Feasible tests for regime switching in autoregressive models

Hajo Holzmann[1] **and Florian Ketterer**

Fachbereich Mathematik und Informatik, Philipps-Universität Marburg, Germany.

Abstract

Autoregressive models with switching regime form a popular and very flexible class of non-linear time-series models, which have been widely applied in finance, econometrics, engineering and other fields. In a first step, a basic yet hard problem is to test whether a regime switch is actually present in the time series, and there still appears to be a need for methods which are sufficiently straightforward to use, and still have good power properties. In this paper we propose penalized quasi-likelihood based tests, which have a simple, nuisance-parameter free limit distribution. Simulations show that the asymptotic approximation is reasonably accurate already for moderate sample sizes, and that the power properties compare favorably to those of ordinary quasi-likelihood ratio tests. We apply our methods to the series of seasonally adjusted quarterly U.S. GNP data.

Key words: switching autoregression; likelihood ratio test; time series

Running Title: Tests for regime switching

1 Introduction

Autoregressive models with switching regime form a very flexible class of non-linear time-series models, that behave locally linear but globally feature structural changes. Often, the regime is taken as a finite-state Markov chain. Hence, a basic methodological issue is to determine the number of states of the underlying regime, or in a first place to test for the existence of at least two states. The asymptotic distribution of likelihood-ratio type tests is quite involved, however, and depends both on the underlying parametric model as well as on the true parameter values. Therefore, in this paper we introduce likelihood-based tests with tractable, nuisance-parameter free asymptotic distributions and good power properties, following the approaches by Chen, Chen and Kalbfleisch (2001) and Chen and Li (2009) for i.i.d. mixtures.

Let us now give a formal definition of the model. An autoregressive process with Markov regime, or Markov-switching autoregression, is a bivariate process $(S_k, X_k)_k$, where $(S_k)_k$ is an unobserved Markov chain on a finite state space $\mathcal{M}$, and where $(X_k)_k$ is the observed time

[1]Address for correspondence: Prof. Dr. Hajo Holzmann, Philipps-Universität Marburg, Fachbereich Mathematik und Informatik, Hans-Meerweinstr. D-35032 Marburg, Germany email: holzmann@mathematik.uni-marburg.de, Fon: + 49 6421 2825454

series, which are related as follows. Given $(S_k)_k$, $(X_k)_k$ is an inhomogeneous p-order Markov chain on a measurable space $\mathcal{X}$, and for each $t \geq 0$ the conditional distribution of X_t only depends on S_t and the lagged X's. We shall assume that $(S_k)_k$ is stationary and ergodic so that its stationary distribution is uniquely defined. The hidden process $(S_k)_k$ is referred to as regime.

Testing for regime switching corresponds to testing the null hypothesis $\mathcal{M} = \{1\}$ of a single state (so that the model reduces to a mere autoregressive process) against the alternative hypothesis $\mathcal{M} = \{1, 2\}$ of (at least) two states. Deriving the asymptotic distribution of the LRT and related test statistics is a difficult task for a variety of reasons. First, under the null hypothesis, parameters of the full model are not identified, and the asymptotic distribution of the corresponding LRT will be highly non-standard. This problem already arises in the closely related problem of testing for homogeneity in two-component mixtures, which has been intensively studied in recent years, see Chen et al. (2001), Dacunha-Castelle and Gassiat (1999) and Liu and Shao (2003). Second, additional difficulties arise if the Markov dependence structure of the regime is incorporated in the test statistic. Indeed, even for compact parameter spaces, Gassiat and Kerebin (2000) show that the LRT for regime switching may not converge in distribution at all.

Therefore, Cho and White (2007) and Dannemann and Holzmann (2008) suggest quasi LRTs which neglect the dependence structure of the regime in the construction of the test statistic. Cho and White (2007) consider an ordinary (quasi) likelihood ratio test and obtain results when testing for regime switching which are analogous to those in Chen et al. (2001) for the LRT and in Chen and Chen (2003) for testing for homogeneity in finite mixtures. Dannemann and Holzmann (2008) extend the test by Chen et al. (2004) for two against more states to a Markov-dependent regime. In this paper we propose several modifications of ordinary quasi LRTs in order to obtain a tractable asymptotic distribution, and maintain good power properties.

In economics, regime switching models are used for business cycle analysis (Hamilton 1989) and for further macroeconomic time series (e.g. Porter 1983 for investigating cartel behavior; Davig 2004 for the U.S. debt-output ratio) and also to study financial time series (Hamilton and Susmel 1994 for stock returns or Cai 1994 for treasury bills). In this paper we apply our methods to the series of seasonally adjusted quarterly U.S. GNP data from 1947(1) to 2002(3). It turns out that there is a regime switch in the variance of the series, which is however, exclusively due to the great moderation starting in 1984. When dividing into two subseries, for the series from 1984(2)-2002(3) the AIC is in favor of a switching model, while the BIC selects a simple AR(2)-model. Using the EM-test, we cannot reject the hypothesis of no regime switch with bootstrap p-value of ≈ 0.1. Thus, there is no clear evidence of an additional regime switch in the series, apart from that due to the great moderation.

The outline of the paper is as follows. In Section 2 we specify the model, give some examples, and discuss consistency properties of penalized quasi maximum likelihood estimators. Section 3.1 has the asymptotic distributions of a penalized (or modified) quasi LRT. In Section 3.2 we consider tests based on fixed proportions under the alternative as well as a variant called the EM-test, introduced by Chen and Li (2009) for i.i.d. normal mixtures, and cover all relevant models with single switching parameter including the switching intercept linear autoregressive model with normal innovations. In Section 4 we report the results of a simulation study, and in Section 5 we apply the methods to the series of U.S. quarterly GNP. The main steps of the proofs are given in Appendix A, further details are deferred to the supplementary material in Appendix B (attached for reviewer information only, not intended for publication).

2 Model specification and estimation

2.1 Markov switching autoregressive models

We specify the model for a two-state chain, which can be written as

$$X_t = F_{\boldsymbol{\omega}}(S_t, X_{t-1}^p; \epsilon_t), \tag{1}$$

where $(\epsilon_k)_{k\geq 0}$ is an independent and identically distributed sequence of random variables with $E(\epsilon_1) = 0$ and $E(\epsilon_1^2) = 1$ that we shall call the innovation process, $X_k^p = (X_k, \ldots, X_{k-p+1})$ and $(F_{\boldsymbol{\omega}})_{\boldsymbol{\omega}}$ is a family of functions indexed by some finite-dimensional parameter $\boldsymbol{\omega}$. We assume that $\boldsymbol{\omega}$ consists of the entries a_{21}, a_{12} of the transition matrix $P = (a_{ij})_{i,j=1,2}$, the switching parameters $\vartheta_1, \vartheta_2 \in \Theta \subset \mathbb{R}^r$ as well as the structural parameters $\boldsymbol{\eta} \in \boldsymbol{H} \subset \mathbb{R}^d$ which are the same for all states. The parameter sets Θ and $\boldsymbol{H}$ are assumed to be compact and convex. Further, we shall restrict our attention to the case of a single switching parameter, i.e. $r = 1$, so that

$$\boldsymbol{\omega}^T = \big(a_{21}, a_{12}, \vartheta_1, \vartheta_2, \boldsymbol{\eta}^T\big).$$

Example 1 (*Linear switching autoregression)***.** 1. The linear switching autoregressive model with switching intercept is given by

$$X_t = \zeta_{S_t} + \sum_{j=1}^{p} \phi_j X_{t-j} + \sigma\epsilon_t \tag{2}$$

where σ is a scale parameter for the innovation distribution, the ϕ_j's are the (non-switching) autoregressive parameters, and the intercept ζ switches according to S_t. Krolzig (1997) and Hamilton (2008) give further motivation and discussion of the properties. For the innovations, the normal distribution is a standard choice; another useful distribution is the t-distribution, which allows for heavier tails which are often observed empirically. In the above notation, we have $\vartheta_i = \zeta_i$, $i = 1, 2$. If σ is fixed, we have $d = p$ and $\boldsymbol{\eta} = (\phi_1, \ldots, \phi_p)^T$, otherwise, $d = p+1$ and $\boldsymbol{\eta} = (\phi_1, \ldots, \phi_p, \sigma)^T$.

2. The linear switching autoregressive model with one switching autoregressive parameter is given by

$$X_t = \zeta + \sum_{j=1}^{j_0-1} \phi_j X_{t-j} + \phi_{j_0,S_t} X_{t-j_0} + \sum_{j_0+1}^{p} \phi_j X_{t-j} + \sigma\epsilon_t. \tag{3}$$

where ζ is the non-switching intercept, ϕ_j, $j = 1, \ldots, j_0 - 1, j_0 + 1, \ldots, p$, are the (non-switching) autoregressive parameters, σ is the scale parameter of the innovation process and ϕ_{j_0,S_t} switches according to S_t. In the above notation, we have $d = p + 1$, $\vartheta_i = \phi_{j_0,i}$, $i = 1, 2$ and $\boldsymbol{\eta} = (\zeta, \phi_1, \ldots, \phi_{j_0-1}, \phi_{j_0+1}, \ldots, \phi_p, \sigma)^T$. Model (3) includes a single switching autoregressive coefficient $X_t = \phi_{S_t} X_{t-1} + \sigma\epsilon_t$.

3. The linear switching autoregressive model with switching variance is given by

$$X_t = \zeta + \sum_{j=1}^{p} \phi_j X_{t-j} + \sigma_{S_t}\epsilon_t, \tag{4}$$

where σ is a scale parameter for the innovation distribution which switches according to S_t, the intercept ζ and the ϕ_j's are the non-switching parameters. This is a very popular model for time series of asset prices (see e.g. Piger 2009). In the above notation, we have $d = p + 1$, $\vartheta_i = \sigma_i$, $i = 1, 2$, and $\boldsymbol{\eta} = (\zeta, \phi_1, \ldots, \phi_p)^T$.

Example 2 (*Switching ARCH*)**.** Regime switching ARCH-models were introduced by Hamilton and Susmel (1994) and by Cai (1994). The model specification by Cai (1994) when neglecting leverage effects is

$$X_t = \sigma_t \epsilon_t, \qquad \sigma_t^2 = \vartheta_{S_t} + \sum_{j=1}^{p} \phi_j X_{t-j}^2 \tag{5}$$

with parameters $\vartheta_i \geq 0$, $i = 1, 2$, and $\phi_j \geq 0$, $j = 1, \ldots, p$. In the above notation, we have $d = p$ and $\boldsymbol{\eta} = (\phi_1, \ldots, \phi_p)^T$. Again, we consider both normal as well as t-distributed innovations.

2.2 Penalized maximum likelihood estimation

Likelihood-based methods play a prominent role for parameter estimation in switching autoregressive models. Suppose that conditional on $X_{k-1}^p = x_{k-1}^p$ and $S_k = i$, X_k has density $g(x_k | x_{k-1}^p; \vartheta_i, \boldsymbol{\eta})$ w.r.t. some σ-finite measure μ on $\mathcal{X}$. Then the conditional likelihood given the initial observations $X_0^p = (X_0, \ldots, X_{-p+1})$ (we start indexing from $-p+1, -p+2, \ldots$) and the initial unobserved state $S_0 = i_0$ is given by

$$\tilde{l}_n(\boldsymbol{\omega}) = \log \Big(\sum_{i_1=1}^{2} \cdots \sum_{i_n=1}^{2} \prod_{k=1}^{n} a_{i_{k-1}, i_k} \prod_{k=1}^{n} g(X_k | X_{k-1}^p; \vartheta_{i_k}, \boldsymbol{\eta}) \Big) \tag{6}$$

The maximizer $\hat{\boldsymbol{\omega}}$ of $\tilde{l}_n(\boldsymbol{\omega})$ is called the (conditional) maximum likelihood estimate. Its asymptotic properties, especially consistency as well as asymptotic normality are well-established by now (cf. Douc et al. 2004).

As indicated in the introduction, instead of using the full-model log likelihood function $\tilde{l}_n(\boldsymbol{\omega})$ we shall base inference on a quasi likelihood which neglects the dependence structure in the regime. Let $\boldsymbol{\psi}^T = (\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta}^T)$,

$$g_{\mathrm{mix}}(x_t | x_{t-1}^p; \boldsymbol{\psi}) = (1 - \alpha) g(x_t | x_{t-1}^p; \vartheta_1, \boldsymbol{\eta}) + \alpha g(x_t | x_{t-1}^p; \vartheta_2, \boldsymbol{\eta}) \tag{7}$$

and consider the *quasi log-likelihood function* given by

$$l_n(\boldsymbol{\psi}) = \sum_{t=1}^{n} \log g_{\mathrm{mix}}(X_t | X_{t-1}^p; \boldsymbol{\psi}). \tag{8}$$

Note that (8) only is the true likelihood function if the regime is independent. Such models are sometimes called mixture autoregressive models (cf. Wong and Li 2000). For the time series model itself, an independent regime may not appear particularly attractive, but it can nevertheless be used for constructing a feasible test for regime switching. For a Markov-dependent regime, the parameter α in (7) corresponds to the stationary distribution of the underlying transition matrix.

Following Chen et al. (2001, 2004) and Chen and Li (2009), in order to obtain a feasible asymptotic distribution we consider a penalized version of l_n, called *modified or penalized quasi likelihood function*, which is defined by

$$pl_n(\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta}) = l_n(\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta}) + p(\alpha), \tag{9}$$

where $p(\alpha)$ is a penalty function fulfilling the following properties:
(i) $p(\alpha)$ attains its maximum at $\alpha = 0.5$, *(ii)* $p(\alpha)$ is continuous on $(0, 1)$, *(iii)* $p(\alpha) = p(1-\alpha)$ and *(iv)* $p(\alpha) \to -\infty$ for $\alpha \to 0$.
We shall use

$$p(\alpha) = \log(1 - |1 - 2\alpha|), \tag{10}$$

for an extensive discussion for reasons for this choice and modifications see Li (2007).
Let $(\widehat{\alpha}, \widehat{\vartheta}_1, \widehat{\vartheta}_2, \widehat{\boldsymbol{\eta}})$ (resp. $(\widehat{\alpha}^*, \widehat{\vartheta}_1^*, \widehat{\vartheta}_2^*, \widehat{\boldsymbol{\eta}}^*)$) be the maximizers of $l_n(\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta})$ (resp. $pl_n(\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta})$) over the parameter space $[0, 1] \times \Theta^2 \times \boldsymbol{H}$, and let $(\widehat{\vartheta}_0, \widehat{\boldsymbol{\eta}}_0)$ be the maximizers of $l_n(1/2, \vartheta, \vartheta, \boldsymbol{\eta})$ or equivalently of $pl_n(1/2, \vartheta, \vartheta, \boldsymbol{\eta})$ over the parameter space $\Theta \times \boldsymbol{H}$. We denote the true parameter under the null hypothesis of no regime switch by $(\vartheta_0, \boldsymbol{\eta}_0)$. If not otherwise specified, we compute the probabilities and expectations with respect to this distribution.
We shall need the following assumptions, which are satisfied in the above models under appropriate assumptions on the distribution of the innovations ϵ_t and the coefficients. See the discussion below Theorem 2 for details.

Assumption 1. The process $(\boldsymbol{Z}_k)_{k\geq 0} = (S_k, X_k, \ldots, X_{k-p+1})_{k\geq 0}$ is a Markov chain on $\mathcal{M} \times \mathcal{X}^p$. Under the null hypothesis, the observable process $(X_k)_k$ is strictly stationary and geometrically ergodic.

Assumption 2. *(Identifiability)* If for parameters $\boldsymbol{\psi}^T = (\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta}^T)$ and $\boldsymbol{\psi}'^T = (\alpha', \vartheta_1', \vartheta_2', \boldsymbol{\eta}'^T)$, $\alpha \notin \{0, 1\}$, one has that

$$g_{\text{mix}}(x|\, y^p; \boldsymbol{\psi}) = g_{\text{mix}}(x|\, y^p; \boldsymbol{\psi}') \text{ for all } x \in \mathcal{X}, y^p \in \mathcal{X}^p,$$

then $\boldsymbol{\eta} = \boldsymbol{\eta}'$ and after possibly permuting the states of the Markov chain $(S_k)_k$ we further have that $\alpha = \alpha'$ and $\vartheta_i = \vartheta_i'$, $i = 1, 2$.

Assumption 3. For all fixed $x \in \mathcal{X}, y^p \in \mathcal{X}^p$, $g(x|y^p; \cdot, \cdot) \in C^{(3)}(\Theta, \boldsymbol{H})$. Further, there exists a nonnegative measurable function $K : \mathcal{X}^{p+1} \to [0, \infty)$ such that

$$EK(X_1^{p+1}) < \infty \qquad \text{and} \quad |\log(g(x_1|x_0^p; \vartheta, \boldsymbol{\eta}))| \leq K(x_1^{p+1})$$

for all $x_1^{p+1} \in \mathcal{X}^{p+1}$ and all $(\vartheta, \boldsymbol{\eta}) \in \Theta \times \boldsymbol{H}$.

Assumptions 2 and 3 are Wald-type conditions, needed for consistency. The next assumption concerns the expressions in the score. For $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d)^T \in \boldsymbol{H}$ and $\vartheta \in \Theta$ we set

$$\begin{aligned} Y_i(\vartheta, \boldsymbol{\eta}) &= \frac{\frac{\partial}{\partial \vartheta} g(X_i|X_{i-1}^p; \vartheta, \boldsymbol{\eta})}{g(X_i|X_{i-1}^p; \vartheta_0, \boldsymbol{\eta}_0)}, \qquad Z_i(\vartheta) = \frac{\partial}{\partial \vartheta} Y_i(\vartheta), \qquad V_{i,j}(\vartheta, \boldsymbol{\eta}) = \frac{\partial}{\partial \eta_j} Y_i(\vartheta, \boldsymbol{\eta}) \\ U_{i,j}(\boldsymbol{\eta}) &= \frac{\frac{\partial}{\partial \eta_j} g(X_i|X_{i-1}^p; \vartheta_0, \boldsymbol{\eta})}{g(X_i|X_{i-1}^p; \vartheta_0, \boldsymbol{\eta}_0)}, \qquad W_{i,j,k}(\boldsymbol{\eta}) = \frac{\partial}{\partial \eta_k} U_{i,j}(\boldsymbol{\eta}), \qquad j, k = 1, \ldots, d. \end{aligned} \tag{11}$$

We require that these derivatives are uniformly bounded by an integrable function. This assumption implies tightness of relevant processes related to the score, which is required in the asymptotic analysis. For ease of notation, we set

$$Y_i = Y_i(\vartheta_0, \boldsymbol{\eta}_0), \qquad Z_i = Z_i(\vartheta_0), \qquad U_{i,j} = U_{i,j}(\boldsymbol{\eta}_0), \qquad j = 1, \ldots, d. \tag{12}$$

Assumption 4. There exists a nonnegative function $\tilde{K} : \mathcal{X}^{p+1} \to [0, \infty)$ such that $E\tilde{K}(X_1^{p+1}) < \infty$ and such that for all $\boldsymbol{\eta} \in \boldsymbol{H}$, $\vartheta \in \Theta$,

$$|Y_1(\vartheta, \boldsymbol{\eta})|^3 \le \tilde{K}(X_1^{p+1}), \quad |Z_1(\vartheta)|^3 \le \tilde{K}(X_1^{p+1}), \quad |U_{1,j}(\boldsymbol{\eta})|^3 \le \tilde{K}(X_1^{p+1}),$$
$$|V_{i,j}(\vartheta, \boldsymbol{\eta})|^2 \le \tilde{K}(X_1^{p+1}), \quad |W_{i,j,k}(\boldsymbol{\eta})|^2 \le \tilde{K}(X_1^{p+1}), \quad |\partial/\partial_\vartheta Z_i(\vartheta)|^2 \le \tilde{K}(X_1^{p+1}).$$

for $j, k = 1, \ldots, d$.

Theorem 1. *Suppose that Assumptions 1 – 4 are satisfied. In case of a single state (i.e. no regime switch), we have that*

(i) $\widehat{\vartheta}_0 - \vartheta_0 = o_P(1)$, $\widehat{\boldsymbol{\eta}}_0 - \boldsymbol{\eta}_0 = o_P(1)$ *and*

(ii) $\widehat{\vartheta}_1^* - \vartheta_0 = o_P(1)$, $\widehat{\vartheta}_2^* - \vartheta_0 = o_P(1)$, $\widehat{\boldsymbol{\eta}}^* - \boldsymbol{\eta}_0 = o_P(1)$.

Thus, under the hypothesis of no regime switch, both estimators $\widehat{\vartheta}_i^*$ are consistent for ϑ_0. This is due to the penalty term $p(\alpha)$ in (9): The estimator $\widehat{\alpha}^*$ is forced away from 0 and 1, so that both $\widehat{\vartheta}_i^*$ need to be consistent. This is not true for the quasi MLEs $\widehat{\vartheta}_i$.

3 Feasible quasi-likelihood based tests for regime switching

3.1 The modified quasi-likelihood ratio test

If $(1-\alpha, \alpha)$ denotes the stationary distribution of $(S_k)_k$, then the hypothesis of no regime switch is equivalent to

$$H_0 : \alpha(1-\alpha)(\vartheta_1 - \vartheta_2) = 0.$$

We propose to test H_0 via the *modified quasi likelihood ratio test* (MQLRT) statistic

$$M_n = 2\{pl_n(\widehat{\alpha}^*, \widehat{\vartheta}_1^*, \widehat{\vartheta}_2^*, \widehat{\boldsymbol{\eta}}^*) - pl_n(1/2, \widehat{\vartheta}_0, \widehat{\vartheta}_0, \widehat{\boldsymbol{\eta}}_0)\}. \tag{13}$$

For the asymptotic distribution of M_n, we shall additionally require

Assumption 5. The covariance matrix of $(U_{1,1}, \ldots, U_{1,d}, Y_1, Z_1)$ (see (12) for the definition) is positive definite.

Assumption 5 does not hold true in model (2) when the innovations are normal, see the discussion of Examples 1 and 2 below. We shall derive the asymptotic distribution of a closely related test statistic in this model in Theorem 4.

Theorem 2. *Under the null hypothesis H_0 of no regime switching, if Assumptions 1 – 5 are satisfied we have that*

$$M_n \xrightarrow{d} \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2,$$

where χ_p^2 denotes the χ^2-distribution with $p > 0$ degrees of freedom, χ_0^2 is the point mass at 0*, and $\xrightarrow{d}$ denotes convergence in distribution.*

As desired, the asymptotic distribution of M_n is easy to handle and does not depend on the underlying parametric model, the actual true parameter values or the choice of the compact set Θ (as long as it contains the true value). This is in contrast to the asymptotic distribution of a quasi likelihood ratio test (QLRT, without penalization) of H_0 based on (8).

Examples 1 and 2 *(continued).* First consider the *identifiability* Assumptions 2 and 5. Suppose that the innovations ϵ_t are real-valued with continuous density $f > 0$ w.r.t. Lebesgue measure, and let $f(x;\mu,\sigma) = f\big((x-\mu)/\sigma\big)/\sigma$ denote the corresponding location-scale family. First consider models (4) and (5).

Lemma 1. *a. If the parameter $(\alpha,\mu,\sigma_1,\sigma_2)$ in a two-component scale mixture $(1-\alpha)f(x;\mu,\sigma_1)+\alpha f(x;\mu,\sigma_2)$ is identified (except for label switching), then Assumption 2 is satisfied for the models* (4) *and* (5)*.*
b. Suppose that for any (μ,σ) and $a_1,a_2,a_3\in\mathbb{R}$,

$$a_1\frac{\partial f(x;\mu,\sigma)}{\partial\mu}+a_2\frac{\partial f(x;\mu,\sigma)}{\partial\sigma}+a_3\frac{\partial^2 f(x;\mu,\sigma)}{\partial^2\sigma}=0 \quad \text{for Lebesgue-a.e. } x \tag{14}$$

implies that $a_1=a_2=a_3=0$. Then Assumption 5 is satisfied for the models (4) *and* (5)*.*

See Appendix A for the proof. Since general finite mixtures of normal and t-distributions (even with variable degrees of freedom) are identifiable (cf. Holzmann et al. 2006), Lemma 1 a. implies that Assumption 2 will be satisfied in these cases. As for b., we give the proof for the normal distribution in Appendix A, for the t-distribution see Appendix B.
For models (2) and (3), we have the general result

Lemma 2. *a. If the parameter $(\alpha,\mu_1,\mu_2,\sigma)$ in a two-component location mixture $(1-\alpha)f(x;\mu_1,\sigma)+\alpha f(x;\mu_2,\sigma)$ is identified (except for label switching), then Assumption 2 is satisfied for the models* (2) *and* (3)*.*
b. Suppose that for any (μ,σ) and $a_1,a_2,a_3\in\mathbb{R}$,

$$a_1\frac{\partial f(x;\mu,\sigma)}{\partial\mu}+a_2\frac{\partial^2 f(x;\mu,\sigma)}{\partial^2\mu}+a_3\frac{\partial f(x;\mu,\sigma)}{\partial\sigma}=0 \quad \text{for Lebesgue-a.e. } x \tag{15}$$

implies that $a_1=a_2=a_3=0$. Then Assumption 5 is satisfied for the models (2) *and* (3)*.*

The proof is similar to that of the above lemma and omitted. In Appendix B we show that (15) is indeed satisfied for the t-distribution.
However, for the normal distribution, since $\sigma\frac{\partial^2 f(x;\mu,\sigma)}{\partial^2\mu}=\frac{\partial f(x;\mu,\sigma)}{\partial\sigma}$ holds condition (15) is not fulfilled. Indeed, for model (2) with normal innovations the MQLRT for testing for homogeneity does not admit the simple asymptotic distribution given in Theorem 2 in case of a variable scale parameter (it does for fixed scale parameter). We shall propose a test and derive its asymptotic distribution in Section 3.2.
For model (3) in case of normal innovations a direct argument shows that Assumption 5 is still fulfilled, see Lemma 5 in Appendix A.

3.2 Fixed proportions and the EM-test

Instead of penalizing small values of α, a simple possibility is to consider a finite set of fixed values $\mathcal{J}=\{\alpha_1,\dots,\alpha_J\}$ for α under the alternative. If the set $\mathcal{J}$ is sufficiently large, this need not lead to a great loss of power. Specifically, let $(\widehat{\vartheta}_{1,\alpha_j}, \widehat{\vartheta}_{2,\alpha_j}, \widehat{\boldsymbol{\eta}}_{\alpha_j})$ be the maximizer of $l_n(\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta})$ subject to $\alpha = \alpha_j$, and set

$$\begin{aligned} R_n(\alpha_j) &= 2\{l_n(\alpha_j, \widehat{\vartheta}_{1,\alpha_j}, \widehat{\vartheta}_{2,\alpha_j}, \widehat{\boldsymbol{\eta}}_{\alpha_j}) - l_n(1/2, \widehat{\vartheta}_0, \widehat{\vartheta}_0, \widehat{\boldsymbol{\eta}}_0)\}, \quad \alpha_j \in \mathcal{J}, \\ R_n(\mathcal{J}) &= \max_{\alpha_j \in \mathcal{J}} R_n(\alpha_j). \end{aligned}$$

Note that no penalty is required when testing against fixed proportions. In order to increase the power, Chen and Li (2009) proposed to perform, starting from each $\alpha \in \mathcal{J}$, a fixed finite number K of EM-steps, for which they require the penalized likelihood function $pl_n(\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta})$. This leads to the EM-test statistic $EM_n^{(K)} = EM_n^{(K)}\big(\mathcal{J}\big)$. We give the details for its computation in Appendix A. Let us point out that $EM_n^{(0)}(\mathcal{J})$ is simply the penalized version of $R_n(\mathcal{J})$

Theorem 3. *Under the assumptions of Theorem 2, if $\mathcal{J} \subset (0,1)$ is a finite subset with $1/2 \in \mathcal{J}$, then*

$$R_n(\mathcal{J}) \xrightarrow{d} \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2, \qquad EM_n^{(K)}(\mathcal{J}) \xrightarrow{d} \frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2. \tag{16}$$

For the proof see Appendix A. Finally, we consider model (2) with normal innovations, i.e.

$$X_t = \zeta_{S_t} + \sum_{j=1}^{p} \phi_j X_{t-j} + \sigma\epsilon_t, \quad \epsilon_t \overset{iid}{\sim} N(0,1). \tag{17}$$

In case of independent mixtures, testing for homogeneity in such a model has been studied in Chen and Chen (2003), Qin and Smith (2004) and Chen and Li (2009). For the MLRT, Chen and Chen (2003) derive an asymptotic upper bound which is strengthened by Qin and Smith (2004) to the $\frac{1}{2}\chi_1^2 + \frac{1}{2}\chi_2^2$ distribution. Qin and Smith (2004) even claim that this is the asymptotic distribution of the MLRT, however, their argument that this bound is attained seems to be incorrect, and the exact asymptotic distribution for the MLRT remains somewhat unclear. We conjecture that it is the same as for $R_n(\mathcal{J})$ as specified below. An outline of the proof of the following theorem, which is similar to the proof of theorem 2 in Chen and Li (2009) is given in Appendix A.

Theorem 4. *Suppose that $(X_k)_k$ follows a stationary AR(p) model with normal innovations. When testing for regime switching as specified by (17), if for the finite set $\mathcal{J} \subset (0,1)$ we have a. $1/2 \in \mathcal{J}$ and b. $\mathcal{J}$ contains a further element $\neq 1/2$, then for the test against fixed proportions that for $x \in \mathbb{R}$ as $n \to \infty$,*

$$P(R_n(\mathcal{J}) \leq x) \to F(x)\big(1_{x>0} + F(x)\big)/2,$$

where F is the cdf of a χ_1^2-variate, and for the EM-test that

$$P(EM_n^{(K)} \leq x) \longrightarrow F(x - \Delta)\big(1_{x>0} + F(x)\big)/2,$$

where

$$\Delta = 2 \max_{\alpha_j \neq 1/2} \{p(\alpha_j) - p(1/2)\}.$$

4 Simulations

Here we present some of the results of an extensive simulation study of the tests proposed in the two previous sections. We conducted the simulations with $\mathcal{J} = \{0.1, 0.3, 0.5\}$ for both R_n and $M_n^{(K)}$. We only present results for normally distributed innovations, for the t-distribution see Ketterer (2011).

4.1 Simulated levels

In this section we simulate the size of the MQLRT, the EM-test and the test for fixed proportions in several settings.

a. Switching Autoregression with switching intercept.

Data-generating process (DGP): $X_t = 0.5X_{t-1} + \epsilon_t$, where $\epsilon_t \overset{iid}{\sim} N(0,1)$.

Model : $X_t = \zeta_{S_t} + \phi X_{t-1} + \sigma\epsilon_t$ with $\epsilon_t \overset{iid}{\sim} N(0,1)$

Note that Theorem 4 applies. The results for various sample sizes can be found in Table 1. The tests are somewhat anticonservative for small sample sizes. Figure 1 shows the ecdf of the EM-test statistic $EM_n^{(2)}$ for sample sizes $n = 200$ as well as $n = 1000$.

b. Switching Autoregression with normally distributed innovations with switching scale.

DGP: $X_t = 0.2X_{t-1} + \epsilon_t$, where $\epsilon_t \overset{iid}{\sim} N(0,1)$.

Model: $X_t = \zeta + \phi X_{t-1} + \sigma_{S_t}\epsilon_t$ with $\epsilon_t \overset{iid}{\sim} N(0,1)$.

The results for various sample sizes are contained in Table 2. The tests have almost identical levels.

c. Switching Autoregression with normally distributed innovations with switching autoregressive parameter.

DGP: $X_t = 0.5X_{t-1} + \epsilon_t$, where $\epsilon_t \overset{iid}{\sim} N(0,1)$.

Model: $X_t = \zeta + \phi_{S_t} X_{t-1} + \sigma\epsilon_t$ with $\epsilon_t \overset{iid}{\sim} N(0,1)$.

The results for various sample sizes are contained in Table 3. The tests are somewhat conservative.

d. Switching ARCH

DGP: $X_t = \sigma_t\epsilon_t$; $\sigma_t^2 = 1 + 0.5X_{t-1}^2 + 0.3X_{t-2}^2$, where $\epsilon_t \overset{iid}{\sim} N(0,1)$,

Model: $X_t = \sigma_t\epsilon_t$; $\sigma_t^2 = \vartheta_{S_t} + \phi_1 X_{t-1}^2 + \phi_2 X_{t-2}^2$ with $\epsilon_t \overset{iid}{\sim} N(0,1)$.

The results are in Table 4, the tests are slightly conservative.

4.2 Power comparison of several tests

We present the results of a power comparison of the various tests. We shall restrict ourselves to the linear autoregressive model with switching intercept and normal innovations, with variable scale. Thus, as a model we maintain

$$X_t = \zeta_{S_t} + \phi X_{t-1} + \sigma\epsilon_t \qquad \text{with} \qquad \epsilon_t \overset{iid}{\sim} N(0,1). \tag{18}$$

For proper estimation of the power we shall use simulated critical values. Precisely, for given alternative, we simulate the critical value of the tests from the distribution (without regime switch) which is fitted to a large sample from the alternative by maximum likelihood, it is (approximately) the closest approximation to the alternative in Kullback-Leibler distance. Note the analogy to a corresponding bootstrap procedure.

As data generating process we maintain $X_t = (-1)^{S_t} + \phi X_{t-1} + \epsilon_t$, where $\epsilon_t \overset{iid}{\sim} N(0, 1)$, with various values of ϕ, a_{12} and a_{21}.

Effect of mixture proportion

First, we investigate the effect of the mixture proportion α on the various tests. We fix $\phi = 0.5$, and for distinct α choose a_{12} and a_{21} to generate an independent regime. Table 5 contains the results. As could be expected, for small α the QLRT and the test for fixed proportions have higher power than the EM-test, which penalizes small α's. For α close to 0.5, the EM-test has highest power.

Effect of dependent regime and autoregression

Next, we investigate the effect of a dependent regime. We consider various combinations of a_{12}, a_{21} which lead to $\alpha = 0.5$. In Table 6 the results are displayed for underlying parameter $\phi = 0$, similar results were obtained for $\phi = 0.5$. The tests have the highest power when the Markov chain reduces to an i.i.d. sample, i.e. $a_{12} = a_{21} = 0.5$, for strongly dependent regime, the power is much smaller. In contrast, if we drop the autoregressive parameter ϕ from the model, i.e. when testing for homogeneity in a hidden Markov-model with state dependent distributions $P(X_t \leq x | S_t = i) = \Phi\big((x - \zeta_i)/\sigma\big)$, $i = 1, 2$, Table 7 shows that there is no loss of power for a dependent regime.

5 Application

In this section, we apply our methods to the series of quarterly, saisonally adjusted U.S. GNP from 1947(1) to 2002(3). The data are Real U.S. Gross National Product in billions of chained 1996 dollars and can be obtained by the Federal Reserve Bank of St. Louis (*http://research.stlouisfed.org/*). Instead of considering the data, say Y_t, we consider the growth rate $X_t = \nabla \log(Y_t) = \log(Y_t) - \log(Y_{t-1})$ (in %) which is plotted in Figure 2 (left).

As indicated by the acf and pacf (see Figure 2, (right)), the series shows autocorrelation. Therefore, we model the data by various switching autoregressive models of orders $p = 1, \ldots, 4$ with normal innovations. Here, $\mathcal{M}_1$ is the ordinary AR(p)-model, $\mathcal{M}_2$ allows all parameters to switch, $\mathcal{M}_3$ has a switching scale parameter (model (4)) and $\mathcal{M}_4$ a switching intercept (model (2)).

Using formal model selection criteria, see Table 8, one chooses model $\mathcal{M}_3$ and $p = 1$ according to BIC and $\mathcal{M}_3$ with $p = 3$ according to AIC. Here, we note, that the AIC and BIC are computed by

$$AIC = -2\tilde{l}_n(\hat{\boldsymbol{\omega}}) + 2 \cdot k(\hat{\boldsymbol{\omega}}) \text{ and } BIC = -2\tilde{l}_n(\hat{\boldsymbol{\omega}}) + \log(n) \cdot k(\hat{\boldsymbol{\omega}}),$$

where $\tilde{l}_n(\boldsymbol{\omega})$ is the log likelihood conditional on the first 4 observations and on state 1 and $k(\hat{\boldsymbol{\omega}})$ denotes the length of the vector $\hat{\boldsymbol{\omega}}$.

Testing for homogeneity in model $\mathcal{M}_3$ via the MQLRT as in Theorem 3, we reject the hypothesis of no regime switch for all orders $p = 1, \ldots, 4$ with p-value < 0.01. However, the resulting ML estimates in model $\mathcal{M}_3$ (see Table 9) for the hidden Markov chain are highly persistent. Computing the most likely sequence of hidden states using the Viterbi algorithm (see Viterbi, 1967) given the fitted model $\mathcal{M}_3$ ($p = 1$), we see that there is only one regime

switch in the variance between 1984(1) and 1984(2). This result corresponds to the 'Great Moderation' of the U.S. GNP growth rate, i.e. the permanent decline in the (variability of the) growth rate of U.S. GNP.
Therefore, we divide our series in two subseries: the first from 1947(1)–1984(1) and the second 1984(2)–2002(3), and fit the models $\mathcal{M}_1, \ldots, \mathcal{M}_4$ for lags $p = 1, \ldots, 4$, to the two subseries. For the series from 1947(1)–1984(1), the BIC (455.78) as well as the AIC (446.87) favor a simple AR(1)-model. Testing for switching intercept or switching scale, neither hypothesis can be rejected by the corresponding EM or MQLR tests. For the subseries from 1984(2)–2002(3), the BIC (119.8) selects a purely linear autoregressive model of order $p = 2$ whereas the AIC selects a switching model of order $p = 2$ (model $\mathcal{M}_2$ has AIC=109.32, $\mathcal{M}_4$ AIC=109.66). Therefore, we test for a switching intercept using the EM-test, for order $p = 2$. The asymptotic p-value of $EM_n^{(2)}$ is 0.04, however, the simulation in Table 1 showed that the test based on the asymptotic distribution is rather anticonservative. Thus, we use a parametric bootstrap based on the fitted AR(2)-model to estimate the p-value, which yields 0.095. Thus, there is no clear evidence in favor of regime switching.

Summarizing, apart from the change point at the beginning of the great moderation, which results in a single switch to a less volatile state, there appears to be no clear evidence of regime switching in the series or the subseries corresponding to the time before and after the great moderation.

Acknowledgements

The authors gratefully acknowledge financial support from the DFG, grant Ho 3260/3-1 and from the Landesstiftung Baden-Württemberg, "Juniorprofessorenprogramm".

6 References

Cai, J. (1994). A Markov model of unconditional variance in ARCH. *Journal of Business and Economic Statistics* **12**, 309–316.

Chen, H. , Chen, J. and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *J. R. Stat. Soc. Ser. B* **63**, 19–29.

Chen, H. and Chen, J. (2003). Tests for homogeneity in normal mixtures in the presence of a structural parameter. *Statistica Sinica* **13**, 351-365.

Chen, H., Chen, J. and Kalbfleisch, J. D. (2004). Testing for a finite mixture model with two components. *J. R. Stat. Soc. Ser. B* **66**, 95–115.

Chen, J. and Li, P. (2009). Hypothesis test for Normal Mixture Models: the EM Approach. *Annals of Statistics* **37**, 2523-2542.

Cho, J. S. and White, H. (2007). Testing for regime switching. *Econometrica* **75**, 1671–1720.

Dacunha-Castelle, D. and Gassiat, E. (1999). Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *Ann. Statist.* **27**, 1178–1209.

Dannemann, J. and Holzmann, H. (2008). Testing for two states in a hidden Markov model.

Canadian J. Statist **36**, 505-520.

Davig, T. (2004). Regime-Switching Debt and Taxation. *Journal of Monetary Economics* **51**, 837–859.

Douc, R., Moulines, E. and Rydén T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Annals of Statistics* **32**, 2254–2304.

Ferguson, T.S. (1996). *A Course in Large Sample Theory*. Chapman & Hall, London.

Fu, Y., Chen, J. and Li P. (2008). Modified likelihood ratio test for homogeneity in a mixture of von Mises distributions. *Journal of Statistical Planning and Inference* **138**, 667–681.

Frydman, R. (1980). A proof of the consistency of maximum likelihood estimators of nonlinear regression models with autocorrelated errors. *Econometrica* **48**, 853–860.

Gassiat E. and Keribin C. (2000). The likelihood ratio test for the number of components in a mixture with Markov regime. *ESAIM P&S* **4**, 25–52.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* **57**, 357–384.

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

Hamilton, J. D. (2008). Regime-Switching Models. In *The New Palgrave Dictionary of Economics, Second Edition* (eds. S. N. Durlauf and L. E. Blume). Palgrave McMillan Ltd.

Holzmann, H. and Ketterer, F. (2011). Feasible tests for regime switching in autoregressive models: Technical Details. Preprint, Marburg University.

Holzmann, H., Munk, A. and Gneiting, T. (2006). Identifiability of finite mixtures of elliptical distributions. *Scandinavian J. Statist* **33**, 753–763.

Ketterer, F. (2011). *Penalized likelihood based tests for regime switching in autoregressive models*. Ph.D. thesis, Marburg University, Germany.

Krolzig, H. M. (1997). *Markov-switching vector autoregressions*. Lecture notes in Economics and Mathematical Systems 454. Springer, Berlin.

Li, P. (2007). *Hypothesis testing in finite mixture models*. Ph.D. thesis, University of Walterloo, Canada.

Liu, X. and Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Ann. Statist.* **31**, 807–832.

Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80**, 267–278

Piger, J. (2009). Econometrics: models of regime changes. In *Encyclopedia of Complexity and Systems Science* (R.A. Meyers, ed.). Springer, New York.

Porter, R. H. (1983). A study of cartel stability: The Joint Executive Committee, 1880-1886. *The Bell Journal of Economics* **14**, 301–314.

Qin, Y. S. and Smith, B. (2004). Likelihood ratio test for homogeneity in normal mixtures in the presence of a structural parameter. *Statistica Sinica* **14**, 1165–1177.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE TRans. Inorm. Th.* **13**, 260–269.

Wong, C. S. and Li, W. K. (2000). On a mixture autoregressive model. *J. Royal Statist. Soc. B* **62**, 95–115.

7 Tables and Figures

Table 1: DGP: $X_t = 0.5X_{t-1} + \epsilon_t$, where $\epsilon_t \overset{iid}{\sim} N(0,1)$, Model: $X_t = \zeta_{S_t} + \phi X_{t-1} + \sigma\epsilon_t$ with $\epsilon_t \overset{iid}{\sim} N(0,1)$; number of replications: 20,000.

Sample Size	Nominal Levels (%)	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	$R_n(\mathcal{J})$
$n = 200$	10%	12.3	12.8	13.1	15.7
	5%	6.7	7.1	7.3	8.4
	1%	1.5	1.7	1.8	1.9
$n = 500$	10%	11.9	12.0	12.2	14.9
	5%	6.1	6.2	6.4	8.0
	1%	1.2	1.3	1.3	1.8
$n = 1000$	10%	10.8	10.9	11.0	13.8
	5%	5.6	5.7	5.7	7.4
	1%	1.2	1.2	1.3	1.6

Table 2: DGP: $X_t = 0.2X_{t-1} + \epsilon_t$, where $\epsilon_t \overset{iid}{\sim} N(0,1)$, Model: $X_t = \zeta + \phi X_{t-1} + \sigma_{S_t}\epsilon_t$ with $\epsilon_t \overset{iid}{\sim} N(0,1)$; number of replications: 20,000.

Sample Size	Nominal Levels (%)	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	$MQLRT$
$n = 200$	10%	11.0	11.1	11.1	11.4
	5%	6.3	6.4	6.4	6.7
	1%	1.9	1.9	2.0	2.1
$n = 500$	10%	9.6	9.7	9.7	9.8
	5%	5.1	5.1	5.1	5.3
	1%	1.1	1.1	1.1	1.2

Table 3: DGP: $X_t = 0.5X_{t-1} + \epsilon_t$, where $\epsilon_t \overset{iid}{\sim} N(0,1)$, Model: $X_t = \zeta + \phi_{S_t} X_{t-1} + \sigma\epsilon_t$ with $\epsilon_t \overset{iid}{\sim} N(0,1)$; number of replications: 20,000.

Sample Size	Nominal Levels (%)	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	$MQLRT$
$n = 200$	10%	7.9	8.0	8.0	8.6
	5%	4.2	4.2	4.2	4.8
	1%	1.0	1.0	1.0	1.2
$n = 500$	10%	8.9	8.9	8.9	9.3
	5%	4.8	4.8	4.8	5.3
	1%	1.0	1.0	1.0	1.3

Table 4: DGP: $X_t = \sigma_t\epsilon_t$; $\sigma_t^2 = 1 + 0.5X_{t-1}^2 + 0.3X_{t-2}^2$, where $\epsilon_t \overset{iid}{\sim} N(0,1)$, Model: $X_t = \sigma_t\epsilon_t$; $\sigma_t^2 = \vartheta_{S_t} + \phi_1 X_{t-1}^2 + \phi_2 X_{t-2}^2$ with $\epsilon_t \overset{iid}{\sim} N(0,1)$; number of replications: 20,000.

Sample Size	Nominal Levels (%)	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	$MQLRT$
$n = 100$	10%	8.0	8.0	8.1	8.4
	5%	4.4	4.4	4.5	4.7
	1%	0.9	0.9	0.9	1.1
$n = 500$	10%	8.8	8.8	8.8	9.1
	5%	4.6	4.6	4.6	4.9
	1%	1.1	1.1	1.1	1.1

Table 5: Nominal level: 5%; DGP: $X_t = \zeta_{S_t} + \phi X_{t-1} + \sigma\epsilon_t$, where $\epsilon_t \overset{iid}{\sim} N(0,1)$, sample size: 500, number of replications: 5,000, Model: $X_t = \zeta_{S_t} + \phi X_{t-1} + \sigma\epsilon_t$, with $\epsilon_t \overset{iid}{\sim} N(0,1)$.

a_{12}	a_{21}	α	ζ_1	ζ_2	ϕ	σ	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	QLRT	$R_n(\mathcal{J})$
0.1	0.9	0.1	-1	1	0.5	1	77.4	77.5	77.5	84.1	86.1
0.3	0.7	0.3	-1	1	0.5	1	92.8	93.0	93.1	91.3	92.9
0.4	0.6	0.4	-1	1	0.5	1	88.9	89.1	89.2	85.9	85.6

Table 6: Nominal level: 5%; DGP: $X_t = (-1)^{S_t}\zeta + \epsilon_t$, where $\epsilon_t \overset{iid}{\sim} N(0,1)$, sample size: 500, number of replications: 5,000, Model: $X_t = \zeta_{S_t} + \phi X_{t-1} + \sigma\epsilon_t$, with $\epsilon_t \overset{iid}{\sim} N(0,1)$. Let $\alpha = a_{12}/(a_{12}+a_{21})$ and $(1-\alpha,\alpha)$ be the stationary distribution of the hidden Markov Chain $(S_k)_k$.

a_{12}	a_{21}	α	ζ	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$	QLRT	$R_n(\mathcal{J})$
0.1	0.1	0.5	1.0	12.9	12.9	12.9	10.0	10.7
0.2	0.2	0.5	1.0	34.3	34.4	34.4	26.0	28.1
0.3	0.3	0.5	1.0	65.4	65.5	65.3	54.4	56.8
0.5	0.5	0.5	1.0	88.1	88.3	88.4	81.9	82.8
0.7	0.7	0.5	1.0	63.9	63.8	64.0	52.7	54.7
0.8	0.8	0.5	1.0	30.2	30.2	30.4	21.3	24.2
0.9	0.9	0.5	1.0	9.3	9.3	9.4	6.4	7.2

Table 7: Nominal level: 5%; DGP: $X_t = (-1)^{S_t}\zeta + \epsilon_t$, where $\epsilon_t \overset{iid}{\sim} N(0,1)$, sample size: 500, number of replications: 5,000, Model: $X_t = \zeta_{S_t} + \sigma\epsilon_t$, with $\epsilon_t \overset{iid}{\sim} N(0,1)$. Let $\alpha = a_{12}/(a_{12}+a_{21})$ and $(1-\alpha,\alpha)$ be the stationary distribution of the hidden Markov Chain $(S_k)_k$.

a_{12}	a_{21}	α	ζ	$EM_n^{(0)}$	$EM_n^{(1)}$	$EM_n^{(2)}$
0.1	0.1	0.5	1.0	87.6	87.6	87.6
0.2	0.2	0.5	1.0	88.0	88.0	88.0
0.3	0.3	0.5	1.0	89.3	89.3	89.3
0.5	0.5	0.5	1.0	89.5	89.5	89.5
0.7	0.7	0.5	1.0	88.6	88.6	88.6
0.8	0.8	0.5	1.0	89.5	89.5	89.5
0.9	0.9	0.5	1.0	89.0	89.0	89.0

Table 8: BIC (left) and AIC (right) for the corresponding models for series 1947(1)–2002(3).

BIC	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$	**AIC**	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
$p=1$	615.05	602.91	**592.56**	622.51	$p=1$	604.90	575.83	572.25	602.21
$p=2$	618.86	607.16	593.71	626.57	$p=2$	605.32	573.32	570.02	602.88
$p=3$	621.54	615.63	597.01	628.85	$p=3$	604.62	575.01	**569.94**	601.77
$p=4$	623.91	623.73	600.71	632.39	$p=4$	603.60	576.35	570.25	601.93

Table 9: Fits for model $\mathcal{M}_3$ (only the variance is allowed to switch, using maximum likelihood estimation) 1947(1)–2002(3).

	$\hat{a}_{12}$	$\hat{a}_{21}$	$\hat{\zeta}$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	$\hat{\phi}_4$	$\hat{\sigma}_1$	$\hat{\sigma}_2$
$p=1$	0.007	0.005	0.51	0.35				0.51	1.12
$p=2$	0.007	0.005	0.44	0.30	0.14			0.50	1.12
$p=3$	0.007	0.005	0.48	0.31	0.17	-0.10		0.50	1.12
$p=4$	0.007	0.005	0.53	0.31	0.19	-0.07	-0.09	0.50	1.11

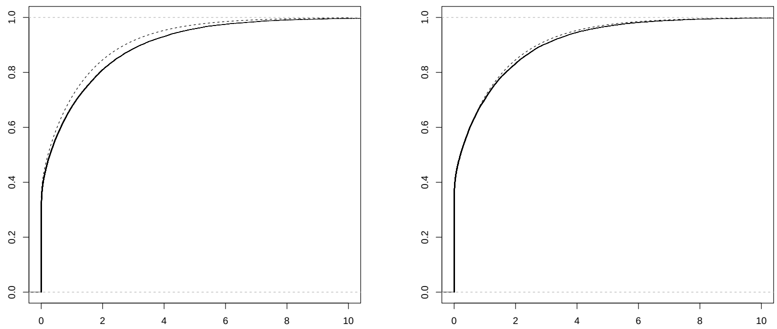

Figure 1: Ecdf of $EM_n^{(2)}$ for testing for homogeneity in model (17) (solid line) for DGP $X_t = 0.5X_{t-1} + \epsilon_t$, $\epsilon_t \overset{iid}{\sim} N(0,1)$, together with the limit distribution (dashed line) for $n = 200$ (left) and $n = 1000$ (right).

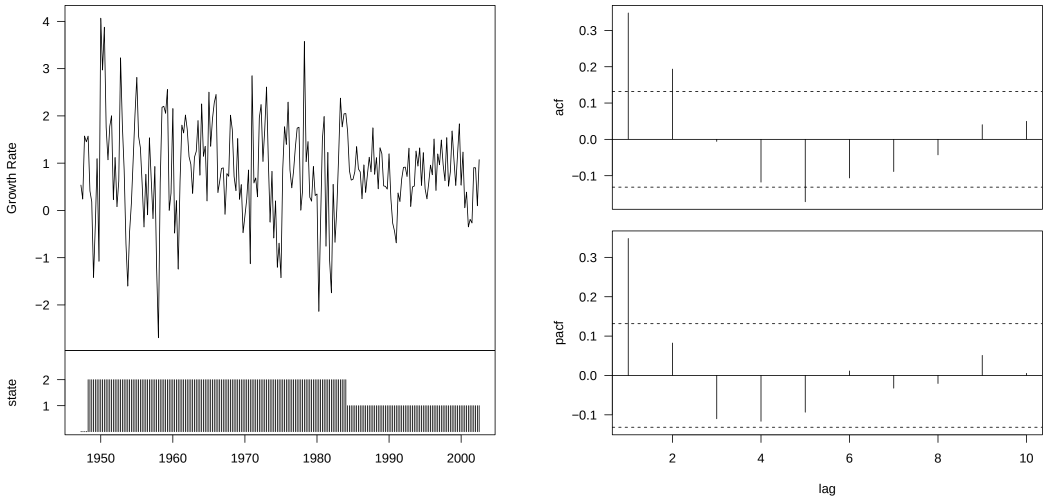

Figure 2: U.S. GNP quarterly growth rate in % with sequence of associated states in the a-posteriori analysis (left) and acf and pacf of the sequence (right).

A Proofs

A.1 Proofs of Sections 2.2 and 3.1

All probabilities and expected values are taken under the hypothesis of no regime switch. Recall that $(\vartheta_0, \boldsymbol{\eta}_0)$ denote the true parameters. To prove Theorem 1, we need the following lemma.

Lemma 3. *Under Assumption 2, if $\delta > 0$ and $\alpha \in [\delta, 1-\delta]$ we have that*

$$E \log \left(\frac{g_{mix}(X_1|X_0^p; \boldsymbol{\psi})}{g(X_1|X_0^p; \vartheta_0, \boldsymbol{\eta}_0)} \right) \leq 0$$

with equality if and only if $g(x_1|x_0^p; \vartheta_i, \boldsymbol{\eta}) = g(x_1|x_0^p; \vartheta_0, \boldsymbol{\eta}_0)$ Leb. – a.s., $i = 1, 2$.

Proof. Using Jensen's inequality and Assumption 2 we get

$$E \log \left(\frac{g_{\text{mix}}(X_1|X_0^p; \boldsymbol{\psi})}{g(X_1|X_0^p; \vartheta_0, \boldsymbol{\eta}_0)} \right) \leq \log E \left(\frac{g_{\text{mix}}(X_1|X_0^p; \boldsymbol{\psi})}{g(X_1|X_0^p; \vartheta_0, \boldsymbol{\eta}_0)} \right) \leq 0$$

with equality if and only if $g(x_1|x_0^p; \vartheta_i, \boldsymbol{\eta}) = g(x_1|x_0^p; \vartheta_0, \boldsymbol{\eta}_0)$ Leb. $-$ *a.s.*, $i = 1, 2$. □

Proof of Theorem 1. (i) Since $(X_t)_t$ is stationary and ergodic, $(g(X_t|X_{t-1}^p; \vartheta, \boldsymbol{\phi}))_t$ is stationary and ergodic (cf. Krengel 1985, Prop. 1.4.3). By Assumption 3 and the ergodic theorem,

$$\frac{1}{n}\{l_n(0.5, \vartheta, \vartheta, \boldsymbol{\eta})\} \to E \log(g(X_1|X_0^p; \vartheta, \boldsymbol{\eta})) \tag{19}$$

holds almost surely for every fixed $(\vartheta, \boldsymbol{\eta}) \in \Theta \times \boldsymbol{H}$. As in Ferguson (1996), one can show that (19) holds almost surely and uniformly over $(\vartheta, \boldsymbol{\eta}) \in \Theta \times \boldsymbol{H}$. The claim follows by theorem 1 in Frydman (1980) using Lemma 3.
(ii) Let

$$Q(\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta}) = E \log \left(\frac{g_{\text{mix}}(X_1|X_0^p; \boldsymbol{\psi})}{g(X_1|X_0^p; \vartheta_0, \boldsymbol{\eta}_0)} \right).$$

From Cho and White (2007) we have

$$R_n = 2\{l_n(\widehat{\alpha}, \widehat{\vartheta}_1, \widehat{\vartheta}_2, \widehat{\boldsymbol{\eta}}) - l_n(1/2, \widehat{\vartheta}_0, \widehat{\vartheta}_0, \widehat{\boldsymbol{\eta}}_0)\} = O_P(1).$$

Using $0 \leq M_n \leq R_n$ and the properties of the penalty function $p(\alpha)$ we get $0 \leq M_n - 2\{p(\widehat{\alpha}^*) - p(0.5)\} \leq R_n$ and therefore $p(\widehat{\alpha}^*) = O_P(1)$. Therefore there exists an $\delta > 0$ for which $P(\delta \leq \widehat{\alpha}^* \leq 1 - \delta) \to 1$, $n \to \infty$, holds and we can suppose that $\alpha \in [\delta, 1 - \delta]$. By the ergodic theorem and Assumption 4 we get under the null distribution

$$\frac{1}{n}\{pl_n(\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta}) - pl_n(0.5, \vartheta_0, \vartheta_0, \boldsymbol{\eta}_0)\} \to Q(\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta}) \tag{20}$$

almost surely and uniformly over $(\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta}) \in [\delta, 1 - \delta] \times \Theta^2 \times \boldsymbol{H}$. Let ω be a point in the sample space for which (20) is true and note that the set of all such points has probability 1.
Suppose for a ω the claim of the theorem is not true and, for example (the procedure for the other parameters is the same), $\widehat{\vartheta}_1^*$ does not converge to ϑ_0. There must exist a subsequence (n') such that $\widehat{\vartheta}_{1n'}^* \to \vartheta' \neq \vartheta_0$. Consider

$$\Omega' = \{(\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta}) : |\vartheta_1 - \vartheta_0| \geq \epsilon, \alpha \in [\delta, 1 - \delta]\},$$

where $\epsilon = |\vartheta' - \vartheta_0|/2$. Then for all large n', $(\widehat{\alpha}^*, \widehat{\vartheta}_1^*, \widehat{\vartheta}_2^*, \widehat{\boldsymbol{\eta}}^*)$ at the sample point ω, belongs to the subset. By Assumption 2 and Lemma 3 $Q(\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta}) < 0$ for all $(\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta}) \in \Omega'$. It then follows that

$$pl_{n'}(\widehat{\alpha}^*, \widehat{\vartheta}_1^*, \widehat{\vartheta}_2^*, \widehat{\boldsymbol{\eta}}^*) - pl_{n'}(0.5, \widehat{\vartheta}_0, \widehat{\vartheta}_0, \widehat{\boldsymbol{\eta}}_0) < 0$$

for all large enough n'. But this is a contradiction to $(\widehat{\alpha}^*, \widehat{\vartheta}_1^*, \widehat{\vartheta}_2^*, \widehat{\boldsymbol{\eta}}^*)$ being modified maximum likelihood estimator and so $\widehat{\vartheta}_{1n'}^* \to \vartheta_0$ on ω. Thus $\widehat{\vartheta}_{1n'}^* \to \vartheta_0$ almost surely. □

Proof of Theorem 2. We have

$$\begin{aligned} M_n &= 2\big(pl_n(\widehat{\alpha}^*, \widehat{\vartheta}_1^*, \widehat{\vartheta}_2^*, \widehat{\boldsymbol{\eta}}^*) - pl_n(1/2, \vartheta_0, \vartheta_0, \boldsymbol{\eta}_0)\big) + 2\big(pl_n(1/2, \vartheta_0, \vartheta_0, \boldsymbol{\eta}_0) - pl_n(1/2, \widehat{\vartheta}_0, \widehat{\vartheta}_0, \widehat{\boldsymbol{\eta}}_0)\big) \\ &= r_{1n} + r_{2n}. \end{aligned}$$

First examine r_{1n} : Write

$$\begin{aligned} r_{1n}(\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta}) &= 2 \sum_{i=1}^n \log(1 + \delta_i) + 2p(\alpha) - 2p(1/2), \\ \delta_i &= (1 - \alpha) \left\{ \frac{g(X_i|X_{i-1}^p; \vartheta_1, \boldsymbol{\eta})}{g(X_i|X_{i-1}^p; \vartheta_0, \boldsymbol{\eta}_0)} - 1 \right\} + \alpha \left\{ \frac{g(X_i|X_{i-1}^p; \vartheta_2, \boldsymbol{\eta})}{g(X_i|X_{i-1}^p; \vartheta_0, \boldsymbol{\eta}_0)} - 1 \right\}. \end{aligned}$$

Then using Assumptions 1 and 3 - 5, we obtain the quadratic expansion (see Appendix B)

$$
\begin{aligned}
r_{1n}(\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta}) \leq & 2\sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} \delta_i^2 + \frac{2}{3}\sum_{i=1}^{n} \delta_i^3 \\
\leq & 2\sum_{i=1}^{n} \big((\eta_1 - \eta_{1,0})U_{i,1} + \ldots + (\eta_d - \eta_{d,0})U_{i,d} + m_1 Y_i + m_2 Z_i\big) \\
& - \sum_{i=1}^{n} \big((\eta_1 - \eta_{1,0})U_{i,1} + \ldots + (\eta_d - \eta_{d,0})U_{i,d} + m_1 Y_i + m_2 Z_i\big)^2 (1 + o_P(1)) \\
& + o_P(1),
\end{aligned}
\tag{21}
$$

where

$$
m_1 = (1-\alpha)(\vartheta_1 - \vartheta_0) + \alpha(\vartheta_2 - \vartheta_0), \qquad m_2 = (1-\alpha)(\vartheta_1 - \vartheta_0)^2 + \alpha(\vartheta_2 - \vartheta_0)^2.
$$

Note that $m_2 \geq 0$. Using Assumption 5 we may orthogonalize

$$
\begin{aligned}
\tilde{U}_{i,1} &= U_{i,1}, \qquad \tilde{U}_{i,2} = U_{i,2} - \frac{E\tilde{U}_{1,1}U_{1,2}}{E\tilde{U}_{1,1}^2}\tilde{U}_{i,1}, \qquad \ldots \\
\tilde{Y}_i &= Y_i - \sum_{j=1}^{d} \frac{E\tilde{U}_{1,j}Y_1}{E\tilde{U}_{1,j}^2}\tilde{U}_{i,j}, \qquad \tilde{Z}_i = Z_i - \frac{EZ_1\tilde{Y}_1}{E\tilde{Y}_1^2}\tilde{Y}_i - \sum_{j=1}^{d} \frac{E\tilde{U}_{1,j}Z_1}{E\tilde{U}_{1,j}^2}\tilde{U}_{i,j}.
\end{aligned}
\tag{22}
$$

By Assumption 5, given $\eta_1, \ldots, \eta_d, m_1, m_2$ there exist unique constants $t_1, \ldots, t_{d+2}$ for which

$$
\begin{aligned}
&(\eta_1 - \eta_{1,0})U_{i,1} + \ldots (\eta_d - \eta_{d,0})U_{i,d} + m_1 Y_i + m_2 Z_i \\
=& t_1\tilde{U}_{i,1} + t_2\tilde{U}_{i,2} + \ldots t_d\tilde{U}_{i,d} + t_{d+1}\tilde{Y}_i + t_{d+2}\tilde{Z}_i
\end{aligned}
\tag{23}
$$

where we have in particular that $t_{d+2} = m_2 \geq 0$. Since $\tilde{U}_{i,j}$, $j = 1, \ldots, d$, $\tilde{Y}_i$ and $\tilde{Z}_i$ are mutually orthogonal, we have for the mixed terms that

$$
\sum_{i=1}^{n} \tilde{U}_{i,j}\, \tilde{U}_{i,k} = o_P\big(\sum_{i=1}^{n} \tilde{U}_{i,j}^2\big), \quad j \neq k,
$$

and similarly for the other mixed terms. Thus, setting

$$
\begin{aligned}
q(t_1, \ldots, t_{d+2}) =& 2\sum_{i=1}^{n} \big(t_1\tilde{U}_{i,1} + \ldots t_d\tilde{U}_{i,d} + t_{d+1}\tilde{Y}_i + t_{d+2}\tilde{Z}_i\big) \\
& - \sum_{i=1}^{n} \big(t_1^2\tilde{U}_{i,1}^2 + \ldots t_d^2\tilde{U}_{i,d}^2 + t_{d+1}^2\tilde{Y}_i^2 + t_{d+2}^2\tilde{Z}_i^2\big)
\end{aligned}
$$

we can conclude from (21) that

$$
r_{1n}(\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta}) \leq q(t_1, \ldots, t_{d+2})\big(1 + o_P(1)\big) + o_P(1),
$$

where $t_1, \ldots t_{d+2}$ are determined as in (23). Maximizing $q(t_1, \ldots, t_{d+2})$ under the sign constraint $t_{d+2} \geq 0$ we get

$$
\begin{aligned}
(\tilde{t}_1, \ldots, \tilde{t}_{d+2}) =& \arg\max_{t_1, \ldots, t_{d+2}} q(t_1, \ldots, t_{d+2}) \\
=& \left(\frac{\sum \tilde{U}_{i,1}}{\sum (\tilde{U}_{i,1})^2}, \ldots, \frac{\sum \tilde{U}_{i,d}}{\sum (\tilde{U}_{i,d})^2}, \frac{\sum \tilde{Y}_i}{\sum \tilde{Y}_i^2}, \frac{(\sum \tilde{Z}_i)^+}{\sum \tilde{Z}_i^2} \right)
\end{aligned}
\tag{24}
$$

and therefore an upper bound for r_{1n} is given by

$$r_{1n}(\widehat{\alpha}^*, \widehat{\vartheta}_1^*, \widehat{\vartheta}_2^*, \widehat{\boldsymbol{\eta}}^*) \leq \frac{(\sum \tilde{U}_{i,1})^2}{\sum (\tilde{U}_{i,1})^2} + \ldots + \frac{(\sum \tilde{U}_{i,d})^2}{\sum (\tilde{U}_{i,d})^2} + \frac{(\sum \tilde{Y}_i)^2}{\sum \tilde{Y}_i^2} + \frac{((\sum \tilde{Z}_i)^+)^2}{\sum \tilde{Z}_i^2} + o_P(1). \tag{25}$$

For $\alpha = 1/2$ and the values $\tilde{\vartheta}_1^*, \tilde{\vartheta}_2^*$ and $\tilde{\boldsymbol{\eta}}^*$ determined by the maximizers $\tilde{t}_1, \ldots, \tilde{t}_{d+2}$ by (24), we see that this upper bound is attained.
Expanding r_{2n} in a similar way as r_{1n}, we obtain

$$-r_{2n} = \frac{(\sum \tilde{U}_{i,1})^2}{\sum (\tilde{U}_{i,1})^2} + \ldots + \frac{(\sum \tilde{U}_{i,d})^2}{\sum (\tilde{U}_{i,d})^2} + \frac{(\sum \tilde{Y}_i)^2}{\sum \tilde{Y}_i^2} + o_P(1).$$

Therefore,

$$M_n = \frac{\left((\sum \tilde{Z}_i)^+\right)^2}{\sum \tilde{Z}_i^2} + o_P(1).$$

Since $(\tilde{Z}_i)_{i\geq 1}$ form a square integrable stationary martingale difference sequences, the result follows from the ergodic theorem (applied to the denominator) and the central limit theorem for stationary ergodic martingale difference sequences (applied to the numerator). □

Proof of Lemma 1. b. First, we consider model (4). Let $\mu(\zeta, \phi_1, \ldots, \phi_p; x_0^p) = \zeta + \phi_1 x_0 + \ldots + \phi_p x_{1-p}$. Then

$$\begin{aligned}
U_1^{\zeta} &= \frac{\frac{\partial}{\partial \mu} f(X_1; \mu(\zeta, \phi_1, \ldots, \phi_p; X_0^p), \sigma)}{f(X_1; \mu(\zeta, \phi_1, \ldots, \phi_p; X_0^p), \sigma)} \\
U_1^{\phi_\tau} &= \frac{\frac{\partial}{\partial \mu} f(X_1; \mu(\zeta, \phi_1, \ldots, \phi_p; X_0^p), \sigma) X_{1-\tau}}{f(X_1; \mu(\zeta, \phi_1, \ldots, \phi_p; X_0^p), \sigma)} = U_1^{\zeta} X_{1-\tau}, \quad \tau = 1, \ldots, p, \\
Y_1 &= \frac{\frac{\partial}{\partial \sigma} f(X_1; \mu(\zeta, \phi_1, \ldots, \phi_p; X_0^p), \sigma)}{f(X_1; \mu(\zeta, \phi_1, \ldots, \phi_p; X_0^p), \sigma)}, \\
Z_1 &= \frac{\frac{\partial^2}{\partial^2 \sigma} f(X_1; \mu(\zeta, \phi_1, \ldots, \phi_p; X_0^p), \sigma)}{f(X_1; \mu(\zeta, \phi_1, \ldots, \phi_p; X_0^p), \sigma)},
\end{aligned}$$

where we denote the $U_{1,j}$ by U_1^{ζ} and $U_1^{\phi_\tau}$. The covariance matrix of $(U_1^{\zeta}, U_1^{\phi_1}, \ldots, U_1^{\phi_p}, Y_1, Z_1)$ is non-degenerate if and only if these random variables are linearly independent (in L_2). Therefore, suppose that for some constants b_j,

$$b_1 Z_1 + b_2 Y_1 + b_3 U_1^{\zeta} + \sum_{\tau=1}^{p} b_{3+\tau} U_1^{\phi_\tau} = 0 \quad a.s. \tag{26}$$

holds. Since the distribution of $X_1, \ldots, X_{1-p}$ is equivalent to Lebesgue measure on $\mathbb{R}^{p+1}$, (26) is equivalent to

$$b_1 \frac{\partial}{\partial^2 \sigma} f + b_2 \frac{\partial}{\partial \sigma} f + b_3 \frac{\partial}{\partial \mu} f + \sum_{\tau=1}^{p} b_{3+\tau} \frac{\partial}{\partial \mu} f x_{1-\tau} = 0 \quad \text{Leb.} - a.s. \tag{27}$$

with $f = f(x_1; \mu(\zeta, \phi_1, \ldots, \phi_p; x_0^p), \sigma)$. From (14) it follows that $b_1 = b_2 = 0$ and

$$b_3 + \sum_{\tau=1}^{p} b_{3+\tau} x_{1-\tau} = 0 \quad \text{Leb.} - a.s.,$$

so that $b_3 = \ldots = b_{3+p} = 0$.

For model (5), let $\sigma(\vartheta, \phi_1, \ldots, \phi_p; x_0^p) = \big(\vartheta + \phi_1 x_0^2 + \ldots + \phi_d x_{1-p}^2\big)^{(1/2)}$. Then setting $\sigma = \sigma(\vartheta, \phi_1, \ldots, \phi_p; X_0^p)$,

$$\begin{aligned} U_1^{\phi_j} &= \frac{\frac{\partial}{\partial \sigma} f\big(X_1;\sigma\big) X_{1-j}^2}{f\big(X_1;\sigma\big)2\sigma}, \quad j = 1, \ldots, p, \\ Y_1 &= \frac{\frac{\partial}{\partial \sigma} f\big(X_1;\sigma\big)/(2\sigma)}{f\big(X_1;\sigma\big)}, \qquad Z_1 = \frac{\frac{\partial^2}{\partial^2 \sigma} f\big(X_1;\sigma\big)/(4\sigma^2) - \frac{\partial}{\partial \sigma} f\big(X_1;\sigma\big)/(4\sigma^3)}{f\big(X_1;\sigma\big)}. \end{aligned}$$

Again, the covariance is non-degenerate if and only if these random variables are linearly independent in L_2. Therefore, suppose that for constants b_j,

$$b_1 Z_1 + b_2 Y_1 + \sum_{j=1}^{p} b_{j+2} U_1^{\phi_j} = 0 \quad a.s. \tag{28}$$

Again, (28) is equivalent to

$$b_1 \Big(\frac{\partial^2}{\partial^2 \sigma} f/(2\sigma) - \frac{\partial}{\partial \sigma} f/(2\sigma) \Big) + b_2 \frac{\partial}{\partial \sigma} f + \sum_{j=1}^{p} b_{j+2} \frac{\partial}{\partial \sigma} f x_{1-j}^2 = 0 \quad \text{Leb.} - a.s.$$

where $f = f\big(x_1; \sigma(\vartheta, \phi_1, \ldots, \phi_p; x_0^p)\big)$. From (14), $b_1 = 0$ (as coefficient of $\frac{\partial^2}{\partial^2 \sigma} f$) and

$$b_2 + \sum_{j=1}^{p} b_{j+2} x_{1-j}^2 = 0 \quad \text{Leb.} - a.s.,$$

so that $b_2 = b_3 = \ldots = b_{p+2} = 0$. □

Lemma 4. *Let $f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\big(-\frac{(x-\mu)^2}{2\sigma^2}\big)$ be the pdf of a normally distributed random variable with expectation μ and standard deviation $\sigma > 0$. If*

$$a_1 \frac{\partial f(x;\mu,\sigma)}{\partial \mu} + a_2 \frac{\partial f(x;\mu,\sigma)}{\partial \sigma} + a_3 \frac{\partial^2 f(x;\mu,\sigma)}{\partial^2 \sigma} = 0 \quad \textit{for Lebesgue-a.e. } x \tag{29}$$

for any (μ,σ) and $a_1, a_2, a_3 \in \mathbb{R}$, then $a_1 = a_2 = a_3 = 0$.

Proof of Lemma 4. Let

$$\varphi(t;\mu,\sigma) = \exp\big(it\mu - \frac{\sigma^2 t^2}{2}\big).$$

Taking the Fourier transform in (29) and interchanging integral and derivative gives

$$a_1 \frac{\partial \varphi(t;\mu,\sigma)}{\partial \mu} + a_2 \frac{\partial \varphi(t;\mu,\sigma)}{\partial \sigma} + a_3 \frac{\partial^2 \varphi(t;\mu,\sigma)}{\partial^2 \sigma} = 0 \quad \text{for all } t \in \mathbb{R}. \tag{30}$$

Since

$$\frac{\partial \varphi(t;\mu,\sigma)}{\partial \sigma} = -\sigma t^2 \varphi(t;\mu,\sigma), \quad \frac{\partial^2 \varphi(t;\mu,\sigma)}{\partial^2 \sigma} = (\sigma^2 t^4 - t^2)\varphi(t;\mu,\sigma), \quad \frac{\partial \varphi(t;\mu,\sigma)}{\partial \mu} = it\varphi(t;\mu,\sigma).$$

we get from (30) after dividing by $\varphi(t;\mu,\sigma)$ that

$$a_1 it - a_2 \sigma t^2 + a_3(\sigma^2 t^4 - t^2) = 0 \quad \text{for all } t \in \mathbb{R},$$

which is equivalent to

$$a_1 it + (-a_2\sigma - a_3)t^2 + a_3 \sigma^2 t^4 = 0 \quad \text{for all } t \in \mathbb{R},$$

from which we easily conclude $a_1 = a_2 = a_3 = 0$. □

Lemma 5. *For the normal distribution, Assumption 5 is satisfied for the model* (3).

Proof of Lemma 5. Arguing as in the proof of Lemma 1, we arrive at

$$b_1 \frac{\partial}{\partial \mu} f + b_2 \frac{\partial}{\partial \sigma} f + \frac{b_3}{\sigma} \frac{\partial}{\partial \sigma} f x_{1-j_0}^2 + \sum_{\tau=1}^{p} b_{\tau+3} \frac{\partial}{\partial \mu} f x_{1-\tau} = 0 \quad \text{Leb.} - a.s., \tag{31}$$

since $\sigma \frac{\partial^2}{\partial^2 \mu} f = \frac{\partial}{\partial \sigma} f$ holds for the normal distribution. From Lemma 4 it follows that

$$b_2 + \frac{b_3}{\sigma} x_{1-j_0}^2 = 0, \qquad b_1 + \sum_{\tau=1}^{p} b_{\tau+3} x_{1-\tau} = 0, \quad \text{Leb.} - a.s.,$$

so that $b_2 = b_3 = 0$ and $b_1 = b_4 = \ldots = b_{p+3} = 0$. □

A.2 EM-test and proofs of Section 3.2

We now describe the EM-test, which is most conveniently accomplished in form of the following algorithm.

Step 0. Choose the initial values $0 < \alpha_1 < \alpha_2 < \ldots < \alpha_J = 0.5$. Compute

$$(\tilde{\vartheta}_0, \tilde{\boldsymbol{\eta}}_0) = \arg\max_{\vartheta, \boldsymbol{\eta}} pl_n(0.5, \vartheta, \vartheta, \boldsymbol{\eta}).$$

Put $j = 1$ and $k = 0$.

Step 1. Put $\alpha_j^{(k)} = \alpha_j$.

Step 2. Compute

$$(\vartheta_{1j}^{(k)}, \vartheta_{2j}^{(k)}, \boldsymbol{\eta}_j^{(k)}) = \arg\max_{\vartheta_1, \vartheta_2, \boldsymbol{\eta}} pl_n(\alpha_j^{(k)}, \vartheta_1, \vartheta_2, \boldsymbol{\eta})$$

and

$$M_n^{(k)}(\alpha_j) = 2\left(pl_n(\alpha_j^{(k)}, \vartheta_{1j}^{(k)}, \vartheta_{2j}^{(k)}, \boldsymbol{\eta}_j^{(k)}) - pl_n(0.5, \tilde{\vartheta}_0, \tilde{\vartheta}_0, \tilde{\boldsymbol{\eta}}_0)\right)$$

Step 3. Compute for $i = 1, \ldots, n$ the weights

$$w_{ij}^{(k)} = \frac{\alpha_j^{(k)} g(X_i | X_{i-1}^p; \vartheta_{2j}^{(k)}, \boldsymbol{\eta}_j^{(k)})}{(1 - \alpha_j^{(k)})\, g(X_i | X_{i-1}^p; \vartheta_{1j}^{(k)}, \boldsymbol{\eta}_j^{(k)}) + \alpha_j^{(k)} g(X_i | X_{i-1}^p; \vartheta_{2j}^{(k)}, \boldsymbol{\eta}_j^{(k)})}.$$

Compute the estimators

$$\begin{aligned}
\alpha_j^{(k+1)} &= \arg\max_{\alpha} \left(\left(n - \sum_{i=1}^n w_{ij}^{(k)}\right) \log(1-\alpha) + \sum_{i=1}^n w_{ij}^{(k)} \log(\alpha) + p(\alpha) \right) \\
\vartheta_{1j}^{(k+1)} &= \arg\max_{\vartheta_1} \left(\sum_{i=1}^n (1 - w_{ij}^{(k)}) \log g(X_i | X_{i-1}^p; \vartheta_1, \boldsymbol{\eta}_j^{(k)}) \right) \\
\vartheta_{2j}^{(k+1)} &= \arg\max_{\vartheta_2} \left(\sum_{i=1}^n w_{ij}^{(k)} \log g(X_i | X_{i-1}^p; \vartheta_2, \boldsymbol{\eta}_j^{(k)}) \right)
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\eta}_j^{(k+1)} = \arg\max_{\boldsymbol{\phi}} \Bigg(&\sum_{i=1}^n (1 - w_{ij}^{(k)}) \log g(X_i | X_{i-1}^p; \vartheta_{1j}^{(k+1)}, \boldsymbol{\eta}) \\
&+ \sum_{i=1}^n w_{ij}^{(k)} \log g(X_i | X_{i-1}^p; \vartheta_{2j}^{(k+1)}, \boldsymbol{\eta}) \Bigg).
\end{aligned}$$

Compute

$$M_n^{(k+1)}(\alpha_j) = 2\left\{pl_n(\alpha_j^{(k+1)}, \vartheta_{1j}^{(k+1)}, \vartheta_{2j}^{(k+1)}, \boldsymbol{\eta}_j^{(k+1)}) - pl_n(0.5, \tilde{\vartheta}_0, \tilde{\vartheta}_0, \tilde{\boldsymbol{\eta}}_0)\right\},$$

put $k = k + 1$ and repeat Step 3 for a fixed number of iterations K.

Step 4. Put $j = j + 1$, $k = 0$ and go to Step 1, until $j = J$.

Step 5. Compute the test statistic

$$EM_n^{(K)}(\mathcal{J}) = \max_{j=1,\dots,J} M_n^{(K)}(\alpha_j).$$

In the above construction of the EM-test, we actually use an ECM algorithm (Meng and Rubin, 1993) since the EM algorithm would require joint maximization to obtain the update $(\vartheta_{1j}^{(k+1)}, \vartheta_{2j}^{(k+1)}, \boldsymbol{\eta}_j^{(k+1)})$. If $\boldsymbol{\eta}$ is highdimensional, this could be further refined by maximizing successively over the components of $\boldsymbol{\eta}$.

Proof of Theorem 3. It is clear that

$$EM_n^{(K)}(\mathcal{J}) \leq M_n \leq \frac{\left((\sum \tilde{Z}_i)^+\right)^2}{\sum \tilde{Z}_i^2} + o_P(1).$$

Since one of the starting values in the EM-test is assumed to be $\alpha_J = 0.5$, and since the ECM algorithm only increases the value of the likelihood (even though applied to a penalized quasi likelihood, see Appendix B), using the same argument as in the end of the proof of Theorem 2, we have

$$EM_n^{(K)}(\mathcal{J}) \geq EM_n^{(K)}(\{1/2\}) \geq \frac{\left((\sum \tilde{Z}_i)^+\right)^2}{\sum \tilde{Z}_i^2} + o_P(1).$$

□

Proof of Theorem 4

We give an outline of the proof which is quite similar to that of theorem 2 in Chen and Li (2009), only the additional linear autoregression must be taken care of. For full details see Ketterer (2011). Since we assume that the innovations $(\epsilon_k)_k$ are independent $N(0, \sigma^2)$-distributed, the conditional density (w.r.t. Lebesgue measure on $\mathbb{R}$) of X_t given $X_{t-1}^p = x_{t-1}^p$ and $S_t = i$ is given by

$$g(x_t|x_{t-1}^p; \zeta_i, \boldsymbol{\phi}, \sigma) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_t - \zeta_i - \sum_{j=1}^p \phi_j x_{t-j})^2}{2\sigma^2}\right).$$

In the following, let $(\overline{\alpha}, \overline{\zeta}_1, \overline{\zeta}_2, \overline{\boldsymbol{\phi}}, \overline{\sigma})$ be estimators of an EM-step. Then we write that a statement holds for example for $\overline{\alpha}$ if and only if it holds for every $\alpha_j^{(k)}$, $j = 1, \dots, J$ and $k = 1, \dots, K$.

Lemma 6. *For each given $\overline{\alpha} \in (0, 0.5]$ we have under the null model*

$$\overline{\sigma} - \sigma_0 = o_P(1), \qquad \overline{\boldsymbol{\phi}} - \boldsymbol{\phi}_0 = o_P(1),$$
$$\overline{\zeta}_1 - \zeta_0 = o_P(1), \qquad \overline{\zeta}_2 - \zeta_0 = o_P(1).$$

Proof. Since we assume $(X_k)_k$ to be a causal AR(p) process under the null model we know that the order of the autoregressive process is uniquely defined and that the parameters are identifiable (cf.

Kreiss and Neuhaus, 2006). Assuming $\sigma_0 \in [\delta, \infty)$, $\delta > 0$, we have

$$\begin{aligned} & E\left|\log\big(g(X_1|X_0^p;\zeta_0,\boldsymbol{\phi}_0,\sigma_0)\big)\right| \\ \le\ & E\left[\frac{1}{2}\frac{(X_1-\zeta_0-\sum_{j=1}^p \phi_{j,0}X_{1-j})^2}{\sigma_0^2}\right] + \left|\frac{1}{2}\log(2\pi\sigma_0^2)\right| \\ =\ & \frac{1}{2\sigma_0^2}E\epsilon_1^2 + \left|\frac{1}{2}\log(2\pi\sigma_0^2)\right| < \infty. \end{aligned}$$

Therefore, the ergodic theorem applies to $(\log g(X_t|X_{t-1}^p;\zeta_0,\boldsymbol{\phi}_0,\sigma_0))_t$. Further, we have

$$l_n(\overline{\alpha},\overline{\zeta}_1,\overline{\zeta}_2,\overline{\boldsymbol{\phi}},\overline{\sigma}) - l_n(0.5,\zeta_0,\zeta_0,\boldsymbol{\phi}_0,\sigma_0) \ge \min_{j=1,\ldots,J}\{p(\alpha_j)-p(0.5)\} > -\infty \tag{32}$$

since $0.5 = \arg\max_\alpha p(\alpha)$ and

$$\begin{aligned} l_n(\alpha_j^{(k)},\zeta_{1j}^{(k)},\zeta_{2j}^{(k)},\boldsymbol{\phi}_j^{(k)},\sigma_j^{(k)}) + p(\alpha_j^{(k)}) &= pl_n(\alpha_j^{(k)},\zeta_{1j}^{(k)},\zeta_{2j}^{(k)},\boldsymbol{\phi}_j^{(k)},\sigma_j^{(k)}) \\ &\ge pl_n(\alpha_j^{(0)},\zeta_{1j}^{(0)},\zeta_{2j}^{(0)},\boldsymbol{\phi}_j^{(0)},\sigma_j^{(0)}) \\ &\ge pl_n(\alpha_j,\zeta_0,\zeta_0,\boldsymbol{\phi}_0,\sigma_0) \\ &= l_n(\alpha_j,\zeta_0,\zeta_0,\boldsymbol{\phi}_0,\sigma_0) + p(\alpha_j) \end{aligned}$$

for every $j = 1,\ldots,J$ and $k = 1,\ldots,K$ by the *EM-property*. Now the result follows using the argument as in theorem 2 for the i.i.d. case in Wald (1949). □

From now on we assume without loss of generality $\zeta_0 = 0$ and $\sigma_0 = 1$, and, to ensure readibility, we restrict attention to the case $p = 1$. Let

$$\begin{aligned} Y_t &:= \left.\frac{\frac{\partial}{\partial\zeta}g(X_t|X_{t-1};\zeta,\boldsymbol{\phi}_0,1)}{g(X_t|X_{t-1};0,\boldsymbol{\phi}_0,1)}\right|_{\zeta=0} = \epsilon_t, \\ Z_t &:= \frac{1}{2}\left.\frac{\frac{\partial^2}{\partial^2\zeta}g(X_t|X_{t-1};\zeta,\boldsymbol{\phi}_0,1)}{g(X_t|X_{t-1};0,\boldsymbol{\phi}_0,1)}\right|_{\zeta=0} = (\epsilon_t^2-1)/2, \\ U_t &:= \frac{1}{6}\left.\frac{\frac{\partial^3}{\partial^3\zeta}g(X_t|X_{t-1};\zeta,\boldsymbol{\phi}_0,1)}{g(X_t|X_{t-1};0,\boldsymbol{\phi}_0,1)}\right|_{\zeta=0} = (\epsilon_t^3-3\epsilon_t)/6, \\ V_t &:= \frac{1}{24}\left.\frac{\frac{\partial^4}{\partial^4\zeta}g(X_t|X_{t-1};\zeta,\boldsymbol{\phi}_0,1)}{g(X_t|X_{t-1};0,\boldsymbol{\phi}_0,1)}\right|_{\zeta=0} = (\epsilon_t^4-6\epsilon_t^2+3)/24, \\ W_t &:= \left.\frac{\partial_\phi g(X_t|X_{t-1};0,\boldsymbol{\phi},1)}{g(X_t|X_{t-1}^p;0,\boldsymbol{\phi}_0,1)}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0} = X_{t-1}\epsilon_t. \end{aligned}$$

Note that by causality, X_{t-1} and ϵ_t are independent, and thus a direct computation shows that Y_t, Z_t, U_t, V_t and W_t are mutually orthogonal.

Lemma 7. *For each $\alpha_j \in (0, 0.5]$, under the null model we have whenever $\overline{\alpha} - \alpha_j = o_P(1)$ that*

$$\begin{aligned} \overline{\sigma}^2 - 1 = O_P(n^{-1/4}), \quad \overline{\phi} - \phi_0 = O_P(n^{-1/2}), \\ \overline{\zeta}_1 = O_P(n^{-1/8}), \quad \overline{\zeta}_2 = O_P(n^{-1/8}). \end{aligned} \tag{33}$$

Proof. First we intend to find an appropriate asymptotic upper bound for

$$2\{pl_n(\overline{\alpha},\overline{\zeta}_1,\overline{\zeta}_2,\overline{\phi},\overline{\sigma}) - pl_n(0.5,0,0,\phi_0,1)\}.$$

To this end, write

$$2\{pl_n(\overline{\alpha}, \overline{\zeta}_1, \overline{\zeta}_2, \overline{\phi}, \overline{\sigma}) - pl_n(0.5, 0, 0, \phi_0, 1)\} = 2\sum_{t=1}^{n} \log(1+\overline{\delta}_t) + 2\{p(\overline{\alpha}) - p(0.5)\},$$

$$\overline{\delta}_t = (1-\overline{\alpha})\frac{g(X_t|X_{t-1}; \overline{\zeta}_1, \overline{\phi}, \overline{\sigma})}{g(X_t|X_{t-1}; 0, \phi_0, 1)} + \overline{\alpha}\frac{g(X_t|X_{t-1}; \overline{\zeta}_2, \overline{\phi}, \overline{\sigma})}{g(X_t|X_{t-1}; 0, \phi_0, 1)}.$$

Using Taylor expansion we get

$$\overline{\delta}_t = \overline{s}_1 Y_t + \overline{s}_2 Z_t + \overline{s}_3 U_t + \overline{s}_4 V_t + \overline{s}_5 W_t + \overline{\epsilon}_{tn} \tag{34}$$

with an appropriate remainder $\overline{\epsilon}_{tn}$, where

$$\begin{aligned} &\overline{s}_1 = \overline{m}_1, \quad \overline{s}_2 = \overline{m}_2 + (\overline{\sigma}^2 - 1), \quad \overline{s}_3 = \overline{m}_3, \\ &\overline{s}_4 = \overline{m}_4 - 3\overline{m}_2^2, \quad \overline{s}_5 = (\overline{\phi} - \phi_0), \\ &\overline{m}_k = (1-\overline{\alpha})\overline{\vartheta}_1^k + \overline{\alpha}\overline{\vartheta}_2^k, \qquad k = 1, \ldots, 6. \end{aligned}$$

Using the orthogonality of Y_t, Z_t, U_t, V_t and W_t, one proves that (see Ketterer 2011)

$$\begin{aligned} &2\{pl_n(\overline{\alpha}, \overline{\zeta}_1, \overline{\zeta}_2, \overline{\phi}, \overline{\sigma}) - pl_n(0.5, 0, 0, \phi_0, 1)\} \\ &\le 2\overline{s}_1 \sum_{t=1}^{n} Y_t - \overline{s}_1^2 \sum_{t=1}^{n} Y_t^2\{1+o_p(1)\} + 2\overline{s}_2 \sum_{t=1}^{n} Z_t - \overline{s}_2^2 \sum_{t=1}^{n} Z_t^2\{1+o_p(1)\} \\ &+2\overline{s}_3 \sum_{t=1}^{n} U_t - \overline{s}_3^2 \sum_{t=1}^{n} U_t^2\{1+o_p(1)\} + 2\overline{s}_4 \sum_{t=1}^{n} V_t - \overline{s}_4^2 \sum_{t=1}^{n} V_t^2\{1+o_p(1)\} \\ &+2\overline{s}_5 \sum_{t=1}^{n} W_{1t} - \overline{s}_5^2 \sum_{t=1}^{n} W_{1t}^2\{1+o_p(1)\} + 2\{p(\alpha_j) + p(0.5)\} + o_P(1). \end{aligned} \tag{35}$$

From (35) we wish to conclude that

$$\overline{s}_j = O_P(n^{-1/2}), \quad j = 1, \ldots, 5. \tag{36}$$

Consider $\overline{s}_1$: By maximizing the quadratic function,

$$2\overline{s}_1 \sum_{t=1}^{n} Y_t - \overline{s}_1^2 \sum_{t=1}^{n} Y_t^2\{1+o_P(1)\} \le \frac{(\sum_{t=1}^{n} Y_t)^2}{\sum_{t=1}^{n} Y_t^2}\{1+o_P(1)\} = O_P(1).$$

where the last equality follows from the CLT (applied to the numerator) and the SLLN (applied to the denominator). Therefore,

$$\begin{aligned} 0 &\le 2\{pl_n(\overline{\alpha}, \overline{\zeta}_1, \overline{\zeta}_2, \overline{\phi}, \overline{\sigma}) - pl_n(0.5, 0, 0, \phi_0, 1)\} \\ &\le 2\overline{s}_1 \sum_{t=1}^{n} Y_t - \overline{s}_1^2 \sum_{t=1}^{n} Y_t^2\{1+o_P(1)\} + O_P(1) \\ &= O_P(1) \end{aligned}$$

where the first inequality is due to the *EM-property*. Hence

$$2\overline{s}_1 \sum_{t=1}^{n} Y_t - \overline{s}_1^2 \sum_{t=1}^{n} Y_t^2\{1+o_P(1)\} = O_P(1)$$

so that the two terms need to be balanced, which leads to $\overline{s}_1 = O_P(n^{-1/2})$. For $\overline{s}_j$, $j = 2, \ldots, 5$, the argument is similar. Analogously, we get

$$\overline{s}_j = O_P(n^{-1/2}), \quad j = 2, 3, 4, 5. \tag{37}$$

By the definition of $\overline{s}_5$ it immediately follows that

$$\overline{\phi} - \phi = O_P(n^{-1/2})$$

It can be shown (see Ketterer 2011) that

$$\overline{\zeta}_1^4 = O_P\Big(\sum_{j=1}^{5} |\overline{s}_j|\Big), \ \overline{\zeta}_2^4 = O_P\Big(\sum_{j=1}^{5} |\overline{s}_j|\Big) \text{ and } (\overline{\sigma}^2 - 1)^2 = O_P\Big(\sum_{j=1}^{5} |\overline{s}_j|\Big)$$

which together with (37) implies (33).

□

Let $(\overline{\alpha}, \overline{\zeta}_1, \overline{\zeta}_2, \overline{\phi}, \overline{\sigma})$ be some EM-estimator. We define

$$\begin{aligned} H_n(\alpha) &= \left(n - \sum_{t=1}^{n} \overline{w}_t\right) \log(1-\alpha) + \sum_{t=1}^{n} \overline{w}_t \log(\alpha) + p(\alpha) \\ &=: R_n(\alpha) + p(\alpha), \end{aligned}$$

where

$$\overline{w}_t = \frac{\overline{\alpha} g(X_t|X_{t-1}; \overline{\zeta}_2, \overline{\phi}, \overline{\sigma})}{(1-\overline{\alpha}) g(X_t|X_{t-1}; \overline{\zeta}_1, \overline{\phi}, \overline{\sigma}) + \overline{\alpha} g(X_t|X_{t-1}; \overline{\zeta}_2, \overline{\phi}, \overline{\sigma})}.$$

Let $\overline{\alpha}^* = \arg\max_{\alpha \in [0,1]} H_n(\alpha)$. The following lemma shows that if $\overline{\alpha} - \alpha_j = O_P(n^{-1/4})$ holds true for any estimator $\overline{\alpha}$ then also for the estimator $\overline{\alpha}^*$ maximizing $H_n(\alpha)$.

Lemma 8. *Let $(\overline{\alpha}, \overline{\zeta}_1, \overline{\zeta}_2, \overline{\phi}, \overline{\sigma})$ be an EM-estimator. If $\overline{\alpha} - \alpha_j = O_P(n^{-1/4})$ for some $\alpha_j \in (0,1)$, then under the null model, we have*

$$\overline{\alpha}^* - \alpha_j = O_P(n^{-1/4}).$$

For the proof see Ketterer (2011).

Lemma 9. *Let $(\overline{\alpha}, \overline{\zeta}_1, \overline{\zeta}_2, \overline{\phi}, \overline{\sigma})$ be an EM-estimator of $(\alpha, \zeta_1, \zeta_2, \phi, \sigma)$. Under the null model the following holds:*

(i) If $\overline{\alpha} - 0.5 = O_P(n^{-1/4})$, then

$$\begin{aligned} &2\{pl_n(\overline{\alpha}, \overline{\zeta}_1, \overline{\zeta}_2, \overline{\phi}, \overline{\sigma}) - pl_n(0.5, 0, 0, \phi_0, 1)\} \\ \le &\frac{(\sum_{t=1}^n Y_t)^2}{\sum_{t=1}^n Y_t^2} + \frac{(\sum_{t=1}^n Z_t)^2}{\sum_{t=1}^n Z_t^2} + \frac{\{(\sum_{t=1}^n V_t)^-\}^2}{\sum_{t=1}^n V_t^2} + \frac{(\sum_{t=1}^n W_{1t})^2}{\sum_{t=1}^n W_{1t}^2} + o_P(1). \end{aligned} \tag{38}$$

where x^- denotes the negative part of a real number x.

(ii) If $\overline{\alpha} - \alpha_j = o_P(1)$ for some $\alpha_j \in (0, 0.5)$, then

$$\begin{aligned} &2\{pl_n(\overline{\alpha}, \overline{\zeta}_1, \overline{\zeta}_2, \overline{\phi}, \overline{\sigma}) - pl_n(0.5, 0, 0, \phi_0, 1)\} \\ \le &\frac{(\sum_{t=1}^n Y_t)^2}{\sum_{t=1}^n Y_t^2} + \frac{(\sum_{t=1}^n Z_t)^2}{\sum_{t=1}^n Z_t^2} + \frac{(\sum_{t=1}^n U_t)^2}{\sum_{t=1}^n U_t^2} + \frac{(\sum_{t=1}^n W_{1t})^2}{\sum_{t=1}^n W_{1t}^2} \\ &+ 2\{p(\alpha_j) - p(0.5)\} + o_P(1). \end{aligned} \tag{39}$$

Proof. (i) One needs to show that $\overline{s}_3 = o_P(n^{-1/2})$, and further that $\overline{s}_4$ is non-positive in probability, which is established by showing that

$$\overline{s}_4 = -2\overline{\zeta}_2^4 + o_P(n^{-1/2}) \tag{40}$$

where $\overline{\zeta}_2^4 = O_P(n^{-1/2})$ by (33). Hence, we can strengthen the upper bound in (35) to (38). See Ketterer (2011) for the details.
(ii) One needs to show that $\overline{s}_4 = o_P(n^{-1/2})$ so that the upper bound in (35) is strengthened to (39). □

Proof of Theorem 4. We know that

$$
\begin{aligned}
&2\{pl_n(0.5, \widehat{\zeta}_0, \widehat{\zeta}_0, \widehat{\phi}_0, \widehat{\sigma}_0) - pl_n(0.5, 0, 0, \phi_0, 1)\} \\
&= \frac{(\sum_{t=1}^n Y_t)^2}{\sum_{t=1}^n Y_t^2} + \frac{(\sum_{t=1}^n Z_t)^2}{\sum_{t=1}^n Z_t^2} + \frac{(\sum_{t=1}^n W_{1t})^2}{\sum_{t=1}^n W_{1t}^2} + o_P(1).
\end{aligned} \tag{41}
$$

Using the results of Lemma 9 we get

$$
M_n^{(K)}(0.5) \leq \frac{\{(\sum_{t=1}^n V_t)^-\}^2}{\sum_{t=1}^n V_t^2} + o_P(1)
$$

and

$$
M_n^{(K)}(\alpha_j) \leq \frac{(\sum_{t=1}^n U_t)^2}{\sum_{t=1}^n U_t^2} + 2\{p(\alpha_j) - p(0.5)\} + o_P(1)
$$

for $\alpha_j \neq 0.5$. Note that this inequality still holds true if we replace $2\{p(\alpha_j) - p(0.5)\}$ by $\Delta = 2\max_{\alpha_j \neq 0.5}\{p(\alpha_j) - p(0.5)\}$ as defined in Theorem 4. Therefore,

$$
EM_n^{(K)} \leq \max\left[\frac{(\sum_{t=1}^n U_t)^2}{\sum_{t=1}^n U_t^2} + \Delta, \frac{\{(\sum_{t=1}^n V_t)^-\}^2}{\sum_{t=1}^n V_t^2}\right] + o_P(1).
$$

It may be shown (see Ketterer 2011) that the upper bound is indeed obtained. To conclude, by the multivariate central limit theorem $(1/\sqrt{n})\sum_{t=1}^n (U_t, V_t)^T$ is bivariate normal. Since U_t and V_t are uncorrelated, see end of this section, $(1/\sqrt{n})\sum_{t=1}^n U_t$ and $(1/\sqrt{n})\sum_{t=1}^n V_t$ are asymptotically independent. Therefore, the limiting distribution is given by $F(x - \Delta)\big(1_{x>0} + F(x)\big)/2$, where F is the cdf of a χ_1^2 variate.

$\square$

B Supplementary material: Technical details, not intended for publication

B.1 Proof of (21) in the proof of Theorem 2

For $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d)^T \in \boldsymbol{H}$ and $\vartheta \in \Theta$, $\vartheta \neq \vartheta_0$ we set

$$\tilde{Y}_i(\vartheta, \boldsymbol{\eta}) = \frac{g(X_i|X_{i-1}^p; \vartheta, \boldsymbol{\eta}) - g(X_i|X_{i-1}^p; \vartheta_0, \boldsymbol{\eta})}{\big(\vartheta - \vartheta_0\big) g(X_i|X_{i-1}^p; \vartheta_0, \boldsymbol{\eta}_0)}, \quad \tilde{Y}_i(\vartheta) = \tilde{Y}_i(\vartheta, \boldsymbol{\eta}_0),$$

$$\tilde{Z}_i(\vartheta) = \frac{\tilde{Y}_i(\vartheta, \boldsymbol{\eta}_0) - \tilde{Y}_i(\vartheta_0, \boldsymbol{\eta}_0)}{\vartheta - \vartheta_0},$$

where $\tilde{Y}_i(\vartheta_0, \boldsymbol{\eta}) = Y_i(\vartheta_0, \boldsymbol{\eta})$ and $\tilde{Z}_i(\vartheta_0) = Z_i(\vartheta_0)$, and for $j = 1, \ldots, d$, and

$$\tilde{U}_{i,j}(\boldsymbol{\eta}) = \frac{g(X_i|X_{i-1}^p; \vartheta_0, \boldsymbol{\eta}) - g(X_i|X_{i-1}^p; \vartheta_0, \eta_1, \ldots, \eta_{j-1}, \eta_{j,0}, \eta_{j+1}, \ldots, \eta_d)}{\big(\eta_j - \eta_{j,0}\big) g(X_i|X_{i-1}^p; \vartheta_0, \boldsymbol{\eta}_0)} \quad \text{if } \eta_j \neq \eta_{j,0},$$

while $\tilde{U}_{i,j}(\boldsymbol{\eta}) = U_{i,j}(\boldsymbol{\eta})$ otherwise.

Lemma 10. *The processes*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\tilde{U}_{i,j}(\boldsymbol{\eta}) - \tilde{U}_{i,j}(\eta_1, \ldots, \eta_{k-1}, \eta_{k,0}, \eta_{k+1}, \ldots, \eta_d)}{\eta_k - \eta_{k,0}}, \quad 1 \leq j, k \leq d,$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\tilde{Y}_i(\vartheta, \boldsymbol{\eta}) - \tilde{Y}_i(\vartheta, \eta_1, \ldots, \eta_{k-1}, \eta_{k,0}, \eta_{k+1}, \ldots, \eta_d)}{\eta_k - \eta_{k,0}}, \quad 1 \leq k \leq d,$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\tilde{Z}_i(\vartheta) - \tilde{Z}_i}{\vartheta - \vartheta_0}$$

are tight.

Proof. Consider

$$U^*_{n,j,k}(\boldsymbol{\eta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{U_{i,j}(\boldsymbol{\eta}) - U_{i,j}(\eta_1, \ldots, \eta_{k-1}, \eta_{k,0}, \eta_{k+1}, \ldots, \eta_d)}{\eta_k - \eta_{k,0}}.$$

Using Billingsley (1968, p.95), see also Klicnarova (2007, prop. 1), for an appropriate multivariate extension, it suffices to show that

$$E\big(U^*_{n,j,k}(\boldsymbol{\eta_1}) - U^*_{n,j,k}(\boldsymbol{\eta_2})\big)^2 \leq C \|\boldsymbol{\eta_1} - \boldsymbol{\eta_2}\|^2$$

To this end, by the mean-value theorem, it suffices that the derivatives of summands

$$\frac{\tilde{U}_{i,j}(\boldsymbol{\eta}) - \tilde{U}_{i,j}(\eta_1, \ldots, \eta_{k-1}, \eta_{k,0}, \eta_{k+1}, \ldots, \eta_d)}{\eta_k - \eta_{k,0}}$$

are uniformly bounded in $\boldsymbol{\eta}$ by a square-integrable random variable depending only on X_i^{p+1}. This follows from the mean-value theorem and Assumption 4, i.e. by the assumption on the $W_{i,j,k}(\boldsymbol{\eta})$. The argument for the other processes is the same. □

Proof of (21). Write

$$r_{1n}(\alpha, \vartheta_1, \vartheta_2, \boldsymbol{\eta}) = 2 \sum_{i=1}^n \log(1 + \delta_i) + 2p(\alpha) - 2p(1/2) \tag{42}$$

with
$$\delta_i = (1-\alpha)\left\{\frac{g(X_i|X_{i-1}^p;\vartheta_1,\boldsymbol{\eta})}{g(X_i|X_{i-1}^p;\vartheta_0,\boldsymbol{\eta}_0)} - 1\right\} + \alpha\left\{\frac{g(X_i|X_{i-1}^p;\vartheta_2,\boldsymbol{\eta})}{g(X_i|X_{i-1}^p;\vartheta_0,\boldsymbol{\eta}_0)} - 1\right\}. \tag{43}$$
Since $2\log(1+x) \le 2x - x^2 + (2/3)x^3$, using the properties of the penalty function we have

$$r_{1n}(\alpha,\vartheta_1,\vartheta_2,\boldsymbol{\eta}) \le 2\sum_{i=1}^n \delta_i - \sum_{i=1}^n \delta_i^2 + \frac{2}{3}\sum_{i=1}^n \delta_i^3. \tag{44}$$

Expand δ_i as follows:

$$\begin{aligned}
\delta_i \quad = \quad & (1-\alpha)(\vartheta_1-\vartheta_0)\frac{g(X_i|X_{i-1}^p;\vartheta_1,\boldsymbol{\eta}) - g(X_i|X_{i-1}^p;\vartheta_0,\boldsymbol{\eta})}{(\vartheta_1-\vartheta_0)g(X_i|X_{i-1}^p;\vartheta_0,\boldsymbol{\eta}_0)} \\
& +\alpha(\vartheta_2-\vartheta_0)\frac{g(X_i|X_{i-1}^p;\vartheta_2,\boldsymbol{\eta}) - g(X_i|X_{i-1}^p;\vartheta_0,\boldsymbol{\eta})}{(\vartheta_2-\vartheta_0)g(X_i|X_{i-1}^p;\vartheta_0,\boldsymbol{\eta}_0)} \\
& +(\eta_1-\eta_{1,0})\frac{g(X_i|X_{i-1}^p;\vartheta_0,\boldsymbol{\eta}) - g(X_i|X_{i-1}^p;\vartheta_0,\eta_{1,0},\eta_2,\ldots,\eta_d)}{(\eta_1-\eta_{1,0})g(X_i|X_{i-1}^p;\vartheta_0,\boldsymbol{\eta}_0)} \\
& +(\eta_2-\eta_{2,0})\frac{g(X_i|X_{i-1}^p;\vartheta_0,\eta_{1,0},\eta_2,\ldots,\eta_d) - g(X_i|X_{i-1}^p;\vartheta_0,\eta_{1,0},\eta_{2,0},\eta_3,\ldots,\eta_d)}{(\eta_2-\eta_{2,0})g(X_i|X_{i-1}^p;\vartheta_0,\boldsymbol{\eta}_0)} \\
& \vdots \\
& +(\eta_d-\eta_{d,0})\frac{g(X_i|X_{i-1}^p;\vartheta_0,\eta_{1,0},\ldots,\eta_{d-1,0},\eta_d) - g(X_i|X_{i-1}^p;\vartheta_0,\boldsymbol{\eta}_0)}{(\eta_d-\eta_{d,0})g(X_i|X_{i-1}^p;\vartheta_0,\boldsymbol{\eta}_0)} \\
= \quad & (1-\alpha)(\vartheta_1-\vartheta_0)\tilde{Y}_i(\vartheta_1,\boldsymbol{\eta}) + \alpha(\vartheta_2-\vartheta_0)\tilde{Y}_i(\vartheta_2,\boldsymbol{\eta}) + \\
& +(\eta_1-\eta_{1,0})\tilde{U}_{i,1}(\boldsymbol{\eta}) + \ldots + (\eta_d-\eta_{d,0})\tilde{U}_{i,d}(\eta_{1,0},\ldots,\eta_{d-1,0},\eta_d),
\end{aligned} \tag{45}$$

where $\tilde{U}_i^{\eta_j}(\cdot)$ is defined in (11). Now, for $j=1,2$,

$$\begin{aligned}
\tilde{Y}_i(\vartheta_j,\boldsymbol{\eta}) \quad = \quad & \tilde{Y}_i(\vartheta_j,\boldsymbol{\eta}) - \tilde{Y}_i(\vartheta_j,\eta_1,\ldots,\eta_{d-1},\eta_{d,0}) \\
& +\tilde{Y}_i(\vartheta_j,\eta_1,\ldots,\eta_{d-1},\eta_{d,0}) - \tilde{Y}_i(\vartheta_j,\eta_1,\ldots,\eta_{d-2},\eta_{d-1,0},\eta_{d,0}) \\
& \vdots \\
& +\tilde{Y}_i(\vartheta_j,\eta_1,\eta_{2,0},\ldots,\eta_{d-1,0},\eta_{d,0}) - \tilde{Y}_i(\vartheta_j,\boldsymbol{\eta}_0) \\
& +(\vartheta_j-\vartheta_0)(\tilde{Z}_i(\vartheta_j) - Z_i) \\
& +(\vartheta_j-\vartheta_0)Z_i + Y_i
\end{aligned} \tag{46}$$

and

$$\begin{aligned}
\tilde{U}_{i,d}(\eta_{1,0},\ldots,\eta_{d-1,0},\eta_d) &= \tilde{U}_{i,d}(\eta_{1,0},\ldots,\eta_{d-1,0},\eta_d) - U_{i,d} + U_{i,d} \\
\tilde{U}_{i,d-1}(\eta_{1,0},\ldots,\eta_{d-2,0},\eta_{d-1},\eta_d) &= \tilde{U}_{i,d-1}(\eta_{1,0},\ldots,\eta_{d-2,0},\eta_{d-1},\eta_d) \\
&\quad -\tilde{U}_{i,d-1}(\eta_{1,0},\ldots,\eta_{d-1,0},\eta_d) \\
&\quad +\tilde{U}_{i,d-1}(\eta_{1,0},\ldots,\eta_{d-1,0},\eta_d) - U_{i,d-1} \\
&\quad +U_{i,d-1} \\
&\vdots \\
\tilde{U}_{i,1}(\boldsymbol{\eta}) &= \tilde{U}_{i,1}(\boldsymbol{\eta}) - \tilde{U}_{i,1}(\eta_1,\ldots,\eta_{d-1},\eta_{d,0}) \\
&\quad +\tilde{U}_{i,1}(\eta_1,\ldots,\eta_{d-1},\eta_{d,0}) \\
&\quad -\tilde{U}_{i,1}(\eta_1,\ldots,\eta_{d-2},\eta_{d-1,0},\eta_{d,0}) \\
&\quad \vdots \\
&\quad +\tilde{U}_{i,1}(\eta_1,\eta_{2,0},\ldots,\eta_{d-1,0},\eta_{d,0}) - U_{i,1} \\
&\quad +U_{i,1}.
\end{aligned} \tag{47}$$

Plugging (46) and (47) into (45), we can write

$$\delta_i = (\eta_1 - \eta_{1,0})U_{i,1} + \ldots + (\eta_d - \eta_{d,0})U_{i,d} + m_1 Y_i + m_2 Z_i + \epsilon_{in}, \tag{48}$$

where

$$m_1 = (1-\alpha)(\vartheta_1 - \vartheta_0) + \alpha(\vartheta_2 - \vartheta_0), \qquad m_2 = (1-\alpha)(\vartheta_1 - \vartheta_0)^2 + \alpha(\vartheta_2 - \vartheta_0)^2$$

and ϵ_{in} is a remainder term. Note at this stage that each of the sequences the variables $(U_{i,j})_{i\geq 1}$, $j = 1, \ldots d$, $(Y_i)_{i\geq 1}$ and $(Z_i)_{i\geq 1}$ form square intergable (Assumption 4) stationary martingale difference sequences w.r.t. the filtration generated by the observations (X_i).

Now plug (48) into (44). To obtain (21), it remains to estimate the remainder terms as well as the cubic term.
Let us shows that the remainder terms in (48) are negligible as compared to the quadratic term in (44). Let $\epsilon_n = \sum_{i=1}^n \epsilon_{in}$. By Lemma 10,

$$\begin{aligned}
\epsilon_n &= \sqrt{n}(\eta_d - \eta_{d,0})^2 O_P(1) \\
&\quad +\sqrt{n}(\eta_{d-1} - \eta_{d-1,0})\Big(\sum_{j=d-1}^{d}(\eta_j - \eta_{j,0})\Big)O_P(1) \\
&\quad \vdots \\
&\quad +\sqrt{n}(\eta_1 - \eta_{1,0})\Big(\sum_{j=1}^{d}(\eta_j - \eta_{j,0})\Big)O_P(1) \\
&\quad +\sqrt{n}(1-\alpha)(\vartheta_1 - \vartheta_0)\Big(\sum_{j=1}^{d}(\eta_j - \eta_{j,0})\Big)O_P(1) \\
&\quad +\sqrt{n}\alpha(\vartheta_2 - \vartheta_0)\Big(\sum_{j=1}^{d}(\eta_j - \eta_{j,0})\Big)O_P(1) \\
&\quad +\sqrt{n}(1-\alpha)(\vartheta_1 - \vartheta_0)^3 O_P(1) + \sqrt{n}\alpha(\vartheta_2 - \vartheta_0)^3 O_P(1).
\end{aligned}$$

We may restrict our attention to a small neighborhood of $(\eta_{1,0},\ldots,\eta_{d,0},\vartheta_0)$ as suggested by the consistency results in Theorem 1(ii). Therefore we may regard $\eta_1 - \eta_{1,0},\ldots,\eta_d - \eta_{d,0}, \vartheta_1 - \vartheta_0, \vartheta_2 - \vartheta_0$

as $o_P(1)$ and we get

$$\begin{aligned}\epsilon_n &= \sqrt{n}(\eta_d-\eta_{d,0})o_P(1)+\sqrt{n}(\eta_{d-1}-\eta_{d-1,0})o_P(1)+\ldots+\sqrt{n}(\eta_1-\eta_{1,0})o_P(1)\\ &\quad+\sqrt{n}(1-\alpha)(\vartheta_1-\vartheta_0)o_P(1)+\sqrt{n}\alpha(\vartheta_2-\vartheta_0)o_P(1)\\ &\quad+\sqrt{n}(1-\alpha)(\vartheta_1-\vartheta_0)^2o_P(1)+\sqrt{n}\alpha(\vartheta_2-\vartheta_0)^2o_P(1).\end{aligned}$$

Since $|x|\le 1+x^2$, we obtain

$$|\epsilon_n|\le n\{(\eta_1-\eta_{1,0})^2+\ldots+(\eta_d-\eta_{d,0})^2+m_1^2+m_2^2\}o_p(1)+o_P(1).$$

On the other hand, by Assumption 5 there is a $\lambda>0$ such that for all $(\alpha_1,\ldots,\alpha_{d+2})\in\mathbb{R}^{d+2}\setminus\{\mathbf{0}\}$ we have that

$$E\{\alpha_1U_{1,1}+\ldots+\alpha_dU_{1,d}+\alpha_{d+1}Y_1+\alpha_{d+2}Z_1\}^2\ge\lambda(\alpha_1^2+\ldots+\alpha_{d+2}^2). \tag{49}$$

Therefore, the remainder term in (48) is negligible as compared to the quadratic term in (44). For the cubic term, by the ergodic theorem, Assumption 4 and (49) imply

$$\begin{aligned}&\frac{\sum_{i=1}^n|(\eta_1-\eta_{1,0})U_{i,1}+\ldots(\eta_d-\eta_{d,0})U_{i,d}+m_1Y_i+m_2Z_i|^3}{\sum_{i=1}^n((\eta_1-\eta_{1,0})U_{i,1}+\ldots(\eta_d-\eta_{d,0})U_{i,d}+m_1Y_i+m_2Z_i)^2}\\ =\;&\frac{E|\eta_1-\eta_{1,0})U_{1,1}+\ldots(\eta_d-\eta_{d,0})U_{1,d}+m_1Y_1+m_2Z_1|^3}{E((\eta_1-\eta_{1,0})U_{1,1}+\ldots(\eta_d-\eta_{d,0})U_{1,d}+m_1Y_1+m_2Z_1)^2}O_P(1),\\ \le\;&\frac{|\eta_1-\eta_{1,0}|^3+\ldots+|\eta_d-\eta_{d,0}|^3+|m_1|^3+|m_2|^3}{(\eta_1-\eta_{1,0})^2+\ldots+(\eta_d-\eta_{d,0})^2+m_1^2+m_2^2}O_P(1)\\ \le\;&\{|\eta_1-\eta_{1,0}|+\ldots+|\eta_d-\eta_{d,0}|+|m_1|+|m_2|\}O_P(1)=o_P(1)\end{aligned}$$

thus, it is also negligible as compared to the quadratic term. This concludes the proof of (21).

□

B.2 Proof of EM property in the proof of Theorem 3

For the argument, given the sample $X_1=x_1,\ldots,X_n=x_n$, we work with a (hypothetic) independent regime $(S_k)_{k\ge0}$. The parameter vector is then given by $\boldsymbol{\psi}^T=(\alpha,\vartheta_1,\vartheta_2,\boldsymbol{\phi}^T)\in\mathbb{R}^{d+3}$, where α is the probability for state 2 for the independent regime. Denote

(i) $\mathbf{S}=(S_1,\ldots,S_n)$, $\mathbf{X}=(X_1,\ldots,X_n)$, $\mathbf{x}=(x_1,\ldots,x_n)$ and $\mathbf{s}=(s_1,\ldots,s_n)$,

(ii) q be the joint pdf of $(\mathbf{X},\mathbf{S})$ given $X_0^p,\boldsymbol{\psi}$ (under this artificial model),

(iii) r be the pdf of $\mathbf{S}$ given $\mathbf{X},X_0^p,\boldsymbol{\psi}$ (also under this artificial model).

so that

$$p(\mathbf{x}|x_0^p,\boldsymbol{\psi})r(\mathbf{s}|\mathbf{x},x_0^p,\boldsymbol{\psi})=q(\mathbf{x},\mathbf{s}|x_0^p,\boldsymbol{\psi}). \tag{50}$$

Explicitely,

$$\begin{aligned}p(\mathbf{x}|x_0^p,\boldsymbol{\psi}) &= \prod_{k=1}^n\{(1-\alpha)g(x_k|x_{k-1}^p;\vartheta_1,\boldsymbol{\eta})+\alpha g(x_k|x_{k-1}^p;\vartheta_2,\boldsymbol{\eta})\}\\ r(\mathbf{s}|\mathbf{x},x_0^p,\boldsymbol{\psi}) &= \prod_{k=1}^n\frac{(1-\alpha)^{\mathbb{1}_{\{s_k=1\}}}\alpha^{\mathbb{1}_{\{s_k=2\}}}g(x_k|x_{k-1}^p;\vartheta_{s_k},\boldsymbol{\eta})}{(1-\alpha)g(x_k|x_{k-1}^p;\vartheta_1,\boldsymbol{\eta})+\alpha g(x_k|x_{k-1}^p;\vartheta_2,\boldsymbol{\eta})}\\ q(\mathbf{x},\mathbf{s}|x_0^p,\boldsymbol{\psi}) &= \prod_{k=1}^n(1-\alpha)^{\mathbb{1}_{\{s_k=1\}}}\alpha^{\mathbb{1}_{\{s_k=2\}}}g(x_k|x_{k-1}^p;\vartheta_{s_k},\boldsymbol{\eta})\end{aligned}$$

Denote by $E_{\boldsymbol{\psi}^{(k)}}$ expectation w.r.t. the (artificial) distribution including the independent regime under the parameter $\boldsymbol{\psi}^{(k)}$. From (50), we get

$$pl_n(\boldsymbol{\psi}) = Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}) - R(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}) + p(\alpha),$$

where

$$\begin{aligned}
\bar{Q}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}) &= E_{\boldsymbol{\psi}^{(k)}}\big(\log(q(\mathbf{X},\mathbf{S}|X_0^p,\boldsymbol{\psi}))|\mathbf{X},X_0^p,\boldsymbol{\psi}^{(k)}\big) + p(\alpha),\\
R(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)}) &= E_{\boldsymbol{\psi}^{(k)}}\{\log(r(\mathbf{S}|\mathbf{X},X_0^p,\boldsymbol{\psi}))|\mathbf{X},X_0^p,\boldsymbol{\psi}^{(k)}\}
\end{aligned}$$

and $\boldsymbol{\psi}^{(k)}$ is the current value of $\boldsymbol{\psi}$. Then

$$\bar{Q}(\boldsymbol{\psi}^{(k+1)}|\boldsymbol{\psi}^{(k)}) \geq \bar{Q}(\boldsymbol{\psi}^{(k)}|\boldsymbol{\psi}^{(k)}) \Longrightarrow pl_n(\boldsymbol{\psi}^{(k+1)}) \geq pl_n(\boldsymbol{\psi}^{(k)}). \tag{51}$$

Proof of (51). Using Jensen's inequality we get:

$$\begin{aligned}
R(\boldsymbol{\psi}^{(k+1)}|\boldsymbol{\psi}^{(k)}) - R(\boldsymbol{\psi}^{(k)}|\boldsymbol{\psi}^{(k)}) &= E_{\boldsymbol{\psi}^{(k)}}\left\{\log\frac{r(\mathbf{S}|\mathbf{X},X_0^p,\boldsymbol{\psi}^{(k+1)})}{r(\mathbf{S}|\mathbf{X},X_0^p,\boldsymbol{\psi}^{(k)})}\middle|\mathbf{X},X_0^d,\boldsymbol{\psi}^{(k)}\right\}\\
&\leq \log E_{\boldsymbol{\psi}^{(k)}}\left\{\frac{r(\mathbf{S}|\mathbf{X},X_0^p,\boldsymbol{\psi}^{(k+1)})}{r(\mathbf{S}|\mathbf{X},X_0^p,\boldsymbol{\psi}^{(k)})}\middle|\mathbf{X},X_0^p,\boldsymbol{\psi}^{(k)}\right\}\\
&= 0,
\end{aligned}$$

and therefore

$$\begin{aligned}
pl_n(\boldsymbol{\psi}^{(k)}) &= \bar{Q}(\boldsymbol{\psi}^{(k)}|\boldsymbol{\psi}^{(k)}) - R(\boldsymbol{\psi}^{(k)}|\boldsymbol{\psi}^{(k)})\\
&\leq \bar{Q}(\boldsymbol{\psi}^{(k+1)}|\boldsymbol{\psi}^{(k)}) - R(\boldsymbol{\psi}^{(k)}|\boldsymbol{\psi}^{(k)})\\
&\leq \bar{Q}(\boldsymbol{\psi}^{(k+1)}|\boldsymbol{\psi}^{(k)}) - R(\boldsymbol{\psi}^{(k+1)}|\boldsymbol{\psi}^{(k)})\\
&= pl_n(\boldsymbol{\psi}^{(k+1)}).
\end{aligned}$$

□

Next we show that $\bar{Q}(\boldsymbol{\psi}^{(k+1)}|\boldsymbol{\psi}^{(k)}) \geq \bar{Q}(\boldsymbol{\psi}^{(k)}|\boldsymbol{\psi}^{(k)})$ holds for the updates obtained by the ECM algorithm (as proposed in Meng and Rubin, 1993). Relabel $\boldsymbol{\psi} = (\psi_1,\dots,\psi_{d+3})$ and $1 \leq r \leq d+3$ let

$$\pi_{\{t_1,\dots,t_r\}} : \mathbb{R}^{d+3} \to \mathbb{R}^r, \quad \pi_{\{t_1,\dots,t_r\}}(\psi_1,\dots,\psi_{d+3}) = (\psi_{t_1},\dots,\psi_{t_r}),$$

$P_1,\dots,P_q$ any partition of $\{1,\dots,d+3\}$ and $-P_j = \{1,\dots,d+3\}\setminus P_j$.
The ECM algorithm proceeds as follows.
Step 1: Compute $\boldsymbol{\psi}^{(k+1/q)} = \arg\max_{\boldsymbol{\psi}} \bar{Q}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})$ subject to $\pi_{-P_1}(\boldsymbol{\psi}) = \pi_{-P_1}(\boldsymbol{\psi}^{(k)})$.
Step 2: Compute $\boldsymbol{\psi}^{(k+2/q)} = \arg\max_{\boldsymbol{\psi}} \bar{Q}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})$ subject to $\pi_{-P_2}(\boldsymbol{\psi}) = \pi_{-P_2}(\boldsymbol{\psi}^{(k+1/q)})$.
$\vdots$
Step q: Compute $\boldsymbol{\psi}^{(k+q/q)} = \arg\max_{\boldsymbol{\psi}} \bar{Q}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})$ subject to $\pi_{-P_q}(\boldsymbol{\psi}) = \pi_{-P_q}(\boldsymbol{\psi}^{(k+(q-1)/q)})$.
The updated value is given by $\boldsymbol{\psi}^{(k+1)} = \boldsymbol{\psi}^{(k+q/q)}$. Then, by construction, we have

$$\bar{Q}(\boldsymbol{\psi}^{(k+1)}|\boldsymbol{\psi}^{(k)}) \geq \bar{Q}(\boldsymbol{\psi}^{(k+(q-1)/(q))}|\boldsymbol{\psi}^{(k)}) \geq \ldots \geq \bar{Q}(\boldsymbol{\psi}^{(k+1/q)}|\boldsymbol{\psi}^{(k)}) \geq \bar{Q}(\boldsymbol{\psi}^{(k)}|\boldsymbol{\psi}^{(k)}).$$

which implies (51).
Since

$$\begin{aligned}
&\bar{Q}(\boldsymbol{\psi}|\boldsymbol{\psi}^{(k)})\\
= &\sum_{i=1}^{n}\{\log((1-\alpha)g(X_i|X_{i-1}^p;\vartheta_1,\boldsymbol{\eta}))(1-w_i^{(k)}) + \log(\alpha g(X_i|X_{i-1}^p;\vartheta_2,\boldsymbol{\eta}))w_i^{(k)}\} + p(\alpha)
\end{aligned}$$

with

$$w_i^{(k)} = \frac{\alpha^{(k)} g(X_i|X_{i-1}^p; \vartheta_1^{(k)}, \boldsymbol{\eta}^{(k)})}{(1-\alpha^{(k)}) g(X_i|X_{i-1}^p; \vartheta_1^{(k)}, \boldsymbol{\eta}^{(k)}) + \alpha^{(k)} g(X_i|X_{i-1}^p; \vartheta_2^{(k)}, \boldsymbol{\eta}^{(k)})},$$

the algorithm in the EM-test is actually the ECM algorithm with $P_1 = \{\alpha, \vartheta_1, \vartheta_2\}$ and $P_2 = \{\boldsymbol{\eta}\}$.

B.3 Details concerning the assumptions

Lemma 11. *For a fixed ν, let $f(x) = \Gamma\left(\frac{\nu+1}{2}\right)\left(\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu}\left(1+\frac{x^2}{\nu}\right)^{(\nu+1)/2}\right)^{-1}$ be the density of the t-distribution with ν degrees of freedom. Then for the associated location-scale family $f(x;\mu,\sigma)$, for any (μ,σ) and $a_1, a_2, a_3, b_1, b_2, b_3 \in \mathbb{R}$,*

a.)

$$a_1 \frac{\partial f(x;\mu,\sigma)}{\partial\mu} + a_2 \frac{\partial^2 f(x;\mu,\sigma)}{\partial^2\mu} + a_3 \frac{\partial f(x;\mu,\sigma)}{\partial\sigma} = 0 \quad \textit{for Leb.-a.e. } x$$

implies that $a_1 = a_2 = a_3 = 0$.

b.)

$$b_1 \frac{\partial f(x;\mu,\sigma)}{\partial\mu} + b_2 \frac{\partial f(x;\mu,\sigma)}{\partial\sigma} + b_3 \frac{\partial^2 f(x;\mu,\sigma)}{\partial^2\sigma} = 0 \quad \textit{for Leb.-a.e. } x$$

implies that $b_1 = b_2 = b_3 = 0$.

Proof of Lemma 11. The characteristic function of the t-distribution is given by (cf. Hurst 1995)

$$\varphi(t) = \frac{K_{\frac{1}{2}\nu}\left(\sqrt{\nu}|t|\right)\left(\sqrt{\nu}|t|\right)^{\frac{1}{2}\nu}}{\Gamma\left(\frac{1}{2}\nu\right) 2^{\frac{1}{2}\nu-1}},$$

where $\Gamma(\cdot)$ is the Gamma function and $K_p(\cdot)$ is the modified Bessel function of the second kind and order p (cf. Andrews 1986, chapter 6). Therefore, the characteristic function of the corresponding location-scale family is

$$\varphi(t;\mu,\sigma) = e^{i\mu t}\ \varphi(\sigma t) = e^{i\mu t} \frac{K_m\left(\sqrt{\nu}\sigma|t|\right)\left(\sqrt{\nu}\sigma|t|\right)^m}{\Gamma(m)\, 2^{m-1}}, \tag{52}$$

where we put $m = \frac{1}{2}\nu$. The partial derivatives are given by

$$\begin{aligned}
\frac{\partial\varphi(t;\mu,\sigma)}{\partial\mu} &= it\ e^{i\mu t}\ \frac{K_m\left(\sqrt{\nu}\sigma|t|\right)\left(\sqrt{\nu}\sigma|t|\right)^m}{\Gamma(m)\, 2^{m-1}}, \\
\frac{\partial^2\varphi(t;\mu,\sigma)}{\partial^2\mu} &= -t^2\ e^{i\mu t}\ \frac{K_m\left(\sqrt{\nu}\sigma|t|\right)\left(\sqrt{\nu}\sigma|t|\right)^m}{\Gamma(m)\, 2^{m-1}}
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial\varphi(t;\mu,\sigma)}{\partial\sigma} &= -|t|\ e^{i\mu t} \frac{K_{m-1}(\sqrt{\nu}\sigma|t|)\sqrt{\nu}\left(\sqrt{\nu}\sigma|t|\right)^m}{\Gamma(m) 2^{m-1}}, \\
\frac{\partial^2\varphi(t;\mu,\sigma)}{\partial^2\sigma} &= \frac{|t|\sqrt{\nu}e^{i\mu t}}{\Gamma(m)2^{m-1}}\ \left(\sqrt{\nu}|t|\right)^m\ \sigma^{m-1}\left(\sqrt{\nu}\sigma|t| K_{m-2}(\sqrt{\nu}\sigma|t|) - K_{m-1}(\sqrt{\nu}\sigma|t|)\right),
\end{aligned}$$

cf. Andrews (1986).

(a). Taking the Fourier transform and interchanging integral and derivative gives

$$a_1 \frac{\partial \varphi(t;\mu,\sigma)}{\partial \mu} + a_2 \frac{\partial^2 \varphi(t;\mu,\sigma)}{\partial^2 \mu} + a_3 \frac{\partial \varphi(t;\mu,\sigma)}{\partial \sigma} = 0 \quad \text{for all } t \in \mathbb{R}. \tag{53}$$

Plugging the partial derivatives into (53) and dividing by $te^{i\mu t}(\sqrt{\nu}|t|)^m \, \sigma^{m-1}/\big(\Gamma(m)2^{m-1}\big)$ gives with $x = \sqrt{\nu}\sigma|t|$

$$a_1 \, i\sigma \; K_m(x) - a_2\sigma t \; K_m(x) - a_3\sqrt{\nu}\sigma \, \mathrm{sign}(t) K_{m-1}(x) = 0, \quad t \in \mathbb{R}. \tag{54}$$

Choosing $t = 1$ and $t = -1$ and adding, we get $a_1 = 0$. Next, dividing by $t\, K_m(x)$ and letting $t \to \infty$ (hence $x \to \infty$), since $K_{m-1}(x)/K_m(x) \to 1$ (Andrews 1986), we get $a_2 = 0$, and finally $a_3 = 0$.
(b). Taking the Fourier transform and interchanging integral and derivative gives

$$b_1 \frac{\partial \varphi(t;\mu,\sigma)}{\partial \mu} + b_2 \frac{\partial \varphi(t;\mu,\sigma)}{\partial \sigma} + b_3 \frac{\partial^2 \varphi(t;\mu,\sigma)}{\partial^2 \sigma} = 0 \quad \text{for all } t \in \mathbb{R} \tag{55}$$

Plugging the partial derivatives into (55) and dividing by $te^{i\mu t}(\sqrt{\nu}|t|)^m \, \sigma^{m-1}/\big(\Gamma(m)2^{m-1}\big)$ gives with $x = \sqrt{\nu}\sigma|t|$

$$b_1 \, i\sigma \; K_m(x) - b_2\sqrt{\nu}\sigma \, \mathrm{sign}(t) K_{m-1}(x) + b_3\sqrt{\nu} \, \mathrm{sign}(t)\big(xK_{m-2}(x) - K_{m-1}(x)\big) = 0, \quad t \in \mathbb{R}. \tag{56}$$

Choosing $t = 1$ and $t = -1$ and adding, we get $b_1 = 0$. Therefore equation (56) reduces to

$$b_2\sigma K_{m-1}(x) - b_3\big(xK_{m-2}(x) - K_{m-1}(x)\big) = 0, \quad t \in \mathbb{R}. \tag{57}$$

Dividing by $x\, K_{m-2}(x)$ and letting $x \to \infty$, since $K_{m-1}(x)/K_{m-2}(x) \to 1$ (Andrews 1986), we get $b_3 = 0$, and therefore $b_2 = 0$.

□