

Chapter 3 – Panel estimation of state dependent adjustment when the target is unobserved

This chapter is forthcoming as Deutsche Bundesbank Discussion Paper Series 1, No. 09/2008, Mai 2008. I thank Jörg Breitung for many important discussions, encouragement and patience. Olympia Bover made a vital comment, reminding me of the quasi-differencing strategy. In a conference discussion, John van Reenen set me on a track that ultimately led to this paper. Vassilis Hajivassiliou discussed an early version. Along the way, I had helpful discussions with George von Fürstenberg and Ben Craig. The paper is accepted for presentation at the 2008 Econometric Society, European Meeting in Mailand and has been presented partly or fully at the 2007 Deutsche Bundesbank / Banque de France Spring Conference “Microdata Analysis and Macroeconomic Implications” in Eltville, 2007 Annual Meeting of the Verein für Socialpolitik in Munich and the 2007 CES-Ifo Conference on “Survey Data in Economics - Methodology and Applications” in Munich. The views expressed in this paper do not necessarily reflect those of the Deutsche Bundesbank. All errors, omissions and conclusions remain the sole responsibility of the author.

Abstract:

Understanding adjustment processes has become central in economics. Empirical analysis is fraught with the problem that the target is usually unobserved. This paper develops, simulates and applies GMM methods for estimating dynamic adjustment models in a panel data context with partially unobserved targets and endogenous, time-varying persistence. In this setup, the standard first difference GMM procedure fails. I propose three estimation strategies. One is based on quasi-differencing, and it leads to two different, but related sets of moment conditions. The second is characterised by a state-dependent filter, while the third is an adaptation of the GMM level estimator.

Keywords: Dynamic panel data models, economic adjustment

JEL-Classification: C23, C15, D21

Chapter 3 – Panel estimation of state dependent adjustment when the target is unobserved

1 Introduction

New Keynesian economics, with its emphasis on real and financial frictions, has introduced a focus on microeconomic adjustment dynamics into the empirical literature. Adjustment dynamics are essential for understanding aggregate behaviour and its sensitivity towards shocks. Important examples range from price adjustment and its significance for the New Keynesian Phillips curve (Woodford 2004), over plant level adjustment and aggregate investment dynamics (Caballero, Engel and Haltiwanger, 1995, Caballero and Engel 1999, Bayer 2006), to aggregate employment dynamics, building from microeconomic evidence (Caballero, Engel and Haltiwanger 1997). In these studies, as in Chapter 2, the adjustment dynamics itself becomes the principal object of analysis, instead of being treated as an important, but burdensome obstacle to understanding equilibrium phenomena.

In a rather general form, economic adjustment can be framed by a "gap equation", as formalised by Caballero, Engel and Haltiwanger (1995):

$$\Delta y_{i,t} = \Lambda(g_{i,t}, \mathbf{x}_{i,t}) \cdot g_{i,t}, \text{ where } g_{i,t} = y_{i,t-1} - y_{i,t}^*.$$

Here, subscripts refer to individual i at time t , and $g_{i,t}$ is the gap between the state $y_{i,t-1}$ inherited from the last period and the target $y_{i,t}^*$ that would be realised if adjustment costs were zero for one period of time. The speed of adjustment, written as a function Λ of the gap itself and additional state variables $\mathbf{x}_{i,t}$, determines the fraction of the gap that is removed within one period of time. The adjustment function will reflect convex or non-convex adjustment costs, irreversibility and indivisibilities, financing constraints or other restrictions, and the uncertainty of expectation formation. With quadratic adjustment costs or Calvo-type probabilistic adjustment, Λ will be a constant.

Estimating the function Λ is inherently difficult. In general, both $y_{i,t}^*$ and $g_{i,t}$ will be not observable. But some measure of the gap is needed for any estimation, and if Λ explicitly depends on $g_{i,t}$, this measure will move to the centre stage. In order to ad-

dress this issue, one may first try to do the utmost to observe the target as exactly as possible. The meticulous measurement work of Bayer (2004) and the controversy between Caballero and Engel (2004) and Cooper and Willis (2004) on interpreting the results of gap equation estimates bear testimony to the problems that may result from imperfect measures of the gap.

There is a second route. In linear dynamic panel estimation, the problem can successfully be addressed by positing an error component structure for the measurement error and eliminating the individual fixed effect by a suitable transformation, such as first differencing. See Bond et al. (2003) and Bond and Lombardi (2007) for an error correction model of capital stock adjustment. The GMM estimator developed by Arellano and Bond (1991) accounts for the presence of lagged endogenous variables, the endogeneity of other explanatory variables, and unobserved individual specific effects. Individual effects (including a possible measurement error in the target) are differenced out. Endogenous explanatory variables can be instrumented using lagged dependent variables if the memory of the error process is limited. Time fixed effects can also be accommodated; the remaining idiosyncratic component of the measurement error needs to be uncorrelated with the instruments.

In the unrestricted, non-linear case, this approach is not feasible, as a host of incidental parameters will preclude identification. But there may be direct qualitative information on the level of $\Lambda(\cdot)$, e.g. from survey data, ratings or market information services. If one is willing to treat the adjustment process as piecewise linear, distinguishing regimes of adjustment, then, as will be shown, this information can be harnessed to eliminate the incidental parameters from the problem completely.

Linear dynamic panel estimation was pioneered by Anderson and Hsiao (1982) and it was developed and perfected by Holtz-Eakin, Newey and Rosen (1988), Arellano and Bond (1991), Arellano and Bover (1995) and Blundell and Bond (1998). This paper shows how these methods can be adapted for the analysis of economic adjustment if the target is (partially) unobserved and the non-linearity takes the form of discrete regimes. This is not straightforward, as the unknown and time varying adjustment coefficient interacts with the equally unknown individual specific measurement error. But the re-

ward is substantial: the well-known array of procedures and tests can be brought to bear on the investigation of economic adjustment.

Section 2 of this paper characterises the stochastic process to be estimated. A continuous scalar and a discrete regime vector are evolving jointly, and the adjustment of the continuous-type variable depends on the regime. It is shown that the standard procedure for estimating linear dynamic panel models is not applicable. Section 3 proposes two estimators on the basis of quasi-differencing, one of them with the virtue of great simplicity, the other being more efficient. Both of them are non-linear, which may lead to a small sample bias if in one of the regimes the adjustment speed is almost zero. Section 4 works out two linear GMM estimators that are immune to this problem. One of them uses state dependent filtering, the other is a level estimator applied to a modified model equation. The latter can also cope with contemporaneously correlated regimes. Section 5 tests and compares the proposed routines in a Monte Carlo study.

2 A regime-specific adjustment process

I examine a situation where a variable $y_{i,t}$ reverts to some target level $y_{i,t}^*$ characteristic of individual i . The speed of adjustment depends on the value of $\mathbf{r}_{i,t}$. This is an L -dimensional column vector of regime indicator variables, with one element taking a value of 1, and all others being zero. The equation is

$$\Delta y_{i,t} = -(1 - \alpha_{i,t-1})(y_{i,t-1} - y_{i,t}^*) + \varepsilon_{i,t} \quad (1)$$

with

$$\alpha_{i,t} = \mathbf{a}' \mathbf{r}_{i,t}.$$

The target level $y_{i,t}^*$ is unobservable. It follows an equation that contains an individual-specific latent term:

$$y_{i,t}^* = \mathbf{x}_{i,t}' \boldsymbol{\beta} + \mu_i.$$

The idiosyncratic component μ_i in the adjustment equation may reflect a measurement error or unobserved explanatory variables. The vector $\mathbf{x}_{i,t}$ may encompass random explanatory variables, deterministic time trends and also time dummies. In its absence,

the target level is entirely unobservable, but static. The vector \mathbf{a} holds the state dependent adjustment coefficients. The adjustment coefficient $\alpha_{i,t}$ varies over time and individuals, and $(1 - \alpha_{i,t-1})$ is the adjustment speed at date t . If the process is stable, it would eventually settle in the target in the absence of shocks. We will start by assuming the error term to be a martingale difference sequence:

$$E(\varepsilon_{i,t} | \Omega_{i,t-1}) = 0, \text{ with} \quad (2)$$

$$\Omega_{i,t-1} = \{\mathbf{r}_{i,t-1}, \mathbf{r}_{i,t-2}, \dots, \mathbf{X}_{i,t-1}, \mathbf{X}_{i,t-2}, \dots, \varepsilon_{i,t-1}, \varepsilon_{i,t-2}, \dots, \mu_i, y_{0i}\}.$$

Accommodation to the more general assumption $E(\varepsilon_{i,t} | \Omega_{i,t-k}) = 0, k \geq 1$, with

$$\Omega_{i,t-k} = \{\mathbf{r}_{i,t-1}, \mathbf{r}_{i,t-2}, \dots, \mathbf{X}_{i,t-k}, \mathbf{X}_{i,t-k-1}, \dots, \varepsilon_{i,t-k}, \varepsilon_{i,t-k-1}, \dots, \mu_i, y_{0i}\},$$

is straightforward in every case that will be discussed. Note, however, that this generalisation maintains the assumption of predetermined regime indicators. The case of endogenous regime indicators will be treated separately in Subsection 4.3.

The regime variable $\mathbf{r}_{i,t}$ is generated by a threshold process:

$$\mathbf{r}(k)_{i,t} = \text{Ind}(c_{k-1} \leq s_{i,t} \leq c_k). \quad (3)$$

The unobserved state variable $s_{i,t}$ may, for example, be an autoregressive process or a moving average process of order q . Generally, there will be a non-zero covariance between the error term and the regime indicators, $\text{cov}(\varepsilon_{i,t}, \mathbf{r}_{i,t}) \neq 0$. If, for example, $\varepsilon_{i,t}$ is the error term in a capital accumulation equation and $\mathbf{r}_{i,t}$ is the regime indicating the degree of financing constraints, there should be a contemporaneous correlation between those two.

As we do not observe the target, we have no direct information on the position of the individual relative to the target. But the panel dimension can help us to identify the adjustment process nonetheless, as it allows us to use an error component approach for modelling the unobserved target. In the adjustment equation, both the individual effect

and $\mathbf{x}_{i,t}$ are interacted with a time varying and endogenous variable. Solving for $y_{i,t}$ yields:

$$y_{i,t} = \alpha_{i,t-1}y_{i,t-1} + (1 - \alpha_{i,t-1}) \underbrace{\mathbf{x}'_{i,t}\boldsymbol{\beta} + \mu_i + \varepsilon_{i,t}}_{\text{latent}}. \quad (4)$$

For later purposes it is useful to work out the backward solution to this stochastic difference equation. For $t \geq 1$ and a given starting value $y_{i,0}$ it is:

$$y_{i,t} = \left[y_{i,0} - \boldsymbol{\beta}'\mathbf{x}_{i,1} - \mu_i \right] \prod_{k=0}^{t-1} \alpha_{i,k} + \mathbf{x}'_{i,t}\boldsymbol{\beta} + \mu_i + A_{i,t}, \quad (5)$$

with

$$A_{i,t} = \sum_{l=1}^{t-1} \left(\varepsilon_{i,l} - \Delta \mathbf{x}'_{i,l+1}\boldsymbol{\beta} \right) \prod_{k=l}^{t-1} \alpha_{i,k} + \varepsilon_{i,t}. \quad (6)$$

The solution has three components. The first term captures the influence of the initial deviation. The second term is the target level at time t , $\mathbf{x}'_{i,t}\boldsymbol{\beta} + \mu_i$. The third term, $A_{i,t}$, represents the effect of shocks and target variations, past and present. In the long run, when the influence of the initial conditions has died out, $A_{i,t}$ is equal to the deviation from the target.

Anderson and Hsiao (1982) have devised the classic strategy for estimating linear dynamic panel equations with fixed effects. Consider a first-order autoregressive equation:

$$y_{i,t} = \gamma y_{i,t-1} + \mu_i + \varepsilon_{i,t}.$$

Obviously, the latent fixed effect μ_i is correlated with the explanatory variable. Transforming the equation by taking first differences eliminates the fixed effect:

$$\Delta y_{i,t} = \gamma \Delta y_{i,t-1} + \Delta \varepsilon_{i,t}.$$

Now the transformed error term $\Delta \varepsilon_{i,t}$ is correlated with the transformed regressor, $\Delta y_{i,t-1}$. This can be accommodated using an instrument variable procedure. Anderson and Hsiao propose using either lagged first differences or lagged levels as instruments. Employing second and further lags of the level as instruments for the differenced equation makes use of the following moment restrictions:

$$E(y_{i,t-s} \cdot \Delta \varepsilon_{i,t}) = 0, \quad s = 2, 3, \dots$$

If these moment restrictions hold, then the lagged levels will be valid instruments, because they are correlated with the regressor variable. The suggestion of Anderson and Hsiao was refined by Holtz-Eakin, Newey and Rosen (1988) and Arellano and Bond (1991), who propose to use an efficient GMM estimator that uses all available moment restrictions optimally, instead of the IV or 2SLS method. Formally, the moment equations are written as a system, in order to be able to use a varying number of instruments according to availability. The instruments are weighted optimally using the Hansen (1982) two-stage procedure.

In order to investigate the feasibility of the standard approach in the context of the model with time-varying coefficients, we look at the first difference of equation (2), also focussing on the simple case of a static target:

$$\Delta y_{i,t} = \alpha' \Delta(\mathbf{r}_{i,t-1} y_{i,t-1}) + \underbrace{(\mathbf{1} - \alpha') \Delta(\mathbf{r}_{i,t-1} \mu_i)}_{\text{latent process}} + \Delta \varepsilon_{i,t}. \quad (7)$$

Unlike the linear case, the expression containing the unobserved μ_i is not differenced out, and we have to deal with a time-varying error component that is correlated with the explanatory variables. Instruments that are uncorrelated with this latent process, but correlated with the explanatory variables in such a way that each of the coefficients is identified are hard to come by. The following sections are devoted to finding moment conditions that make estimation feasible in practice.

3 Two non-linear moment conditions based on quasi-differencing

This section discusses two nonlinear transformations of the adjustment equation that eliminate the unobserved heterogeneity. Holtz-Eakin, Newey and Rosen (1988) proposed quasi-differencing as a strategy in a case where fixed effects are subject to time varying shocks that are common across individuals.¹ We explore whether this method can be generalised to the more complicated case at hand, where coefficients are endogenous and vary over time and individuals.

¹ See also Chamberlain (1983), p. 1263-64. I thank Olympia Bover for reminding me of this 'classical' strategy.

Literally, the quasi-differencing procedure as proposed by these authors involves lagging equation (1), multiplying both sides by $\frac{1-\alpha_{i,t-1}}{1-\alpha_{i,t-2}}$ and subtracting the result from equation (1). After reordering coefficients, this gives:

$$\Delta y_{i,t} - \frac{1-\alpha_{i,t-1}}{1-\alpha_{i,t-2}} \alpha_{i,t-2} \Delta y_{i,t-1} - (1-\alpha_{i,t-1}) \Delta \mathbf{x}'_{i,t} \boldsymbol{\beta} = \varepsilon_{i,t} - \frac{1-\alpha_{i,t-1}}{1-\alpha_{i,t-2}} \varepsilon_{i,t-1}. \quad (8)$$

The unobserved heterogeneity has duly been eliminated, but this equation is difficult to deal with, because in general $\alpha_{i,t-1}$ is correlated with $\varepsilon_{i,t-1}$ and $\alpha_{i,t-2}$. The underlying idea nonetheless leads to useful moment conditions, actually in two different ways. First, dividing equation (8) by $(1-\alpha_{i,t-1})$ gives

$$\frac{1}{1-\alpha_{i,t-1}} \Delta y_{i,t} - \frac{\alpha_{i,t-2}}{1-\alpha_{i,t-2}} \Delta y_{i,t-1} - \Delta \mathbf{x}'_{i,t} \boldsymbol{\beta} = \psi_{i,t}, \quad (9)$$

with

$$\psi_{i,t} = \frac{\varepsilon_{i,t}}{1-\alpha_{i,t-1}} - \frac{\varepsilon_{i,t-1}}{1-\alpha_{i,t-2}}. \quad (10)$$

This transformation – which shall be referred to as "QD1" – corresponds to solving equation (1) for the expression $y_{i,t-1} - \mathbf{x}'_{i,t} \boldsymbol{\beta} - \mu_i$, then solving the lagged version of (1) for $y_{i,t-2} - \mathbf{x}'_{i,t-1} \boldsymbol{\beta} - \mu_i$ and ultimately differencing μ_i out. Second, we may multiply equation (9) by $1-\alpha_{i,t-2}$, to obtain:

$$\frac{1-\alpha_{i,t-2}}{1-\alpha_{i,t-1}} \Delta y_{i,t} - \alpha_{i,t-2} \Delta y_{i,t-1} - (1-\alpha_{i,t-2}) \Delta \mathbf{x}'_{i,t} \boldsymbol{\beta} = \xi_{i,t}, \quad (11)$$

with

$$\xi_{i,t} = \frac{1-\alpha_{i,t-2}}{1-\alpha_{i,t-1}} \varepsilon_{i,t} - \varepsilon_{i,t-1}. \quad (12)$$

This transformation shall be labelled "QD2". It corresponds to multiplying equation (1) by $(1-\alpha_{i,t-2})/(1-\alpha_{i,t-1})$ and subtracting the lag of the original adjustment equation.

Proposition 1: Under assumption (2), the levels $y_{i,t-p}$, $\mathbf{x}_{i,t-p}$ and the regime indicators $\mathbf{r}_{i,t-p}$, $p \geq 2$, are instruments in equations (9) and (11), that is:

$$\begin{aligned} E(y_{i,t-p} \boldsymbol{\psi}_{i,t}) &= E(y_{i,t-p} \boldsymbol{\xi}_{i,t}) = \mathbf{0} \quad , \\ E(\mathbf{x}_{i,t-p} \boldsymbol{\psi}_{i,t}) &= E(\mathbf{x}_{i,t-p} \boldsymbol{\xi}_{i,t}) = \mathbf{0} \quad , \\ E(\mathbf{r}_{i,t-p} \boldsymbol{\psi}_{i,t}) &= E(\mathbf{r}_{i,t-p} \boldsymbol{\xi}_{i,t}) = \mathbf{0} . \end{aligned}$$

Proof: If $E(\boldsymbol{\varepsilon}_{i,t} | \Omega_{i,t-1}) = \mathbf{0}$, with $\Omega_{i,t-1}$ some information set that varies over individuals and time, then any function $f(\Omega_{i,t-1})$ will be orthogonal to $\boldsymbol{\varepsilon}_{i,t}$, because

$$E[f(\Omega_{i,t-1}) \boldsymbol{\varepsilon}_{i,t}] = E[f(\Omega_{i,t-1}) \boldsymbol{\varepsilon}_{i,t} | \Omega_{i,t-1}] = E[f(\Omega_{i,t-1}) E(\boldsymbol{\varepsilon}_{i,t} | \Omega_{i,t-1})] = \mathbf{0} . \quad (13)$$

Consider first $E(y_{i,t-p} \boldsymbol{\psi}_{i,t})$, with $p \geq 2$. By iterating equation (1), $y_{i,t-p}$ is a function of

$(\mathbf{r}_{i,t-p-1}, \mathbf{r}_{i,t-p-2}, \dots, \mathbf{x}_{i,t-p}, \mathbf{x}_{i,t-p-1}, \dots, \boldsymbol{\varepsilon}_{i,t-p}, \boldsymbol{\varepsilon}_{i,t-p-1}, \dots, \mu_i, y_{i,0})$. The expressions $\frac{1}{1-\alpha_{i,t-1}}$ and $\frac{1}{1-\alpha_{i,t-2}}$ are functions of $\mathbf{r}_{i,t-1}$ and $\mathbf{r}_{i,t-2}$. Applying (13) to the products $\frac{y_{i,t-p}}{1-\alpha_{i,t-1}} \boldsymbol{\varepsilon}_{i,t}$ and $\frac{y_{i,t-p}}{1-\alpha_{i,t-2}} \boldsymbol{\varepsilon}_{i,t-1}$ yields $E(y_{i,t-p} \boldsymbol{\psi}_{i,t}) = \mathbf{0}$. The other orthogonalities follow likewise. \square

If assumption (2) is replaced by $E(\boldsymbol{\varepsilon}_{i,t} | \Omega_{i,t-k}) = \mathbf{0}$, then the set of valid instruments is pushed backward in time accordingly.

To discuss estimation on the basis of the two sets of moment conditions, it is useful to rewrite the transformations (9) and (11) somewhat. Equation (9) has the convenient feature that $\Delta \mathbf{x}'_{i,t} \boldsymbol{\beta}$ enters additively. Collecting terms, we can write:

$$\begin{aligned} \boldsymbol{\psi}_{i,t} &= \Delta y_{i,t-1} + \left(\frac{1}{1-\alpha_{i,t-1}} \Delta y_{i,t} - \frac{1}{1-\alpha_{i,t-2}} \Delta y_{i,t-1} \right) - \Delta \mathbf{x}'_{i,t} \boldsymbol{\beta} \\ &= \Delta y_{i,t-1} + \Delta(\boldsymbol{\gamma}' \mathbf{r}_{i,t-1} \Delta y_{i,t}) - \Delta \mathbf{x}'_{i,t} \boldsymbol{\beta} \\ &= \Delta y_{i,t-1} + \boldsymbol{\gamma}' \Delta(\mathbf{r}_{i,t-1} \Delta y_{i,t}) - \Delta \mathbf{x}'_{i,t} \boldsymbol{\beta} , \end{aligned} \quad (14)$$

with

$$\boldsymbol{\gamma}' = \left(\frac{1}{1-\alpha_1} \quad \cdots \quad \frac{1}{1-\alpha_L} \right). \quad (15)$$

Equation (14) is linear in the coefficient vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, and can be estimated by linear GMM using the moment conditions of Proposition 1. It relates the structural coefficients $\boldsymbol{\alpha}$ to the elements of $\boldsymbol{\gamma}$ by a nonlinear one-to-one transformation, see equation (15). Inverting this transformation therefore gives a nonlinear GMM estimator of $\boldsymbol{\alpha}$. Standard deviations and covariances can be assessed using the delta method.

Putting QD2 to use for GMM estimation is trickier. Let $\mathbf{d}(\mathbf{r}_{i,t-2}, \mathbf{r}_{i,t-1})$ be an $L^2 \times 1$ indicator vector, where each element is a dummy variable indicating one of the possible switches from $\mathbf{r}_{i,t-2}$ to $\mathbf{r}_{i,t-1}$. Let $\boldsymbol{\lambda}$ be the vector of coefficients $(1-\alpha_{i,t-2})/(1-\alpha_{i,t-1})$ corresponding to the elements of $\mathbf{d}(\cdot)$:

$$\boldsymbol{\lambda}' = \left(1 \quad \frac{1-\alpha_1}{1-\alpha_2} \quad \frac{1-\alpha_1}{1-\alpha_3} \quad \cdots \quad \cdots \quad \frac{1-\alpha_L}{1-\alpha_{L-2}} \quad \frac{1-\alpha_L}{1-\alpha_{L-1}} \quad 1 \right).$$

Let furthermore $\boldsymbol{\delta}$ be a vector of products of the adjustment coefficients, $(\mathbf{1}-\boldsymbol{\alpha})$, and $\boldsymbol{\beta}$:

$$\boldsymbol{\delta} = (\mathbf{1}-\boldsymbol{\alpha}) \otimes \boldsymbol{\beta} = \begin{pmatrix} (1-\alpha_1)\boldsymbol{\beta} \\ (1-\alpha_2)\boldsymbol{\beta} \\ \vdots \\ (1-\alpha_L)\boldsymbol{\beta} \end{pmatrix}.$$

Finally, let

$$\boldsymbol{\pi} = \begin{pmatrix} \boldsymbol{\lambda} \\ -\boldsymbol{\alpha} \\ -\boldsymbol{\delta} \end{pmatrix} = \mathbf{h}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \quad (16)$$

be an $L(L+1+K) \times 1$ vector of reduced form coefficients, of which $L(L+K)$ are unknown. Then we can write:

$$\begin{aligned}\xi_{i,t} &= \lambda' \mathbf{d}(\mathbf{r}_{i,t-2}, \mathbf{r}_{i,t-1}) \Delta y_{i,t} - \boldsymbol{\alpha}' \mathbf{r}_{i,t-2} \Delta y_{i,t-1} - \boldsymbol{\delta}' \mathbf{r}_{i,t-2} \Delta \mathbf{x}_{i,t} \\ &= \left[\mathbf{d}(\mathbf{r}_{i,t-2}, \mathbf{r}_{i,t-1})' \Delta y_{i,t} \quad \mathbf{r}_{i,t-1}' \Delta y_{i,t-1} \quad \mathbf{r}_{i,t-1}' \Delta \mathbf{x}_{i,t} \right] \boldsymbol{\pi}.\end{aligned}$$

As with QD1, this equation is non-linear in the structural parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and linear in a transformed coefficient vector. However, here there is no convenient one-to-one transformation from $\boldsymbol{\pi}$ to the structural parameter. The nonlinearity of the problem therefore has to be treated explicitly. Consider the simplest case, with two states and no explanatory variables $\mathbf{x}_{i,t}$. Then λ and $\boldsymbol{\alpha}$ have two elements each and we can write:

$$\boldsymbol{\pi}' = \mathbf{h}(\boldsymbol{\alpha})' = \begin{pmatrix} 1 & \frac{1-\alpha_1}{1-\alpha_2} & \frac{1-\alpha_2}{1-\alpha_1} & 1 & -\alpha_1 & -\alpha_2 \end{pmatrix}.$$

In principle, there are two ways of estimating the structural parameters. First, we may estimate the coefficients $\boldsymbol{\pi}$ together with the covariance matrix, and then go to the structural parameters using (16). As the reduced form has more parameters than the structural equation, the structural parameters are over-determined. The information can be aggregated efficiently using the classical minimum distance (CMD) estimator. Second, we may treat the transformed equation directly as a nonlinear estimation problem in the structural parameters. These alternatives shall be discussed in turn.

Let $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_0' \quad \boldsymbol{\beta})'$ be the true vector of structural coefficients and $\boldsymbol{\pi}_0 = \mathbf{h}(\boldsymbol{\theta}_0)$ be the true vector of reduced form coefficients. We assume that there is a consistent and asymptotically normal estimator $\hat{\boldsymbol{\pi}}_N$ of $\boldsymbol{\pi}_0$, with $\text{Avar} \sqrt{N} (\hat{\boldsymbol{\pi}}_N - \boldsymbol{\pi}_0) = \boldsymbol{\Xi}_0$. The vector $\hat{\boldsymbol{\pi}}_N$ could, for example, be a GMM estimate of the reduced form equation. Any hypothetical value $\boldsymbol{\theta}$ of the structural coefficients implies a vector of reduced form coefficients $\mathbf{h}(\boldsymbol{\theta})$. The CMD estimator determines $\hat{\boldsymbol{\theta}}$ in such a way that the weighted deviations of $\mathbf{h}(\hat{\boldsymbol{\theta}})$ from their counterparts $\hat{\boldsymbol{\pi}}$ resulting from the unconstrained estimation is minimised.² That is, $\hat{\boldsymbol{\theta}}$ is to solve:

$$\min_{\hat{\boldsymbol{\theta}}} \left(\hat{\boldsymbol{\pi}} - \mathbf{h}(\hat{\boldsymbol{\theta}}) \right)' \boldsymbol{\Omega} \left(\hat{\boldsymbol{\pi}} - \mathbf{h}(\hat{\boldsymbol{\theta}}) \right),$$

² See Wooldridge (2001) and Newey and McFadden (1994) for a discussion of CMD estimation.

with $\mathbf{\Omega}$ a possibly data dependent positive definite weighting matrix. Under these assumptions, the CMD estimator is consistent and asymptotically normal.

A weighting matrix is efficient if it leads to a CMD estimator with a "smaller" variance than what could be obtained from any other weighting matrix, in the sense that the difference between their asymptotic covariance matrices is positive semi-definite. It can be shown that an efficient weighting matrix is given by $\mathbf{\Omega} = \hat{\mathbf{\Xi}}^{-1}$, with $\hat{\mathbf{\Xi}}$ any matrix such that $\text{plim}_{N \rightarrow \infty} \hat{\mathbf{\Xi}} = \mathbf{\Xi}_0$, provided that $\mathbf{\Xi}_0$ has full rank. Therefore, the inverse of any consistent estimator of $\text{Avar} \sqrt{N}(\hat{\boldsymbol{\pi}}_N - \boldsymbol{\pi})$ is an efficient weighting matrix. Let $\mathbf{H}(\boldsymbol{\theta})$ be the $L^2 \times L$ matrix of partial derivatives of $\mathbf{h}(\boldsymbol{\theta})$:

$$\mathbf{H}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathbf{h}(\boldsymbol{\theta}) = \left(\frac{\partial \mathbf{h}(\boldsymbol{\theta})}{\partial \theta_1}, \dots, \frac{\partial \mathbf{h}(\boldsymbol{\theta})}{\partial \theta_{L+K}} \right).$$

The i 'th column of $\mathbf{H}(\boldsymbol{\theta})$ is the derivative of $\mathbf{h}(\boldsymbol{\theta})$ with respect to θ_i . Using an efficient weighting matrix leads to:

$$\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \text{N}\left(0, [\mathbf{H}(\boldsymbol{\theta}_0)' \mathbf{\Xi}_0^{-1} \mathbf{H}(\boldsymbol{\theta}_0)]^{-1}\right).$$

The appropriate estimator for the covariance matrix of $\hat{\boldsymbol{\theta}}$ then is:

$$\text{Est var}(\hat{\boldsymbol{\theta}}) = [\mathbf{H}(\hat{\boldsymbol{\theta}})' \hat{\mathbf{\Xi}}^{-1} \mathbf{H}(\hat{\boldsymbol{\theta}})]^{-1} / N = [\mathbf{H}(\hat{\boldsymbol{\theta}})' (\text{Est var}(\hat{\boldsymbol{\pi}}))^{-1} \mathbf{H}(\hat{\boldsymbol{\theta}})]^{-1}.$$

Linear restrictions, such as the equality of coefficients, can be subjected to a standard Wald-test. Alternatively, a criterion function test statistic is available.³

There is a drawback to the CMD procedure in the given context. For asymptotic efficiency we need the matrix $\text{plim}_{N \rightarrow \infty} \hat{\mathbf{\Xi}} = \mathbf{\Xi}_0$ to be of full rank, such that the inverse matrix can be formed. If the reduced form is to be estimated by linear GMM, this requires that each of the reduced form parameters is separately identified by the moment conditions. This will not always be possible. As we have seen, $\boldsymbol{\pi}$ represents four reduced form parameters in the case of two states and no explanatory variables. With three states, it is already nine parameters. Each explanatory variable adds L parameters to the reduced

form coefficients vector. In practice, the information content of the available instruments may not be sufficient to identify all of the many reduced form parameters separately. And although CMD estimation can be performed on the basis of any positive definite matrix, the weighting matrix $\mathbf{\Omega} = \hat{\mathbf{\Xi}}^{-1}$ for efficient CMD would then cease to exist in the limit.

Therefore we may prefer to estimate directly in terms of the underlying structural parameters:

$$\xi_{i,t} = \left[\mathbf{d}(\mathbf{r}_{i,t-2}, \mathbf{r}_{i,t-1})' \Delta y_{i,t} \quad \mathbf{r}_{i,t-1}' \Delta y_{i,t-1} \quad \mathbf{r}_{i,t-1}' \Delta \mathbf{x}_{i,t} \right] \cdot \mathbf{h}(\boldsymbol{\alpha} \quad \boldsymbol{\beta}) \quad (17)$$

Though nonlinear in the parameters, this equation is linear in the transformed variables. This makes it easy to use the Gauss-Newton method for solving the optimisation problem inherent in GMM estimation, using routines for linear GMM in performing the iteration steps. Appendix B elaborates on the Gauss-Newton method in the context of non-linear GMM problems. As initial values for iteration, we can either use CMD estimates or the results from nonlinear indirect estimation exposed earlier in this section.

The transformations QD1 and QD2 are nonlinear, and the stochastic properties of the transformed residuals depend on the adjustment parameters. Consider the transformed

residuals $\psi_{i,t} = \frac{\varepsilon_{i,t}}{1 - \alpha_{i,t-1}} - \frac{\varepsilon_{i,t-1}}{1 - \alpha_{i,t-2}}$ on the one hand and $\xi_{i,t} = \frac{1 - \alpha_{i,t-2}}{1 - \alpha_{i,t-1}} \varepsilon_{i,t} - \varepsilon_{i,t-1}$ on the

other. The variance of $\psi_{i,t}$ will become large if one or both alpha-coefficients are in the neighbourhood of 1, creating problems in small samples. An adjustment coefficient approaching 1 will affect $\xi_{i,t}$ to a lesser degree. First, only one of the two components of the difference is affected. Second, the effect is mitigated by the denominator, $1 - \alpha_{i,t-2}$.

The random factor $(1 - \alpha_{i,t-2}) / (1 - \alpha_{i,t-1})$ in $\xi_{i,t}$ can take three values, of which only one is larger than 1. Indeed, if the alpha coefficients are of similar size, the random factor will stay in the neighbourhood of 1. Therefore, when the alpha coefficients are high (i.e. adjustment speed is low), efficiency gains can be expected from using QD2. I will investigate this in a simulation study below.

³ For this test, and a criterion function specification test, see Wooldridge (2002).

4 Two moment conditions based on linear transformations

4.1 Forward differences and generalised differences

Being nonlinear, the transformation we just investigated may lead to poor results if in one or more of the regimes the adjustment speed is very low. They cannot be used at all if one of the regimes is characterised by an adjustment speed of exactly zero. This is a case of considerable theoretical interest, as the presence of fixed adjustment costs or irreversibility leads to bands around the target where no adjustment takes place – the solution to the stochastic control problem triggers adjustment when some threshold level is surpassed. In a literal sense, this sort of behaviour is to be expected only when decisions on single projects are considered, as opposed to entire firms or sectors. But it is certainly useful to explicitly consider regimes of no adjustment, as have done parts of the literature, eg. Caballero, Engel and Haltiwanger (1995)

To this end, it may be worth asking whether there is a linear transformation that could be brought to bear on the problem at hand, in the spirit of the first differencing procedure. As we shall see, there is such a transformation if the regime indicator has limited memory with respect to $\varepsilon_{i,t}$. We start by looking again at the first difference of $y_{i,t}$:

$$\Delta y_{i,t} = \boldsymbol{\alpha}' \Delta(\mathbf{r}_{i,t-1} y_{i,t-1}) + (\mathbf{1} - \boldsymbol{\alpha}') \Delta(\mathbf{r}_{i,t-1} \mathbf{x}'_{i,t}) \boldsymbol{\beta} + (\mathbf{1} - \boldsymbol{\alpha}') \Delta(\mathbf{r}_{i,t-1} \boldsymbol{\mu}_i) + \Delta \varepsilon_{i,t}.$$

For unchanging adjustment regimes, $\mathbf{r}_{i,t-1} = \mathbf{r}_{i,t-2}$, this simplifies to

$$\Delta y_{i,t} = \boldsymbol{\alpha}' \mathbf{r}_{i,t-1} \Delta y_{i,t-1} + (\mathbf{1} - \boldsymbol{\alpha}') \mathbf{r}_{i,t-1} \Delta \mathbf{x}'_{i,t} \boldsymbol{\beta} + \Delta \varepsilon_{i,t}.$$

This expression looks very much like the first difference in the linear case, although there is more than one adjustment coefficient to estimate. Taking first differences of observations that belong to *different* regimes leads to a latent term $(\mathbf{1} - \boldsymbol{\alpha}') \Delta \mathbf{r}_{i,t-1} \boldsymbol{\mu}_i$ that will be correlated with the lagged dependent variable under a variety of circumstances.

As it is this term that makes the use of the standard technique difficult, the following strategy comes to mind: Differences are only formed for observations with $\mathbf{r}_{i,t-2} = \mathbf{r}_{i,t-1}$. On the basis of cases where two consecutive observations belong to the first regime, we could estimate a_1 , and using differences of observations that both belong to the second regime, we could infer on a_2 , etc. In this straight fashion, however, the idea will not

work. The transformed residual $\Delta\varepsilon_{i,t}$ has an expectation different from zero in the two groups of observations. This is because $\mathbf{r}_{i,t-1}$ and $\varepsilon_{i,t-1}$ are correlated by assumption. The expectation $E(\varepsilon_{i,t-1} | \mathbf{r}(1)_{i,t-1} = 1)$ is not equal to zero, and neither is $E(\varepsilon_{i,t-1} | \mathbf{r}(2)_{i,t-1} = 1)$. Selecting residuals according to regimes will lead to biased estimators.

If $\varepsilon_{i,t}$ is uncorrelated with past regime indicators, $\mathbf{r}_{i,t-1}, \mathbf{r}_{i,t-2}, \dots$, then we are able to use a modified differencing approach. Autocorrelation of $\varepsilon_{i,t}$ is permitted if the usual requirement of limited memory is satisfied. The following two principles will generate moment conditions involving the use of lagged endogenous variables as instruments:

1. Let q be the maximum τ for which there is a correlation between $\mathbf{r}_{i,t}$ and $\varepsilon_{i,t-\tau}$, eg. as a consequence of an MA structure of the state driving the regime indicator as exemplified in Assumption 2. Then the observation is to be transformed subtracting past observations of the same regime with a lag of at least $s = 2 + q$.
2. If an observation is not matched by a $2 + q$ -lag in the same regime, it may be transformed using any other lag $s > q + 2$.

The second principle avoids the loss of many observations in cases where regimes in t and $t+q$ do not match because of regime switches. What I propose here is a dynamic filter, which varies according to regimes.

Similar to (7) we obtain for the s 'th difference:

$$(y_{i,t} - y_{i,t-s}) = \boldsymbol{\alpha}'(\mathbf{r}_{i,t-1}y_{i,t-1} - \mathbf{r}_{i,t-s-1}y_{i,t-s-1}) + (1 - \boldsymbol{\alpha}')(\mathbf{r}_{i,t-1} - \mathbf{r}_{i,t-s-1})\boldsymbol{\mu}_i + (\varepsilon_{i,t} - \varepsilon_{i,t-s}),$$

which simplifies to

$$(y_{i,t} - y_{i,t-s}) = \boldsymbol{\alpha}'\mathbf{r}_{i,t-1}(y_{i,t-1} - y_{i,t-s-1}) + (\varepsilon_{i,t} - \varepsilon_{i,t-s}),$$

if the two observations are characterised by the same regime, such that $\mathbf{r}_{i,t-1} = \mathbf{r}_{i,t-s-1}$.

When does the conditional expectation of the residual term, $(\varepsilon_{i,t} - \varepsilon_{i,t-s})$, become zero?

It is sufficient that $\varepsilon_{i,t}$ and $\varepsilon_{i,t-s}$ are both uncorrelated with the conditioning variables, which are $\mathbf{r}_{i,t-1}$ and $\mathbf{r}_{i,t-s-1}$. Now assume $\varepsilon_{i,t}$ to be uncorrelated with $\mathbf{r}_{i,t-1}$ and $\mathbf{r}_{i,t-s-1}$.

Then the same is true with respect to $\varepsilon_{i,t-s}$ and $\mathbf{r}_{i,t-s-1}$. Therefore, by choosing s , we have only to make sure that $\varepsilon_{i,t-s}$ and $\mathbf{r}_{i,t-1}$ are uncorrelated. This will never happen with $s = 1$, as we have seen before. However, if $\mathbf{r}_{i,t}$ is uncorrelated with all lags of $\varepsilon_{i,t}$, then $s = 2$ will ensure that

$$\mathbb{E}\left(\varepsilon_{i,t} - \varepsilon_{i,t-s} \mid \mathbf{r}_{i,t-1} = \mathbf{r}_{i,t-s-1}\right) = 0, \quad (18)$$

regardless of whatever value $\mathbf{r}_{i,t-1}$ and $\mathbf{r}_{i,t-s-1}$ take. More generally, if there is correlation between $\mathbf{r}_{i,t}$ and $\varepsilon_{i,t-\tau}$ up to lag $\tau = q$, the difference that guarantees the above equation to hold will be at least of order $s = 2 + q$. GMM estimation on the basis of this transformation may be called *forward difference estimation*. But we are not restricted to using only differences of the order that is "just right", i.e. $2 + q$. Any other difference of order $s \geq 2 + q$ will fulfil eq. (18) just as well. Therefore I construct the difference using the most proximate observation of the same regime with lag $s \geq 2 + q$. With respect to admissibility and validity of instruments, the rules of the classic approach apply: the instruments need to be uncorrelated with the earlier of the two observations that make up the difference. In the following, this procedure will be called the *generalised difference estimator*.

To state the moment condition, I have to strengthen assumption (2). In addition to the variables in the conditioning set $\Omega_{i,t-1}$, $\varepsilon_{i,t}$ must also be uncorrelated to the future regimes $\mathbf{r}_{i,t+q+1}, \mathbf{r}_{i,t+q+2}, \dots$

Proposition 2: Let the conditional expectation of $\varepsilon_{i,t}$ satisfy

$$\mathbb{E}\left(\varepsilon_{i,t} \mid \Omega_{i,t-1}, \mathbf{r}_{i,t+q+1}, \mathbf{r}_{i,t+q+2}, \dots\right) = 0. \quad (19)$$

with $\Omega_{i,t-1}$ defined as in (2). Then the levels $y_{i,t-s-p}$, $p \geq 1$ are valid instruments for the equations transformed by taking the s 'th difference, with $s \geq 2 + q$:

$$\mathbb{E}\left(\left(\varepsilon_{i,t} - \varepsilon_{i,t-s}\right) y_{i,t-s-p} \mid \mathbf{r}_{i,t-1} = \mathbf{r}_{i,t-s-1}\right) = 0. \quad (20)$$

Proof: The proposition follows from the law of iterated expectations:

$$\begin{aligned} & \mathbb{E}\left(y_{i,t-s-p} \left(\varepsilon_{i,t} - \varepsilon_{i,t-s}\right) \mid \mathbf{r}_{i,t-1}, \mathbf{r}_{i,t-s-1}\right) \\ &= \mathbb{E}_{y_{i,t-s-p}} \left(\mathbb{E}\left(y_{i,t-s-p} \left(\varepsilon_{i,t} - \varepsilon_{i,t-s}\right) \mid \mathbf{r}_{i,t-1}, \mathbf{r}_{i,t-s-1}, y_{i,t-s-p}\right) \right) \\ &= \mathbb{E}_{y_{i,t-s-p}} \left(y_{i,t-s-p} \cdot \mathbb{E}\left(\varepsilon_{i,t} - \varepsilon_{i,t-s} \mid \mathbf{r}_{i,t-1}, \mathbf{r}_{i,t-s-1}, y_{i,t-s-p}\right) \right) = 0, \end{aligned}$$

because the conditional expectation within the brackets is zero for $s \geq 2 + q$. The backward solution (3) and (4) decomposes $y_{i,t}$ into $y_{i,0}$, μ_i , and the history of $\varepsilon_{i,t}$ and $\mathbf{r}_{i,t}$. Condition (19) ensures that the expected values of $\varepsilon_{i,t}$ and $\varepsilon_{i,t-s}$ do not depend on the components of $y_{i,t-s-p}$. The additional conditioning on $y_{i,t-s-p}$ can have no influence on the expected value. \square

As in the case of the two nonlinear estimators, $\mathbf{r}_{i,t-1}$ must be uncorrelated with the current error term. The generalised difference approach cannot work if the regime indicator is contemporaneous with respect to the current error term. Furthermore, it is an identifying assumption for the process that drives the regime indicator to have finite memory with respect to innovations $\varepsilon_{i,t}$. This is a limitation of the approach. If $\mathbf{r}_{i,t}$ were correlated with all past values of $\varepsilon_{i,t}$, the conditional expectation of the transformed error term resulting from a difference of two observations from the same regime would not disappear. The resulting bias can be expected to wane if the minimum lag length is chosen to be large. But doing so would result in losing many observations, exacerbating another weakness of the estimation strategy.

4.2. Testing the validity of the length of memory

In order to use generalised differencing, we need to decide on the length of the memory of the process driving the regime with respect to $\varepsilon_{i,t}$. This is difficult to do on an *a priori* basis. There are two simple solutions. The first is to use the Sargan-Hansen test to check the appropriateness of the transformation. This is straightforward, as the Sargan-Hansen test is a test of the validity of the moment conditions. The drawback is that the Sargan-Hansen test is generally used as an omnibus test of the specification, including the choice of the instruments. If we employ the Sargan-Hansen test as a means of finding the correct lag length, then estimation will be conditional on the test statistic being insignificant. For further purposes, the test is spent.

Alternatively, we may base a test on the fact that the expected value of the residual will not disappear if the lag length chosen is too short. In that case, as we have seen, the choice of observations belonging to one regime or the other will select positive or negative outcomes of $\varepsilon_{i,t}$, because of the correlation between the regime variable and the error component $\varepsilon_{i,t}$. If we enter *regime dummies* into our specification, they will be estimated as positive or negative quantities according to the direction of selectivity, although they should be zero according to the basic specification. Furthermore, we know how these estimates for regime constants are distributed under the null of a correct specification. Using a GMM estimator, they are asymptotically normal, with mean zero, and their standard deviation is given by the standard deviation of the coefficient. Therefore, the t-value on these coefficients is a valid test statistic.

It may be argued that this test ignores the possibility that the regime-specific constants truly belong into the equation. Consider a trend in the term in the brackets of equation (1) that makes the target level of $y_{i,t}$ change over time:

$$\Delta y_{i,t} = -(1 - \alpha_{i,t-1})(y_{i,t-1} - \kappa t - \mu_i) + \varepsilon_{i,t}.$$

Solving for $y_{i,t}$, we get:

$$y_{i,t} = \alpha_{i,t-1} y_{i,t-1} + (1 - \alpha_{i,t-1}) \kappa t + (1 - \alpha_{i,t-1}) \mu_i + \varepsilon_{i,t}.$$

After transforming the equation by subtracting an observation belonging into the same regime, lagged λ periods, we have

$$y_{i,t} - y_{i,t-\lambda} = \alpha_{i,t-1} (y_{i,t-1} - y_{i,t-\lambda-1}) + (1 - \alpha_{i,t-1}) \kappa \lambda + (\varepsilon_{i,t} - \varepsilon_{i,t-\lambda}).$$

Regime-specific constants may thus be the result of a trending target variable. However, in this case they should be proportional to each other, with a factor of proportionality given by 1 minus the regime-specific coefficient on the lagged dependent variable. Using the delta method to test this restriction is relatively straightforward. More generally, they should not be of different sign, as it will be the case if the coefficient on the regime dummy collects the residuals selected for their high or low value.

4.3. Moment restrictions for the equation in levels

Both estimation methods discussed above – the two moment conditions for quasi-differences as well as the generalised differences approach – require the regime variable to be predetermined with respect to the current shock term. This may hold in many cases, specifically if there are long planning and gestation lags as in investment decisions. In other circumstances, the error term in the adjustment equation and the threshold variable governing the adjustment regime may be contemporaneously correlated. I will investigate an approach that can be brought to bear in this case. For greater clarity, the adjustment equation shall be rewritten as follows:

$$\Delta y_{i,t} = -(1 - \alpha_{i,t}) (y_{i,t-1} - \mathbf{x}'_{i,t} \boldsymbol{\beta} - \mu_i) + \varepsilon_{i,t}, \quad (21)$$

or
$$y_{i,t} = \alpha_{i,t} y_{i,t-1} + (1 - \alpha_{i,t}) (\mathbf{x}'_{i,t} \boldsymbol{\beta} + \mu_i) + \varepsilon_{i,t}. \quad (22)$$

The dating of the adjustment coefficient has been changed, to highlight the possibility of a contemporaneous correlation between the speed of adjustment and $\varepsilon_{i,t}$.

It turns out that this structure can be accessed by means of *level estimation*, relying on a type of moment condition that was introduced by Arellano and Bover (1995) and Blundell and Bond (1998) as a response to a specific problem arising in the standard autoregressive model. If the coefficient of the lagged dependent variable is in the neighbourhood of one, the level behaves like a random walk and will be a weak instrument in the differenced equation. Under certain conditions, the following moment equation can be used in the estimation of the standard autoregressive model, as stated in Section 2 above:

$$E\left[\Delta y_{i,t-s}(\mu_i + \varepsilon_{i,t})\right] = 0,$$

with $s \geq 1$. If $\varepsilon_{i,t}$ is serially uncorrelated, it is sufficient that $y_{i,t}$ is mean stationary and displays a constant correlation with μ_i for the moment equation to hold. Blundell and Bond (1998) have shown that this implies a requirement on the initial conditions: the deviation of the starting value from the stationary level needs to be uncorrelated with the stationary level itself.

The latent term of equation (22) is given by $(1 - \alpha_{i,t})\mu_i + \varepsilon_{i,t}$. In the attempt to use first differences as instruments for levels, we first look at

$$E\left(\Delta y_{i,t-s} \left((1 - \alpha_{i,t})\mu_i + \varepsilon_{i,t} \right)\right).$$

This expectation will be zero if, first, $E\Delta y_{i,t-s} = 0$, and second, $\Delta y_{i,t-s}$ is uncorrelated with both $(1 - \alpha_{i,t-1})\mu_i$ and $\varepsilon_{i,t}$. The first condition requires the process to be mean stationary, as in the derivation of Blundell/Bond and Arellano/Bover. The second condition is hard to fulfil. To see why, we adjust the backward solution to the modified dating:

$$y_{i,t} = \left[y_{i,0} - \mathbf{x}'_{i,1} \boldsymbol{\beta} - \mu_i \right] \prod_{k=1}^t \alpha_{i,k} + \mathbf{x}'_{i,t} \boldsymbol{\beta} + \mu_i + A_{i,t}.$$

Plugging this back into (21) we obtain:

$$\Delta y_{i,t} = -(1 - \alpha_{i,t}) \left(\left[y_{i,0} - \mathbf{x}'_{i,1} \boldsymbol{\beta} - \mu_i \right] \prod_{k=0}^{t-1} \alpha_{i,k} + A_{i,t-1} - \Delta \mathbf{x}'_{i,t} \boldsymbol{\beta} \right) + \varepsilon_{i,t}. \quad (23)$$

The difference $\Delta y_{i,t-s}$ is a function of all $\varepsilon_{i,k}$, $\Delta \mathbf{x}_{i,k}$ and $\alpha_{i,k}$ and, $k \geq s$, as well as of the initial condition. One of the requirements for the covariance of $\Delta y_{i,t-s}$ and $(1 - \alpha_{i,t})\mu_i$ to disappear is therefore a limited memory of $\alpha_{i,t} = \boldsymbol{\alpha}' \mathbf{r}_{i,t}$ with respect to its own past. This excludes all sorts of fixed effects in $\mathbf{r}_{i,t}$. For the estimation problem at hand, a direct adaptation of the Arellano/Bover and Blundell/Bond strategy therefore does not look very promising.

We can weaken the requirements considerably by decomposing the target level, μ_i , into its expectations over all individuals, μ^e , and the individual-specific deviation μ_i^* . I define:

$$\mu_i = \mu^e + \mu_i^*, \text{ with } \mu^e = E_i \mu_i.$$

The parameter μ^e is the expected value over all individuals i , and μ_i^* is the individual deviation from this expectation. By definition, $E \mu_i^* = 0$. Rewriting the adjustment equation, we arrive at:

$$y_{i,t} = \boldsymbol{\alpha}'(\mathbf{r}_{i,t} y_{i,t-1}) + (\mathbf{1} - \boldsymbol{\alpha}') \mathbf{r}_{i,t} \mathbf{x}_{i,t}' \boldsymbol{\beta} + \mu^e (\mathbf{1} - \boldsymbol{\alpha}') \mathbf{r}_{i,t} + \underbrace{\mu_i^* (\mathbf{1} - \boldsymbol{\alpha}') \mathbf{r}_{i,t}}_{\text{latent term}} + \varepsilon_{i,t}. \quad (24)$$

This equation contains a new, regime-specific shift term $\mu^e (\mathbf{1} - \boldsymbol{\alpha}') \mathbf{r}_{i,t}$. In estimation, this term can be taken into account by introducing the regime vector $\mathbf{r}_{i,t-1}$ as a regressor into the equation. Investigating under what condition $\Delta y_{i,t-s}$ is an instrument for the rewritten equation, we arrive at:

Proposition 3: In order to estimate equation (24), we can make use of the moment restriction

$$E\left(\Delta y_{i,t-s} \left((1 - \alpha_{i,t}) \mu_i^* + \varepsilon_{i,t} \right)\right), \quad s \geq k, \quad (25)$$

under the following two sufficient conditions:

- a) $E\left(\varepsilon_{i,t} \mid \varepsilon_{i,t-k}, \varepsilon_{i,t-k-1}, \dots, \Delta \mathbf{x}_{i,t-k}, \Delta \mathbf{x}_{i,t-k-1}, \dots, \mathbf{r}_{i,t-k}, \mathbf{r}_{i,t-k-1}, \dots, y_{i,0} - \mathbf{x}_{i,1}' \boldsymbol{\beta} - \mu_i\right) = 0$;
- b) $E\left(\mu_i^* \mid \{\varepsilon_{i,t}\}, \{\mathbf{r}_{i,t}\}, \{\Delta \mathbf{x}_{i,t}\}, (y_{i,0} - \mathbf{x}_{i,1}' \boldsymbol{\beta} - \mu_i)\right) = 0$,

where a term in curly brackets, $\{\cdot\}$, denotes an entire time series. In both parts of the condition, the invariance with respect to the initial value can be dispensed with if the process has been running "long enough" for $E(y_{i,t})$ to have converged.

Proof: Moment condition (25) holds if, first,

$$E \Delta y_{i,t-k} \varepsilon_{i,t} = E\left(\Delta y_{i,t-k} \cdot E\left(\varepsilon_{i,t} \mid \Delta y_{i,t-k}\right)\right) = 0, \quad (26)$$

and second,

$$E(\Delta y_{i,t-k} (1 - \alpha_{i,t}) \mu_i^*) = 0. \quad (27)$$

Given the backward solution (23), condition a) is sufficient for the expectation in the bracket of (26) to be identically zero, as $\{\Delta y_{i,t-k}\}$ is a coarser information set than $\{\varepsilon_{i,t-k}, \varepsilon_{i,t-k-1}, \dots, \mathbf{r}_{i,t-k-1}, \mathbf{r}_{i,t-k-2}, \dots, y_{i,0} - \mu_i\}$. Similarly, we can write:

$$E(\Delta y_{i,t-k} (1 - \alpha_{i,t}) \mu_i^*) = E(\Delta y_{i,t-k} (1 - \alpha_{i,t}) \cdot E(\mu_i^* | \Delta y_{i,t-k} (1 - \alpha_{i,t}))).$$

Again, if, as in condition b), the expectation of μ_i^* is zero conditional on all random variables that may enter $\Delta y_{i,t-k}$ according to its reduced form, the expectation in (27) is zero, too. \square

It goes without saying that, if the conditions for its use are met, the moment condition can also be used in the case of a predetermined regime indicator. It is natural that we have to impose conditions on μ_i , now that we leave it in the equation instead of differencing it out. The invariance of expected μ_i with respect to the time path $\{\varepsilon_{i,t}\}$ is quite unproblematic. It accords well with the basic structure of the error component model. The irrelevance of the regime process is less innocuous. It is well conceivable that a real-world data generating process for $\mathbf{r}_{i,t}$ may contain a fixed effect that is correlated with μ_i . This would invalidate the moment equation (25). Similar reservations apply with respect to the required irrelevance of $\{\Delta \mathbf{x}_{i,t}\}$. Lastly, the necessity of having an expected value of μ_i that is independent of the initial deviation, $(y_{i,0} - \mu_i)$ was also found by Blundell and Bond (1998) when investigating the use of moment equations for levels in a linear context. The condition is not innocuous either: it excludes an initial condition such as $y_{i,0} = 0$. As already stated in the proposition, we can replace it by the requirement that the process has been running for a "very long" time, as the first term inside the bracket of equation (23) will disappear asymptotically.

It is interesting to compare the conditions for Propositions 1, 2 and 3. All of them require the expected value of $\varepsilon_{i,t}$ to be invariant with respect to past values

$\varepsilon_{i,t-s}, \varepsilon_{i,t-s-1}, \dots$, the levels or first differences of $\mathbf{x}_{i,t-s}, \mathbf{x}_{i,t-s-1}, \dots$ as well as to $\mu_i + \mathbf{x}_{i,1}'\boldsymbol{\beta}$ and $y_{i,0}$. Propositions 1 and 2 also need $\varepsilon_{i,t}$ to be uncorrelated with $\mathbf{r}_{i,t-1}$, the regime indicator figuring in the current date adjustment equation, whereas for Proposition 3, invariance of $\varepsilon_{i,t}$ with respect to lag s and earlier of the regime indicator is sufficient. As an additional identifying assumption for the generalised differencing approach, we need the memory of $\mathbf{r}_{i,t}$ to be finite with respect to $\varepsilon_{i,t}$. This excludes, for example, an autoregressive equation for the threshold variable driving the regime indicator if the current shocks are correlated. The level estimator, for his part, needs the expected value of the individual effect μ_i to be unrelated to the rest of the process, including the initial deviation $(y_{i,0} - \mu_i)$. Both of these restrictions can be burdensome. But the two linear estimators based on Propositions 2 and 3 are able to fulfil special tasks. The generalised difference estimator will be unbiased even if some of the alpha coefficients are large – in fact it still works if one of them is exactly equal to 1. The level estimator, on the other hand, will discern differential adjustment speeds also if the regime indicator is contemporaneous. In order to better understand the comparative advantages, the next section shows simulation results.

5. Implementing and simulating the estimators

5.1 Setting up the simulation

In the simulation study, the three sets of moment conditions are used separately for estimation. For the regime indicator, I specify a threshold process. The k 'th element of $\mathbf{r}_{i,t}$ is given by

$$\mathbf{r}(k)_{i,t} = \text{Ind}(c_{k-1} \leq s_{i,t} \leq c_k).$$

The numbers c_0, \dots, c_L are thresholds, with the first and the last element being equal to $-\infty$ and ∞ , respectively. As an example for a threshold process with infinite memory with respect to the error term we use an AR(1):

$$s_{i,t} = k s_{i,t-1} + v_{i,t},$$

where the current shock $v_{i,t}$ is contemporaneously correlated with the error term $\varepsilon_{i,t}$. Alternatively, as an example of a process with finite memory, it is assumed that the threshold process be driven by an MA(q):

$$s_{i,t} = a + \sum_{j=0}^q b_j \eta_{i,t-j}, \text{ with } b_0 = 1.$$

The elements of the moving average conform to:

$$E\eta_{i,t} = 0, E\eta_{i,t}\eta_{i,t-k} = 0 \quad \forall k > 0, E\eta_{i,t}\varepsilon_{i,t} \neq 0, E\eta_{i,t}\varepsilon_{i,t-k} = 0 \quad \forall k > 0.$$

Concretely, the two interrelated processes $\{\mathbf{r}_{i,t}, y_{i,t}\}$ are simulated as follows:

Regime-dependent error correction process: $\varepsilon_{i,t}$ is standard normal, μ_i follows an $N(1,1)$ process, $\varepsilon_{i,t}$ and μ_i are independent. As a benchmark I use $\alpha_0 = 0.3$ and $\alpha_1 = 0.8$. Note that the larger of these coefficients is not far from 1.

Regime indicator process: If the threshold process is driven by an AR(1), I set $E v_{i,t}^2 = 1, E v_{i,t} \varepsilon_{i,t} = 0.8$, $v_{i,t}$ being calculated as a weighted sum of $\varepsilon_{i,t}$ and an independent Gaussian process. The AR-parameter k is 0.8. Likewise, for the MA(q), the stochastic structure is chosen as $E \eta_{i,t}^2 = 1, E \eta_{i,t} \varepsilon_{i,t} = 0.8$, with $\eta_{i,t}$ being calculated as a weighted sum of $\varepsilon_{i,t}$ and an independent Gaussian process. The threshold level is set equal to zero, resulting in an equal number of observations in each regime on average. I experiment with a MA(0) (uncorrelated regimes states) and a MA(1) with $b_1 = 0.8$. Note the high contemporaneous correlation between the shocks in the regime equation and the error term.

Panel structure: The panel is unbalanced, with individuals carrying either 8, 9 or 10 observations, 1,000 of each type, that is 3,000 in total. For each individual, the process is simulated for 50 periods, and only the last 8, 9 or 10 observations are used for estimation.

All estimators are implemented by first calculating the transformed observations and the instruments and then adapting and using the routines supplied with the DPD module for

Ox written by Doornik, Arellano and Bond to perform GMM estimates and tests.⁴ Details on the estimation routines follow.

a) Quasi-Difference estimations

I assume an AR(1) as a process driving the threshold variable that constitutes the regime. The estimation equations are transformed in the way described above. The first version of the quasi-differencing approach, QD1, is implemented by estimating the transformed equation using a standard linear GMM estimator and then calculating the structural parameters by inverting equation (13). The more complicated QD2 estimation is performed by treating the moment as a non-linear function of the structural parameters, as in equation (25). Both CMD estimations on the basis of the linear reduced form and direct non-linear GMM estimation of the structural parameters were used. The latter was implemented using the iterative Gauss-Newton method. The results are rather similar. For the conceptual reasons mentioned in the text, the nonlinear GMM procedure is preferred, and only these results are shown. It has to be mentioned though that the CMD procedure is clearly faster than the iterative nonlinear GMM estimation procedure, without being less efficient.

The procedure used for QD2 estimation is explained in some detail in Appendix 2. The Gauss-Newton method iterates on a linearised moment function calculated for preliminary estimates, sequentially improving the estimation. Calculating pseudo-observations for each step, the estimation problem can be solved using routines for the estimation of linear econometric models. As initial values, I use parameter estimates on the basis of the QD1 transformation. CMD estimates on the basis of the same moment condition could also be used. Indeed, they yield better initial values, but were not chosen here because of the conceptual problems with the asymptotics. As instruments, I use levels lagged twice. It turns out that the instruments are more informative (the estimates being more precise) if they are separated out in regimes. That is: For purposes of instrumentation, the lags of $y_{i,t-2}$ are interacted with regime dummies, $\mathbf{r}_{i,t-2}$.

⁴ Ox is an object-oriented matrix programming language. For a complete description of Ox see Doornik (2001).

b) Forward Differences and Generalised Difference estimation

When implementing the difference estimator, we can use the moment conditions in a specific way that greatly facilitates the calculation of moments. Proposition 1 requires us to calculate the λ 'th difference of every observation, with $\lambda \geq q + 2$ and differences being taken using only observations in the same regime. Then we may use levels lagged $\lambda + 1, \lambda + 2, \dots$ as instruments. It seems that this requires us to make the set of instruments for a specific observation dependent on whether or not there are two observations in the same regime within a specific time distance. By taking the *earlier* of the two observations as a point of reference $y_{i,t}$ and assigning to it the nearest *lead* $y_{i,t+\lambda}$ of the same regime with $\lambda \geq 2 + q$, the definition of suitable instruments is straightforward. We can uniformly use lags $y_{i,t-1}, y_{i,t-2}$ and earlier as instruments.

I have experimented with two variants of the differencing approach. The "Forward Difference Estimator" uses a fixed lead of $s = 2 + q$. This transformation preserves the correlation structure. However, it also leads to a heavy loss of observations, as only observations fulfilling $\mathbf{r}_{i,t-1} = \mathbf{r}_{i,t+\lambda-1}$ can be transformed. The "Generalised Difference Estimator" uses the fact that the moment condition for differenced observation presented in Proposition 1 holds for all leads $s \geq 2 + q$. The transformation thus is carried out using the *nearest* lead $s \geq 2 + q$ with $\mathbf{r}_{i,t-1} = \mathbf{r}_{i,t+s-1}$. Due to the larger number of valid observations, this results in much more precise estimations. Only the results for this estimator are shown. As in Quasi-Difference estimation, I interacted the lagged levels $y_{i,t-1}$ with regime indicators $\mathbf{r}_{i,t-1}$. In order to test the validity of the transformation, regime dummies are included as additional RHS variables. They also enter the instrument set.

c) Level estimation

As described above, the level estimator is implemented by specifying an artificial equation that contains the vector $\mathbf{r}_{i,t-1}$ as an additional RHS variable. Instruments are the levels of $\mathbf{r}_{i,t}y_{i,t-1}$ (i.e. two interaction terms) and current regime dummies in those columns where the regime variable is predetermined, and $\mathbf{r}_{i,t-1}y_{i,t-1}$ plus lagged regime dummies where the regime variable is contemporaneous.

5.2 Simulation results

Tables 1 and 2 show estimates on the basis of quasi-difference transformations (1,000 runs). The theoretical discussion has shown that the finite sample properties of the estimators may depend on the size of the regime specific coefficients, notably on their difference from 1. Therefore estimations for a whole range of parameters are shown. The true value for α_1 is set as 0.3, whereas the value for α_2 ranges from 0.3 to 0.9. Larger ranges and finer steps are plotted in Figures 1 and 2.

For the construction of Table 1 and Figure 1, the simpler QD1 transformation was used. Whereas for smaller coefficient values the estimator performs well and yields correct estimates with a good precision, it is less reliable if one of the regime specific coefficients is large. For $\alpha_1 = \alpha_2 = 0.3$, the mean bias is only of the order of -0.0004 for both parameters. It will be 0.0177 for $\hat{\alpha}_2$ when α_2 is raised to 0.7, and for $\alpha_2 = 0.9$ the finite sample bias of $\hat{\alpha}_2$ becomes a non-negligible -0.0415. The estimates $\hat{\alpha}_1$ also deteriorate, although less markedly. The table also gives t-values and Sargan statistics. The bias leads the t-tests for the true value of individual coefficients reject too often when one of the coefficients is too high: In the extreme case of $\alpha_2 = 0.9$, the true value is rejected 77.9% of the times. The same is true for the Sargan test of instrument validity: with large regime specific coefficients, it rejects the instruments too often. We can conclude that slow speeds of adjustment (high persistence) create a problem for QD1 estimation.

Table 2 and Figure 2 give results for the QD2 transformation, moment condition 2. As was expected, the estimator performs better for large values of regime specific coefficients than its counterpart based on QD1. In the extreme case of $\alpha_1 = 0.3$ and $\alpha_2 = 0.9$, the bias is 0.015 and 0.017. In terms of absolute value, this is about half of what resulted from QD1. For smaller values of regime specific coefficients, there is hardly any bias at all. Sargan statistics and t-values remain reliable but for very high values of α_2 .

Figure 1: Mean bias for estimates on the basis of QD1, with $\alpha_1 = 0.3$ and α_2 varying

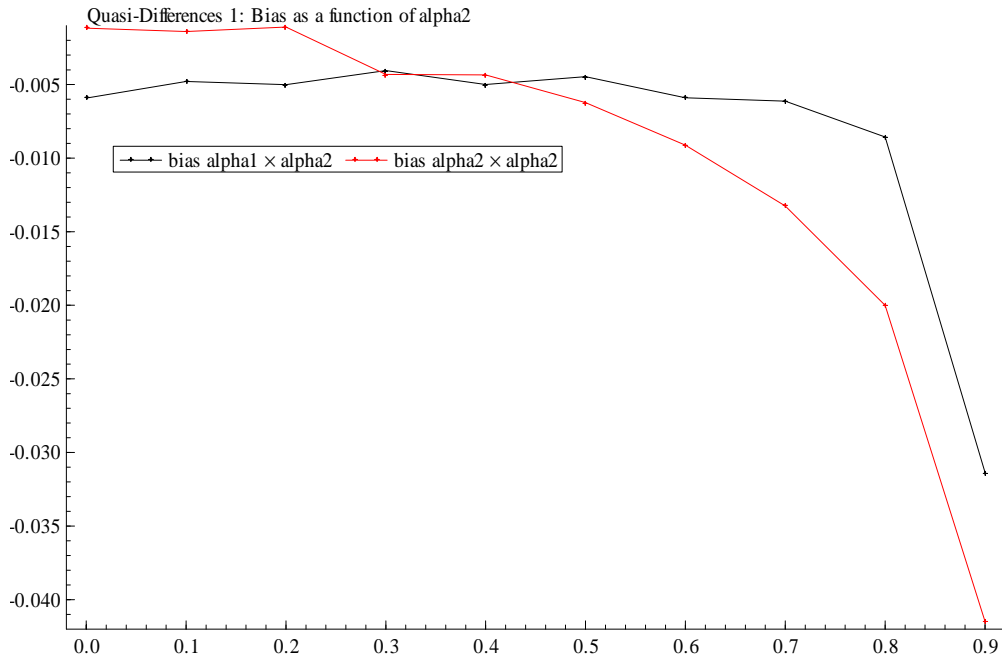


Figure 2: Mean bias for estimates on the basis of QD2, with $\alpha_1 = 0.3$ and α_2 varying

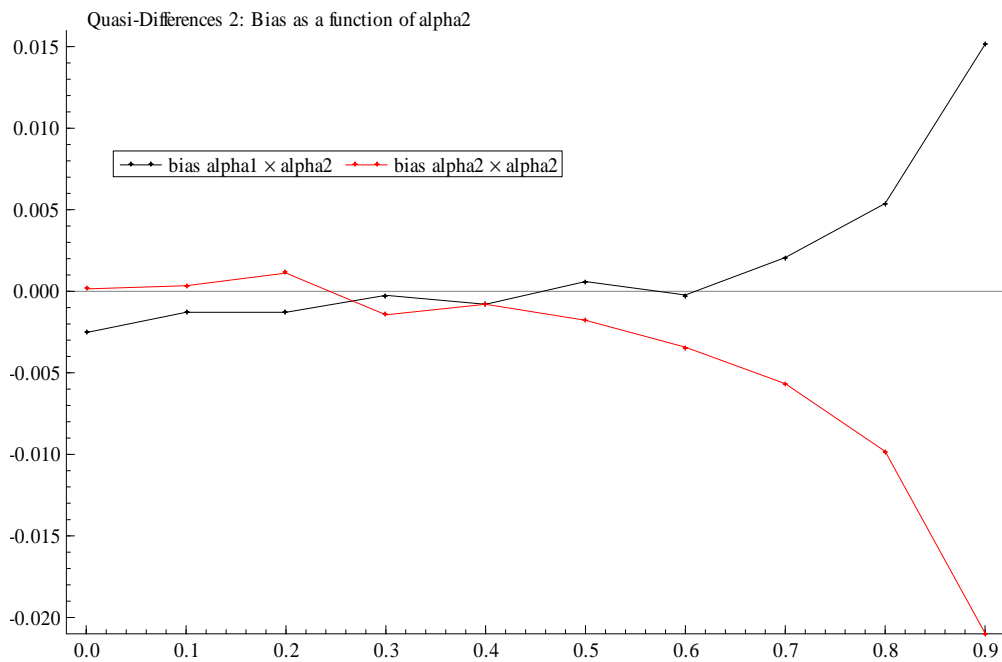


Figure 3: Mean bias for estimates on the basis of QD1, with $\alpha_1 = 0.8$ and α_2 varying

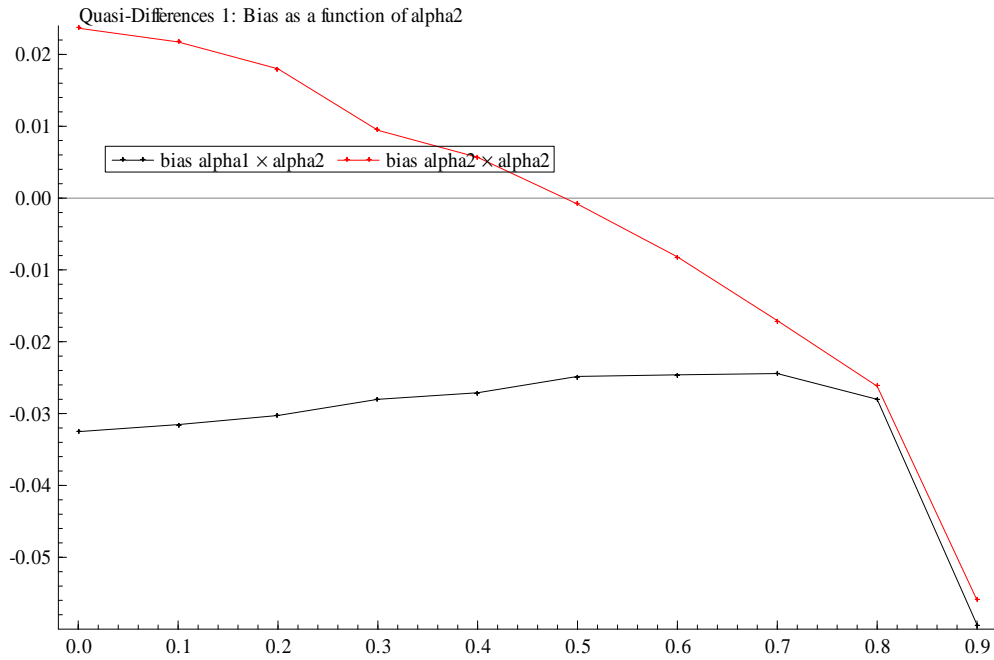
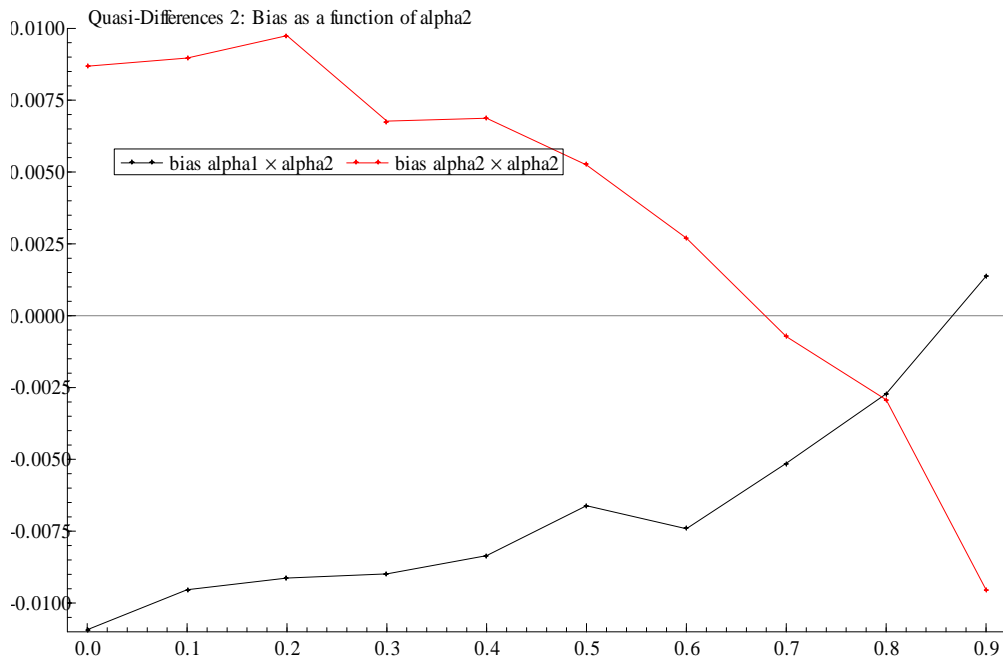


Figure 4: Mean bias for estimates on the basis of QD2, with $\alpha_1 = 0.8$ and α_2 varying



The theoretical discussion showed that the precision of the QD2 estimator should depend on the *difference* between the regime specific coefficients. If both of them are high, but of similar size, the ratio $(1-\alpha_{i,t-2})/(1-\alpha_{i,t-1})$ in the definition of the transformed error term $\xi_{i,t}$ cancels out, see eq. (12). The error term in QD1, in contrast, depends on the absolute distance of the regime specific coefficients from 1. To study this issue, the simulations of QD1 and QD2 estimation are performed using a value of $\alpha_1 = 0.8$ as a platform and varying over α_2 . The result is shown in Figures 3 (QD1 estimation) and 4 (QD2 estimation) Here, the QD1 estimates are biased throughout the range. The bias of $\hat{\alpha}_2$ switches from positive to negative, whereas the bias of $\hat{\alpha}_1$ is negative throughout. In contrast, with QD2 the bias practically disappears when both parameters are large, to be noticeable only when α_1 is small.

Table 3 and Figures 5 and 6 give the results using GMM on observations transformed by Generalised Differences. In Columns 1 and 2 the estimator is correctly used. The memory of the regime process is restricted – Column (1) assumes uncorrelated regimes, and Column (2) assumes a threshold process driven by an MA(1). The minimum leads used in transformation are 2 and 3, respectively. In both cases, the Generalised difference estimator performs well. The estimates are unbiased. The standard deviations are similar to what can be obtained from the quasi-difference estimates for the smaller of the two coefficients and actually somewhat lower for the higher coefficient. In the case of an MA(1) regime process, standard deviations are higher, as less observations can be used. Column (1), with a minimum lead of 2, yields an average of 15.058 valid observations per estimation. This number decreases to 11.277 in Column (2), when a minimum lead of 3 is imposed. On the same set of simulated data, the estimates based on quasi-differencing can use 21.000 observations each run. Figure 5 shows that the bias of the Generalised Difference estimator is very small when the conditions for its use are met and does not depend systematically on the size of the adjustment coefficients. Even regime specific coefficients equal to or larger than 1 can be accommodated, as long as the overall process remains stable. Columns (3) and (4) do "the wrong thing". For Column (3), a minimum lead of 2 is used on data generated with a regime process generated by an MA(1), where a lead of $\lambda \geq 3$ would be warranted. Column (4) assumes an AR(1) process driving the threshold variable: this process has infinite memory. Unex-

pectedly, in both cases the estimator turns out to be biased. However, in spite of a strong correlation between the shock in the regime variable and the error term, the bias is moderate. In Column (3), only the estimates $\hat{\alpha}_2$ are biased, to a degree that is similar to the performance of the QD2 estimator under the same (unfavourable) parameter values. When, as assumed in Column (4), the regime process is driven by a process with infinite memory, the resulting bias is larger, similar in size to the weak performance of the QD1 estimator when one of the coefficients is large. Figure 6 shows how in this latter case the bias depends on the alpha-parameters.

The specification tests do not fail to detect the erroneous use of the estimator. In both cases, the regime constant test rejects the specification in 100% of the cases. As the estimated coefficients are of opposite sign, they cannot be caused by trending target values. The regime dummies have "captured" the regime-specific non-zero expectations of the differenced residuals $E(\varepsilon_{i,t} - \varepsilon_{i,t-2} | \mathbf{r}_{i,t-1} = \mathbf{r}_{i,t-3})$ for the two values that $\mathbf{r}_{i,t-1}$ can take. The Sargan test is sensitive for the misspecification in Column (3) where the wrong lead is used, rejecting 91.9% of the estimates. Detecting an infinite memory of the regime variable is harder for the Sargan-test: only 23.2% of estimates in Column (4) are rejected.

Table 4, together with Figures 7 and 8, show simulation results for the level estimator, both for the case of a predetermined regime and a contemporaneous regime. In both cases, a regime process with infinite memory is assumed. The table and the figures vary α_2 for a fixed value of $\alpha_1 = 0.3$. In the predetermined case, there is little bias for the whole range of parameters, with the possible exception of the $\alpha_2 = 1$, where the value of the bias of $\hat{\alpha}_1$ assumes a moderate 0.01. Standard deviations are similar to what was obtained with the other estimators. If α_2 assumes a value larger than 1, the estimates become extremely exact.

Figure 5: Mean bias for Generalised Differences estimates, with $\alpha_1 = 0.8$ and α_2 varying. Here: regime process uncorrelated over time, correct lead of 2

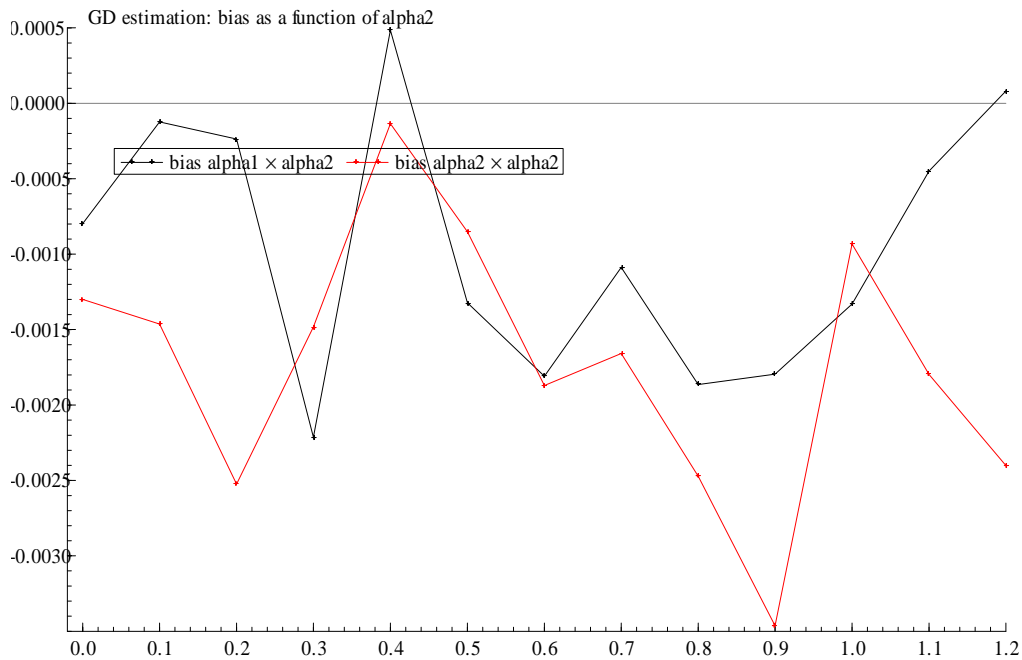


Figure 6: Mean bias for Generalised Differences estimates, with $\alpha_1 = 0.8$ and α_2 varying. Here: regime process unlimited memory AR(1), misspecified lead of 2

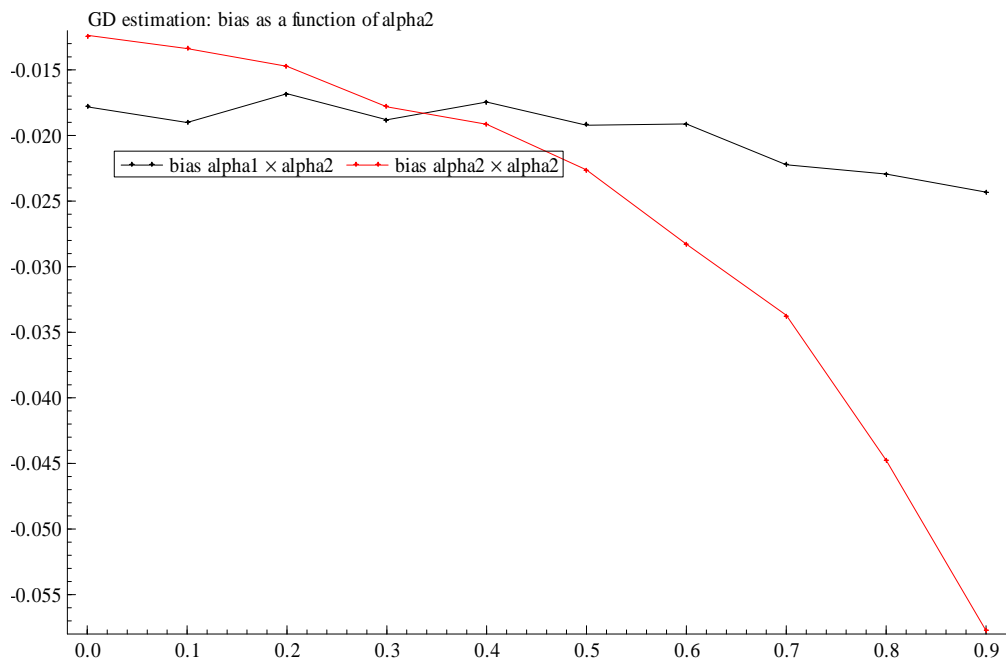


Figure 7: Level estimation with predetermined regimes. Mean bias for estimates on the basis of moment condition 4, with $\alpha_1 = 0.3$ and α_2 varying

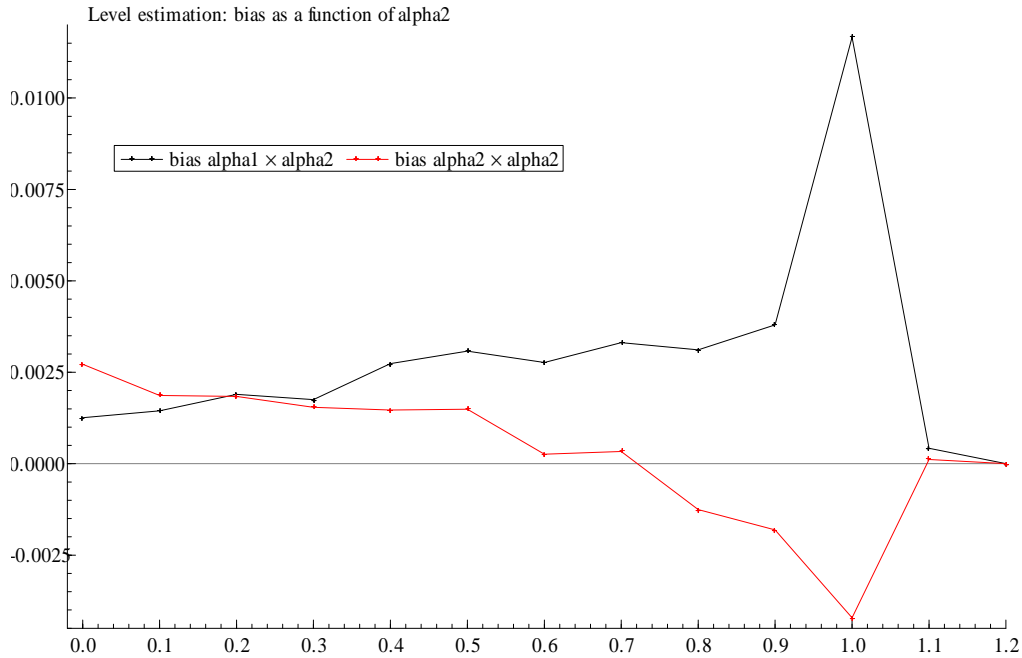
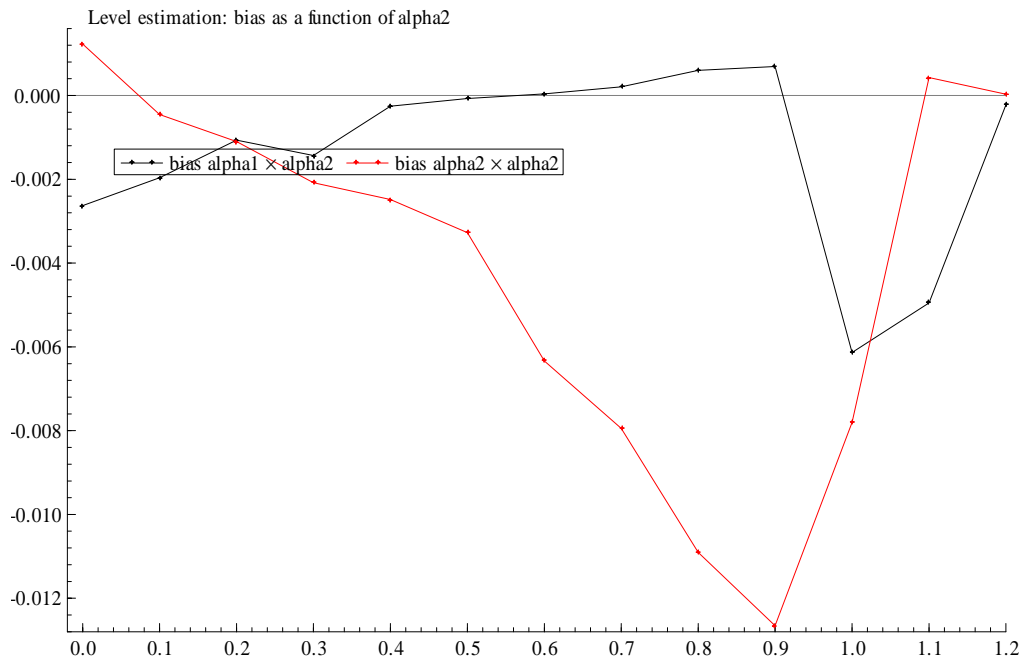


Figure 8: Level estimation with contemporaneous regimes. Mean bias for estimates on the basis of moment condition 4, with $\alpha_1 = 0.3$ and α_2 varying



Columns (3) and (4), as well as Figure 8 show that the level estimator indeed successfully copes with contemporaneous regime variables, a problem that cannot be solved by any of the other approaches. There is a moderate bias that peaks 0.012 for $\hat{\alpha}_2$ when $\alpha_2 = 0.9$, and the standard deviations are higher than with a predetermined regime for $\alpha_2 < 1$. Again, for $\alpha_2 > 1$ the level estimates become very exact. In all columns, the regime dummy that is artificially introduced into the equation is very near the theoretical value of $E(1 - \alpha_{i,t-1})\mu^e$, a term that is introduced into equation (24) by splitting up the firm fixed effect into its expectation and a deviation uncorrelated with the shocks in the other processes.

5.3 Comparing the estimators

I have presented four different ways of estimating an adjustment equation with time-varying persistence, all within a GMM framework, albeit with a different set of moment conditions.

Two estimation techniques rely on transforming the original equation using quasi-differences. Both quasi-differences estimators are very precise when all coefficients are small. When both coefficients are large and of similar size (high persistence throughout the regimes), the results of QD1 estimation have been shown to be unusable in simulation, whereas the QD2 approach continues to deliver correct results. In Chapter 4, the QD2 estimator is successfully employed for estimating differential adjustment speeds for the capital stock. The most difficult parameterisation is when coefficients are widely different, while one of them is large. While not unaffected by small sample problems, the QD2 estimator performs clearly better in this situation. In direct comparison, the major virtue of the QD1 estimator lies in its surprising simplicity, while still being consistent in a wider range of circumstances.

The third method involves transformation using Forward Differences or Generalised Differences, with a lead that is long enough to overcome the memory in the process driving the regime indicator for the $\varepsilon_{i,t}$ -shocks. This method is applicable only when the memory of the regime process is limited. I have shown how to test this requirement. Although a limited memory may be a good approximation in a number of circumstances, such as investment under financing constraints, the requirement will not always

be fulfilled. This method leads to a linear estimator which remains unbiased even if some of the coefficients are in the neighbourhood of 1 or even larger. The fourth method leaves the equation untransformed, and past differences are used as instruments. Regime dummies are employed to capture and neutralise the time-varying non-zero expected value of the residual process. The memory of the regime process is irrelevant for this technique. However, we have to assume the individual-specific deterministic equilibrium to be independent of the shock parameters of the other relevant processes. The level estimator is very precise with regard to larger coefficients. This is not really surprising: the use of level equations has originally been proposed to overcome the problem of weak instruments in cases where the autoregressive parameter approaches 1. More important is another virtue of the fourth method: the level estimator is the sole procedure that can be used when the regime indicator is contemporaneous to the error term in the adjustment equation.

To sum up, the two quasi-differencing methods should be regarded as the standard procedure, with the QD1 method apt for quick specification search and the QD2 transformation leading to efficient results in small samples. QD1 must not be used if one or more of the regime specific autoregressive coefficients are large. A preliminary estimation constraining the coefficients to be equal across regimes may provide a helpful warning – it can be done using the standard methodology devised by Arellano/Bond, Blundell/Bond and Arellano/Bover. The two linear techniques are of great value in cases where the quasi-differencing techniques do not work properly: near unit roots in at least some regimes and – concerning the level estimator – contemporaneous regimes.

Table 1: Quasi-differences, QD1 transformation, 1000 runs

Simulation #	(1)	(2)	(3)	(4)
Specification state variable underlying regimes			AR(1)	
True α_1	0.3	0.3	0.3	0.3
True α_2	0.3	0.5	0.7	0.9
α_1 Mean parameter estimate	0.2930	0.2955	0.2939	0.2687
Mean bias	-0.0041	-0.0045	-0.0061	-0.0313
Mean estimated std. deviation	0.0220	0.0236	0.0276	0.0351
Std. dev. parameter estimate	0.0218	0.0247	0.0298	0.0533
RMSE	0.0222	0.0251	0.0304	0.0618
Freq. rejections of true value on 5% conf. level	4.6%	6.8%	5.9%	25.7%
α_2 Mean parameter estimate	0.2957	0.4938	0.6868	0.8586
Mean bias	-0.0043	-0.0062	-0.0133	-0.0414
Mean estimated std. deviation	0.0194	0.0189	0.0177	0.0139
Std. dev. parameter estimate	0.0197	0.0190	0.0188	0.0203
RMSE	0.0202	0.0200	0.0230	0.0262
Freq. rejections of true value on 5% conf. level	6.0%	5.4%	12.3%	77.9%
Freq. rejection by Sargan- Hansen on 5% conf. level	8.1%	9.4%	16.4%	81.6%
Valid obs. in estimation	21,000	21,000	21,000	21,000

Notes: The table shows estimates of α_1 and α_2 on the basis of moment condition 1. Columns vary by parameters α_1 and α_2 used for generating the panels according to eq. (1). Each column represents 1000 repetitions of two stage GMM estimates using an unbalanced panel of 3000 individuals with 10, 9 and 8 observations (1000 individuals each). The number of valid observations is reduced by the need to transform variables. Instruments are the levels of $\mathbf{r}_{i,t-2}y_{i,t-2}$ (i.e. two interaction terms) and a constant. Estimated standard deviations are derived from reduced form estimates using the delta method. Estimation is executed using DPD package version 1.2 on Ox version 3.30 and additional, user written routines.

Table 2: Quasi-differences, QD2 transformation, 1000 runs

Simulation #	(1)	(2)	(3)	(4)
Specification state variable underlying regimes			AR(1)	
True α_1	0.3	0.3	0.3	0.3
True α_2	0.3	0.5	0.7	0.9
α_1 Mean parameter estimate	0.2998	0.3006	0.3021	0.3152
Mean bias	-0.0002	0.0006	0.0021	0.0152
Mean estimated std. deviation	0.0221	0.0229	0.0261	0.0418
Std. dev. parameter estimate	0.0217	0.0235	0.0270	0.0463
RMSE	0.0217	0.0235	0.0271	0.0487
Freq. rejections of true value on 5% conf. level	4.7%	5.8%	5.8%	9.5%
α_2 Mean parameter estimate	0.2985	0.4982	0.6943	0.8790
Mean bias	-0.0014	-0.0018	-0.0057	-0.0209
Mean estimated std. deviation	0.0195	0.0194	0.0187	0.0174
Std. dev. parameter estimate	0.0195	0.0192	0.0188	0.0170
RMSE	0.0196	0.0193	0.0197	0.0269
Freq. rejections of true value on 5% conf. level	5.9%	4.5%	5.9%	23.0%
Freq. rejection by Sargan- Hansen on 5% conf. level	5.2%	6.0%	6.0%	22.9%
Valid obs. in estimation	21,000	21,000	21,000	21,000

Notes: The table shows estimates of α_1 and α_2 on the basis of moment condition 2. Columns vary by parameters α_1 and α_2 used for generating the panels according to eq. (1). Each column represents 1000 repetitions of a two stage GMM procedure iterating on pseudoregressors, using an unbalanced panel of 3000 individuals with 10, 9 and 8 observations (1000 individuals each). As an initial value, an estimate on the basis of moment condition 1 was used, see the results in Table 2. The number of valid observations is reduced by the need to transform variables. Instruments are the levels of $\mathbf{r}_{i,t-2}y_{i,t-2}$ (i.e. two interaction terms) and a constant. Estimated standard deviations are calculated as a by-product from the final Gauss-Newton iteration step. Estimation is executed using DPD package version 1.2 on Ox version 3.30 and additional, user written routines.

Table 3: Generalised Differences Estimation, $(\alpha_1, \alpha_2) = (0.3, 0.8)$ 1,000 runs

		... using appropriate leads		... using inappropriate leads	
Specification state variable underlying regimes		(1) MA(0)	(2) MA(1)	(3) MA(1)	(4) AR(1)
		lead = 2	lead = 3	lead = 2	lead = 2
α_1	Mean estimate (true value 0.3)	0.2990	0.2994	0.2950	0.2767
	Mean est. std. dev.	0.0215	0.0286	0.0243	0.0223
	Mean bias	-0.0010	-0.0006	-0.0050	-0.0232
	RMSE	0.0118	0.0276	0.0239	0.0322
	Freq. rejections of true value on 5% conf. level	6.4%	3.8%	5.1%	18.0%
α_2	Mean estimate (true value 0.8)	0.7978	0.7995	0.7736	0.7568
	Mean est. std. dev.	0.0261	0.0304	0.0298	0.0276
	Mean bias	-0.0021	-0.0005	-0.0264	-0.0432
	RMSE	0.0113	0.0296	0.0399	0.0518
	Freq. rejections of true value on 5% conf. level	3.7%	4.5%	14.2%	34.6%
Specification tests					
G_1	Mean estimate	-0.0001	0.0000	-0.0765	-0.0977
	Mean est. std. dev.	0.0116	0.0145	0.0114	0.0102
	Freq. rejections of zero value on 5% conf. level	5.8%	5.4%	100%	100%
G_2	Mean estimate	-0.0007	-0.0010	0.0822	0.0977
	Mean est. std. dev.	0.0114	0.0141	0.0114	0.0102
	Freq. rejection of zero value on 5% conf. level	4.9%	4.7%	100%	100%
	Freq. rejection by Sargan-Hansen on 5% conf. level	5.1%	4.8%	91.9%	23.2%
Av. no. of valid observations		15.058	11.277	14.107	15.117

Note: The table shows estimates of α_1 and α_2 on the basis of moment condition 3 (Generalised Difference Estimator). Columns vary by the stochastic specification of the regime indicator and by the lead used for transformation. Columns (1), (2), and (3) specify processes where the memory of the regime variable is limited over time and the state variable that underlies the regime indicator follows an MA process. In column (4), the regime process is supposed to have infinite memory. In all columns, $\alpha_1 = 0.3$ and $\alpha_2 = 0.8$. Each column represents 1000 repetitions of two stage GMM estimates using an unbalanced panel of 3000 individuals with 10, 9 and 8 observations (1000 individuals each). The number of valid observations is reduced by the need to transform variables. Instruments are the levels of $\mathbf{r}_{i,t-1}y_{i,t-1}$ (i.e. two interaction terms) and a constant. Estimation is executed using DPD package version 1.2 on Ox version 3.30 and additional, user written routines.

Table 4: Level Estimation 1000 runs

Simulation #		(1)	(2)	(3)	(4)
Regime indicator		Predetermined		Contemporaneous	
State variable underlying regimes		AR(1)			
True α_1		0.3	0.3	0.3	0.3
True α_2		0.8	1.1	0.8	1.1
α_1	Mean parameter estimate	0.3031	0.3004	0.3006	0.2951
	Mean bias	0.0031	0.0004	0.0006	-0.0049
	Mean estimated std. deviation	0.0197	0.0074	0.0255	0.0073
	Std. dev. parameter estimate	0.0187	0.0078	0.0252	0.0079
	RMSE	0.0190	0.0078	0.0252	0.0094
	Freq. rejections of true value on 5% conf. level	4.3%	4.2%	4.7%	10.5%
	α_2	Mean parameter estimate	0.7987	1.1001	0.7891
Mean bias		-0.0013	0.0001	-0.0109	0.0004
Mean estimated std. deviation		0.0188	0.0013	0.0283	0.0017
Std. dev. parameter estimate		0.0191	0.0013	0.0285	0.0017
RMSE		0.0192	0.0013	0.0305	0.0017
Freq. rejections of true value on 5% conf. level		5.7%	5.2%	7.0%	5.6%
<i>Auxiliary regime constants</i>					
G_1	Mean estimate	0.6985	0.698	0.6795	0.6922
	Theoretically expected	0.7	0.7	0.7	0.7
G_2	Mean estimate	0.2030	-0.0984	0.2434	-0.0852
	Theoretically expected	0.2	-0.1	0.2	-0.1
	Freq. rejection by Sargan-Hansen on 5% conf. level	4.1%	3.0%	5.2%	4.9%
	Valid obs. in estimation	24,000	24,000	24,000	24,000

Note: The table shows estimates of α_1 and α_2 on the basis of moment condition 4 (Level Estimator). Columns vary by parameters α_1 and α_2 used for generating the panels according to eq. (1) and by the stochastic specification of the regime indicator. In all cases, the regime process is supposed to have infinite memory, following an AR(1) process. Columns (1) and (2) relate to processes where the regime variable is predetermined in the adjustment equation, and Columns (3) and (4) results for regime variable that are contemporaneously correlated with the error term. In all columns, $\alpha_1 = 0.3$. Whereas columns (1) and (3) specify $\alpha_2 = 0.8$, columns (2) and (4) show results for $\alpha_2 = 1.1$. Each column represents 1000 repetitions of two stage GMM estimates using an unbalanced panel of 3000 individuals with 10, 9 and 8 observations (1000 individuals each). Instruments are the levels of $\mathbf{r}_{i,t}y_{i,t-1}$ (i.e. two interaction terms) and current regime dummies in those columns where the regime variable is predetermined, and $\mathbf{r}_{i,t-1}y_{i,t-1}$ plus lagged regime dummies where the regime variable is contemporaneous. Estimation is executed using DPD package version 1.2 on Ox version 3.30 and additional, user written routines.

Appendix A: A state dependent error correction model

Formally, the state dependent adjustment equation considered in this paper involves a lagged dependent variable and a forcing term $\mathbf{x}_{i,t}$. But also higher order adjustment processes can be accommodated, by redefining states appropriately.

Consider a linear autoregressive process with distributed lags in a forcing term $\mathbf{x}_{i,t}$ and an individual specific constant μ_i :

$$A(L)y_{i,t} = B(L)\mathbf{x}_{i,t} + \mu_i + \varepsilon_{i,t},$$

where $A(L)$ and $B(L)$ are lag polynomials. As is well known, the process can always be written in the error correction format. If, for example, $A(L)$ and $B(L)$ are of order 2, this leads to

$$\begin{aligned} \Delta y_{i,t} = & -\phi \left(y_{i,t-1} - \boldsymbol{\beta}' \mathbf{x}_{i,t-1} - \mu_i^* \right) \\ & + \boldsymbol{\gamma}^0' \Delta \mathbf{x}_{i,t} + \boldsymbol{\gamma}^1' \Delta \mathbf{x}_{i,t-1} + \omega \Delta y_{i,t-1} + \varepsilon_{i,t}. \end{aligned}$$

In the first line, the term in brackets is the deviation from static equilibrium, where $\boldsymbol{\beta}$ may be interpreted as a cumulative long run effect of a shock in $\mathbf{x}_{i,t}$. The transformed constant μ_i^* is equal to $[A(L)]^{-1} \mu_i$. The term ϕ is the speed of adjustment. If the process is stable, then $|\phi| < 1$. The second line depicts the transitional dynamics, which is not directly related to the deviation from equilibrium. With $A(L)$ or $B(L)$ of higher order than 2, the transitional dynamics in the error correction format would involve higher order lags of differences $\Delta \mathbf{x}_{i,t}$ and $\Delta y_{i,t}$.

A straightforward generalisation of the adjustment process considered hitherto makes ϕ , $\boldsymbol{\gamma}^0$, $\boldsymbol{\gamma}^1$, and ω state dependent, while leaving the transformed constant μ_i^* and the long run effect $\boldsymbol{\beta}$ time invariant. The latter imposes a constraint on the time varying coefficients. For simplicity, I consider all adjustment coefficients as predetermined:

$$\Delta y_{i,t} = -\phi_{i,t-1} \left(y_{i,t-1} - \boldsymbol{\beta}' \mathbf{x}_{i,t-1} - \mu_i^* \right) + \boldsymbol{\gamma}_{i,t-1}^0' \Delta \mathbf{x}_{i,t} + \boldsymbol{\gamma}_{i,t-1}^1' \Delta \mathbf{x}_{i,t-1} + \omega_{i,t-1} \Delta y_{i,t-1} + \varepsilon_{i,t}. \quad (\text{A1})$$

Now let again $\mathbf{r}_{i,t}$ be an indicator variable for the state of adjustment. As the adjustment process is parameterised over two lags, it is straightforward to model the time varying parameters as a function involving the state variables in two periods, $t-1$ and $t-2$.

Finally, let $\mathbf{d}_{i,t-1}$ be an indicator vector of dummies for all the possible values $(\mathbf{r}_{i,t-1}, \mathbf{r}_{i,t-2})$ can take. Then we can write:

$$\phi_{i,t-1} = \boldsymbol{\varphi}' \mathbf{d}_{i,t-1}, \quad \omega_{i,t-1} = \boldsymbol{\omega}' \mathbf{d}_{i,t-1}, \quad \gamma_{i,t-1}^0 = \boldsymbol{\Gamma}^0 \mathbf{d}_{i,t-1}, \quad \gamma_{i,t-1}^1 = \boldsymbol{\Gamma}^1 \mathbf{d}_{i,t-1},$$

with $\boldsymbol{\varphi}$, $\boldsymbol{\omega}$, $\boldsymbol{\Gamma}^0$ and $\boldsymbol{\Gamma}^1$ vectors and matrices of state dependent adjustment coefficients to be estimated. Written this way, the problem is fully equivalent to the one I have treated in this paper, with $\mathbf{d}_{i,t-1}$ taking the place of $\mathbf{r}_{i,t-1}$ with respect to the adjustment speed, $\phi_{i,t-1}$, and using appropriate interaction terms for all the other state dependent coefficients. With the help of quasi-differencing or generalised differencing, we can eliminate the fixed effect from equation (A1). With contemporaneous adjustment coefficients, we may use the level estimator. It has to be noted though that – compared to a first order adjustment process – the generalised difference estimator will be difficult to use, as there are L^2 states to be considered here, and only pairs of observations belonging to the same regime with a given minimum time distance can be used. The other two estimation principles are not affected by this profusion of states, except for the fact that the number of coefficients is higher.

Appendix B: Nonlinear GMM estimation using the Gauss-Newton Method

The Gauss-Newton method has been developed for Nonlinear Least Squares problems. Its use in GMM estimation is much less frequent and shall therefore be exposed. See Davidson and McKinnon (1993) on the use of Gauss-Newton in Nonlinear Least Squares and Instrumental Variables Estimation, Hayashi (2000) on GMM estimation, and Judge et al. (1985) on numerical methods in maximisation.

A GMM estimator $\hat{\theta}$ maximises an objective function $Q_n(\theta)$, given as

$$Q_n(\theta) = -\frac{1}{2} \mathbf{g}_n(\theta)' \hat{\mathbf{W}} \mathbf{g}_n(\theta).$$

$(1 \times K)$ $(K \times K)$ $(1 \times K)$

The function $\mathbf{g}_n(\theta)$ represents an empirical moment, calculated for some P -dimensional parameter vector θ . $\hat{\mathbf{W}}$ is a possibly data-dependent matrix weighting the K moments. I will assume the specific case of a generalised nonlinear instrumental variables estimation problem⁵ where the moment function $\mathbf{g}_n(\theta)$ can be written as a product of the vector of instruments and an error term:

$$\mathbf{g}_n(\theta) = \frac{1}{n} \sum_i^n \mathbf{z}_i (y_i - f(\mathbf{x}_i, \theta)).$$

$(1 \times K)$ n \sum_i^n $(K \times 1)$

Here, y_i is a scalar, \mathbf{x}_i is a vector of Q explanatory variables and \mathbf{z}_i is a vector of instruments. The double indexation is dropped, and i characterises an observation, not an individual. The first order condition for maximising the objective function is

$$\frac{\partial}{\partial \theta} Q_n(\hat{\theta}) = -\mathbf{G}_n(\hat{\theta})' \hat{\mathbf{W}} \mathbf{g}_n(\hat{\theta}) = \mathbf{0},$$

$(P \times 1)$ $(P \times K)$ $(K \times K)$ $(K \times 1)$

where

$$\mathbf{G}_n(\hat{\theta}) = \frac{\partial}{\partial \theta} \mathbf{g}_n(\hat{\theta}) = -\frac{1}{n} \sum_i^n \left(\mathbf{z}_i \cdot \frac{\partial}{\partial \theta} f(\mathbf{x}_i, \hat{\theta}) \right).$$

$(P \times K)$ $\frac{\partial}{\partial \theta}$ $(K \times 1)$ $(1 \times P)$

⁵ This is the most important case and the only one of relevance here. Actually, with the exception of the pseudo-data iteration technique, the following does not depend on this specific structure of the moment function.

is the Jacobian matrix of $\mathbf{g}_n(\boldsymbol{\theta})$, the derivative of the vector of moments with respect to the parameters. If the equation is nonlinear in variables only, as in the case of the quasi-differencing approach, we have:

$$f(\mathbf{x}_i, \boldsymbol{\theta}) = \mathbf{x}_i' \mathbf{h}(\boldsymbol{\theta}),$$

$(1 \times Q) \quad (Q \times K)$

with $\mathbf{h}(\cdot)$ a vector-valued function, and thus:

$$\mathbf{g}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_i^n (\mathbf{z}_i y_i - \mathbf{z}_i \mathbf{x}_i' \mathbf{h}(\boldsymbol{\theta}))$$

$(1 \times K)$

and

$$\mathbf{G}_n(\hat{\boldsymbol{\theta}}) = \frac{\partial}{\partial \boldsymbol{\theta}'} \mathbf{g}_n(\hat{\boldsymbol{\theta}}) = -\frac{1}{n} \sum_i^n \left(\mathbf{z}_i \cdot \mathbf{x}_i' \frac{\partial}{\partial \boldsymbol{\theta}'} \mathbf{h}(\hat{\boldsymbol{\theta}}) \right).$$

$(P \times K) \quad (K \times 1) \quad (1 \times Q) \quad (Q \times P)$

A Gauss-Newton estimation step minimises the objective function $Q_n(\boldsymbol{\theta})$ with the function $\mathbf{g}_n(\boldsymbol{\theta})$ replaced by a linearised version. This is the core of the iterative optimisation procedure, but the Gauss-Newton estimation is also useful for generating test statistics. The first order Taylor-expansion of $\mathbf{g}_n(\boldsymbol{\theta})$ around some preliminary estimator $\hat{\boldsymbol{\theta}}_j$ is

$$\begin{aligned} \mathbf{g}_n(\boldsymbol{\theta}) &\cong \mathbf{g}_n(\hat{\boldsymbol{\theta}}_j) + \mathbf{G}_n(\hat{\boldsymbol{\theta}}_j)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_j) = \left(\mathbf{g}_n(\hat{\boldsymbol{\theta}}_j) - \mathbf{G}_n(\hat{\boldsymbol{\theta}}_j)\hat{\boldsymbol{\theta}}_j \right) + \mathbf{G}_n(\hat{\boldsymbol{\theta}}_j)\boldsymbol{\theta} \\ &= \tilde{\mathbf{g}}_n(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}_j). \end{aligned}$$

$(K \times P) \quad (P \times 1)$

The gradient of this linearised moment function with expansion point $\hat{\boldsymbol{\theta}}_j$ is a matrix constant:

$$\tilde{\mathbf{G}}_n(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}_j) = \frac{\partial}{\partial \boldsymbol{\theta}'} \tilde{\mathbf{g}}_n(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}_j) = \mathbf{G}_n(\hat{\boldsymbol{\theta}}_j).$$

Replacing $\mathbf{g}_n(\boldsymbol{\theta})$ in the original objective function by $\tilde{\mathbf{g}}_n(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}_j)$ renders a quadratic function. The first order conditions imply a linear GMM estimator:

$$\begin{aligned} \tilde{\mathbf{G}}_n(\boldsymbol{\theta}^* | \hat{\boldsymbol{\theta}}_j)' \hat{\mathbf{W}} \tilde{\mathbf{g}}_n(\boldsymbol{\theta}^* | \hat{\boldsymbol{\theta}}_j) &= \mathbf{G}_n(\hat{\boldsymbol{\theta}}_j)' \hat{\mathbf{W}} \left[\left(\mathbf{g}_n(\hat{\boldsymbol{\theta}}_j) - \mathbf{G}_n(\hat{\boldsymbol{\theta}}_j)\hat{\boldsymbol{\theta}}_j \right) + \mathbf{G}_n(\hat{\boldsymbol{\theta}}_j)\boldsymbol{\theta}^* \right] = \mathbf{0} \\ \boldsymbol{\theta}^* &= \left[\mathbf{G}_n(\hat{\boldsymbol{\theta}}_j)' \hat{\mathbf{W}} \mathbf{G}_n(\hat{\boldsymbol{\theta}}_j) \right]^{-1} \mathbf{G}_n(\hat{\boldsymbol{\theta}}_j)' \hat{\mathbf{W}} \left(\mathbf{g}_n(\hat{\boldsymbol{\theta}}_j) + \mathbf{G}_n(\hat{\boldsymbol{\theta}}_j)\hat{\boldsymbol{\theta}}_j \right). \end{aligned}$$

Here, $\boldsymbol{\theta}^*$ denotes the solution of the modified problem. All elements on the right hand side of this equation are evaluated at the expansion point $\hat{\boldsymbol{\theta}}_j$. Gauss-Newton optimisation iterates on a sequence of these linearised estimation problems, with the updating equation:

$$\hat{\boldsymbol{\theta}}_{j+1} = \hat{\boldsymbol{\theta}}_j + s(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_j).$$

The step length s may be chosen less than 1 in order to ensure convergence in cases where the objective function is flat in the neighbourhood of the solution.

To perform the Gauss-Newton iteration, we may define pseudo-observations:

$$\begin{aligned} \mathbf{x}_i^* &= \frac{\partial}{\partial \boldsymbol{\theta}} f(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_j), \text{ and} \\ y_i^* &= y_i - f(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_j) + \frac{\partial}{\partial \boldsymbol{\theta}} f(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_j) \hat{\boldsymbol{\theta}}_j = y_i - f(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_j) + \mathbf{x}_i^* \hat{\boldsymbol{\theta}}_j \end{aligned}$$

to obtain:

$$\tilde{\mathbf{g}}_n(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}_j) = \frac{1}{n} \sum_i^n \mathbf{z}_i (y_i^* - \mathbf{x}_i^* \boldsymbol{\theta})$$

$$\text{and } \tilde{\mathbf{G}}_n(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}_j) = -\frac{1}{n} \sum_i^n \mathbf{z}_i \mathbf{x}_i^*.$$

This is the format of linear GMM estimation. The first order conditions lead to

$$\boldsymbol{\theta}^* = \left[\left(\sum \mathbf{z}_i \mathbf{x}_i^* \right)' \hat{\mathbf{W}} \left(\sum \mathbf{z}_i \mathbf{x}_i^* \right) \right]^{-1} \left(\sum \mathbf{z}_i \mathbf{x}_i^* \right)' \hat{\mathbf{W}} \left(\sum \mathbf{z}_i y_i^* \right),$$

which is the standard linear GMM estimator when applied to the pseudo-observations. This is identical to the procedure in nonlinear least squares estimation.

The solution of the non-linear estimation problem, $\hat{\boldsymbol{\theta}}$, is a fixed point in the Gauss-Newton iterations. Evaluated at the expansion point, i.e. with $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_j$, the moment function $\mathbf{g}_n(\boldsymbol{\theta})$ and its gradient $\mathbf{G}_n(\boldsymbol{\theta})$ are equal to their respective linearised counterparts $\tilde{\mathbf{g}}_n(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}_j)$ and $\tilde{\mathbf{G}}_n(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}_j)$. Therefore, given that $\hat{\boldsymbol{\theta}}$ satisfies the first order condition for the nonlinear problem, it will also fulfil the first order conditions for the corresponding linearised problem, if $\hat{\boldsymbol{\theta}}$ is chosen at the expansion point. The residuals of the Gauss-

Newton estimates are identical to the residuals of the original nonlinear problem at the point of convergence. For $\hat{\boldsymbol{\theta}}_j = \hat{\boldsymbol{\theta}}$, the Gauss-Newton residuals are

$$y_i^* - \mathbf{x}_i^* \hat{\boldsymbol{\theta}} = y_i - f(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) + \frac{\partial}{\partial \boldsymbol{\theta}} f(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) \hat{\boldsymbol{\theta}} - \frac{\partial}{\partial \boldsymbol{\theta}} f(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) \hat{\boldsymbol{\theta}} = y_i - f(\mathbf{x}_i, \hat{\boldsymbol{\theta}}).$$

Similarly, the covariance matrix of $\hat{\boldsymbol{\theta}}$ can be computed as the covariance matrix of the Gauss-Newton estimation at the point of convergence. This follows directly from comparing the asymptotic covariance of nonlinear GMM estimation with its linearised counterparts. In fact, the asymptotic covariance is computed using the very same linearisation of $\mathbf{g}_n(\boldsymbol{\theta})$ that also defines the Gauss-Newton regression above, see Hayashi (2000), Section 7.3.

The Gauss-Newton procedure is a gradient method. We can write:

$$\boldsymbol{\theta}^* = \hat{\boldsymbol{\theta}}_j + \left[\mathbf{G}_n(\hat{\boldsymbol{\theta}})' \hat{\mathbf{W}} \mathbf{G}_n(\hat{\boldsymbol{\theta}}) \right]^{-1} \left[-\mathbf{G}_n(\hat{\boldsymbol{\theta}})' \hat{\mathbf{W}} \mathbf{g}_n(\hat{\boldsymbol{\theta}}_j) \right].$$

The second expression in brackets is the gradient of the objective function $Q_n(\boldsymbol{\theta})$, evaluated at $\hat{\boldsymbol{\theta}}_j$. It is multiplied by the inverse of a quadratic form in the gradient of the moment function. This latter expression takes the role of the negative inverted Hessian in the Newton-Raphson algorithm. This matrix will be positive definite in the neighbourhood of $\boldsymbol{\theta}_0$, provided that $E(\mathbf{G}_n(\boldsymbol{\theta}_0))$ has full column rank and the number of observations is large. Thus, if s is chosen small enough, the value of the objective function will increase each iteration.