NBER WORKING PAPER SERIES

MISSING AGGREGATE DYNAMICS: ON THE SLOW CONVERGENCE OF LUMPY ADJUSTMENT MODELS

David Berger
Ricardo J. Caballero
Eduardo Engel

Working Paper 9898
http://www.nber.org/papers/w9898

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2003, Revised June 2017

We are grateful to Filippo Altissimo, Fernando Alvarez, Luigi Bocola, William Brainard, Jeff Campbell, Larry Christiano, Olivier Coibion, Ian Dew-Becker, Marty Eichenbaum, Xavier Gabaix, Pablo García, Marc Giannoni, Robert Hall, Fabiano Schiaverdi, Jon Steinsson, Eric Swanson, Yuta Takahashi, Harald Uhlig, Joe Vavra and seminar participants at Chicago Fed, Columbia University University of Chicago, FGV (Rio and Sao Paulo), Humboldt Universität, IEIS Stockholm, MIT, NBER SI (EFCE), PUC (Chile), Universidad de Chile (CEA and FEN), University of Maryland, University of Paris, University of Pennsylvania, Uppsala University, Yale University, NBER EFG Meeting, 2nd ECB/IMOP Workshop on Dynamic Macroeconomics, Hydra and the Central Bank of Chile's "The Micro and the Macro of Price Rigidities" workshop for their comments on an earlier version of this paper. We thank Juan Daniel Díaz for outstanding research assistance. Financial support from NSF is gratefully acknowledged. This paper is an extensively revised version of "Adjustment is Much Slower than You Think," NBER WP #9898. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Missing Aggregate Dynamics: On the Slow Convergence of Lumpy Adjustment Models
David Berger, Ricardo J. Caballero, and Eduardo Engel
NBER Working Paper No. 9898
August 2003, Revised June 2017
JEL No. C22,C43

ABSTRACT

The estimated persistence of macro aggregates involving lumpy microeconomic adjustment is biased downward when inferred from VAR estimates. The extent of this "missing persistence bias" decreases with the level of aggregation, yet convergence is very slow. Paradoxically, while idiosyncratic shocks smooth away microeconomic non-convexities and are often used to justify approximating aggregate dynamics with linear models, their presence exacerbates the bias. We propose a method to estimate the true speed of adjustment and illustrate its effectiveness via simulations and applications to real data.

The missing persistence bias is relevant for macroeconomists on many grounds. First, when calibrating or estimating models via simulation based methods, macroeconomists should pay attention to the number of agents used in simulations for otherwise they are likely to obtain systematic biases in their parameter estimates. Second, results purporting to find persistence measures that vary systematically with levels of aggregation should be examined with care since the differential speeds may disappear when using estimation methods robust to the missing persistence bias. To illustrate the latter, we show that the difference in the speed with which inflation responds to sectoral and aggregate shocks (Boivin et al 2009; Mackoviak et al 2009) disappears once we correct for the missing persistence bias.

David Berger
Department of Economics
Northwestern University
2001 Sheridan Road
Evanston, IL 60208
and NBER
david.berger@northwestern.edu

Eduardo Engel
University of Chile
Department of Economics
Diagonal Paraguay 257 Piso 14
Santiago
CHILE
eengel@econ.uchile.cl

Ricardo J. Caballero
Department of Economics, E52-528
MIT
77 Massachusetts Avenue
Cambridge, MA 02139
and NBER
caball@mit.edu

1 Introduction

Measuring the dynamic response of aggregate variables to shocks is one of the central concerns of applied macroeconomics. The main procedure used to measure these dynamics consists in estimating a vector autoregression (VAR). In non- or semi-structural approaches, the characterization of dynamics stops there. In other, more structural approaches, researchers wish to uncover underlying parameters from the estimated VAR and use the implied response to shocks as the benchmark against which the success of the calibration exercise, and the need for further theorizing, is assessed.

The main point of this paper is that when the microeconomic adjustment underlying an aggregate variable is lumpy, conventional VAR procedures often lead the researcher to conclude that there is less persistence than there really is. We refer to this as the “missing persistence bias". The extent to which persistence is underestimated decreases with the level of aggregation: linear models miss any persistence that might be present when applied to an individual series while the bias vanishes completely when they are applied to a series that aggregates infinitely many agents. However, convergence is very slow: the bias is likely to be present in general for sectoral data and, quite often, for aggregate series as well. For example, the response of inflation to a monetary shock in a Calvo-type model, as measured by the half-life of the shock, will be overestimated by a factor of 5 with sectoral data (1000 effective price setters) and by a factor of 1.5 with aggregate data (15,000 effective price setters).[2]

The bias has significant implications for applied macroeconomic research. It implies that estimated impulse response functions will be biased when computed in the standard way – as nonlinear functions of the parameters from the estimated VAR system. More generally, care must be taken when conducting any structural tests that derive from dynamic systems estimated in VAR style models such as the common practice of estimating DSGE models by indirect inference.[3] For example, when using simulations based methods to calibrate or estimate model parameters, the common practice of simulating a very large (continuum) number of agents will lead to systematic biases in parameter estimations when the underlying data aggregate has many fewer underlying observations.

We show that the missing persistence bias is present in VAR-based estimates of impulse response functions and propose a procedure immune to the bias that estimates the true speed of adjustment. We provide two detailed applications where we correct for the bias. In the first application, we explain why estimates for the speed of adjustment of sectoral prices obtained using direct measures are much lower than those obtained with standard linear time-series models, thereby potentially solving a puzzling finding in Bils and Klenow (2004). In this application we can measure

[2] "Effective" is defined as the inverse of the Herfindahl index (see Section 2). This qualifier is important because when individual data observations have different weights there can be a large difference between the number of observations and the effective number of observations (which is what matters for magnitude of the bias). For example, in the U.S. CPI, the median total number of observations per month used in its construction over the period 1988-2007 is 65,938 while the median effective number of observations per month during this period is 15,503.

[3] For example, Christiano, Eichenbaum and Evans (2005).

the size of the bias and find that our bias correction procedure works well in practice: linear time series models deliver estimates in line with those obtained with unbiased nonlinear methods once the linear methods are applied correcting for the missing persistence bias.

Our second, more substantial, application revisits Boivin, Giannoni and Mihov's (2009) finding that sectoral inflation responds much faster to sectoral shocks than to aggregate shocks (see also Mackowiak, Moench and Wiederholt, 2009). This widely cited finding has been interpreted as strong evidence in favor of models in which agents choose how much information they acquire because in these models the amount firms respond to shocks depends on the relative variance of the shocks.[4] While these models may still capture an important aspect of price-setting, we show that Boivin, Giannoni and Mihov's (2009) persistence measure is subject to the missing persistence bias and that once we correct for it, the responses of sectoral inflation to both types of shocks look very similar. This application illustrates the general point that results purporting to find persistence measures that vary systematically with levels of aggregation should be examined with care since the differences in estimated speeds of adjustment may be manifestations of the missing persistence bias.

The intuition underlying our main result is best explained by comparing the impulse response of the true nonlinear model that includes lumpy adjustment with the impulse response computed by a linear approximation to the true, nonlinear dynamics. In the simple case of *one* agent and i.i.d. shocks, the agent's optimal response every time it acts is to adjust by the sum of shocks that accumulated since the last time it adjusted. We then have that the agent responds in period $t+k$ to a shock that took place in period t only if the agent adjusted in $t+k$ and did not adjust in all periods between t and $t+k-1$. It follows that the average response in $t+k$ to a shock that took place in t is equal to the probability of having to wait exactly k periods after the shock takes place until the first opportunity to adjust. In the simple case where the arrival process that determines when adjustments take place follows a geometric distribution, as in the discrete time version of the Calvo (1983) model, the nonlinear impulse response will be identical to that of an AR(1) process, with persistence parameter equal to the probability of not adjusting in a given period.

Consider next the impulse response obtained using a linear time-series model. This response will depend on the correlations between the agent's actions at different points in time. If the agent did not adjust in one of the periods under consideration, there is no correlation since at least one of the variables entering the correlation computation is exactly zero. The correlation will also be zero when the agent adjusted at both points in time because the agent's actions reflect shocks in non-overlapping periods and shocks are uncorrelated. This implies that the impulse response obtained via linear methods will be zero at all strictly positive lags, suggesting immediate adjustment to shocks and therefore no persistence, independent of the true degree of persistence. That is, even though the nonlinear IRF recovers the Rotemberg (1987) result, according to which the aggregate of

[4]For example, classic rational inattention models such as Mackowiak or Wiederholt (2006) or recently rational inattention/imperfect information hybrids such as Stevens (2016) and Baily and Blanco (2016).

interest follows an AR(1) process with first-order autocorrelation equal to the fraction of units that remain inactive, the linear IRF implies an i.i.d. process which corresponds to the above mentioned AR(1) process when all units adjust in every period and wrongly suggests instantaneous adjustment to shocks.

The bias falls as aggregation rises because the correlations at leads and lags of the adjustments across individual units are non-zero. That is, the common components in the adjustments of different agents at different points in time provides the correlation that allows the econometrician using linear time-series methods to recover the nonlinear impulse response. The more important this common component is —as measured either by the variance of aggregate shocks relative to the variance of idiosyncratic shocks or the frequency with which adjustments take place— the faster the estimate converges to the value of the persistence parameter as the number of agents grows. While idiosyncratic productivity and demand shocks smooth away microeconomic non-convexities and are often used as a justification for approximating aggregate dynamics with linear models, their presence exacerbates the bias. The fact that in practice idiosyncratic uncertainty is many times larger than aggregate uncertainty, suggests that the problem of missing aggregate dynamics is likely to be prevalent in empirical and quantitative macroeconomic research.

The remainder of the paper is organized as follows. Section 2 presents the Rotemberg (1987) equivalence result that justifies using linear time-series methods to estimate the dynamics for aggregates with lumpy microeconomic adjustment, as long as the number of units in the aggregate is infinite. Section 3 begins by presenting the missing persistence bias that arises when the number of units considered is finite. Next we describe approaches to correct for the bias. In Section 4, we show the robustness of the bias to many extensions of the baseline model. Section 5 studies two detailed applications and Section 6 concludes. Several appendices follow.

2 Linear time-series models and the Calvo-Rotemberg limit

Regardless of whether the final goal is to have a reduced form characterization of aggregate dynamics, or whether this is an intermediate step in identifying structural parameters, or whether it is just a metric to assess the performance of a calibrated model, it is common that researchers in macroeconomics at some key stage estimate an equation of the form:

$$a(L)\Delta y_t = \varepsilon_t, \tag{1}$$

where Δy represents the change in the log of some aggregate variable of interest, such as a price index, the level of employment, or the stock of capital; ε is an i.i.d. innovation and $a(L) \equiv 1 - \sum_{k=1}^{p} a_k L^k$, where L is the lag operator and the a_is are fixed parameters.

The question that concerns us here is whether the estimated $a(L)$ captures the true dynamics of the system when the underlying microeconomic variables exhibit lumpy adjustment behavior. We

show that unless the effective number of underlying micro units is very large, the answer is 'no'.

We setup the basic environment by constructing a simple model of microeconomic lumpy adjustment. Let y_{it} denote the variable of concern at time t for agent i and y^*_{it} be the level the agent chooses if it adjusts in period t (the 'reset value' of y). We will have that:

$$\Delta y_{it} = \xi_{it}(y^*_{it} - y_{it-1}), \tag{2}$$

where $\xi_{it} = 1$ if the agent adjusts in period t and $\xi_{it} = 0$ if not.

From a modeling perspective, discrete adjustment entails two distinct features. First, periods of inaction are followed by abrupt adjustments to accumulated imbalances. Second, the likelihood of an adjustment increases with the size of the imbalance and is therefore state dependent. While the second feature is central for the macroeconomic implications of state-dependent models, it is the first feature of discrete adjustment that is crucial to generating to missing persistence bias. Since the focus of this paper is the genesis of this bias and because we want to highlight the features which drive this bias, we start by focusing on a model that only has the first feature of discrete adjustment. This is the well-known Calvo model (1983).[5]

In this model:

$$\begin{aligned} \Pr\{\xi_{it} = 0\} &= \rho, \\ \Pr\{\xi_{it} = 1\} &= 1 - \rho. \end{aligned} \tag{3}$$

It follows from (3) that the *expected* value of ξ_{it} is $1-\rho$. When ξ_{it} is zero, the agent experiences inaction; when its value is one, the unit adjusts so as to eliminate the accumulated imbalance. We assume that ξ_{it} is independent of $(y^*_{it} - y_{it-1})$ —this is the simplification that Calvo (1983) makes vis-a-vis more realistic state dependent models— and therefore have:

$$\mathrm{E}[\Delta y_{it} \mid y^*_{it},\, y_{it-1}] \;=\; (1-\rho)(y^*_{it} - y_{it-1}), \tag{4}$$

so that ρ represents the degree of *inertia* of Δy_{it}. When ρ is large, the unit adjusts on average by a small fraction of its current imbalance and the expected half-life of shocks is large. Conversely, when ρ is small, the unit is expected to react promptly to any imbalance.

Let us now consider the behavior of aggregates. Given a set of weights w_i, $i = 1, 2, .., n$, with $w_i > 0$ and $\sum_{i=1}^n w_i = 1$, we define the *effective number of units*, N, as the inverse of the Herfindahl index:

$$N \equiv \frac{1}{\sum_{i=1}^n w_i^2}.$$

When all units contribute the same to the aggregate ($w_i = 1/n$) we have $N = n$, otherwise the effective number of units can be substantially smaller than the actual number of units.

[5] In Section 4, we return to state-dependent price models and demonstrate that the bias is also large in quantitatively relevant versions of these models.

We can now write the aggregate at time t, y_t^N, as:

$$y_t^N \equiv \sum_{i-1}^{n} w_i y_{it}.$$

Similarly we define the value of the aggregate reset value, y_t^{N*}, as

$$y_t^{N*} \equiv \sum_{i-1}^{n} w_i y_{it}^*.$$

Technical Assumptions (Shocks)

Let $\Delta y_{it}^* \equiv v_t^A + v_{it}^I$, where the absence of a subindex i denotes an element common to all units. We assume:

1. The v_t^A's are i.i.d. with mean μ_A and variance $\sigma_A^2 > 0$.
2. The v_{it}^I's are independent (across units, over time, and with respect to the v^A's), identically distributed with zero mean and variance $\sigma_I^2 > 0$.
3. The ξ_{it}'s are independent (across units, over time, and with respect to the v^A's and v^I's), identically distributed Bernoulli random variables with probability of success $\rho \in (0,1]$. ∎

As Rotemberg (1987) showed, when N goes to infinity, equation (4) for Δy^∞ becomes:

$$\Delta y_t^\infty = (1-\rho)(y_t^{\infty *} - y_{t-1}^\infty). \tag{5}$$

Taking first differences yields

$$\Delta y_t^\infty = \rho \Delta y_{t-1}^\infty + (1-\rho)\Delta y_t^{\infty *}, \tag{6}$$

which is the analog of Euler equations derived from a simple quadratic adjustment cost model applied to a representative agent.[6]

This is a powerful result which lends substantial support to the standard practice of approximating the aggregates as if they were generated by a simple linear model. What we show below, however, is that while this approximation may be good for some purposes, it can be particularly bad when it comes to motivating VAR estimation of aggregate dynamics.

Before doing so, let us close the loop by recovering equation (1) in this setup. For this, let us momentarily relax the Technical Assumptions 1 and 2, allowing for persistence in the v_t^A and v_{it}^I's, so that the change in the aggregate reset value of y, $\Delta y^{\infty *}$, is generated by:

$$b(L)\Delta y_t^{\infty *} = \varepsilon_t,$$

[6] For the proof, see Appendix E.

where the ε_t's are i.i.d. and $b(L) \equiv 1 - \sum_{i=1}^{q} b_i L^i$ defines a stationary AR(q) for $\Delta y^{\infty *}$. Assuming Technical Assumption 3 holds we have

$$\Delta y_t^{\infty} = \rho \Delta y_{t-1}^{\infty} + (1-\rho)\Delta y_t^{\infty *},$$

which combined with the AR(q) specification for $\Delta y^{\infty *}$ yields

$$(1-\rho L)b(L)\Delta y_t^{\infty} = (1-\rho)\varepsilon_t.$$

Comparing this expression with (1) we conclude that

$$a(L) = b(L)\frac{(1-\rho L)}{1-\rho}.$$

The bias we highlight in this paper comes from a severe downward bias in the (explicit or implicit) estimate of ρ, resulting in an estimate for $a(L)$ that misses significant dynamics. In the next section we simplify the exposition and set $b(L) \equiv 1$, as in the case considered by the Technical Assumptions. We consider the general case in Section 4.

3 The missing persistence bias

The effective number of units, N, in any real world aggregate is not infinity. The question that concerns us in this section is whether N is sufficiently large so that the limit result provides a good approximation.

Our main proposition states that the answer to this question depends on parameter values, in particular, on the relative importance of aggregate and idiosyncratic shocks, the effective number of agents, and the frequency of adjustment. When any of these is small, the bias can remain significant even at the economy-wide level. We argue that this is likely to be the case for various aggregates with lumpy microeconomic adjustment in the U.S. and, by extension, for smaller economies and sectoral data.

3.1 The theory

We ask whether estimating (6) with an effective number of units equal to N instead of infinity yields a consistent (as T goes to infinity) estimate of ρ, when the true microeconomic model is described by (2) and (3). The following proposition answers this question by providing an explicit expression for the bias as a function of the parameters characterizing adjustment probabilities and shocks (ρ, μ_A, σ_A and σ_I) and N.

Proposition 1 (Aggregate Bias)

Let $\hat{\rho}$ denote the OLS estimator of ρ in

$$\Delta y_t^N = \text{const.} + \rho \Delta y_{t-1}^N + e_t. \tag{7}$$

Let T denote the time series length. Then, under the Technical Assumptions, $\text{plim}_{T\to\infty}\hat{\rho}$ *depends on the weights w_i only through N and*

$$\text{plim}_{T\to\infty}\hat{\rho}^N = \frac{K}{1+K}\rho, \tag{8}$$

with

$$K \equiv \frac{\frac{1-\rho}{1+\rho}(N-1) - \left(\frac{\mu_A}{\sigma_A}\right)^2}{1 + \left(\frac{\sigma_I}{\sigma_A}\right)^2 + \frac{1+\rho}{1-\rho}\left(\frac{\mu_A}{\sigma_A}\right)^2}. \tag{9}$$

It follows that:

$$\lim_{N\to\infty} \text{plim}_{T\to\infty}\hat{\rho}^N = \rho. \tag{10}$$

Proof See Appendix C. ∎

Equation (10) in the proposition restates Rotemberg's (1987) result. Yet here we are interested in the value of $\hat{\rho}$ before the limit is reached and how structural parameters affect the magnitude of the missing persistence bias. That is, we would like to assess how K varies with underlying parameters.

Examination of equation (10) reveals that the bias drops as the effective number of units in the aggregate being considered increases and as the relative importance of aggregate to idiosyncratic shocks rises. Other factors that contribute to slow convergence is a larger drift (in absolute value) in the process driving the reset variable y^*, and a larger degree of inertia as captured by the fraction of agents that do not adjust in any given period, ρ. Before we provide more intuition for this, we first argue that this bias is relevant in many empirical applications.

3.2 The bias is large in practice

To put the relevance of this non-limit result in perspective, next we consider three macroeconomic variables where lumpy microeconomic adjustment has been well established —employment, prices, and investment— and use our theory to provide back-of-the-envelope estimates of the magnitude of the missing persistence bias in each of these cases. Table 1 reports how the estimated ρ and half-life of shocks varies for these aggregates with the effective number of units, N. We focus on the T = ∞ case for two important reasons: the missing persistence bias is conceptually distinct from the well-known AR(1) finite sample bias[7] and in most realistic applications (including our empiri-

[7] See Hamilton 1994 pp 216 for a textbook treatment.

cal applications in Section 5) the missing persistence bias is an order of magnitude larger than the finite sample bias.[8]

Table 1: SLOW CONVERGENCE

$\hat{\rho}$ and Half-Life of Shocks

Persistence measure	Aggregate	Freq	Effective number of agents (N)						
			100	400	1,000	4,000	15,000	40,000	∞
$\hat{\rho}$	Prices	M	0.070	0.232	0.415	0.679	**0.803**	0.838	0.860
	Employment	Q	0.156	0.352	0.468	**0.560**	0.589	0.596	0.600
	Investment	A	0.033	0.209	**0.402**	0.671	0.794	0.828	0.850
Half-life	Prices	M	0.261	0.475	0.788	1.788	**3.157**	3.913	4.596
	Employment	Q	0.373	0.663	0.913	**1.197**	1.309	1.339	1.357
	Investment	M	0.203	0.442	**0.760**	1.737	3.006	3.677	4.265

This table documents how the bias varies with the effective number of units. We consider two different persistence measures. The first three rows show results for estimated ρ, $\hat{\rho}$, which is computed using equation (8) and the parameter values listed below. Parameters for prices: $\rho = 0.86$, $\mu_A = 0.003$, $\sigma_A = 0.0054$, $\sigma_I = 0.048$. Parameters for employment: $\rho = 0.60$, $\mu_A = 0.005$, $\sigma_A = 0.03$, $\sigma_I = 0.25$. Parameters for investment: $\rho = 0.85$, $\mu_A = 0.12$, $\sigma_A = 0.056$, $\sigma_I = 0.50$. Numbers in boldface correspond, approximately, to the effective number of units for U.S. aggregates (CPI for prices, non-farm business sector for employment and investment). The fourth to sixth rows show the reported half-life. The half-life is inferred from the estimated ρ's in the first three rows and is computed using the following formula: $-\log 2/\log\hat{\rho}$.

Table 1 provides persistence estimates for three different aggregate series: prices, employment and investment and two different measures of persistence. The first measure of persistence (rows 1-3) is our AR(1) estimate $\hat{\rho}$, computed using equations (8) and (9) for a given set of parameter values and underlying agents. The second measure (rows 4-6) is the half-life computed as $-\log 2/\log\hat{\rho}$. For each aggregate series, we use parameter values for the underlying shock processes which have been disciplined by microdata. Then, holding these underlying parameter values fixed, we vary the underlying number of agents, N, to highlight how the magnitude of the bias depends on the degree of aggregation.

We begin with prices, which are reported in the first row in Table 1. We assume $\rho = 0.86$, in line with the median frequency of price adjustments for regular prices reported in Klenow and Kryvtsov (2008).[9] Values for μ_A and σ_A are taken from Bils and Klenow (2004), while σ_I is consistent with the value estimated in Caballero et al (1997).[10] The first result that jumps out is that the bias is very large for small values of effective units, with the estimated half-life being biased downward by more

[8] Monte-Carlo simulations confirming this statement are available upon request.

[9] The average over the eight median frequencies reported by Nakamura and Steinsson (2008) for regular price changes suggest taking $\rho = 0.89$ which leads to a larger bias.

[10] To go from the σ_I computed for employment in Caballero et al. (1997) to that of prices, we note that if the demand faced by a monopolistic competitive firm is isoelastic, its production function is Cobb-Douglas, and its capital fixed

than 80% (1.788/4.596) when $N = 1,000$. Consistent with equation (8), the magnitude of the bias is strongly decreasing in N, however, the table shows that the bias remains significant (30%) even for $N = 15,000$. This is the empirically relevant number since this corresponds, approximately, to the effective number of prices used to calculate the entire CPI.[11] This suggests the bias might be significant even in the published aggregate inflation series. The main reason for the persistence of the bias even for large N is the high value of σ_I/σ_A.

The second row in Table 1 reports the results for aggregate U.S. employment. We use the parameters estimated by Caballero, Engel, and Haltiwanger (1997) with quarterly Longitudinal Research Datafile (LRD) data for μ_A, σ_A, σ_I and ρ. The second row in Table 1 suggests that with $N = 3,683$, which is the effective size of employment in the non-farm business sector in 2001, the bias is only slightly above 10%. However, note that when $N = 100$, which corresponds to the average effective number of establishments in a typical two-digit sector of the LRD, the estimate half-life of shocks is less than one third of the actual half-life. The main reason the bias is smaller is the high value of the frequency of adjustment.

Finally, the third row in Table 1 reports the estimates for equipment investment, the most sluggish of the three series. The estimate of ρ, μ_A and σ_A, are from Caballero, Engel, and Haltiwanger (1995), and σ_I is consistent with that found in Caballero et al. (1997).[12] Here the bias remains very large and significant throughout. In particular, when $N = 986$, which corresponds to the effective number of establishments for capital weights in the U.S. Non-Farm Business sector in 2001, the estimated half-life of a shock is only 14% of the true half-life or, equivalently, the estimated frequency of adjustment, $1-\rho$, is more than four times the true frequency. The reasons for this is the combination of a high ρ, a high μ_A (mostly due to depreciation) and a large σ_I (relative to σ_A).

Summing up, our back-of-the-envelope estimates indicate that the missing persistence bias is quantitatively large at the sectoral level for inflation, employment and investment. Furthermore, linear time-series models will miss a substantial part of the dynamic behavior of U.S. inflation and investment at the aggregate level as well. The true half-life of a shock is close to 150% its estimate for inflation and more than seven times its estimate for investment. Even though the setting we have used to gauge the magnitude of the bias is stylized, in Section 4 we show that these conclusions

(which is nearly correct at high frequency), then (up to a constant):

$$p^*_{it} = (w_t - a_{it}) + (1-\alpha_L) l^*_{it}$$

where p^* and l^* denote the logarithms of frictionless price and employment, w_t and a_{it} are the logarithm of the nominal wage and productivity, and α_L is the labor share. It is straightforward to see that as long as the main source of idiosyncratic variance is demand, which we assume, $\sigma_{I_{p^*}} \simeq (1-\alpha_L)\sigma_{I_{l^*}}$. This approach gives similar numbers to the values used by Nakamura and Steinsson (2010) and Klenow and Kryvtsov (2008).

[11] Recall from Section 2 that the number of effective observations is given by the inverse of the Herfindahl index. For the CPI, the median (mean) total number of observations per month between 1988:02 and 2007:12 is 65,938 (66,822). The median (mean) *effective* number of observations per month during this period is 15,503 (15,276). The large difference comes from the fact that some items have much larger expenditure weights than other items.

[12] To go from the σ_I computed for employment in Caballero et al. (1997) to that of capital, we note that if the demand faced by a monopolistic competitive firm is isoelastic and its production function is Cobb-Douglas, then $\sigma_{I_{k^*}} \simeq \sigma_{I_{l^*}}$.

extend to more general settings.

3.3 What is behind the bias and slow convergence?

Having established the proposition and the practical relevance of the bias, let us turn to the intuition behind the proof of the proposition. We do this in two steps. We first describe the genesis of the bias, which can be seen most clearly when $N = 1$. We then show why, for realistic parameter values, the extreme bias identified for $N = 1$ vanishes very slowly as N grows.

3.3.1 The genesis of the bias

Let us set $\mu_A = 0$. From (8) we have that when $N = 1$, regardless of the true value of ρ,

$$\text{plim}_{T\to\infty}\hat{\rho} = 0. \tag{11}$$

That is, a researcher that uses a linear model to infer the speed of adjustment from the series for one unit will conclude that adjustment is infinitely fast independent of the true value of ρ. Of course, few would estimate a simple AR(1) for a series of one agent with lumpy adjustment, but the point here is not to discuss optimal estimation strategies for lumpy models but to illustrate the source of the bias step-by-step. The case $N = 1$ is a convenient starting point in this process.

The key point to notice is that when adjustment is lumpy, the correlation between this period's and the previous period's adjustment is zero, independent of the true value of ρ. To see why this is so, consider the covariance of Δy_t and Δy_{t-1}, noting that, because adjustment is complete whenever it occurs, we may re-write (2) as:

$$\Delta y_t \;=\; \xi_t \sum_{k=0}^{l_t-1} \Delta y^*_{t-k} \;=\; \begin{cases} \sum_{k=0}^{l_t-1} \Delta y^*_{t-k} & \text{if } \xi_t = 1, \\ \\ 0 & \text{if } \xi_t = 0, \end{cases} \tag{12}$$

where l_t denotes the number of periods, as of period t, since the last adjustment took place. So that $l_t = 1$ if the unit adjusted in period $t-1$, 2 if it did not adjust in $t-1$ and adjusted in $t-2$, and so on.

Table 2: CONSTRUCTING THE MAIN COVARIANCE

Adjust in $t-1$	Adjust in t	Δy_{t-1}	Δy_t	Contribution to $\text{Cov}(\Delta y_t, \Delta y_{t-1})$
No	No	0	0	$\Delta y_t \Delta y_{t-1} = 0$
No	Yes	0	Δy^*_t	$\Delta y_t \Delta y_{t-1} = 0$
Yes	No	$\sum_{k=0}^{l_{t-1}} \Delta y^*_{t-1-k}$	0	$\Delta y_t \Delta y_{t-1} = 0$
Yes	Yes	$\sum_{k=0}^{l_{t-1}} \Delta y^*_{t-1-k}$	Δy^*_t	$\text{Cov}(\Delta y_{t-1}, \Delta y_t) = 0$

There are four scenarios to consider when constructing the key covariance (see Table 2). If there

is no adjustment in this and/or the last period (three scenarios), then the product of this and last period's adjustment is zero, since at least one of the adjustments is zero. This leaves the case of adjustments in both periods as the only possible source of non-zero correlation between consecutive adjustments. Conditional on having adjusted both in t and $t-1$, we have

$$\text{Cov}(\Delta y_t, \Delta y_{t-1} \mid \xi_t = \xi_{t-1} = 1) = \text{Cov}(\Delta y_t^* \, , \, \Delta y_{t-1}^* + \Delta y_{t-2}^* + \cdots + \Delta y_{t-l_{t-1}-1}^*) = 0.$$

When a unit adjusts in consecutive periods the covariance between adjustments equals the covariance between shocks occurring during non-overlapping time intervals and is therefore equal to zero. Every time the unit adjusts, it catches up with all previous shocks it had not adjusted to and starts accumulating shocks anew. Thus, adjustments at different moments in time are uncorrelated.

The case $N = 1$ is also useful to compare the impulse responses inferred from linear models with those obtained from first principles. We define the latter via:

$$I_k \equiv \text{E}_t \left[\frac{\partial \Delta y_{t+k}}{\partial \Delta y_t^*} \right].$$

It follows from Proposition 1 that the impulse response of Δy to Δy^* inferred from a linear time-series model estimated for an individual series of Δy will be equal to one upon impact and zero for higher lags.

To calculate the correct impulse response, we note that Δy_{t+k} responds to Δy_t^* if and only if the first time the unit adjusted after the period t shock was in period $t+k$. It also follows from our Technical Assumptions that in this event the response is one-for-one. Thus

$$I_k = \Pr\{\xi_t = 0, \xi_{t+1} = 0, ..., \xi_{t+k-1} = 0, \xi_{t+k} = 1\} = (1-\rho)\rho^k. \tag{13}$$

This is the IRF for an AR(1) process obtained for *aggregate* inflation in the standard Calvo model (see, for example, Section 3.2 in Woodford, 2003).[13]

What happened to Wold's representation, according to which any process that is stationary and non-deterministic admits an (eventually infinite) MA representation? Why is Wold's representation in this case an i.i.d. process, suggesting an infinitely fast response to shocks, independent of the true persistence of shocks?

In general, Wold's representation is a distributed lag of the one-step-ahead *linear* forecast errors for the process. In the case we consider here we have $\text{E}[\Delta y_t \Delta y_{t+1}] = 0$ and therefore $\Delta y_{t+1} - \text{E}[\Delta y_{t+1} | \Delta y_t] = \Delta y_{t+1}$ so that the Wold innovation at time $t+1$, Δy_{t+1}, differs from the innovation of economic interest, Δy_{t+1}^*.

Wold's representation does not capture the entire process but only its first two moments. If

[13] As discussed in Caballero and Engel (2007), the impulse response for an individual unit and the corresponding aggregate will be the same for a broad class of macroeconomic models, including the one specified by the Technical Assumptions in Section 2.

higher moments are relevant, as is generally the case when working with variables that involve lumpy adjustment, the response of the process to the innovation process in Wold's representation will not capture the response to the economic innovation of interest. This misidentification will be present in any VAR model including variables with lumpy adjustment.

This fact has wide-ranging implications for applied macroeconomic researchers, which we explore in Section 3.5. In particular, it implies that estimated impulse response functions, and more generally, any structural test that derives from dynamic systems estimated in VAR style models will be biased.

3.3.2 Slow convergence

We have characterized the two extremes. When $N = 1$, the bias is maximum; when $N = \infty$ there is no bias. Next we explain how aggregation reduces the bias, and then study the speed at which convergence occurs.

For this purpose, we begin by writing $\hat{\rho}$ as an expression that involves sums and quotients of four different terms:

$$\text{plim}_{T\to\infty}\hat{\rho} = \frac{\text{Cov}(\Delta y_t^N, \Delta y_{t-1}^N)}{\text{Var}(\Delta y_t^N)} = \frac{\sum_i w_i^2 \text{Cov}(\Delta y_{1,t}, \Delta y_{1,t-1}) + \sum_{i\neq j} w_i w_j \text{Cov}(\Delta y_{1,t}, \Delta y_{2,t-1})}{\sum_i w_i^2 \text{Var}(\Delta y_{1,t}) + \sum_{i\neq j} w_i w_j \text{Cov}(\Delta y_{1,t}, \Delta y_{2,t})},$$

and since $N = 1/\sum_i w_i^2$ and $\sum_i w_i = 1$:

$$\text{plim}_{T\to\infty}\hat{\rho} = \frac{N\text{Cov}(\Delta y_{it}, \Delta y_{i,t-1}) + N(N-1)\text{Cov}(\Delta y_{it}, \Delta y_{j,t-1})}{N\text{Var}(\Delta y_{it}) + N(N-1)\text{Cov}(\Delta y_{it}, \Delta y_{jt})}, \tag{14}$$

where the subindices i and j in Δy denote two different units. Table 3 provides the expressions for the four terms that enter in the calculation of $\hat{\rho}$.

Table 3: CONSTRUCTING THE FIRST ORDER CORRELATION

	$\text{Cov}(\Delta y_{it}, \Delta y_{i,t-1})$	$\text{Cov}(\Delta y_{it}, \Delta y_{j,t-1})$	$\text{Var}(\Delta y_{it})$	$\text{Cov}(\Delta y_{it}, \Delta y_{jt})$
Lumpy ($\mu_A = 0$):	0	$\frac{1-\rho}{1+\rho}\rho\sigma_A^2$	$\sigma_A^2 + \sigma_I^2$	$\frac{1-\rho}{1+\rho}\sigma_A^2$
Lumpy ($\mu_A \neq 0$):	$-\rho\mu_A^2$	$\frac{1-\rho}{1+\rho}\rho\sigma_A^2$	$\sigma_A^2 + \sigma_I^2 + \frac{2\rho}{1-\rho}\mu_A^2$	$\frac{1-\rho}{1+\rho}\sigma_A^2$

If $N = 1$, only the two within-agent terms remain, one in the numerator and one in the denominator. Since the covariance in the numerator is zero,[14] $\hat{\rho}$ is zero as well. This drag on $\hat{\rho}$ remains present as N grows, but its relative importance declines since the between-agents covariances in the numerator and denominator are multiplied by terms of order N^2. This means that the reduction of the bias must come from the between-agents correlations at leads and lags, captured by the

[14] For simplicity we continue assuming $\mu_A = 0$.

second expressions, both in the numerator and denominator. The expression in the numerator is positive because not all individual units react to common shocks at the same time. The expression in the denominator is positive, because some do react at the same time. Either way, it is clear that these expressions are proportional to the variance in aggregate shocks only. In fact, as summarized in the first row of Table 3:

$$\text{Cov}(\Delta y_{it}, \Delta y_{i,t-1}) = \frac{1-\rho}{1+\rho}\rho\sigma_A^2,$$

$$\text{Cov}(\Delta y_{it}, \Delta y_{jt}) = \frac{1-\rho}{1+\rho}\sigma_A^2,$$

and we see that the ratio of the two between-agents covariance terms is indeed ρ. When N goes to infinity, it is this ratio that dominates $\hat{\rho}$.

While these between-agents terms are proportional to the variance of aggregate shocks only, the within-agent responsible for the biases are proportional to total variance. In particular, the denominator of (14) is

$$\text{Var}(\Delta y_{1,t}) = \sigma_A^2 + \sigma_I^2,$$

which cannot be compensated by the within-agent covariance in the numerator since this is equal to zero for the reasons described earlier. Thus $\hat{\rho}$ remains small even for large values of N when σ_I^2 is large.

Aside from the role played by the relative importance of idiosyncratic shocks for the bias, we see from the expression for K in Proposition 1 that the bias is larger when the drift is different from zero and when persistence is high. The latter is intuitive: When ρ is high, the between-agents covariances are small since adjustments across units are further apart, thus a larger number of units are required for these terms to dominate in the calculation of $\hat{\rho}$.

To understand the impact of the drift on convergence, we must explain why the covariance between Δy_t and Δy_{t-1} for a given unit is negative when $\mu_A \neq 0$ and why the variance term increases with $|\mu_A|$ (see the second row in Table 3). To provide the intuition for the negative covariance, assume $\mu_A > 0$ (the argument is analogous when $\mu_A < 0$) and note that the unconditional expectation of Δy_t is equal to μ_A, which corresponds to expected adjustment when adjusting in consecutive periods (the intuition is straightforward, see Appendix C for a formal proof). The expected adjustment when adjusting after more than one period is larger than μ_A. It follows that a value of Δy_t above average indicates that it is likely that the agent did not adjust in $t-1$, implying that Δy_{t-1} is likely to be smaller than average. Similarly, a value of Δy_t below average indicates that it is likely that the agent adjusted in period $t-1$, and Δy_{t-1} is likely to be larger than average in this case.

The reason why the variance term increases when $\mu_A \neq 0$ is that the dispersion of accumulated shocks is larger in this case, because by contrast with the case where $\mu_A = 0$, conditional on adjusting, the average adjustment increases with the number of periods since the unit last adjusted (it is equal to μ_A times the number of periods).

Summing up, linear time-series models use a combination of self- and cross-covariance terms

involving units' adjustments to estimate the microeconomic speed of adjustment. Inaction biases the self-covariance terms toward infinitely fast adjustment (and beyond when $\mu_A \neq 0$). It follows that the speed with which we recover the true value of ρ depends on the extent to which the cross-covariance terms play a dominant role. Since these terms recover ρ thanks to the common components in the adjustment of different units in consecutive periods, their contribution when estimating ρ will be smaller when adjustment is less frequent (larger ρ), and when idiosyncratic uncertainty is large relative to aggregate uncertainty.

3.4 Bias correction

This section studies an approach to correct for the missing persistence bias, based on using a proxy for the reset value y^*. In Appendix A we discuss two alternative approaches—one based on an ARMA representation of Δy_t^N and the other on instrumental variables.

So far we have assumed that the sluggishness parameter ρ is estimated using only information on the economic series of interest, y. Yet often the econometrician can resort to a proxy for the reset value y^*. Instead of (7), the estimating equation, which is valid for $N = \infty$, becomes:

$$\Delta y_t^N = \text{const.} + \rho \Delta y_{t-1}^N + (1-\rho)\Delta y_t^{*N} + e_t, \tag{15}$$

with some proxy available for the regressor Δy^*.

Equation (15) suggests correcting for the bias by using a proxy for the shock Δy^*. Since the regressors are orthogonal, from Proposition 1 we have that the coefficient on Δy_{t-1} will be biased downward. By contrast, the true speed of adjustment can be estimated directly from the parameter estimate associated with Δy_t^*, as long as the constraint that the sum of the coefficients on both regressors add up to one is *not* imposed. Of course, the estimate of ρ will be biased if the econometrician imposes the latter constraint. We summarize these results in the following proposition.

Proposition 2 (Bias with Regressors)

With the same notation and assumptions as in Proposition 1, consider the following equation:

$$\Delta y_t^N = \text{const.} + b_0 \Delta y_{t-1}^N + b_1 \Delta y_t^{*N} + e_t, \tag{16}$$

where Δy_t^{*N} denotes the average shock in period t, $\sum w_i \Delta y_{it}^*$. Then, if (16) *is estimated via OLS, and K defined as in (9),*

(i) without any restrictions on b_0 and b_1:

$$plim_{T\to\infty} \hat{b}_0 = \frac{K}{1+K}\rho, \tag{17}$$

$$plim_{T\to\infty} \hat{b}_1 = 1-\rho; \tag{18}$$

(ii) imposing $b_0 = 1 - b_1$*:*

$$plim_{T\to\infty}\hat{b}_0 = \rho - \frac{(1-\rho)^2}{K+1-\rho}.$$

Proof See Appendix C. ∎

Proposition 2 entails the general message that constructing a proxy for the reset variable y^* can be very useful when estimating the dynamics of a macroeconomic variable with lumpy microeconomic adjustment. This proposition also suggests not imposing constraints that hold only when $N = \infty$.

There is a third lesson implicit in Proposition 2, which is explained best in the more general setting we consider next. Assume that the actual process when $N = \infty$ satisfies

$$\Delta y_t^{\infty} = \sum_{k=1}^{p} a_k \Delta y_{t-k}^{\infty} + b\Delta y_t^{*\infty},$$

and suppose we want to estimate the impulse response of Δy for lags $0, 1, 2, ..., K$. A first possibility is to estimate via OLS the regression

$$\Delta y_t^N = \sum_{k=1}^{p} a_k \Delta y_{t-k}^N + b\Delta y_t^{*N} + e_t.$$

As we saw, the correlation between Δy_t^N and Δy_{t-k}^N is a biased estimator for the corresponding correlation when $N = \infty$,[15] which implies that the coefficients of the autoregressive polynomial will be biased as well, leading to a downward biased estimate of the adjustment speed.

By contrast, if we estimate

$$\Delta y_t^N = \sum_{k=0}^{K} I_k \Delta y_{t-k}^{*N} + e_t, \tag{19}$$

the estimated I_ks will be proportional to the true values, since the correlation between Δy_t^N and Δy_{t-k}^{N*} converges to the correlation when $N = \infty$ as the time-period under consideration tends to infinity. This leads to a consistent estimator for the speed of response. We could also include lags of Δy^N among regressors without biasing the estimate of the speed of response, as long as the lags involved are larger than K.[16]

Alternatively, since the regressors in (19) are orthogonal, we could estimate I_k from

$$\Delta y_t^N = I_k \Delta y_{t-k}^{*N} + e_t.$$

[15] More precisely, the correlation with N does not converge to the correlation with $N = \infty$ when the time series length, T, tends to infinity.

[16] Proposition 2 corresponds to the particular case of this principle when $K = 1$ and when the entire IRF can be inferred from the estimated coefficient for Δy^*, which stops being true when $K > 1$.

This involves estimating $K+1$ regressions via OLS. This approach, which is a particular case of Jorda's (2005) methodology for estimating IRFs, also leads to a consistent estimate of the speed of adjustment. It is more robust than using (19) but less parsimonious, since we can impose that the I_k are a function of a small number of parameters only when using (19).

The first application we consider in Section 5 applies Proposition 2 directly. The second application considers the regressiones similar to (19) to obtain estimates of the IRF that are immune to the missing persistence bias.

3.5 Implications for empirical reseachers

This section studies the implications of the missing persistence for two important tools in the applied macroeconomic toolkit: the estimation of impulse response functions and simulation based estimators.

3.5.1 Estimating IRFs

There are two main methods for estimating impulse response functions (IRFs) to an identified structural shock (Ramey 2016). The standard method, which we refer to as the "VAR approach," is to estimate a vector autoregression and use the estimated system of equations to compute the IRF. A second method, which we refer to as the "MA" approach, is to regress the series of interest, for example π_t, on k lags of the structural shock, $\epsilon_t, \epsilon_{t-1}, \epsilon_{t-2}, ..., \epsilon_{t-k}$ where each estimated coefficient is an element of this impulse response function. This approach is closely related to Jorda's (2005) local projection approach to estimating IRFs.

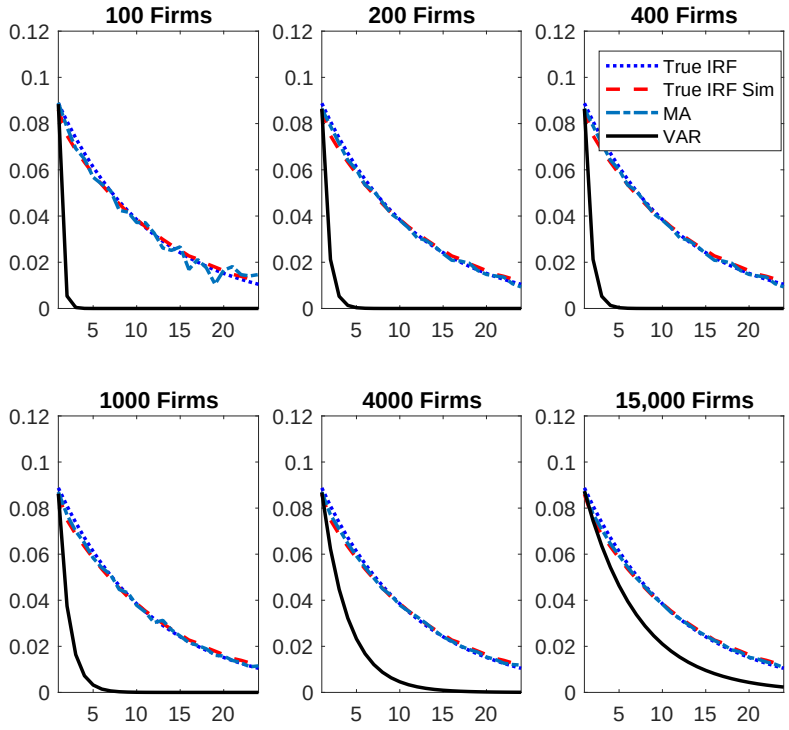
There are many reasons why the VAR approach is, by far, the most commonly used method to estimate IRFs, despite the fact that these two methods of computing IRFs are equivalent in infinitely long samples (T = ∞) for linear models.[17] Prominent among them is parsimony: in most applications the VAR approach can achieve similar precision with fewer parameters than the MA approach. When using a VAR with p lags the number of parameters that are estimated will be proportional to p, independently of the number of lags of interest in the IRF. By contrast, under the MA approach, the number of parameters' estimated is proportional to the number of lags desired for the IRF. This inefficiency is particularly costly to applied macroeconomists because their samples are often small. Also, since the MA approach imposes essentially no restriction on the shape of the IRF, which is not the case for a low order VAR, IRFs estimated with the MA approach are less precisely estimated and can behave erratically.

Despite these limitations, the MA approach has some merits. Ramey (2016) argues that the MA approach is more robust when the estimated VAR is misspecified, which might happen if the true dynamics are non-linear. In this case, the VAR approach will compound these specification errors

[17] See Christiano, Eichenbaum and Evans (1999) for details.

at each horizon of the IRF. Here we highlight a second reason to prefer to the MA approach: it is robust to the missing persistence bias.

Figure 1: Response of Inflation to a Nominal Shock in a GE Calvo Model



This figure shows the IRF of inflation to a nominal shock computed in four separate ways. 1) Using the analytical expression in equation 13 (blue dots); 2) The average (across 100 simulations) of the true non-linear IRF in the model computed via simulation (red dash); 3) Using our MA methodology (light blue dot-dashed) 4) Using our VAR methodology (black solid line).

Consider the following simple example. A policy maker wishes to estimate the response of inflation to a monetary policy shock ignoring the fact that price setting is subject to lumpy adjustment. The standard VAR approach would be to estimate a parsimonious VAR using data on output, inflation and interest rates and impose a timing assumption (e.g. Cholesky) in order to identify the structural shock to monetary policy. One would then estimate this series of equations by OLS and use the entire system of equations to compute the IRFs, which are non-linear function of the estimated VAR equations. Because of lumpy adjustment, the coefficient on lagged inflation would be biased downwards, biasing the estimates of all of the IRFs. In contrast, since the MA approach never regresses a variable subject to lumpy adjustment on lags of itself,[18] the missing persistence bias would not be relevant. Thus we would expect that the VAR and MA approaches would give

[18]This assumes there is no bias in the estimation of the interest rate, which will be the case if the structural interest rate shock is identified as the residual from the interest rate equation and then inflation is regressed on this shock and its lags.

different results if the missing persistence bias were present and important.

We explore the scenario described above using the standard Calvo model of price setting where the only novelty is that we vary the number of underlying agents in the economy instead of assuming an infinite number as is usually done. We then compute IRFs using both methods and examine whether the missing persistence bias is present in the VAR approach. For comparability with our menu cost results (Section 4.1) and our empirical examples (Section 5), we use the calibration of Nakamura and Steinsson (2010), which is chosen to match relevant moments of CPI microdata.[19] We use four methods to compute IRFs and Figure 1 shows the average IRFs from 100 simulations.

The first method is analytical. . As derived in (13), given our assumptions the response of inflation in period $t+k$ to a nominal shock ϵ_t is:

$$E_t\left[\frac{\Delta\pi_{t+k}}{\partial\epsilon_t}\right] = (1-\rho)\rho^k$$

This is shown in the dotted line. The second procedure, shown in the dashed line, uses a simple Monte Carlo ("Simulation") method where the IRF is the response of π to a one grid point increment of Δ of the nominal shock at time t relative to a world where this shock did not occur. In particular, we compute the IRF as

$$E_t\left[\frac{\partial\pi_{t+k}}{\partial\epsilon_t}\right] = (E_t[\pi_t|\epsilon_t=\Delta] - E_t[\pi_t|\epsilon_t=0])/\Delta$$

Given that the Monte Carlo method is not polluted by lumpy adjustment if we use the true number of agents in the simulations, the estimated IRFs will not be biased.[20] Finally, we estimate IRFs using both the VAR and MA approaches. They are the solid and dashed-dot lines respectively in Figure 1.

As expected, the Monte Carlo method closely approximates the true response for all N. Two other results jump out. First and consistent with the results in Table 5, the bias is substantial for the VAR approach, particularly for small N. The estimated IRF using this approach is always below the true response. Thus researchers using this approach will infer much faster adjustment to nominal shocks than exists in the model. Second, the MA approach does a good job of estimating the true IRF even in small samples. This suggests that this methodology is a robust way of dealing with the missing persistence bias. Overall, this exercise provides support for using the Jorda (2005) methodology, as it is robust to both misspecification and the missing persistence bias.

[19] We follow their calibration exactly including allowing the idiosyncratic shock to be AR(1) rather than a random walk (a deviation from our baseline assumptions). The parameter values are: $\mu_A = 0.0021$, $\sigma_A = 0.0032$, $\sigma_I = 0.0425$, $\rho_I = 0.66$ and $K = 0.0245$ which implies that $\rho = 0.91$. Results are very similar if we use our baseline calibration.

[20] This method will be more useful later on, when we compute IRFs that lack analytical solutions, such as IRFs for Ss models.

3.5.2 Simulation based estimators

Simulation based estimators are a common way of estimating macroeconomic models because inference only requires the ability to simulate data from the economic model rather than needing to deal with an often analytically intractable or difficult to evaluate likelihood function. Indirect inference is an approach used frequently in this context (Smith, 2008). The goal of indirect inference is to choose the parameters of the economic model so that the observed data and the simulated data look the same from the vantage point of some moments or "auxiliary model", which are both informative about the underlying structural parameters and can easily be computed in both the model and the data. The parameters of the underlying economic model are then chosen so as to minimize the difference between the parameter estimates of the auxiliary model in the model and in the data. Under mild assumptions, this approach will identify the structural parameters of interest.

A good example of this approach is the classic Christiano, Eichenbaum and Evans (2005) paper,[21] which seeks to explain the dynamic response of inflation to an identified monetary policy shock. In the language of indirect inference, their auxiliary model is the IRF of eight macroeconomic variables to a monetary policy shock where these IRFs are computed from an identified VAR.[22] They then estimate six parameters of their medium scale DSGE model by minimizing the distance between these eight impulse response functions and their counterparts in the model.

While indirect inference has many virtues, this methodology must be applied with care if the missing persistence bias is present. Consider the above example. We know from the previous subsection that when an underlying variable has lumpy adjustment and IRFs are estimated using the "VAR" approach, the estimates of the IRF will be biased. This bias in the estimation of the auxiliary equation can translate into bias in the estimates of the underlying structural parameters.

One solution to this issue is to estimate IRFs using a methodology that is robust to the missing persistence bias such as Jorda (2005). A more general solution is to simulate data in exactly the same form as the researcher has access to in reality. In particular, it is crucial to use actual sample sizes when estimating the auxiliary model: if the researcher simulates much larger samples of data in the model then one would eliminate the missing persistence bias in the model but not in the data, potentially biasing the estimates of the parameters of interest.

Table 4 illustrates this point for a simple Monte Carlo simulation that builds on our previous Calvo model. Consider an applied researcher who wants to estimate the frequency of adjustment (the structural parameter) by SMM using the impulse response function of inflation to a nominal shock as the auxiliary model. This IRF is a sensible choice since the k^{th} element of the IRF is equal to $\rho^k(1-\rho)$.[23] Assume that there are 400 price setting firms in the data who all use Calvo pricing with the same frequency of adjustment, 1-ρ, equal to 0.25. The data moment is the IRF of inflation

[21] A similar estimation procedure can be found in Rotemberg and Woodford (1997), Amato and Laubach (2003), Gilchrist and Williams (2000) and Boivin and Giannoni (2006).

[22] This is the "VAR" approach discussed in the previous sub-section.

[23] Obviously, this is a highly stylized example – in more complicated frameworks this IRF would depend on more than one structural parameter. The example is kept deliberately simple to illustrate the main point.

Table 4: SMM TABLE

Monte Carlo example: matching IRFs by simulated method of moments (SMM)

	Estimator	Weight Matrix	**Model moments** Effective number of agents (N) in simulation			
			400	1,000	4,000	15,000
Data	VAR	Identity	0.250	0.730	0.820	0.840
($N = 400$)		Proportional	0.250	0.510	0.760	0.770
($1-\rho = 0.25$)		Optimal	0.250	0.710	0.820	0.840
Data	MA	Identity	0.250	0.250	0.250	0.250
($N = 400$)		Proportional	0.250	0.250	0.250	0.250
($1-\rho = 0.25$)		Optimal	0.250	0.250	0.250	0.250

This table documents that it is important to treat real and simulated data similarly when the missing persistence bias is present using a simple Monte-Carlo. The number of underlying agents is 400 in the "Data". We compute the IRF of inflation to a nominal shock in two ways: the VAR approach (top panel) and MA approach (bottom panel). The true frequency of adjustment, $1-\rho = 0.25$. We compute the analogous model implied IRF by simulation. The only difference across the simulations is the number of underlying agents used to calculate this IRF: we vary the number of units from 400 to 15,000 to allow for comparability with Figure 1. All rows show the estimated $1-\hat{\rho}$ from the SMM estimation and all results are averages across 100 simulations.

to a nominal shock computed in this model.

The top panel of Table 4 illustrates the case when both the data and model IRFs are computed using the standard VAR approach. Each row shows the results from the SMM estimation for three different weight matrices, while each column varies the number of underlying firms when the researcher estimates the IRF. In all cases we compute averages of the model moments across 100 simulations. Two results are clear. The first column shows that the SMM estimator provides an unbiased estimator of the frequency of adjustment when the researchers simulation has the same number of firms in the model as are in the data. This gives support for the folk wisdom that researchers should treat real and simulated data similarly.

The perils of not doing this are shown in the other three columns. Since the underlying data has 400 firms, the missing persistence bias is severe. If a researcher tried to match this IRF using a simulation with 15,000 firms, she would infer a much faster speed of adjustment as shown by the last column of Table 4. The reason is that the VAR approach is subject to the missing persistence bias and this bias diminishes with the number of effective firms (compare the top left panel of Figure 1 which shows the IRF for 100 firms to the bottom right panel which shows the IRF with 15,000 firms). The only way to match the biased data estimate with an unbiased estimate is by increasing the frequency of adjustment – this is why the estimated frequency increases as one moves from left to right across the columns. In contrast, the bottom panel shows that no such issue exists if IRFs are

estimated by the MA approach. This is because this approach is immune to the missing persistence bias.

4 Robustness

The Technical Assumptions we made so far (Calvo adjustment, no strategic complementarities and i.i.d. innovations, see Section 2) allowed for closed form expressions and simple intuitions for the missing persistence bias. In this section, we show that the bias is significant under more general assumptions. We focus on two departures from our baseline that are motivated by empirical realism: allowing for the probability of adjustment to be state-dependent (Ss pricing) and allowing for agents' decisions to be strategic complements. In Appendix B we consider two further extensions. There we relax the assumption that y^* follows a random walk and we allow for time to build. We show that the missing persistence bias continues to be present (and significant) in all of these cases.

4.1 State-dependent models

The intuition we provided in Section 3 for the missing persistence bias is based on three assumptions: adjustment is lumpy, there are no strategic complementarities and innovations (the Δy^*) are independent across periods. Thus the correlation between Δy_t and Δy_{t-1} for a unit is zero either because the agent did not adjust in one of the periods or because adjustments at different points in time are independent. This intuition does not depend on whether agents' adjustments are determined by an exogenous process (as in the Calvo model considered so far) or state-dependent (as with Ss-type models) since in both pricing models agents fully adjust to all shocks they have faced since they last adjusted.[24] In other words, Table 2 in Section 3.3.1 continues to be valid when adjustment policies are state-dependent because in these models we also have that shocks in non-overlapping time periods are independent when y^* follows a random walk.[25]

Thus the main ingredient for the missing persistence bias is valid both for models with constant and state-dependent adjustment hazards, all that matters is that consecutive adjustments are uncorrelated. Of course, the statistics of interest will be different across both types of models and we no longer have closed form expressions for our missing persistence bias sufficient statistic (Equation (9)). However, we can examine the magnitude of the bias numerically. Specifically, we generate an analogous table for Ss adjustment (Table 5) to the one we used to provide back-of-the-envelop estimates of the bias for Calvo adjustment (Table 1).

We focus on the case of prices for brevity and because it maps closely to our empirical application in Section 5. Just like Table 1, Table 5 provides two different estimates of persistence. The first

[24] Here the assumption of no strategic complementarities is crucial. We consider the case with strategic complementarities in Section 4.2.

[25] Jorda (1997) provides a general characterization of these models in terms of random point processes (processes with highly localized data distributed randomly in time).

Table 5: SLOW CONVERGENCE AND SS ADJUSTMENT

$\hat{\rho}$ and Half-Life with Ss adjustment

Persistence measure	Calibration	Effective number of agents (N)					
		100	400	1,000	4,000	15,000	40,000
$\hat{\rho}$	Baseline	0.074	0.175	0.253	0.317	0.343	0.346
	NS 2010	0.141	0.352	0.469	0.552	0.585	0.589
Half-life	Baseline	0.560	0.674	0.755	0.825	0.850	0.866
	NS 2010	0.563	0.721	0.878	1.120	1.251	1.345

This table documents how the bias varies with the effective number of units for a model with Ss adjustment. We consider two different persistence measures. The first two rows show results for estimated ρ, $\hat{\rho}$, which is computed using equation (8). The third and fourth rows report the corresponding half-life. The half-life is inferred directly from the average simulated IRF computed by the VAR method in Figure 2. We consider two calibrations. The first (baseline) uses the same parameter values as our baseline Calvo calibration ($\mu_A = 0.003$, $\sigma_A = 0.0054$, $\sigma_I = 0.048$) and picks the size of the menu cost, $\kappa = 0.062$ to match $\rho = 0.86$. The second calibration uses the same parameter values used by Nakamura and Steinsson (2010) to calibrate a menu cost model. We follow their calibration exactly including allowing the idiosyncratic shock to be AR(1) rather than a random walk. The parameter values are: $\mu_A = 0.0021$, $\sigma_A = 0.0032$, $\sigma_I = 0.0425$, $\rho_I = 0.66$ and $\kappa = 0.0245$ which implies that $\rho = 0.91$.

two rows display the estimated AR(1) persistence, $\hat{\rho}$ computed using equations (8) and (9), while rows three and four report results for the implied half-life.

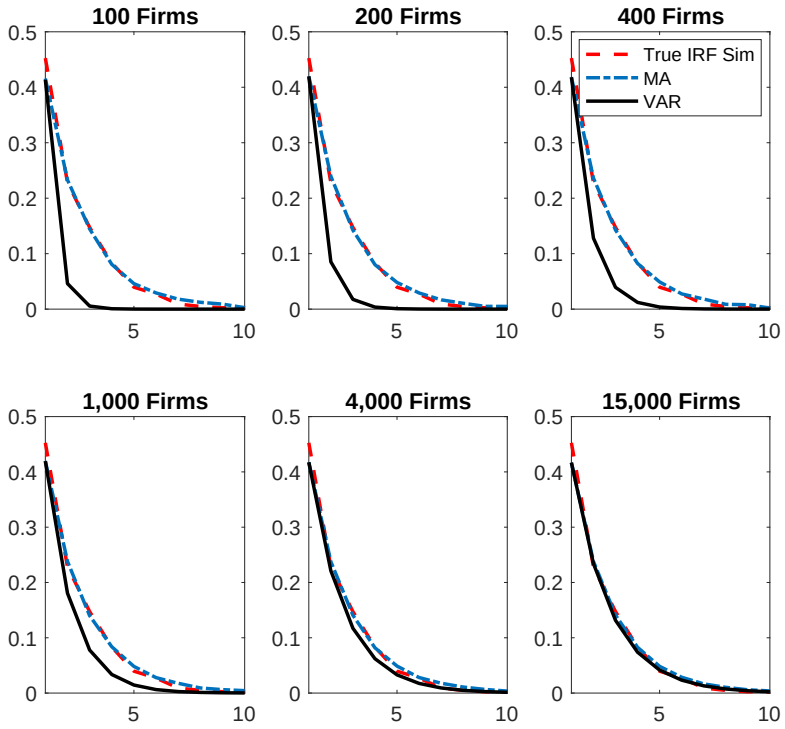
We consider two calibrations. The first (baseline) uses the same parameter values as our baseline Calvo calibration ($\mu_A = 0.003$, $\sigma_A = 0.0054$, $\sigma_I = 0.048$) with one exception. With Ss adjustment the frequency of adjustment is no longer a primitive so we calibrate the size of the menu cost, $\kappa = 0.062$, to match the frequency of adjustment ($\rho = 0.86$) we used in Table 1. For our second calibration, we take parameters directly from Nakamura and Steinsson (2010). We follow their calibration exactly including allowing the idiosyncratic shock to be AR(1) rather than a random walk (a deviation from our baseline assumptions). The parameter values are: $\mu_A = 0.0021$, $\sigma_A = 0.0032$, $\sigma_I = 0.0425$, $\rho_I = 0.66$ and $\kappa = 0.0245$ which implies that $\rho = 0.91$. These parameters were chosen to match moments of the CPI microdata.

Examining Table 5, the bias is substantial for small N for both calibrations. When the number of agents is 400 the estimated half-life is biased downward by 47% and remains biased by 35% even when $N = 1,000$. Similar to the Calvo case, the bias decreases substantially with N. Table 5 shows that one difference between the Calvo and Ss case is that the bias diminishes more quickly and thus is less quantitatively relevant for large N (only 8%). Overall this suggests that in the Ss case, the bias remains large at the sectoral level but not in the aggregate.[26]

[26]Another difference between the Calvo and Ss cases is that the overall level of persistence is lower in the Ss model. This is unsurprising since it is well-known that the Calvo model implies significantly more monetary non-neutrality than an equivalent Ss model (Golosov and Lucas, 2008).

The bias in estimated IRFs also remains in the Ss case. To see this, we calibrate a standard menu cost model of Nakamura and Steinsson (2010) using their calibration and compute IRFs using both the VAR and MA methods. The details of the model are given in Appendix G.1. We compare the true IRF estimated by Monte Carlo methods, to the VAR and MA approaches discussed in Section 3.5.

Figure 2: Response of Inflation to a Nominal Shock in a GE Menu Cost Model



This figure shows the IRF of inflation to a nominal shock computed in three separate ways. 1) The average (across 100 simulations) of the true non-linear IRF in the model computed via simulation (red dash); 2) Using our MA methodology (light blue dot-dashed) 3) Using our VAR methodology (black solid line).

The results are shown in Figure 2. Consistent with the results in Table 5, the bias is substantial for small and medium N. As we document in the next section, the mean, median and maximum number of effective observations in each of our 66 CPI sectors[27] is 187, 142 and 980 respectively. A visual inspection of Figure 2 suggests the bias is large for N of this size. It is only when sample sizes approach numbers that are representative of the entire CPI ($N = 15{,}000$) that the bias becomes small.

Which benchmark model is closer to the data? Recent work by Alvarez, Le Behan and Lippi (2016) concludes that a model in between a Calvo and Ss model best matches the relevant micro evidence.[28] This conclusion is consistent with previous empirical work by Nakamura and Steinsson

[27] Our definition of sectors is close to a two digit level of disaggregation.

[28] Specifically, they "review empirical measures of kurtosis and frequency and conclude that a model that successfully

(2008), Klenow and Kryvtsov (2008) and Midrigan (2011) which all found evidence consistent with a hybrid model in U.S. microdata.

Thus, the main message of Section 3 remains in the presence of Ss adjustment. Our results also suggest that researchers need to be careful when using simulated methods of moments or indirect inference to calibrate or estimate parameters for a DSGE model when lumpy adjustment is present. Using the correct number of agents is important, otherwise the parameters that are obtained and IRFs are likely to be biased.

4.2 Strategic complementarities

Under the Technical Assumptions from Section 2, agents' decision variables are neither strategic complements nor strategic substitutes. This may not be a reasonable assumption. For example, in the pricing literature many authors have argued that strategic complementarities are a central element to match the persistence suggested by VAR evidence (Woodford, 2003; Christiano, Eichebaum and Evans, 1999, 2005; Clarida, Gali and Gertler, 2000; Gopinath and Itskhoki, 2010).

This observation motivates considering the case where the y^* are strategic complements. Following Woodford (2003, section 3.2) we assume that log-nominal income follows a random walk with innovations ε_t. Aggregate inflation, π_t, then follows an AR(1) process

$$\pi_t = \phi\pi_{t-1} + (1-\phi)\varepsilon_t$$

with $\phi > \rho$ when prices are strategic complements. In line with the strategic complementarity parameters advocated by Woodford, we assume $\phi = 0.944$. The true half-life of shocks increases from 4.6 to 12.1 months and the expected response time from 6.1 to 16.9 months.

Under these assumptions, $\Delta \log p_t^*$ follows the following ARMA(1,1) process:

$$\Delta \log p_t^* = \phi \Delta \log p_{t-1}^* + c(\varepsilon_t - \rho\varepsilon_{t-1}),$$

with $c = (1-\phi)/(1-\rho)$.[29]

The second and fourth rows in Table 6 present the AR(1) persistence measure, $\hat{\rho}$, and estimated half-life, respectively, in this setting. The first and third rows reproduce the values for the case with no strategic complementarities (Table 1). The bias is larger with strategic complementarities: With 15,000 units, which corresponds to approximately the effective number of prices considered when calculating the CPI, the estimated half-life is approximate one-third of its true value, compared with 60 percent of its true value in the case with no complementarities.

The intuition is the following. Section 3.3.2 showed that identification of $\hat{\rho}$ comes from cross-item terms with the speed of convergence to the true ρ increasing in the aggregate signal, σ_A, and

matches the micro evidence on kurtosis and frequency produces real effects that are about four times larger than in the Golosov-Lucas model, and about 30 percent below those of the Calvo model."

[29] In the notation of Section 2 we have $b(L) = (1-\phi L)/(1-\rho L)$.

Table 6: SLOW CONVERGENCE AND STRATEGIC COMPLEMENTARITIES

$\hat{\rho}$ and Half-Life of Shocks with Strategic Complementarities

Persistence measure	$\underline{\rho}$	$\underline{\phi}$	Effective number of agents (N)						
			100	400	1,000	4,000	10,000	40,000	∞
$\hat{\rho}$	0.8600	0.8600	0.070	0.232	0.415	0.679	0.777	0.838	0.860
	0.8600	0.9442	0.029	0.115	0.246	0.555	0.738	0.882	0.944
Half-life	0.8600	0.8600	0.261	0.475	0.788	1.788	2.748	3.913	4.596
	0.8600	0.9422	0.196	0.321	0.495	1.177	2.277	5.542	12.072

First two rows show the estimated ρ, $\hat{\rho}$, which is computed using equation (8) and the parameter values listed below. Parameters: $\rho = 0.86$, $\mu_A = 0.003$, $\sigma_A = 0.0054$, $\sigma_I = 0.048$. Rows 3-4 show results when the half-life is the measure of persistence. The half-life is inferred from the estimated ρ's in the first two rows and is computed using the following formula: $-\log 2/\log \hat{\rho}$.

decreasing in the idiosyncratic noise, σ_I. All other things equal, this means that strategic complementarities weaken the strength of the aggregate signal, slowing convergence.[30] When strategic complementarities are present and agents adjust, they no longer adjust fully to the aggregate shocks that accumulated since the last time they adjusted. This decreases the strength of these cross-item terms, leading to slower convergence.

5 Applications

The pricing literature is a natural context in which to study the relevance of the missing persistence bias because numerous studies over the last decade have shown that at the item level prices adjust infrequently.[31] The two applications we present next provide evidence of the presence of the bias and correct for it using the approach outlined in Section 3.4, based on an estimate for the aggregate and sectoral shocks facing retail price-setters, obtained from establishment level prices.

Our first example shows that accounting for the missing persistence bias explains a puzzling finding in Bils and Klenow's now classic 2004 paper (henceforth BK). Figure 2 in BK shows that the response of sectoral prices to shocks estimated from a linear time-series model is much faster than suggested by the Calvo model, which raises the question of whether this difference may be due to the missing persistence bias. this paper— we We start with this example because (i) the assumptions are identical to those underlying the results in Section 3 (ii) it highlights that the missing

[30] There's a counterveiling effect because the firm's own-price-change correlation now is positive. Yet the impact of this effect on aggregate inflation decreases fast as the number of firms grows.

[31] For evidence based on the micro database used to calculate the CPI see Bils and Klenow (2004), Nakamura and Steinsson (2008) and Klenow and Kryvtsov (2008).

persistence bias is relevant in U.S. pricing data at the sectoral level and (iii) we are able to calculate the exact magnitude of the bias in this case from the CPI micro database. We show that bias is substantial and then proceed to correct it, finding that this solves the puzzle.

In our second application, we turn to recent empirical work using sectoral price data to argue that firms respond faster to sectoral shocks than to aggregate shocks (Boivin, Giannoni and Mihov, 2009; Mackoviak, Moench and Wiederholt, 2009). These results have been interpreted as evidence in favor of rational inattention or imperfect information models of price setting, because they suggest that firms respond more to bigger, more salient shocks. However, we show that once the missing persistence bias is accounted for, there is little evidence that sectoral prices respond faster to sectoral shocks than to aggregate shocks.

5.1 A simple test of the Calvo model

In Bils and Klenow's influential 2004 paper the authors conduct a simple test of the Calvo model using CPI microdata (see Figure 2 in their paper). They start by using the micro data to estimate the frequency of price adjustment in each sector, λ_s. Next, they estimate the following regression by OLS:

$$\pi_{st} = \rho_s \pi_{s,t-1} + e_{st}, \tag{20}$$

where π_{st} is inflation in sector s at time t.

Under the assumptions of the Calvo pricing model considered in Section 3 with $N = \infty$, which happen to be the same assumptions considered by BK, we should find that $\hat{\rho}_s$ is approximately equal to $1 - \hat{\lambda}_s$. In contrast, BK find that in all sectors $\hat{\rho}_s$ is substantially smaller than $1 - \hat{\lambda}_s$ and interpret this as strong evidence against the Calvo model.

Our paper suggests a more cautious interpretation of that finding. What BK show is that the persistence of shocks inferred from a linear time-series model estimated on sectoral data is considerably smaller than the true persistence parameter inferred from microeconomic retail pricing data. Since price adjustment is lumpy and small samples underly the construction of the sectoral inflation series, the missing persistence bias could also explain BK's result.

Next we test this assertion. Notice also that BK's estimating equation, equation 20, is identical to the situation considered in Proposition 1 in Section 3.1. This means that we can test whether the missing persistence bias is responsible for BK's result using the bias correction approach outlined in Section 3.4. We implement this approach using the BLS microdata and show that once we correct for this bias the systematic difference between $\hat{\rho}_s$ and $1 - \hat{\lambda}_s$ disappears.[32]

We proceed in two steps. First, we use the reset price inflation methodology of Bils, Klenow and Malin (2012) to estimate sector specific reset price inflation series, v_{st}, using the CPI micro data. Bils, Klenow and Malin (2012) show that reset price inflation is an unbiased estimate of sectoral shocks in a variety of standard models. Second, we then use the bias correction approach from

[32] For an alternative explanation for the bias see Le Bihan and Matheron (2012)

Section 3.4 to obtain estimates for ρ_s that are immune to the missing persistence bias. We find that the bias correction method does a good job, that is, we find that $\hat{\rho}_s \simeq 1 - \hat{\lambda}_s$.

The basic idea behind reset price inflation is to make inferences about the underlying shocks using information contained only in observed price changes where the implicit assumption is that when a firm adjusts it is adjusting ("resetting") to its optimal price. Specifically, define $p_{i,t}$ as the log price of item i and time t and define a price change indicator as:

$$I_{i,t} = \begin{cases} 1 & \text{if} \quad p_{i,t} \neq p_{i,t-1}, \\ 0 & \text{if} \quad p_{i,t} = p_{i,t-1}. \end{cases}$$

The reset price, $p_{i,t}^{\text{reset}}$, for prices that do not change is simply the current price. The reset price for non-changers is then updated using the rate of reset price inflation estimated from the price changers in the current period:

$$p_{i,t}^{\text{reset}} = \begin{cases} p_{i,t} & I_{i,t} = 1, \\ p_{i,t-1} + \pi_t^{\text{reset}} & I_{i,t} = 0. \end{cases}.$$

Given $p_{i,t-1}^{\text{reset}}$, define reset price inflation, π_t^{reset}, as:

$$\pi_t^{\text{reset}} = \frac{\sum_i \omega_{i,t} \left(p_{i,t} - p_{i,t-1}^{\text{reset}} \right) I_{i,t}}{\sum_i \omega_{i,t} I_{i,t}},$$

where $\omega_{i,t}$ denote i's relative expenditure weight at time t. Thus reset price inflation is the "inflation rate" conditional on the price adjustment. With Calvo price setting and assuming that the technical assumptions in Section 3 hold, it is easy to show that reset price inflation reduces to the following formula:[33]

$$\pi_t^{\text{reset}} = \frac{\pi_t - \rho \pi_{t-1}}{(1-\rho)} = \nu_t^A$$

This justifies using reset price inflation as an estimate of sectoral shocks. In Appendix G.3 we present simulation results showing that reset price inflation is also a good method to recover the true shock innovations in both more realistic Calvo environments with large idiosyncratic shocks and Ss-type settings.[34]

We implement both the reset price inflation methodology and our bias correction approach using micro data on prices from the BLS. We use the CPI research database which contains individual price observations for the thousands of non-shelter items underlying the CPI over the sample period 1988:03-2007:12. Prices are collected monthly for all items only in New York, Los Angeles

[33] This holds in the limit as the number of price setters becomes large so that the frequencies are exact and the idiosyncratic shocks average out.

[34] We also tried estimating the shocks using a repeat-price-change approach (similar to the Case-Shiller index) and found similar results.

and Chicago, and we restrict our analysis to these cities to ensure the representativeness of our sample.[35] The database contains thousands of individual "quote-lines" with price observations for many months. In our data set, an average month contains approximately 12,000-15,000 different quote-lines. Quote-lines are the highest level of disaggregation possible and correspond to an individual item at a particular outlet. An example of a quote-line collected in the research database is a16 oz bag of frozen corn at a particular Chicago outlet.

Much of the recent literature has discussed the difference between sales, regular price changes and product substitutions. We exclude sales following Eichenbaum, Jaimovich, and Rebelo (2012) and Kehoe and Midrigan (2016), who argue that the behavior of sales is often significantly different from that of regular or reference prices and that regular prices are likely to be the object of interest for aggregate dynamics. We exclude product substitutions because these require a judgement on what portion of a price change is due to quality adjustment and which component is a pure price change. This introduces measurement error in the calculation of price changes at the time of product substitution. Bils (2009) shows that these errors can be substantial.[36]

We work with the two-digit or "Expenditure class" level of aggregation rather than the ELI level of aggregation used in BK because we will need to estimate underlying shocks when correcting for the bias and this level of aggregation provides a good balance between having a sufficiently large number of sectors and being able to obtain good estimates for underlying shocks.[37] This leaves us with 66 sectors.

Once we have our 66 reset price inflation estimates, we implement our bias correction procedure by including our measure of the sectoral shock, υ_{st}, as an additional control in equation (16):

$$\pi_{st} = \beta_s \pi_{s,t-1} + \gamma_s \upsilon_{st} + e_{st}. \tag{21}$$

Proposition 2 from Section 3.4 implies that if we estimate β_s and γ_s in the above equation without imposing any constraints across them, then $\widehat{\gamma}_s$ will be an unbiased estimate of the actual fraction of adjusters λ_s. We then examine how close $\widehat{\gamma}_s$ is to λ_s.

As a first step we replicate BK's results using our 66 sectors. In particular, we estimate equation (20) using the micro data, and denote the implied frequency of adjustment estimates as $\lambda_s^{\text{VAR}} = 1-\widehat{\beta}_s$. As in BK, we find that $\widehat{\beta}_s \ll 1-\lambda_s^{\text{micro}}$, where λ_s^{micro} denotes the true frequency of adjustment, estimated from the micro level quote-lines. Across all 66 sectors, the mean (median) estimate of $\widehat{\beta}_s$ is 0.08 (0.06) compared to 0.88 (0.93) for $1-\lambda_s^{\text{micro}}$ and $\widehat{\beta}_s < 1-\lambda_s^{\text{micro}}$ in all sectors, with the exception of only one. Now that we have established that BK's baseline result holds in our dataset, we

[35] The most representative sample would be to use all bimonthly observations, but then many price changes are potentially missing. Some items are sampled monthly outside of New York, Los Angeles and Chicago, but these items are not representative, so we restrict our monthly analysis to these three cities.

[36] Nevertheless, we have also repeated the analysis including product substitutions and found similar results.

[37] We only use representative monthly pricing data in constructing our price indices to be able to measure monthly shocks, which cuts down our underlying sample sizes significantly when compared to using bimonthly data as well. Also, we only chose those sectors for which we could have data for the entire sample period.

implement our bias correction procedure by estimating equation (21) using our constructed shock measure, v_{st}.

We start with some definitions. Denote the coefficient on our sectoral reset price inflation measure by $\lambda_s^c = \widehat{\gamma}_s$, where the superindex c stands for "corrected". Define $\lambda_s^{\text{VAR}} = 1 - \widehat{\beta}_s$ where $\widehat{\beta}_s$ is estimated using equation (20). To gauge the extent to which the λ_s^c correct the missing persistence bias, we regress the change in estimated speed of adjustment we achieve in a given sector, $\lambda_s^c - \lambda_s^{\text{VAR}}$, on the magnitude of the bias, $\lambda_s^{\text{micro}} - \lambda_s^{\text{VAR}}$. That is, since we are in a rare situation where we actually know the bias, we are able to estimate by OLS the following equation:

$$(\lambda_s^c - \lambda_s^{\text{VAR}}) = \alpha + \eta \text{bias}_s + \epsilon_s, \tag{22}$$

with $\text{bias}_s \equiv \lambda_s^{\text{micro}} - \lambda_s^{\text{VAR}}$. Here η is the coefficient of interest as it captures the extent to which our bias correction actually decreases the bias. If the bias reduction is large but unrelated to the magnitude of the bias, the estimated value of α will be large while η won't be significantly different from zero. By contrast, if the bias reduction is proportional to the actual bias, we expect an estimate of η that is significantly positive, taking values close to one if the bias completely disappears.

Table 7: Missing Persistence Bias: Cross-sectional Evidence

	CPI	Ss	Calvo	CPI	Ss	Calvo
	(Bias Correction)			(Bias reduction)		
	(1)	(2)	(3)	(4)	(5)	(6)
η	1.004	1.028	1.028			
	(0.028)	(0.027)	(0.008)			
Frequency				-1.091	-0.194	-1.014
				(0.133)	(0.156)	(0.165)
μ_A				-22.615	-10.666	-9.444
				(9.392)	(9.480)	(10.545)
N				-0.285	-0.093	-0.242
				(0.122)	(0.138)	(0.153)
Constant	-0.063	0.024	0.001	1.022	0.575	0.571
	(0.024)	(0.015)	(0.003)	(0.030)	(0.034)	(0.038)
Observations	66	66	66	66	66	66
R-squared	0.951	0.957	0.997	0.664	0.084	0.480

The first three columns estimate equation (22) in the CPI microdata, a calibrated Ss model and in a calibrated Calvo model respectively. The main coefficient of interest is η, which captures the extent to which our proposed estimator does a reducing the missing persistence bias. Columns 4-6 document how the magnitude of the bias across sectors, measured by the gap between the VAR implied frequency and the true frequency of adjustment, $\lambda_s^{\text{VAR}} - \lambda_s^{\text{micro}}$, varies with observables (the frequency of adjustment, the mean of the sectoral inflation process and the number of effective observations) which Proposition 1 suggests should be related to the magnitude of the bias.

The first column of Table 7 shows the results for the CPI. Since the estimated value of η is not statistically different from one and the constant term is close to zero, these results suggest that our

Table 8: SLOW CONVERGENCE IN THE DATA

$\hat{\rho}$ and Half-Life from Sub-Sampling the U.S. CPI Microdata

Persistence measure	Number of agents (N)					
	90	359	898	4490	8980	17960
$\hat{\rho}$	0.051	0.127	0.200	0.289	0.307	0.316
	(0.066)	(0.064)	(0.051)	(0.023)	(0.016)	(0.000)
Half-Life	0.233	0.336	0.431	0.558	0.587	0.602
	(0.102)	(0.081)	(0.069)	(0.036)	(0.025)	(0.000)

This table is generated by repeatedly randomly sampling N price change observations in each month (including zeros) from the microdata and and using this sub-sample to compute a time-series for inflation, $\hat{\pi}_t$. We then estimate an AR(1) on this inflation series as a measure of persistence and repeat this process 500 times. First row shows our AR measure of persistence, $\hat{\rho}$, using one lag (using two lags gave similar results). Bootstrapped standard errors are computing by sampling with replacement from the underlying data simulations 500 times. The third row shows the implied half-life, $-\log 2/\log\hat{\rho}$. Standard errors are computed using the delta method.

bias correction strategy comes very close to eliminating the bias entirely. This suggests that an alternative interpretation for BK's finding is that it provides evidence for the empirical importance of the missing persistence bias rather than as a rejection of the Calvo model. Columns 2 and 3 report the estimates for (21) in calibrated multi-sector Ss and Calvo models. These multi-sector models provide a useful laboratory to test in a controlled setting whether the missing persistence bias is relevant and whether our bias correction approach works.[38] Since a crucial element in these calibration is to work with the correct number of price setters in each sector, we set the number of effective price-setters in each sector equal to the number of effective price-setters in the relevant sector of the CPI microdata.[39]

We first establish that the missing persistence bias is present. In particular, we find that $\widehat{\beta}_s \ll 1-\lambda_s^{\text{micro}}$ for the vast majority of sectors in both models.[40] Columns 2 and 3 show that our bias correction procedure works well in both models. This is not surprising in the Calvo example since the assumptions in Section 3.1 are satisfied, however, the fact that it works for the Ss case suggests the procedure works in more general settings.[41]

[38] Our calibration is standard and so the details are relegated to Appendix G.2.

[39] In particular, we use item level expenditure weights w_i, $i = 1,2,..,n$, with $w_i > 0$ and $\sum_{i=1}^{n} w_i = 1$ within each sector. Then the effective number of units in each sector, N_s, is definied as the inverse of the Herfindahl index:

$$N_s \equiv \frac{1}{\sum_{i=1}^{n} w_i^2}.$$

[40] 64 of 66 sectors in the Calvo simulation; 57 of 66 in the Ss Simulation.

[41] Another implication is that Bils and Klenow's test of the Calvo model is not a useful way of discriminating between these two models since our bias correction procedure identifies the true frequency of adjustment in both cases.

Columns 4-6 of Table 7 provides further evidence that the missing persistence bias is at work by explicitly examining the comparative statics discussed in Section 3.3.2 (see equation (9)). In particular, we use cross-sector variation to explore how the magnitude of bias, $\lambda_s^{\text{VAR}} - \lambda_s^{\text{micro}}$, varies with underlying parameters that we can directly measure using sector level microdata: the frequency of adjustment, λ_s^{micro}, the effective number of observations, N_s, and the time-series mean of sectoral inflation, μ_A. We find evidence that the frequency of adjustment and the number of observations are both significantly negatively related to level of the bias. While the coefficient on the drift in inflation has the wrong sign, this coefficient is not significantly different from zero in two of the three cases. Overall, this example shows that the bias is relevant at the sectoral level and that through the use of microeconomic data one can implement our bias correction procedure (see Section 3.4) in practice.

Table 8 presents further evidence that the missing persistence bias is present in the CPI micro data. We generate this table, an empirical version of Table 1, by sub-sampling the microdata and then estimating the persistence of inflation using these sub-samples. In particular, we randomly sample N price change observations in each month (including zeros) and use this sub-sample to compute a time-series of inflation rates, $\hat{\pi}_t$.[42] We then estimate an AR(1) on this inflation series as a measure of persistence. We repeat this process 500 times and display the mean estimate and standard error.

Reading across Table 8 it is clear that estimated persistence increases sharply as we increase N, which indicates the presence of the missing persistence bias. Similar to both Tables 1 (Calvo) and 5 (Ss), the bias reduction is concave in N. This suggests that the missing persistence bias is most relevant at the 2-digit sectoral level in the CPI micro data where the underlying number of observations is relatively small.

5.2 Does inflation respond more quickly to sectoral shocks than aggregate shocks?

It is well-known from the theoretical literature on sticky-information and costly observation models that there is no reason why prices should adjust equally fast to different types of shocks as agents may optimally choose to focus on the shocks that matter more to them. Boivin, Giannoni and Mihov (2009) (henceforth BGM) use BLS microdata[43] and find that sectoral inflation responds much faster to sectoral shocks than to aggregate shocks and interpret this result as evidence in favor of these models. However, when lumpy adjustment is present, differential speed of adjustment to shocks at different levels of aggregation could also signal the presence of the missing persistence bias. We explore this possibility next and show that the difference in speed of adjustment disappears once we correct for the bias.

To understand BGM's approach, we must first introduce some terminology. Define Π_t as a column vector with monthly sectoral inflation rates in period t, for sectors 1 through S, where S de-

[42] We make sure the implied frequency of adjustment is similar across samples.

[43] The underlying inflation series in the PCE come from the CPI.

notes the number of sectors. BGM assume that Π_t can be decomposed into the sum of a small number R of common factors, C_t, and a sectoral component, e_t:

$$\Pi_t = \Lambda C_t + e_t, \tag{23}$$

where Λ denotes an SxR matrix of factor loadings that are allowed to differ across sectors, while C_t and e_t are Rx1 and Sx1 matrices. This formulation allows them to disentangle the fluctuations in sectoral inflation rates due to the macroeconomic factors—represented by the common components C_t with sector specific weights—from those due to sector-specific conditions represented by the term e_t.

BGM extract R principal components from the large data set Π_t to obtain consistent estimates of the common factors.[44] Next, they regress each sectoral inflation series on these common factors,[45] denoting the predicted aggregate component, $\lambda_i' C_t$, by π_{st}^{agg}, and the residual that captures the sector-specific component, e_{st}, by π_{st}^{sect}. This methodology decomposes each sectoral inflation series into aggregate and sectoral components that are orthogonal:

$$\pi_{st} = \lambda_s' C_t + e_{st} = \pi_{st}^{\text{agg}} + \pi_{st}^{\text{sect}}. \tag{24}$$

To calculate IRFs with respect to the common and sectoral shocks, BGM fit separate AR(13) processes to the π_{st}^{agg} and π_{st}^{sect} series and measure the persistence of shocks by the sum of the 13 AR coefficients. This is a standard method for estimating IRFs and is motivated by the observation that if there is a lot of persistence in the data then the sum of the AR coefficients should be close to one. For example, if the underlying microdata were generated by a Calvo model with $N = \infty$, then this sum is equal to one minus the frequency of adjustment. Decreases in the adjustment frequency increase actual persistence and this method of measuring IRFs reflects this accurately.

However, since this method of estimating persistence is the standard VAR methodology discussed in Section 3, if N is small and adjustment is lumpy then the missing persistence bias is a concern. The reason is that if adjustment is lumpy then we know from Section 3 that the AR coefficients will be biased downwards, leading econometricians to find less persistence (faster response to shocks) than there actually is. Since the underlying prices adjust infrequently in the CPI and there are fewer prices underlying the sectoral component, π_{st}^{sect}, relative to the aggregate component, π_{st}^{agg}, BGM's results could be driven by the missing persistence bias. We explore this possibility next.

We start by reproducing BGM's benchmark results using the CPI data. There are a few differences between our sample and BGM's.[46] The first two columns show results for BGM's baseline

[44] Stock and Watson (2002) show that the principal components consistently recover the space spanned by the factors when S is large and the number of principal components used is at least as large as the true number of factors.

[45] BGM allow C_t to follow an AR process. Therefore we allow C_t to have 6 lags in our baseline estimation. We have also tried different specifications where we allow for either 0 or 12 lags of C_t and found similar results.

[46] First, BGM use information on both prices and quantities whereas we just use information on prices. Second, BGM

Table 9: BGM (2009): ESTIMATED PERSISTENCE TO AGGREGATE AND SECTORAL SHOCKS

Sum of AR coefficients for AR(13)

	BGM Sample (Baseline)		BGM Sample (PCE + 88-05)		BLS Sample (CPI + 88-07)	
	π_{st}^{agg}	π_{st}^{sect}	π_{st}^{agg}	π_{st}^{sect}	π_{st}^{agg}	π_{st}^{sect}
Mean	0.92	−0.07	0.58	−0.02	0.45	−0.11
Median	0.94	−0.01	0.66	0.09	0.64	−0.04

sample taken directly from Table 1 in their paper. The third and fourth columns show results using BGM's methodology on the data sample that is closest to our setting: using only PCE inflation series to construct the aggregate factors (Equation 23) and the 1988-2005 time period. The last two columns show our results when we implemented BGM's methodology with CPI data.

Table 9 shows that despite differences in the data used, we find similar results to BGM when we replicate their methodology with CPI data.[47] In all cases there is clear evidence of significant persistence to aggregate shocks and negligible persistence to sectoral shocks. While the amount of persistence to aggregate shocks is smaller in the CPI relative to BGM's baseline, a comparison between the third and fifth columns shows that these differences disappear once we use similar underlying data and time periods.[48] Overall, then, BGM's methodology robustly delivers the result that inflation responds faster to sectoral than aggregate shocks. However, given that price adjustment is lumpy and sample sizes are small for the sectoral series, the missing persistence bias could also explain this result. We explore this possibility next.

Our approach is simple. Since the bias only manifests itself when a researcher regresses a lumpy variable on lags of itself, as the VAR methodology does, we must use a different method for measuring persistence to both shocks. In particular, we take a cue from the end of Section 4.1, which showed that if one had measures of both types of shocks, one could regress each sectoral inflation series, π_{st}, on lags of these shocks to recover an unbiased estimate the IRF to each type of shock, even for small N. We referred to this as the MA approach. In contrast, our results in Section 4.1 showed the VAR approach was significantly biased in small samples. In Appendix G.3, we provide simulation results showing that our procedure accurately recovers the true underlying amount of persistence, whereas the VAR methodology infers that inflation responds more slowly to aggregate shocks than sectoral shocks.

use a longer sample period (1976-2005) than we have (1988-2007). Finally, BGM use more data (BGM use 653 series, half of which are price series) whereas we use 66.

[47] We report results that assume there are 5 common factors.

[48] Reassuringly, Mackoviak, Moench and Wiederholt (2011) reach a similar to conclusion to BGM using the CPI data and a different methodology.

To implement the MA approach we need estimates of both aggregate, m_t, and sectoral shocks, x_{st}, for each sector s. To get each we use our sectoral reset price shock measures, v_{st}'s, from Section 5.1. These were computed from CPI microdata over the period 1988:03-2007:12. Define V_t as the Sx1 vector with the period t sectoral shock measures. Our proxy for aggregate shocks is the first R principal components of V, denoted by m_t^k, $k = 1,2,...,R$. The logic for this approach is that aggregate shocks are the common component of the v_{st}'s since by definition they affect each of these series.

We compute the pure sectoral shock as a residual. In particular, we decompose v_{st} into the sum of an aggregate and a sectoral component and we recover the sectoral shocks by regressing each sectoral reset price series on our estimated aggregate shocks. Since we are using retail data, we include lags of the aggregate shocks in order to allow for some delay in these shocks propagating up the supply chain. Denote the pure sectoral shock as x_{st}.[49] Concretely:

$$v_{st} = \sum_{k=1}^{R} \sum_{j=0}^{J} \gamma_{sj}^k m_{t-j}^k + x_{st}, \tag{25}$$

where the term with double sums on the r.h.s. is the component driven by aggregate shocks, while the residual x_{st} is the component driven by sectoral shocks.

Now that we have our R aggregate shocks, m_t^k, and a sectoral shock, x_{st}, for each of our 66 sectors, we can implement our MA approach to estimate IRFs. We do this by regressing each sectoral inflation series on distributed lags of the aggregate and sectoral shocks:

$$\pi_{st} = \sum_{k=1}^{R} \eta_s^k(L) m_t^k + \nu_s(L) x_{s,t}, \tag{26}$$

where $\eta_s^k(L) = \sum_{j\geq 0} \eta_{sj} L^j$ and $\nu_s(L) = \sum_{j\geq 0} \nu_{sj} L^j$ denote lag polynomials. In order to parsimoniously estimate these lag polynomials, we model each $\eta_s^k(L)$ and $\nu_s(L)$ as quotients of two second degree polynomials.[50] This allows us to flexibly approximate a variety of possible shapes for our IRFs while also maintaining parsimony. The results we obtain are robust to reasonable variations in the order of these polynomials.[51]

We use the expected response time as our measure of persistence because it is more robust than the half-life to noise in the estimation process since it more naturally accommodates IRFs which contain both negative and positive values. Appendix D provides more details and shows that in the AR(1) case discussed in Section 3.1, the expected response time is equal to $\frac{\hat{\rho}}{1-\hat{\rho}}$, so that more persistence implies a higher expected response time. We compute the expected response time for

[49] Our results are robust to ignoring these distributed lags of common components yet we believe it is more realistic to include them so they are including in our baseline.

[50] We do not have enough data to estimate an unrestricted version of this equation given that we only have 254 observations for each series and R*number of lags in each lag polynomial coefficients.

[51] These robustness results are available upon request. We implemented this estimation using the polyest command in Matlab. See http://jp.mathworks.com/help/ident/ref/polyest.html for details.

Table 10: THE RESPONSE OF SECTORAL INFLATION RATES TO AGGREGATE AND IDIOSYNCRATIC SHOCKS

Median of estimated expected response times to shocks

PCs	nlags	agg	sec
		(1)	(2)
2	0	3.63	3.03
		(0.84)	(0.56)
2	3	2.57	2.71
		(0.77)	(0.55)
2	6	3.05	1.77
		(0.86)	(0.51)
2	12	2.79	2.86
		(0.91)	(0.56)
4	0	2.72	2.56
		(0.44)	(0.53)
4	3	1.98	2.53
		(0.44)	(0.54)
4	6	2.12	1.99
		(0.34)	(0.50)
4	12	1.72	2.17
		(0.45)	(0.54)
6	0	1.87	2.51
		(0.38)	(0.50)
6	3	2.00	2.83
		(0.46)	(0.64)
6	6	1.97	2.56
		(0.33)	(0.55)
6	12	2.14	2.24
		(0.33)	(0.56)

each of the R aggregate shocks and summarize the R response times to aggregate shocks by their median. In particular:

$$\begin{aligned} \tau_s^{\text{sec}} &\equiv \sum_{j\geq 0} j\nu_{sj}^k / \sum_{j\geq 0} \nu_{sj}^k, \\ \tau_s^{\text{agg},k} &\equiv \sum_{j\geq 0} j\eta_{sj}^k / \sum_{j\geq 0} \eta_{sj}^k, \\ \tau_s^{\text{agg}} &\equiv \text{median}_k \tau_{s,k}. \end{aligned}$$

Crucially for our procedure, because we have a direct proxy for both shocks, our measures of persistence to these shocks are not susceptible to the missing persistence bias.

The results are shown in Table 10. The numbers we report are medians across sectors. The

interquartile ranges (divided by the square root of the number of sectors) are shown in parentheses. We consider 12 possible combinations for the number of principal components (PC) and number of lags (nlags) used on the r.h.s. of (25).

Columns (1) and (2) show that after correcting for the missing persistence bias using the procedure outlined above, the average response to aggregate and sectoral shocks are 2.38 and 2.48 months, respectively. That is, sectoral prices adjust faster, on average, to aggregate shocks than to sectoral shocks yet this difference is not significant. We conclude that once one corrects for the missing persistence bias, there is no longer evidence that firms respond differently to aggregate and sectoral shocks.

6 Conclusion

While many microeconomic actions are infrequent and lumpy, large idiosyncratic shocks map these discrete microeconomic series into smooth aggregated counterparts. The presumption then is that standard linear time series analyses can be applied to these smooth aggregated time series to gage their dynamic behavior. The main result of this paper is to qualify and challenge this presumption. We show that while it holds with an infinite number of agents, convergence is extremely slow, precisely because idiosyncratic shocks are usually large. Moreover, we show that away from this limit the bias is systematic, leading to faster estimated responses of aggregate time series to aggregate shocks than is actually the case. We also find that the magnitude of the bias is relevant for sectoral series and may be present in some aggregate series as well.

On the constructive side, we discuss various procedures to correct for the bias. All of them have in common that they include estimates for the shocks among regressors while being careful about which lags of the response variable to include (or avoiding them altogether). We also demonstrate the usefulness of correction procedures with two applications. In the first one we show that the bias provides an alternative explanation for the persistence-gap reported in Bils and Klenow's (2004). In the second one we show that the difference in the speed with which inflation responds to sectoral and aggregate shocks (Boivin et al 2009; Mackoviak et al 2009) disappears once we correct for the missing persistence bias.

References

[1] Alvarez, Fernando, Le Bihan, Herve, and Francesco Lippi, “The real effects of monetary shocks in sticky price models: a sufficient statistic approach," *American Economic Review*, **106** (10), October 2016, 2817–51.

[2] Ash, Robert B. and Melvin F. Gardner, *Topics in Stochastic Processes*, New York: Academic Press, 1975.

[3] Bils, Mark, “Do Higher Prices for New Goods Reflect Quality Growth or Inflation,” *The Quarterly Journal of Economics*, **124**(2), May 2009, 637–675.

[4] Bils, Mark and Peter J. Klenow, “Some Evidence on the Importance of Sticky Prices,” *J. of Political Economy*, **112**, 2004, 947–985.

[5] Bils, Mark, Peter J. Klenow, and Ben Malin, “Reset Price Inflation and the Impact of Monetary Policy Shocks”, *American Economic Review*, **102** (2), October 2012, 2798–2825.

[6] Boivin, Jean, Marc P. Giannoni, and Illian Mihov, “Sticky Prices and Monetary Policy: Evidence from Disaggregated US Data", *American Economic Review*, **102** (2), March 2009, 350–384.

[7] Caballero, Ricardo J., Eduardo M.R.A. Engel, “Price stickiness in *Ss* models: New Interpretations of old results”, *Journal of Monetary Economics*, **12**, 2007, 100–121.

[8] Caballero, Ricardo J., Eduardo M.R.A. Engel, and John C. Haltiwanger, “Plant-Level Adjustment and Aggregate Investment Dynamics”, *Brookings Papers on Economic Activity*, 1995 (2), 1–39.

[9] Caballero, Ricardo J., Eduardo M.R.A. Engel, and John C. Haltiwanger, “Aggregate Employment Dynamics: Building from Microeconomic Evidence”, *American Economic Review*, **87** (1), March 1997, 115–137.

[10] Calvo, Guillermo, “Staggered Prices in a Utility-Maximizing Framework,” *Journal of Monetary Economics*, **12**, 1983, 383–398.

[11] Carlsson, Mikael, and Oskar Nordstrom Skans. “Evaluating Microfoundations for Aggregate Price Rigidities: Evidence from Matched Firm-Level Data on Product Prices and Unit Labor Cost." *American Economic Review*, 102(4), June 2012: 1571-95.

[12] Christiano, Eichenbaum and Evans. “Monetary Policy Shocks: What Have We Learned and to What End?." *Handbook of Macroeconomics*, 1999.

[13] Christiano, Eichenbaum and Evans. “Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy" *J. of Political Economy*, **113**, 2005, 1-45.

[14] Clarida, Richard, Jordi Gali, and Mark Gertler, "Monetary Policy Rules and Macroeconomic Stability: Evidence and Some Theory," *Quarterly Journal of Economics 115*, 2000, 147-180.

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

[16] Dynan, Karen E., "Habit Formation in Consumer Preferences: Evidence from Panel Data," *American Economic Review*, **90**(3), June 2000, 391–406.

[17] Eichenbaum, Martin, Nir Jaimovich, and Sergio Rebelo, "Reference Prices, Costs and Nominal Rigidities", *American Economic Review*, **101** (1), February 2011, 234–262.

[18] Engel, Eduardo M.R.A., "A Unified Approach to the Study of Sums, Products, Time-Aggregation and other Functions of ARMA Processes", *Journal Time Series Analysis*, **5**, 1984, 159–171.

[19] Gertler, Mark and John Leahy, "A Phillips Curve with an Ss Foundation," *J. of Political Economy*, **116**, 2008, 533–572.

[20] Goodfriend, Marvin, "Interest-rate Smoothing and Price Level Trend Stationarity," *Journal of Monetary Economics*, **19**, 1987, 335–348.

[21] Hamilton, James *Time Series Analysis*, Princton University Press, 1994.

[22] Holt, Charles, Franco Modigliani, John Muth and Herbert A. Simon, *Planning Production Inventories, and Work Force*, Prentice-Hall, 1960.

[23] Jorda, Oscar, "Random Time Aggregation in Partial Adjustment Models," *Journal of Business and Economic Statistics*, **7**(3), July 1999, 382–396.

[24] Kehoe, Patrick J., and Vrigiliu Midrigan, "Prices are Sticky After All", *Federal Reserve Bank of Minneapolis Research Department Staff Report*, **413**, June 2012.

[25] Klenow, Peter J., and Oleksiy Kryvtsov, "State-Dependent or Time-Dependent Price: Does it Matter for Recent U.S. Inflation", *Quarterly Journal of Economics 123*, 2008, 863–904.

[26] Le Bihan, Herve, and Julien Matheron, "Price Stickiness and sectoral Inflation Persistence: Additional Evidence," *Journal of Money, Credit and Banking*, **44** (7), October 2012, 1427–1422 .

[27] Mackowiak, Bartosz, Emanuel Moench and Mirko Wiederholt, "Sectoral price data and models of price setting," *J. of Monetary Economics*, **56** October 2009, S78–S99.

[28] Mackowiak, Bartosz and Mirko Wiederholt, "Optimal Sticky Prices under Rational Inattention," *American Economic Review*, **99** (3), June 2009, 769–803.

[29] Nakamura, Emi, and Jon Steinsson, "Five facts about prices: A reevaluation of menu cost models", *Quarterly Journal of Economics 123*, 2008, 147–180.

[30] Majd, Saman and Robert S. Pindyck, "Time to Build, Option Value, and Investment Decisions," *Journal of Financial Economics*, **18**, March 1987, 7–27.

[31] Midrigan, Virgiliu, "Menu Costs, Multiproduct Firms, and Aggregate Fluctuations," *Econometrica*, Vol. 79(4), 2011, 1139–1180.

[32] Ramey, Valerie. "Macroeconomic Shocks and Their Propagation ." *Handbook of Macroeconomics*, 2016.

[33] Rotemberg, Julio J., "The New Keynesian Microfoundations," in O. Blanchard and S. Fischer (eds), NBER Macroeconomics Annual, 1987, 69–104.

[34] Sack, Brian, "Uncertainty, Learning, and Gradual Monetary Policy," Federal Reserve Board Finance and Economics Discussion Series Paper 34, August 1998.

[35] Sargent, Thomas J., "Estimation of Dynamic Labor Demand Schedules under Rational Expectations," *Journal of Political Economy*, **86**, 1978, 1009–1044.

[36] Sims, Christopher, "Output and Labour Input inManufacturing," *Brookings Papers on Economic Activity*, 1974, No. 3, 695–735.

[37] Stock, Jim, H., and Mark W. Watson, "Has the Business Cycle Changed and Why?", *NBER Macroeconomics Annual*, 2002, Vol. 17, 159–218.

[38] Tinsley, Peter A., "A Variable Adjustment Model of Labour Demand," *International Economic Review*, 1971, Vol. 12(3), 482–510.

[39] Woodford, Michael, "Optimal Monetary Policy Inertia," NBER WP # 7261, July 1999.

[40] Woodford, Michal, *Interest and prices: Foundations of a theory of monetary policy*, New York: Cambridge University Press, 2005.

APPENDIX

A Additional Bias Correction Methods

In the main text we studied an approach to correct for missing persistence bias using a proxy for y^*, which is the approach we used in Section 5. Here we provide two additional approaches.

A.1 ARMA Correction

The second correction we propose is based on a simple ARMA representation for Δy_t^N.

Proposition 3 (ARMA Representation)

Consider the assumptions and notation of Proposition 1. We then have that Δy_t^N follows the following ARMA(1,1) process:

$$\Delta y_t^N = \rho \Delta y_{t-1}^N + (1-\rho)[\varepsilon_t - \theta \varepsilon_{t-1}], \tag{27}$$

where ε_t is an i.i.d. innovation process and $\theta = (S - \sqrt{S^2-4})/2 > 0$ with $S = [2 + (1-\rho^2)(K-1)]/\rho$.[52]

Proof See Appendix C. ∎

Using (27) to write Δy_t^N as an infinite moving average shows that its impulse response to ε-shocks satisfies:

$$I_k = \begin{cases} 1-\rho & \text{if } k = 0 \\ \\ (1-\rho)(\rho-\theta)\rho^{k-1} & \text{if } k \geq 1. \end{cases}$$

Yet this is not the impulse response to the aggregate shock v_t^A, because ε_t in (27) is not v_t^A. As in section 3.3.1, the innovation of the Wold representation is not the innovation of economic interest. The derivation of the impulse response from section 3.3.1 for the case where $N = 1$ carries over to the case with $N > 1$ and the true impulse response is equal to $(1-\rho)\rho^k$, that is, it corresponds to the case where $\theta = 0$ in (27).

This suggests a straightforward approach to estimating the adjustment speed parameter, ρ: Estimate an ARMA(1,1) process (27) and read off the estimate of ρ (and the true impulse response) from the estimated AR-coefficient. That is, first estimate an ARMA model, next drop the MA polynomial and then make inferences about the implied dynamics using only the AR polynomial.

This approach runs into two difficulties when applied in practice. First, for small values of N we have that Δy_t^N is close to an i.i.d. process which means that θ and ρ will be similar. It is well known that estimating an ARMA process with similar roots in the AR and MA polynomials leads to imprecise estimates, resulting in an imprecise estimate for the parameter of interest, ρ.

Second, to apply this approach in a more general setting like the one described by equation (1) in Section 2, the researcher will need to estimate a time-series model with a complex web of AR and MA polynomials and then "drop" the MA polynomial before making inference about the implied dynamics. This strategy is likely to be sensitive to the model specification, for example, the number of lags in the AR-polynomial $b(L)$ in the case of (1).

[52] Scaling the right hand side term by $(1-\rho)$ is inoccuous but useful in what follows.

A.2 Instrumental Variables

Equation (27) in Proposition 1 suggests that lagged values of Δy and Δy^* (or components thereof) may be valid instruments to estimate ρ in a regression of the form

$$\Delta y_t^N = \text{const.} + \rho \Delta y_{t-1}^N + e_t.$$

More precisely, if $v_t = \Delta y_t^{*N}$, then Δy_{t-k} and Δy_{t-k}^{*N} will be valid instruments for $k \geq 2$. Yet things are a bit more complicated, since $v_t = \Delta y_t^{*N}$ holds only for $N = \infty$. As shown in the following proposition, the set of valid instruments is larger than suggested above and also includes Δy_{t-1}^{*N}.

Proposition 4 (Instrumental Variables)

With the same notation and assumptions as in Proposition 1, we will have that Δy_{t-k}^N, $k \geq 2$ *and* Δy_{t-j}^{*N}, $j \geq 1$ *are valid instruments when estimating* ρ *from*

$$\Delta y_t^N = \text{const.} + \rho \Delta y_{t-1}^N + e_t.$$

By contrast, Δy_{t-1}^N *is not a valid instrument.*

Proof See Appendix C. ∎

B Extensions

B.1 Relaxing the i.i.d. Assumption

In Section 3 we assumed that Δy^* is i.i.d. Even though this assumption is a good approximation in many settings (nominal output follows a random walk in Woodford [2003, sect. 3.2], nominal marginal costs follow a random walk in Bils and Klenow [2004]) it is worth exploring what happens when we relax this assumption. When doing so, the cross correlations between contiguous adjustments are no longer zero, but the missing persistence bias typically remains.

We consider first the case where both components of Δy^*, v_t^A and v_{it}^I, follow AR(1) processes with the same first-order autocorrelation ϕ. The case we considered in the main text corresponds to $\phi = 0$. We show in Appendix E that, with a continuum of agents, Δy_t^∞ follows the following stationary ARMA(2,1) process:

$$\Delta y_t^\infty = (\rho + \phi)\Delta y_{t-1}^\infty - \rho\phi \Delta y_{t-2}^\infty + \varepsilon_t - \beta\rho\phi\varepsilon_{t-1},$$

with ε_t proportional to v_t^A and β denoting the agent's discount factor.[53]

Table 11 shows the measures of speed of convergence considered in Table 1, for the case of prices, once the i.i.d. assumption is relaxed. The first half of the table reports the estimated half-life of a shock, the second half the expected response time. The reported estimates assume that the researcher not only is aware that Δy^* is not i.i.d. but also knows the exact value of the first order autocorrelation, ϕ, as well as β, and estimates ρ via maximum likelihood from

$$(\Delta y_t^N - \phi \Delta y_{t-1}^N) = \text{const.} + \rho(\Delta y_{t-1}^N - \phi \Delta y_{t-2}^N) + e_t - \beta\phi\rho e_{t-1}.$$

[53] With the notation of Section 2 we have $b(L) = (1 - \phi L)/(1 - \beta\rho\phi L)$.

Table 11: SLOW CONVERGENCE

Estimated Half-Life and Expected Response Time Δy^* follows an AR(1)

	Effective number of agents (N)						
ϕ	100	400	1,000	4,000	10,000	40,000	True
0	0.252	0.466	0.769	1.724	2.639	3.794	4.596
0.1	0.246	0.440	0.723	1.683	2.659	3.841	4.615
0.2	0.296	0.426	0.686	1.671	2.646	3.852	4.644
0.3	0.379	0.459	0.661	1.615	2.651	3.882	4.690
0.4	0.529	0.564	0.662	1.589	2.697	3.993	4.764
0.5	0.751	0.767	0.801	1.416	2.704	4.064	4.887
0	0.068	0.292	0.684	2.021	3.329	4.988	6.143
0.1	0.069	0.247	0.587	1.932	3.339	5.045	6.160
0.2	0.139	0.246	0.522	1.874	3.290	5.039	6.186
0.3	0.277	0.332	0.509	1.745	3.251	5.050	6.225
0.4	0.514	0.533	0.596	1.661	3.255	5.158	6.288
0.5	0.865	0.870	0.885	1.424	3.183	5.177	6.393

First six rows report the average estimate of the half-life of a shock. The parameter ρ is estimated via maximum likelihood from $(\Delta y_t^N - \phi \Delta y_{t-1}^N) = \text{const.} + \rho(\Delta y_{t-1}^N - \phi \Delta y_{t-2}^N) + e_t - \beta\phi\rho e_{t-1}$ with β and ϕ known. The estimated half-life is obtained by finding k that solves $\sum_{j=0}^{k} d_k = \frac{1}{2}\sum_{j=0}^{\infty} d_k$, where $\Delta y_t^N = \sum_{k\geq 0} \psi_k v_{t-k}$ is the (infinite) MA representation of Δy_t^N assumed by the researcher. Estimates based on 100 simulations of length 1,000 each. Rows 7-12 are analogous to rows 1-6 with expected response time instead of estimated half-life. The expected response time is calculated from $(\phi+\rho-2\phi\rho)/(1-\phi-\rho+\rho\phi) - \beta\rho\phi/(1-\beta\rho\phi)$ (see Appendix D). Parameters (monthly pricing data): $\rho = 0.86$, $\mu_A = 0.003$, $\sigma_A = 0.0054$, $\sigma_I = 0.048$, $\beta = 0.96^{1/12}$.

The only source of bias is that the researcher ignores the fact that because the actual aggregate considers a finite number of agents, using the linear specification valid for an infinite number of agents will bias the estimated speed of adjustment upwards.[54]

It follows from Table 11 that the bias is generally larger when the Δy^* are correlated than in the i.i.d. case, even though the increase in the bias is small. For example, for $N = 10,000$, the estimated half-life is biased downward by 44.7% when $\phi = 0.5$ as compared with 42.6% when $\phi = 0$. Similarly, the bias for the corresponding expected response times are 45.8 and 50.2%, respectively.

In Section 3 we assumed that y^* is not stationary, we consider next the stationary case. Here we consider a stationary case by assuming that both the aggregate and idiosyncratic components of y^*_{it} follow stationary AR(1) processes with the same first-order autocorrelation ϕ, in previous sections we assumed $\phi = 1$. The innovations for these processes are the v^A_t and v^I_{it}, respectively. The remaining assumptions are unchanged.

It follows from Appendix E that, with a continuum of agents, y^∞_t follows the following stationary AR(2) process:

$$y^\infty_t = (\rho + \phi) y^\infty_{t-1} - \rho\phi y^\infty_{t-2} + \varepsilon_t,$$

with ε_t proportional to v^A_t.

Table 12 revisits Table 1, for annual investment data, this time assuming y^* follows an AR(1) process instead of a random walk. We consider investment, instead of prices as we did in Table 11, because the stationarity assumption for y^* is more reasonable in the case of investment.[55]

Table 12: SLOW CONVERGENCE

Estimated Fraction of Adjusters, $1-\rho$, when y^* follows an AR(1)

	Effective number of agents (N)						
ϕ	100	400	1,000	4,000	10,000	40,000	True
0.6	0.493	0.374	0.287	0.198	0.172	0.158	0.150
0.7	0.599	0.448	0.328	0.210	0.177	0.158	0.150
0.8	0.712	0.533	0.385	0.231	0.186	0.161	0.150
0.9	0.843	0.646	0.469	0.269	0.205	0.169	0.150
1.0	0.982	0.856	0.697	0.410	0.279	0.188	0.150

Parameter ρ estimated based on (28), 100 simulations with series of length 1,000. Parameters (annual investment data): $\rho = 0.85$, $\mu_A = 0.12$, $\sigma_A = 0.056$, $\sigma_I = 0.5$, $\beta = 0.96$.

Table 12 reports the estimated fraction of firms, not the estimated half-life or the expected response time. The reason for reporting a persistence measure different from those reported earlier is that when y is stationary the half-life and expected response time for Δy become infinite.[56]

[54] Simulations show that the bias disappears if we estimate $(\Delta y^N_t - \phi\Delta y^N_{t-1}) = \text{const.} + \rho(\Delta y^N_{t-1} - \phi\Delta y^N_{t-2}) + e_t - \gamma_1 e_{t-1} - \gamma_2 e_{t-2}$ with no constraints on γ_1 and γ_2. This suggests that the random walk assumption can be relaxed in Proposition 3. We thank Juan Daniel Díaz for this insight.

[55] Nonetheless, results are qualitatively similar if we work with prices.

[56] Also, if we report the half-life and expected response time for y instead of Δy, these persistence measures will be finite but cannot be meaningfully compared with the measures in Table 1 because the latter do not converge to the former when ϕ tends to one.

Reported estimates assume the researcher knows the value of ϕ in the AR(1) process but believes $N = \infty$, and therefore estimates ρ via OLS from

$$y_t^N - \phi y_{t-1}^N = \rho(y_{t-1}^N - \phi y_{t-2}^N) + e_t. \tag{28}$$

Table 12 shows that the bias is still present when $\phi < 1$ but decreases as ϕ becomes smaller. We show in Appendix F that there is no bias when $\phi = 0$. Because the parameters in Table 12 correspond to annual investment data, the first order autocorrelation parameter ϕ is likely to be around 0.8, suggesting the bias will be large. For example, for $N = 1{,}000$ (which corresponds roughly to the effective number of firms for the U.S. non-farm business sector) and $\phi = 0.8$, the researcher concludes, on average, that 38.5% of firms adjust in any given year, when the true value is 15%.

B.2 Adding smooth adjustment

Suppose now that in addition to the infrequent adjustment pattern described above, once adjustment takes place, it is only gradual. Such behavior is observed, for example, when there is a time-to-build feature in investment (e.g., Majd and Pindyck (1987)) or when policy is designed to exhibit inertia (e.g., Goodfriend (1987), Sack (1998), or Woodford (1999)). Our main result here is that the econometrician estimating a linear ARMA process —a Calvo model with additional serial correlation— will only be able to extract the gradual adjustment component but not the source of sluggishness from the infrequent adjustment component. That is, again, the estimated speed of adjustment will be too fast, for exactly the same reason as in the simpler model.

Let us modify our basic model so that equation (2) now applies for a new variable $\tilde{y}_t$ in place of y_t, with $\Delta\tilde{y}_t$ representing the *desired* adjustment of the variable that concerns us, Δy_t. This adjustment takes place only gradually, for example, because of a time-to-build component. We capture this pattern with the process:

$$\Delta y_t = \sum_{k=1}^{K} \phi_k \Delta y_{t-k} + (1 - \sum_{k=1}^{K} \phi_k)\Delta\tilde{y}_t. \tag{29}$$

Now there are two sources of sluggishness in the transmission of shocks, Δy_t^*, to the observed variable, Δy_t. First, the agent only acts intermittently, accumulating shocks in periods with no adjustment. Second, when the agent adjusts, it does so only gradually.

By analogy with the simpler model, suppose the econometrician approximates the lumpy component of the more general model by:

$$\Delta\tilde{y}_t = \rho\Delta\tilde{y}_{t-1} + \upsilon_t. \tag{30}$$

Replacing (30) into (29), yields the following linear equation in terms of the observable, Δy_t:

$$\Delta y_t = \sum_{k=1}^{K+1} a_k \Delta y_{t-k} + \varepsilon_t, \tag{31}$$

with

$$\begin{array}{lcll} a_1 & = & \phi_1 + \rho, & \\ a_k & = & \phi_k - \rho\phi_{k-1}, & k = 2, ..., K, \\ a_{K+1} & = & -\rho\phi_K, & \end{array} \tag{32}$$

and $\varepsilon_t \equiv (1-\rho)(1-\sum_{k=1}^{K}\phi_k)\Delta y_t^*$.

By analogy to the simpler model, we now show that the econometrician will miss the source of persistence stemming from ρ.

Proposition 5 (Omitted Source of Sluggishness)

Let all the assumptions in Proposition 1 hold, with $\tilde{y}$ in the role of y. Also assume that (29) applies, with all roots of the polynomial $1-\sum_{k=1}^{K}\phi_k z^k$ outside the unit circle. Let $\hat{a}_k, k=1,...,K+1$ denote the OLS estimates of equation (31).

Then:

$$\begin{array}{lcll} \text{plim}_{T\to\infty}\hat{a}_k & = & \phi_k, & k=1,...,K, \\ \text{plim}_{T\to\infty}\hat{a}_{K+1} & = & 0. & \end{array} \tag{33}$$

Proof See Appendix C. ∎

Comparing (32) and (33) we see that the proposition simply reflects the fact that the (implicit) estimate of ρ is zero.

C Proof of Propositions

Proof of Proposition 1

In this appendix we prove Proposition 1. The proof uses an auxiliary variable, x_{it}, equal to how much unit i adjusts in period t if it adjusts that period (that is, the value of Δy_{it} conditional on adjustment). Because of the Technical Assumptions, x_{it} equals the unit's accumulated shocks since it last adjusted. The following dynamic dynamic definition of x_{it} is what we use in the proof:

$$x_{i,t+1} = (1-\xi_{it})x_{it} + \Delta y_{i,t+1}^*, \tag{34}$$

$$\Delta y_{it} = \xi_{it} x_{it}. \tag{35}$$

In what follows, subindices i and j denote *different* units.

We first derive the following unconditional expectations:

$$\mathrm{E}x_{it} = \frac{\mu_A}{1-\rho}, \tag{36}$$

$$\mathrm{E}[\Delta y_{it}] = \mu_A, \tag{37}$$

$$\mathrm{E}[\Delta y_t^N] = \mu_A, \tag{38}$$

$$\mathrm{E}[x_{it}x_{jt}] = \frac{1}{1-\rho^2}\left[\sigma_A^2 + \frac{1+\rho}{1-\rho}\mu_A^2\right], \tag{39}$$

$$\mathrm{E}[x_{it}^2] = \frac{1}{1-\rho}\left[\sigma_A^2 + \sigma_I^2 + \frac{1+\rho}{1-\rho}\mu_A^2\right]. \tag{40}$$

From (34) and the Technical Assumption in the main text we have:

$$\mathrm{E}x_{i,t+1} = \rho \mathrm{E}x_{it} + \mu_A.$$

The above expression leads to (36) once we note that the stationarity of x_{it} implies $\mathrm{E}x_{i,t+1} = \mathrm{E}x_{it}$.

Equation (37) follows from (36) and Technical Assumption 3. Equation (38) follows directly from (37).

To derive (39), we note that, from (34)

$$\begin{aligned}
\mathrm{E}[x_{i,t+1}x_{j,t+1}] &= \mathrm{E}[\{(1-\xi_{it})x_{it}+\Delta y^*_{i,t+1}\}\{(1-\xi_{jt})x_{jt}+\Delta y^*_{j,t+1}\}] \\
&= \mathrm{E}[(1-\xi_{it})x_{it}(1-\xi_{jt})x_{jt}] + \mathrm{E}[\Delta y^*_{i,t+1}(1-\xi_{jt})x_{jt}] \\
&\qquad + \mathrm{E}[(1-\xi_{it})x_{it}\Delta y^*_{j,t+1}] + \mathrm{E}[\Delta y^*_{i,t+1}\Delta y^*_{j,t+1}] \\
&= \rho^2\mathrm{E}[x_{it}x_{jt}] + 2\frac{\rho}{1-\rho}\mu_A^2 + (\mu_A^2+\sigma_A^2),
\end{aligned}$$

where we used the Technical Assumptions, (36) and $i \neq j$. Noting that $x_{it}x_{jt}$ is stationary and therefore $\mathrm{E}[x_{it}x_{jt}] = \mathrm{E}[x_{i,t-1}x_{j,t-1}]$, the above expression leads to (39).

Finally, to prove (40), we note that, from (34) we have

$$\begin{aligned}
\mathrm{E}[x^2_{i,t+1}] &= \mathrm{E}[(1-\xi_{it})x^2_{it}] + 2\mathrm{E}(1-\xi_{it})x_{it}\Delta y^*_{i,t+1}] + \mathrm{E}[(\Delta y^*_{i,t+1})^2] \\
&= \rho\mathrm{E}[x^2_{it}] + 2\frac{\rho}{1-\rho}\mu_A^2 + (\sigma_A^2+\sigma_I^2+\mu_A^2),
\end{aligned}$$

where we used that $(1-\xi_{it})^2 = 1-\xi_{it}$, (36) and the Technical Assumptions. Stationarity of x_{it} (and therefore x^2_{it}) and some simple algebra complete the proof.

Next we use the five unconditional expectations derived above to obtain the four expressions in the second row of Table 3. The expression for the OLS estimate $\hat{\rho}$ in (8) then follows from tedious but otherwise straightforward algebra.

We have:

$$\begin{aligned}
&\mathrm{Cov}(\Delta y_{i,t+1},\Delta y_{it}) = \mathrm{E}[\Delta y_{i,t+1}\Delta y_{it}] - \mu_A^2 = \mathrm{E}[\xi_{i,t+1}x_{i,t+1}\xi_{it}x_{it}] - \mu_A^2 = (1-\rho)\mathrm{E}[x_{i,t+1}\xi_{it}x_{it}] - \mu_A^2 \\
&\quad = (1-\rho)\mathrm{E}[\{(1-\xi_{it})x_{it}+\Delta y^*_{i,t+1}\}\xi_{it}x_{it}] - \mu_A^2 = (1-\rho)\mathrm{E}[\{(1-\xi_{it})\xi_{it}x^2_{it}] + (1-\rho)\mathrm{E}[\Delta y^*_{i,t+1}\xi_{it}x_{it}] - \mu_A^2 \\
&= (1-\rho)\times 0 + (1-\rho)\mu_A^2 - \mu_A^2 = -\rho\mu_A^2,
\end{aligned}$$

where in the crucial step we used that $(1-\xi_{it})\xi_{it}$ always equals zero.

We also have the cross-covariance terms ($i \neq j$):

$$\begin{aligned}
\mathrm{Cov}(\Delta y_{i,t+1},\Delta y_{jt}) &= \mathrm{E}[\xi_{i,t+1}x_{i,t+1}\xi_{jt}x_{jt}] - \mu_A^2 = (1-\rho)\mathrm{E}[x_{i,t+1}\xi_{jt}x_{jt}] - \mu_A^2 \\
&= (1-\rho)\mathrm{E}[\{(1-\xi_{it})x_{it}+\Delta y^*_{i,t+1}\}\xi_{jt}x_{jt}] - \mu_A^2 = \rho(1-\rho)^2\mathrm{E}[x_{it}x_{jt}] + (1-\rho)\mu_A^2 - \mu_A^2 = \frac{1-\rho}{1+\rho}\rho\sigma_A^2. \\
\mathrm{Cov}(\Delta y_{it},\Delta y_{jt}) &= \mathrm{E}[\xi_{it}x_{it}\xi_{jt}x_{jt}] - \mu_A^2 = (1-\rho)^2\mathrm{E}[x_{it}x_{jt}] - \mu_A^2 = \frac{1-\rho}{1+\rho}\sigma_A^2.
\end{aligned}$$

Finally, the variance term is obtained as follows:

$$\mathrm{Var}(\Delta y_{it}) = \mathrm{E}[\xi^2_{it}x^2_{it}] - \mu_A^2 = \mathrm{E}[\xi_{it}x^2_{it}] - \mu_A^2 = (1-\rho)\mathrm{E}[x^2_{it}] - \mu_A^2 = \sigma_A^2+\sigma_I^2+\frac{2\rho}{1-\rho}\mu_A^2. \quad \blacksquare$$

Proof of Proposition 2

Part (i) follows trivially from Proposition 1 and the fact that both regressors are uncorrelated. To prove (ii) we first note that:

$$\text{plim}_{T\to\infty}\hat{b}_1 = \frac{\text{Cov}(\Delta y_t - \Delta y_{t-1}\,,\, \Delta y_t^* - \Delta y_{t-1})}{\text{Var}(\Delta y_t^* - \Delta y_{t-1})}.$$

We therefore need expressions for $\text{Cov}(\Delta y_t^N, \Delta y_t^{N*})$, $\text{Cov}(\Delta y_t^N, \Delta y_{t-1}^N)$ and $\text{Var}(\Delta y_t^N)$.
We have

$$\text{Cov}(\Delta y_t^N, \Delta y_t^{N*}) = \frac{1}{N}\text{Cov}(\Delta y_{it}, \Delta y_{it}^*) + \left(1 - \frac{1}{N}\right)\text{Cov}(\Delta y_{it}, \Delta y_{jt}).$$

Both covariances on the r.h.s. are calculated using (34), yielding $\sigma_A^2 + \sigma_I^2$ and σ_A^2, respectively. Expressions for $\text{Cov}(\Delta y_t^N, \Delta y_{t-1}^N)$ and $\text{Var}(\Delta y_t^N)$ are obtained using an analogous decomposition and the covariances and variances from Table 3. We have all the terms for the expression above for $\hat{b}_1$, the remainder of the proof is some tedious but otherwise straightforward algebra. ∎

Proof of Proposition 3

To prove that Δy_t^N follows an ARMA(1,1) process with autoregressive coefficient ρ, it suffices to show that the process' autocorrelation function, γ_k, satisfies:[57]

$$\gamma_k = \rho\gamma_{k-1}, \qquad k \geq 2. \tag{41}$$

We prove this next and derive the moving average parameter θ by finding the unique θ within the unit circle that equates the first-order autocorrelation of this process, which by Proposition 1 is given by (8), with the following well known expression for the first order autocorrelation of an ARMA(1,1) process:

$$\gamma_1 = \frac{(1-\phi\theta)(\phi-\theta)}{1+\theta^2-2\phi\theta}.$$

Proving that θ tends to zero as N tends to infinity is straightforward.

We have:

$$\begin{aligned}
\text{E}[\Delta y_{t+k}^N \Delta y_t^N] &= \sum_{i=1}^n\sum_{j=1}^n w_i w_j \text{E}[\xi_{i,t+k} x_{i,t+k} \xi_{jt} x_{jt}] \\
&= (1-\rho)\sum_{i=1}^n\sum_{j=1}^n w_i w_j \text{E}[x_{i,t+k}\xi_{jt}x_{jt}] \\
&= (1-\rho)\sum_{i=1}^n\sum_{j=1}^n w_i w_j \text{E}[\{(1-\xi_{i,t+k-1})x_{i,t+k-1} + \Delta y_{i,t+k}^*\}\xi_{jt}x_{jt}] \\
&= (1-\rho)\rho\sum_{i=1}^n\sum_{j=1}^n w_i w_j \text{E}[x_{i,t+k-1}\xi_{jt}x_{jt}] \;+\; (1-\rho)\mu_A\sum_{i=1}^n\sum_{j=1}^n w_i w_j \text{E}[\xi_{jt}x_{jt}] \\
&= \rho\sum_{i=1}^n\sum_{j=1}^n w_i w_j \text{E}[\xi_{i,t+k-1}x_{i,t+k-1}\xi_{jt}x_{jt}] \;+\; (1-\rho)\mu_A^2 \\
&= \rho\text{E}[\Delta y_{t+k-1}^N \Delta y_t^N] + (1-\rho)\mu_A^2,
\end{aligned}$$

[57] Here we are using Theorem 1 in Engel (1984) characterizing ARMA processes in terms of difference equations satisfied by their autocorrelation function.

where in the fourth step we assumed $k \geq 2$, since we used that $\xi_{i,t+k-1}$ and ξ_{jt} are independent even when $i = j$. Noting that $\gamma_k = (\mathrm{E}[\Delta y^N_{t+k} \Delta y^N_t] - \mu_A^2)/\mathrm{Var}(\Delta y_t)$ and using the above identity yields (41) and concludes the proof. ∎

Proof of Proposition 4

We have:

$$\Delta y^N_t = \sum_i w_i \xi_{it} x_{it} = \sum_i w_i \xi_{it} (y^*_{it} - y_{i,t-1}) = \sum_i w_i (1-\rho)(y^*_{it} - y_{i,t-1}) + \sum_i w_i (\xi_{it} - 1 + \rho)(y^*_{it} - y_{i,t-1}).$$

Similarly

$$\Delta y^N_{t-1} = \sum_i w_i (1-\rho)(y^*_{i,t-1} - y_{i,t-2}) + \sum_i w_i (\xi_{i,t-1} - 1 + \rho)(y^*_{i,t-1} - y_{i,t-2}).$$

Subtracting the latter from the former and rearranging terms yields

$$\Delta y^N_t = \rho \Delta y^N_{t-1} + (1-\rho)\Delta y^{*N}_t + \epsilon^N_t \tag{42}$$

with

$$\epsilon^N_t = \sum_i w_i \left[(\xi_{it} - 1 + \rho)(y^*_{it} - y_{i,t-1}) - (\xi_{i,t-1} - 1 + \rho)(y^*_{i,t-1} - y_{i,t-2}) \right]. \tag{43}$$

The extra term ϵ^N_t on the r.h.s. of (43) explains why Δy^N_{t-1} is not a valid instrument: Δy^N_{t-1} is correlated with ϵ^N_t because both include $\xi_{i,t-1}$ terms. Of course, ϵ^N_t tends to zero as N tends to infinity: its mean is zero and a calculation using many of the expressions derived in the proof of Proposition 1 shows that

$$\mathrm{Var}(\epsilon_t) = \frac{2\rho}{N} \left[\sigma_A^2 + \sigma_I^2 + \frac{1+\rho}{1-\rho} \mu_A^2 \right].$$

It follows from (42), (43) and Technical Assumption 3 that ϵ_t is uncorrelated with Δy^*_s, for all s, which implies that Δy^*_{t-s} is a valid instrument for $s \geq 1$. And since $\Delta y_{i,t-k}$ are uncorrelated with ξ_{it} and $\xi_{i,t-1}$ for $k \geq 2$, we have that lagged values of Δy, with at least two lags, are valid instruments as well. ∎

Proof of Proposition 5

The equation we estimate is:

$$\Delta y_t = \sum_{k=1}^{K+1} a_k \Delta y_{t-k} + \varepsilon_t, \tag{44}$$

while the true relation is that described by (29) and (30).

It is easy to see that the second term on the right hand side of (29) denoted by w_t in what follows, is uncorrelated with Δy_{t-k}, $k \geq 1$. It follows that estimating (44) is equivalent to estimating (29) with error term

$$w_t = = (1 - \sum_{k=1}^{K} \phi_k) \xi_t \sum_{k=0}^{l_t - 1} \Delta y^*_{t-k},$$

and therefore:

$$\mathrm{plim}_{T \to \infty} \hat{a}_k = \begin{cases} \phi_k & \text{if } k = 1, 2, ..., K, \\ 0 & \text{if } k = K+1. \end{cases}$$

This concludes the proof. ∎

D The Expected Response Time Index: τ

We define the expected response time of Δy to Δy^* as:

$$\tau \equiv \frac{\sum_{k\geq 0} k I_k}{\sum_{k\geq 0} I_k}, \tag{45}$$

with

$$I_k \equiv \mathrm{E}_t\left[\frac{\partial \Delta y_{t+k}}{\partial \epsilon_t}\right].$$

Where $\mathrm{E}_t[\cdot]$ denotes expectations conditional on information (that is, values of Δy and Δy^*) known at time t. This index is a weighted sum of the components of the impulse response function, with weights proportional to the number of periods that elapse until the corresponding response is observed. For example, an impulse response with the bulk of its mass at low lags has a small value of τ, since Δy responds relatively fast to shocks.

Lemma A1 (τ for an Infinite MA) *Consider a second order stationary stochastic process*

$$\Delta y_t = \sum_{k\geq 0} \psi_k \epsilon_{t-k},$$

with $\psi_0 = 1$, $\sum_{k\geq 0} \psi_k^2 < \infty$, the ϵ_t's uncorrelated, and ϵ_t uncorrelated with $\Delta y_{t-1}, \Delta y_{t-2}, \ldots$ Assume that $\Psi(z) \equiv \sum_{k\geq 0} \psi_k z^k$ has all its roots outside the unit disk.

Then:

$$I_k = \psi_k \quad \textit{and} \quad \tau = \frac{\Psi'(1)}{\Psi(1)} = \frac{\sum_{k\geq 1} k\psi_k}{\sum_{k\geq 0} \psi_k}.$$

Proof That $I_k = \psi_k$ is trivial. The expressions for τ then follow from differentiating $\Psi(z)$ and evaluating at $z = 1$. ∎

Proposition A1 (τ for an ARMA Process) *Assume Δy_t follows an ARMA(p,q):*

$$\Delta y_t - \sum_{k=1}^{p} \phi_k \Delta y_{t-k} = \epsilon_t - \sum_{k=1}^{q} \theta_k \epsilon_{t-k},$$

where $\Phi(z) \equiv 1 - \sum_{k=1}^{p} \phi_k z^k$ and $\Theta(z) \equiv 1 - \sum_{k=1}^{q} \theta_k z^k$ have all their roots outside the unit disk. The assumptions regarding the ϵ_t's are the same as in Lemma A1.

Define τ as in (45)*. Then:*

$$\tau = \frac{\sum_{k=1}^{p} k\phi_k}{1 - \sum_{k=1}^{p} \phi_k} - \frac{\sum_{k=1}^{q} k\theta_k}{1 - \sum_{k=1}^{q} \theta_k}.$$

Proof Given the assumptions we have made about the roots of $\Phi(z)$ and $\Theta(z)$, we may write:

$$\Delta y_t = \frac{\Theta(L)}{\Phi(L)} \epsilon_t,$$

where L denotes the lag operator. Applying Lemma A1 with $\Theta(z)/\Phi(z)$ in the role of $\Psi(z)$ we then have:

$$\tau = \frac{\Theta'(1)}{\Theta(1)} - \frac{\Phi'(1)}{\Phi(1)} = \frac{\sum_{k=1}^{p} k\phi_k}{1 - \sum_{k=1}^{p} \phi_k} - \frac{\sum_{k=1}^{q} k\theta_q}{1 - \sum_{k=1}^{q} \theta_k}. \quad ∎$$

Proposition A2 (τ for a Lumpy Adjustment Process) *Consider Δy_t in the simple lumpy adjustment model (12) and τ defined in* (45)*. Then $\tau = \rho/(1-\rho)$.*

Proof $\partial \Delta y_{t+k}/\partial \Delta y_t^*$ is equal to one when the unit adjusts at time $t+k$, not having adjusted between times t and $t+k-1$, and is equal to zero otherwise. Thus:

$$I_k \equiv \mathrm{E}_t\left[\frac{\partial \Delta y_{t+k}}{\partial \Delta y_t^*}\right] = \Pr\{\xi_{t+k}=1\,,\,\xi_{t+k-1}=\xi_{t+k-2}=...=\xi_t=0\} = (1-\rho)\rho^k. \tag{46}$$

The expression for τ now follows easily. ∎

E Rotemberg's Equivalence Result

Proposition 6 (Rotemberg's Equivalence Result)

Agent i controls y_{it}, $i=1,...,N$. The aggregate value of y is defined as $y_t^N \equiv \frac{1}{N}\sum_{i=1}^N y_{it}$. In every period, the cost of changing y is either infinite (with probability ρ) or zero (with probability $1-\rho$) (Calvo Model). When the agent adjusts, it chooses y_{it} equal to $\tilde{y}_t$ that solves

$$\min_{\tilde{y}_t} \mathrm{E}_t \sum_{k\geq 0} (\beta\rho)^k (y_{t+k}^* - \tilde{y}_t)^2,$$

where β denotes the agent's discount factor and y_t^* denotes an exogenous process.[58] *We then have*

$$\tilde{y}_t = (1-\beta\rho)\sum_{k\geq 0}(\beta\rho)^k \mathrm{E}_t y_{t+k}^*. \tag{47}$$

It follows that, as N tends to infinity, y_t^∞ satisfies:

$$y_t^\infty = \rho y_{t-1}^\infty + (1-\rho)\tilde{y}_t. \tag{48}$$

Consider next an alternative adjustment technology (Quadratic Adjustment Costs) where in every period agent i choose y_{it} that solves:

$$\min_{y_{it}} \mathrm{E}_t \sum_{k\geq 0} \beta^k [(y_{t+k}^* - y_{it})^2 + c(y_{it} - y_{i,t-1})^2],$$

where $c>0$ captures the relative importance of quadratic adjustment costs. We then have that there exists $\rho' \in (0,1)$ and $\delta \in (0,1)$ s.t.[59]

$$y_t^\infty = \rho' y_{t-1}^\infty + (1-\rho')\hat{y}_t, \tag{49}$$

with

$$\hat{y}_t = (1-\delta)\sum_{k\geq 0}\delta^k \mathrm{E}_t y_{t+k}^*. \tag{50}$$

Finally, and this is Rotemberg's contribution, a comparison of (47)-(48) *and* (49)-(50) *shows that an*

[58] This formulation can be extended to incorporate idiosyncratic shocks.

[59] The expression that follows is equivalent to the partial adjustment formulation:

$$\Delta y_t^\infty = (1-\rho')(\hat{y}_t - y_{t-1}^\infty),$$

econometrician working with aggregate data cannot distinguish between the Calvo model and the Quadratic Adjustment Costs model described above: ρ' plays the role of ρ and δ the role of $\beta\rho$.

Proof See Rotemberg (1987). ∎

Corollary 1 *Under the assumptions of the Calvo Model in Proposition 6.*

a) *Consider the case where y_t^* follows an AR(1):*

$$y_t^* = \psi y_{t-1}^* + e_t,$$

with $|\psi| < 1$. We then have that $\mathrm{E}_t y_{t+k}^* = \psi^k y_t^*$ and y_t^∞ follows the following AR(2) process:

$$y_t^\infty = (\rho + \psi) y_{t-1}^\infty - \rho\psi y_{t-2}^\infty + \frac{(1-\rho)(1-\beta\rho)}{1-\beta\rho\psi} e_t. \tag{51}$$

b) *Consider the case where Δy_t^* follows an AR(1):*

$$\Delta y_t^* = \phi \Delta y_{t-1}^* + e_t,$$

with $|\phi| < 1$. We then have that

$$\mathrm{E}_t y_{t+k}^* = \frac{\phi(1-\phi^k)}{1-\phi} \Delta y_t^* + y_t^*$$

and Δy_t^∞ follows the following ARMA(2,1) process:

$$\Delta y_t^\infty = (\rho + \phi)\Delta y_{t-1}^\infty - \rho\phi \Delta y_{t-2}^\infty + \frac{1-\rho}{1-\beta\rho\phi}[e_t - \beta\rho\phi e_{t-1}].$$

Proof Straightforward. ∎

F The case where y^* is i.i.d.

Assume that

$$y_{it}^* = y_t^{*A} + y_{it}^{*I}$$

with y_t^{*A} i.i.d. with mean μ_A and variance σ_A^2 and y_{it}^{*I} i.i.d. with zero mean and variance σ_I^2. The y_{it}^{*I} processes are independent across agents and independent from the aggregate shock process y_t^{*A}. The remaining assumptions are the same as in the Technical Assumptions we made in Section 2.

For simplicity we assume $\mu_A = 0$, the case where $\mu_A \neq 0$ just adds a constant to the expressions that follow. Equation (51) then implies that:

$$y_t^\infty = \rho y_{t-1}^\infty + (1-\rho)(1-\beta\rho) y_t^{*A}. \tag{52}$$

We show next that the OLS estimator of ρ in the regression

$$y_t^\infty = \rho y_{t-1}^\infty + e_t \tag{53}$$

provides a consistent estimator of ρ even when N is finite. That is, when the driving processes y^* are i.i.d., there is no missing persistence bias.

Extending the analysis (and notation) from Appendix E to incorporate idiosyncratic shocks, we obtain

$$\tilde{y}_{it} = (1-\beta\rho)\, y^*_{it}.$$

Using the notation we introduced in Appendix C this implies that

$$y^N_t = \frac{1}{N}\sum_{i=1}^{N}(1-\xi_{it})\, y_{i,t-1} + (1-\beta\rho)\frac{1}{N}\sum_{i=1}^{N}\xi_{it} y^*_{it}.$$

Following a similar logic to the one we used in the proof of Proposition 4, we can rewrite the above expression as

$$y^N_t = \rho\, y^N_{t-1} + \varepsilon_t \tag{54}$$

with

$$\varepsilon_t = \frac{1}{N}\sum_{i=1}^{N}(1-\xi_{it}-\rho)\, y_{i,t-1} + (1-\beta\rho)\frac{1}{N}\sum_{i=1}^{N}\xi_{it} y^*_{it}.$$

Even though ε_t differs from the error term in (52), it also is uncorrelated with the regressor y^N_{t-1} which is all we need for $\hat{\rho}$ estimated via OLS from (54) to be a consistent estimator for ρ.

G Simulation details

G.1 Menu cost model

This baseline menu cost is a single sector version of the menu cost model in Nakamura and Steinsson (2010). The household side of the model is straightforward:

$$\max_{n_t, c^i_t} E_0 \sum_{t=0}^{\infty} \beta^t \left[\log C_t - \omega n_t\right],$$

subject to

$$\int_0^1 p^i_t c^i_t\, di \leq W_t n_t + \int_0^1 \pi^i_t,$$

where

$$C_t = \left(\int_0^1 \left(c^i_t\right)^{\frac{\theta-1}{\theta}} di\right)^{\frac{\theta}{\theta-1}}$$

is a Dixit-Stiglitz aggregator of consumption goods c^i_t, p^i_t is the price of good i, n_t is the household's labor supply, ω is the disutility of labor, W_t is the nominal wage, π^i_t is the profits the household receives from owning firm i, and θ is the elasticity of substitution.

Given firm prices, household demand is given by:

$$c^i_t = \left(\frac{p^i_t}{P_t}\right)^{-\theta} C_t,$$

where P_t is the Dixit-Stiglitz price index:

$$P_t = \left(\int_0^1 \left(p^i_t\right)^{1-\theta} di\right)^{\frac{1}{1-\theta}}.$$

The first order condition for labor supply gives:

$$\omega = \lambda_t W_t$$

where λ is the multiplier on the budget constraint. The consumption first order condition implies that

$$\left(c_t^i\right)^{\frac{-1}{\theta}} \left(\int_0^1 \left(c_t^i\right)^{\frac{\theta-1}{\theta}} di\right)^{-1} = \lambda p_t^i$$

Going through a bit more algebra, we get that $\lambda_t = \frac{1}{C_t P_t}$ so the real wage is given by $\frac{W_t}{P_t} = \omega C_t$.
Turning to firm's problem, firms produce using a linear production function in labor

$$y_t^i = z_t^i l_t^i,$$

where firm i's idiosyncratic productivity z_t^i evolves according to

$$\log z_t^i = \rho_z \log z_{t-1}^i + \sigma_z \varepsilon_t^i; \quad \varepsilon_t^i \sim N(0,1)$$

Firms pay a fixed menu cost f in units of labor in order to adjust their nominal price. Given these constraints, the firm i's problem is to choose prices to maximize discounted profits

$$\max_{p_t^i} E_t \sum_{t=0}^{\infty} Q^t \pi_t^i,$$

where $Q = \beta \frac{U'(C')}{U'(C')} = \beta \frac{C}{C'}$ and flow firm profits are given by:

$$\pi_t^i = \left(\underbrace{\frac{p_t^i}{P_t}}_{\text{Unit Revenues}} - \underbrace{\frac{W_t}{z_t^i P_t}}_{\text{Unit Cost}} \right) \underbrace{\left(\frac{p_t^i}{P_t}\right)^{-\theta} C_t}_{\text{Demand}} - \underbrace{f \frac{W_t}{P_t} I_{p_t^i \neq p_{t-1}^i}}_{\text{Menu Cost if Adjusting}}$$

Nominal Demand is assumed to be a random walk in logs: $\log S_{t+1} = \log S_t + \mu + \varepsilon_t$. The aggregate price level will be a function of aggregate spending and the initial distribution of firms $P_t = \varphi\left(\chi\left(p_{-1}, z\right), S\right)$. Given the density of firms χ, φ, and the evolution of χ, we can still write down the firm problem as:

$$\begin{aligned}
V(p_{-1}, z; \chi(p_{-1}, z), S) &= \max\{V^a(z; \chi(p_{-1}, z), S), V^n(p_{-1}, z; \chi(p_{-1}, z), S)\} \\
V^a(z; \chi(p_{-1}, z), S) &= \max_p \left(\frac{p}{P} - \frac{\omega \frac{S}{P}}{z}\right)\left(\frac{p}{P}\right)^{-\theta} \frac{S}{P} - f\omega \frac{S}{P} \\
&\quad + \beta E_{z,S'} \frac{\frac{S}{P}}{\frac{S'}{P'}} V\left(\frac{p}{P'}, \rho_z z + \varepsilon; \chi'\left(p'_{-1}, z'\right), S'\right) \\
V^n(p_{-1}, z; \chi(p_{-1}, z), S) &= \left(\frac{p_{-1}}{P} - \frac{\omega \frac{S}{P}}{z}\right)\left(\frac{p_{-1}}{P}\right)^{-\theta} \frac{S}{P} \\
&\quad + \beta E_{z,S'} \frac{\frac{S}{P}}{\frac{S'}{P'}} V\left(\frac{p_{-1}}{P}, \rho_z z + \varepsilon; \chi'\left(p'_{-1}, z'\right), S'\right) \\
\text{with } P &= \varphi(\chi(p_{-1}, z), S) \ \& \ \chi'\left(p'_{-1}, z'\right) = \Gamma(\chi(p_{-1}, z), S')
\end{aligned}$$

In order to make this problem tractable, we follow Krusell-Smith (1998) and guess that we can accurately forecast how the aggregate price level evolves using the simple log-linear equation:

$$\log \frac{P}{S} = \gamma_0 + \gamma_1 \log \frac{P_{-1}}{S}$$

Consistent with Nakamura and Steinsson (2010), we find that this update rule works well in practice and delivers R^2 in excess of 99%.

G.2 Calibration details

The details of our multi-sector Calvo and Ss models calibration are as follows. We calibrate a 66 sector version of each pricing model. For each sector, we set the average sectoral inflation rate to what is observed in the CPI micro data. We choose the standard deviation of the sectoral inflation rate series, the persistence and standard deviation of the sectoral idiosyncratic shock series (assumed to be an AR(1) in logs) to match the following four moments: the average size of price increases, and decreases, the fraction of price changes that are price increases and the standard deviation of the sectoral inflation rate. In the model, the number of firms in each sector is given by the median (across time) number of firms for that sector in the micro BLS data and each firm was simulated for 238 periods, which is the number of periods in the underlying data.

Table 13 shows basic descriptive statistics for the simulated model. The reported statistics are medians across the 66 sectors, suggesting that both models do a good job matching moments across sectors.

Table 13: CALIBRATION DETAILS: MULTI-SECTOR CALVO AND SS

Calibration Results: Basic Statistics

	CPI	Calvo	Ss
Frequency of monthly adjustment	0.068	0.068	0.068
Fraction of price changes	0.669	0.563	0.668
Average size of price increases	7.997	8.435	9.087
Average size of price decreases	9.073	7.720	8.986
Std deviation of sectoral inflation	0.005	0.004	0.005

G.3 Monte-Carlo evidence: do we recover the true shock In practice?

In order to verify that our shock measure recovers the true shock, we simulate both a Calvo and an Ss model with the following standard parameter values: the frequency of adjustment = 0.2, $\mu_{\text{agg}} = 0.002$, $\sigma_{\text{agg}} = 0.003$, $\rho_I = 0.97; \sigma_I = 0.04$ (we also tried something farther from a random walk: $\rho_I = 0.7$) These economies were simulated for T=300 periods with a burn in of 100 periods. Notice that there are two types of shocks: aggregate shocks that affect everyone and idiosyncratic shocks that are firm specific. In each simulation we ran the following regression:

$$v_t = \alpha + \beta z_t + e_t$$

where v_t is our shock measure (reset price inflation) and z_t is the true shock innovation from each simulation. The level and fit of this regression is informative of how well our shock measure proxies for the true shock. It is an important robustness check because we want to make sure that we can recover an unbiased estimate of the true aggregate shock in a situation where idiosyncratic shocks are realistically large relative to aggregate shocks. The results (averaged across 100 simulations) are comforting and shown below:

Table 14: DOES RESET PRICE INFLATION RECOVER THE TRUE SHOCKS?

REGRESSION OF ESTIMATED SHOCK ON TRUE SHOCK: RESET PRICE INFLATION

			CALVO			SS	
	NFIRMS	INTERCEPT	SLOPE	R^2	INTERCEPT	SLOPE	R^2
$\rho = .7$	500	-0.00	1.02	0.34	-0.00	3.07	0.41
		(0.00)	(0.08)	(0.04)	(0.00)	(0.19)	(0.04)
	5000	-0.00	1.04	0.76	-0.00	3.05	0.67
		(0.00)	(0.03)	(0.02)	(0.00)	(0.18)	(0.04)
	25000	-0.00	1.04	0.85	-0.00	3.07	0.72
		(0.00)	(0.02)	(0.02)	(0.00)	(0.10)	(0.03)
$\rho = .97$	500	-0.00	0.99	0.07	-0.00	2.97	0.28
		(0.00)	(0.21)	(0.03)	(0.00)	(0.26)	(0.04)
	5000	-0.00	1.02	0.35	-0.00	3.00	0.45
		(0.00)	(0.07)	(0.05)	(0.00)	(0.20)	(0.04)
	25000	-0.00	1.01	0.51	-0.00	3.00	0.48
		(0.00)	(0.06)	(0.04)	(0.00)	(0.22)	(0.03)

Unsurprisingly, the overall fit improves in terms of R^2 as the sample sizes increase. Most importantly, we recover the true innovations in the Calvo case and an affine transformation of the innovations in the Ss case for all sample sizes.

G.4 Monte-Carlo evidence

In this section we verify that the methodology we proposed in Section 5.2 for recovering the persistence of sectoral inflation to aggregate and sectoral shocks is an improvement over the standard VAR methodology, which is subject to the missing persistence bias. We test this using a multi-sector Calvo model as a laboratory with both aggregate and sectoral shocks. In this model, the assumptions of Section 3.1 hold so that we know that for a given frequency of adjustment (1-ρ), the estimated response time is equal to $\frac{\rho}{1-\rho}$ to *both* aggregate and sectoral shocks. In other words, in this model we know both what the true level of persistence is and that it is the same to both aggregate and sectoral shocks.

In order to be consistent with our previous work we use our baseline Calvo calibration where $\mu_A = 0.003$, $\sigma_A = 0.0054$, and $\sigma_I = 0.048$ and $\rho = 0.86$. We also consider a second calibration with a higher frequency of adjustment ($\rho = 0.80$) in order to show that our results work for a variety of frequencies. We then simulate data from a version of this model that has 50 sectors, with 200 firms

and 1000 periods per sector. We then implement the two methodologies discussed in Section 5.2. using this simulated data. In particular, we estimate the persistence of sectoral inflation, π_{st} to both aggregate and sectoral shocks. We use the estimated response time as our measure of persistence since we know it's exact value in our simulations and it is what we reported in Table 10. We run this experiment 100 times and average across simulations.

The results are shown in Table 15. The last two columns ("Theory"), show what the true level of persistence in the model. This is equal to $6.14 = \frac{0.86}{0.14}$ in the first calibration and $4.00 = \frac{0.80}{0.20}$ in the second. The first two columns show the results from using VAR's methodology while the second two columns show the results from using our methodology. Two results stick out. Comparing (BEC) to (VAR), we see that our methodology (BEC) does a good job of recovering the true level of persistence to both aggregate and sectoral shocks. The estimated level of persistence to both shocks are (a) similar to each other and (b) close to the true value. This is not true if one uses the VAR methodology. In this case one would infer that inflation responds much more slowly to aggregate shocks than sectoral shocks despite the fact that the true persistence in the model is the same to both shocks.

Table 15: COMPARING METHODS FOR RECOVERING PERSISTENCE

	VAR		BEC		Theory	
	Agg	Sec	Agg	Sec	Agg	Sec
$\rho = 0.86$						
Mean	5.090	1.345	5.779	6.082	6.143	6.143
Median	5.090	1.334	5.843	6.076	6.143	6.143
Std. Deviation	0.000	0.139	0.249	0.033	0.000	0.000
$\rho = 0.80$						
Mean	3.853	1.576	4.026	4.051	4.000	4.000
Median	3.853	1.563	4.029	4.056	4.000	4.000
Std. Deviation	0.000	0.143	0.024	0.033	0.000	0.000

This table documents how different methods of estimating persistence do at recovering the true persistence to nominal shocks. We consider two methodologies: the standard VAR methodology and the one described in Section 5.2 of this paper (BEC). The measure of persistence is the expected response time, which under the assumptions of Section 3.1 (Calvo assumptions) is equal to $\frac{\rho}{1-\rho}$. We consider two calibrations. The first (baseline) uses the same parameter values as our baseline Calvo calibration ($\mu_A = 0.003$, $\sigma_A = 0.0054$, $\sigma_I = 0.048$ and $\rho = 0.86$). The second calibration uses the same parameter values except for $\rho = 0.80$.