University of Leicester

DEPARTMENT OF ECONOMICS

A Nonlinear Panel Data Model of Cross-Sectional Dependence

James Mitchell, University of Leicester, UK
George Kapetanios, University of London, UK
Yongcheol Shin, University of York, UK

Working Paper No. 12/01
January, 2012

A Nonlinear Panel Data Model of Cross-Sectional Dependence*

George Kapetanios
Queen Mary, University of London

James Mitchell
University of Leicester

Yongcheol Shin
University of York

January 11, 2012

Abstract

This paper proposes a nonlinear panel data model which can generate endogenously both 'weak' and 'strong' cross-sectional dependence. The model's distinguishing characteristic is that a given agent's behaviour is influenced by an aggregation of the views or actions of those around them. The model allows for considerable flexibility in terms of the genesis of this herding or clustering type behaviour. At an econometric level, the model is shown to nest various extant dynamic panel data models. These include panel AR models, spatial models, which accommodate weak dependence only, and panel models where cross-sectional averages or factors exogenously generate strong, but not weak, cross sectional dependence. An important implication is that the appropriate model for the aggregate series becomes intrinsically nonlinear, due to the clustering behaviour, and thus requires the disaggregates to be simultaneously considered with the aggregate. We provide the associated asymptotic theory for estimation and inference. This is supplemented with Monte Carlo studies and two empirical applications which indicate the utility of our proposed model as both a structural and reduced form vehicle to model different types of cross-sectional dependence, including evolving clusters.

JEL Classification: C31, C33, C51, E31, G14.

Keywords: Nonlinear Panel Data Model, Clustering, Cross-section Dependence, Factor Models, Monte Carlo Simulations, Application to Stock Returns and Inflation Expectations

*We would like to thank seminar participants at Universities of City, Essex, London, Piraeus, Queen Mary, Seokang, Yonsei, the Bank of Korea, the Melbourne Institute, conference delegates at the Econometric Society World Congress 2010 at Shanghai International Convention Center during 16-22 August, and especially Charlie Cai, Jinseo Cho, In Choi, Matthew Greenwood-Nimmo, Jinwook Jeong, Taehwan Kim, Kevin Lee, Myunghwan Seo, Seungju Song, James Stock for helpful comments and suggestions. Yongcheol Shin gratefully acknowledges partial financial support from the ESRC (grant No. RES-000-22-3161). The usual disclaimer applies.

1 Introduction

In many theoretical models economic agents learn from each other. Whether in herding models, where agents are assumed fully rational but have incomplete information sets (e.g., Banerjee (1992) and Bikhchandani, Hirshleifer, and Welch (1992)),[1] or in adaptive models where agents learn or form their expectations based on recent experience (see, e.g., Timmermann (1994) and Chevillon, Massmann, and Mavroeidis (2010)), agents are affected by past outcomes or the views of groups of other agents. Carroll (2003), for example, sets out a model whereby agents update their views probabilistically by looking at media reports, as opposed to forming full-information rational expectations. Similarly, ideas from cognitive psychology might be used to explain the contagion of views which leads to herd or imitating behaviour (see, e.g., Jegadeesh and Kim (2010)). See Akerlof and Shiller (2009) for a popular textbook discussion. More generally, it is widely observed that all kinds of economic unit (firms, consumers or countries, say) are influenced by their peers, and other economic units, in a wide variety of ways. Therefore, the models, econometric or otherwise, used to model the variables that measure aspects of the behaviour of economic units, need to take into account these influences.

In this paper, motivated by these considerations, we develop a general econometric modelling framework, that allows cross-sectional dependence, of many forms, among large numbers of economic variables, in the form of panels, to arise endogenously. In contrast, popular factor models, that are used for similar modelling purposes, view cross-sectional dependence as an exogenous feature of the data. The proposal, discussion and econometric analysis of the proposed class of models, which is shown to nest many extant models as a special case, forms the main aim of this paper.

The models proposed in this paper are nonlinear panel data models. The distinguishing characteristic of this class of models is the use of unit-specific aggregates/summaries of past values of variables relating to other units that are 'close' in some sense to a given unit, to model that unit. The nature of the models is dynamic, in the sense that the past values of aggregates determine the present. It is instructive at this point to present a generic form for the model given by

$$x_{i,t} = \sum_{j=1}^{N} w_{ij}\left(x_{-i,t-1}, x_{i,t-1}; \gamma\right) x_{j,t-1} + \epsilon_{i,t}, \quad i = 1, ...N, \quad t = 1, ...T, \tag{1}$$

where $x_{-i,t} = (x_{1,t}, x_{2,t}, ..., x_{i-1,t}, x_{i+1,t}, ..., x_{Nt})'$ and $\sum_{j=1}^{N} w_{ij}\left(x_{-i,t-1}, x_{i,t-1}; \gamma\right) = 1$. This form of the model is extremely general and simply signifies that $x_{i,t}$ depends, possibly in a

[1] Information-driven herding can sometimes be classified as "clustering" to differentiate it from herding due to extraneous incentive structures (e.g., Trueman (1994) and Hirshleifer and Teoh (2003)).

nonlinear fashion depending on how w_{ij} is parameterised, on weighted averages of past values of $x_t = (x_{1,t}, ..., x_{Nt})'$, where the weights depend on x_{t-1}. We split x_{t-1} into $x_{-i,t-1}$ and $x_{i,t-1}$ to emphasise the potentially special role of the own lag of $x_{i,t}$ in the specification. One particular motivation for the above model is, in a sense, structural and follows from the claim that it mimics structural interactions between economic units. Another, more econometric, justification simply notes that this model can accommodate generic forms of cross-sectional dependence, including evolving clusters.

The model in (1) is extremely general as it encompasses a wide variety of nonlinear specifications. We consider a number of particular nonlinear specifications for the construction of the unit specific aggregates. We place particular emphasis on specifications where the weights depend on x_{t-1} only through distances of the form $|x_{j,t-1} - x_{i,t-1}|$. We choose a particular specification of this type that is easy to analyse, based on a threshold mechanism, to illustrate the class of models we focus on. This model nests a variety of dynamic panel data models, such as panel data AR models and panel models where cross-sectional averages are used to pick up cross-sectional dependence (e.g., see Pesaran (2006)). Interestingly, it is also closely related to factor models, that have received considerable attention recently following work by Bai and Ng (2002), Stock and Watson (2002) and Bai (2003).

Our models provide an intuitive means by which many forms of cross-sectional dependence can arise in a large panel dataset comprised of variables of a 'similar' nature that relate to different agents/units. These variables might be the disaggregates underlying often studied macroeconomic or financial aggregates, such as economy-wide inflation or the S&P500 index. In particular, the model allows these different economic units to cluster; and for these clusters (including their number) to evolve over time. Such clustering, while of independent interest when interest rests with understanding the behaviour of the individual units or perhaps forecasting them, also has implications when modelling and forecasting the aggregate of these units. In particular, even if concerned only with modelling and forecasting the aggregate, the nonlinearity means that the appropriate aggregate model should not be specified only in terms of aggregated variables; the disaggregate or individual units should be considered simultaneously too.

The degree of cross-sectional dependence can vary, from a case where it is similar to standard factor models, for which the largest eigenvalue of the variance covariance matrix of the data tends to infinity at a rate N, to the case of very weak or no factor structure where this eigenvalue is bounded as $N \to \infty$. Of course, all intermediate cases can arise as well. In this sense, our work is closely related to the work of Chudik and Pesaran (2010) and Chudik, Pesaran, and Tosetti (2009). These papers discuss the concepts of weak and strong cross-sectional dependence based on the characteristics of the variance-covariance matrix of the data and are dynamic in nature, being instances of large dimensional VAR models. Our

work can be viewed as a particular instance of a large dimensional VAR, but for the fact that our model is intrinsically nonlinear in nature.

Our work has precedents in the system engineering literature. However, all the work in that literature relates to simple deterministic models whose limit behaviour is a fixed point that represents clustering. A discussion of the asymptotic behaviour of the deterministic version of our basic model can be found in Blondel, Hendrickx, and Tsitsiklis (2009), following Krause (1997). Another literature that is closely related to our work is the 'similarity' literature as exemplified by Gilboa, Lieberman, and Schmeidler (2006); and references therein. This work relates to univariate processes. It suggests that forecasts for y_t, at time T, can be based on a model which places heavier weights on those past observations of y_t, for which a given vector of variables, x_t, is close to x_T with respect to some metric. In other words, observations y_t, $t \leq T$, for which $||x_t - x_T||$ is small, for some metric $||.||$, have a larger weight for constructing forecasts of y_{T+1} at time T. Gilboa, Lieberman, and Schmeidler (2006) provide powerful theoretical economic justifications for this approach. Our work can be thought of as an extension of this analysis to a multi-agent panel framework, where similarity between agents takes the place of similarity between circumstances.

We provide a comprehensive analysis of the stochastic version of the model; and allow for both threshold and smooth transition type nonlinearities. Our model constitutes, to the best of our knowledge, the first attempt to introduce endogenous cross-sectional dependence and correlation into a panel modelling framework. From an econometric point of view, we establish a number of properties of this new model. First, the basic model (introduced in (2) below) displays the strong form of cross-sectional dependence common to factor models. But, surprisingly, the cross-sectional average model, obtained as a special case of the basic model, exhibits a weaker form of cross-section dependence; this contrasts the apparently similar cross-sectional average augmentation scheme employed by Pesaran (2006). Interestingly, we can also extend the basic model so that it resembles spatial AR or MA models, where dependence is again weak. Secondly, we establish the limiting estimation theory for the model. When the threshold mechanism is used to select which group of units affect a given unit, we use a grid search to estimate consistently both slope and threshold parameters; but only the slope estimator follows the normal distribution asymptotically. The asymptotic distribution of the threshold parameter is non-standard and complex, as in Chan (1993) and Hansen (1999). To overcome this complexity, we follow Gonzalo and Wolf (2005) and undertake inference about the threshold parameter using robust subsampling-based methods. These are proven to be valid for our proposed panel threshold model. When smooth, rather than threshold, transitions are considered we establish that both slope and transition parameters asymptotically follow normal distributions. Finally, and importantly, in the presence of unobserved effects commonly employed in (dynamic) panel data models, we show that the 'Nickel' bias (Nickell (1981)),

familiar to the traditional within-group estimator of a panel data AR model where T (the number of time periods) is fixed, does not arise in our model specifications. This obviates the need for less efficient GMM estimators, which rely on taking first-differences.

Monte Carlo studies confirm that the proposed estimators are reliable, even in samples with small T. We also provide two empirical applications. The first models a panel dataset of inflationary expectations from the Survey of Professional Forecasters in the U.S.; and sheds light on how expectations are formed, as well as casting doubt on the validity of traditional means of extracting a 'consensus' forecast from a panel dataset of individual forecasts. The second application estimates and then forecasts individual stock returns from the S&P500 aggregate index, at a weekly frequency, finding that the proposed nonlinear model offers superior fit relative to benchmark linear autoregressive models, which are well known to be tough to beat when examining stock returns. Both applications, therefore, indicate the utility of our proposed model as both a structural and reduced form vehicle to model cross sectional dependence.

The structure of the paper is as follows: Section 2 presents the basic specification of the model and discusses in detail its theoretical properties. Section 3 presents a number of extensions and discusses their properties. Section 4 discusses the issue of how to test for the presence of nonlinearity in the data. Section 5 presents extensive Monte Carlo simulation evidence. Section 6 provides two empirical illustrations for analysing nonlinearity and cross-section dependence of stock returns and inflation expectations, which demonstrate the utility of our proposed models. Section 7 concludes. All proofs are relegated to an Appendix.

2 The Theoretical Model

In the introduction, see (1), we proposed a general model, which can be given a behavioural interpretation, based on the familiar idea that agents consider the views or behaviour of those around them and aggregate them in some way in order to decide on their own expectations or behaviour. This interaction or mimicking may be explicit, in the sense that agents know what the other agents experienced or expect; or it could be implicit, in the sense that groups of agents happen to behave similarly, even though they do not interact formally. This might be because they are subject to the same environment and/or have similar loss functions and information sets when forming expectations. Alternatively, (1) can and will be motivated as a mechanism for capturing cross-sectional dependence in a reduced form, econometric sense.

So to formalise more clearly the motivating ideas, we propose a particular dynamic panel model for a multitude of agents. Let $x_{i,t}$ denote the value of the variable of interest, such as the agent's income or the agent's view of the future value of some macroeconomic variable, at time t, for agent i. We assume a sample of T observations for each of N agents. Then, we

specify that

$$x_{i,t} = \frac{\rho}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) x_{j,t-1} + \epsilon_{i,t}, \quad t = 2, ..., T, \quad i = 1, ..., N, \tag{2}$$

where

$$m_{i,t} = \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right),$$

$\{\epsilon_{i,t}\}_{t=1}^{T}$ is an error process whose properties will be further discussed below, $\mathcal{I}(.)$ is the indicator function and $-1 < \rho < 1$. Verbally, the above model states that $x_{i,t}$ is influenced by the cross-sectional average of a selection of past x_j and in particular that the relevant x_j are those that lie closest to $x_{i,t-1}$. This formalises the intuitive idea that people are affected more by those with whom they share common views or behaviour. The model may be equally viewed as a descriptive model of agents' behaviour, reflecting the fact that 'similar' agents are affected by 'similar' effects, or as a structural model of agents' views whereby agents use the past views of other agents, similar to them in some respect, to form their own views. The interaction term in (2) may then be thought to capture the (cross-sectional) local average or common component of their views. This idea of commonality has various clear, motivating, concrete examples in a variety of social science disciplines, such as psychology and politics. In economics and finance, the herding could be rational (imitative herding: see Devenow and Welch (1996)) or irrational.

A deterministic form of the above model has been analysed previously in the mathematical and system engineering literature. In particular, Blondel, Hendrickx, and Tsitsiklis (2009) have analysed a continuous form of the restricted version of (2) given by

$$x_{i,t} = \frac{1}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq 1\right) x_{j,t-1}, \quad t = 2, ..., T, \quad i = 1, ..., N, \tag{3}$$

where $m_{i,t} = \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq 1\right)$. To the best of our knowledge, we are the first both to introduce a stochastic term to this type of model and to allow for an unknown value of the threshold parameter.

(2) bears considerable resemblance to threshold autoregressive (TAR) models analysed in the time-series literature. However, unlike straightforward extensions of TAR models to a panel setting, whereby individual units/agents would not interact through the nonlinear specification, the nonlinearity in (2) is inherently cross-sectional in nature; this provides for the development of a dynamic network effect. In deterministic contexts this has been shown to generate interesting behaviour, such as clustering.

Before concluding the introduction to the first of the particular instances of the generic model (1) that we analyse, it is worth addressing a point of statistical importance. Both (1)

and (2) are modelling instances of conditional means of $\left\{\{x_{i,t}\}_{i=1}^{\infty}\right\}_{t=1}^{\infty}$. But $\left\{\{x_{i,t}\}_{i=1}^{\infty}\right\}_{t=1}^{\infty}$ is a two-dimensional random field. Random fields are multidimensional extensions of stochastic processes that are indexed by vectors of the form (i,t), rather than scalars. Such fields are considerably more complex to analyse than simple stochastic processes. While the formal analysis of random fields has not been considered in panel data analysis, we note that concepts such as the existence and uniqueness of random fields that follow particular conditional distributions have been the topic of an extensive literature in statistics. For now, it is sufficient to note the work of Durlauf (1992), who discusses in detail issues arising in the analysis of random fields. Further, by Theorem 1 of Dobruschin (1968), it is obvious that there exist random fields that follow (1) and therefore all particular instances of (1) analysed in the rest of the paper. However, proving uniqueness of random fields that satisfy (1) is much more complicated, as noted in Dobruschin (1968), and is beyond the scope of this paper.

2.1 Clustering

To appreciate more concretely the dynamic behaviour that can be captured by the model, (2), we report some graphical results. We start by showing the dynamic behaviour of the deterministic model (i.e., setting $\epsilon_{i,t} = 0$). In particular we set $N = 100$, $T = 20$. We set the initial conditions to $x_{i,0} \sim N(0, 25)$ and report the evolution of the system for $\rho = 1$, $r = 0.5$ and $r = 3$, in Figure 1. As we see, the system settles quickly to a steady state with a number of clusters. The number of clusters declines with the size of the threshold parameter, as one would intuitively expect. Obviously, for a large value of r, only one cluster will arise.

Of course, the dynamic behaviour of the stochastic model is expected to be quite different. To explore this, we simulate realisations from the stochastic system. We set $N = 100$, $T = 500$, with the initial conditions set as before. For the remaining parameters, we set $r = 0.5$, $\rho = 0.999$, and $\epsilon_{i,t} \sim N(0, 0.1)$. As we shall discuss below, the model is stationary when $|\rho| < 1$. But nonstationarity is of interest, too, and has been explored extensively in the factor model literature. The most interesting behaviour of the model can be obtained when ρ is high enough for the model to be quite persistent. We report two sets of realisation from this model in Figure 2. The first realisation shows emerging cluster structures in the first 100 observations. Then, there are clearly two clusters that persist throughout the rest of the sample. A number of units are outlying and do not join any cluster for the whole sample. The second realisation has one dominant cluster. There is a second cluster which starts at the beginning of the sample and fizzles out by observation 250. At that point a new cluster emerges and by the end of the sample becomes as dominant as the original major cluster.

Clearly the model (2) can model flexibly all sorts of clustering behaviour. It is tempting to attempt to characterise the behaviour of the model as a function of the parameters; it is clear

that for persistent ρ, the interplay of r and the variance of $\epsilon_{i,t}$ is crucial. For instance, a small variance for $\epsilon_{i,t}$ relative to r implies that units do not escape clusters easily. Similarly, *ceteris paribus*, a larger r leads to fewer clusters and dynamically to faster consolidation towards clusters. This needs to be tempered with the finding, discussed in detail later, that when the value of r tends to infinity the model has a smaller degree of cross-sectional dependence. So, overall, it seems that the model can behave in distinct ways depending sensitively on all its parameters, including higher moments of $\epsilon_{i,t}$, as we discuss below.

Next, we allow for fat tails in the distribution of $\epsilon_{i,t}$. We set $\epsilon_{i,t} \sim t_3$, and subsequently normalise $\epsilon_{i,t}$ to have variance equal to 0.1. We report a realisation of this model in Figure 3. Here, it is clear that more clusters arise. There is cluster consolidation but at the same time cluster bifurcation (see the cluster made up of units with high values that bifurcates around observation 400 only to re-emerge as a single cluster by the end of the sample). Overall, it is clear that the new model can generate complex dynamic behaviour across units.

2.2 Special cases

It is interesting to note the nature of restricted versions of the above model, obtained by taking extreme values of the threshold parameter. By setting $r = 0$, we obtain a simple panel autoregressive model of the form

$$x_{i,t} = \rho x_{i,t-1} + \epsilon_{i,t} \tag{4}$$

On the other hand letting $r \to \infty$, we obtain the model

$$x_{i,t} = \frac{\rho}{N} \sum_{j=1}^{N} x_{j,t-1} + \epsilon_{i,t} \tag{5}$$

where past cross-sectional averages of opinions inform, in similar fashions, current opinions. Recently, the use of such cross-sectional averages has been advocated by Pesaran (2006), Chudik and Pesaran (2010) and Chudik, Pesaran, and Tosetti (2009) as a means of modelling cross-sectional dependence in the form of unobserved factors. However, unlike these models where the use of cross-sectional averages is an approximation to the unknown model, in our case this is a limiting case of a structural nonlinear model.

A graphical comparison of these restricted versions of the nonlinear model is also instructive. In Figure 4, we report comparable realisations to those in Figure 1; but setting $r = 0$ in the upper panel and $r = \infty$ in the lower panel. These are, of course, just single realisations; but repeated realisations suggest a very similar picture. While the upper panel depicts independent and very persistent series evolving with little regard to other series in the panel, the lower panel depicts a closely linked set of series behaving similarly. It is interesting to note that this similarity, reminiscent of factor structures, can be proven to arise only for finite N

when $|\rho| < 1$, as we will discuss in more detail below. Neither of these pictures compares in terms of complexity and flexibility to the realisations of the nonlinear model seen in Figures 2-3. It is clear that neither of these two restricted versions of the model can accommodate clustering or evolving herding behaviour.

It is important to investigate the statistical properties of our model. A number of results, stated and proved in the appendix, provide help in this respect. Intuitively, as we show in Lemma 1, (2) is geometrically ergodic, and therefore asymptotically stationary, if $|\rho| < 1$. This allows for the analysis of estimators along traditional lines, as discussed below.

2.3 Cross-sectional dependence and factor models

It is of interest to examine the cross-sectional dependence properties of the model. This is slightly complicated by the need to define cross-sectional dependence in our context. We choose to follow an approach which is used in the analysis of factor models. In the factor literature, the behaviour of the covariance matrix of $x_t = (x_{1,t}, ..., x_{N,t})'$ is considered. Factor models have the property that both the maximum eigenvalue and the row/column sum norm of the covariance matrix tend to infinity at rate N, as $N \to \infty$. In contrast, for other models of cross-sectional dependence such as, for example, spatial AR or MA models, these quantities are bounded, implying that they exhibit much lower degrees of cross-sectional dependence than factor models.[2] It is useful to see where our model fits in this nomenclature. Lemma 4 shows that the column sum norm of the variance covariance matrix of x_t when x_t follows (2) is $O(N)$. Thus, the model is much more similar to factor models than spatial AR or MA models. Interestingly, as we will see in the next section that discusses extensions to the basic model (2), there are versions of (2) that resemble spatial models, more than factor models. Another interesting finding is that (5) implies a variance covariance matrix for x_t with a column sum norm that is $O(1)$. This is surprising, given the similarity that cross-sectional average schemes have with factor models as detailed in Pesaran (2006). However, this result and the analysis of Pesaran (2006) are not directly comparable. Pesaran (2006) assumes the prior existence of factors and uses cross-sectional averages to approximate them. These pre-existing exogenous factors generate high cross-sectional dependence and herding. In our case no exogenous factors exist and the cross-sectional average is a primitive term that exists in the structure of the model. Our surprising result is proven in Lemma 3.[3]

[2] A useful discussion of the various concepts of cross-sectional dependence can be found in Chudik and Pesaran (2010)

[3] Further interesting interactions arise if we let $\rho = 1$. This unit root behaviour counteracts the tendency of the cross-sectional average to disappear asymptotically as $N \to \infty$. The behaviour of both the variances and the covariances of x_t, depends on the limit of $\frac{T}{N}$, as both N and $T \to \infty$. For example, as long as $\frac{T}{N}$ remains bounded, so do the variances of x_t, despite the unit root structure of the model. We feel that a detailed investigation of this issue is beyond the scope of the present paper.

Given the above, it is of interest to examine the analogy with factor models in more detail. We do this by simulating data using (2) and the parametrisation used to construct the realisations in Figure 2. Using the simulated dataset we then extract factor estimates using principal components. We extract 8 principal components and subsequently examine the proportion of the variance of the dataset explained by these principal components. Our previous pictorial analysis suggests that factor like behaviour emerges in the form of clusters of series moving together. The first column of Table 1 presents the average cumulative proportion of the dataset variance explained by successive principal components, over 100 replications. As we can see there is behaviour reminiscent of factor analysis. The first factor explains about 40% of the total dataset variance, rising to about 77% when all 8 factors are considered.

For comparability, we also consider simulations from the same model but setting $r = \infty$. Results are reported in the second column of Table 1. As we see, while the first factor explains roughly the same proportion of the variance in the two parametrisations, the rest of the factors explain little further. This is reasonable. In this case there is only one cluster arising around the cross-sectional mean. As we noted above, there is a crucial difference between (2) and (5). This relates to the fact that while the column sum norm of x_t for (2) is $O(N)$, it is $O(1)$ for (5). This result is asymptotic with respect to N and as noted in footnote 3, the distinction can be difficult to discern for values of ρ close to 1. As a result, we consider a further simulation along the same lines but setting higher values for N ($N = 100, 200, 400, 800, 1000$ and 1500) and a lower value for ρ ($\rho = 0.8$). Results on the average cumulative proportion of the dataset variance explained by successive principal components, over 100 replications, are reported in Tables 2 and 3. It is clear that data from (2) are more cross-sectionally dependent than data from (5). More pertinently, while it is clear that as N increases principal components can explain a decreasing proportion of the data variance for (5), the proportion remains constant for (2).

It is important to restate here differences between our model and a factor model. When a dataset has pronounced cross-sectional dependence exhibited by, say, exploding eigenvalues or the column sum norm associated with its covariance matrix, then a factor model should offer some fit, irrespective of the structural form giving rise to this cross-sectional dependence. Principal components, in particular, nonparametrically construct linear combinations of the variables that capture (strong) cross-sectional dependence, whatever its genesis. But when the data generating process resembles our structural model, such that clusters emerge endogenously and their number varies over time, a large number of factors may be required; and the number needed may also have to change over time. Factor models are intrinsically reduced form; they focus on modelling cross-sectional dependence using an exogenously given number of unobserved factors. Since our model nests (5), it is not surprising that it can approximate a factor model when $r \to \infty$; cf. Pesaran (2006). On the other hand, our model has a

clear parametric structure, such that the slope parameters can be given a structural/economic interpretation; this is a feature shared by some classes of dynamic spatial model; see, e.g., Korniotis (2010). But our models are more general than spatial models, in the sense that the weighting schemes are estimated endogenously, rather than assumed *ex ante*. Furthermore, it is worth noting that the factor model cannot accommodate the weak cross-sectional dependence seen in spatial models, in contrast to the extensions of our nonlinear model described in Section 3 below. These extensions demonstrate that the nonlinear model can, in general, be seen to lie between the two extremes characterised by weakly cross-sectionally dependent spatial models and strongly cross-sectionally dependent factor models.

2.4 Estimation

In this section we explore estimation of the nonlinear model in (2). We consider the standard estimation procedure for a threshold model, whereby a grid of values for r is constructed. Then for all values on that grid the model is estimated by least squares to obtain estimates of the autoregressive parameter, ρ. More specifically, denoting $\tilde{x}_{i,t} = \frac{1}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) x_{j,t-1}$, $\tilde{x}_i = (\tilde{x}_{i,1}, ..., \tilde{x}_{i,T-1})'$, $\tilde{x} = (\tilde{x}_1', ..., \tilde{x}_N')'$, $x_i = (x_{i,2}, ..., x_T)'$ and $x = (x_1', ..., x_N')'$, x is regressed on $\tilde{x}$ using OLS to give an estimate for ρ, for a given value of r in the grid. The value of r that minimises the sum of squared residuals, $\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\epsilon}_{i,t}^2(\rho, r)$, where

$$\hat{\epsilon}_{i,t}(\rho, r) = x_{i,t} - \frac{\rho}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) x_{j,t-1}$$

is the estimator of r. We denote the least squares estimator of (ρ, r) by $(\hat{\rho}, \hat{r})$. We make the following assumption about the error term, $\epsilon_{i,t}$.

Assumption 1 *$\epsilon_{i,t}$ is i.i.d. across t and independent across i. $E(\epsilon_{i,t}^2) = \sigma_{\epsilon_i}^2$. $E(\epsilon_{i,t}^4) < \infty$. For all i, the density of $\epsilon_{i,t}$ is bounded and positive over all compact subsets of $\mathbb{R}$.*

Then, we have the following theorems:

Theorem 1 *Let Assumption 1 hold for $\epsilon_{i,t}$ in (2). Then, as long as $|\rho| < 1$, the least squares estimator of (ρ, r) is consistent as $N, T \to \infty$.*

Theorem 2 *Let Assumption 1 hold for $\epsilon_{i,t}$ in (2). Let (ρ^0, r^0) denote the true value of (ρ, r). Then, as long as $|\rho| < 1$, $NT(\hat{r} - r^0) = O_p(1)$. Further, as long as $|\rho| < 1$, $(NT)^{1/2}(\hat{\rho} - \rho^0)$ has the same asymptotic distribution as if r^0 was known.*

These theorems are intuitive, as they accord with the work and theoretical analysis of Chan (1993) who was the first to analyse, theoretically, the estimator for the univariate

threshold autoregressive model. There exist a number of possible theoretical extensions of this estimation problem. One obvious one relates to the fact that the asymptotic distribution of $NT(\hat{r} - r^0)$ is non-normal and depends on unknown parameters, as discussed in Chan (1993). The work of Hansen (2000) is of great use here, since by assuming that the model asymptotically is linear, a tractable distributional theory can be obtained for $\hat{r}$. We feel that it is perhaps more appropriate to allow for the nonlinearity to persist asymptotically and, therefore, we do not pursue further this interesting avenue of research.

2.5 Unbalanced panels

The model, (1) can be adjusted to allow for unbalanced panels. In this case (2) takes the form

$$x_{i,t} = \rho \tilde{x}_{i,t}^{up} + \epsilon_{i,t}, \quad t = 2, ..., T, \quad i = 1, ..., N_t, \tag{6}$$

as long as both $x_{i,t}$ and $\tilde{x}_{i,t}^{up}$ are observable, where N_t is the number of observable pairs, $\left(x_{i,t}, \tilde{x}_{i,t}^{up}\right)$, at time t. The definition of $\tilde{x}_{i,t}^{up}$ depends on the application at hand. An obvious definition is

$$\tilde{x}_{i,t}^{up} = \frac{\rho}{m_{i,t}} \sum_{j=1}^{N_{t-1}} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) x_{j,t-1} \tag{7}$$

where $m_{i,t} = \sum_{j=1}^{N_{t-1}} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right)$ and $(x_{i,t}, x_{i,t-1})$ is observable.

Alternative specifications can be used to increase the number of available observations. For example, if $x_{i,t-1}$ is not observed, the latest available observation for the i-th unit prior to time t could be used. More specifically, letting $s_{i,t}$ denote the latest time period, prior to t, in which x is observable for unit i, we can define $\tilde{x}_{i,t}^{up}$ as either

$$\tilde{x}_{i,t}^{up} = \frac{\rho}{m_{i,t}} \sum_{j=1}^{N_{t-1}} \mathcal{I}\left(\left|x_{i,s_{i,t}} - x_{j,t-1}\right| \leq r\right) x_{j,t-1} \tag{8}$$

where $m_{i,t} = \sum_{j=1}^{N_{t-1}} \mathcal{I}\left(\left|x_{i,s_{i,t}} - x_{j,t-1}\right| \leq r\right)$ or

$$\tilde{x}_{i,t}^{up} = \frac{\rho}{m_{i,t}} \sum_{j=1}^{N_{s_{i,t}}} \mathcal{I}\left(\left|x_{i,s_{i,t}} - x_{j,s_{i,t}}\right| \leq r\right) x_{j,s_{i,t}} \tag{9}$$

where $m_{i,t} = \sum_{j=1}^{N_{s_{i,t}}} \mathcal{I}\left(\left|x_{i,s_{i,t}} - x_{j,s_{i,t}}\right| \leq r\right)$, respectively. The specifications in (8) and (9) allow for a larger set of available observations to be used than in (7). Estimation of this model can then be carried out similarly to the case where the number of cross-sectional units is fixed over time. In this case, the effective number of observations is equal to the number of observable pairs of $\left(x_{i,t}, \tilde{x}_{i,t}^{up}\right)$ over i and t, rather than NT, and the statements of Theorems 1 and 2 need to be amended accordingly.

Model (2) can be extended in a large variety of ways. We explore a number of extensions in the next section.

3 Extensions

The model given in (2), while interesting from the perspective of analysing cross-sectional dependence or studying phenomena, such as herding, in an empirical context is quite restrictive in a number of senses. This section therefore provides some extensions. Given that our benchmark model is a panel model it is natural to include constant terms. The basic model then becomes

$$x_{i,t} = \nu_i + \frac{\rho}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) x_{j,t-1} + \epsilon_{i,t} \tag{10}$$

where $\nu_i \sim i.i.d.(0, \sigma_\nu)$. Of course, more general versions of the above model can be accommodated, such as

$$x_{i,t} = \nu_i \zeta_t + \frac{\rho}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) x_{j,t-1} + \epsilon_{i,t} \tag{11}$$

for an $r \times 1$ vector of observable variables, ζ_t.

We now examine the properties of the least squares estimator for (10). As is well known, the presence of ν_i induces endogeneity in standard panel AR models, leading to biased estimation of the autoregressive parameter for finite T, when standard panel least squares estimators, such as the within group estimator, are used. It is easiest to see the problem for standard AR models, and its relation to our model, by noting that the endogeneity arises because unbiasedness, for least squares estimators, requires that

$$E\left(x_{i,t-1}\left(\epsilon_{i,t} - \frac{1}{T}\sum_{t=1}^{T}\epsilon_{i,t}\right)\right) = 0 \tag{12}$$

Obviously, the expectation in (12) is not zero but $O\left(\frac{1}{T}\right)$. One would expect a similar problem to arise for (10). However, surprisingly, this is not the case. As is shown in Lemma 9 in the Appendix

$$E\left(\left(\frac{1}{m_{i,t}}\sum_{j=1}^{N}\mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) x_{j,t-1}\right)\left(\epsilon_{i,t} - \frac{1}{T}\sum_{t=1}^{T}\epsilon_{i,t}\right)\right) = O\left(\frac{1}{NT}\right) \tag{13}$$

which implies that Theorems 1 and 2 hold for (10). As a result the standard within group estimator can be used for (10), thus removing the need for less efficient GMM estimation as is usually the case.

It is straightforward to allow for higher order, say p, lags in (2), such that

$$x_{i,t} = \sum_{s=1}^{p}\left[\frac{\rho_s}{m_{i,t,s}}\sum_{j=1}^{N}\mathcal{I}\left(|x_{i,t-s} - x_{j,t-s}| \leq r\right) x_{j,t-s}\right] + \epsilon_{i,t} \tag{14}$$

where $m_{i,t,s} = \sum_{j=1}^{N} I\left(|x_{i,t-s} - x_{j,t-s}| \leq r\right)$. Alternatively, and more importantly, we can introduce multiple, say q, 'regimes', such that

$$x_{i,t} = \sum_{s=1}^{q} \left[\frac{\rho_s}{m_{i,t,s}} \sum_{j=1}^{N} \mathcal{I}\left(r_s \leq |x_{i,t-1} - x_{j,t-1}| < r_{s+1}\right) x_{j,t-s} \right] + \epsilon_{i,t} \tag{15}$$

where $m_{i,t,s} = \sum_{j=1}^{N} I\left(r_s \leq |x_{i,t-1} - x_{j,t-1}| < r_{s+1}\right)$. Both (14) and (15) can be estimated similarly to (2). However, sufficient conditions for their geometric ergodicity are different to those for (2), and are given in Lemmas 10 and 11, respectively. Suppose that $q = 1$ in (15), then we have two regimes:

$$x_{i,t} = \rho_1 \tilde{x}_{i,t-1} + \rho_2 x^c_{i,t-1} + \epsilon_{i,t}, \tag{16}$$

where $\tilde{x}_{i,t-1} = \frac{1}{m_{i,t}} \sum_{j=1}^{N} I\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) x_{j,t-1}$, $x^c_{i,t-1} = \frac{1}{N - m_{i,t}} \sum_{j=1}^{N} I\left(|x_{i,t-1} - x_{j,t-1}| > r\right) x_{j,t-1}$ are the cross-section averages associated with the group of neighbours and non-neighbours, respectively. This model may be more relevant when modelling heterogeneous interactions, since it is more general than (2), where the restriction, $\rho_2 = 0$, is imposed in (2).

Furthermore, another important issue is how best to modify the basic model to decompose the slope parameter, ρ, into the own effect and a neighbour effect. One obvious candidate is to consider the following extension:[4]

$$x_{i,t} = \rho_0 x_{i,t-1} + \rho_1 x^*_{i,t-1} + \epsilon_{i,t} \tag{17}$$

where $x^*_{i,t-1} = \frac{1}{m_{i,t}-1} \sum_{j=1, j\neq i}^{N} I\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) x_{j,t-1}$, and more generally

$$x_{i,t} = \rho_0 x_{i,t-1} + \rho_1 x^*_{i,t-1} + \rho_2 x^c_{i,t-1} + \epsilon_{i,t} \tag{18}$$

Notice that the model, (17), is similar to the time-space recursive model considered in Korniotis (2010) for investigating the issue of internal versus external consumption habit formation

$$x_{i,t} = \rho_0 x_{i,t-1} + \rho_1 \sum_{j=1, j\neq i}^{N} w_{ij} x_{j,t-1} + \epsilon_{i,t}, \tag{19}$$

where ρ_0 captures the time-series dependence in x_{it} and ρ_1 captures time-space autoregressive dependence. The crucial difference between our model, (17), and the time-space recursive model, (19), is that the selection mechanism for the distance is endogenous in our model; whilst the spatial weights, w_{ij}, in (19) are given exogenously in essentially an *ad hoc* manner.

More generally, we can allow the individual weights to be inversely proportional to the distance, $|x_{i,t-1} - x_{i,t-1}|$, in which case we consider the following extension

$$x_{i,t} = \frac{\rho}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(d_{ij} \leq r\right) w_{ij} x_{j,t-1} + \epsilon_{i,t} \tag{20}$$

[4]Another potentially interesting approach is to modify the approach of Sias (2004) to an analysis of herding; and find an efficient way to decompose $\rho = \rho_{own} + \rho_{neighbour}$.

where the weights are given by

$$w_{ij} = \frac{d_{ij}^{-2}}{\sum_{j=1}^{N} d_{ij}^{-2}},\ d_{ij} = |x_{i,t-1} - x_{j,t-1}| \text{ with } w_{ii} = 1. \tag{21}$$

The estimation of (20) can be conducted practically in two steps. First, the consistent estimate of r is obtained from (2); then construct the weights using (21) and estimate the model, (20), by least squares.

Up until now we have considered only threshold mechanisms for constructing the unit-specific cross-sectional averages. But, as we discussed in the introduction, the class of models we wish to propose is much more general. In particular, we next envisage models of the form

$$x_{i,t} = \rho \sum_{j=1}^{N} \frac{w(|x_{i,t-1} - x_{j,t-1}|\,;\gamma) x_{j,t-1}}{\sum_{j=1}^{N} w(|x_{i,t-1} - x_{j,t-1}|\,;\gamma)} + \epsilon_{i,t} \tag{22}$$

where $w(x;\gamma)$ is a positive twice differentiable integrable function such as, e.g., the exponential function $\exp(-\gamma x^2)$ or the normal cdf, $\Phi(x)$. By now, the properties of this model should be reasonably clear. Lemma 6 shows that the model is geometrically ergodic if $|\rho| < 1$ and similarly to model (2), the column sum norm of the covariance matrix of x_t, when x_t follows (22) is $O(N)$, as shown in Lemma 7. The model in its simple form given by (22) can be estimated by nonlinear least squares; and we have the following Theorem concerning the asymptotic properties of this estimator.

Theorem 3 *Let Assumption 1 hold for $\epsilon_{i,t}$ in (22). Then, as long as $|\rho| < 1$, the nonlinear least squares estimator of (ρ, γ) is $(NT)^{1/2}$-consistent and asymptotically normal as $N, T \to \infty$.*

Similarly to Lemma 9, it can also be shown that

$$E\left(\left(\sum_{j=1}^{N} \frac{w(|x_{i,t-1} - x_{j,t-1}|\,;\gamma) x_{j,t-1}}{\sum_{j=1}^{N} w(|x_{i,t-1} - x_{j,t-1}|\,;\gamma)}\right)\left(\epsilon_{i,t} - \frac{1}{T}\sum_{t=1}^{T} \epsilon_{i,t}\right)\right) = O\left(\frac{1}{NT}\right) \tag{23}$$

which implies that a 'within' estimator is valid for estimating (22), when fixed effects are incorporated in (22).

Another obvious extension to the set of models we have been developing is to introduce other variables to the model, either linearly as in

$$x_{i,t} = \frac{\rho}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) x_{j,t-1} + \beta z_{i,t} + \epsilon_{i,t} \tag{24}$$

or nonlinearly as in, e.g.,

$$x_{i,t} = \frac{\rho}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) x_{j,t-1} + \frac{\beta}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) z_{j,t-1} + \epsilon_{i,t}; \tag{25}$$

or indeed to introduce other switch variables, giving rise to a model of the form

$$x_{i,t} = \frac{\rho}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r_1\right) x_{j,t-1} + \frac{\beta}{m_{z,i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|z_{i,t-1} - z_{j,t-1}| \leq r_2\right) x_{j,t-1} + \epsilon_{i,t} \tag{26}$$

where $m_{z,i,t} = \sum_{j=1}^{N} \mathcal{I}\left(|z_{i,t-1} - z_{j,t-1}| \leq r_2\right)$. It is also clear from the work of Kapetanios (2001) that information criteria can be used to choose the switch variables. The theoretical properties of the models in (24)-(26) should be obvious from the preceding analysis. For example, geometric ergodicity of (26) holds if $|\rho + \beta| < 1$.

The extension presented in (26) is very important. While it is intuitive that it is likely that there exists some variable which can be used to order units (denoted by $z_{i,t}$ in (26)), it is not clear why one would want to set $z_{i,t} = x_{i,t}$ as we did in the first version of the model we presented in (2). A main reason for doing so, in the first instance, was because then the model was self-contained and could be analysed along the lines seen in Section 2. But there is another reason why one may wish to focus on (2), rather than the more general model (26). To see why, let us provide a simple analogy in terms of an univariate time series model, before analysing the case at hand. Let

$$x_t = s_t + u_t$$

where

$$s_t = \gamma s_{t-1} + v_t$$

and u_t and v_t are serially uncorrelated. It is well known that this model has a univariate $ARMA(1,1)$ representation. Therefore, it is straightforward to see that a good approximation for this model can be provided by fitting an $AR(1)$ model to x_t. Similarly, let the true model for $x_{i,t}$ be given by a slight variation of (26) of the form

$$x_{i,t} = s_{i,t} + \epsilon_{i,t} \tag{27}$$

where

$$s_{i,t} = \frac{\beta}{m_{z,i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|z_{i,t-1} - z_{j,t-1}| \leq r_2\right) q_{j,t-1} \tag{28}$$

and let

$$z_{i,t} = \gamma z_{i,t-1} + v_{i,t} \quad \text{and} \quad q_{i,t} = \delta q_{i,t-1} + \xi_{i,t}$$

By the fact that the $z_{i,t}$ and $q_{i,t}$ are serially correlated, it follows that the $s_{i,t}$ are serially correlated; since units which cluster together along the z dimension at time t will be more likely to cluster together along the z dimension at time $t+1$. Therefore, the serial correlation in $q_{i,t}$ will be transmitted onto $s_{i,t}$. Furthermore, units which cluster along the z dimension will tend to have more correlated $s_{i,t}$ over i. But, of course, this means that units that cluster

along the z dimension will also cluster along the x dimension, in the same order as across the z dimension, since they will have $s_{i,t}$ that are more correlated across i than units which do not cluster along the z dimension. The ensuing clustering along the x dimension then implies that a term of the form $\frac{\rho}{m_{z,i,t}}\sum_{j=1}^{N}\mathcal{I}\left(|x_{i,t-1}-x_{j,t-1}|\leq r_2\right)q_{j,t-1}$ will have explanatory power for $x_{i,t}$, justifying the use of model (2). So, just as (27) can be approximated by an $AR(1)$,

$$x_{i,t}=\frac{\beta}{m_{i,t}}\sum_{j=1}^{N}\mathcal{I}\left(|z_{i,t-1}-z_{j,t-1}|\leq r\right)x_{j,t-1}+\epsilon_{i,t}, \tag{29}$$

can be approximated by (2), which has an 'AR' structure in the distance/trigger variable. Of course, all the above observations remain valid if we replace $q_{i,t}$ with $x_{i,t}$, resulting in a model whose form is closer to our original specification (2). The utility of this approximation becomes more apparent if one notes the possibility of having cross-sectional averages defined through intersections of triggering events with more than one trigger variables, such as

$$x_{i,t}=\frac{\beta}{m_{z,i,t}}\sum_{j=1}^{N}\mathcal{I}\left(\cap_{s=1}^{p}\left\{\left|z_{i,t-1}^{(s)}-z_{j,t-1}^{(s)}\right|\leq r_s\right\}\right)x_{j,t-1}+\epsilon_{i,t} \tag{30}$$

where $\left(z_{i,t-1}^{(1)},...,z_{i,t-1}^{(p)}\right)'$ is a vector of trigger variables and $\mathcal{I}\left(\cap_{s=1}^{p}\left\{\left|z_{i,t-1}^{(s)}-z_{j,t-1}^{(s)}\right|\leq r_s\right\}\right)=1$ if and only if $\mathcal{I}\left(\left|z_{i,t-1}^{(s)}-z_{j,t-1}^{(s)}\right|\leq r_s\right)=1$ for all s. Further, it is also clear that even if there is structural change, whereby the identity of the trigger variables changes over time, the model with the 'AR' structure in the distance/trigger variable, can still approximate the true unknown and changing model.

It is reasonable to expect that there are further sources of cross-sectional dependence in the panel. For example, the endogenously determined cross-sectional dependence exemplified by model (2) can be coupled with exogenous cross-sectional dependence, such as common shocks arising in the macroeconomy. Such exogenous cross-sectional dependence can be modelled by linear factor structures. Further cross-sectional dependence, of the factor variety, can be introduced by considering the following extension of (2)

$$x_{i,t}=\frac{\rho}{m_{i,t}}\sum_{j=1}^{N}\mathcal{I}\left(|x_{i,t-1}-x_{j,t-1}|\leq r\right)x_{j,t-1}+\eta_{i,t} \tag{31}$$

where

$$\eta_{i,t}=\lambda_i'f_t+\epsilon_{i,t} \tag{32}$$

and f_t is an unobserved factor. The estimation of (31) is of particular interest. If the factor is serially uncorrelated, estimation of this model along the lines suggested for estimation of (2) is possible. However, if the factor is serially correlated, it is clear that $\eta_{i,t}$ and $\tilde{x}_{i,t}=\sum_{j=1}^{N}\mathcal{I}\left(|x_{i,t-1}-x_{j,t-1}|\leq r\right)x_{j,t-1}$ are correlated. Then, we suggest estimating a parametric

factor model, whereby the factor is modelled as a VAR process

$$\bar{x}_{i,t} = x_{i,t} - \tilde{x}_{i,t} = \lambda_i' f_t + \epsilon_{i,t}$$

$$f_t = A f_{t-1} + v_t$$

The resulting state space model is then estimated by pseudo-MLE using the Kalman filter. If one entertains (22) as the chosen model, then estimation may be carried out by nonlinear least squares.

It is interesting to consider the behaviour of this extended model. Therefore, we reconsider the model underlying the realisations reported in Figure 2, but allow for a factor which is i.i.d. and distributed as $f_t \sim t_1$. The loadings are given by $\lambda_i \sim U(0,1)$. We are explicitly aiming to introduce extreme behaviour through the factor. We consider two values of ρ, given by 0.9 and 0.999. The realisations from these two different values of r are reported in Figure 5. In the first case, there is clearly a single cluster but, as expected, the factor can generate abrupt shifts in all units. We see this around observation 130; and again around observation 170. Moving onto the very persistent case, yet more interesting behaviour arises. Here it is clear that big shocks attributed to the factor can lead to the destruction or creation of new clusters. For example, a shock around observation 260 leads to consolidation of three clusters into two. Conversely, the shock at observation 325 leads to the emergence of three clusters from the two which existed before the shock.

While our main focus is on the dynamic characteristics of the model, (2), it is also interesting to simultaneously capture contemporaneous cross-sectional dependence effects that might be very important in fields such as financial asset pricing, where dynamics may be less prevalent, at least when modelling the conditional mean. For example, the CAPM specifies that individual asset excess returns depend contemporaneously on a market excess return index which can be viewed as an aggregate of individual excess returns. Alternatively, one can think of opinions (e.g., fund manager opinions) on variables such as asset return prospects, as being determined contemporaneously by agents considering the opinions of similar agents. This motivates us to consider the following extension of the basic model, (2):

$$x_{i,t} = \frac{\rho_0}{m_{0,i,t}} \sum_{j=1, j\neq i}^{N} \mathcal{I}\left(|x_{i,t} - x_{j,t}| \leq r_0\right) x_{j,t} + \frac{\rho_1}{m_{1,i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r_1\right) x_{j,t-1} + \epsilon_{i,t}, \quad (33)$$

where $m_{0,i,t}$ and $m_{1,i,t}$ are defined in an obvious way. This extended model incorporates a complex mechanism for the determination of x_t since each $x_{i,t}$ depends in a complicated way on every other $x_{j,t}$. The complex nature of this extension can be best understood by noting that simulating (33) involves solving N nonlinear simultaneous equations at each point in time, where the nonlinearity has discontinuities arising from the threshold nature of the relevant

functions. This is a non-trivial mathematical problem. A linear simplification may help clarify further the issue. A simplified linear version of (33) is given by

$$x_{i,t} = \frac{\rho_0}{N}\sum_{j=1}^{N} x_{j,t} + \frac{\rho_1}{N}\sum_{j=1}^{N} x_{j,t-1} + \epsilon_{i,t}$$

In the case where $\rho_1 = 0$, the model decouples temporally and the solution at each point in time is given by

$$x_t = \left(I - \frac{\rho}{N}\iota\iota'\right)^{-1}\epsilon_t$$

where $x_t = (x_{1,t}, ..., x_{N,t})'$, $\epsilon_t = (\epsilon_{1,t}, ..., \epsilon_{N,t})'$ and $\iota = (1, ..., 1)'$. It is worth noting that $\left(I - \frac{\rho}{N}\iota\iota'\right)^{-1}$ does not exist when $\rho = 1$.

The final extension generalises further the gamut of weighted averages that can inform the evolution of agent opinion formation or agent actions to a very general class of models which take the form

$$x_{i,t} = \frac{\rho}{m^{\mathcal{S}}_{i,t}}\sum_{j=1}^{N} I\left(j \in \mathcal{S}_{i,t-1}\right) x_{j,t} + \epsilon_{i,t}, \tag{34}$$

where $\mathcal{S}_{i,t-1}$ denotes a set of unit indices for unit i at time $t-1$ and $m^{\mathcal{S}}_{i,t} = \sum_{j=1}^{N} I\left(j \in \mathcal{S}_{i,t-1}\right)$. This opens up a wide variety of modelling options, such as the existence of a leader unit or set of units whose behaviour is mimicked by other units. For example, a specific instance of (34), where

$$\mathcal{S}_{i,t-1} = \mathcal{S}_{t-1} = \arg\max_{j=1,...,N}\sum_{s=1}^{p} q_{j,t-s} \tag{35}$$

might be used to model fund managers that follow the best performing manager in the near past. In this case $x_{i,t}$ would denote the holdings of a given asset by manager i at time t, while $q_{i,t}$ would denote a performance measure of manager i at time t. Of course, multivariate extensions to describe the evolution of holdings for multiple assets are obvious. Similarly

$$\mathcal{S}_{i,t-1} = \mathcal{S}_{t-1} = median\left(\sum_{s=1}^{p} q_{j,t-s}\right) \tag{36}$$

might be used to proxy the behaviour of fund managers that conforms to forms of benchmarking. Obviously schemes such as (35) or (36) imply a factor like covariance matrix for $x_{i,t}$. Note that specifications such as (35) or (36) are significantly different to schemes that *a priori* specify units that are dominant such as, e.g., macroeconometric panel models that give a leading status to U.S. variables. The present specifications describe a *mechanism* that allocates leader status to a given unit or set of units endogenously.

Alternatively, we modify the selection mechanism as follows

$$x_{i,t} = \frac{\rho}{m_{i,t}}\sum_{j=1}^{N}\mathcal{I}\left(\left|x^{\max}_{t-1} - x_{j,t-1}\right| \leq r\right) x_{j,t-1} + \epsilon_{i,t} \tag{37}$$

where $x_{t-1}^{\max} = \max_j x_{j,t-1}$, such that the distance is measured with respect to the best performer rather than unit i. Alternative functional forms, based on the median or mean, might also be considered.

This extension completes the set of extensions that we think are both interesting and relevant for the effects we attempt to capture through our basic model (2). In section 5 we report Monte Carlo results to assess the performance of the estimators proposed in this section.

4 Testing Linearity

In this section, we discuss how to test if the data support the nonlinear representation contained in the proposed models. We start by recalling what parameter values imply linearity both for the basic model, (2), and the leading case of the smooth version of the model given by (22), where $w(x;\gamma) = \exp(-\gamma x^2)$.

As we noted in section 2, setting $r = 0$ reduces (2) to the panel autoregression (4), while setting $r = \infty$ gives the model (5). Both are linear models. We also see that these two linear models are nested in (22). Setting $\gamma = 0$, gives (5); whereas setting $\gamma = \infty$, gives (4). As a result, and unlike standard time series models, there is no unique test of linearity. Which test one carries out very much depends on the null hypothesis of interest.

The differences with linearity tests for standard nonlinear time series models do not stop here. A well-known problem with linearity testing in time series relates to the fact that because there invariably exist underidentified nuisance parameters, the test statistics do not have standard distributions. For example, when two regime threshold (TAR) models are considered, the specifications usually include two autoregressive parameters and the threshold. Linearity is obtained by setting the two autoregressive parameters equal to each other, in which case the threshold parameter is not identified under the null. Further, in the case of threshold models, the problem is compounded by the fact that the threshold parameter does not, in any case, have a standard asymptotic distribution.

A cursory analysis of the panel threshold model suggests that no underidentified parameter problem arises here. Both linear models nested by the nonlinear models, (2) and (22), have the same number of parameters as the nonlinear models, apart from the actual parameter being restricted by the null hypothesis. As a result, testing in the context of the panel model is considerably easier. In the case of (22) and using Theorem 3, one can use the normal asymptotic approximation to carry out tests on γ.

Inference in the threshold model is more difficult due to the nonstandard distribution of $\hat{r}$. Although we have not established this distribution formally, the results in Chan (1993) and Hansen (1999) suggest that it should be nonstandard and very difficult to use in practice.

Note that for standard time series TAR models the standard bootstrap was shown by Yu (2009) to be invalid for inference on the threshold parameter; while the parametric bootstrap was shown to be valid by Yu (2007). Since our model is likely to suffer from a number of potential misspecification issues, which would invalidate the use of the parametric bootstrap, we suggest a simulation approach for conducting inference on this parameter; namely the use of subsampling approach. Gonzalo and Wolf (2005) consider the use of subsampling methods for inference in time series threshold models. Subsampling, as advanced by Politis and Romano (1994), is similar in a number of respects to bootstrapping, and is based on resamples of a smaller dimension than the original sample. Subsampling is more robust, in the sense that subsampling is valid for the overwhelming majority of cases where the bootstrap is invalid, as discussed in Politis, Romano, and Wolf (1999).

In our case, the application of subsampling carries added complications, because the sample grows in two dimensions. Following Politis, Romano, and Wolf (1999) and Kapetanios (2010), we suggest the following algorithm. Set the temporal and cross-sectional subsample sizes to $b_T = T^{\zeta}$ and $b_N = N^{\zeta}$, respectively, for some $0 < \zeta < 1$. Construct initial subsamples by sampling blocks of data temporally. These are given by $\{\tilde{x}_{1,b_T}, \tilde{x}_{2,b_T+1}, ..., \tilde{x}_{T-b_T+1,T}\}$ where $\tilde{x}_{t_1,t_2} = (x_{t_1}, ..., x_{t_2})'$. Then, for each $\tilde{x}_{t_1,t_2}$, select b_N cross-sectional units randomly to construct the B-th subsample, x_{t_1,t_2}, $t_1 = 1, .., T - b_T + 1$, $t_2 = b_T, ..., T$, $B = 1, .., T - b_T + 1$. Notice that the cross-sectional units can be different across subsamples. Although this is of no importance theoretically, it makes sense to employ information contained in as many cross-sectional units as possible. ζ is a tuning parameter related to block size. There exists no theory on its determination, but usual values range between 0.7 and 0.8. Then, r is estimated for each subsample created. The empirical distribution of the set of estimates, denoted by $\hat{r}^{*,(i)}$, $i - 1, ...B$, can be used for inference with the empirical distribution given by

$$L_{b_T,b_N}(x) = \frac{1}{B}\sum_{s=1}^{B} 1\left\{b_N b_T\left(\hat{r}^{*,(s)} - \hat{r}\right) \leq x\right\}. \tag{38}$$

The following theorem justifies the use of subsampling for the nonlinear panel threshold model.

Theorem 4 *Let Assumption 1 hold for $\epsilon_{i,t}$ in (2). Then, as long as $|\rho| < 1$, $L_{b_T,b_N}(x)$ is a consistent estimate of $\Pr_P\left(NT\left(\hat{r} - r^0\right) \leq x\right)$ where P denotes the unknown joint probability distribution of the idiosyncratic errors $\epsilon_{i,t}$.*

As a final point it is worth noting some cases where the need for testing arises for reasons that are specific to the panel nature of the model. One leading case is when one wishes to use this model to draw inference about aggregate variables. Let $\bar{x}_t = \frac{1}{N}\sum_{j=1}^{N} x_{j,t}$ denote the aggregate. Further, consider the case where the model is of the form (4) but with the presence of an exogenous factor. This model is given by

$$x_{i,t} = \rho x_{i,t-1} + \eta_{i,t} \tag{39}$$

where $\eta_{i,t}$ is given by (32). Then, it follows that

$$\bar{x}_t = \rho\bar{x}_{t-1} + \frac{1}{N}\sum_{j=1}^{N}\eta_{i,t} = \rho\bar{x}_{t-1} + \left(\frac{1}{N}\sum_{j=1}^{N}\lambda_i'\right) f_t + \frac{1}{N}\sum_{j=1}^{N}\epsilon_{i,t} \tag{40}$$

Assuming that λ_i does not have zero mean and that $\epsilon_{i,t}$ are zero mean and i.i.d. across i, the above implies that $\bar{x}_t$ follows a linear $AR(1)$ representation whose error tends to f_t as $N \to \infty$. Similarly, letting the model be of the form (5), but allowing for factors, gives

$$x_{i,t} = \rho\frac{1}{N}\sum_{j=1}^{N}x_{j,t-1} + \eta_{i,t} \tag{41}$$

where again $\eta_{i,t}$ is given by (32). Then,

$$\bar{x}_t = \rho\frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N}x_{j,t-1}\right) + \frac{1}{N}\sum_{j=1}^{N}\eta_{i,t} = \rho\bar{x}_{t-1} + \left(\frac{1}{N}\sum_{j=1}^{N}\lambda_i'\right) f_t + \frac{1}{N}\sum_{j=1}^{N}\epsilon_{i,t}$$

which, under the same assumptions as for (40), again implies that $\bar{x}_t$ accepts a linear $AR(1)$ representation whose error term tends to f_t as $N \to \infty$. This appears to justify the widespread use of autoregressive models for aggregate variables. But, if the basic model for $x_{i,t}$ is given by (2), and there is endogenous rather than exogenous cross-sectional dependence, there is no justification for a linear AR model for the aggregate variable. Further, and this has more general and important implications for the modelling of the aggregate variable, if (2) holds then the aggregate variable cannot be modelled in terms of lags of the aggregate variable alone. The constituents of the aggregate variable enter the aggregate equation in complicated ways which imply that the appropriate model for the aggregate variable is based on a model for the whole panel, even if one only cares about the aggregate variable. Therefore, a test of linearity is crucial in determining the model which should be used with aggregated variables.

5 Monte Carlo Study

In this section we undertake a detailed Monte Carlo study of the new model and a number of its extensions. The Monte Carlo study focuses on the small sample properties of the estimators of the nonlinear model.

5.1 Monte Carlo setup

We consider three different sets of Monte Carlo experiment. The first focuses on the main model given by (2); the second considers (10); while the third uses (22). Of course, given the number of extensions considered in the previous section, additional Monte Carlo experiments could be considered, but we feel that these three give a crucial and informative impression of

the performance of the estimators. They enable one to have some confidence in the fact that estimation of the model can be carried out effectively with relatively small samples.

The first set of experiments uses (2), where we set $\rho = 0.9$, $r = 0.5$ and $\sigma^2_{\epsilon_i} = 0.5$; $\epsilon_{i,t} \sim N.I.I.D.(0, \sigma^2_{\epsilon_i})$. We let $N, T = 5, 10, 20, 50, 100, 200$. The grid for determining r is $0.10, 0.11, 0.12, ..., 1.09, 1.10$. The second set of experiments is like the first, but we set $\eta_i \sim N.I.I.D.(0, 1)$ and use within group estimation, which simply involves demeaning both RHS and LHS variables prior to applying least squares. Finally, the third set of experiments uses the model given by (22) where $w(x, \gamma) = e^{-\gamma x^2}$ and $\gamma = 0.5$. The rest of the settings are as with the first set of experiments. The estimation method used is nonlinear least squares. We carry out 1000 replications for all experiments. The bias and variance of the estimators over the Monte Carlo replications (multiplied by 100) are reported in Tables 4-6.

5.2 Monte Carlo results

Results make interesting reading. We start by examining the results for the first set of experiments, reported in Table 4. We look at the estimator for ρ first. The biases for this estimator are extremely small, at less than 0.01 even for $N, T = 5$. Given the very small size of the bias it is not surprising to note that there is not really a clear pattern as the number of observations increases. The bias does not reduce further as N increases, for small values of T, but it does reduce as either T increases or N increases, for moderate and large values of T. Overall, for the largest sample size ($N, T = 200$), the bias is negligible. The variance of $\hat{\rho}$ is reduced at equal rates when either N or T increases, as we should expect from Theorem 2. Moving on to $\hat{r}$, we note that the biases are much larger for very small sample sizes, but reduce very rapidly, again consistent with our expectations given Theorem 2. The most rapid declines occur as N, T increase from their smallest values. Both biases and variances are reduced with either N or T increasing. Overall, it is clear that even when $N, T = 10$ one can be reasonably confident that reliable estimation of (2) can be carried out.

Next, we consider results for the second set of experiments, reported in Table 5. Here, the biases related to $\hat{\rho}$ are considerably larger. The biases are reduced as both N and T rise; but they are reduced much faster with T. The variances for $\hat{\rho}$ are again much larger compared to the first set of experiments, but are reduced quite quickly as the number of observations increases. Moving on to $\hat{r}$, we note that unlike $\hat{\rho}$, the estimation of r is hardly affected by the presence of individual effects. If anything, the performance of the estimator is better. This is a surprising result, but as there is little work on the small sample properties of estimators of nonlinear panel models with individual effects, our prior about the performance of this estimator was not very strong.

Finally, we consider the third set of experiments; results are reported in Table 6. The

biases and variances for $\hat{\rho}$ are comparable but slightly larger than those for the first set of experiments. However, the absolute performance for this estimator is very good even for very small samples, such as $N, T = 5$. Estimation of γ in very small samples is problematic. But, as long as both N and T equal or exceed 10, estimation improves greatly. The size of the bias and variance becomes comparable to that seen for r in the first two sets of experiments.

Overall, we conclude that estimators of both the autoregressive coefficient and the parameters of the nonlinear terms are quite reliable, in terms of bias and variance. The time dimension does not have to be large for reliable inference, in contrast to when linear time series models are estimated. This is helpful given that many panel datasets, to which this model might be applied, have a short time dimension.

6 Empirical Illustrations

In this section, we provide two empirical applications that illustrate the potential utility of the proposed modelling approach.

6.1 Inflation Expectations

In this section we consider a widely exploited dataset that can be usefully analysed with the new nonlinear panel model. This is the Survey of Professional Forecasters (SPF) carried out from 1968 to 1990 by the American Statistical Association and the NBER and, since 1990, by the Federal Reserve Bank of Philadelphia. We should expect macroeconomic forecasts, such as those from the SPF, to be correlated among forecasters and estimation of the new nonlinear panel model is instructive in determining empirically the nature of the cross-sectional dependence. In turn, this is helpful in understanding further the nature of expectation formation. As Carroll (2003) stressed, there have been few attempts to model actual expectations data. Moreover, there have been even fewer studies of expectational data at the micro-economic level. Souleles (2004), who found considerable heterogeneity across individuals, is a notable exception. Other work, more interested in the forecasting properties of these expectational data than in testing alternative models of expectation formation, has restricted attention to modelling any dependence among the agents using factor models (see Gregory, Smith, and Yetman (2001)). Therefore it does not admit the possibility of alternative ways to model dependence, such as our nonlinear model, that may offer an insight into the nature of the dependence. Determining the nature of the dependence among a panel of forecasters also has a practical importance given that Gregory, Smith, and Yetman (2001) motivate use of the mean (across forecasters) forecast as a summary statistic, to be used for policymaking etc., when there is forecast "consensus". Forecast consensus is defined as when individual forecasts are both determined by a latent variable (a factor) subject to an idiosyncratic mean zero error,

and when each forecaster places the same weight on the common component. But the (linear) mean forecast is not a valid measure of consensus under the nonlinear model (see, e.g., Manski (2010)).

In our application we focus on the one-quarter ahead CPI inflation rate forecasts from the SPF. While our model, as discussed in Section 2.5, can accommodate missing data, given there is so much in the SPF we conduct our analysis on a subsample of regular SPF respondents. This is common practice with the SPF and indeed any forecaster panel given that respondents come and go from the survey, for various reasons, so frequently. We focus on responses for the period 1990Q1-2010Q1, a total of 81 quarters. Over this period we have records of 18 professional forecasters, giving a total of 1458 potential observations. However, there remain significant gaps in the dataset which leave a total of 1079 actual observations. We consider the simple model given by (10), with includes constant terms, in this case. This model generalises the model of Gregory, Smith, and Yetman (2001).

The current application provides a number of practical challenges. First, we have to deal with the considerable number of missing observations; we assume that the pattern of missing observations is random. Secondly, we wish to allow for the joint presence of a nonlinear herding mechanism of the form we advocate, as well as the possibility of a factor structure similar to that of Gregory, Smith, and Yetman (2001). To handle missing observations we use the formulation given in (8).

Noting that the inflation rate data are expressed as annualised quarter-over-quarter percentage points, the threshold is estimated to be 0.99 while the estimated autoregressive coefficient is estimated to be 0.5303, with an associated t-statistic of 18.44. If we fit a panel AR model of the form of (4) with constant terms, we get an AR coefficient of 0.4589; whereas if we fit a cross-sectional average model, (5), the coefficient becomes 0.6154. The strongest lagged (inertia) effect is therefore observed in the cross-sectional average model, while the weakest persistence is found in the panel AR model. This suggests that inflation forecasters react to the lagged opinion of the group average (as herding behaviour implies) more strongly than to their own personal opinion in the previous period. This is consistent with the view that individuals may set their forecasts close to the previous averaged opinion, in the hope that if their forecasts are wrong then they are not the only forecaster to make a mistake. Interestingly, the results for the nonlinear panel model lie between these two bounds. Inflation forecasters prefer to set their current forecast close to the average lagged forecast from an endogenously selected peer group, with this peer group identified as those forecasters whose lagged forecast lay within 1% point of their own previous forecast.

Use of this model also has implications when modelling the aggregate forecast. As noted in Section 4, it is not appropriate to assume a linear model for the aggregate forecast given these results. This supports the use of nonlinear models for the conditional mean, perhaps in

conjunction with ARCH structures for the conditional variance. For example, the volatility associated with the spread of forecasts is often used as an important source of information at the aggregate level.

Next, we consider an extension in which the model is augmented by an exogenous factor structure such as (31)-(32). Due to the missing data, we consider a different estimation approach to that suggested when the factor extension was discussed earlier. An additional advantage of the estimation method described below is that we do not need to specify a parametric model for the unobserved factor.

Specifically, we consider an EM type algorithm, whereby we initialise estimation by obtaining some factor estimate and using it as an observed variable in a model of the form

$$x_{i,t} = \frac{\rho}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) x_{j,t-1} + \lambda_i' f_t + \epsilon_{i,t} \tag{42}$$

which is estimated as if the factor were observed, and then the residuals, given by

$$\hat{\epsilon}_{i,t} = x_{i,t} - \frac{\hat{\rho}}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq \hat{r}\right) x_{j,t-1}$$

are used to extract a new estimate of the factor. The whole approach is iterated to convergence. The actual factor is estimated, accommodating missing observations, by introducing a second estimation loop where for a given set of observed residuals and a given pattern of missing residuals, both the factor and the missing residuals are estimated. This is done by conditioning on a factor estimate to get estimated missing residuals using the factor and estimated loadings $\hat{\lambda}_i'$. Once these estimates are obtained one can estimate a new factor estimate. This two step estimation is again iterated to convergence.

When this estimation is carried out we find minimal changes in the parameter estimates for the nonlinear model. The threshold is again estimated to be 0.99, while the estimated autoregressive coefficient is also again given by 0.5305, with an associated t-statistic of 18.46. This suggests that a factor structure is redundant in the presence of the nonlinear cross-sectional average.

One alternative way to see this, that is of independent interest, is to compute both a measure and test of cross-sectional dependence. We use the following statistic, which is a slight modification of the sphericity test statistic of Ledoit and Wolf (2002)

$$cd(x) = \frac{1}{N} tr\left(\left(C(x) - I\right)\left(C(x) - I\right)\right)$$

where $C(x)$ denotes the estimated correlation matrix of a given dataset, x. When the SPF data $x_{i,t}$ are used to compute cd, $cd(x) = 14.69$, while if the residuals from the nonlinear cross-sectional average model, without factors, are used the statistic is 1.76. The statistic obtained

when residuals, from the nonlinear cross-sectional average model with factors, are used, is 1.75. Once again the difference is minimal suggesting that our model is capable of capturing the cross-sectional dependence of the data quite adequately. As a final check, we consider the statistic associated with using only a factor model without a nonlinear structure. The associated statistic is then larger at 3.30, further illustrating the superior fit of the nonlinear model.

6.2 Stock Returns and Realised Volatilities

Perhaps surprisingly, given that our model is one that models the dynamics of the conditional mean, for our second application we consider a dataset of stock returns. We motivate this as follows. Firstly, market returns are important for individual stock returns, albeit contemporaneously, in a number of theoretical models. Our model, with its emphasis on forms of cross-sectional averages, provides a useful vehicle to model them. Second, an autoregressive specification, which is a special case of our model, is used routinely as a benchmark for modelling, and especially forecasting, stock returns. Thirdly, although a linear dynamic specification has a poor track record for modelling stock returns, a common finding in the literature is that nonlinearity has a role to play in this respect (e.g., Guidolin, Hyde, McMillan, and Ono (2009)). This is a common finding when stock return indices are analysed. Given our discussion at the end of section 4, on aggregating processes that follow our model, which implies that the aggregate has a nonlinear structure, our model can offer interesting insights. Finally, as noted in Section 3, a model of the form of (2), which uses the own lag of the dependent variable to define the dimension along which the cross-sectional averaging is carried out, can approximate models which have other variables defining distance. So, in the case of returns, the model we use approximates models that may define distance in terms of industrial sector, profitability or other characteristics. As noted earlier the approximation properties of this model are likely to be retained to a certain extent even if the identity of the variables that regulate the distance undergoes structural change over time. In this sense our model is a 'reduced form' approximation for more structural explanations for cross-sectional correlations in returns.

We consider constituent stock return data from the S&P500 at a weekly frequency. The data are from 1993W1 through 2007W52. In our dataset, only 364 companies are present throughout the period and these are the ones that we analyse.

We first estimate the simple nonlinear model given by (10). We estimate $\hat{\rho} = -0.0995$, $\hat{r} = 0.08$. The t-test associated with $\hat{\rho}$ is -39.37, which is extremely significant given Theorem 2. The panel R^2 associated with the model is 0.0058, which is of course extremely low, but expected, given that we analyse stock returns. The average R^2 across the cross-sectional

equations is 0.0063. Next, we introduce two comparator models: a panel data AR and a model where the lagged cross-sectional average is used as an explanatory variable, i.e., the nonlinear model for $r = \infty$. For the panel data AR, $\hat{\rho} = -0.066$ with t-test given by 37.08, the panel $R^2 = 0.0052$ and the average $R^2 = 0.0053$, while for the cross-sectional average model the respective numbers are: -0.107, -28.40, 0.0033 and 0.0036. The nonlinear model has better fit, as measured by the R^2, than the comparator models. Of course, the nonlinear model has an extra parameter (the threshold) which needs to be penalised. A multivariate information criterion is not possible since the dimension of the model is so large that the determinant of the covariance matrix of the residuals, needed to construct the information criterion, is found to be numerically indistinguishable from zero. We choose to construct information criteria for each cross-sectional equation, where the penalty parameter is set to $1/N$ since the threshold parameter is shared by all cross-sectional equations. Table 7 reports the proportion of companies for which each criterion chooses the nonlinear model over the two comparator models. Again we see that the nonlinear model is preferred over its comparators.

Next, we carry out a variety of tests on the residuals of the models. In particular, for every stock return series, we obtain its residuals, from the nonlinear model and the comparator models, and test them for the following: normality (Jarque-Bera test), residual serial correlation (LM test with 1 and 4 lags), ARCH effects (LM test with 1 and 4 lags) and neglected dynamic nonlinearity (Teräsvirta, Lin, and Granger (1993) RESET type test with a third order polynomial approximation and one lag). We report the number of rejections, at the 5% significance level, in Table 8. It seems that all residuals are non-normal, as one would expect. There is some limited evidence of further serial correlation. There is significant evidence of ARCH effects. There is considerable evidence of neglected nonlinearity. It seems that the cross-sectional model displays considerably more evidence of further serial correlation compared to the other models. The most interesting finding relates to neglected nonlinearity. The nonlinear model has about 10% fewer cases of rejection than the other models. This supports the case for the presence of the effect our model is designed to pick up.

Next, we add idiosyncratic AR(1) components to every cross-sectional equation. This makes the specification more flexible and allows for an own-lag effect whose inclusion has a compelling rationale given the existing literature. We do not consider the panel AR model in this case for obvious reasons. In this case, $\hat{\rho} = -0.083$ with t-test given by -14.81, the panel $R^2 = 0.0098$ and the average $R^2 = 0.0103$ while for the cross-sectional average model the respective numbers are: -0.049, -11.12, 0.0095 and 0.0098. Tables 9 and 10 report the respective information criteria and test results. These again make clear that the nonlinear model is preferred. In particular, the favourable evidence from the neglected nonlinearity test is, if anything, even stronger.

As a further extension we add to the model a set of macroeconomic variables commonly

used in the existing literature to model stock returns. Specifically we consider: a set of US T-bill yields (3-month, 6-month, 1-year, 2-year and 10-year), oil prices (Brent crude), effective exchange rates, industrial production, unemployment rate and CPI inflation. We consider our model augmented with these macroeconomic regressors, and the two restricted versions of the model (the panel AR model and the cross-sectional average model) which, in turn, are both augmented with the set of macroeconomic variables. Estimation then reveals $\hat{\rho} = -0.1106$ and $\hat{r} = 0.06$. The t-test associated with $\hat{\rho}$ is -47.96, which is again very significant given Theorem 2. The panel R^2 associated with the model is 0.02429, which is considerably higher than previously. The average R^2 for the nonlinear model, across cross-sectional equations, is 0.02495. Looking at the two comparator models, for the panel AR $\hat{\rho} = -0.083$ with t-test given by 45.87, the panel $R^2 = 0.02366$ and the average $R^2 = 0.02385$. These results suggest that in-sample the nonlinear model improves fit by at least 4% compared to the linear panel AR model. For the cross-sectional average model the respective numbers are: -0.134, -34.39, 0.0207 and 0.021. Clearly, the nonlinear model has better fit as measured by the R^2 compared to this model as well. Finally, we note that, once again, nonlinearity is less prevalent in the residuals of the nonlinear model, with the nonlinearity test rejecting 138 times, while the equivalent number for the panel AR is 153 and, for the cross-sectional average model, 148.

We undertake a final and crucial test of the stock return nonlinear model. We carry out an extensive out-of-sample forecasting exercise. We focus on the simple nonlinear model given by (10), augmented with idiosyncratic AR components. We compare the one-step-ahead forecasting performance of this model to that of individual AR(1) models fitted to every stock return. We use the relative root mean square error (RMSFE) as our performance criterion and consider the last three years of the data as the forecast evaluation period. We also consider the Diebold-Mariano test of equal predictive ability to evaluate the significance of our findings. The results are supportive of the nonlinear model. Out of 364 stock return series, the nonlinear model outperforms the simple AR models in 206 cases. We have 32 stock returns for which the nonlinear model gives a relative RMSFE compared to the simple AR(1) models of 0.98 or less, with a minimum RMSFE of 0.961. The equivalent numbers for the simple AR(1) models are 6 and 0.972. The Diebold-Mariano tests indicate that the test rejects in favour of the nonlinear model in 24 cases, at a significance level of 5%; the number of rejections in favour of the simple AR(1) model is 4. The equivalent numbers for a 10% significance level are 52 and 18. It is clear that these results suggest that the nonlinear model has a significant advantage in terms of forecasting performance compared to a time series model which is commonly believed to provide a good benchmark when forecasting.

Given the aforementioned fact, that stock returns do not have a large dynamic conditional mean component, we also consider data on realised volatility. The data used in this paper are extracted and compiled from the Trade and Quote (TAQ) database provided by Wharton

Research Data Services. Thirty stocks from the S&P500 components are used; to select the stocks, we rank the 500 component stocks of the S&P500 Index by market capitalization as of March, 2011. The sample period covers almost 18,976 data points, starting from early January 2010 and ending in March 2011. The data record the last price observed during every five minute interval within each working day. Following the literature, we clean the data as follows. First, trades before 9:30 AM or after 4:00 PM are removed to deal with the jumps and days that contain long strings of zero or constant returns (caused by data feed problems) are also eliminated. Finally, any trade that has a 137 price increase (decrease) of more than 5% followed by a price decrease (increase) of more than 5% is removed. We use the previous-tick interpolation method, in order to obtain a regularly spaced sequence of mid-quotes, which are thus sampled at the 5-minute and daily frequency, from which 5-minute and daily log returns are computed. Thus we obtain for each day a total of 78 intra-day observations which are used to compute the realised volatility series.[5] We fit the simple nonlinear model given by (10), augmented with idiosyncratic AR components. Again results suggest the presence of the nonlinear term in our model. Our results indicate that $\hat{\rho} = 0.401$, $\hat{r} = 0.33$ and a t-test associated with $\hat{\rho}$ of 3.27, which is significant, once again. Finally, looking at the t-test associated with the cross-sectional average model augmented with AR components, we get a value of 1.47, associated with the coefficient of the lagged cross-sectional average, which is insignificant providing some further final support for our nonlinear model.

7 Conclusions

In economics and finance fundamental modelling assumptions, such as full-information rational expectations, are increasingly being questioned in favour of bounded forms of rationality and learning, whereby agents interact and form their own views by looking at other agents' views. This *groupthink* can explain herding or clustering, as commonly observed in financial markets, for example; but this type of clustering can also be expected when modelling many types of disaggregate variables. While the theoretical analysis of these forms of rationality has become relatively commonplace, econometric studies and empirical models that complement these theoretical advances are rather less developed. Therefore, in this paper, we propose and analyse a nonlinear dynamic panel data model that in an intuitive manner, that might also be given a structural interpretation, accommodates endogenous cross sectional dependence, whereby agents react to the average view of an endogenously determined group of 'similar' agents.

From an economic point of view, the local cross-sectional averages that appear in the proposed dynamic panel regressions might be interpreted as 'shortcuts' that agents take to

[5]We thank Alev Atak for carrying out the requisite data manipulations.

form views and expectations (cf. Carroll (2003)). This type of interpretation relates our work to the extensive literature on bounded rationality and behavioural explanations for economic behaviour. From an econometric point of view, our model provides, to the best of our knowledge, the first attempt to introduce endogenous cross-sectional correlation into a dynamic panel framework, where units share commonalities in terms of parameters but typically remain stochastically uncorrelated. We link our model to a variety of existing models, such as nonlinear time series models, factor models and dynamic spatial panel data models. We also propose numerous extensions, which indicate the flexibility of our model and its ability to model various types of interaction within the panel, including both strong and weak cross-sectional dependence.

We should hope that the proposed model, given its ability to model evolving clusters among the cross-sectional units, will be useful in various applications in economics and finance; both when modelling and forecasting the disaggregate time-series themselves as well as the aggregated variable. Endogenous cross-sectional dependence, as accommodated by our model, implies that even if interest rests with the aggregate variable the appropriate model for the aggregate is intrinsically nonlinear and requires the disaggregates to be simultaneously considered. The increasing availability and use of micro or disaggregate datasets in economics and finance, where we might expect the micro units to interact whether implicitly or explicitly and thereby cluster, means that we hope that our model will be a useful tool when modelling and forecasting with panels.

Finally, in future research, one might render the model yet more flexible, in terms of capturing interactions among agents, by employing neural-network type selection mechanisms in conjunction with the local cross-sectional averages proposed in this paper. Future applications might use these models to identify the possibly asymmetric effect of 'differences of opinion' on stock prices and volumes (e.g., Banerjee, Kaniel, and Kremer (2009) and Banerjee and Kremer (2010)).

References

Akerlof, G., and R. Shiller (2009): *Animal spirits: How Human Psychology Drives the Economy, and Why It Matters for Global Capitalism*. Princeton University Press.

Bai, J. (2003): "Inferential Theory for Factor Models of Large Dinensions," *Econometrica*, 71, 135–173.

Bai, J., and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221.

Banerjee, A. V. (1992): "A Simple Model of Herd Behavior," *The Quarterly Journal of Economics*, 107(3), 797–817.

Banerjee, S., R. Kaniel, and I. Kremer (2009): "Price Drift as an Outcome of Differences in Higher Order Beliefs," *Review of Financial Studies*, 22, 3707–3734.

Banerjee, S., and I. Kremer (2010): "Disagreement and Learning: Dynamic Patterns of Trade," *Journal of Finance*, Forthcoming.

Bikhchandani, S., D. Hirshleifer, and I. Welch (1992): "A Theory of Fads, Fashion, Custom, and Cultural Change in Informational Cascades," *Journal of Political Economy*, 100(5), 992–1026.

Blondel, V. D., J. M. Hendrickx, and J. N. Tsitsiklis (2009): "Continuous Time Average Preserving Opinion Dynamics with Opinion Dependent Communications," *Working Paper, arXiv No. 0907 4662 v1*.

Carroll, C. D. (2003): "Macroeconomic Expectations Of Households And Professional Forecasters," *The Quarterly Journal of Economics*, 118(1), 269–298.

Chan, K. S. (1989): "A Note on the Geometric Ergodicity of a Markov chain," *Advances in Applied Probability*, 21, 702–704.

——— (1993): "Consistency and Limiting Distribution of the Least Squares Estimator of a Threshold Autoregressive Model," *The Annals of Statistics*, 21(1), 520–533.

Chan, K. S., and H. Tong (1985): "On the Use of the Deterministic Lyapunov Function for the Ergodicity of Stochastic Difference Equations," *Advances in Applied Probability*, 17, 666–678.

Chevillon, C., M. Massmann, and S. Mavroeidis (2010): "Inference in Models with Adaptive Learning," *Journal of Monetary Economics*, 57, 341–351.

Chudik, A., and M. H. Pesaran (2010): "Infinite Dimensional VARs and Factor Models," *Journal of Econometrics*, p. Forthcoming.

Chudik, A., M. H. Pesaran, and E. Tosetti (2009): "Weak and Strong Cross Section Dependence and Estimation of Large Panels," *Mimeo, University of Cambridge*.

Devenow, A., and I. Welch (1996): "Rational herding in financial economics," *European Economic Review*, 40(3-5), 603–615.

Dobruschin, P. L. (1968): "The Description of a Random Field by Means of Conditional Probabilities and Conditions of its Regularity," *Theory of Probability and its Applications*, 13(2), 197–224.

Durlauf, S. N. (1992): "Nonergodic Economic Growth," *Review of Economic Studies*, 60(2), 349–366.

Gilboa, I., O. Lieberman, and D. Schmeidler (2006): "Empirical Similarity," *Review of Economics and Statistics*, 88(3), 433–444.

Gonzalo, J., and M. Wolf (2005): "Subsampling Inference in Threshold Autoregressive Models," *Journal of Econometrics*, 127, 201–224.

Gregory, A. W., G. W. Smith, and J. Yetman (2001): "Testing for Forecast Consensus," *Journal of Business and Economic Statistics*, 19, 34–43.

Guidolin, M., S. Hyde, D. McMillan, and S. Ono (2009): "Non-Linear Predictability in Stock and Bond Returns: When and Where Is It Exploitable," *International Journal of Forecasting*, 21(2), 373–399.

Hansen, B. E. (2000): "Sample Splitting and Threshold Estimation," *Econometrica*, 68, 575–603.

Hayashi, F. (2000): *Econometrics*. Princeton University Press.

Hirshleifer, D., and S. H. Teoh (2003): "Herd Behavior and Cascading in Capital Markets: A Review and Synthesis," *Review of Financial Studies*, 9, 25–66.

Jegadeesh, N., and W. Kim (2010): "Do Analysts Herd? An Analysis of Recommendations and Market Reactions," *Review of Financial Studies*, 23, 901–937.

Kapetanios, G. (2001): "Model Selection in Threshold Models," *Journal of Time Series Analysis*, 22, 733–754.

——— (2010): "A Testing Procedure for Determining the Number of Factors in Approximate Factor Models with Large Datasets," *Journal of Business and Economic Statistics*, 28, 397–409.

Korniotis, G. M. (2010): "Estimating Panel Models With Internal and External Habit Formation," *Journal of Business and Economic Statistics*, 28, 145–158.

Krause, U. (1997): "Soziale Dynamiken mit vielen Interakteuren. Eine Problemskizze," *Modellierung und Simulation von Dynamiken mit vielen interagierenden Akteuren*, pp. 37–51.

Ledoit, O., and M. Wolf (2002): "Some Hypotheses Tests for the Covariance Matrix when the Dimension is Large Compared to the Sample Size," *Annals of Statistics*, 30(4), 1081–1102.

Manski, C. F. (2010): "Interpreting and Combining Heterogeneous Surveys," in *Forecasts, the Oxford Handbook on Economic Forecasting*, ed. by M. Clements, and D. Hendry. Oxford University Press.

Nickell, S. J. (1981): "Biases in Dynamic Models with Fixed Effects," *Econometrica*, 49(6), 1417–26.

Pesaran, M. H. (2006): "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure," *Econometrica*, 74(4), 967–1012.

Politis, D. N., and J. P. Romano (1994): "Large Sample Confidence Regions Based on Subsamples Under Minimal Assumptions," *Annals of Statistics*, 22, 2031–2050.

Politis, D. N., J. P. Romano, and M. Wolf (1999): *Subsampling*. Springer Verlag.

Schwarz, H., H. R. Rutishauser, and E. Stiefel (1973): *Numerical Analysis of Symmetric Matrices*. Prentice Hall.

Sias, R. W. (2004): "Institutional Herding," *Review of Financial Studies*, 17, 165–206.

Souleles, N. S. (2004): "Expectations, Heterogeneous Forecast Errors, and Consumption: Micro Evidence from the Michigan Consumer Sentiment Surveys," *Journal of Money, Credit and Banking*, 36(1), 39–72.

Stock, J. H., and M. W. Watson (2002): "Macroeconomic Forecasting Using Diffusion Indices," *Journal of Business and Economic Statistics*, 20, 147–162.

Teräsvirta, T., C. F. Lin, and C. W. J. Granger (1993): "Power of the Neural Network Linearity Test," *Journal of Time Series Analysis*, 14, 209–220.

Timmermann, A. (1994): "Can Agents Learn to Form Rational Expectations? Some Results on Convergence and Stability of Learning in the UK Stock Market," *Economic Journal*, 104, 777–798.

Tong, H. (1995): *Nonlinear time series: A dynamical system approach.* Oxford University Press.

Trueman, B. (1994): "Analyst Forecasts and Herding Behavior," *Review of Financial Studies*, 7, 97–124.

Tweedie, R. L. (1975): "Sufficient Conditions for Ergodicity and Recurrence of Markov Chains on a General State Space," *Stochastic Processes Appl.*, 3, 385–403.

Yu, P. (2007): "Likelihood-based Estimation and Inference in Threshold Regression," *Working Paper, University of Auckland*.

——— (2009): "Bootstrap in Threshold Regression," *Working Paper, University of Auckland*.

Appendix

Lemmas

In what follows, we develop some theoretical results that form the basis of our analysis. As noted earlier, we aim to analyse the general case where both N and T tend to infinity. Therefore, without loss of generality we let $N(T)$ be an unspecified function of T. For notational convenience we suppress the dependence of N on T. We have the following Lemmas.

Lemma 1 *Let* $\left\{\{x_{i,t}\}_{i=1}^{N}\right\}_{t=1}^{T}$ *follow (2). Then, for all* $N_0 \leq N$, *there exists* T_0 *such that for all* $T > T_0$, $\left\{\{x_{i,t}\}_{i=1}^{N_0}\right\}_{t=T_0}^{T}$ *is geometrically ergodic and asymptotically stationary, as long as* $|\rho| < 1$. *Further, if* $\sup_{i\leq N_0} E\left(\epsilon_{i,t}^4\right) < \infty$, *then* $\sup_{i\leq N_0} E\left(x_{i,t}^4\right) < \infty$.

Proof: We can write the part of (2) relevant for $\{x_{i,t}\}_{i=1}^{N_0}$, as

$$x_t^{(N_0)} = \Phi_t^{(N_0)} x_{t-1}^{(N_0)} + \epsilon_t \tag{43}$$

where $x_t^{(N_0)} = (x_{1,t}, ..., x_{N_0,t})'$, $\epsilon_t^{(N_0)} = (\epsilon_{1,t}, ..., \epsilon_{N_0,t})$ and $\Phi_t^{(N_0)} = [\Phi_{i,j,t}]$ where

$$\Phi_{i,j,t} = \frac{\rho}{m_{i,t}} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right).$$

Then, by Theorem A1.5 of Tong (1995), using the work of Tweedie (1975), the Lemma follows if $\sup_t \lambda_{\max}(\Phi_t^{(N_0)}) < 1$, where $\lambda_{\max}(\Phi_t^{(N_0)})$ denotes the maximum eigenvalue of $\Phi_t^{(N_0)}$ in absolute value. By Schwarz, Rutishauser, and Stiefel (1973), $\sup_t \lambda_{\max}(\Phi_t^{(N_0)})$ is bounded from above by the supremum over t of the row sum norm of $\Phi_t^{(N_0)}$. But, by the definition of $m_{i,t}$ this row sum norm is equal to ρ for all t. Therefore, the result for the first part of the Lemma follows. The second part of the Lemma, follows by the discussions in Remark B of Chan (1993), Chan and Tong (1985) and Chan (1989).

Lemma 2 *Let* $\left\{\{x_{i,t}\}_{i=1}^{N}\right\}_{t=1}^{T}$ *be given by*

$$x_{i,t} = q_{i,t-1} + \epsilon_{i,t}$$

such that the column sum norm of the variance covariance matrix of $\epsilon_t^{(N)}$ *is* $O(1)$ *as* $N \to \infty$. *The column sum norm of the variance covariance matrix of* $x_t^{(N)}$ *is* $O(N)$ *if (i)* $q_{i,t-1}$ *is stationary, (ii) there is* $\delta > 0$ *such that for all* N, *there exist units* $i, j = 1, ..., \delta N$ *such that (a)* $0 < \lim_{N\to\infty} \sup_{i=1,...,\delta N} Var\left(q_{i,t-1}\right)$ *and (b)* $\lim_{N\to\infty} \sup_{i=1,...,\delta N} Var\left(q_{i,t-1}\right) < \infty$ *and (iii) there is* $\delta > 0$ *such that for all* N, *there exist units* $i, j = 1, ..., \delta N$, *such that* $Cov\left(q_{i,t-1}, q_{j,t-1}\right) \neq 0$. *If (ii)(a) does not hold then the column sum norm of the variance covariance matrix of* $x_t^{(N)}$ *is* $O(1)$.

Proof: The proof is immediate once the definition of the column sum norm is taken into account.

Lemma 3 *Let* $\left\{\{x_{i,t}\}_{i=1}^{N}\right\}_{t=1}^{T}$ *follow (5). The column sum norm of the variance covariance matrix of* $x_t^{(N)}$ *is* $O(1)$.

Proof: To prove this theorem we will use the second part of Lemma 2. (5) can be written as

$$x_t = \nu + \rho\iota\bar{x}_{t-1} + \epsilon_t = \nu + \rho\Phi x_{t-1} + \epsilon_t \tag{44}$$

where $\nu = (\nu_1, ..., \nu_N)'$, $\bar{x}_{t-1} = \frac{1}{N}\sum_{j=1}^{N} x_{j,t-1}$, $\Phi = \frac{1}{N}\iota\iota'$ and $\iota = (1, ..., 1)'$. Note that Φ is idempotent. This implies that

$$x_t = \frac{(1-\rho^{t-1})}{(1-\rho)}\Phi\nu + \rho^t\Phi x_0 + \epsilon_t + \Phi\sum_{i=1}^{t-1}\rho^i\epsilon_{t-i} = \rho^t\Phi x_0 + \epsilon_t + \iota\left[\frac{1}{N}\sum_{j=1}^{N}\xi_{j,t}\right] \tag{45}$$

where

$$\xi_{j,t} = \sum_{i=1}^{t-1}\rho^i\epsilon_{j,t-i}$$

But, it is straightforward to show that

$$\lim_{N\to\infty} Var\left(\frac{1}{N}\sum_{j=1}^{N}\xi_{j,t}\right) = 0$$

this proving the Lemma.

Lemma 4 *Let* $\left\{\{x_{i,t}\}_{i=1}^{N}\right\}_{t=1}^{T}$ *follow (2). The column sum norm of the variance covariance matrix of* $x_t^{(N)}$ *is* $O(N)$.

Proof: We use Lemma 2. The model can be written as

$$x_{i,t} = q_{i,t-1} + \epsilon_{i,t}$$

where

$$q_{i,t-1} = \frac{\rho}{m_{i,t}}\sum_{j=1}^{N}\mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right)x_{j,t-1}.$$

We need to verify the three conditions of Lemma 2. Condition (i) follows from Lemma 1. Next, we establish Condition (ii). By Lemma 1 it follows that it is sufficient to show that

$$0 < \lim_{N\to\infty} Var\left(\frac{\rho}{m_{i,t}}\sum_{j=1}^{N}\mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right)x_{j,t-1}\right), \text{ for all } j. \tag{46}$$

By Assumption 1, we know that

$$\Pr\left(|\epsilon_{i,t} - \epsilon_{i,t-1}| > r\right) > 0$$

for all j and $r < \infty$. This implies that

$$\Pr\left(|x_{i,t} - x_{i,t-1}| > r\right) > 0$$

for all j and $r < \infty$. From this it follows that there exists $\epsilon > 0$ such that

$$\Pr\left(\left|\frac{\rho}{m_{i,t}}\sum_{j=1}^{N}\mathcal{I}\left(|x_{i,t} - x_{j,t}| \leq r\right)x_{j,t} - \frac{\rho}{m_{i,t}}\sum_{j=1}^{N}\mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right)x_{j,t-1}\right| > \epsilon\right) > 0 \quad (47)$$

But since by Markov's inequality

$$\Pr\left(\left|\frac{\rho}{m_{i,t}}\sum_{j=1}^{N}\mathcal{I}\left(|x_{i,t} - x_{j,t}| \leq r\right)x_{j,t} - \frac{\rho}{m_{i,t}}\sum_{j=1}^{N}\mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right)x_{j,t-1}\right| > \epsilon\right) <$$

$$\frac{1}{\epsilon^2}E\left|\frac{\rho}{m_{i,t}}\sum_{j=1}^{N}\mathcal{I}\left(|x_{i,t} - x_{j,t}| \leq r\right)x_{j,t} - \frac{\rho}{m_{i,t}}\sum_{j=1}^{N}\mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right)x_{j,t-1}\right|^2$$

(47) implies (46). The final condition to be checked is Condition (iii) of Lemma 2. We need to show that there is $\delta > 0$ such that for all N, there exist units $i, j = 1, ..., \delta N$, such that

$$E\left[\left(\frac{\rho}{m_{i,t}}\sum_{s=1}^{N}\mathcal{I}\left(|x_{i,t} - x_{s,t}| \leq r\right)x_{s,t}\right)\left(\frac{\rho}{m_{j,t}}\sum_{s=1}^{N}\mathcal{I}\left(|x_{j,t} - x_{s,t}| \leq r\right)x_{s,t}\right)\right] \neq 0 \quad (48)$$

Let $\mathcal{M}_{j,t}$ denote the set of j such that $\mathcal{I}\left(|x_{i,t} - x_{j,t}| \leq r\right) = 1$. By the geometric ergodicity of $x_t^{(N_0)}$ for all N_0, established in Lemma 1, and the fact that the stationary density of $x_t^{(N_0)}$ is strictly positive over all compact sets in $\mathbb{R}^{N_0}$ for all N_0, which is implied by our assumption that the density of $\epsilon_t^{(N_0)}$ is strictly positive over all compact sets in $\mathbb{R}^{N_0}$ for all N_0, we have that there is a non-zero proportion of units, that lie in both $\mathcal{M}_{i,t}$ and $\mathcal{M}_{j,t}$ for a non-zero proportion of j,$k = 1, ..., N$. This implies that (48) holds for some $\delta > 0$ and units $i, j = 1, ..., \delta N$, proving the result of the Lemma.

Lemma 5 *Let* $\left\{\{x_{i,t}\}_{i=1}^{N}\right\}_{t=1}^{T}$ *follow (2). Then,*

$$\sup_i Var\left(\frac{\rho}{m_{i,t}}\sum_{j=1}^{N}\mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right)x_{j,t-1}\right) = O(1) \quad (49)$$

and

$$\inf_i Var\left(\frac{\rho}{m_{i,t}}\sum_{j=1}^{N}\mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right)x_{j,t-1}\right) = O(1) \quad (50)$$

Proof: We examine (49) which involves simply a form of cross-sectional averaging. (50) can be analysed similarly. It is easy to see that

$$Var\left(\frac{1}{m_{i,t}}\sum_{j\in\mathcal{M}_{i,t}}x_{j,t-1}\right)\sim\frac{1}{m_{i,t}^2}\left(\sum_{j\in\mathcal{M}_{i,t}}\sigma_{x_j}^2+2\sum_{j\in\mathcal{M}_{i,t}}\sum_{k\in\mathcal{M}_{i,t}}\sigma_{x_j,x_k}\right)$$

where $\sim$ denotes equality in order of magnitude and $\sigma_{x_j}^2$ and σ_{x_j,x_k} denote the variance of $x_{j,t-1}$ and the covariance of $x_{j,t-1}$ and $x_{k,t-1}$ respectively. The result of the Lemma follows immediately by Lemma 2.

Lemma 6 *Let* $\left\{\{x_{i,t}\}_{i=1}^{N}\right\}_{t=1}^{T}$ *follow (22). Then, for every* $N_0\leq N$*, there exists* T_0 *such that for all* $T>T_0$*,* $\left\{\{x_{i,t}\}_{i=1}^{N_0}\right\}_{t=T_0}^{T}$ *is geometrically ergodic and asymptotically stationary, as long as* $|\rho|<1$*.*

Proof: Proceeding as in the proof of Lemma 1, we can write part of (22) relevant for $\{x_{i,t}\}_{i=1}^{N_0}$, as

$$x_t=\Phi_t^{w,(N_0)}x_{t-1}+\epsilon_t \tag{51}$$

where $\Phi_t^{w,(N_0)}=[\Phi_{i,j,t}^{w}]$ where

$$\Phi_{i,j,t}^{w}=\frac{\rho w(|x_{i,t-1}-x_{j,t-1}|;\gamma)}{\sum_{j=1}^{N}w(|x_{i,t-1}-x_{j,t-1}|;\gamma)}.$$

Then, the result follows along very similar lines to the proof of Lemma 1.

Lemma 7 *Let* $\left\{\{x_{i,t}\}_{i=1}^{N}\right\}_{t=1}^{T}$ *follow (22). The column sum norm of the variance covariance matrix of* $x_t^{(N)}$ *is* $O(N)$*.*

Proof: Let $x_t=x_t^{(N)}$and $\Phi_t^w=\Phi_t^{w,(N)}$. We have the following MA representation of x_t.

$$x_t=\left(\prod_{j=1}^{t}\Phi_{t-j}^{w}\right)x_0+\epsilon_t+\sum_{i=1}^{t-1}\left(\prod_{j=1}^{i-1}\Phi_{t-j}^{w}\right)\epsilon_{t-i} \tag{52}$$

By Lemma 6,

$$\left\|\prod_{j=1}^{t}\Phi_{t-j}^{w}\right\|=O_{a.s.}\left(\rho^t\right) \tag{53}$$

Noting that ϵ_t is a i.i.d. sequence gives,

$$E\left(x_tx_t'\right)=\Sigma_\epsilon+\sum_{i=1}^{t-1}E\left(\left(\prod_{j=1}^{i-1}\Phi_{t-j}^{w}\right)\epsilon_{t-i}\epsilon_{t-i}'\left(\prod_{j=1}^{i-1}\Phi_{t-j}^{w}\right)'\right) \tag{54}$$

It is sufficient to show that

$$\left\|E\left(\Phi_t^w \epsilon_{t-1}\epsilon_{t-1}'\Phi_t^{w\prime}\right)\right\|_c = O(N)$$

where $||.||_c$ denotes column sum norm. We have that

$$\Phi_t\epsilon_{t-1} = \left(\rho\sum_{j=1}^{N}\frac{w(|x_{1,t-1}-x_{j,t-1}|;\gamma)}{\sum_{j=1}^{N}w(|x_{i,t-1}-x_{j,t-1}|;\gamma)}\epsilon_{j,t-1},...,\rho\sum_{j=1}^{N}\frac{w(|x_{N,t-1}-x_{j,t-1}|;\gamma)}{\sum_{j=1}^{N}w(|x_{i,t-1}-x_{j,t-1}|;\gamma)}\epsilon_{j,t-1}\right)'$$

and it follows that every element of $\Phi_t^w\epsilon_{t-1}\epsilon_{t-1}'\Phi_t^{w\prime}$ has nonzero expectation by the geometric ergodicity of $x_t^{(N_0)}$ established in Lemma 6. As a result, $\left\|E\left(\Phi_t\epsilon_{t-1}\epsilon_{t-1}'\Phi_t'\right)\right\|_c = O(N)$, thus establishing the result of the Lemma. It can again be similarly established that for any ordering of the units and any choice of N_0, the Lemma holds for $x_t^{(N_0)}$.

Lemma 8 *Let* $\left\{\{x_{i,t}\}_{i=1}^N\right\}_{t=1}^T$ *follow (22). Then,*

$$\sup_i Var\left(\sum_{j=1}^{N}\frac{\rho w(|x_{i,t-1}-x_{j,t-1}|;\gamma)x_{j,t-1}}{\sum_{j=1}^{N}w(|x_{i,t-1}-x_{j,t-1}|;\gamma)}\right) = O(1)$$

and

$$\inf_i Var\left(\sum_{j=1}^{N}\frac{\rho w(|x_{i,t-1}-x_{j,t-1}|;\gamma)x_{j,t-1}}{\sum_{j=1}^{N}w(|x_{i,t-1}-x_{j,t-1}|;\gamma)}\right) = O(1)$$

Proof: The proof follows very similarly to that of Lemma 3.

Lemma 9 *Let* $\left\{\{x_{i,t}\}_{i=1}^N\right\}_{t=1}^T$ *follow (10).Let* $\bar{\epsilon}_{i,t} = \epsilon_{i,t} - \bar{\epsilon}_i$*, where* $\bar{\epsilon}_i = \frac{1}{T}\sum_{j=1}^{T}\epsilon_{j,t}$ *Then, there exists* T_0 *such that for all* $T > T_0$*,*

$$E\left(\left[\frac{\rho}{m_{i,t}}\sum_{j=1}^{N}\mathcal{I}\left(|x_{i,t-1}-x_{j,t-1}|\leq r\right)x_{j,t-1}\right](\epsilon_{i,t}-\bar{\epsilon}_i)\right) = O\left(\frac{1}{NT}\right).$$

Proof: We establish the result for $r = \infty$ (i.e. the linear model given by (5)). Then, the result follows by Lemma 1 and the assumption that the stationary density of $\{x_{i,t}\}_{i=1}^{N_0}$ is positive uniformly over N_0, since this implies that there exists T_0 such that for all $T > T_0$, and uniformly over i, the number of j such that $\mathcal{I}\left(|x_{i,t-1}-x_{j,t-1}|\leq r\right) = 1$ for any t, is a non-zero proportion of N_0, for all N_0.

To show the result for the linear model, let, as before, $x_t = x_t^{(N)}$. As before, (5) can be written as

$$x_t = \nu + \rho\iota\bar{x}_{t-1} + \epsilon_t = \nu + \rho\Phi x_{t-1} + \epsilon_t \tag{55}$$

where $\nu = (\nu_1, ..., \nu_N)'$, $\bar{x}_{t-1} = \frac{1}{N}\sum_{j=1}^{N} x_{j,t-1}$, $\Phi = \frac{1}{N}\iota\iota'$ and $\iota = (1, ..., 1)'$. Note that Φ is idempotent. This implies that

$$x_t = \frac{(1-\rho^{t-1})}{(1-\rho)}\Phi\nu + \rho^t\Phi x_0 + \epsilon_t + \Phi\sum_{i=1}^{t-1}\rho^i\epsilon_{t-i} = \rho^t\Phi x_0 + \epsilon_t + \iota\sum_{i=1}^{t-1}\rho^i\left(\frac{1}{N}\sum_{j=1}^{N}\epsilon_{j,t-i}\right) \quad (56)$$

For simplicity, we assume that $x_0 = \nu = 0$. We need to show that

$$E\left(\left[\frac{\rho}{N}\sum_{j=1}^{N}x_{j,t-1}\right](\epsilon_{i,t} - \bar{\epsilon}_i)\right) = E\left(\rho\bar{x}_{t-1}(\epsilon_{i,t} - \bar{\epsilon}_i)\right) = O\left(\frac{1}{NT}\right).$$

We have that

$$\bar{x}_{t-1} = \frac{1}{N}\sum_{j=1}^{N}\epsilon_{j,t-1} + \sum_{i=1}^{t-2}\rho^i\left(\frac{1}{N}\sum_{j=1}^{N}\epsilon_{j,t-i-1}\right)$$

Then,

$$\bar{x}_{t-1}(\epsilon_{i,t} - \bar{\epsilon}_i) = \left(\frac{1}{N}\sum_{j=1}^{N}\epsilon_{j,t-1}\right)(\epsilon_{i,t} - \bar{\epsilon}_i) + \sum_{i=1}^{t-2}\rho^i\left(\frac{1}{N}\sum_{j=1}^{N}\epsilon_{j,t-i-1}\right)(\epsilon_{i,t} - \bar{\epsilon}_i) \quad (57)$$

Looking at the expectation of the first term on the RHS of (57), we have

$$E\left(\left(\frac{1}{N}\sum_{j=1}^{N}\epsilon_{j,t-1}\right)\left(\epsilon_{i,t} - \frac{1}{T}\sum_{j=1}^{T}\epsilon_{j,t}\right)\right) = \frac{1}{NT}\sigma_\epsilon^2 \quad (58)$$

For the expectation of the second term on the RHS of (57), using (58), we have

$$E\left(\sum_{i=1}^{t-2}\rho^i\left(\frac{1}{N}\sum_{j=1}^{N}\epsilon_{j,t-i-1}\right)\left(\epsilon_{i,t} - \frac{1}{T}\sum_{j=1}^{T}\epsilon_{j,t}\right)\right) = \frac{1}{NT}\sum_{i=1}^{t-2}\rho^i\sigma_\epsilon^2 = \frac{(1-\rho^{t-1})\sigma_\epsilon^2}{(1-\rho)NT}$$

which proves the result.

Lemma 10 *Let* $\left\{\{x_{i,t}\}_{i=1}^{N}\right\}_{t=1}^{T}$ *follow (14). Then, for all* $N_0 \leq N$*, there exists* T_0 *such that for all* $T > T_0$*,* $\left\{\{x_{i,t}\}_{i=1}^{N_0}\right\}_{t=T_0}^{T}$ *is geometrically ergodic and asymptotically stationary, as long as* $p\sum_{i=1}^{p}|\rho_s| < 1$*.*

Proof: As is usual for autoregressive models with more than one lag, we write the model in companion form. So, we can write the part of (14) relevant for $\{x_{i,t}\}_{i=1}^{N_0}$, as

$$x_t^{(p,N_0)} = \Phi_t^{(p,N_0)}x_{t-1}^{(p,N_0)} + \epsilon_t^{(p,N_0)} \quad (59)$$

where $x_t^{(p,N_0)} = (x_{1,t}, ..., x_{N_0,t}, ..., x_{1,t-p}, ..., x_{N_0,t-p})'$, $\epsilon_t^{(N_0)} = (\epsilon_{1,t}, ..., \epsilon_{N_0,t}, 0, ..., 0)'$,

$$\Phi_t^{(p,N_0)} = \begin{pmatrix} \tilde{\Phi}_t^{(1,N_0)} & \tilde{\Phi}_t^{(2,N_0)} & ... & \tilde{\Phi}_t^{(p,N_0)} \\ I & 0 & ... & 0 \\ ... & ... & ... & ... \\ 0 & ... & I & 0 \end{pmatrix},$$

$\tilde{\Phi}_t^{(s,N_0)} = [\tilde{\Phi}_{i,j,t}^{(s)}]$, $s = 1, ..., p$, and

$$\tilde{\Phi}_{i,j,t}^{(s)} = \frac{\rho_s}{m_{i,t,s}} \mathcal{I}\left(|x_{i,t-s} - x_{j,t-s}| \leq r\right) x_{j,t-s}.$$

Then, similarly to the proof of Lemma 1 it is sufficient that the row sum norm of $\left(\tilde{\Phi}_t^{(1,N_0)} \quad \tilde{\Phi}_t^{(2,N_0)} \quad ... \quad \tilde{\Phi}_t^{(p,N} \right.$ is bounded from above by one. But for this, it sufficient that $p \sum_{i=1}^{p} |\rho_s| < 1$ proving the result.

Lemma 11 *Let* $\left\{ \{x_{i,t}\}_{i=1}^{N} \right\}_{t=1}^{T}$ *follow (15). Then, for all* $N_0 \leq N$*, there exists* T_0 *such that for all* $T > T_0$*,* $\left\{ \{x_{i,t}\}_{i=1}^{N_0} \right\}_{t=T_0}^{T}$ *is geometrically ergodic and asymptotically stationary, as long as* $q \sum_{i=1}^{q} |\rho_s| < 1$*.*

Proof: The proof follows along very similar lines to that of Lemma 10.

Proof of Theorem 1

We prove consistency of the least squares estimator of ρ and r. We define $x_{ij,t-s} = |x_{i,t-s} - x_{j,t-s}|$ and $\mathcal{F}_{t-1} = \sigma(x_{1,t-1}, ..., x_{N,t-1}, x_{1,t-2}, ..., x_{N,t-2}, ...)$. Recall that ρ^0 and r^0 denote the true value of ρ and r, and denote the respective expectation conditional on $\mathcal{F}_{t-1}$ by $E_{\rho,r}(.|t-1)$. We proceed as in Chan (1993). Following the proof of consistency of the threshold parameter estimates by Chan (1993), we see that three conditions need to be satisfied for consistency. Firstly, we need to show that the data $x_{i,t}$ are geometrically ergodic and hence asymptotically covariance stationary (Condition C1). Secondly, we need to show that (Condition C2)

$$E\left(x_{i,t} - E_{\rho^0,r^0}(x_{i,t}|t-1)\right)^2 < E\left(x_{i,t} - E_{\rho,r}(x_{i,t}|t-1)\right)^2 \quad \forall \rho \neq \rho^0, \quad \forall r \neq r^0, \, i = 1, .., N, \tag{60}$$

is satisfied and, thirdly, we need to show that (Condition C3)

$$\lim_{\delta \to 0} E\left(\sup_{(\rho,r) \in B((\rho^0,r^0),\delta)} |E_{\rho^0,r^0}(x_{i,t}|t-1) - E_{\rho,r}(x_{i,t}|t-1)| \right) = 0, \tag{61}$$

where $B(a,b)$ is an open ball of radius b centered around a, is satisfied. These three conditions together imply the uniform convergence of the objective function given

$$S(\rho, r) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(x_{i,t} - \frac{\rho}{m_{i,t}} \sum_{j=1}^{N} \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) x_{j,t-1} \right)^2$$

to the limit objective function which is the key to establishing consistency. C1 is needed for obtaining a law of large numbers needed for Claim 1 of Chan (1993), and hence for convergence of the objective function. C3 is needed for uniformity of the convergence and, finally, C2 is needed to show that the limiting objective function is minimized at the true parameter values. C1 can be seen to follow from Lemma 1. We establish C2 and C3.

For C2 we have that

$$E\left(x_{i,t} - E_{\rho^0,r^0}(x_{i,t}|t-1)\right)^2 = \sigma^2_{\epsilon_i} \tag{62}$$

Letting $m^0_{i,t} = \sum_{j=1}^N \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r^0\right)$, and assuming, without loss of generality, that $r \geq r^0$, we also have

$$E\left(x_{i,t} - E_{\rho,r}(x_{i,t}|t-1)\right) = \epsilon_{i,t} + \frac{\rho^0}{m^0_{i,t}} \sum_{j=1}^N \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r^0\right) x_{j,t-1} - \tag{63}$$

$$\frac{\rho}{m_{i,t}} \sum_{j=1}^N \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) x_{j,t-1} =$$

$$\epsilon_{i,t} + \frac{(\rho^0 - \rho)}{m_{i,t}} \sum_{j=1}^N \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r^0\right) x_{j,t-1} -$$

$$\frac{\rho}{m_{i,t}} \sum_{j=1}^N \left(\mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) - \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r^0\right)\right) x_{j,t-1} = \epsilon_{i,t} + h_{i,t-1}$$

But, under our assumption that $\epsilon_{i,t}$ is i.i.d. across i and t, $E(\epsilon_{i,t} h_{i,t-1}) = 0$, thus implying that

$$E\left(\epsilon_{i,t} + h_{i,t-1}\right)^2 > \sigma^2_{\epsilon_i}$$

and thereby establishing C2. For C3, we have, using (63),

$$E_{\rho^0,r^0}(x_{i,t}|t-1) - E_{\rho,r}(x_{i,t}|t-1) = \frac{(\rho^0 - \rho)}{m_{i,t}} \sum_{j=1}^N \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r^0\right) x_{j,t-1} - \tag{64}$$

$$\frac{\rho}{m_{i,t}} \sum_{j=1}^N \left(\mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) - \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r^0\right)\right) x_{j,t-1}$$

We examine the first term of the RHS of (64). By Lemma 3,

$$\frac{1}{m_{i,t}} \sum_{j=1}^N \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r^0\right) x_{j,t-1} = O_{m.s.}(1)$$

and so

$$\lim_{\delta \to 0} E\left(\sup_{(\rho,r) \in B((\rho^0,r^0),\delta)} \left| \frac{(\rho^0 - \rho)}{m_{i,t}} \sum_{j=1}^N \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r^0\right) x_{j,t-1} \right| \right) = 0$$

Moving to the second term on the RHS of (64), we have, using the fact that the stationary density of $\{x_{i,t}\}_{i=1}^{N_0}$ is positive and bounded, uniformly over N_0, which follows from Assumption 1 on the density of $\{\epsilon_{i,t}\}_{i=1}^{N_0}$, that

$$\lim_{\delta \to 0} E\left(\sup_{(\rho,r) \in B((\rho^0,r^0),\delta)} \left| \frac{\rho}{m_{i,t}} \sum_{j=1}^N \left(\mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r\right) - \mathcal{I}\left(|x_{i,t-1} - x_{j,t-1}| \leq r^0\right)\right) x_{j,t-1} \right| \right) \leq$$

$$\lim_{\delta \to 0} \sup_{i,j} \sup_{(\rho,r) \in B((\rho^0,r^0),\delta)} Pr(|x_{i,t-1} - x_{j,t-1}| \in (r, r^0)) = 0$$

proving the result.

Proof of Theorem 2

We prove the rate of convergence of $\hat{r}$ to r^0. We focus on the pooled least squares estimator. Since we know that $(\hat{\rho}, \hat{r})$ is consistent, we restrict the parameter space to a neighborhood of (ρ^0, r^0), given by

$$\vartheta(\Delta) = \left\{(\rho, r) \in \Omega,\ \left|\rho - \rho^0\right| < \Delta;\ \left|r - r^0\right| < \Delta,\ 0 < \Delta < 1\right\}$$

It is sufficient to prove that for any ε, there exists K, such that for $(\rho, r) \in \vartheta(\Delta)$, and $r > K/(NT)$,

$$\Pr\left(S(\rho, r) - S(\rho, r^0) > 0\right) > 1 - \varepsilon. \tag{65}$$

Recall that $x_{ij,t-s} = |x_{i,t-s} - x_{j,t-s}|$. Define $Q_{ij}(r) = E\left(\mathcal{I}\left(x_{ij,t} < r\right)\right)$. By Claim 1 of Proposition 1 of Chan (1993), it follows that (65) holds if for any $\varepsilon > 0$, $\eta > 0$, there exists $K > 0$ such that for all N, T

$$\inf_{1 \le i,j \le N} \Pr\left(\sup_{\Delta \ge r > K/(NT)} \left|\sum_{j=1}^{N}\sum_{t=2}^{T} \frac{\mathcal{I}\left(x_{ij,t-1} < r\right)}{NTQ_{ij}(r)} - 1\right| < \eta\right) > 1 - \varepsilon, \tag{66}$$

$$\inf_{1 \le i,j \le N} \Pr\left(\sup_{\Delta \ge r > K/(NT)} \left|\sum_{j=1}^{N}\sum_{t=2}^{T} \frac{\epsilon_{i,t}\mathcal{I}\left(x_{ij,t-1} < r\right)}{NTQ_{ij}(r)}\right| < \eta\right) > 1 - \varepsilon \tag{67}$$

and

$$\inf_{1 \le i,j \le N} \Pr\left(\sup_{\Delta \ge r > K/(NT)} \left|\sum_{j=1}^{N}\sum_{t=2}^{T} \frac{x_{j,t-1}\epsilon_{i,t}\mathcal{I}\left(x_{ij,t-1} < r\right)}{NTQ_{ij}(r)}\right| < \eta\right) > 1 - \varepsilon \tag{68}$$

By Claim 2 of Proposition 1 of Chan (1993), (66)- (68) hold if there exists $H < \infty$, such that

$$\sup_{1 \le i,j \le N} Var\left(\sum_{j=1}^{N}\sum_{t=2}^{T} \mathcal{I}\left(x_{ij,t-1} < r\right)\right) \le NTH \sup_{1 \le i,j \le N} Q_{ij}(r), \tag{69}$$

$$\sup_{1 \le i,j \le N} Var\left(\sum_{j=1}^{N}\sum_{t=2}^{T} |x_{j,t-1}\epsilon_{i,t}|\,\mathcal{I}\left(r_1 < x_{ij,t-1} < r_2\right)\right) \le NTH \sup_{1 \le i,j \le N} \left(Q_{ij}(r_2) - Q_{ij}(r_1)\right) \tag{70}$$

and

$$\sup_{1 \le i,j \le N} Var\left(\sum_{j=1}^{N}\sum_{t=2}^{T} x_{j,t-1}\epsilon_{i,t}\mathcal{I}\left(x_{ij,t-1} < r\right)\right) \le NTH \sup_{1 \le i,j \le N} Q_{ij}(r) \tag{71}$$

But, by Lemma 1and the boundedness of the indicator function, it follows that there exists $0 < m < M < \infty$ such that

$$mr \le \sup_{1 \le i,j \le N} Q_{ij}(r) \le Mr \tag{72}$$

Then, by (72), the uniform boundedness of the indicator function and the second part of Lemma 1, (69)-(71) follow, thus proving the result for the rate of convergence. The second part of the theorem follows similarly to the proof of Theorem 2 and (4.11) of Chan (1993).

Proof of Theorem 3

We wish to prove that the NLS estimator of (ρ^0, γ^0), denoted by $(\hat{\rho}, \hat{\gamma})$ is consistent and asymptotically normal. For consistency, we need to establish conditions (60) and (61) but for the model given by (22). These follow along very similar lines to those for the threshold model and are therefore omitted. These conditions together with geometric ergodicity imply consistency.

Let

$$Q(\rho, \gamma) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=2}^{T} \left(x_{i,t} - \rho \sum_{j=1}^{N} \frac{w(|x_{i,t-1} - x_{j,t-1}|; \gamma) x_{j,t-1}}{\sum_{j=1}^{N} w(|x_{i,t-1} - x_{j,t-1}|; \gamma)} \right)^2$$

For asymptotic normality, we note that using, e.g., Proposition 7.8 of Hayashi (2000), and noting that, under our assumptions, (ρ^0, γ^0) lies in the interior of the parameter space and $w(.;.)$ is twice differentiable and integrable, it is sufficient to show that

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=2}^{T} \left(\sum_{j=1}^{N} \frac{\rho^0 \frac{\partial w}{\partial \gamma}(|x_{i,t-1} - x_{j,t-1}|; \gamma^0) x_{j,t-1} \epsilon_{i,t}}{\sum_{j=1}^{N} w(|x_{i,t-1} - x_{j,t-1}|; \gamma^0)} \right) \xrightarrow{d} N(0, W_1), \tag{73}$$

$$\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=2}^{T} \left(\sum_{j=1}^{N} \frac{w(|x_{i,t-1} - x_{j,t-1}|; \gamma^0) x_{j,t-1} \epsilon_{i,t}}{\sum_{j=1}^{N} w(|x_{i,t-1} - x_{j,t-1}|; \gamma^0)} \right) \xrightarrow{d} N(0, W_2). \tag{74}$$

where

$$W_1 = \lim_{N \to \infty} E \left(\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \frac{\rho^0 \frac{\partial w}{\partial \gamma}(|x_{i,t-1} - x_{j,t-1}|; \gamma^0) x_{j,t-1} \epsilon_{i,t}}{\sum_{j=1}^{N} w(|x_{i,t-1} - x_{j,t-1}|; \gamma^0)} \right) \right)^2 \right),$$

$$W_2 = \lim_{N \to \infty} E \left(\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \frac{w(|x_{i,t-1} - x_{j,t-1}|; \gamma^0) x_{j,t-1} \epsilon_{i,t}}{\sum_{j=1}^{N} w(|x_{i,t-1} - x_{j,t-1}|; \gamma^0)} \right) \right)^2 \right)$$

and that

$$p \lim_{N,T \to \infty} \left(\nabla^2 Q(\rho, \gamma) \right)^{-1} \tag{75}$$

exists, where

$$\nabla^2 Q(\rho, \gamma) = \begin{pmatrix} \frac{\partial^2 Q}{\partial \rho^2} & \frac{\partial^2 Q}{\partial \rho \partial \gamma} \\ \left(\frac{\partial^2 Q}{\partial \rho \partial \gamma} \right)' & \frac{\partial^2 Q}{\partial \gamma' \partial \gamma} \end{pmatrix}.$$

We prove (73). (74) and (75) follow similarly. We examine $w_{i,t} \epsilon_{i,t}$ where

$$w_{i,t} = \sum_{j=1}^{N} \frac{\rho^0 \frac{\partial w}{\partial \gamma}(|x_{i,t-1} - x_{j,t-1}|; \gamma^0) x_{j,t-1}}{\sum_{j=1}^{N} w(|x_{i,t-1} - x_{j,t-1}|; \gamma^0)}.$$

By Lemma 8 which implies that $w_{i,t}$ has finite variance, uniformly over i, the fact that $w_{i,t}$ and $\epsilon_{i,t}$ are independent, and the fact that $\epsilon_{i,t}$ has finite variance, uniformly over i, by assumption, it follows that $\{w_{i,t} \epsilon_{i,t}\}_{i=1}^{N}$ is a martingale difference with finite second moments. Therefore,

$w_t = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_{i,t}\epsilon_{i,t}$ has zero mean and finite second moments for all N. By the independence of $\epsilon_{i,t}$ across t, it follows that $\{w_t\}_{t=1}^{T}$ is a martingale difference sequence. Hence, a martingale difference CLT holds for w_t proving (73).

Proof of Theorem 4

Define

$$J_{T,N}(x,P) = \Pr_P\left\{NT\left(\hat{r} - r^0\right) \leq x\right\}. \tag{76}$$

Denote by $J(x,P)$ the limit of $J_{T,N}(x,P)$ as $N,T \to \infty$.The subsampling approximation to $J(x,P)$ is given by $L_{b_T,b_N}(x)$. For x_α, where $J(x_\alpha,P) = \alpha$, we need to prove that

$$L_{b_T,b_N}(x_\alpha) \to J(x_\alpha,P)$$

for the theorem to hold. But,

$$E(L_{b_T,b_N}(x_\alpha)) = J_{T,N}(x,P)$$

because as discussed in Section 4, the subsample is a sample from the true model, retaining the temporal ordering of the original sample. Hence, it suffices to show that $Var(L_{b_T,b_N}(x_\alpha)) \to 0$ as $N,T \to \infty$. Let

$$1_{b_T,b_N,s} = 1\left\{b_N b_T\left(\hat{r}^{*,(s)} - \hat{r}\right) \leq x_\alpha\right\}, \tag{77}$$

$$v_{B,h} = \frac{1}{B}\sum_{s=1}^{B} Cov\left(1_{b_T,b_N,s}, 1_{b_T,b_N,s+h}\right). \tag{78}$$

Then,

$$Var\left(L_{b_T,b_N}(x_\alpha)\right) = \frac{1}{B}\left(v_{B,0} + 2\sum_{h=1}^{B} v_{B,h}\right) = \tag{79}$$

$$\frac{1}{B}\left(v_{B,0} + 2\sum_{h=1}^{Cb_T-1} v_{B,h}\right) + \frac{2}{B}\sum_{h=Cb_T}^{B} v_{B,h} = V_1 + V_2.$$

for some $C > 1$. We first determine the order of magnitude of V_1. By the boundedness of $1_{b_T,b_N,s}$, it follows that $v_{B,h}$ is uniformly bounded across h. Hence, $|V_1| \leq \frac{Cb_T}{B}\max_h |v_{B,h}|$, from which it follows that $V_1 = O(Cb_T/B) = o(1)$. We next examine V_2. For this we have that

$$|V_2| \leq \frac{2}{B}\sum_{h=Cb_T}^{B-1} |v_{B,h}|, \tag{80}$$

But, by Lemma 1, it follows that

$$v_{B,h} = o(1), \text{ uniformly across } h. \tag{81}$$

Note that this follows by the geometric ergodicity and, hence β-mixing of the process. Further, note that (81) follows for any random selection of cross sectional units undertaken to construct the subsamples. Hence,

$$\frac{2}{B}\sum_{h=Cb_T}^{B-1}|v_{B,h}| = o(1),$$

proving the convergence of $L_{b_T,b_N}(x_\alpha)$ to $J(x_\alpha, P)$.