Neglecting Parameter Changes in Autoregressive Models

Eric Hillebrand

Louisiana State University*

April 16, 2004

Abstract

We study situations in which autoregressive models are estimated on time series that contain switches in the data generating parameters and these switches are not accounted for. The geometry of this estimation problem causes estimated vector autoregressive models to display a unit eigenvalue, and the sum of the estimated autoregressive parameters of ARMA and GARCH models to be close to one. This artefact is a confounding factor in the analysis of persistence. If the existence of parameter changes in a time series cannot be ruled out, autoregressive models are an inadequate research tool to capture the dynamics of the series. Data must be analyzed for possible change-points before the sample period for an autoregressive model can be specified.

JEL codes: C22, C51

Keywords: spurious unit roots, persistence, ARMA, VAR, GARCH

*Department of Economics, 2126 CEBA Bldg., Baton Rouge, LA 70803, phone: (225) 578-3795, fax (225) 578-3807, erhil@lsu.edu

1 Introduction: Estimation of Autoregressive Parameters in the Presence of Neglected Parameter Changes

Many economic time series are subject to structural changes. These changes can be modeled as switches in the parameter vector of the time series model, which imply shifts in the level of the series, or as shifts in an exogenous level process. While the literature on the detection of these shifts has grown substantially in recent years (for example, Andrews 1993, Bai 1994 and 1997, Bai and Perron 1998, Kokoszka and Leipus 1999, Altissimo and Corradi 2003), the literature on the effects of neglected parameter changes on the parameter estimates is sparser.

Perron (1989) shows that if a Dickey-Fuller test is carried out on a series that contains shifts in the level or in the trend, the estimate of the first-order autoregressive coefficient converges to one. Chen and Tiao (1990) show that the estimate of the first-order sample autocorrelation coefficient will converge to one if the time series was generated by an ARMA process plus a level process that undergoes changes. Perron (1990) shows that if a first-order autoregressive model is estimated on data that contains a shift in a level process, the slope coefficient converges to one. Hendry and Neale (1991) demonstrate that in the presence of structural breaks, unit root tests will too often fail to reject the null of a unit root.

The autoregressive parameters are a measure of the persistence of a time series. In the GARCH literature, Diebold (1986) conjectures that parameter changes may cause the sum of the estimated autoregressive parameters to take values close to one, indicating high volatility persistence when in fact the persistence within regimes of constant parameters is low. Simulation and real data evidence for this conjecture is provided by Lamoureux and Lastrapes (1990), Hamilton and Susmel (1994), and Francq et al. (2001). Mikosch and Star-

ica (2000) consider the Whittle estimator of the ARMA(1,1) representation of GARCH(1,1) and demonstrate that the autoregressive coefficient will be estimated close to one if there are neglected change-points in the data. The relation between structural breaks and spurious estimation of high persistence is also discussed in the literature on long memory models (for example, Lobato and Savin 1998, Granger and Hyung 1999, Diebold and Inoue 2001, Granger and Teräsvirta 2001).

The situation is, therefore, that we have evidence from real and from simulated data that neglected changes in the parameters of an autoregressive time series cause an overestimation of the autoregressive parameters in the proximity of a unit root. We also have analytical results for the first-order autoregressive model and for the first-order sample autocorrelation coefficient for the case of an exogenous level process that undergoes changes. Recently, Hillebrand (2004) has provided an argument how neglected changes in the data-generating parameters of a GARCH(1,1) process, which imply changes in the level of volatility, cause the convergence of the sum of the estimated autoregressive parameters to unity.

This paper shows that this argument describes in fact a general phenomenon in the estimation of all autoregressive models. We provide an encompassing theory of parameter estimation of autoregressive models in the presence of neglected parameter changes, including ARMA, VAR, and GARCH specifications. The central result is that changes in the parameters of an autoregressive process, if not accounted for in the estimation, result in an estimated sum of autoregressive parameters close to one. In the case of vector autoregressions, the largest eigenvalue of the sum of the estimated autoregressive coefficient matrices converges to one in modulus. This result is a consequence of the geometry of the estimation problem. Therefore, it is not restricted to particular estimation methods, changes in specific parameters, or to specific change-point structures (single vs. multiple or deterministic vs. stochastic). In the presence of parameter changes

that are not accounted for, autoregressive models are incapable of capturing the dynamics of the series correctly. They will indicate high persistence even though the persistence within segments of constant parameters may be low.

The outline of the paper is as follows. Section 2 reviews the AR(1) model to provide intuition of the situation. Section 3 provides the result that neglected parameter changes in VAR models lead to a unit eigenvalue in the sum of the estimated autoregressive matrices. Section 4 shows that this result applies to ARMA(p,q) models. Section 5 covers GARCH(p,q) models. The contribution of this paper in the area of GARCH beyond Hillebrand (2004) is that we show the close connection to other autoregressive models. In fact, we derive the main result about GARCH as a corollary to the VAR case. Further, we extend the proof from GARCH(1,1) to the GARCH(p,q) case. Section 6 provides simulation evidence for VAR and ARMA models. We show that the convergence stated in Sections 3 and 4 substantially distorts the estimation of persistence in finite samples and for realistic parameter values. Section 7 discusses a common class of change-point detectors as a possible remedy for the problem and relates them to the theory developed here. Section 8 concludes.

2 First-Order Autoregressive Models

For illustration, consider the first-order autoregressive model. The sample period runs from 1 to T. Assume that there is a single parameter change occurring at time T_1, where the constant changes from c_1 to c_2:

$$x_t = \begin{cases} c_1 + \phi x_{t-1} + \varepsilon_t, & t = 1, \ldots, T_1 \\ c_2 + \phi x_{t-1} + \varepsilon_t, & t = T_1 + 1, \ldots, T, \end{cases} \tag{1}$$

where $c_1, c_2 \in \mathbb{R}$, $c_1 \neq c_2$, $\phi \in (-1, 1)$, and ε_t is white noise. The segment lengths T_1 and $T - T_1$ be large enough in a sense that will be made precise later.

If the change-point T_1 is known, the two segments are estimated separately

by OLS and the slope ϕ is captured consistently. Contrary to that, if the change-point is unknown and the model

$$x_t = c + \phi x_{t-1} + \varepsilon_t \tag{2}$$

is estimated on the entire sample that was generated by (1), ϕ will be estimated close to one. This effect has a simple geometric reason. In the (x_{t-1}, x_t)-plane, the points of each of the two segments cluster around the means $\mu_1 := c_1/(1-\phi)$ and $\mu_2 := c_2/(1-\phi)$. Given that the segment lengths are not too short, the sample mean of the series $\{x_t\}$, $t = 1, \ldots, T$ is approximately the same as the sample mean of the series $\{x_{t-1}, t = 2, \ldots, T\}$. Therefore, if we plot the x_t against the x_{t-1}, we find two clusters centered at two different points (μ_1, μ_1) and (μ_2, μ_2) on the identity line. Thus, given that μ_1 and μ_2 are sufficiently different, the estimator $\hat{\phi}$ of the slope in model (2) will pick up the slope of the identity, not the in-segment dynamics ϕ of the data generating process (1). Figure 1 illustrates the situation.

FIGURE 1 ABOUT HERE

This effect is very general. It is not restricted to a change in the intercept c but also occurs in the case of a change in ϕ, as this leads to different means $\mu_1 = c/(1-\phi_1)$ and $\mu_2 = c/(1-\phi_2)$. Likewise, the effect is not restricted to a single change-point. If there are several parameter changes, say $k-1$, they will induce k different means $\mu_1, \mu_2, \ldots, \mu_k$, which all lie on the identity line in the (x_{t-1}, x_t)-plane. Ignoring the changes and estimating (2) on the entire sample will force the estimated line to go through all these means, thereby exhibiting slope equal to one. Also, as it is a geometrical phenomenon, it is not confined to a particular estimation method. The only assumption we will have to make is that the asymptotic variance of the estimator vanishes with growing sample size. This includes all common estimation methods for autoregressive models.

3 Vector Autoregressive Models

Consider the first-order vector autoregression

$$x_t = c + \Phi x_{t-1} + \varepsilon_t, \tag{3}$$

where $c, x_t \in \mathbb{R}^N$, $\Phi \in \mathbb{R}^{N\times N}$, ε_t is N-dimensional white noise and the sample size is T.

Assume that there are $k-1$ points in the sample where the parameters change (k regimes)

$$x_t^{(i)} = c_i + \Phi_i x_{t-1}^{(i)} + \varepsilon_t,\ t = T_{i-1}+1, \ldots, T_i, \tag{4}$$

for $i = 1, \ldots, k$, $T_0 = 0$ and $T_k = T$. All eigenvalues of the Φ_i matrices are inside the unit circle. Within each regime, the data points will cluster around

$$\mu_i := (I - \Phi_i)^{-1} c_i.$$

Denote by $E_i(\cdot) := \mathbb{E}(\cdot | x_{T_{i-1}})$ the expectation conditional on the initial value of segment i, and define $\mathrm{Var}_i(\cdot)$, $\mathrm{cov}_i(\cdot,\cdot)$ analogously. We will often use that the expected value of x_t, $t = T_{i-1}+1, \ldots, T_i$, conditional on the initial value in segment i has the representation

$$\begin{aligned}\mathbb{E}_i x_t &= (I-\Phi_i)^{-1}(I - \Phi_i^{t-T_{i-1}})c_i + \mathbb{E}_i[\sum_{j=0}^{t-T_{i-1}-1} \Phi_i^j \varepsilon_{t-j}] + \Phi_i^{t-T_{i-1}} x_{T_{i-1}} \\ &= \mu_i + O(\Phi_i^{t-T_{i-1}}),\end{aligned} \tag{5}$$

where $O(\Phi_i^{t-T_{i-1}})$ is a deterministic term that vanishes with growing $t - T_{i-1}$, $t < T_i$, as the eigenvalues of Φ_i are all inside the unit circle.

If the changing regimes are not accounted for and the model (3) is estimated on the entire sample $\{x_1, x_2, \ldots, x_T\}$, an effect analogous to the AR(1) case occurs: The geometrical fact that the different segments i imply different means μ_i that lie on the identity of the $(N \times N)$-dimensional space (x_{t-1}, x_t) leads to

the phenomenon that the global estimate $\hat{\Phi}$ will have a unit eigenvalue. This holds true provided that the segment lengths are not too short and the μ_i are sufficiently distinct.

In order to show this, a technical lemma is necessary:

Lemma 1. *Let x_t be a time series that has a representation* (5)*. Then, the global sample mean $\bar{x} = \sum_{t=1}^{T} x_t/T$ has the representation*

$$\bar{x} = \frac{1}{T}\sum_{j=1}^{k}(T_i - T_{i-1})\mu_i + o(1)_T.$$

Here, $o(1)_T$ denotes a term that vanishes as the sample size T becomes large. It is assumed that as T becomes large, all segment sizes $T_i - T_{i-1}$ become large.

The proof is provided in the Appendix. In order for the main result of this section to hold, it is sufficient that the sizes of at least two of the segments grow with T to infinity.

In the case of VAR models, ordinary least squares estimation is equivalent to maximum likelihood or generalized method of moments estimation and the estimates are asymptotically normally distributed:

$$\sqrt{T}(\hat{\Phi} - \Phi) \overset{T\to\infty}{\sim} \mathcal{N}(0, \Sigma), \tag{6}$$

where Σ is a symmetric positive definite matrix.

We will use this property of standard estimators in a slightly different way. Consider the covariance of the estimator $\hat{\Phi}$ with a single observation x_t and apply the Cauchy-Schwarz inequality

$$\operatorname{cov}(\hat{\Phi}, x_t) = \mathbb{E}\left[(\hat{\Phi} - \mathbb{E}\hat{\Phi})(x_t - \mathbb{E}x_t)\right] \leq \sqrt{\operatorname{Var}(\hat{\Phi})\operatorname{Var}(x_t)}, \tag{7}$$

The second moment of a covariance-stationary VAR process x_t is finite. From (6) we have that the variance $\operatorname{Var}(\hat{\Phi})$ of the estimator, which is the diagonal of Σ divided by T, vanishes with order T. Therefore, the property (6) translates to

$$\operatorname{cov}(\hat{\Phi}, x_t) = o(1)_T. \tag{8}$$

In other words, the influence of a single observation on the estimator $\hat{\Phi}$ vanishes with growing sample size.

Together with Lemma 1, we can state the main result for vector autoregressions.

Theorem 1. *If a first-order vector autoregressive model* (3) *is estimated on a time series that has a representation* (5) *and that underwent* $(k-1)$ *parameter changes, the matrix* $\mathbb{E}_i(\hat{\Phi}) = \mathbb{E}(\hat{\Phi}|x_{T_{i-1}})$ *has a unit eigenvalue with corresponding eigenvector* $e_i := \mu_i - 1/T \sum_{j=1}^{k}(T_j - T_{j-1})\mu_j$. *This holds true in all segments* $i = 1, 2, \ldots, k$, *up to terms that vanish with growing segment sizes* $T_i - T_{i-1}$.

The proof is provided in the Appendix.

Consider the VAR(p) model

$$\tilde{x_t} = \tilde{c} + \Phi_1 \tilde{x}_{t-1} + \Phi_2 \tilde{x}_{t-2} + \ldots + \Phi_p \tilde{x}_{t-p} + \tilde{\varepsilon}_t, \tag{9}$$

where $\tilde{x}_t, \tilde{c}, \tilde{\varepsilon}_t \in \mathbb{R}^N$ and $\Phi_i \in \mathbb{R}^{N \times N}$. The unconditional mean of this process is $\tilde{\mu} = (1 - \sum_{j=1}^{p} \Phi_j)^{-1}\tilde{c}$. The model has the VAR(1) representation

$$x_t = c + M x_{t-1} + \varepsilon_t, \tag{10}$$

where

$$x_t = \begin{bmatrix} \tilde{x}_t \\ \tilde{x}_{t-1} \\ \vdots \\ \tilde{x}_{t-p+1} \end{bmatrix}, \quad c = \begin{bmatrix} \tilde{c} \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad M = \begin{bmatrix} \Phi_1 & \Phi_2 & \ldots & \Phi_{p-1} & \Phi_p \\ I & 0 & \ldots & 0 & 0 \\ 0 & I & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & I & 0 \end{bmatrix}, \quad \varepsilon_t = \begin{bmatrix} \tilde{\varepsilon}_t \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

and I is the $N \times N$ identity matrix. Consider the case in which the process underwent $(k-1)$ parameter changes:

$$x_t^{(i)} = c_i + M_i x_{t-1}^{(i)} + \varepsilon_t, \; t = T_{i-1} + 1, \ldots, T_i, \tag{11}$$

for $i = 1, \ldots, k$; $T_0 = 0$, and $T_k = T$. From Theorem 1 we have the following corollary.

Corollary 1. *Consider a VAR(p) model that is estimated on a time series that underwent* $(k-1)$ *parameter changes, and these changes are not accounted for. Then, by the VAR(1) representation* (10) *and thereby representation* (5)*, where* M *replaces* Φ*, Theorem 1 applies. In that case, the matrix* $\mathbb{E}_i(\hat{\Phi}_1 + \hat{\Phi}_2 + \ldots + \hat{\Phi}_p)$ *has an asymptotic unit eigenvalue with corresponding eigenvector* $\tilde{e}_i = \tilde{\mu}_i - 1/T \sum_{j=1}^{k} (T_j - T_{j-1})\tilde{\mu}_j$*.*

The statement is proved in the Appendix.

4 Autoregressive Moving Average Models

Consider the ARMA(p,q) model

$$\tilde{x}_t = \tilde{c} + \phi_1 \tilde{x}_{t-1} + \ldots + \phi_p \tilde{x}_{t-p} + \epsilon_t + \psi_1 \epsilon_{t-1} + \ldots + \psi_q \epsilon_{t-q}, \tag{12}$$

or

$$\Phi(L)\tilde{x}_t = \tilde{c} + \Psi(L)\epsilon_t,$$

where ϵ is white noise, $\tilde{c} \in \mathbb{R}$, and $\Phi(L)$ and $\Psi(L)$ have roots outside the unit circle. Write (12) in vector notation

$$x_t = c + \Phi x_{t-1} + \Psi \varepsilon_t, \tag{13}$$

where $x_t = (\tilde{x}_t, \tilde{x}_{t-1}, \ldots, \tilde{x}_{t-m+1})'$, $c = (\tilde{c}, 0, \ldots, 0)' \in \mathbb{R}^m$, $\varepsilon_t = (\epsilon_t, \epsilon_{t-1}, \ldots, \epsilon_{t-m})' \in \mathbb{R}^{m+1}$, and $m = \max\{p, q\}$. The matrix Φ of autoregressive coefficients is given by

$$\Phi = \begin{bmatrix} \phi_1 & \phi_2 & \ldots & \phi_{p-1} & \phi_p & [0] \\ 1 & 0 & \ldots & 0 & 0 & [0] \\ 0 & 1 & \ldots & 0 & 0 & [0] \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 & [0] \\ [0] & [0] & \ldots & [0] & [1] & [0] \end{bmatrix} \in \mathbb{R}^{m \times m}.$$

The matrix Ψ of moving average coefficients is given by

$$\Psi = \begin{bmatrix} \psi_0 & \psi_1 & \dots & \psi_q & [0] \\ 0 & 0 & \dots & 0 & [0] \\ 0 & 0 & \dots & 0 & [0] \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & [0] \\ [0] & [0] & \dots & [0] & [0] \end{bmatrix} \in \mathbb{R}^{m+1 \times m+1},$$

where $\psi_0 = 1$. The symbols [0] and [1] mean that in the respective cases where $q < m$ or $p < m$, the matrix is filled with zeros and lower diagonal ones in the case of Φ. Stack the intercept c and the autoregressive coefficients into a parameter vector $\theta = (c, \text{vech}\Phi)$ and define[1]

$$\mu(\theta) := (I - \Phi)^{-1} c = \frac{\tilde{c}}{1 - \phi_1 - \ldots - \phi_p}.$$

Assume that there are $k-1$ points where θ changes in the data-generating process:

$$x_t^{(i)} = c_i + \Phi_i x_{t-1}^{(i)} + \Psi \varepsilon_t, \; t = T_{i-1} + 1, \ldots, T_i, \tag{14}$$

for $i = 1, \ldots, k$ where $T_0 = 0$ and $T_k = T$. Within each regime, the data points will cluster around

$$\mu_i := \mu(\theta_i) = \frac{c_i}{1 - \phi_{1,i} - \ldots - \phi_{p,i}}.$$

The expected value of x_t conditional on the initial value in segment i is given by

$$\begin{aligned} \mathbb{E}_i(x_t) &= (I - \Phi_i)^{-1}(I - \Phi_i^{t-T_{i-1}})c_i + \Psi \mathbb{E}_i \left[\sum_{j=0}^{t-T_{i-1}-1} \Phi_i^j \varepsilon_{t-j} \right] + \Phi_i^{t-T_{i-1}} x_{T_{i-1}} \\ &= \mu_i + O(\Phi_i^{t-T_{i-1}}), \; t = T_{i-1} + 1, \ldots, T_i, \end{aligned}$$

same as in (5). Therefore, Lemma 1 applies to processes generated by (14). Common estimators of ARMA models have an asymptotic distribution with a

[1] If ϕ_j is the j-th column of the matrix Φ, the column vector vechΦ is given by $(\phi_1', \phi_2', \ldots, \phi_n')'$.

symmetric positive definite covariance matrix that vanishes with growing sample size (e.g., Brockwell and Davis 1991, p 258 for the MLE, and Harris 1999 for alternative GMM estimators). Therefore, (8) applies and the following corollary is proven.

Corollary 2. *If an ARMA model* (13) *is estimated on a time series that has a representation* (5) *and that underwent* $(k-1)$ *parameter changes, the matrix* $\mathbb{E}_i(\hat{\Phi}) = \mathbb{E}(\hat{\Phi}|x_{T_{i-1}})$ *has a unit eigenvalue with corresponding eigenvector* $e_i := \mu_i - 1/T \sum_{i=1}^{k}(T_i - T_{i-1})\mu_i$. *This holds true in all segments* $i = 1, 2, \ldots, k$, *up to terms that vanish with growing segment sizes* $T_i - T_{i-1}$.

This result shows that in the presence of neglected parameter changes, ARMA models exhibit the same artefact error in the estimation of persistence as VAR models. Before we proceed to give an illustrating example, we note that Theorem 1 and Corollary 2 provide a new way to prove the result of Chen and Tiao (1990). They show that for an ARMA model with random level shifts, the k-th sample autocorrelation coefficient

$$\hat{\rho}(k) = \frac{\sum_{t=1}^{T-k} y_t y_{t+k}}{\sum_{t=1}^{T} y_t^2} \longrightarrow_{T\to\infty} 1 \text{ in probability.}$$

The ordinary least squares estimator $\hat{\phi}_k$ of the simple regression of the centered series

$$x_t - \bar{x} = \phi_k(x_{t-k} - \bar{x}) + \varepsilon_t,$$

where $\bar{x} = \sum_{t=1}^{T} x_t/T$, is identical to $\hat{\rho}(k)$ for $y_t = x_t - \bar{x}$. Theorem 1 and Corollary 2 show that for the case $k = 1$, $\hat{\phi}_1$ converges to one with increasing segment sizes. It is straightforward to extend the proof of Theorem 1 to the case of a general lag k, because only the terms involving the initial values of the segments will change. The estimator $\hat{\phi}_k$ still captures the unity slope of the identity.

We consider an example that illustrates Corollary 2. Assume that we have a data set that we assume to be generated by a simple ARMA(1,1) process

$$x_t + \phi x_{t-1} = c + \epsilon_t + \psi \epsilon_{t-1}.$$

In fact, it is generated by the process

$$x_t - 0.3x_{t-1} = \begin{cases} 0.05 + \epsilon_t + 0.1\epsilon_{t-1} & \text{for } t \in \{1, \ldots, T_1\} \\ 0.10 + \epsilon_t + 0.1\epsilon_{t-1} & \text{for } t \in \{T_1 + 1, \ldots, T_2\} \\ 0.15 + \epsilon_t + 0.1\epsilon_{t-1} & \text{for } t \in \{T_2 + 1, \ldots, T\}, \end{cases} \tag{15}$$

where $\epsilon_t \sim \mathcal{N}(0, 1\text{e-}4)$, $T_1 = 3000, T_2 = 6000, T = 9000$. The situation is shown in Figure 2. The data are centered at zero by subtracting the global sample mean. Again, the alignment of the clusters along the identity of the subspace (x_{t-1}, x_t) causes the estimate of the autoregressive coefficient to pick up slope one. The three small hyperplanes correspond to the estimations when the correct segmentation is considered and when an ARMA(1,1) model is estimated on each segment. The large hyperplane is the result of a global fit of a single ARMA(1,1) model to the whole data set. The locally estimated persistence parameters are all of the magnitude $\hat{\phi}_{local} \approx 0.30$ whereas the globally estimated slope is $\hat{\phi}_{global} = 0.98$.

FIGURE 2 ABOUT HERE.

5 Generalized Autoregressive Conditional Heteroskedasticity

Consider the GARCH(p,q) model. Let r_t be the returns of a financial instrument, then

$$\begin{aligned} r_t &= \mathbb{E}(r_t|\mathcal{F}_{t-1}) + \epsilon_t = f(b) + \epsilon_t \; t = 1, \ldots, T, \\ \epsilon_t|\mathcal{F}_{t-1} &\sim \mathcal{N}(0, \mathtt{h}_t), \\ \mathtt{h}_t &= \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \mathtt{h}_{t-i}, \end{aligned} \tag{16}$$

where $f(b)$ is some conditional mean function with parameters b, for example a linear model $x_t'b$ with exogeneous variables x_t, and ϵ_t is a conditionally heteroskedastic disturbance. The parameters of the conditional variance function are $\omega \in \mathbb{R}$, $\alpha_i \in (0,1)$, and $\beta_i \in (0,1)$ such that $\sum^q \alpha_i + \sum^p \beta_i \leq 1$. In the case where equality holds, we have the integrated GARCH or IGARCH(p,q) model.

The conditional variance equation in (16) can be written in vector form

$$h_t = \omega + A\varepsilon_{t-1}^2 + Bh_{t-1}, \tag{17}$$

where $h_t = (\mathtt{h}_t, \mathtt{h}_{t-1}, \ldots, \mathtt{h}_{t-m+1})'$ and $m = \max\{p, q\}$, $\varepsilon_t^2 = (\epsilon_t^2, \epsilon_{t-1}^2, \ldots, \epsilon_{t-m+1}^2)'$, and $\omega = (\alpha_0, 0, \ldots, 0)'$. The coefficient matrices A and B are given by

$$A = \begin{bmatrix} \alpha_1 & \alpha_2 & \ldots & \alpha_q & [0] \\ 0 & 0 & \ldots & 0 & [0] \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & [0] \\ [0] & [0] & \ldots & [0] & [0] \end{bmatrix} \in \mathbb{R}^{m \times m},$$

$$B = \begin{bmatrix} \beta_1 & \beta_2 & \dots & \beta_{p-1} & \beta_p & [0] \\ 1 & 0 & \dots & 0 & 0 & [0] \\ 0 & 1 & \dots & 0 & 0 & [0] \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 & [0] \\ [0] & [0] & \dots & [0] & [1] & [0] \end{bmatrix} \in \mathbb{R}^{m \times m}.$$

The symbols [0] and [1] mean that in the respective cases where $q < m$ or $p < m$, the matrix is filled with zeros and, in the case of B, lower diagonal ones until it has the dimension $m \times m$.

Unlike the case of ARMA(p,q) models, there are no observations of h_t and ε_t available, the r_t are the only data. Therefore, in order to evaluate the log-likelihood

$$L(\theta) = -\frac{T}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^{T} \log \mathtt{h}_t - \frac{1}{2} \sum_{t=1}^{T} \frac{\epsilon_t^2}{\mathtt{h}_t},$$

the residual ε_t and the conditional variance $\mathtt{h}_t$ must be estimated from an initial guess for the parameter vector

$$\theta := (b, \alpha_0, \alpha_1, \alpha_2, \dots, \alpha_q, \beta_1, \beta_2, \dots, \beta_p)'$$

and then updated at every iteration step of the optimization. If we had direct observations, this would clearly be better information about the market's volatility. For the purposes of this analysis, we will assume that there are direct observations of h_t and ε_t. We will show that when there are neglected changes in the variance parameters of θ in the data-generating process, the sum $\sum^q \hat{\alpha}_i + \sum^p \hat{\beta}_i$ will be equal to one in the limit, because one of the eigenvalues of the matrix $\hat{\Phi} = \hat{A} + \hat{B}$ will be one in modulus, up to terms that vanish with growing segment lengths. We make the conjecture that in the case where the h_t have to be estimated, the same effect will occur. Hillebrand (2004) provides simulation evidence for the validity of this conjecture.

Consider the case where there are $k-1$ points in time where the variance parameters in the data-generating θ change:

$$h_t(\theta_i) = \omega_i + A_i \varepsilon_{t-1}(\theta_i)^2 + B_i h_{t-1}(\theta_i),\ t = T_{i-1}+1, \ldots, T_i, \tag{18}$$

where $i = 1, \ldots, k$ and setting $T_0 = 0$ and $T_k = T$. The parameter vector within segments is denoted θ_i. It contains the parameters b of the conditional mean equation, which do not change from segment to segment, and the segment-dependent parameters of the conditional variance $\alpha_{0,i}$, $\alpha_{1,i}$, $\ldots$, $\alpha_{q,i}$, $\beta_{1,i}$, $\ldots$, $\beta_{p,i}$. We will establish that within each parameter regime, the h_t cluster around

$$\mu_i := \mathbb{E}h_t(\theta_i) = (1 - A_i - B_i)^{-1}\omega_i = \left(\frac{\alpha_{0,i}}{1-\lambda_i}, 0, \ldots, 0\right)' \in \mathbb{R}^m, \tag{19}$$

where $\lambda_i = \sum_{m=1}^{q} \alpha_{m,i} + \sum_{n=1}^{p} \beta_{n,i}$.

This mean-changing structure of the data-generating process is not accounted for in the estimation. The estimated variance model is

$$h_t = \hat{\omega} + \hat{A}\varepsilon_{t-1}^2 + \hat{B}h_{t-1}. \tag{20}$$

Let $\mathbb{E}_i h_t$ denote the expected value with respect to the start value in segment $i = 1, \ldots, k$:

$$\mathbb{E}_i h_t := \mathbb{E}(h_t | \mathcal{F}_{T_{i-1}}).$$

In order to show that Theorem 1 applies to GARCH, we need

1. a representation of $\mathbb{E}_i h_t$ and $\mathbb{E}_i \varepsilon_t^2$ as in (5), that is, as the unconditional mean plus vanishing terms;

2. the applicability of Lemma 1 to h_t and ε_t^2, that is, a representation of the global sample means as a weighted average of in-segment means plus vanishing terms;

3. an equivalent of property 8, that is, an asymptotic distribution of the estimator with symmetric positive definite covariance matrix that vanishes with growing sample size.

The following lemma provides the representation (5) for GARCH.

Lemma 2. *The expected values* $\mathbb{E}_i h_t$ *and* $\mathbb{E}_i \varepsilon_t^2$ *of a stationary GARCH(p,q) model conditional on the initial value* $h_{T_{i-1}}$ *have the representation*

$$\mathbb{E}_i \varepsilon_t^2 = \mathbb{E}_i h_t = \mu_i + O(\Phi_i^{t-T_i}), \tag{21}$$

where $\Phi_i = A_i + B_i$, $t \in \{T_{i-1}+1, \ldots, T_i\}$, *and* $\mu_i = (1 - A_i - B_i)^{-1}\omega_i$.

The proof is provided in the Appendix. With this representation, Lemma 1 applies to the sample mean of the conditional variance process $\bar{h}$ and to the sample mean of the squared error process $\overline{\varepsilon_t^2}$.

$$\begin{aligned} \bar{h} &= \frac{1}{T}\sum_{i=1}^{k}(T_i - T_{i-1})\mu_i + o(1)_T, \\ \overline{\varepsilon^2} &= \frac{1}{T}\sum_{i=1}^{k}(T_i - T_{i-1})\mu_i + o(1)_T. \end{aligned} \tag{22}$$

The analogue to property (8) for GARCH models is as follows.

Assumption 1. *The influence of a single realization of the processes* ε_t^2 *and* h_t *on the estimator* $\hat{\theta}$ *vanishes with growing segment size:*

$$\begin{aligned} cov_i(\hat{\theta}, \varepsilon_t^2) &= cov(\hat{\theta}, \varepsilon_t^2)|_{\mathcal{F}_{T_{i-1}}} = o(1)_{T_i - T_{i-1}} \,\forall\, t \\ cov_i(\hat{\theta}, h_t) &= cov(\hat{\theta}, h_t)|_{\mathcal{F}_{T_{i-1}}} = o(1)_{T_i - T_{i-1}} \,\forall\, t. \end{aligned}$$

Applying the Cauchy-Schwarz inequality as in the derivation of (8), we obtain that the assumption is tantamount to assuming that the product of the variances of the estimator $\hat{\theta}$ and of ε_t^2 and h_t, respectively, vanishes with the sample size. The variances of ε_t^2 and h_t are finite (Bollerslev 1986), so in order for the assumption to hold, the variance of $\hat{\theta}$ must vanish with growing sample size.

For instance, the asymptotic distribution of the maximum likelihood estimator $\hat{\theta}$ of the GARCH(1,1) parameters $\theta = (\mu, \alpha_0, \alpha_1, \beta)^T$ is given by

$$\sqrt{T}(\hat{\theta} - \theta) \sim_{T\to\infty} \mathcal{N}(0, \Sigma)$$

where Σ is a symmetric positive definite matrix that arises from the outer product of the likelihood score (Weiss 1986, Bollerslev and Wooldridge 1992, Lumsdaine 1996). Hence, the variance of $\hat{\theta}$ vanishes with order T and the assumption is satisfied. For GARCH(p,q), however, no asymptotic distribution theory is available yet. Therefore, the property is stated as an assumption here. The conjecture that the estimator behaves as in the GARCH(1,1) case is commonly made and most software packages return t-statistics as if it did.

We have all prerequisites to show the following corollary of Theorem 1.

Corollary 3. *If a GARCH model* (20) *is estimated on a time series that is generated by* (18) *and that underwent* $(k-1)$ *parameter changes, the matrix* $\mathbb{E}_i(\hat{\Phi}) = \mathbb{E}(\hat{A} + \hat{B}|\mathcal{F}_{T_{i-1}})$ *has a unit eigenvalue with corresponding eigenvector* $e_i := \mu_i - 1/T \sum_{j=1}^{k}(T_j - T_{j-1})\mu_j$. *This holds true in all segments* $i = 1, 2, \ldots, k$, *up to terms that vanish with growing segment sizes* $T_i - T_{i-1}$.

The proof is provided in the Appendix.

Corollary 4. *It follows immediately that in the case where* $\mathbb{E}_i\hat{\Phi}$ *has a unit eigenvalue, the characteristic polynomial*

$$z^m - \mathbb{E}_i(\hat{\alpha}_1 + \hat{\beta}_1)z^{m-1} - \mathbb{E}_i(\hat{\alpha}_2 + \hat{\beta}_2)z^{m-2} - \ldots - \mathbb{E}_i(\hat{\alpha}_m + \hat{\beta}_m) = 0$$

has a unit root and therefore

$$\mathbb{E}_i\hat{\lambda} = \mathbb{E}_i\left(\sum_{l=1}^{m}\hat{\alpha}_l + \sum_{j=1}^{m}\hat{\beta}_j\right) = 1.$$

Figure 3 shows a synthetic GARCH(1,1) series with a single change in α_0. The two different data-generating parameter vectors induce two distinct expected values $\mu_1 = \mathbb{E}h_t(\theta_1)$ and $\mu_2 = \mathbb{E}h_t(\theta_2)$. The spheres in Figure 3 are centered at these expected values. The data points $(h_{t-1}, \varepsilon_{t-1}^2, h_t)$ of the segments cluster around these respective means. The clusters exhibit slopes in both subspaces, reflecting the data-generating α_1 in the $(\varepsilon_{t-1}^2, h_t)$-subspace and the data-generating β in

the $(h_{t-1},\, h_t)$-subspace. These slopes cannot be captured by the single estimation hyperplane that has to go through both segments. The relative position of the two means dominates.

FIGURE 3 ABOUT HERE.

As the mean of the $\{h_t\}$ and the mean of the $\{h_{t-1}\}$ is equal for sufficiently long segments, a line connecting two different means in the $(h_t,\, h_{t-1})$-subspace has slope equal to one. Therefore, β will be estimated close to one. The remaining autoregressive parameter $\hat{\alpha}_1$ is chosen residually such that $\hat{\alpha}_1 + \hat{\beta} < 1$ as the estimated process $\hat{h}_t$ would explode otherwise.

FIGURE 4 ABOUT HERE.

In real estimation problems, the ε_t and the h_t cannot be observed but have to be estimated along with the parameters: $\hat{\varepsilon}_t = \hat{\varepsilon}_t(\hat{b})$ and $\hat{h}_t = \hat{h}_t(\hat{\alpha}_0,\, \hat{\alpha}_1,\, \hat{\beta})$. Figures 4 and 5 show the estimated data points $(\hat{h}_{t-1}(\hat{\theta}),\, \hat{\varepsilon}^2_{t-1}(\hat{\theta}),\, \hat{h}_t(\hat{\theta}))$ for the same synthetic series shown in Figure 3. The estimation hyperplane is the same as in Figure 3. By construction, all points lie on the estimation hyperplane. However, the two-cluster structure is still visible, the data points still cluster around the unconditional means. The figures show that the phenomenon carries over to the case where h is unobservable.

6 Simulations

We simulate parameter changes in VAR and ARMA models in this section. It is demonstrated that the convergence analyzed in Sections 3 and 4 causes severe distortions in the estimates of autoregressive parameters in finite samples. For simulations of the GARCH model, we refer to Hillebrand (2004).

6.1 VAR Simulations

To explore the implications of Theorem 1, we conduct an experiment motivated by the VAR study of Bernanke and Mihov (1998). To distinguish between different monetary policy target hypotheses, Bernanke and Mihov consider a structural VAR system for total reserves and non-borrowed reserves, the federal funds rate, real GDP, the GDP deflator, and the Dow Jones index of spot commodity prices. All data are monthly and the sample ranges from 1965:1 through 1996:12.[2] Besides the entire sample, the authors consider the sub-samples 1965:1–1979:9 (11), 1979:10–1996:12 (12), 1984:2–1996:12 (7), and 1988:9–1996:12 (11). The numbers in parentheses are the highest lags that the authors find to be significant, the lag order for the entire sample is 13.

We estimate the reduced-form VAR model

$$x_t = c + \sum_{j=1}^{p} \Phi_j x_{t-j} + \varepsilon_t, \tag{23}$$

where $x_t, c, \varepsilon_t \in \mathbb{R}^6$ and $\Phi_j \in \mathbb{R}^{6\times 6}$. The lag order and the ordering of the variables in x is as listed above. According to Corollary 1, we consider the largest eigenvalue of the matrix $\sum_{j=1}^{p} \hat{\Phi}_j$. For the entire sample and all sub-samples, the largest eigenvalue is of the order of 1 in modulus (between 0.996 and 1.095), except for the 1988:9–1996:12 sample, where it is estimated at 3.24. The estimated processes are all non-stationary or almost non-stationary. This result is not sensitive to the lag order of the VARs, the largest eigenvalues are close to one for lower lag orders as well. This finding does not indicate that neglected parameter changes are in the data, even though they may contribute. As the VAR is specified in levels, however, it is more likely that genuine unit roots in the data are picked up by the VARs.

We transform the data to obtain stationary estimated VAR processes. For non-borrowed reserves, total reserves, real GDP, and the Dow Jones index of

[2]Ben Bernanke kindly provides the data set on his web site.

spot commodities prices, we use log-differences. For the federal funds rate and the GDP deflator, we use first differences. We scale all variables so that they are of the order of magnitude 0.01. The Akaike information criterion favors a VAR(3) specification while the Bayes information criterion favors VAR(1). For both specifications and all samples, the largest eigenvalue of $\sum_{j=1}^{p} \hat{\Phi}_j$ is well below one. For example, for the entire sample and the VAR(1) specification it is 0.402 and for the VAR(3) specification it is 0.504.

We set up a simulation experiment to study the effects of change-points in the data. The Dow Jones commodity index price series shows a change-point in the 1970's that is caused by the oil price shock. Before 1974, the series oscillates around a mean of 100; after 1978, it oscillates around a mean of 250. We take this regime changing behavior as a prototype for a synthetic time series that we add to the data mentioned above. For 108 observations, corresponding to 1965:1 through 1973:12, we generate white noise with mean 0.01 and standard deviation 0.001. We construct an equidistant grid of 11 points between 0.01 and 0.025 for the mean of the second segment corresponding to 1979:1 through 1996:12. Call this grid $\mu_2(k)$, $k = 1, \ldots, 11$, where $\mu_2(1) = 0.01$ and $\mu_2(11) = 0.025$. The observations of the second segment are then white noise with mean $\mu_2(k)$ and standard deviation 0.001. This gives eleven different jump size scenarios k, including zero, the stationary case. For the 60 observations corresponding to 1974:1 through 1978:12, we construct a smooth transition from $\mu_1 = 0.01$ to $\mu_2(k)$ plus zero-mean white noise with the same standard deviation of 0.001, so that there is no discontinuous jump in the series.

For every jump size scenario, we generate 10,000 runs. In each run, a synthetic time series of the type described above is simulated and added to the set of differenced time series of real data. Then, a VAR(1) and a VAR(3) specification are estimated on the data set including the synthetic series. As the stochastic process that underlies the synthetic time series is white noise, the lag structure

of the VAR will change within the segments of constant means. We store the modulus of the largest eigenvalue of $\sum_{j=1}^{p} \hat{\Phi}_j$. Thus, in every jump size scenario, the experiment results in two series of 10,000 observations each, one for the VAR(1) case and one for the VAR(3) case.

FIGURE 6 ABOUT HERE.

Figure 6 shows the means and the two-sided 95-percent quantiles of the modulus of the largest eigenvalue of $\sum_{j=1}^{p} \hat{\Phi}_j$ for the 11 jump size scenarios. The estimated largest eigenvalue grows to one with increasing jump size. For the jump size of 0.01 to 0.025, which was obtained from the level series of the Dow Jones commodity price index (divided by 10,000), the result is almost a unit root. As the largest eigenvalue is well below one for the real data set (VAR(1): 0.402, VAR(3): 0.504), this result is caused by the synthetic time series. As the stochastic process of the synthetic series is white noise, the estimated almost unit root is an artefact that is caused by the deterministic change in the mean of the white noise process. This illustrates the point of Theorem 1.

6.2 ARMA Simulations

For the ARMA case, simulation experiments in three persistence environments will be carried out: low, medium, and high persistence. The data-generating process is ARMA(2,2); all model orders up to ARMA(2,2) will be estimated on the data.

The basic element of the experiments is an ARMA(2,2) time series of 5000 observations. The generator is

$$x_t - \phi_1 x_{t-1} - \phi_2 x_{t-2} = \left\{ \begin{array}{c} c_1 \\ c_2 \end{array} \right\} + \eta_t + \psi_1 \eta_{t-1} + \psi_2 \eta_{t-2}, \tag{24}$$

where c_1 holds for the first 2500 observations and c_2 for the second 2500 observations. The respective persistence environment defines the values of ϕ_1 and ϕ_2

according to Table 1. The values of ψ_1 and ψ_2 are fixed at 0.20 and 0.70 in all experiments and environments. The constant c_1 is fixed at 1e-5 in all experiments and environments.

In each environment, ten experiments are carried out. The value of c_2 is set according to Table 2, causing a jump in the constant of the data-generating process. Starting from c_2 = 1e-5 (no jump) in the first experiment, c_2 is increased linearly to 1e-2. For every experiment, we generate 10,000 series of (24) and estimate ARMA(2,2), ARMA(1,1), ARMA(1,2), and ARMA(2,1) on it.

TABLES 1 and 2 ABOUT HERE.

Figures 7 through 9 present the results of the 10 experiments in every persistence environment. In all environments, the mean of the estimate of the sum of the autoregressive parameters grows close to one with increasing jump size in the constant.

The solid line in each of the Figures 7 through 9 shows the simulation sample means $\overline{\hat{\phi}_1 + \hat{\phi}_2}$ of the sum of the estimates of the autoregressive parameters of the ARMA(2,2) specification. The error bars are the 0.95 quantiles of these estimations.

As the estimates $\overline{\hat{\phi}_1 + \hat{\phi}_2}$ grow to one with increasing jump size in the constant, the 0.95 quantiles decrease monotonically in all three environments. They are smallest in the high persistence environment and largest in the low persistence environment. In the high and medium persistence environments, the estimation of an ARMA (2,1) model gives almost the same result as the estimation of ARMA(2,2). As is expected, ARMA(1,1) and ARMA(1,2) estimates are upward biased because of the misspecification of the autoregressive lag structure (see in particular the first experiment in each environment, where there is no jump). Again, the low persistence environment is an exception, as ARMA(2,2) is downward biased in the first experiment, ARMA(2,1) appears unbiased, and

ARMA(1,1) and ARMA(1,2) are upward biased.

Regardless of bias and misspecification, the effect under study dominates with increasing jump size as the distance in the respective means of the segments grows larger. This illustrates the point of Corollary 2.

7 Change-Point Detection

From a practical perspective, what can be done to avoid the error in the estimation that stems from parameter change-points? Before the data sample for an autoregressive model can be specified, a change-point detection study has to be carried out. A comprehensive review of the literature on change-point detection is beyond the scope of this study. However, we will briefly discuss a detection method that is of particular interest in our context.

Bai (1994, 1997) proposes the following test statistic to detect parameter changes:

$$S_k^2 := \sum_{t=1}^{k}(y_t - \bar{y}_k)^2 + \sum_{t=k+1}^{T}(y_t - \bar{y}_k^*)^2,$$

where k is the hypothetical change-point, $\bar{y}_k = \sum_{t=1}^{k} y_t/k$ is the sample average in the first segment implied by k and $\bar{y}_k^* = \sum_{t=k+1}^{T} y_t/(T-k)$ is the sample mean in the second segment. Then, $\hat{k} = \text{argmin}_k(S_k^2)$ is the estimator of a single change-point in the series. The detector can be applied sequentially to cover the multiple-change-points case. It can be shown that the estimator can be defined equivalently as $\hat{k} = \text{argmax}_k|V_k|$, where

$$V_k = \frac{\sqrt{k(T-k)}}{T}(\bar{y}_k^* - \bar{y}_k).$$

This statistic is a re-scaled measure of the distance between the two segment sample means implied by a hypothetical change-point k. The change-point is estimated where this difference is maximal. Denote $c_k = \sqrt{k(T-k)}/T$ and let k^* be a single parameter change-point. Neglecting terms that correct for

initial conditions and assuming for simplicity that $\mu_2 > \mu_1$, it is an elementary observation that

$$\mathbb{E}|V_k| \approx \begin{cases} c_k \frac{k^*}{k}(\mu_2 - \mu_1) & \text{for } k > k^*, \\ c_k \frac{T-k^*}{T-k}(\mu_2 - \mu_1) & \text{for } k < k^*, \\ c_k(\mu_2 - \mu_1) & \text{for } k = k^*. \end{cases}$$

The distorting factors k^*/k and $(T-k^*)/(T-k)$ when the hypothetical change-point k is not equal to the actual change-point k^* are both less than one. Therefore, $\mathbb{E}|V_k|$ is maximized, up to terms that vanish with growing segment sizes, at the actual single change-point k^*.

In the univariate case with a single change-point, the "eigenvectors" e_1 and e_2 in Theorem 1 are scalars and given by

$$\begin{aligned} e_1 &= \mu_1 - \frac{k^*}{T}\mu_1 - \frac{T-k^*}{T}\mu_2 = \frac{T-k^*}{T}(\mu_1 - \mu_2) \\ e_2 &= \mu_2 - \frac{k^*}{T}\mu_1 - \frac{T-k^*}{T}\mu_2 = \frac{k^*}{T}(\mu_2 - \mu_1). \end{aligned}$$

Thus, we see that at the actual change-point k^*, Bai's test statistic has the expected value

$$\mathbb{E}|V_{k^*}| \approx \sqrt{|e_1|\,|e_2|},$$

where we neglected initial value terms.

Intuitively, the difference in the in-sample means causes e_1 and e_2 to be non-zero. In the case where there are no parameter changes, equation (28) in the proof of Theorem 1 reads zero equals zero. Therefore, it is self-suggesting to look at a statistic that searches for the point of maximal distance in the segment sample means.

Kokoszka and Leipus (1999) suggested a similar detector for parameter changes in ARCH models. The detector statistic is defined as

$$V_k = \frac{k(T-k)}{T^2}\left(\frac{1}{k}\sum_{j=1}^{k} r_t^2 - \frac{1}{T-k}\sum_{j=k+1}^{T} r_t^2\right),$$

where r_t is the log return from a financial asset and the estimate of the change-point is given by $\hat{k} = \text{argmax}_k |V_k|$, as before. This statistic considers the distance in the means of the squared return series of a financial instrument. Here, at the actual change-point k^*,

$$\mathbb{E}|V_{k^*}| \approx \frac{\sqrt{k^*(T-k^*)}}{T} \sqrt{|e_{1,1}|\,|e_{2,1}|},$$

where the $e_{i,1}$ are the first entries in the eigenvectors e_i from Corollary 3.

8 Conclusion

We consider situations where autoregressive models are estimated on data that contain unknown switches in the data-generating parameters such that the entire time series has different local means. In this case, the sum of the estimated autoregressive parameters (or the largest eigenvalue of the sum of the estimated autoregressive coefficient matrices) is close to one.

The reason for this error in the estimation is that the local means of the different segments are aligned on the identity hyperplane in the $(x_t, x_{t-1}, x_{t-2}, \ldots)$ space of an autoregressive time series x. The estimators of the autoregressive coefficients take on the unit slope of the identity. The phenomenon is geometric and therefore not confined to a specific estimator. Neither does it depend on a specific stochastic structure of the parameter switches; a single jump can suffice, given that it is large enough.

The apparent unit root indicates high persistence when in fact the persistence within segments of constant parameters may be low. Thus, the estimation error is a confounding factor in persistence analysis. For example, the fact that the sum of the estimated autoregressive parameters of GARCH models sum up to almost one regardless of the financial asset under study was taken as evidence of high persistence in volatility. It is still possible that high persistence arises from

genuine unit roots or fractional integration, for example, and not from structural breaks. The analysis presented here shows, however, that in the presence of neglected structural breaks, autoregressive models are incapable of correctly capturing the persistence. This holds for higher-order autoregressive models in the same way as for simple first-order specifications.

The results also explain earlier findings that in the presence of structural breaks in ARMA processes, the first order sample autocorrelation coefficient converges to one (Chen and Tiao 1990) and unit root tests fail to reject the null hypothesis too often (Perron 1989, Hendry and Neale 1991).

The direct practical implication of the result is that when the existence of parameter changes in the time series under study cannot be ruled out, autoregressive models are an inadequate research tool to capture the dynamics of the series. Therefore, a careful change-point detection study has to be undertaken before autoregressive models can be fitted to a data set.

Acknowledgments

The paper benefited from discussions with and comments from George Papanicolaou, Carter Hill, Atsushi Inoue, Remigijus Leipus, Caio Almeida, Knut Sølna, and the participants of the Statistics Seminar at Stanford. Any remaining errors are mine.

References

Altissimo, F. and V. Corradi, 2003. Strong Rules for Detecting the Number of Breaks in a Time Series. Journal of Econometrics 117: 207–244.

Andrews, D. W. K. 1993. Tests for Parameter Instability and Structural Change with Unknown Change Point. Econometrica 61(4): 821–856.

Bai, J. 1994. Least Squares Estimation of a Shift in Linear Processes. Journal of Time Series Analysis 15(5): 453–472.

Bai, J. 1997. Estimating Multiple Breaks One at a Time. Econometric Theory 13: 315–352.

Bai, J. and P. Perron, 1998. Estimating and Testing Linear Models with Multiple Structural Changes. Econometrica 66(1): 47–78.

Bernanke, B. S. and I. Mihov, 1998. Measuring Monetary Policy. Quarterly Journal of Economics. 869–902.

Bollerslev, T. 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327.

Bollerslev, T., Wooldridge, J.M. 1992. Quasi-maximum likelihood estimation and inference in dynamic models with time-varying covariances. Econometric Reviews 11, 143–172

Brockwell, P. J. and R. A. Davis, 1991. Time Series: Theory and Methods. Springer: New York.

Chen, C. and Tiao, G.C. 1990. Random Level-Shift Time Series Models, ARIMA Approximations, and Level-Shift Detection. Journal of Business and Economic Statistics 8(1): 83–97.

Diebold, F.X. 1986. Modeling the persistence of conditional variances: A comment. Econometric Reviews 5, 51–56.

Diebold, F.X., Inoue, A. 2001. Long memory and regime switching. Journal of Econometrics 105, 131–159.

Francq, C., Roussignol, M., Zakoïan, J.-M. 2001. Conditional heteroskedasticity driven by hidden Markov chains. Journal of Time Series Analysis 22, 197–220.

Granger, C.W.J., Hyung, N. 1999. Occasional structural breaks and long memory. UC San Diego Discussion Paper 99-14.

Granger, C.W.J., Teräsvirta, T. 2001. A simple nonlinear time series model with misleading linear properties. Economics Letters 62, 161–165.

Hamilton, J.D., Susmel, R. 1994. Autoregressive conditional heteroskedasticity and changes in regime. Journal of Econometrics 64, 307–333.

Harris, D. 1999. GMM Estimation of Time Series Models. In: Matyas, L. (ed.) Generalized Method of Moments Estimation. Cambridge University Press, Cambridge, pp 149–170.

Hendry, D.F. and A.J. Neale 1991. A Monte Carlo study of the effects of structural breaks on tests for unit roots. In: P. Hackl and A.H. Westlund (Eds.) Economic Structural Change, Analysis and Forecasting. pp. 95-119, Berlin: Springer-Verlag.

Hillebrand, E. 2004. Neglecting Parameter Changes in GARCH Models. Forthcoming in the Annals Issue of the Journal of Econometrics on Modeling Structural Breaks, Long Memory and Stock Market Volatility.
`http://www.bus.lsu.edu/economics/faculty/ehillebrand/personal/`

Kokoszka, P. and R. Leipus, 1999. Testing for Parameter Changes in ARCH Models. Lithuanian Mathematical Journal 39(2): 182–195.

Lamoureux, C.G., Lastrapes, W.D. 1990. Persistence in variance, structural change, and the GARCH model. Journal of Business and Economic Statistics 8, 225–234.

Lobato, I.N., Savin, N.E. 1998. Real and spurious long-memory properties of stock-market data. Journal of Business and Economic Statistics 16, 261–268.

Lumsdaine, R. 1996. Consistency and asymptotic normality of the quasi-maximum likelihood estimator in IGARCH(1,1) and covariance stationary GARCH(1,1) models. Econometrica 64, 575–596.

Mikosch, T., Starica, C. 2000. Change of structure in financial time series, long range dependence and the GARCH model. University of Aarhus, Aarhus School of Business, Centre for Analytical Finance, Working Paper No. 58.

Perron, P. 1989. The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis. Econometrica 57(6): 1361–1401.

Perron, P. 1990. Testing for a Unit Root in a Time Series With a Changing Mean. Journal of Business and Economic Statistics 8(2): 153–162.

Weiss, A. A. 1986. Asymptotic theory for ARCH models: Estimation and testing. Econometric Theory 2, 107–131.

Appendix

Proof of Lemma 1. Write $x_t = \mathbb{E}_i x_t + z_t$, $t = T_{i-1}+1, \ldots, T_i$, where z_t is the deviation from the expected value conditional on the initial value of segment i. Then, by the law of large numbers,

$$\frac{1}{T_i - T_{i-1}} \sum_{T_{i-1}+1}^{T_i} z_t = o(1)_{T_i - T_{i-1}}.$$

The sample mean can be written as

$$\begin{aligned}
\bar{x} &= \frac{1}{T}\sum_{t=1}^{T} x_t, \\
&= \frac{1}{T}\sum_{t=T_0+1}^{T_1} \mathbb{E}_{(1)} x_t + \frac{1}{T}\sum_{t=T_0+1}^{T_1} z_t \\
&\quad + \ldots + \frac{1}{T}\sum_{t=T_{k-1}+1}^{T_k} \mathbb{E}_{(k)} x_t + \frac{1}{T}\sum_{t=T_{k-1}+1}^{T_k} z_t, \\
&= \frac{1}{T}\sum_{t=T_0+1}^{T_1} \mathbb{E}_{(1)} x_t + \ldots + \frac{1}{T}\sum_{t=T_{k-1}+1}^{T_k} \mathbb{E}_{(k)} x_t + o(1)_T,
\end{aligned}$$

given that as T becomes large, all segment sizes $T_i - T_{i-1}$ become large. With (5), we obtain

$$\begin{aligned}
\bar{x} = &\frac{T_1 - T_0}{T}\mu_1 + \frac{T_2 - T_1}{T}\mu_2 + \ldots + \frac{T_k - T_{k-1}}{T}\mu_k \\
&+ \frac{1}{T}\sum_{i=1}^{k}\sum_{t=T_{i-1}+1}^{T_i} O(\Phi_i^{t-T_{i-1}}) + o(1)_T.
\end{aligned}$$

Observing that $1/T \sum_{i=1}^{k} \sum_{t=T_{i-1}+1}^{T_i} O(\Phi_i^{t-T_{i-1}}) = o(1)_T$, Lemma 1 is proven. □

Proof of Theorem 1. Consider observation $x_t^{(i)}$ in segment i. For notational brevity, the superscript i on x_t, $t = T_{i-1}+1, \ldots, T_i$, will be suppressed. Subtract the global sample mean $\bar{x} = 1/T \sum_{t=1}^{T} x_t = \hat{c} + \hat{\Phi}\bar{x}$ and take expectations conditional on the initial value of the segment:

$$\mathbb{E}_i(x_t - \bar{x}) = \mathbb{E}_i(\hat{\Phi}(x_{t-1} - \bar{x})). \tag{25}$$

At this point, the idea of the proof becomes clear already. The common property of the estimator $\hat{\Phi}$ stated in (8) together with the representation of $\bar{x}$ in Lemma 1 as a deterministic sum plus vanishing terms will allow us to decompose the expected value of the product on the right-hand side into the product of the expected values plus vanishing terms. As $\mathbb{E}_i(x_t - \bar{x}) \approx \mathbb{E}_i(x_{t-1} - \bar{x}) \neq 0$, there must be eigenvectors of $\mathbb{E}_i\hat{\Phi}$ corresponding to the eigenvalue one.

Hence, apply (8) to the right-hand side of (25):

$$\mathbb{E}_i(\hat{\Phi}(x_{t-1} - \bar{x})) = \mathbb{E}_i\hat{\Phi}\mathbb{E}_i x_{t-1} + o(1)_{T_i - T_{i-1}} - \mathbb{E}_i(\Phi\bar{x}).$$

Apply Lemma 1 using the decomposition $x_t = \mathbb{E}_i x_t + z_t$ employed in its proof.

$$\begin{aligned}\mathbb{E}_i(\hat{\Phi}(x_{t-1} - \bar{x})) = \\ \mathbb{E}_i\hat{\Phi}\mathbb{E}_i x_{t-1} + o(1)_{T_i - T_{i-1}} - \frac{1}{T}\sum_{j=1}^{k}(T_j - T_{j-1})\mu_j\mathbb{E}_i\hat{\Phi} \\ -\frac{1}{T}\sum_{j=1}^{k}\sum_{t=T_{j-1}+1}^{T_j}\mathbb{E}_i(\hat{\Phi}z_t) - \frac{1}{T}\sum_{j=1}^{k}\sum_{t=T_{j-1}+1}^{T_j}\mathbb{E}_i(\hat{\Phi}\Phi_j^{t-T_{j-1}}x_{T_{j-1}}) \qquad (26)\end{aligned}$$

From (8) we have that

$$\mathrm{cov}_i(\hat{\Phi}, x_t) = \mathrm{cov}_i(\hat{\Phi}, z_t) = o(1)_{T_i - T_{i-1}},$$

so that $1/T\sum_{j=1}^{k}\sum_{t=T_{j-1}+1}^{T_j}\mathbb{E}_i(\hat{\Phi}z_t) = o(1)_T$. The influence $\Phi_i^{t-T_{i-1}}x_{T_{i-1}}$ of the initial value of the segment on x_t is deterministic and therefore $\mathbb{E}_i(\hat{\Phi}\Phi_i^{t-T_{i-1}}x_{T_{i-1}}) = \mathbb{E}_i(\hat{\Phi})\Phi_i^{t-T_{i-1}}x_{T_{i-1}}$. Thus,

$$\frac{\mathbb{E}_i(\hat{\Phi})}{T}\sum_{j=1}^{k}\sum_{t=T_{j-1}+1}^{T_j}\Phi_j^{t-T_{j-1}}x_{T_{j-1}} = O(1/T),$$

as all eigenvalues of the Φ_j are inside the unit circle and the segment sizes $T_j - T_{j-1}$ grow as T grows. Therefore, we have so far that

$$\begin{aligned}\mathbb{E}_i(x_t - \bar{x}) = \mathbb{E}_i(\hat{\Phi}(x_{t-1} - \bar{x})) = \\ \mathbb{E}_i\hat{\Phi}\left[\mathbb{E}_i x_{t-1} - \frac{1}{T}\sum_{j=1}^{k}(T_j - T_{j-1})\mu_j\right] + o(1)_{T_i - T_{i-1}} + o(1)_T + O(1/T). \qquad (27)\end{aligned}$$

Applying Lemma 1 to the left-hand side of (25), we get

$$\mathbb{E}_i(x_t - \bar{x}) = \mathbb{E}_i x_t - \frac{1}{T}\sum_{j=1}^{k}(T_j - T_{j-1})\mu_j + o(1)_T.$$

The only difference to the expression in brackets on the right-hand side of (27) is that we take conditional expectations of the lag of x_t there. From (5), we have

$$\mathbb{E}_i x_{t-1} = \mu_i + O(\Phi_i^{t-T_{i-1}-1})$$
$$\mathbb{E}_i x_t = \mu_i + O(\Phi_i^{t-T_{i-1}}).$$

Using this and plugging in from (27), (25) becomes

$$\left[\mu_i - \frac{1}{T}\sum_{j=1}^{k}(T_j - T_{j-1})\mu_j\right] = \mathbb{E}_i\hat{\Phi}\left[\mu_i - \frac{1}{T}\sum_{j=1}^{k}(T_j - T_{j-1})\mu_j\right] + O(\Phi_i^{t-T_{i-1}}) + O(\Phi_i^{t-T_{i-1}-1}) + o(1)_{T_i - T_{i-1}} + o(1)_T + O(1/T), \quad (28)$$

or

$$e_i = \mathbb{E}_i(\hat{\Phi})e_i + O(\Phi_i^{t-T_{i-1}-1}) + o(1)_{T_i-T_{i-1}} + o(1)_T + O(1/T).$$

The vector e_i does not vanish for all $i = 1, \ldots, k$ as long as at least two of the μ_j are not equal, that is, there is at least one parameter change. Therefore, e_i is an eigenvector of $\mathbb{E}_i\hat{\Phi}$ corresponding to the eigenvalue one, provided that the sample size T and the segment size $T_i - T_{i-1}$ are large enough. □

Proof of Corollary 1. From Theorem 1 we know that if the regime changing structure is not accounted for in the estimation, the matrix $\mathbb{E}_i\hat{M}$ has an asymptotic unit eigenvalue with corresponding eigenvector $e_i = \mu_i - 1/T\sum_{j=1}^{k}(T_j - T_{j-1})\mu_j$, where

$$\mu_i = (I - M_i)^{-1}c_i = \begin{bmatrix} (I - \sum_{j=1}^{p}\Phi_{i,j})^{-1}\tilde{c}_i \\ (I - \sum_{j=1}^{p}\Phi_{i,j})^{-1}\tilde{c}_i \\ \vdots \\ (I - \sum_{j=1}^{p}\Phi_{i,j})^{-1}\tilde{c}_i \end{bmatrix} = \begin{bmatrix} \tilde{\mu}_i \\ \tilde{\mu}_i \\ \vdots \\ \tilde{\mu}_i \end{bmatrix}.$$

Then, the first N rows of the equation $e_i = \mathbb{E}_i(\hat{M})e_i + o(1)_T$ read

$$\tilde{e}_i = \tilde{\mu}_i - \frac{1}{T}\sum_{j=1}^{k}(T_j - T_{j-1})\tilde{\mu}_j = \mathbb{E}_i(\sum_{j=1}^{p}\hat{\Phi}_j)\tilde{e}_i + o(1)_T,$$

and Corollary 1 is proven. □

Proof of Lemma 2. The expected value of (18) conditional on the start value of the segment is given by

$$\begin{aligned}\mathbb{E}_i h_t &= \omega_i + A_i(\mathbb{E}_i\eta_{t-1}^2 \times \mathbb{E}_i h_{t-1}) + B_i\mathbb{E}_i h_{t-1} \\ &= \omega_i + \Phi_i\mathbb{E}_i h_{t-1},\end{aligned}$$

where $\Phi_i = A_i + B_i$, η_t is an m-vector of standard normal random variables and "$\times$" denotes element-wise multiplication. The GARCH(p,q) model is stationary if and only if the roots of

$$z^m - (\alpha_{1,i} + \beta_{1,i})z^{m-1} - (\alpha_{2,i} + \beta_{2,i})z^{m-2} - \ldots - (\alpha_{m,i} + \beta_{m,i}) = 0,$$

lie inside the unit circle. The roots are the eigenvalues of Φ_i. Here, $\alpha_j = 0$ for $j > q$ in the case where $m > q$ or $\beta_j = 0$ for $j > p$ in the case where $m > p$.

From the stationarity of the GARCH process it follows therefore that $1 - \Phi_i$ is invertible and

$$\begin{aligned}\mathbb{E}_i h_t &= (1 - \Phi_i)^{-1}(1 - \Phi_i^{t-T_{i-1}})\omega_i + \Phi_i^{t-T_{i-1}}h_{T_{i-1}} \\ &= \mu_i + \Phi_i^{t-T_{i-1}}(h_{T_{i-1}} - \mu_i) = \mu_i + O(\Phi_i^{t-T_{i-1}}).\end{aligned}$$

The conditional term is the autoregressive matrix operating on the distance of the initial value from the unconditional mean. It vanishes with growing sample size. □

Proof of Corollary 3. Subtract the global sample mean from (20):

$$h_t - \bar{h} = \hat{A}(\varepsilon_{t-1}^2 - \overline{\varepsilon^2}) + \hat{B}(h_{t-1} - \bar{h}). \tag{29}$$

Apply the representation in Lemma 2, Assumption 1, and $E_i\varepsilon_t^2 = \mathbb{E}_i h_t$ from the distribution assumption in (16) to obtain

$$\begin{aligned}\mathbb{E}_i(h_t - \bar{h}) &= \mathbb{E}_i\left[\hat{A}(\varepsilon_{t-1}^2 - \overline{\varepsilon^2}) + \hat{B}(h_{t-1} - \bar{h})\right] \\ &= \mathbb{E}_i(\hat{A} + \hat{B})\,\mathbb{E}_i\left(h_{t-1} - \frac{1}{T}\sum_{i=1}^{k}(T_i - T_{i-1})\mu_i\right) + o(1)_T + o(1)_{T_i - T_{i-1}},\end{aligned}$$

where the $o(1)_T$-term stems from the application of Lemma 2 and the $o(1)_{T_i - T_{i-1}}$-term from the application of Assumption 1. This is the equivalent of equation (27) in the proof of Theorem 1 with h_t replacing x_t and $\Phi = A + B$. The rest of the proof proceeds in exactly the same way as the proof of Theorem 1. □

Table 1: Values of the autoregressive parameters of the data generating ARMA(2,2) processes in the three persistence environments. The moving average parameters were $\psi_1 = 0.20$ and $\psi_2 = 0.70$ in all three environments.

persistence:	high	low	medium
ϕ_1	0.50	0.20	0.20
ϕ_2	0.45	0.10	0.65
$\phi_1 + \phi_2$	0.95	0.30	0.85

Table 2: Ten jump sizes of the intercept of the data generating ARMA(2,2) processes in all three persistence environments.

experiment	1	2	3	4	5	6	7	8	9	10
c	1e-5	1e-3	2e-3	3e-3	4e-3	6e-3	7e-3	8e-3	9e-3	1e-2

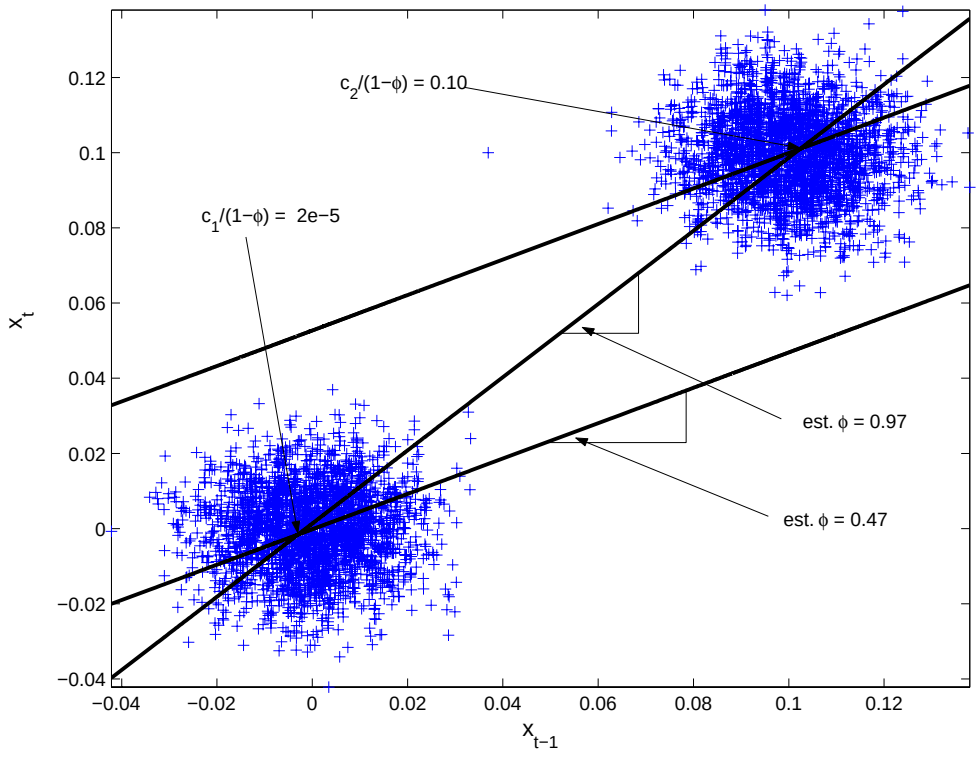

Figure 1: Graph of data generated from model (1) with parameters $\phi = 0.50$, $\sigma = 0.01$, $c_1 =$ 1e-5, $c_2 = 0.05$, $N = 5000$, and $N_1 = 2500$. The points cluster around the unconditional in-segment means 2e-5 and 0.10. If an autoregressive model is fitted globally to the whole data set, the estimated slope will be close to one, the slope of the identity.

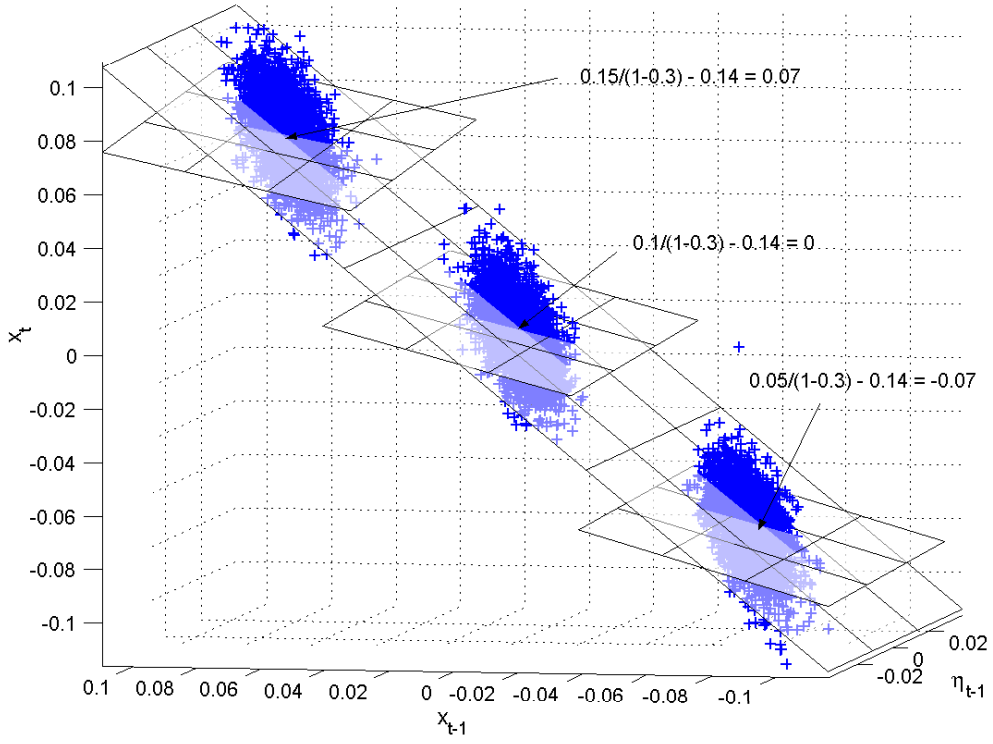

Figure 2: Data generated from model (15) and plotted in the $(x_{t-1}, \epsilon_{t-1}, x_t)$ space. The data were centered by subtracting the global sample mean of 0.14. The three separate hyperplanes are the estimation hyperplanes when the data are correctly segmented and three separate ARMA(1,1) models are estimated. The large hyperplane with slope close to one in the (x_{t-1}, x_t)-subspace is the estimation hyperplane when the parameter switches are ignored and a single ARMA(1,1) model is estimated on the entire time series. Exactly as in Figure 1 for the AR(1) case, the alignment of the segment clusters along the identity line in the (x_{t-1}, x_t)-subspace causes the global estimate of the autoregressive parameter ϕ to take a value close to one.

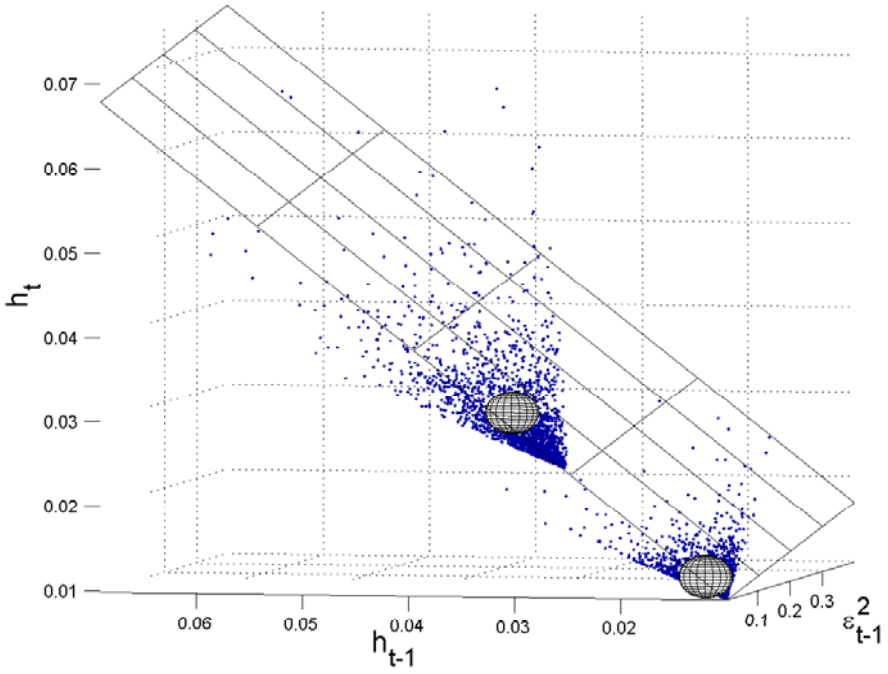

Figure 3: Plot of the data points $(h_{t-1}, \varepsilon^2_{t-1}, h_t)$ for an annualized synthetic GARCH(1,1) series with a single change-point in α_0. The $\{\varepsilon_t\}$ and $\{h_t\}$ are generated by the parameters $\alpha_{0,1}$ =2e-5 and $\alpha_{0,2}$ =5e-5, $\alpha_1 = 0.10$ and $\beta = 0.50$. The length of the entire series is $T = 4200$ and the changepoint T_1 is set at one half of T. The spheres are centered at the unconditional, stationary expected values $\mathbb{E}h_t^{(1)} = 250 * 2\text{e-}5/(1 - 0.1 - 0.5) = 0.0125$ and $\mathbb{E}h_t^{(2)} = 250 * 5\text{e-}5/(1-0.1-0.5) = 0.03125$. The fact that a single hyperplane is fitted through both segments, reflected by the two point clusters, leads to a sum of estimated autoregressive parameters close to one. It can be seen that the slope of the clusters with respect to the (h_{t-1}, h_t)-subspace, which is $\beta = 0.5$, is largely overestimated. The estimator $\hat{\beta}$ picks up the slope of the identity. The slope of the clusters with respect to the $(\varepsilon^2_{t-1}, h_t)$-subspace, which is $\alpha_1 = 0.1$, is underestimated. The estimated parameters are $\hat{\alpha}_0 = 2.6\text{e-}5$, $\hat{\alpha}_1 = 0.018$, and $\hat{\beta} = 0.981$. The estimation hyperplane was obtained by standard quasi-maximum likelihood GARCH estimation.

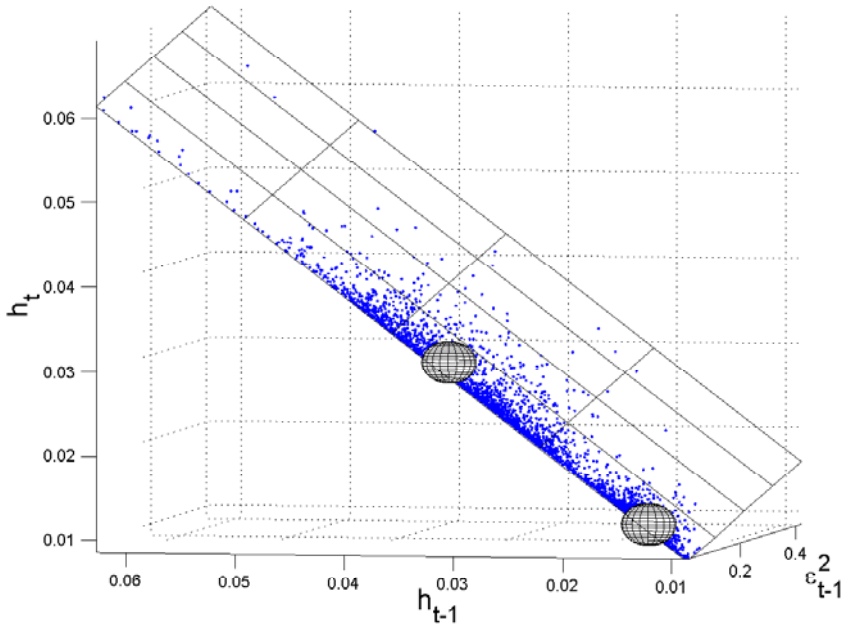

Figure 4: Plots of the estimated data points $(\hat{h}_{t-1}(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\beta}), \hat{\varepsilon}^2_{t-1}(\hat{\mu}), \hat{h}_t(\hat{\alpha}_0, \hat{\alpha}_1, \hat{\beta}))$, for the same synthetic data series considered in Figure 3. By construction, all the points are lying on the hyperplane according to the estimates $\hat{\alpha}_0 = 2.6\text{e-}5$, $\hat{\alpha}_1 = 0.018$, and $\hat{\beta} = 0.981$. However, the two-cluster structure is still visible. The geometry of the situation is similar, with point clusters around the unconditional means.

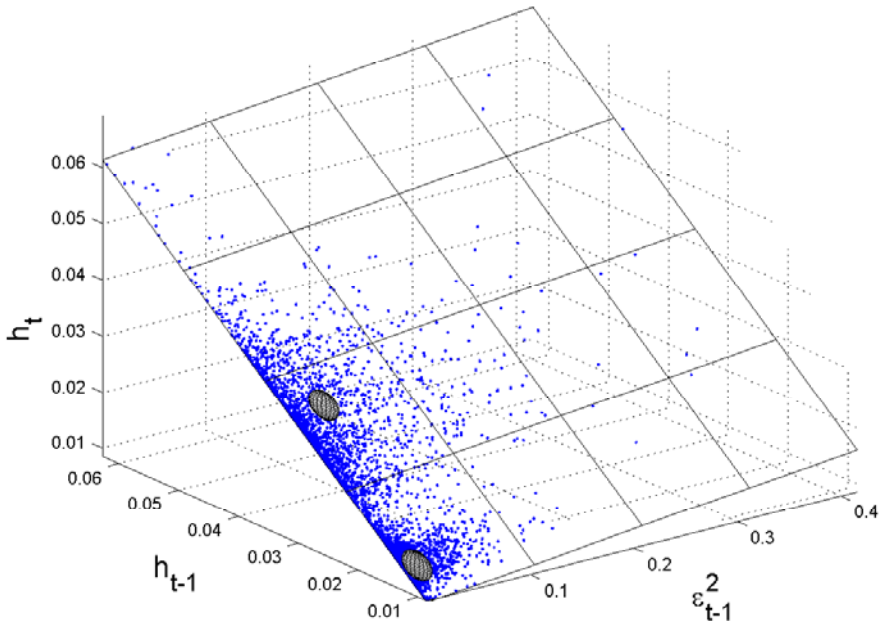

Figure 5: Same situation as in Figure 4 from a different viewpoint than in Figures 3 and 4. From this viewpoint, the two clusters can be seen more distinctly.

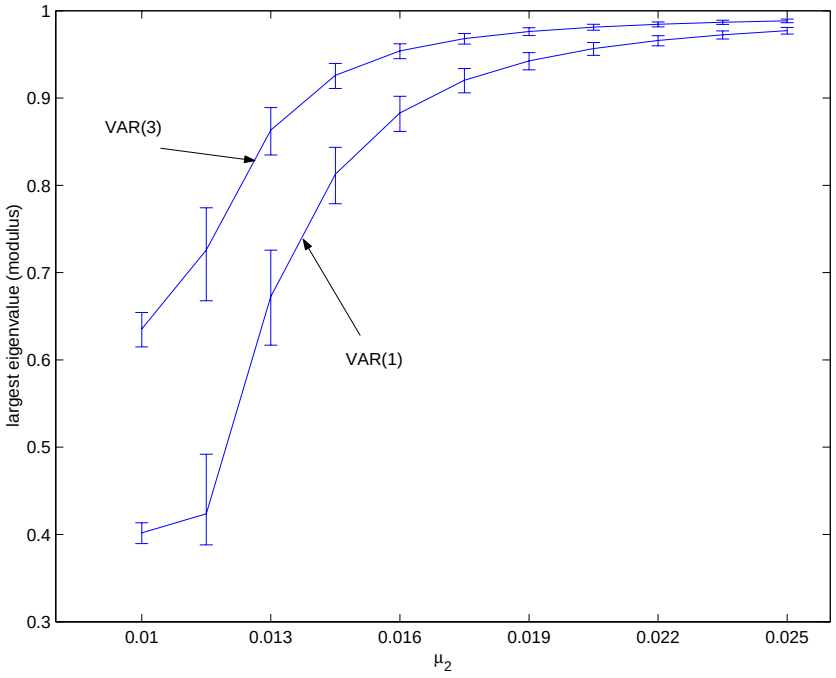

Figure 6: The simulation object is a VAR of 7 variables. Six variables are from the Bernanke and Mihov (1998) data, the seventh is synthetic white noise with a jump in the mean. The ordinate shows the modulus of the largest eigenvalue of the sum of the estimated autoregressive coefficient matrices $\sum_{j=1}^{p} \hat{\Phi}_j$. The curves plot the means and 95 percent quantiles of 10,000 simulations for each of 11 different jump sizes scenarios (abscissa). The largest eigenvalue grows to one in modulus with growing jump size.

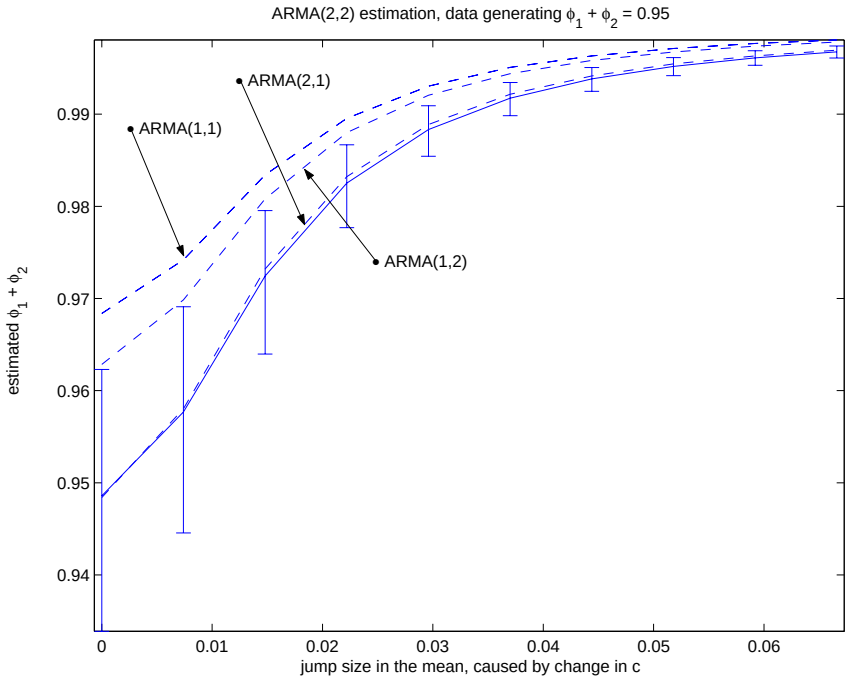

Figure 7: Plot of the simulation results in the high persistence environment according to Tables 1 and 2. The solid line shows the mean of the sum of the autoregressive parameters of the ARMA(2,2) model over the 10000 simulation runs for each jump size. The error bars are the 95 percent quantiles of the estimator. The dashed lines give the means of the sum of the autoregressive parameters when other model specifications are estimated on the simulated data. It is clearly visible for all model orders that the larger the jump size, the closer the sum of the estimated autoregressive parameters is pushed towards one.

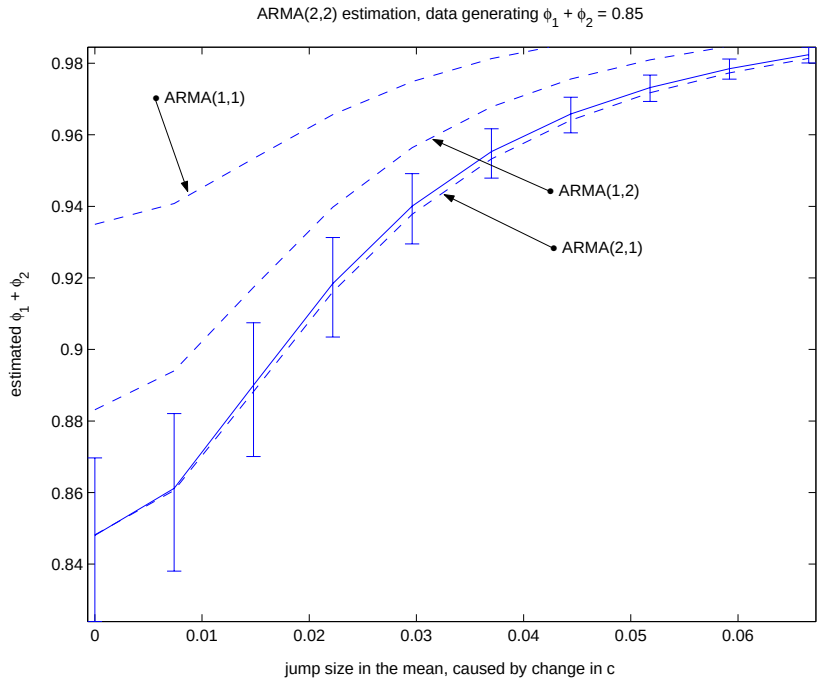

Figure 8: Plot of the simulation results in the medium persistence environment according to Tables 1 and 2. The solid line shows the mean of the sum of the autoregressive parameters of the ARMA(2,2) model over the 10000 simulation runs for each jump size. The error bars are the 95 percent quantiles of the estimator. The dashed lines give the means of the sum of the autoregressive parameters when other model specifications are estimated on the simulated data. The estimated mean of the sum of the autoregressive parameters increases towards one with increasing jump size.

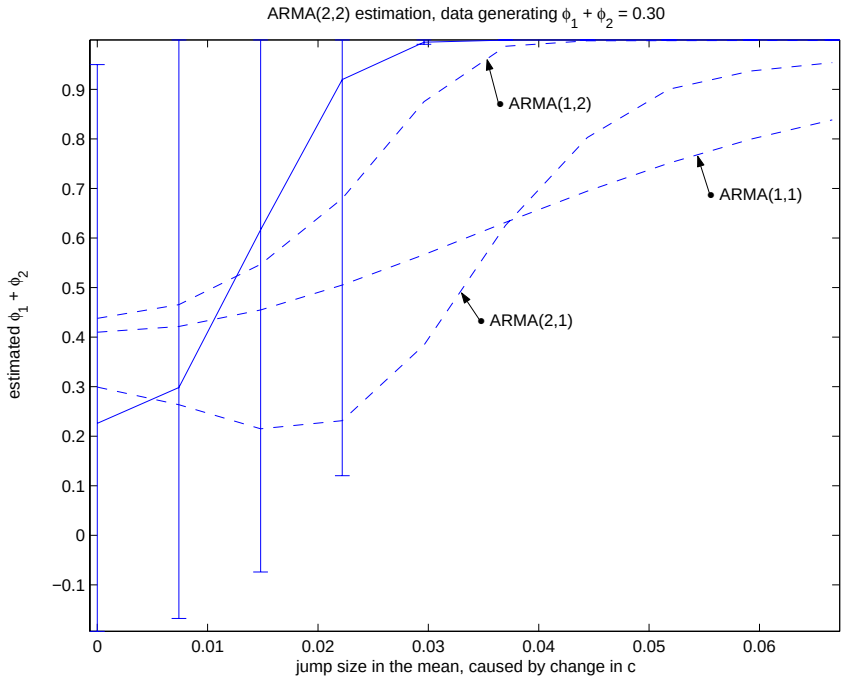

Figure 9: Plot of the simulation results in the low persistence environment according to Tables 1 and 2. The solid line shows the mean of the sum of the autoregressive parameters of the ARMA(2,2) model over the 10000 simulation runs for each jump size. The error bars are the 95 percent quantiles of the estimation. The dashed lines give the means of the sum of the autoregressive parameters when other model specifications are estimated on the simulated data. The estimated mean of the sum of the autoregressive parameters increases towards one with increasing jump size.