represent the histograms for the distributions of regular price changes observed in the data for a given item. The black solid lines the kernel density estimates of the distribution of price changes over 50 panels of simulated data. Each panel was simulated using the estimated parameter values for the first stage for each item and the same primitive shocks used for estimation. When considering the price changes, the first price change in each quote line is excluded. The items chosen are those for which the first-stage SMM objective function evaluated at the estimated parameters displayed the *smallest* values.

Figure 2.A.12: Regular price versus frictionless inflation in the UK



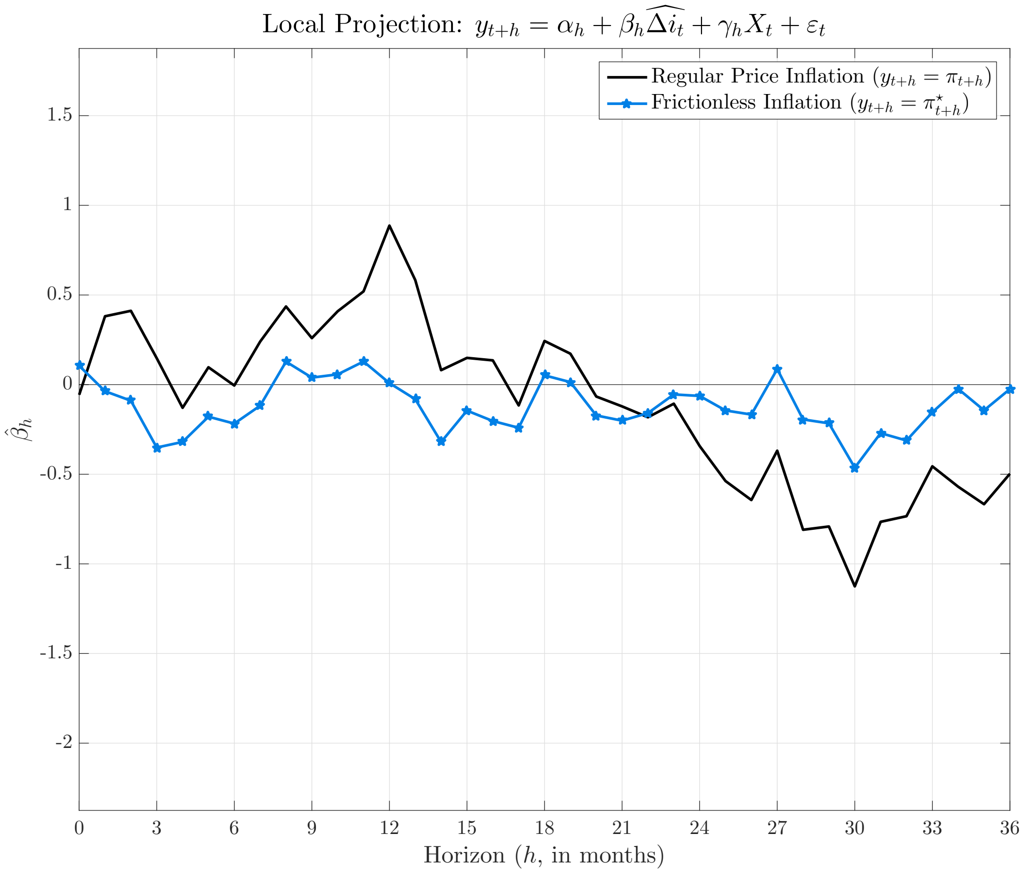
Regular price and frictionless price indexes are computed at a monthly frequency from weighted averages of the elementary aggregates in (2.9). Year over year inflation is computed as the percentage variation in the index of a given month against the same month of the previous year. The grey shaded area denotes the great recession.

Figure 2.A.13: Model implied IRFs to a monetary policy shock



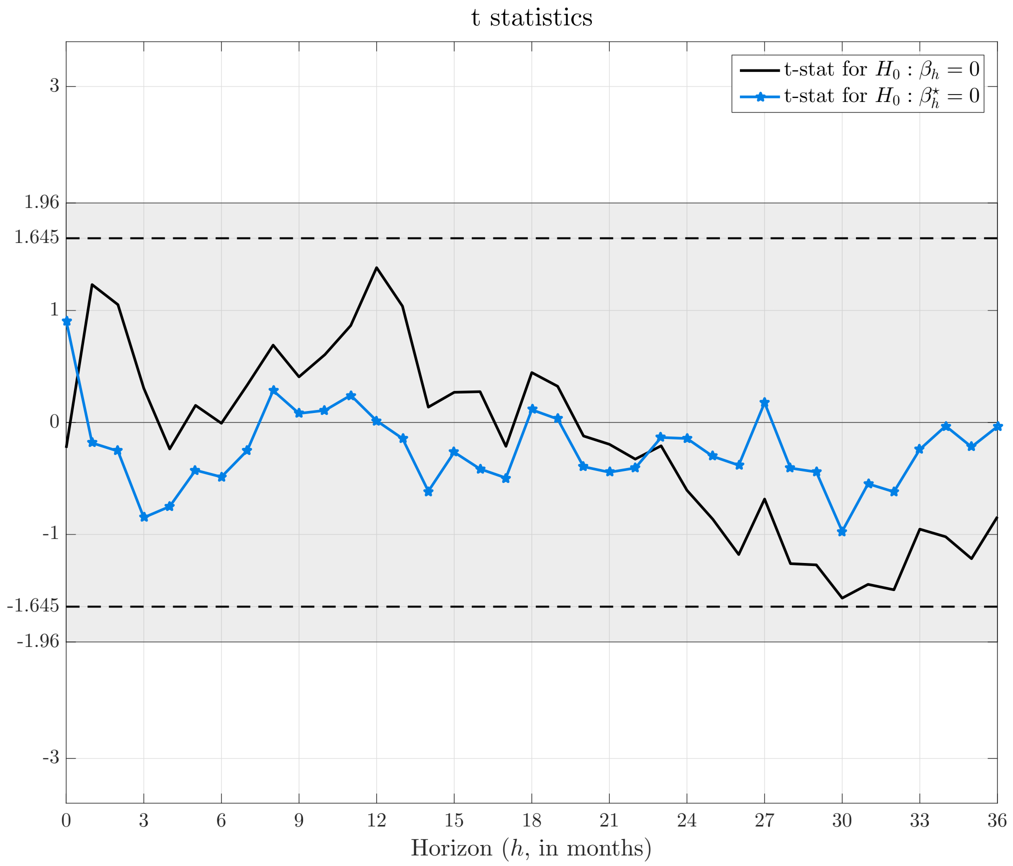
Impulse responses obtained under the baseline parameter values described in Galí (2008, p.53) except for the persistence of the monetary policy shock, ρ_v, that is set to 0.75 (instead of 0.5).

Figure 2.A.14: Estimated IRFs to a monetary policy shock



Impulse responses obtained from a local projection of the form $y_{t+h} = \alpha_h + \beta_h \widehat{\Delta i_t} + \gamma_h X_t + \varepsilon_t$ for horizons $h = 0, 1, \ldots, 36$. The change in the nominal Bank of England interest rate is instrumented with the series of high-frequency identified monetary surprises by Cesa-Bianchi, Thwaites and Vicondoa (2020) and the vector of controls X_t includes four lags of regular price inflation and frictionless inflation.

Figure 2.A.15: Significance tests of estimated IRFs to a monetary policy shock



For a local projection of the form $y_{t+h} = \alpha_h + \beta_h \widehat{\Delta i_t} + \gamma_h X_t + \varepsilon_t$ this plot contains the t-statistics for the null hypothesis $H_0 : \beta_h = 0$ over different horizons. For a given horizon if the line is outside the grey shaded (dashed) area indicates the null can be rejected at the 5% (10%) against a double sided alternative.

2.B. Tables

Table 2.B.1: Descriptive statistics for estimated parameters

	$\hat{\underline{x}}_j$	$\hat{\bar{x}}_j$	$\hat{\mu}_j$	$\hat{\sigma}_{\varepsilon,j}$	$\hat{\lambda}_j$
Mean	-2.691	2.271	0.001	0.09	0.139
Median	-0.596	0.568	0.002	0.072	0.114
Std Dev	8.633	6.676	0.011	0.06	0.109
IQR	1.87	1.466	0.004	0.064	0.092
5th Percentile	-9.377	0.129	-0.014	0.023	0.024
10th Percentile	-5.586	0.194	-0.007	0.032	0.039
25th Percentile	-2.104	0.31	0	0.05	0.073
75th Percentile	-0.234	1.776	0.004	0.114	0.165
90th Percentile	-0.056	4.925	0.008	0.179	0.266
95th Percentile	-0.02	8.815	0.012	0.213	0.37
Minimum	-130.84	0	-0.137	0.006	0.001
Maximum	0	130.262	0.058	0.4	0.868
N	979	979	979	979	979

Descriptive statistics calculated over different items. All the numbers rounded to three decimal places and any number smaller than 5×10^{-4} is displayed as a zero.

Table 2.B.2: Inflation wedge and changes in the output gap

Deviation from	**HP trend**	**Cubic trend**	**Hamilton trend**
Monthly GDP			
	-5.52	-10.62*	-3.06
	(4.56)	(5.93)	(4.18)
Industrial Production			
	-3.11	-4.6*	-3.77*
	(2.44)	(2.45)	(2.21)

Each cell contains the slope coefficient from a bivariate regression of the form $(\pi_t - \pi_t^\star) = \alpha + \beta\Delta\tilde{y}_t + w_t$. In parenthesis are the Newey and West (1987) HAC standard errors with 12 lags. In all the specifications the dependent variable is computed as the difference between year over year regular price inflation and its frictionless counterpart (as depicted in figure 2.A.12). The independent variable varies across specifications: output is measured by either the log of a monthly GDP index (top panel) or the log of industrial production index (bottom panel) and the output gap is computed as the log deviation of the given output measure from an HP trend (second column), from a cubic trend (third column) or from a Hamilton (2018) trend (fourth column). * indicates significance at the 10% level, ** indicates significance at the the 5% level and *** indicates significance at the 1% level. All the numbers are rounded to two decimal places.

Table 2.B.3: Forecasting Inflation - Frictionless versus Headline

Forecasting horizon	**6 Months**	**12 Months**	**24 Months**	**36 Months**
In-sample standard error				
Frictionless	0.67	0.95	1.54	2.09
Headline	0.69	0.99	1.78	2.54
Both	0.66	0.94	1.54	2.09
Out-of-sample root mean squared error				
Frictionless	0.81	1.16	1.99	2.40
Headline	0.83	1.30	2.79	4.27
Both	0.77	1.09	1.69	2.00
Multivariate regression coefficients				
Frictionless	$0.26^{\star}$	$0.47^{\star}$	$1.32^{\star\star}$	$2.14^{\star\star\star}$
	(0.14)	(0.25)	(0.52)	(0.67)
Headline	0.08	0.20	0.08	0.08
	(0.12)	(0.19)	(0.45)	(0.69)

Each cell in the table is derived from a regression of the form $\pi_{t,t+h} = \alpha + \mathbf{X}\beta + \varepsilon_t$ where h denotes a particular forecasting horizon. For each forecasting horizon, $\mathbf{X}$ contains headline published inflation over the previous 12 months ($\pi_{t-12,t}$) or frictionless inflation over the previous 12 months ($\pi^{\star}_{t-12,t}$) or both. The in-sample standard error in the top panel is computed from the whole sample 1997m2 to 2018m1. The out-of-sample root mean squared error in the second panel is computed by estimating the regression using data from 1997m2 to 2008m1 and using the forecasting errors from then to the end of the sample. The third panel reports the regression coefficients of the multivariate regression containing both $\pi_{t-12,t}$ and $\pi^{\star}_{t-12,t}$ over the whole sample 1997m2 to 2018m1. In parenthesis are the respective Newey and West (1987) standard errors with a lag length choice equal to 12. * indicates significance at the 10% level, ** indicates significance at the the 5% level and *** indicates significance at the 1% level. All the numbers are rounded to two decimal places.

2.C. Proof of results in the main text

This appendix provides the proofs for the two propositions in the main text. For expositional purposes, the proof of proposition 2.2 is presented first followed by the proof of proposition 2.1. The underlying model and the associated notation are identical to the basic new Keynesian model in Galí (2008, chapter 3).

Proof of proposition 2.2. Under monopolistic competition and a demand function arising from a CES aggregator with elasticity of substitution ε, frictionless prices satisfy:

$$P_t^\star = \mathcal{M}\,\psi_{t|t} \tag{2.17}$$

where $\psi_{t|t}$ are the nominal marginal costs of a firm changing prices at time t and $\mathcal{M} \equiv \varepsilon/(\varepsilon - 1)$ is a constant markup that monopolist would charge at every time period in the absence of constraints on the frequency of price adjustment, also referred to as the *desired* or *frictionless* markup. Dividing both sides of (2.17) by P_t and taking logs yields,

$$p_t^\star - p_t = mc_t - mc \tag{2.18}$$

where $mc = -\log(\mathcal{M})$ is the steady state value of marginal cost and mc_t is the log of the economy's average real marginal cost. Since the log of deviation of real marginal cost from steady state is proportional to the log deviation of output from its flexible price counterpart, use equation (20) in Galí (2008, p. 48) to obtain,

$$p_t^\star - p_t = \left(\sigma + \frac{\varphi + \alpha}{1 - \alpha}\right)\tilde{y}_t \tag{2.19}$$

where $\tilde{y}_t$ is the *output gap*, defined as the log deviation of output from its flexible part counterpart, σ is the elasticity of intertemporal substitution, φ is the Frisch elasticity of labor supply and $1 - \alpha \in [0, 1]$ is the exponent of labor in the production function. Lagging (2.19) by one period and subtracting from (2.19),

$$\pi_t^\star - \pi_t = \left(\sigma + \frac{\varphi + \alpha}{1 - \alpha}\right) \Delta \tilde{y}_t \tag{2.20}$$

Finally, multiplying both sides of (2.20) by minus one yields the result in (2.14).

□

Proof of proposition 2.1. I consider the responses of inflation and frictionless inflation under an interest rate rule as in Galí (2008, section 3.4.1). The stochastic component in the interest rate is v_t and it is assumed to follow an AR(1) process, that is, $v_t = \rho_v v_{t-1} + \varepsilon_t^m$ where $\rho_v \in [0, 1)$ and ε_t^m is the monetary policy shock. From (2.20) and any given horizon $h \geqslant 0$ we have that,

$$\frac{\partial \pi_{t+h}^\star}{\partial \varepsilon_t^v} - \frac{\partial \pi_{t+h}}{\partial \varepsilon_t^m} = \left(\sigma + \frac{\varphi + \alpha}{1 - \alpha}\right) \frac{\partial \Delta \tilde{y}_{t+h}}{\partial \varepsilon_t^m} \tag{2.21}$$

Since $\left(\sigma + \frac{\varphi+\alpha}{1-\alpha}\right) > 0$ the relationship between the impulse responses of frictionless inflation and inflation can be inferred from the sign of the impulse response of the changes in the output gap. Using the method of undetermined coefficients, the solution for the output gap is given by,

$$\tilde{y}_{t+h} = -(1 - \beta\rho_v)\, \Lambda_v \, v_{t+h} \tag{2.22}$$

where β is the representative household discount factor and Λ_v is a convolution of structural parameters that and takes only positive values.[31] Finally, in terms of impulse responses it is the case that,

$$\frac{\partial \Delta \tilde{y}_{t+h}}{\partial \varepsilon_t^m} = -(1 - \beta\rho_v)\, \Lambda_v \left(\frac{\partial v_{t+h}}{\partial \varepsilon_t^m} - \frac{\partial v_{t+h-1}}{\partial \varepsilon_t^m}\right) \tag{2.23}$$

[31] For values of structural parameters that ensure equilibrium uniqueness which is maintained assumption, see Galí (2008, equation 27).

Note that for the case $h = 0$, the last term in brackets in (2.23) is equal to one since $\partial v_{t-1}/\partial \varepsilon_t^m = 0$. Since the term $(1 - \beta\rho_v)\Lambda_v$ is positive, it follows that the expression above is *negative* and, hence, $\partial\pi_t^\star/\partial\varepsilon_t^v < \partial\pi_t/\partial\varepsilon_t^m$. Moreover, since the solution for inflation is given by $\pi_t = -\kappa\,\Lambda_v\,v_t$ where $\kappa > 0$ is the slope of the new Keynesian Phillips curve and, hence, $\partial\pi_t/\partial\varepsilon_t^m < 0$. Therefore, (2.10) holds. Finally, for any $h > 0$ expression (2.23) simplifies to,

$$\frac{\partial\Delta\tilde{y}_{t+h}}{\partial\varepsilon_t^m} = -\underbrace{(1 - \beta\rho_v)\,\Lambda_v\rho_v^{h-1}}_{>0}\;\underbrace{(\rho_v - 1)}_{<0}\; > 0 \tag{2.24}$$

Therefore, $\partial\pi_{t+h}^\star/\partial\varepsilon_t^v > \partial\pi_{t+h}/\partial\varepsilon_t^m$ as stated in (2.11). As an endnote, notice that the sign of $\pi_{t+h}/\partial\varepsilon_t^m$ cannot be determined unambiguously. More precisely using the solutions for π_t and $\Delta\tilde{y}_t$ and the relationship in (2.20),

$$\frac{\partial\pi_{t+h}^\star}{\partial\varepsilon_t^v} = \kappa\Lambda_v\rho_v^{h-1}\left[\frac{(1 - \beta\rho_v)(1 - \rho_v)}{\lambda} - \rho_v\right] \tag{2.25}$$

where λ is a positive constant that is a convolution of structural parameters. The sign of the last term in square brackets will depend on the specific calibration of structural parameters. □

Chapter 3

Disaggregated Impulse Responses via the classifier-Lasso

3.1. Introduction

Ever since the seminal contributions of Frisch (1933) and Slutzky (1937), macroeconomists have looked at random shocks as the primary source of business cycle fluctuations (Ramey, 2016). For most of the twentieth century, the bulk of the work in empirical macroeconomics was devoted to identification of macroeconomic shocks and estimation of impulse response functions (IRFs) which summarised the expected dynamic responses of aggregate variables to those shocks. More recently, with the increasing availability of micro data, the focus of the literature has gradually shifted towards understanding how different firms and households respond to these aggregate macroeconomic shocks. The estimation of *disaggregated IRFs* is important not only to quantify what drives the responses of aggregate variables but also as a means to empirically test the predictions from alternative transmission theories.

This paper studies the estimation of disaggregated IRFs in a setting where there is *latent group heterogeneity*. This type of setting is characterised by two main features: (i) each individual belongs to a group within a broadly heterogeneous population and the individual IRFs are the same *within* a group but differ *between* groups and (*ii*) the researcher has disaggregated data on the outcome of interest and a strictly exogenous macroeconomic shock, but does *not* know either how many groups are in the population or which group each individual belongs to. This setting is illustrated in figure 3.A.1. The estimated individual-specific IRFs (grey lines) are obtained from a dataset in which half of the individuals belong to a group for which the true IRF is given by the green-solid line whilst for the other half the true IRF is given by the red-solid line. The problem considered in this paper is that of a researcher that observes a dataset that yields the "cloud" of grey-lines and, based on that dataset, would like to estimate: (*i*) the number of latent groups in the population, (*ii*) the associated group-specific IRFs and (*iii*) which individuals belong to which group.

In the existing literature, the common approach to the estimation of disaggregated IRFs involves first grouping individuals in the sample according to some external classification or observable explanatory variables and subsequently estimating the group-specific IRFs by pooling together individuals that are assumed to belong to the same group. For example, when studying the responses of consumption expenditures to monetary policy changes households can be grouped according to their housing tenure status (Cloyne, Ferreira and Surico, 2019) or according to their relative position in the wealth distribution (Coibion, Gorodnichenko, Kueng and Silvia, 2017). This paper shows that, in the presence of latent group heterogeneity, this ex-ante grouping approach can lead to misleading conclusions. More precisely, it is shown that there is a bias-variance tradeoff between the IRF estimates obtained by ex-ante grouping and the IRFs estimated for each individual separately. The IRF estimates based on the ex-ante grouping of individuals are more precise but are subject to a form of bias, which is here labeled *misclassification bias*, that arises whenever the grouping of individuals imposed by the researcher pools together individuals that do not react in the same way to aggregate shocks.

Motivated by this theoretical result, this paper introduces an alternative methodology to estimate disaggregated IRFs in the presence of latent group heterogeneity. The methodology uses penalized estimation techniques in order to simultaneously estimate the unknown group-specific IRFs and classify individuals to groups whereas the number of latent groups is subsequently estimated via a BIC-type information criterion. This methodology is an application of the classifier-Lasso (C-Lasso) framework developed in Su, Shi and Phillips (2014, 2016) to the estimation of disaggregated IRFs. The theoretical results in Su, Shi and Phillips (2014, 2016) imply that, under a suitable set of assumptions, the group-specific IRFs estimator based on the C-Lasso have the same asymptotic properties as the *group-oracle* IRFs, that is, the IRF estimates that would result from an ex-ante grouping of individuals that exactly matches the true unknown grouping of individuals. Most importantly, and in sharp contrast with ex-ante grouping approach, the C-Lasso IRF estimator achieve this property in a completely data-driven way that does not require the researcher to take a stance on either the number of latent groups or the individual group membership.

To illustrate the finite sample performance of the C-Lasso based classification and estimation procedure, this paper uses a Monte Carlo experiment in which artificial datasets are generated from the same DGP used to generate the IRFs in figure 3.A.1. For each Monte Carlo sample the C-Lasso framework is used to estimate the IRFs both by estimating the whole moving average representation directly and by local projections (Jordà, 2005). The results from this Monte Carlo experiment complement the evidence presented in Su, Shi and Phillips (2014, 2016) and illustrate the good performance of the C-Lasso in terms of determination of the number of latent groups, classification of individuals into different groups and estimation of group-specific IRFs in samples of similar size than those typically used in the existing literature estimating disaggregated IRFs.

As an empirical application, the C-Lasso framework is used to revisit the dynamic responses of firm-level debt to an aggregate investment specific technology (IST) shock from Drechsel (2020). A theoretical prediction from the model in Drechsel (2020) is that the debt of firms that tend to borrow against collateral should decrease following a positive IST shock whereas the debt of firms that tend to borrow against their future earnings should increase. When applied to a subset of the Drechsel (2020) dataset, the C-Lasso framework identifies two latent groups, one for which the response of firm-level debt to an IST shock is positive and other for which debt reacts negatively. The group of firms that increase their debt following an IST shock is composed of firms that are relatively smaller, have a higher share of intangible assets, tend to be earnings-based borrowers and do not belong to the consumer staples or utilities sectors. Altogether, these findings are in line with the theoretical predictions from Drechsel (2020), but also suggest that, on top of whether a firm tends to borrow against earnings or collateral, the specific *sector* that the firm operates also plays a role in determining whether it will increase or decrease its debt in response to aggregate IST shocks.

Relation to the literature. This paper relates and contributes to two strands of literature. First and foremost, it relates to the empirical macroeconomics literature that focuses on the estimation of impulse response functions and, in particular, to several empirical applications that estimate heterogeneous impulse responses functions to aggregate shocks (see section 3.2.2 for a review of some of these applications). This paper contributes to this literature in two ways. First, by formally showing that there is a bias-variance tradeoff between the estimator based on ex-ante individual classification and the estimator based on individual-specific impulse responses. Second, by introducing an alternative methodology that produces estimates that achieve a smaller mean squared error without the need to ex-ante take a stance on individual group membership.

Second, it relates to an extensive literature on variable-coefficient models in panel data (see, for instance, Hsiao, 2014, chapter 6) and, in particular, to panel structure models where individuals are assumed to belong to a number of homogeneous groups within a broadly heterogeneous population and the regression parameters are the same *within* each group but differ *across* groups. Different approaches have been proposed to determine an unknown group structure in modeling unobserved slope heterogeneity in panels, including finite mixture models (e.g. Sun, 2005; Kasahara and Shimotsu, 2009; Browning and Carro, 2014), variants of the K-means algorithm (e.g. Lin and Ng, 2012; Sarafidis and Weber, 2015; Bonhomme and Manresa, 2015) and penalized estimation techniques (e.g. Su, Shi and Phillips, 2014, 2016; Su and Ju, 2018). This paper is the first paper to apply the methods developed in this literature, and in particular the C-Lasso framework developed in (Su, Shi and Phillips, 2014, 2016), to the estimation of impulse response functions.[1]

Structure of the paper. Section 3.2 introduces the data generating process and discusses of the some empirical applications it can accommodate. Section 3.3 discusses the statistical properties of the common approach used to estimate heterogeneous impulse responses in the literature. Section 3.4 introduces a C-Lasso based methodology to estimate heterogeneous impulse responses in the presence of latent group heterogeneity and discusses its asymptotic properties. Section 3.5 uses a Monte Carlo experiment to illustrate the finite sample properties of the proposed methodology. Section 3.6 applies this methodology to revisit the Drechsel (2020) IRF estimates of firm level debt to an aggregate IST shock. Section 3.7 concludes and discusses some promising avenues for future research.

Notation. In all that follows, bold letters are used to denote vectors or matrices and non-bold fonts denote scalars. For a given matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{A}'$ denotes its transpose, $\mathbf{a}_{i,*}$ denotes its i-th row, $\mathbf{a}_{*,j}$ denotes its j-th column and $a_{i,j}$ denotes its (i,j)-th element. Moreover, $\mathbf{1}_{\boldsymbol{m\times n}}$ and $\mathbf{0}_{\boldsymbol{m\times n}}$ denote $m \times n$ matrices of ones and zeros, $\mathbb{1}\left\{\cdot\right\}$ denotes the indicator function, $\|\cdot\|$ denotes the Frobenius norm, $\otimes$ denotes the Kronecker product and $\oplus$ the direct sum of matrices. For any two real numbers $a < b$, denote by $\mathbb{Z}_{[a,b]}$ the set of all integers in $[a,b]$.

[1] The idea of combining some of the methods developed in this literature to estimate heterogenous reactions to a common aggregate shock is also present in Lewis, Melcangi and Pilossoph (2019), where the authors use a variant of the K-means algorithm to estimate the marginal propensity to consume from the 2008 tax rebate. Two key differences between Lewis, Melcangi and Pilossoph (2019) and the present paper are that they focus on the estimation of the on impact effect of a shock (*i.e.* the marginal propensity to consume) whilst this paper focuses on the estimation of the whole impulse response functions and, methodologically, they build from the K-means algorithm in Bonhomme and Manresa (2015) whereas the present paper builds from the C-Lasso framework.

3.2. Impulse response functions under latent group heterogeneity

This section introduces the process that is assumed to generate the panel data set observed by the researcher and discusses how it can accommodate some existing empirical applications that estimate disaggregated IRFs.

3.2.1. The data generating process

In a nutshell, the assumed data generating process can be characterised as a distributed lag model in which the coefficients are allowed to vary across different groups of individuals. For expositional purposes, consider first the distributed lag assumption formalised as follows,

***Assumption* 3.1** *The researcher observes a panel data set $\{(y_{i,t}, \mathbf{x}_{i,t})\}$ for $i = 1, \dots, N$ and $t = 1, \dots, T$ for which the data generating process can be represented as,*

$$y_{i,t} = \mathbf{x}'_{i,t}\,\boldsymbol{\beta}_i + \varepsilon_{i,t} \tag{3.1}$$

where $\mathbf{x}'_{i,t} \equiv [x_{i,t}, x_{i,t-1}, \dots, x_{i,t-H}]$ and $\boldsymbol{\beta}_i = [\beta_{i,0}, \dots, \beta_{i,H}]$. Moreover, let $\mathbf{X}_i = [\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,T}]'$ and $\mathbf{X} = \{\mathbf{X}_i\}_{i=1}^{n}$ and assume the following three conditions hold: (a) *rank* $(\mathbf{X}_i'\mathbf{X}_i) = H+1$, $\forall i$; (b) $\mathbb{E}\,(\varepsilon_{i,t} \mid \mathbf{X}) = 0$, $\forall i, t$ *and* (c) $\mathbb{C}\mathrm{ov}\,(\varepsilon_{i,s}, \varepsilon_{j,t} \mid \mathbf{X}) = \mathbb{1}\{i = j\}\mathbb{1}\{s = t\}\,\sigma^2$.

This assumption is a panel version of a distributed lag model.[2] For the macroeconomic applications that are the focus of the present paper, the independent variables typically consist of a macroeconomic shock and its lags, for instance, a monetary policy shock or an aggregate productivity shock, in which case (3.1) is a finite moving average representation and the vector $\boldsymbol{\beta}_i$ is the im-

[2] See, for instance, Greene (2003, chapter 19) or Baltagi (2008, chapter 6).

pulse response function of the i-th individual in the panel to that shock.[3] In that context, it is assumed that the researcher observes a panel containing the dependent variable of interest for different individuals over time and a time-series of the macroeconomic shock of interest.[4] The error term in (3.1) captures any other individual-specific factors that affect the dependent variable of interest and it is assumed those factors are mean independent of the macroeconomic shock and conditionally homoskedastic.[5] In order to ensure the estimators considered in this paper are well-defined, it is further assumed that the macroeconomic shock is not perfectly collinear with any of its lags.

In addition to assumption 3.1, the individual impulse response functions to the macroeconomic shock are assumed to follow a group pattern of the form,

***Assumption* 3.2** *The individual specific coefficients in* (3.1) *follow a group pattern of the form,*

$$\boldsymbol{\beta}_i = \sum_{k=1}^{K_0} \boldsymbol{\alpha}_k \mathbb{1}\left\{i \in G_k\right\} \tag{3.2}$$

where $\boldsymbol{\alpha}_j \neq \boldsymbol{\alpha}_k$ *for any* $j \neq k$, $\cup_{k=1}^{K_0} G_k = \{1, 2, \dots, N\}$ *and* $G_j \cap G_k = \varnothing$ *for any* $j \neq k$.

Assumption 3.2 imposes the same form of coefficient heterogeneity that is assumed in Su, Shi and Phillips (2016) and, in the present context, it can be

[3] There can be different data generating processes that can be represented by a moving average representation. For instance, it can arise from the inversion of a panel vector autoregression or an autoregressive distributed lag model. If that is the case, the implicit assumption in (3.1) is that the coefficients of the autoregressive component are such that that the process admits a linear moving average representation in which the moving average coefficients decay sufficiently fast to ensure that they can be truncated at a finite horizon H.

[4] Over the last three decades, the empirical macroeconomics literature has come up with a wide range of methods to identify macroeconomic shocks, including several identification schemes for structural VARs, narrative identification (e.g., Romer and Romer, 2004) and high-frequency identification (e.g., Gertler and Karadi, 2015). Different methods of identifying macroeconomic shocks are surveyed in Ramey (2016).

[5] For expositional purposes it is useful to focus on the parsimonious moving average representation in (3.1). Notice, however, that all the IRF estimators presented in this paper are either based on OLS or Penalized least squares the discussion could be extended to include additional independent variables in (3.1) including a constant, individual fixed-effects or time-fixed effects. In those cases, one would simply need to use residualized versions of the dependent and independent variables.

justified by the idea that there is some form of *group sparsity* in the way different individuals react to macroeconomic shocks. Put differently, assumption 3.2 represents a middle-ground between two extreme scenarios. One in which every single individual reacts in a different way to a macroeconomic shock and other where all individuals react in the same way to that shock. Instead, according to (3.2), individuals can belong to one among K_0 groups and the impulse responses differ *between* groups but are common across individuals *within* the same group.

In some aspects, assumption 3.2 is flexible since not only it nests the full heterogeneity and the no heterogeneity scenarios as special cases (when $K_0 = N$ and $K_0 = 1$, respectively) but also it does not impose any particular structure on the process that determines group membership and, hence, group membership can be an arbitrary function of both observable or unobservable variables. However, in other aspects assumption 3.2 is also restrictive, particularly, it imposes that impulse responses of each group do not vary over time and that individuals cannot switch between groups over time. It is crucial to notice that from the researcher's perspective both the true number of groups and which individuals belong to which group are *unknown* and, therefore, the problem of estimating individual impulse responses and understanding what drives their heterogeneity is equivalent to estimating the number of groups (K_0), the individual group membership ($\{G_1, \ldots, G_{K_0}\}$) and the group specific impulse responses ($\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{K_0}$).

3.2.2. Empirical applications

Before turning to the estimation of individual specific impulse responses in the presence of latent group heterogeneity, it is useful to review some heterogeneity analysis conducted in the existing literature both to illustrate some of the settings where the methods developed in this paper could be applied to and to understand what is the "common approach" in the literature to estimate disaggregated IRFs to a common aggregate shock.

Heterogeneous impulse responses to a monetary policy shock. There is an extensive list of heterogeneity analysis that have been conducted to investigate to what extent monetary policy shocks have heterogeneous effects across different individuals, firms or regions. For instance, Coibion, Gorodnichenko, Kueng and Silvia (2017) investigate the effects of monetary policy shocks on consumption of individuals depending on their relative position in the wealth distribution, Wong (2019) investigates whether monetary policy affects differently consumption expenditures of households depending on their age whilst Cloyne, Ferreira and Surico (2019) analyze the responses of consumption expenditures depending on whether the household is a home owner, a renter or a mortgagor. Moreover, some papers have investigated the responses of inflation perceived by different individuals in the population, for instance, Cravino, Lan and Levchenko (2020) investigate the effects of monetary policy on the inflation experienced by individuals in different percentiles of the income distribution whereas Clayton, Jaravel and Schaab (2018) find that monetary policy stabilizes sectors that matter relatively more for college-educated households. Differently, Bernanke and Gertler (1995) investigate the effects of monetary policy on different components of final demand, Carlino and Defina (1999) investigate the effects of monetary policy on the state-level economic activity across US states. Numerous papers have also looked at the effects of a monetary policy shoc across different types of firms. For instance, Gertler and Gilchrist (1994) find that small firms account for a significantly disproportionate share of the manufacturing declines that follows tightening of monetary policy, Kashyap, Lamont and Stein (1994) find that during the 1981-82 recession bank-dependent liquidity constrained firms cut their inventories by significantly more than their nonbank-dependent counterparts. More recently, Ottonello and Winberry (2020) find that the investment of firms with low default risk is the most responsive to monetary shocks, Jeenas (2019) finds that are the firms with fewer liquid assets that tend to reduce investment relative to others and Cloyne, Ferreira, Froemel and Surico (2020) find that younger firms paying no-dividends adjust both their capital expenditure and borrowing significantly more than older firms paying dividends in response to a monetary policy shock.

Other shocks. In the same spirit, heterogeneity analysis have been conducted to understand the reactions of different groups of individuals or firms to other aggregate shocks. In particular, Drechsel (2020) investigates the effects of an aggregate investment specific technology shock on firm level debt depending on whether firms tend to borrow against their collateral or future earnings. This application will be revisited in section 3.6.

3.3. The common approach to estimation of heterogeneous impulse responses

Even though existing heterogeneity analyses focus on different dimensions, methodologically they mostly follow an *ex-ante classification approach*, that is, they first group individuals according to some external classification or observable explanatory variables and then estimate and compare the resulting group specific impulse responses. This section analyses the properties of the estimator based on this approach and shows that in the presence of latent group heterogeneity there is, in general, a *bias-variance tradeoff* between this estimator and estimating individual-specific impulse responses.

3.3.1. Ex-ante classification and individual-specific impulse responses

To introduce the estimator based on the ex-ante classification approach, let a *grouping scheme* be denoted by $\tilde{\mathcal{G}}^K$ which stands for any collection of K non-empty sets satisfying $\cup_{k=1}^{K}\tilde{G}_k = \{1, 2, \ldots, N\}$ and $\tilde{G}_i \cap \tilde{G}_j = \varnothing$ for any $i \neq j$. In practice, the choice of individual group membership might be a function of other observable variables (*e.g.* the income distribution decile that the individual belongs, the household house-tenure status or whether a firm is a flow or a collateral borrower). Most importantly, $\tilde{\mathcal{G}}^K$ is a researcher's choice and both the number of groups chosen and the individual classification can differ from the true number of groups and group membership in assumption 3.2. Given a grouping scheme $\tilde{\mathcal{G}}^K$, the impulse response estimator based on the ex-ante classification approach is defined by,

$$\widetilde{\boldsymbol{\beta}}_i(\tilde{\mathcal{G}}^K) = \sum_{k=1}^{K} \widetilde{\boldsymbol{\alpha}}_k \mathbb{1}\left\{i \in \tilde{G}_k\right\} \tag{3.3}$$

where,

$$\left(\widetilde{\boldsymbol{\alpha}}_1(\tilde{\mathcal{G}}^K), \ldots, \widetilde{\boldsymbol{\alpha}}_K(\tilde{\mathcal{G}}^K)\right) = \underset{\mathbf{a}_1,\ldots,\mathbf{a}_K}{\arg\min}\ \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{i,t} - \mathbf{x}'_{i,t}\sum_{k=1}^{K}\mathbf{a}_k\mathbb{1}\{i \in \tilde{G}_k\}\right)^2 \tag{3.4}$$

The group estimates obtained from (3.4) are nothing more than ordinary least squares estimates obtained from pooling all the individuals in the panel and interacting the shock with a dummy variable for group membership. Once the group estimates are obtained, (3.3) uses the grouping scheme to assign impulse response estimates to each individual in the panel.

As a benchmark, it will be useful to consider the estimator of impulse responses that would be obtained if the researcher did not take a stance on the grouping scheme and instead allowed for complete heterogeneity in the individual responses to the shock. That estimator is defined by,

$$\left(\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_N\right) = \underset{\mathbf{b}_1,\ldots,\mathbf{b}_N}{\arg\min}\ \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{i,t} - \mathbf{x}'_{i,t}\mathbf{b}_i\right)^2 \tag{3.5}$$

where the individual-specific impulse response estimates obtained from (3.5) are the same one would obtain by estimating the moving average representation using the time-series for each individual in the panel separately.

3.3.2. A bias-variance tradeoff

In the presence of latent group heterogeneity the choice between the estimator based on the ex-ante classification approach and the estimator allowing for complete heterogeneity entails a bias-variance tradeoff. This tradeoff is formalised by the following proposition,

***Proposition* 3.1** *Suppose assumptions 3.1 and 3.2 hold. For a given* $\tilde{\mathcal{G}}^K$ *suppose* $i \in \tilde{G}_a \cap G_b$ *for some* $a \in \mathbb{Z}_{[1,K]}$ *and* $b \in \mathbb{Z}_{[1,K^0]}$. *Let* $\widetilde{\boldsymbol{\beta}}_i(\tilde{\mathcal{G}}^K)$ *denote the estimator obtained from* (3.3) *and* (3.4) *and* $\widehat{\boldsymbol{\beta}}_i$ *denote the estimator obtained from* (3.5). *Then,*

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}}_i\right) = \boldsymbol{\beta}_i \tag{3.6}$$

$$\mathbb{E}\left(\widetilde{\boldsymbol{\beta}}_i(\tilde{\mathcal{G}}^K)\right) = \boldsymbol{\varphi}_{a,b}\,\boldsymbol{\beta}_i + \sum_{\substack{k=1\\k\neq b}}^{K^0} \boldsymbol{\varphi}_{a,k}\,\boldsymbol{\alpha}_k \tag{3.7}$$

where $\boldsymbol{\varphi}_{a,b} \equiv \mathbb{E}\left(\left(\sum_{i\in\tilde{G}_a}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\sum_{i\in\tilde{G}_a\cap G_b}\mathbf{X}_i'\mathbf{X}_i\right)$. *Moreover, for any non-zero* $H+1$*-dimensional vector* $\mathbf{r}$ *it holds that,*

$$\mathbf{r}'\,\mathbb{V}\mathrm{ar}\left(\widehat{\boldsymbol{\beta}}_i \mid \mathbf{X}\right)\mathbf{r} \;\geqslant\; \mathbf{r}'\,\mathbb{V}\mathrm{ar}\left(\widetilde{\boldsymbol{\beta}}_i(\tilde{\mathcal{G}}^K) \mid \mathbf{X}\right)\mathbf{r} \tag{3.8}$$

Proof. See appendix 3.C. □

In simple terms, proposition 3.1 states that the researcher faces a fundamental tradeoff when deciding how to estimate individual impulse response functions in the presence of latent group heterogeneity. On the one hand, disaggregating too much could yield a set of estimated IRFs that are largely uninformative because the variability across individuals reflects not only the latent group heterogeneity but also a large share of sampling variability. On the other hand, grouping together individuals that do not share the same responses lead to biased impulse response estimates. This tradeoff is illustrated in figure 3.A.1. The estimation of fully heterogeneous impulse responses yields the “cloud” of grey-lines from which it is almost impossible to infer the true heterogeneity pattern, which comes from the fact that half of the individuals in the sample have their true impulse responses given by the solid-green line whereas the other half have their impulse response given by the solid-red line. On the other hand, if individuals were incorrectly grouped and the ex-ante classification approach was adopter then patterns of heterogeneity could be mistakenly inferred from the data. For example, if all individuals where grouped together would yield the wrong conclusion that the individuals true IRF for all the individuals in the sample is given by the black-dashed line and the “cloud” of grey-lines is the result of sampling variability and not the result of heterogeneity in the true impulse responses.

Misclassification bias. For a given individual i, from (3.7) there is only one case where $\widetilde{\boldsymbol{\beta}}_i(\tilde{\mathcal{G}}^K)$ is not biased: when all the individuals assigned to the same group as i by the researcher indeed belong to the same latent group as individual i. This implies that for the ex-ante classification approach to yield unbiased estimates for *all* individuals in the sample requires that the ex-ante grouping of individuals proposed by the researcher exactly matches the the true individual grouping in assumption 3.2. Whenever this is not the case, the estimator based on the ex-ante classification approach suffers from *misclassification bias.* Expression in (3.7) states that on average the impulse responses estimated for a given individual are equal to a matrix weighted average between the impulse response of the group that individual belongs to and the impulse responses of other groups. For the case where $x_{i,t}$ is an aggregate shock, expression (3.7) becomes,

$$\mathbb{E}\left(\widetilde{\boldsymbol{\beta}}_i(\tilde{\mathcal{G}}^K)\right) = \frac{N_{\tilde{G}_a \cap G_b}}{N_{\tilde{G}_a}}\boldsymbol{\beta}_i + \sum_{\substack{k=1 \\ k \neq b}}^{K^0} \frac{N_{\tilde{G}_a \cap G_k}}{N_{\tilde{G}_a}}\boldsymbol{\alpha}_k \tag{3.9}$$

where $N_{\tilde{G}_a}$ denotes the cardinality of the set $\tilde{G}_a$ and $N_{\tilde{G}_a \cap G_b}$ denotes the cardinality of the set $\tilde{G}_a \cap G_b$. In this case, the weight assigned to each latent group true impulse response is given by the share of individuals from that group that where assigned by the researcher to the same group as individual i.

Scope for efficiency gains. Given the ex-ante classification approach is prone to suffer from a misclassification bias, a natural question is whether the gains in precision obtained by ex-ante grouping of individuals are sufficiently large to outweigh the risk of ending up with biased estimates. According to (3.8) at any horizon considered the sampling variance of the IRFs obtained from the ex-ante classification approach have smaller (or equal) variance than their counterparts obtained from estimating fully heterogeneous IRFs. For the case where $x_{i,t}$ is an aggregate shock it can be shown that,

$$\mathbb{V}\mathrm{ar}\left(\widetilde{\boldsymbol{\beta}}_i(\tilde{\mathcal{G}}^K) \mid \mathbf{X}\right) = \frac{1}{N_{\tilde{G}_a}} \mathbb{V}\mathrm{ar}\left(\widehat{\boldsymbol{\beta}}_i \mid \mathbf{X}\right) \tag{3.10}$$

In other words, by grouping together ten individuals, a relatively small number relative to the typical cross-sectional dimension in datasets used to estimate heterogeneous impulse responses, the sampling variance of impulse responses decreases by 90% which suggest that, in general, the efficiency gains from pooling individuals together can be sizable.

3.4. Impulse response estimation via the classifier-Lasso

Motivated by the bias-variance tradeoff in proposition 3.1, this section introduces an alternative way to estimate group-specific impulse responses that is designed to eliminate this tradeoff *without* the requirement that the researcher correctly specifies ex-ante the group membership. The estimation is an application of the classifier-Lasso (C-Lasso) developed in Su, Shi and Phillips (2014, 2016) to estimate group-specific impulse response functions in the presence of latent heterogeneity. The fundamental insight underlying the C-Lasso is that it builds on penalized techniques to replace ex-ante classification of individuals into groups by a data-driven way of estimating both the individual group membership and the number of latent groups. This section briefly reviews the C-Lasso, showing how it can be applied to the estimation of heterogeneous impulse responses and its asymptotic properties.

3.4.1. Determination of individual group membership

Consider the problem of determining individual group membership taking the number of latent groups as given. First, define the following objective function (Su, Shi and Phillips, 2014, equation 2.4),

$$\mathcal{Q}^{(K)}_{NT,\lambda_1}\left(\mathbf{b},\mathbf{a}\right)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{i,t}-\mathbf{x}'_{i,t}\mathbf{b}_i\right)^2+\frac{\lambda_1}{N}\sum_{i=1}^{N}\prod_{k=1}^{K}\|\mathbf{b}_i-\mathbf{a}_k\| \tag{3.11}$$

where λ_1 is a tunning parameter that converges to zero as $(N,T)\to\infty$. Notice that the first term on the right-hand-side of (3.11) is exactly the same objective function that is used to obtain impulse response estimates under complete individual heterogeneity in (3.5). The second-term on the right-had-side of (3.11) is the distinctive feature of the C-Lasso and its mixed additive-multiplicative form shrinks the individual impulse responses $(\mathbf{b}_i)$ to a particular unknown group-level parameter vector $(\mathbf{a}_k)$. Minimising (3.11) with respect to $\mathbf{b}$ and $\mathbf{a}$ produces the C-Lasso estimates $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ which are henceforth denoted by $\widehat{\boldsymbol{\beta}}^{\text{C-Lasso}}$ and $\widehat{\boldsymbol{\alpha}}^{\text{C-Lasso}}$. Given this set of estimates, the C-Lasso group classifier is given by: $i\in\widehat{G}_k$ if $\widehat{\boldsymbol{\beta}}_i^{\text{C-Lasso}}=\widehat{\boldsymbol{\alpha}}_k^{\text{C-Lasso}}$ for some $k\in\mathbb{Z}_{[1,K]}$, otherwise, $i\in\widehat{G}_l$ for some $l\in\mathbb{Z}_{[1,K]}$ if $||\widehat{\boldsymbol{\beta}}_i^{\text{C-Lasso}}-\widehat{\boldsymbol{\alpha}}_l^{\text{C-Lasso}}||=\{\,||\widehat{\boldsymbol{\beta}}_i^{\text{C-Lasso}}-\widehat{\boldsymbol{\alpha}}_1^{\text{C-Lasso}}||\,,\ldots,\,||\widehat{\boldsymbol{\beta}}_i^{\text{C-Lasso}}-\widehat{\boldsymbol{\alpha}}_K^{\text{C-Lasso}}||\,\}$ and $\sum_{k=1}^{K}\mathbb{1}\{\widehat{\boldsymbol{\beta}}_i^{\text{C-Lasso}}-\widehat{\boldsymbol{\alpha}}_k^{\text{C-Lasso}}\}=0$.[6]

[6] This group classifier achieves in large samples the same properties as the simpler classification rule $\widehat{G}_k=\{i\in\mathbb{Z}_{[1,N]}:\widehat{\boldsymbol{\beta}}_i^{\text{C-Lasso}}=\widehat{\boldsymbol{\alpha}}_k^{\text{C-Lasso}}\}$ for $k\in\mathbb{Z}_{[1,K]}$. Nonetheless, the classifier in the text is preferred since it ensures that all the individuals are classified into one of the K groups in finite samples (see Su, Shi and Phillips, 2016, remark 2).

3.4.2. Determination of the number of groups

The C-Lasso estimates from minimising (3.11) are obtained for a given number of groups (K). In practice, however, the true number of latent groups is unknown and has to be estimated along with the group membership. Following Su, Shi and Phillips (2014, 2016) it is assumed that the true number of groups is bounded from above by a finite integer $K_{\max}$ and the number of groups is estimated through an information criterion (IC).[7] Making the dependence on K and λ_1 explicit, the group classification implied by the C-Lasso can be written as $\widehat{G}(K, \lambda_1) = \left\{ \widehat{G}_1(K, \lambda_1), \ldots, \widehat{G}_K(K, \lambda_1) \right\}$. The information criterion used to determine the number of latent groups is given by Su, Shi and Phillips (2016, equation 2.10),

$$\mathrm{IC}(K, \lambda_1) = \ln\left(\hat{\sigma}^2_{\widehat{G}(K,\lambda_1)}\right) + \rho_{NT}\,(H+1)\,K \tag{3.12}$$

where ρ_{NT} is a tuning parameter and $\hat{\sigma}^2_{\widehat{G}(K,\lambda_1)} = \frac{1}{NT}\sum_{k=1}^{K}\sum_{i\in\widehat{G}(K,\lambda_1)}\sum_{t=1}^{T}\left(y_{i,t} - \mathbf{x}'_{i,t}\widehat{\boldsymbol{\alpha}}_{\widehat{G}_k}\right)^2$ with $\widehat{\boldsymbol{\alpha}}_{\widehat{G}_k} = \left(\sum_{i\in\widehat{G}_k}\sum_{t=1}^{T}\mathbf{x}_{i,t}\mathbf{x}'_{i,t}\right)^{-1}\left(\sum_{i\in\widehat{G}_k}\sum_{t=1}^{T}\mathbf{x}_{i,t}y_{i,t}\right)$. Finally, for a given value of the tunning parameter λ_1, the number of groups is chosen such that the IC in (3.12) is minimized, that is, $\hat{K}(\lambda_1) = \arg\min_{1\leqslant k\leqslant K_{max}} \mathrm{IC}(k, \lambda_1)$.

3.4.3. Post-Lasso impulse responses

Given the estimated group classification based on the C-Lasso this paper focuses on the *post-Lasso* estimates of the impulse responses. For a given group $\widehat{G}_k$ the post-Lasso group impulse response estimates are given by $\widehat{\boldsymbol{\alpha}}_{\widehat{G}_k}$ and the post-Lasso individual impulse responses are given by $\widehat{\boldsymbol{\beta}}_{i,\widehat{G}_k} = \sum_{k=1}^{K}\widehat{\boldsymbol{\alpha}}_{\widehat{G}_k}\mathbb{1}\left\{i \in \widehat{G}_k\right\}$.

3.4.4. Asymptotic properties

The asymptotic properties of the C-Lasso in the context of linear models are formally shown in Su, Shi and Phillips (2014, 2016). Under a suitable set of assump-

[7] An alternative way of determining the number of latent groups is to use the residual-based Lagrange multiplier-type test proposed by Lu and Su (2017).

tions, the authors show that: (i) the classifier proposed in section 3.4.1 is *uniformly consistent* which, in simple terms, means that the proposed C-Lasso group classifier classifies each individual to the correct group with probability approaching 1 as $(N, T) \to \infty$ (see Su, Shi and Phillips, 2016, theorem 2.2); (ii) the selector criterion for K proposed in section 3.4.2 is such that $\mathbb{P}\left(\hat{K}(\lambda_1) = K_0\right) \to 1$ as $(N, T) \to \infty$ (see Su, Shi and Phillips, 2016, theorem 2.6) and (iii) the post-Lasso estimator of $\boldsymbol{\alpha}_k$ defined in section 3.4.3 enjoy the *asymptotic oracle property*, which means that as $(N, T) \to \infty$ it achieves the same limiting distribution as the *oracle estimator* which is the group IRF estimator one would obtain if the true group membership was known (see Su, Shi and Phillips, 2016, theorem 2.5).

3.5. Monte Carlo Experiment

This section uses a Monte Carlo experiment to inspect the finite sample performance of the classification and estimation procedure introduced in section 3.4 when applied to estimate group-specific impulse responses in the presence of latent group heterogeneity.

3.5.1. Data generating process

Each Monte Carlo sample consist of consists of a panel data $\{(y_{i,t}, \mathbf{x}_t)\}$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$ that is generated according to,

$$y_{i,t} = \sum_{h=0}^{12} x_{t-h}\beta_{i,h} + \varepsilon_{i,t} \tag{3.13}$$

where x_t is an aggregate shock such that $x_t \sim \mathcal{N}(0, 1)$ and i.i.d. across t, the idiosyncratic shocks $\varepsilon_{i,t} \sim \mathcal{N}(0, 1)$ are i.i.d. across i and t and x_t and $\varepsilon_{i,t}$ are mutually independent. There are two latent groups $(K_0 = 2)$ and the group-specific impulse responses are parametrised using a Gaussian basis function as in

Barnichon and Mathes (2018). In particular,

$$\beta_{i,h} = \begin{cases} 0.15 \times \exp\left\{-\left(\frac{h-4}{25}\right)^2\right\}, \text{if } i \in G_1 \\ -0.15 \times \exp\left\{-\left(\frac{h-4}{25}\right)^2\right\}, \text{if } i \in G_2 \end{cases} \tag{3.14}$$

which results in the symmetric impulse responses for groups 1 and 2 depicted in figure 3.A.1. Individuals are assigned to group 1 if they are indexed by an odd number and assigned to group 2 if they are indexed by an even number so that, for each sample generated, half of the individuals belong to each group. Sample sizes of size $N = \{100, 200\}$ and time spans $T = \{40, 80\}$ are considered.[8] For each possible combination of N and T, 250 Monte Carlo samples are generated.

3.5.2. Estimation and Classification

For each Monte Carlo sample generated, two alternative ways of estimating the impulse responses are considered. The first one is by focusing directly on the moving average representation and, in that case, the C-Lasso objective function is given by (3.11) with $\mathbf{x}'_{i,t} = [x_t, x_{t-1}, \dots, x_{t-H}]$. The second way of estimating impulse responses is trough the use of local projections (Jordà, 2005) which use a sequence of regressions of $y_{i,t}$ on $x_{i,t-h}$ to estimate the impulse response for individual i at horizon h. The case of local projections can be accommodated in the C-Lasso framework described in section 3.4 by replacing (3.11) by the following modified C-Lasso objective function,

$$\widetilde{Q}^{(K)}_{NTH,\lambda_1}(\mathbf{b}, \mathbf{a}) = \frac{1}{N(H+1)T} \sum_{i=1}^{N} \sum_{h=0}^{H} \sum_{t=1}^{T} (y_{i,t} - x_{i,t-h}\mathrm{b}_{h+1,i})^2 + \frac{\tilde{\lambda}_1}{N} \sum_{i=1}^{N} \prod_{k=1}^{K} \|\mathbf{b}_i - \mathbf{a}_k\| \tag{3.15}$$

where $\mathrm{b}_{h+1,i}$ is the $(h+1, i)$-*th* element of the matrix $\mathbf{b}$ and $\tilde{\lambda}_1$ is a tunning

[8] Even a value of $T = 80$ is still relatively small compared to what is typically used in the literature using local projections to estimate impulse responses. As documented by Herbst and Johannsen (2020) the median value for T across the 100 "most relevant" papers citing Jordà (2005) in Google scholar is around 95.

parameter that tends to zero as $(N, T) \to \infty$.[9] The local projections C-Lasso estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are obtained by minimising (3.15) with respect to $\mathbf{b}$ and $\mathbf{a}$. Given these estimates the individual classification, determination of number of groups and post-Lasso estimates are obtained in an analogous way as described in sections 3.4.1 to 3.4.3, except the tunning parameters that are adjusted to reflect the effective number of observations per cross-sectional unit that is different for the local projections case.

Estimation of impulse responses through local projections has become increasingly popular over the last decade. Among the advantages of local projections cited in Jordà (2005) are their flexibility and the fact that they are more robust to misspecification of the moving average representation if it arises from the inversion of a misspecified vector autoregression. Notice, however, that in the present Monte Carlo experiment the moving average representation is *not* misspecified and, hence, it is not expected that the impulse responses estimated from local projections to display better statistical properties than those estimated directly from the moving average representation. The purpose of including impulse response estimation through local projections in the present exercise is simply to illustrate that they can be accommodated by the C-Lasso framework.

Tuning parameters. Determination of the number of groups and individual classification requires the researcher to specify the tuning parameters λ_1 in (3.11) and ρ_{NT} in (3.12). The assumptions on λ_1 and ρ_{NT} needed to derive the asymptotic properties highlighted in section 3.4.4 are satisfied for any λ_1 such that $\lambda_1 \propto T^{-a}$ for any $a \in (0, -1/2)$ and any ρ_{NT} that can be written as $\rho_{NT} \propto (NT)^{-b}$ for any $b \in (0, 1)$. Even though asymptotically the choice of the tuning parameters is irrelevant as long as they satisfy these conditions, in finite samples their choice can be crucial. In this Monte Carlo experiment the values of the tuning

[9] The modification needed to estimate impulse responses through local projections using the C-Lasso framework is more easily seen in matrix form. Let $\mathbf{X}_i$ be defined as in assumption 1.1 and $\mathbf{y}_i = [y_{i,1}, \ldots, y_{i,T}]'$. The first term on the right-hand-side of (3.11) is given by $\frac{1}{NT} \sum_{i=1}^{N} (\mathbf{y}_i - \mathbf{X}_i \mathbf{b}_i)' (\mathbf{y}_i - \mathbf{X}_i \mathbf{b}_i)$. Estimation through local projections simply requires replacing this term by $\frac{1}{N\tilde{T}} \sum_{i=1}^{N} \left(\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \mathbf{b}_i\right)' \left(\tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i \mathbf{b}_i\right)$ where $\tilde{T} = (H+1)T$, $\tilde{\mathbf{y}}_i = \mathbf{1}_{N \times 1} \otimes \mathbf{y}_i$ and $\tilde{\mathbf{X}}_i = \oplus_{j=1}^{H+1} \mathbf{x}_{*,j}$ and $\mathbf{x}_{*,j}$ denotes the j-*th* column of $\mathbf{X}_i$.

parameters used to estimate impulse responses through the moving average representation are $\lambda_1 = c_\lambda s_Y^2 T^{-1/3}$ and $\rho_{NT} = \frac{2}{3}(NT)^{-\frac{2}{3}}$ where s_Y^2 denotes the sample variance of $y_{i,t}$ and c_λ is set equal to 2.[10,11] For the estimation via local projections the tuning parameters are adjusted to reflect the effective number of observations per cross-sectional unit, that is, $\tilde{\lambda}_1 = \tilde{c}_\lambda s_Y^2 \tilde{T}^{-1/3}$ and $\tilde{\rho}_{NT} = \frac{2}{3}(N\tilde{T})^{-\frac{2}{3}}$ where $\tilde{c}_\lambda$ is equal to 2 and $\tilde{T} = (H+1)T$.

3.5.3. Monte Carlo Results

The results from the Monte Carlo experiment are reported in tables 3.B.1 to 3.B.3 and in figures 3.A.2 and 3.A.3. They can be summarised as follows:

Determination of the number of latent groups. For each DGP considered, table 3.B.1 shows the frequency that different number of groups is chosen across Monte Carlo replications. When the number of groups is based on the estimation of the moving average representation the IC-based group determination procedure always identifies the correct number of latent groups except in 1% percent of the samples for $N = 100$ and $T = 40$ where the IC picks one latent group. When the number of groups is based on the estimation of local projections the performance of the group determination procedure deteriorates, specially for the two DGPs with $T = 40$ where the IC selects one latent group more often than two latent groups. As expected, as N and T increase the frequency that the true number of latent groups increases and, in particular, for $N = 200$ and $T = 80$ the IC selects the correct number of latent groups in 82% of the Monte Carlo samples.

[10] Su, Shi and Phillips (2016) use $c_\lambda \in [0.125, 0.25, 0.5, 1, 2]$ and select the value of c_λ *jointly* with the number of latent groups k to minimise the information criterion in (3.12). For a small number of Monte Carlo replications it was found that jointly determining c_λ with the number of groups did not affect the results whilst substantially increasing the computational costs. For this reason, in each Monte Carlo replication c_λ is kept fixed equal to 2. In the empirical application in section 3.6, the value of c_λ is grid search over the same grid used in Su, Shi and Phillips (2016) and jointly chosen with the number of groups to minimize the information criterion.

[11] For linear models Su, Shi and Phillips (2016) use $\rho_{NT} = \frac{2}{3}(NT)^{-\frac{1}{2}}$. In numerical experiments, for the DGP here considered I found this value for the tuning parameter tends to over-select number of groups that is *smaller* than the true number of groups. I have experimented for values $\rho_{NT} = c_1(NT)^{-c_2}$ for $c_1, c_2 \in (0, 1)$ and found that $\rho_{NT} = \frac{2}{3}(NT)^{-\frac{2}{3}}$ tends to select the correct number of groups with a higher frequency.

Individual classification. The average individual misclassification rates across Monte Carlo replications for each DGP is reported in figures 3.A.2 and 3.A.3. This figure consists of the share of individuals that are assigned to a group they do *not* belong averaged across Monte Carlo samples. From figure 3.A.2, when impulse responses are estimated through the moving average representation this figure is under 4% for $T = 40$ and under 1% for $T = 80$ which is suggestive that the classifier proposed in 3.4.1 tends to classify individuals to the correct group in finite samples too. From figure 3.A.3, when estimating impulse responses through local projections are of the order of 30% for $T = 40\%$ and of the order of 10% when $T = 80$. This inferior performance for the local projections case is justified by the fact that coefficient estimates from local projection regressions have much higher sampling variability since the error term in those regressions includes not only the original error term from the moving average representation ($\varepsilon_{i,t}$) but also all the other leads and lags that are not included.[12] With higher sampling variability the estimated individual impulse responses for individuals from group 1 (group 2) can end up being closer to the impulse response from group 2 (group 1), and in those cases when applying the classifier leads that individual to be assigned to the wrong group.

Post-Lasso impulse responses. The estimated post-Lasso impulse responses are plotted against the true group impulse responses in figures 3.A.2 and 3.A.3. For the case where impulse responses are estimated directly through the moving average representation (figure 3.A.2) the estimated impulse responses almost overlap the true impulse responses which indicates the absence of bias. For the case where $T = 40$ there is some small discrepancies that can be justified by the slightly higher misclassification rate than for the $T = 80$ case. For the case of impulse responses estimated via local projections (figure 3.A.3), there is some bias specially for $T = 40$ where the misclassification rate is of the order of 30%,

[12] Notice that the true data generating process is given by $y_{i,t} = \sum_{h=0}^{12} x_{t-h}\beta_{i,h} + \varepsilon_{i,t}$. For a given horizon h, the local projection regression is given by $y_{i,t} = x_{t-h}\beta_{i,h} + \tilde{\varepsilon}_{i,t}$ where $\tilde{\varepsilon}_{i,t} = \varepsilon_{i,t} + \sum_{j\neq h}^{12} x_{t-j}\beta_{i,j}$. Since the aggregate shocks are mean zero and iid, the fact that they are omitted does *not* cause bias or inconsistency in the local projection estimates of impulse responses but it *does* increase their sampling variance vis-a-vis the estimates obtained through the estimation of the moving average representation.

however, and as expected from (3.9), the bias substantially decreases for the $(N, T) = (200, 80)$ case when the average misclassification rate drops to 7%. Moreover, the difference between the 90th and the 10th percentiles of the sampling distribution is always smaller in figure 3.A.2 than in 3.A.3 which echoes the fact that the sampling variance of the impulse responses estimated via local projections is higher than those estimated directly through the moving average representation (see footnote [12]). The analysis of figures 3.A.2 and 3.A.3 is complemented with the figures in tables 3.B.2 and 3.B.3 that compare bias, variance and mean squared error for the impulse response estimates for the two groups at horizon $h = 4$ (*i.e.* the peak of the impulse responses). The figures in both tables numerically illustrate the theoretical results from proposition 3.1. The full heterogeneity estimator has essentially no bias but a higher variance than the post-Lasso estimator, whilst the post-Lasso estimator has some bias, since in finite samples it does not achieve perfect classification of individuals into groups, but a smaller variance. Most importantly, in MSE terms the post-Lasso estimator is always preferred to the full heterogeneity one. In particular, for the estimation through the moving average representation the decrease in MSE of the post-Lasso estimator vis-à-vis the full heterogeneity over one full order of magnitude. For the estimation through local projection the MSE of the post-Lasso estimator is one to two thirds smaller than the MSE of the full heterogeneity estimator.

3.6. Aggregate IST shocks and Firm level debt revisited

This section uses the C-Lasso classification and estimation procedure introduced in section 3.4 to revisit the estimation of the IRFs of firm-level debt to an aggregate investment specific shock originally studied by Drechsel (2020).

3.6.1. Background

Motivated by microeconomic evidence on corporate borrowing in the US that unveals a direct connection between firms' current earnings and their access to debt, Drechsel (2020) studies the macroeconomic implications of the so-called *earnings-based borrowing constraints.* First, in a prototype business cycle model the author shows that depending on the type of borrowing constraint used firm-level debt responds differently to a permanent investment shock. More precisely, in a setting where firms face a standard collateral constraint their debt decreases in response to a positive investment investment shock whereas if they face an earnings constraint their debt increases following that same shock (see Drechsel, 2020, figure 2).

To empirically test this model prediction, the author uses a time-series of investment specific technology (IST) shocks identified from an SVAR using long-run restrictions combined with a firm-level panel containing firm characteristics and information on the types of covenants included in their debt contracts.[13] The baseline specification to test the model predictions is given by,

$$\begin{aligned}\log\left(b_{i,t+h}\right) = \alpha_h + \beta_h \hat{u}_{IST,t} + \boldsymbol{\gamma}\boldsymbol{X_{i,t-1}} \\ + \beta_h^{earn}\mathbb{1}_{i,t,earn} \times \hat{u}_{IST,t} + \alpha_h^{earn}\mathbb{1}_{i,t,earn} \\ + \beta_h^{coll}\mathbb{1}_{i,t,earn} \times \hat{u}_{IST,t} + \alpha_h^{coll}\mathbb{1}_{i,t,coll} + \delta t + \eta_{i,t+h}\end{aligned} \tag{3.16}$$

where $b_{i,t}$ is the quarterly level of firms' debt liabilities, $\mathbb{1}_{i,t,earn}$ and $\mathbb{1}_{i,t,coll}$ are dummy variables that capture whether the firm is subject to earnings-related covenants or uses collateral, $\hat{u}_{IST,t}$ is the IST shock identified based on long-run restrictions, t is a linear time trend and $\boldsymbol{X_{i,t-1}}$ is vector of controls that includes $\log\left(b_{i,t-1}\right)$, 3-digit industry-level fixed effects, firm size, firm-level real

[13] This panel is obtained by merging the Dealscan dataset with Compustat data. For more details on the construction of this dataset and which variables are used in which specification refer to section 4.3.2 in Drechsel (2020).

sales growth and a variable constructed from SVAR residuals that is meant to capture macroeconomic shocks other than investment shocks.[14] At a given horizon h, the impulse response of an "earnings-based borrower" ("collateral-based borrower") is given by the sum of the coefficients $\beta_h + \beta_h^{earn}$ ($\beta_h + \beta_h^{coll}$) and, hence, in terms of regression coefficients the model predictions to be tested are $\beta_h + \beta_h^{earn} > 0$ and $\beta_h + \beta_h^{coll} < 0$.

The results based on (3.16) are presented in Drechsel (2020, figure 7) and are largely in line with the model implied impulse responses. In the data, across a wide range of specifications, debt of earnings-based borrowers reacts positively to an IST shock whereas debt for collateral-based borrowers declines in response to that same shock.

3.6.2. Revisiting the responses of debt to IST shock

Following the bulk of the existing literature that has conducted heterogeneity analysis on impulse response functions, Drechsel (2020) adopts the *ex-ante classification approach* described in section 3.3. Following the analysis of micro data on firm-level debt issuances and theoretically grounded by model predictions, the author groups firms depending on whether they are "earnings-based borrowers" or "collateral-based borrowers" and estimates the impulse responses to an IST shocks for each of these two groups of firms based on (3.16). In light of proposition 3.1, this approach can be subject a misclassification bias if the ex-ante grouping of firms, based on whether their borrow against collateral or earnings, does not coincide with the true grouping that underlies the heterogeneity firm-level impulse responses observed in the data. To investigate whether this is the case this section re-estimates the impulse responses of firm level debt to an IST shock using the C-Lasso classification and estimation procedure proposed in section 3.4. In

[14] More specifically, $\mathbb{1}_{i,t,earn}$ is equal to 1 if a given firm issues a loan with at least one earnings covenant and $\mathbb{1}_{i,t,earn}$ is equal to one if the debt issued by the firm is secured by specific assets.

order to do so, the following moving average version of (3.16) is considered,

$$\tilde{b}_{i,t} = \sum_{h=0}^{H} \beta_{i,h} \tilde{\hat{u}}_{IST,t} + \epsilon_{i,t} \tag{3.17}$$

where $\tilde{b}_{i,t}$ denotes the residuals of a regression of $\log(b_{i,t})$ on a constant and a linear time trend and, similarly, $\tilde{\hat{u}}_{IST,t}$ denotes the residuals of $\hat{u}_{IST,t}$ on a constant and a linear time trend. The identified IST shock is assumed to be strictly exogenous. Since all the theoretical results for the C-Lasso are derived for a balanced panel and all the empirical applications in Su, Shi and Phillips (2016) are also focused on balanced panels, specification (3.17) is estimated using the C-Lasso approach based on a balanced version of the Drechsel (2020) dataset. The final balanced panel contains 746 firms and 76 quarters spanning the period from 1997Q1 to 2015Q4.[15]

3.6.3. Results

Number of latent groups and post-Lasso IRFs. As illustrated in figure 3.A.4, the IC-based group determination procedure identifies two-latent groups. The post-Lasso IRF estimates for each of these groups along with firm-specific IRFs and the IRF estimated by pooling all the firms are depicted in figure 3.A.5. In line with the theoretical predictions in Drechsel (2020), one of the two latent groups responds *positively* to an IST shocks whilst the second group responds *negatively*. In addition, two points are also worth noting from figure 3.A.5. First, there is significant heterogeneity among individual-specific impulse responses and without any ex-ante theoretical reason to group individuals it would be difficult to identify the group pattern identified by the C-Lasso just by looking at the cloud formed by estimated individual IRFs. Second, despite the differences in the specifications (3.16) and (3.17) and the sample composition, it is reassuring

[15] To reach this balanced panel from the original Drechsel (2020) dataset, I first exclude the 12 periods of the dataset that are lost due to the lagging of the IST shock then, on the resulting sample, I remove all the firms that have missing values of debt for any quarter between 1997Q1 to 2015Q4.

to see that the pooled IRF in figure 3.A.5 has the same shape as the pooled impulse responses depicted in figure 6 of Drechsel (2020), in which the response of debt is increasing up until 2 years after the shock and then starts declining.[16]

Firm characteristics across the two groups. Despite the two latent groups and their associated responses being in line with predictions from Drechsel (2020), the most important aspect to be tested is whether indeed the group that responds positively to an IST shock contains a disproportionately higher share of earnings-based borrowers vis-à-vis the group that responds negatively. In total there are 746 firms, 224 of which where classified into the "positive response group" (group 1) whereas the remaining 522 where classified into the "negative response group" (group 2). Table 3.B.4 looks at four different firm characteristics across these two groups. Panel A looks at the relative proportions of earnings and collateral borrowers across the two groups. To compute these proportions a firm is classified as earnings-based borrower if over the whole sample it has more earnings-based debt issuances and it is classified as a collateral-based borrower if over the whole sample it has more collateral-based debt issuances than earnings based ones.[17] The majority of firms in both groups are earnings-based borrowers, however, in line in the theoretical predictions in Drechsel (2020) the share of earnings-borrowers is *larger* in the group that responds positively to an IST shock although the difference is not statistically significant. Panel B looks at the two measures of the share of the intangible assets as a share of total assets. Theoretically, one would expect firms that have larger share of intangible assets to borrow more against earnings since intangible assets cannot be used as collateral and, hence,

[16] Despite the similar shape of the IRFs there are some differences in terms of magnitudes. The on impact response from the pooled specification in Drechsel (2020) is zero and the peak of the response, that occurs 7 quarters after the shock, is around 2.5%. In figure 3.A.5, the on impact response of debt is roughly -2% whereas the peak response is almost 6%.

[17] Notice this criteria is slightly different than the one used in Drechsel (2020) since the dummies $\mathbb{1}_{i,t,earn}$ and $\mathbb{1}_{i,t,coll}$ are only defined for quarters where a debt issuance for firm i appears in the Dealscan dataset. These dummies are defined relative to a specific debt issuance and, hence, they can vary over time (*e.g.* a given firm can very well issue debt against collateral in a given date and issue another debt contract with earnings covenants in other date). To compute the shares reported in Panel A of table 3.B.4 a firm is classified as an earnings-based borrower if $\sum_{t=1}^{T} \mathbb{1}_{i,t,earn} > \sum_{t=1}^{T} \mathbb{1}_{i,t,coll}$ and classified as a collateral-based borrower if $\sum_{t=1}^{T} \mathbb{1}_{i,t,earn} < \sum_{t=1}^{T} \mathbb{1}_{i,t,coll}$. Otherwise, if $\sum_{t=1}^{T} \mathbb{1}_{i,t,earn} = \sum_{t=1}^{T} \mathbb{1}_{i,t,coll}$ the firm is not classified.

the larger the share of intangible assets the more likely the firm should be to respond positively to an IST shock. Qualitatively this is indeed the case, as the average of both measures of intangibility are higher for group 1, yet the difference is quantitatively small and not statistically significant. Panel C looks at three different measures of firm size as in Dang, Li and Yang (2018). In theory, one would expect smaller and younger firms to have less collateral to pledge and, hence, to borrow more against their future earnings. Again this is confirmed in the data, since for the three measures considered the average firm size is smaller in the group 1 and the difference is statistically significant at the 10% level when size is measured by firm's total assets. Finally, panel D looks at sectorial composition of each of the two groups. In this respect, group 1 has a statistically significant higher share of firms in the materials and industrial sectors whereas group 2 has a significantly higher share of firms in the consumer staples and utilities sectors.

Explaining group membership. To estimate the impact of each of the variables in table 3.B.4 on the probability that a given firm belongs to each of the two groups, a Logit model is estimated using as dependent variable a dummy that is equal to 1 if the firm belongs to group 1. The average marginal effects of each variable across different specifications are reported in table 3.B.5. The results in this table when including each group of explanatory variables at a time (columns 1 to 4) largely corroborate the conclusions based on the analysis of group characteristics in table 3.B.4. The specification in column 5 includes all the covariates at the same time. In this specification the effects of the first three variables are quantitatively small and not statistically significant. The only two statistically significant covariates are the dummies for the consumer staples and utilities sectors. In particular, a firm that is classified as consumer staples (utilities) is, on average, 40 p.p. (21 p.p.) *less* likely to belong to the group for which debt increases following an IST shock. In summary, the group that responds positively to an IST shock is composed by firms that: (*i*) that are relatively smaller; (*ii*) have higher share of *intangible assets*, (*iii*) tend to be *earnings-based borrowers* and (*iv*) do *not* belong to the *consumer staples* or *utilities* sectors.[18] This findings are largely in line with the theoretical predictions in Drechsel (2020), but also suggest that, on top of whether a firm tends to borrow against earnings or collateral, the specific *sector* that the firm operates also plays a role in determining whether it will respond positively or negatively to an aggregate IST shock.[19]

3.7. Concluding remarks and future research

This paper studied the estimation of heterogeneous impulse responses in the presence of latent group heterogeneity. It showed that the common approach to estimate disaggregated IRFs based on the ex-ante grouping of individuals according to some external criteria or observable explanatory variables can lead to misleading conclusions. More precisely, the choice between an estimator of group-specific impulse responses based on an ex-ante grouping of individuals and estimating individual-specific impulse responses entails a bias-variance tradeoff. Motivated by this tradeoff, this paper proposed an alternative methodology based on the C-Lasso to estimate group-specific impulse responses. A Monte Carlo experiment demonstrated good finite-sample performance of this methodology both in classification of individuals into different groups and estimation of group specific impulse responses. An application of this methodology to study firm level debt responses to an aggregate IST shock based on Drechsel (2020) identified two latent groups. One group of firms for which firm-level debt responds positively to an IST shock and other for which the response is negative. The group of

[18] The type of exercise explaining group membership is in spirit to a principal component analysis where the data alone selects the orthogonal factors that explain the correlations observed in the data and ex-post the researcher searches for appropriate names for these factors. In the present context, the C-Lasso approach selects the number of groups, the individual classification and the group-specific IRFs based solely on the data and it is based on the analysis of individual characteristics across characteristics that a name for each group is determined.

[19] In addition to the specification in (3.17), an alternative specification based on local projections is also estimated. In that specification, the IC-based group determination procedure identifies only one latent group. Given the superior performance of the moving average representation in identifying the correct number of groups in the Monte Carlo experiment, the discussion in the main text focuses on the moving average representation. Nonetheless, the estimates of the local projection specification *conditional* on two latent groups yields post-Lasso IRFs that are similar to the post-Lasso IRFs depicted in figure 3.A.5. Moreover, the group classification based on the local projection specification has an overlap of 91% with the individual classification based on the moving average representation.

firms for which debt increases in response to a positive IST shock is composed by firms that are relatively *smaller*, have higher share of *intangible assets*, tend to be *earnings-based borrowers* and do *not* belong to the *consumer staples* or *utilities* sectors.

I conclude by highlighting two dimensions along which the results from the present paper can be extended and applied. On the methodological side, the methodology proposed could be extended to include the possibility that the shocks are used as instruments and a more thorough Monte Carlo study could be conducted to search for values of the fine tuning parameter that improve finite-sample performance in terms of group determination when using local projections. On the applications front, the methodology here introduced could be used to either test alternative transmission channels from aggregate shocks to the cross-section (in the spirit of the application in section 3.6) or to identify the most important dimensions that drive heterogeneous responses to aggregate shocks and use that information to inform the theoretical modeling of DSGE models featuring heterogeneity across firms and/or households.

3.8. References

Baltagi, B. H. (2008). *Econometrics*. Springer, Berlin, Heidelberg, 4th edn.

Barnichon, R. and Mathes, C. (2018). Functional approximation of impulse responses. *Journal of Monetary Economics*, **99**, 41–55.

Bernanke, B. S. and Gertler, M. (1995). Inside the black box: The credit channel of monetary policy transmission. *Journal of Economic Perspectives*, **9** (4), 27–48.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, **83** (3), 1147–1184.

Browning, M. and Carro, J. M. (2014). Dynamic binary outcome models with maximal heterogeneity. *Journal of Econometrics*, **178** (2), 805–823.

Carlino, G. and Defina, R. (1999). The differential regional effects of monetary policy: Evidence from the u.s. states. *Journal of Regional Science*, **39** (2), 339–358.

Clayton, C., Jaravel, X. and Schaab, A. (2018). Heterogeneous price rigidities and monetary policy.

Cloyne, J., Ferreira, C., Froemel, M. and Surico, P. (2020). Monetary policy, corporate finance and investment.

—, — and Surico, P. (2019). Monetary policy when households have debt: New evidence on the transmission mechanism. *Review of Economic Studies*, **87**, 102–129.

Coibion, O., Gorodnichenko, Y., Kueng, L. and Silvia, J. (2017). Innocent bystanders? monetary policy and inequality. *Journal of Monetary Economics*, **88**, 70–89.

Cravino, J., Lan, T. and Levchenko, A. A. (2020). Price stickiness along the income distribution and the effects of monetary policy. *Journal of Monetary Economics*, **110**, 19–32.

Dang, C., Li, Z. and Yang, C. (2018). Measuring firm size in empirical corporate finance. *Journal of Banking and Finance*, **86**, 159–176.

Drechsel, T. (2020). Earnings-based borrowing constraints and macroeconomic fluctuations, working Paper.

Frisch, R. (1933). Propagation problems and impulse problems in dynamic economics. In *Economic Essays in Honor of Gustav Cassel*, London: Allen & Unwin, pp. 171–205.

Gertler, M. and Gilchrist, S. (1994). Monetary policy, business cycles, and the behavior of small manufacturing firms. *Quarterly Journal of Economics*, **109** (2), 309–340.

— and Karadi, P. (2015). Monetary policy surprises, credit costs, and economic activity. *American Economic Journal: Macroeconomics*, **7** (1), 44–76.

Greene, W. H. (2003). *Econometric Analysis*. Upper Saddle River, New Jersey 07458: Prentice Hall, 5th edn.

Hayashi, F. (2000). *Econometrics*. Princeton University Press, 1st edn.

Herbst, E. P. and Johannsen, B. K. (2020). Bias in local projections, finance and Economics Discussion Series 2020-010. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2020.010.

Hsiao, C. (2014). *Analysis of Panel Data*, vol. Econometric Society Monographs. Cambridge University Press.

Jeenas, P. (2019). Firm balance sheet liquidity, monetary policy shocks, and investment dynamics.

Jordà, O. (2005). Estimation and inference of impulse responses by local projections. *American Economic Review*, **95** (1), 161–182.

Kasahara, H. and Shimotsu, K. (2009). Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica*, **77** (1), 135–175.

Kashyap, A., Lamont, O. and Stein, J. (1994). Credit conditions and the cyclical behavior of inventories. *Quarterly Journal of Economics*, **109** (3), 565–592.

Lewis, D., Melcangi, D. and Pilossoph, L. (2019). Latent heterogeneity in the marginal propensity to consume.

Lin, C.-C. and Ng, S. (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods*, **1** (1).

Lu, X. and Su, L. (2017). Determining the number of groups in latent panel structures with an application to income and democracy. *Quantitative Economics*, **8**, 729–760.

Ottonello, P. and Winberry, T. (2020). Financial heterogeneity and the investment channel of monetary policy.

Ramey, V. A. (2016). Macroeconomic shocks and their propagation. In *Handbook of Macroeconomics*, vol. 2, *2*, Elsevier.

Romer, C. and Romer, D. (2004). A new measure of monetary shocks: Derivation and implications. *American Economic Review*, **94** (4), 1055–1084.

Sarafidis, V. and Weber, N. (2015). A partially heterogeneous framework for analyzing panel data. *Oxford Bulletin of Economics and Statistics*, **77** (2), 274–296.

Slutzky, E. (1937). The summation of random causes as the source of cyclic processes. *Econometrica*, **5** (2), 105–146.

Su, L. and Ju, G. (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics*, **206** (2), 554–573.

—, Shi, Z. and Phillips, P. C. B. (2014). Identifying latent structures in panel data, cowles Foundation Discussion Paper No. 1965.

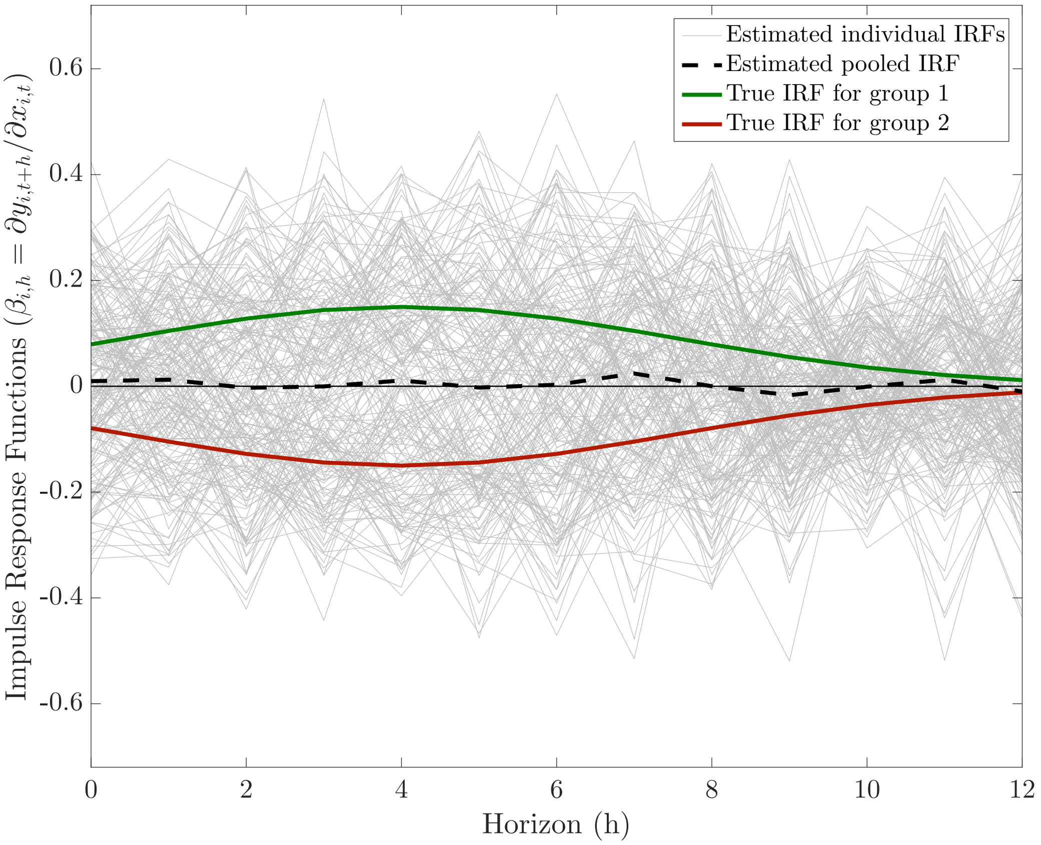
—, — and — (2016). Identifying latent structures in panel data. *Econometrica*, **84** (6), 2215–2264.

Sun, Y. (2005). Estimation and inference in panel structure models, working Paper, Dept. of Working Paper, Dept. of Economics, UCSD.

Wong, A. (2019). Refinancing and the transmission of monetary policy to consumption.

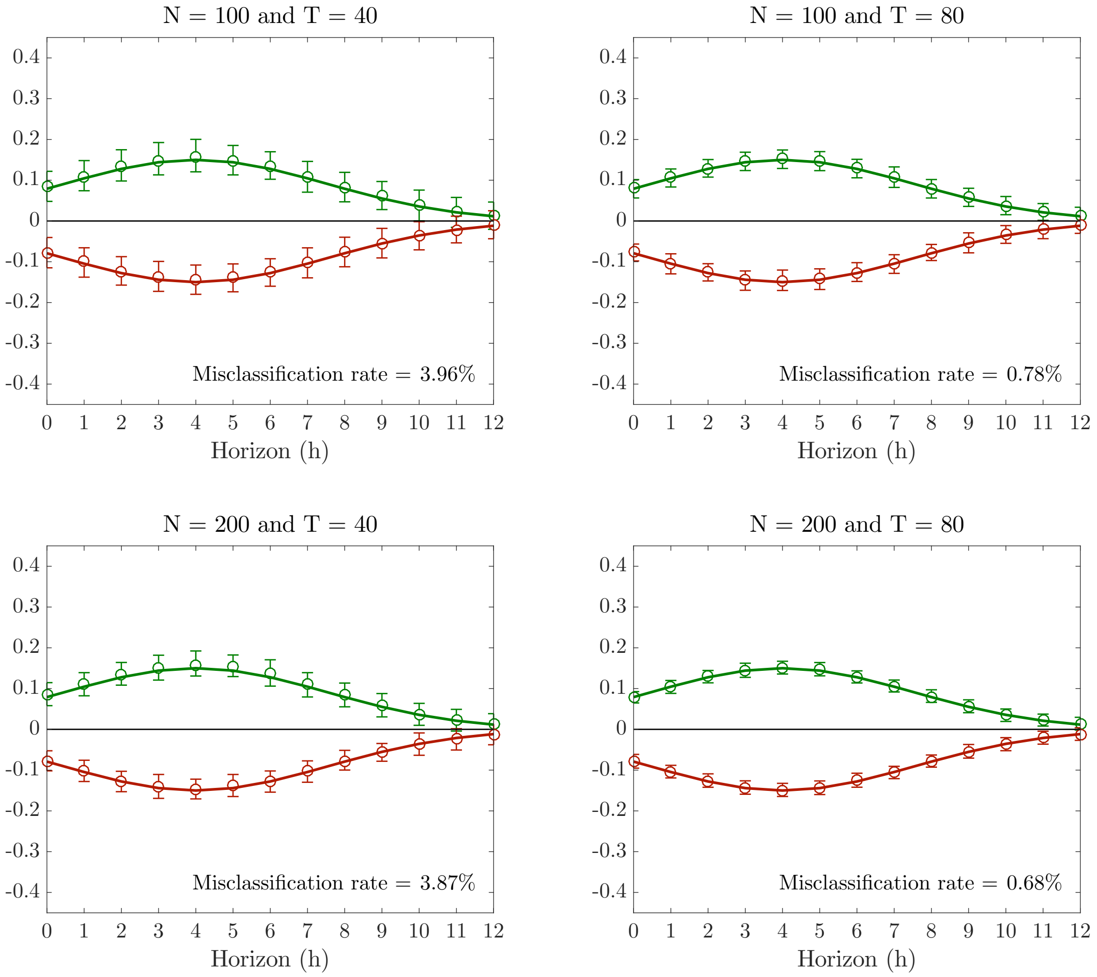
3.A. Figures

Figure 3.A.1: Heterogeneous IRFs under latent group heterogeneity



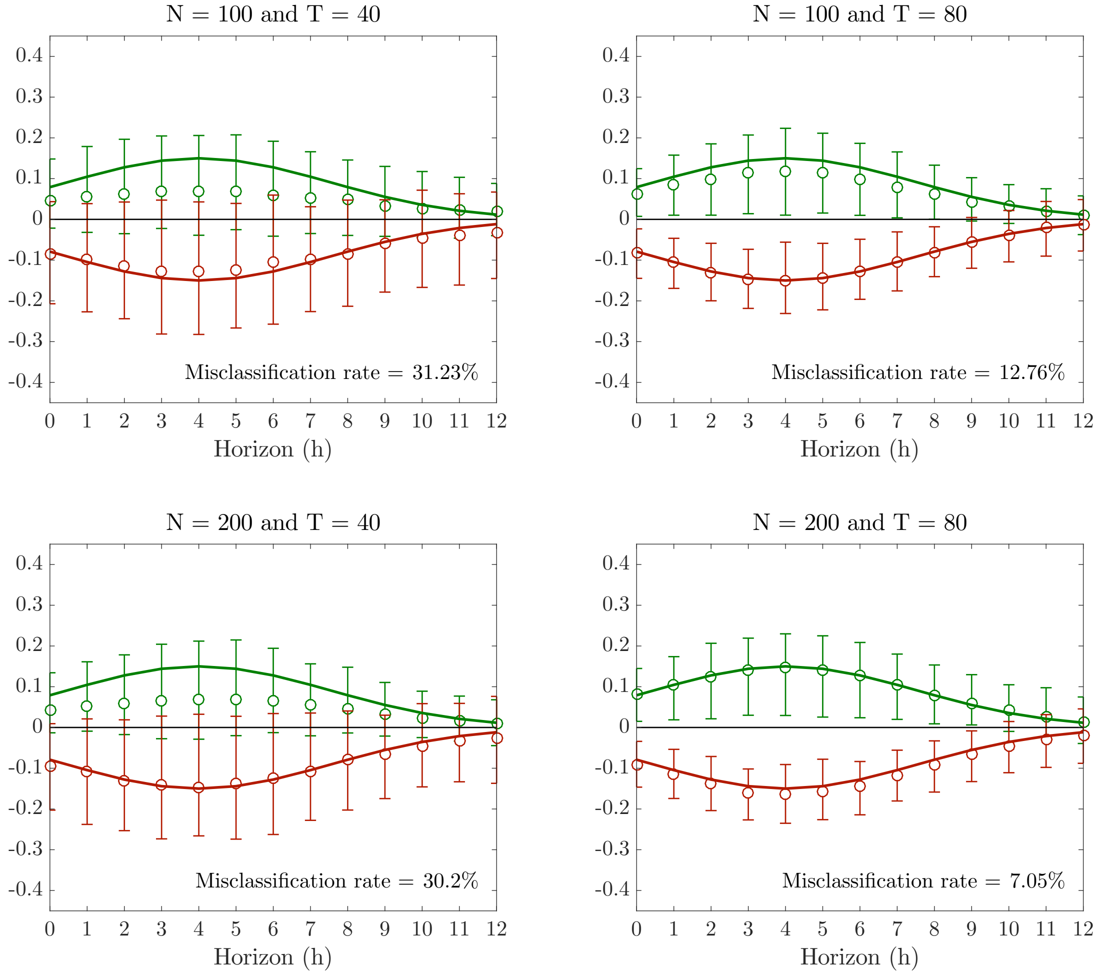
The plot is generated based on artificially generated panel data set with 200 individuals and 40 time periods. Half of the individuals have their true impulse responses given by the green line and the other half by the red line. The grey lines are individual-specific estimated impulse responses. The dashed-black line is the estimated impulse response obtained by pooling all the individuals and ignoring coefficient heterogeneity. The data generating process is described in section 3.5.1.

Figure 3.A.2: Post-Lasso IRFs based on moving average representation



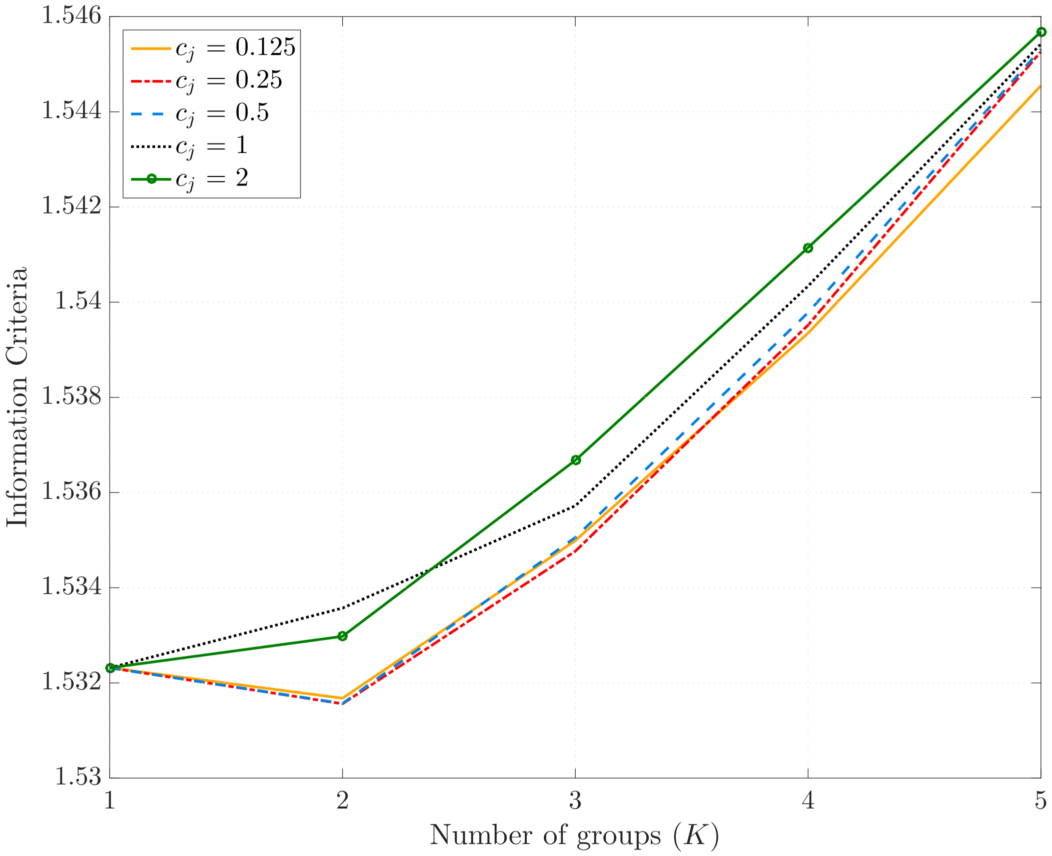
The solid lines are the true impulse responses for each group as defined in (3.14). The hollow circles represent the means of the sampling distribution of post-Lasso impulse response computed across Monte Carlo replications. The vertical line contains the interval from the 10th to 90th percentile of the post-Lasso impulse response estimates computed across Monte Carlo replications. For each Monte Carlo sample the misclassification rate is computed as $\frac{1}{N}\sum_{i=1}^{N}(\mathbb{1}\{i \in G_1 \cap \widehat{G}_2\} + \mathbb{1}\{i \in G_2 \cap \widehat{G}_1\})$ and the misclassification rates reported are the average misclassification rate across Monte Carlo samples.

Figure 3.A.3: Post-Lasso IRFs based on local projections



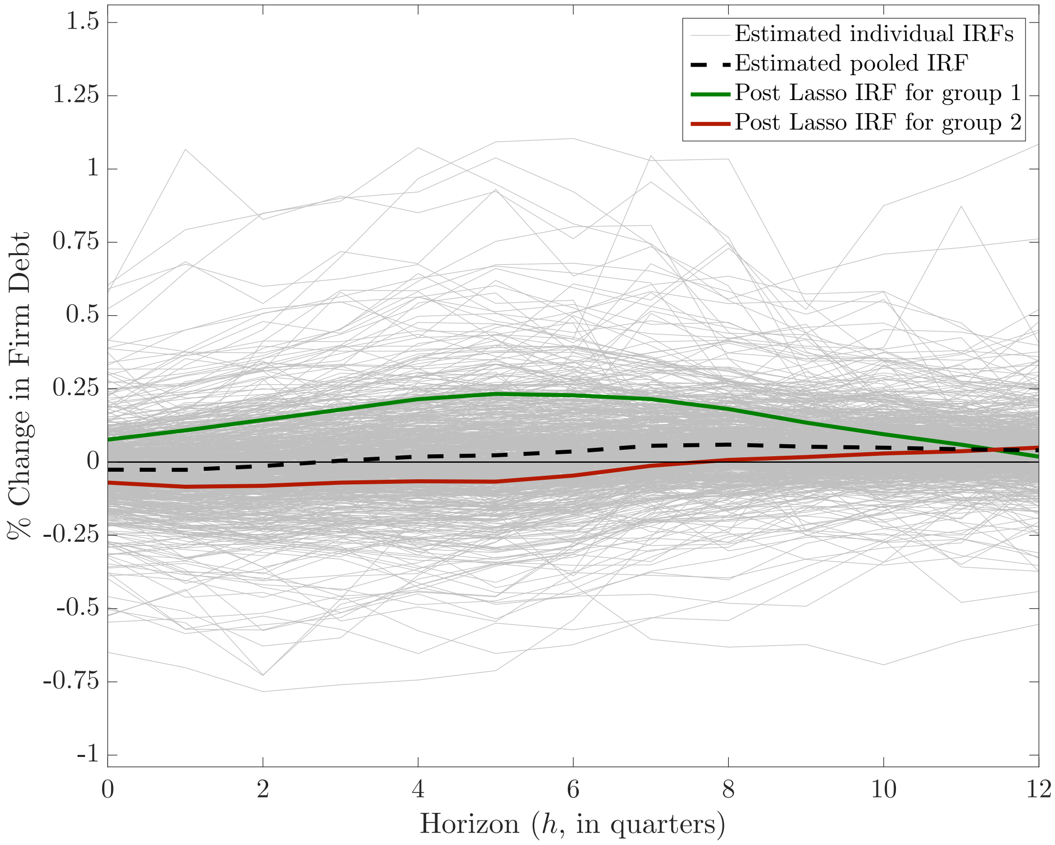
The solid lines are the true impulse responses for each group as defined in (3.14). The hollow circles represent the means of the sampling distribution of post-Lasso impulse response computed across Monte Carlo replications. The vertical line contains the interval from the 10th to 90th percentile of the post-Lasso impulse response estimates computed across Monte Carlo replications. For each Monte Carlo sample the misclassification rate is computed as $\frac{1}{N}\sum_{i=1}^{N}(\mathbb{1}\{i \in G_1 \cap \widehat{G}_2\} + \mathbb{1}\{i \in G_2 \cap \widehat{G}_1\})$ and the misclassification rates reported are the average misclassification rate across Monte Carlo samples.

Figure 3.A.4: Group determination for Drechsel (2020) dataset



Each line reports values of $\text{IC}(K, \lambda_1)$, as defined in (3.12), computed from the Drechsel (2020) dataset for alternative values of K and for $\lambda_1 = c_j s_Y^2 T^{\frac{1}{3}}$. The combination of (K, c_j) that minimises the IC is given by $(2, 0.25)$.

Figure 3.A.5: Post-Lasso IRFs of firm-level debt to IST shock



The solid green and red lines are the estimated post-Lasso impulse responses for the two latent groups identified in the Drechsel (2020) dataset. The dashed black line plot the impulse response obtained by pooling all the firms together. Each grey line plots a firm-specific estimated impulse response function. There is a total of 746 firms of which 224 are classified as belonging to group 1 and 522 are classified as belonging to group 2.

3.B. Tables

Table 3.B.1: Frequency of selecting $K = 1, \ldots, 5$ when $K^0 = 2$

		Moving Average estimation				
N	T	1	2	3	4	5
100	40	0.01	0.99	0	0	0
100	80	0	1	0	0	0
200	40	0	1	0	0	0
200	80	0	1	0	0	0
		Local Projection estimation				
N	T	1	2	3	4	5
100	40	0.59	0.3	0.11	0	0
100	80	0.3	0.67	0.03	0	0
200	40	0.41	0.29	0.29	0.01	0
200	80	0.1	0.82	0.08	0	0

For each DGP identified by a combination of N and T in the first two columns, this table reports $\frac{1}{250}\sum_{m=1}^{250} \mathbb{1}\{\widehat{K}_m = k\}$ where $\widehat{K}_m$ is the number of groups that minimizes the information criterion defined in (3.12) for the *m-th* Monte Carlo sample. In the top panel the C-Lasso estimates are obtained from minimising (3.11) whereas in the bottom panel they are obtained from minimising (3.15).

Table 3.B.2: $\beta_{i,4}$ estimators based on Moving Average representation

Group 1										
DGP		Full Heterogeneity			Post Lasso			Group Oracle		
N	T	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
100	40	0.0015	0.0359	0.0361	-0.0084	0.0058	0.0059	0.0015	0.0008	0.0008
100	80	0.0006	0.0152	0.0153	-0.0018	0.0015	0.0015	0.0006	0.0003	0.0003
200	40	-0.0009	0.0379	0.038	-0.0104	0.0055	0.0056	-0.0009	0.0003	0.0004
200	80	0.0003	0.0151	0.0152	-0.0016	0.0012	0.0012	0.0003	0.0002	0.0002

Group 2										
DGP		Full Heterogeneity			Post Lasso			Group Oracle		
N	T	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
100	40	0.0001	0.0355	0.0357	0.0099	0.0026	0.0027	0.0001	0.0007	0.0007
100	80	0.0022	0.0153	0.0153	0.0045	0.0005	0.0005	0.0022	0.0004	0.0004
200	40	-0.0006	0.0381	0.0383	0.0089	0.0023	0.0024	-0.0006	0.0004	0.0004
200	80	-0.0003	0.0153	0.0154	0.0016	0.0003	0.0003	-0.0003	0.0001	0.0001

The *Full Heterogeneity* estimator is obtained as the minimizer of (3.5), the *Post Lasso* estimator is obtained as described in section 3.4.3 and the *Group Oracle* is obtained as the ex-ante classification estimator from (3.3) and (3.4) under the true group membership. For a given estimator $\widehat{\beta}_{i,4}$: (i) the *bias* column for group j is computed as $\frac{1}{N_{G_j}} \sum_{i \in G_j} \mathbb{B}_{i,4}$ where $\mathbb{B}_{i,4} = \frac{1}{250} \sum_{m=1}^{250} (\widehat{\beta}_{i,4}^{m} - \beta_{i,4})$ and $\widehat{\beta}_{i,4}^{m}$ denotes the estimates for $\beta_{i,4}$ obtained from the m-*th* Monte Carlo sample of the respective DGP; (ii) the *variance* column for group j is computed as $(1/N_{G_j}) \sum_{i \in G_j} \mathbb{V}_{i,4}$ where $\mathbb{V}_{i,4} = (1/250) \sum_{m=1}^{250} (\widehat{\beta}_{i,4}^{m} - \overline{\beta}_{i,4})^2$ where $\overline{\beta}_{i,4} = \frac{1}{250} \sum_{m=1}^{250} \widehat{\beta}_{i,4}^{m}$ and (iii) the *MSE* column for group j is computed as $(1/N_{G_j}) \sum_{i \in G_j} \left(\mathbb{B}_{i,4}^2 + \mathbb{V}_{i,4}\right)$.

Table 3.B.3: $\beta_{i,4}$ estimators based on Local Projections

Group 1

DGP		Full Heterogeneity			Post Lasso			Group Oracle		
N	T	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
100	40	0.0003	0.0307	0.0308	-0.0748	0.0094	0.015	0.0003	0.0054	0.0054
100	80	-0.0024	0.0162	0.0162	-0.0335	0.0067	0.0078	-0.0024	0.0034	0.0034
200	40	-0.0001	0.0316	0.0317	-0.0744	0.0096	0.0151	-0.0001	0.0059	0.0059
200	80	0.0131	0.0157	0.0159	-0.0041	0.0057	0.0057	0.0131	0.003	0.0032

Group 2

DGP		Full Heterogeneity			Post Lasso			Group Oracle		
N	T	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
100	40	-0.0002	0.0315	0.0316	0.0748	0.0178	0.0234	-0.0002	0.0057	0.0057
100	80	0.0009	0.0161	0.0162	0.0321	0.0101	0.0111	0.0009	0.0033	0.0033
200	40	-0.003	0.0311	0.0312	0.0713	0.0181	0.0232	-0.003	0.0056	0.0056
200	80	-0.0127	0.0155	0.0158	0.0045	0.0077	0.0077	-0.0127	0.0031	0.0033

The *Full Heterogeneity* estimator is obtained as the minimizer of (3.5), the *Post Lasso* estimator is obtained as described in section 3.4.3 and the *Group Oracle* is obtained as the ex-ante classification estimator from (3.3) and (3.4) under the true group membership. For a given estimator $\widehat{\beta}_{i,4}$: (i) the *bias* column for group j is computed as $\frac{1}{N_{G_j}}\sum_{i\in G_j}\mathbb{B}_{i,4}$ where $\mathbb{B}_{i,4}=\frac{1}{250}\sum_{m=1}^{250}(\widehat{\beta}^m_{i,4}-\beta_{i,4})$ and $\widehat{\beta}^m_{i,4}$ denotes the estimates for $\beta_{i,4}$ obtained from the *m-th* Monte Carlo sample of the respective DGP; (ii) the *variance* column for group j is computed as $(1/N_{G_j})\sum_{i\in G_j}\mathbb{V}_{i,4}$ where $\mathbb{V}_{i,4}=(1/250)\sum_{m=1}^{250}(\widehat{\beta}^m_{i,4}-\overline{\beta}_{i,4})^2$ where $\overline{\beta}_{i,4}=\frac{1}{250}\sum_{m=1}^{250}\widehat{\beta}^m_{i,4}$ and (iii) the *MSE* column for group j is computed as $(1/N_{G_j})\sum_{i\in G_j}\left(\mathbb{B}^2_{i,4}+\mathbb{V}_{i,4}\right)$.

Table 3.B.4: Firm summary statistics for the two IRF groups

Panel A: Share of Collateral and Flow Borrowers

Variable	Group 1	Group 2	p-value
Collateral Borrowers	17.05%	22.14%	0.15
Earnings Borrowers	82.95%	77.86%	0.15

Panel B: Average share of Intangible Assets

Variable	Group 1	Group 2	p-value
Intangible Assets/Total Assets	17.44%	16.89%	0.67
Goodwill/Total Assets	13.31%	12.43%	0.41

Panel C: Firm Size (in billions USD)

Variable	Group 1	Group 2	p-value
Total Assets	6.64	9.85	0.05
Total Sales	1.73	2.01	0.54
Market capitalization	7.86	9.44	0.47

Panel D: Group Composition by GICS Sectors

Sector	Group 1	Group 2	p-value
Energy	7.14%	6.32%	0.69
Materials	15.63%	8.24%	0.01
Industrials	30.80%	20.88%	0.01
Consumer Discretionary	16.07%	19.92%	0.20
Consumer Staples	3.57%	9.77%	0.00
Health Care	8.48%	6.13%	0.27
Information Technology	10.71%	8.24%	0.30
Communication Services	0.89%	1.72%	0.33
Utilities	5.80%	18.58%	0.00

The columns *Group 1* and *Group 2* contain the average value of each variable computed across firms that are classified as belonging to groups 1 and 2 by the C-Lasso. The p-value column contains the p-value for the null hypothesis that the mean of a variable in group 1 is equal to the mean in group 2 against a two-sided alternative. The GICS sectors Financials and Real Estate were excluded from the table since there is only one firm in the Financial sector and two firms for real estate in the final sample.

Table 3.B.5: Average marginal effects on group membership

	(1)	(2)	(3)	(4)	(5)
Flow Borrower	0.069				0.042
	(0.049)				(0.049)
Intangibles share		0.000			0.000
		(0.001)			(0.001)
Total Assets			-0.002*		-0.002
			(0.001)		(0.002)
Materials				0.085	0.050
				(0.073)	(0.085)
Industrials				0.035	-0.004
				(0.065)	(0.077)
Consumer Disc.				-0.084	-0.109
				(0.069)	(0.079)
Consumer Staples				-0.240**	-0.396***
				(0.094)	(0.137)
Health Care				0.023	-0.033
				(0.081)	(0.100)
IT				0.010	-0.047
				(0.076)	(0.091)
Communication				-0.172	-0.076
				(0.164)	(0.181)
Utilities				-0.271***	-0.212**
				(0.080)	(0.093)
N	578	745	746	746	577

Each column reports the estimated average marginal effects for a Logit specification where the dependent variable is a dummy variable equal to one if the firm belongs to group 1 and 0 if if belongs to group 2. Standard errors in parenthesis. *, ** and *** denote marginal effects that are significant at 10%, 5% and 1% significance levels, respectively.

3.C. Proof of results in the main text

Proof of Proposition 3.1. Consider first the fully heterogeneous estimator,

$$\widehat{\boldsymbol{\beta}}_i = \left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\mathbf{X}_i'\mathbf{y}_i \tag{3.18}$$

where $\mathbf{y}_i \equiv [y_{i,1}, \ldots, y_{i,T}]'$. The proof of (3.6) is a standard textbook proof of OLS unbiasedness under the Gauss-Markov assumptions and it follows that $\mathbb{V}\mathrm{ar}\left(\widehat{\boldsymbol{\beta}}_i \mid \mathbf{X}\right) = \sigma^2\left(\mathbf{X}_i'\mathbf{X}_i\right)^{-1}$.[20] The estimator based on the ex-ante group classification approach defined in (3.3) and (3.4) is given by,

$$\widetilde{\boldsymbol{\beta}}_i(\tilde{\mathcal{G}}^K) = \left(\sum_{i\in\tilde{G}_a}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\sum_{i\in\tilde{G}_a}\mathbf{X}_i'\mathbf{y}_i \tag{3.19}$$

To derive (3.7) let $\boldsymbol{\varepsilon}_i \equiv [\varepsilon_{i,1}, \ldots, \varepsilon_{i,T}]'$ and use assumptions 1.1 and 3.2 to obtain,

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}_i(\tilde{\mathcal{G}}^K) &= \left(\sum_{i\in\tilde{G}_a}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\left(\sum_{i\in\tilde{G}_a}\mathbf{X}_i'\left(\mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i\right)\right)\\
&= \left(\sum_{i\in\tilde{G}_a}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\left(\sum_{i\in\tilde{G}_a}\mathbf{X}_i'\mathbf{X}_i\boldsymbol{\beta}_i + \sum_{i\in\tilde{G}_a}\mathbf{X}_i'\boldsymbol{\varepsilon}_i\right)\\
&= \left(\sum_{i\in\tilde{G}_a}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\left(\sum_{i\in\tilde{G}_a}\sum_{k=1}^{K_0}\mathbf{X}_i'\mathbf{X}_i\boldsymbol{\alpha}_k\mathbb{1}\{i\in G_k\} + \sum_{i\in\tilde{G}_a}\mathbf{X}_i'\boldsymbol{\varepsilon}_i\right)\\
&= \left(\sum_{i\in\tilde{G}_a}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\left(\sum_{i\in\tilde{G}_a\cap G_b}\mathbf{X}_i'\mathbf{X}_i\boldsymbol{\beta}_i + \sum_{\substack{k=1\\k\neq b}}^{K_0}\sum_{i\in\tilde{G}_a\cap G_k}\mathbf{X}_i'\mathbf{X}_i\boldsymbol{\alpha}_k + \sum_{i\in\tilde{G}_a}\mathbf{X}_i'\boldsymbol{\varepsilon}_i\right)
\end{aligned} \tag{3.20}$$

[20] See, for instance, Hayashi (2000, section 1.3).

Define $\tilde{\boldsymbol{\varphi}}_{a,b} \equiv \left(\sum_{i \in \tilde{G}_a} \mathbf{X}_i' \mathbf{X}_i\right)^{-1} \left(\sum_{i \in \tilde{G}_a \cap G_b} \mathbf{X}_i' \mathbf{X}_i\right)$ and simplify (3.20) to obtain,

$$\widetilde{\boldsymbol{\beta}}_i(\tilde{\mathcal{G}}^K) = \tilde{\boldsymbol{\varphi}}_{a,b}\, \boldsymbol{\beta}_i + \sum_{\substack{k=1 \\ k \neq b}}^{K_0} \tilde{\boldsymbol{\varphi}}_{a,k} \boldsymbol{\alpha}_k + \left(\sum_{i \in \tilde{G}_a} \mathbf{X}_i' \mathbf{X}_i\right)^{-1} \sum_{i \in \tilde{G}_a} \mathbf{X}_i' \boldsymbol{\varepsilon}_i \tag{3.21}$$

Taking conditional expectations yields,

$$\begin{aligned} \mathbb{E}\left(\widetilde{\boldsymbol{\beta}}_i(\tilde{\mathcal{G}}^K) \mid \mathbf{X}\right) &= \tilde{\boldsymbol{\varphi}}_{a,b}\, \boldsymbol{\beta}_i + \sum_{\substack{k=1 \\ k \neq b}}^{K_0} \tilde{\boldsymbol{\varphi}}_{a,k} \boldsymbol{\alpha}_k + \left(\sum_{i \in \tilde{G}_a} \mathbf{X}_i' \mathbf{X}_i\right)^{-1} \sum_{i \in \tilde{G}_a} \mathbf{X}_i'\, \mathbb{E}\left(\boldsymbol{\varepsilon}_i \mid \mathbf{X}\right) \\ &= \tilde{\boldsymbol{\varphi}}_{a,b}\, \boldsymbol{\beta}_i + \sum_{\substack{k=1 \\ k \neq b}}^{K_0} \tilde{\boldsymbol{\varphi}}_{a,k} \boldsymbol{\alpha}_k \end{aligned} \tag{3.22}$$

where the second equality follows from the strict exogeneity in assumption 1.1. Finally, using the law of iterated expectations we obtain (3.7),

$$\mathbb{E}\left(\widetilde{\boldsymbol{\beta}}_i(\tilde{\mathcal{G}}^K)\right) = \mathbb{E}\left(\mathbb{E}\left(\widetilde{\boldsymbol{\beta}}_i(\tilde{\mathcal{G}}^K) \mid \mathbf{X}\right)\right) = \boldsymbol{\varphi}_{a,b}\, \boldsymbol{\beta}_i + \sum_{\substack{k=1 \\ k \neq b}}^{K_0} \boldsymbol{\varphi}_{a,k} \boldsymbol{\alpha}_k \tag{3.23}$$

where $\boldsymbol{\varphi}_{a,b} \equiv \mathbb{E}\left(\tilde{\boldsymbol{\varphi}}_{a,b}\right)$. Proving (3.8) requires showing that $\mathbb{V}\text{ar}\left(\widehat{\boldsymbol{\beta}}_i \mid \mathbf{X}\right) - \mathbb{V}\text{ar}\left(\widetilde{\boldsymbol{\beta}}_i(\tilde{\mathcal{G}}^K) \mid \mathbf{X}\right)$ is positive semi-definite. Start by using (3.20) to derive $\mathbb{V}\text{ar}\left(\widetilde{\boldsymbol{\beta}}_i(\tilde{\mathcal{G}}^K) \mid \mathbf{X}\right)$,

$$\begin{aligned} \mathbb{V}\text{ar}\left(\widetilde{\boldsymbol{\beta}}_i(\tilde{\mathcal{G}}^K) \mid \mathbf{X}\right) &= \left(\sum_{i \in \tilde{G}_a} \mathbf{X}_i' \mathbf{X}_i\right)^{-1} \mathbb{V}\text{ar}\left(\sum_{i \in \tilde{G}_a} \mathbf{X}_i' \boldsymbol{\varepsilon}_i\right) \left(\sum_{i \in \tilde{G}_a} \mathbf{X}_i' \mathbf{X}_i\right)^{-1} \\ &= \left(\sum_{i \in \tilde{G}_a} \mathbf{X}_i' \mathbf{X}_i\right)^{-1} \left(\sum_{i \in \tilde{G}_a} \mathbf{X}_i' \mathbb{V}\text{ar}\left(\boldsymbol{\varepsilon}_i \mid \mathbf{X}\right) \mathbf{X}_i\right) \left(\sum_{i \in \tilde{G}_a} \mathbf{X}_i' \mathbf{X}_i\right)^{-1} \\ &= \sigma^2 \left(\sum_{i \in \tilde{G}_a} \mathbf{X}_i' \mathbf{X}_i\right)^{-1} \end{aligned} \tag{3.24}$$

where the second and third equalities follow from the conditional homoskedasticity and no autocorrelation assumption. Combining the expressions for $\mathbb{V}\text{ar}\left(\widehat{\boldsymbol{\beta}}_i \mid \mathbf{X}\right)$ and $\mathbb{V}\text{ar}\left(\widetilde{\boldsymbol{\beta}}_i(\tilde{\mathcal{G}}^K) \mid \mathbf{X}\right)$ yields,

$$\mathbb{V}\text{ar}\left(\widehat{\boldsymbol{\beta}}_i \mid \mathbf{X}\right) - \mathbb{V}\text{ar}\left(\widetilde{\boldsymbol{\beta}}_i(\tilde{\mathcal{G}}^K) \mid \mathbf{X}\right) = \sigma^2 \underbrace{\left((\mathbf{X}_i'\mathbf{X}_i)^{-1} - \left(\sum\nolimits_{i\in\tilde{G}_a} \mathbf{X}_i'\mathbf{X}_i\right)^{-1}\right)}_{[*]} \tag{3.25}$$

The term $[*]$ is positive semi definite if and only if $\sum_{i\in\tilde{G}_a} \mathbf{X}_i'\mathbf{X}_i - \mathbf{X}_i'\mathbf{X}_i$ is positive semi definite. Finally,

$$\sum\nolimits_{i\in\tilde{G}_a} \mathbf{X}_i'\mathbf{X}_i - \mathbf{X}_i'\mathbf{X}_i = \sum\nolimits_{j\in\tilde{G}_a\setminus\{i\}} \mathbf{X}_j'\mathbf{X}_j \tag{3.26}$$

If $\tilde{G}_a\setminus\{i\} = \varnothing$ then $\sum_{i\in\tilde{G}_a} \mathbf{X}_i'\mathbf{X}_i - \mathbf{X}_i'\mathbf{X}_i = \mathbf{0}$. If $\tilde{G}_a\setminus\{i\} \neq \varnothing$, then $\sum_{i\in\tilde{G}_a} \mathbf{X}_i'\mathbf{X}_i - \mathbf{X}_i'\mathbf{X}_i$ is the sum of positive definite matrices and, hence, positive definite. Therefore, $[*]$ is positive semi-definite. □