Asymptotic Properties of the Maximum Likelihood Estimator of the Mixture Autoregressive Model with Applications to Financial Risk

Mary I.Akinyemi* Georgi N.Boshnakov†

Abstract
Mixture autoregressive models provide a flexible framework for modelling time series. These models capture conditional heterogeneity, multi-modality, skewness, kurtosis and heavy tails using only standard distributions as building blocks. We show that the maximum likelihood estimator (MLE) of this class of models is consistent and asymptotically normal. We also give applications to estimation of financial risk.

Key Words: Mixture autoregressive models ; maximum likelihood estimator (MLE) ; financial risk ; Asymptotic Properties

1. Introduction

Mixture autoregressive models provide a flexible framework for modelling time series. They capture conditional heterogeneity, multi-modality, skewness, kurtosis and heavy tails using only standard distributions as building blocks. A maximum likelihood estimate associated with a sample of observations is a choice of parameters that maximizes the probability density function of the sample, called in this context the likelihood function. MLE is of fundamental importance in the theory of inference and it forms the basis of many inferential techniques in statistics (Myung, 2003). There is vast literature on MLE and it's applications as well as it's properties available among which is (Wald, 1949), (Andersen, 1970) among others. When a model has a higher value of the maximized loglikelihood than other models, the model becomes more viable for further investigation than the other models. Various studies have been done on exploring the asymptotic properties of MLE. (Nguyen and McLachlan, 2015) prove that the ML estimators of the LRC parameters are consistent and asymptotically normal, like their natural counterparts. They also show that the LRC allows for simple handling of singularities in the ML estimation of GMMs. Using numerical simulations in the R programming environment, they demonstrate that the MM algorithm can be faster than the EM algorithm in various large data situations, where sample sizes range in the tens to hundreds of thousands and for estimating models with up to 16 mixture components on multivariate data with up to 16 variables.

(Jin et al., 2015) explore properties of pseudo-maximum likelihood (PML) estimators for pooled data. They compared resulting asymptotic efficiency of the PML estimators of factor loadings with that of the multi-group maximum likelihood estimators. The effect of pooling was investigated through a two-group factor model. They found that the variances of factor loadings for the pooled data are underestimated under the normal theory when error variances in the smaller group are larger and that underestimation is due to dependence between the pooled factors and pooled error terms. Small-sample properties of the PML estimators were also investigated using a Monte Carlo study.

We examine here the asymptotic properties of the Maximum-Likelihood Estimator (MLE) of the MAR model. We leverage the results of (Douc et al., 2004) whose assumptions/proofs are hinged on the paper by (Wald, 1949) which says that there exists a deter-

*Department of Mathematics, University of Lagos, Lagos Nigeria
†School of Mathematics, University of Manchester, Manchester, United Kingdom

ministic asymptotic criterion function $l(\theta)$ such that $n^{-1}l_n(\theta, z_0) \to l(\theta)$ a.s. uniformly with respect to $\theta \in \Theta$. We show that the MLE of the MAR model is both consistent and asymptotically normal. In addition, we propose the mixture autoregressive model (MAR) model as an alternative approach to evaluating VaR and ES. We do this by considering the one-step ahead out of sample Value at Risk (VaR) and Expected Shortfall (ES) measures for the Standard and Poor (S&P500). We then compare the results based on the MAR models with both Gaussian and Student-t innovations to the results based on Risk metrics, the Gaussian GARCH, Student-t GARCH, the AR-GARCH with Gaussian innovations, the AR-GARCH with student-t innovations, the Empirical Quantile, the Traditional Extreme value theory and the Point over Threshold EVT. We find that the MAR based models perform predominantly better than the others. The rest of this paper is structured as follows: In Section 1.1, we define the Mixture Autoregressive (MAR) model in its variation and also outline its properties. We give use useful notations and assumptions in Section 1.2. We present the consistency and asymptotic normality of the MLE of the MAR models in Sections 2.1 and 2.2 respectively. Then we show the application of the model to estimating VaR and ES for the S&P500 returns series. Finally, Section 4 concludes.

1.1 The mixture autoregressive model

The mixture autoregressive model of (Wong and Li, 2000) is defined as follows.

Definition 1.1 (Mixture autoregresssive model)**.** A process $\{y_t\}$ is said to be a mixture autoregressive (MAR) process if the conditional distribution function of y_t given past information is given by

$$F_{t|t-1}(x) = \sum_{k=1}^{g} \pi_k F_k \left(\frac{x - \phi_{k,0} - \sum_{i=1}^{p_k} \phi_{k,i} y_{t-i}}{\sigma_k} \right), \qquad (1.1.1)$$

where

g is a positive integer representing the number of components in the model and the kth component of the model, for $k = 1, \dots, g$, is specified by its mixing proportion $\pi_k > 0$, scale parameter $\sigma_k > 0$, autoregressive order p_k, intercept $\phi_{k,0}$, autoregressive coefficients $\phi_{k,i}$, $i = 1, \dots, p_k$, and cumulative distribution function $F_k(\cdot)$. The mixing proportions π_k define a discrete distribution $\boldsymbol{\pi}$, so $\sum_{k=1}^{g} \pi_k = 1$.

We denote by MAR($g; p_1, \dots p_g$) a g-component MAR model whose components are of orders $p_1, \dots, p_g$. The noise distribution functions F_k, $k = 1, \dots, p$, are typically taken to be standard Gaussian (Wong and Li, 2000) or (standardised) Student-t (Wong et al., 2009). We will denote by $f_k(\cdot)$ the corresponding probability density functions. It is also convenient to set $p = \max_{1 \le k \le g} p_k$ and $\phi_{k,i} = 0$ for $i > p_k$.

We do not discuss estimation theory in this paper but it can be developed under relatively mild conditions, usually met in practice. The noise probability densities, $f_k(.)$, need to be continuous and positive everywhere, non-periodic and bounded on compacts sets for all k. Detailed study of the asymptotic theory is given by (Akinyemi, 2013).

A useful interpretation of the MAR model is that at each time t one of g autoregressive-like equations is picked at random to generate y_t. Namely, let $\{z_t\}$ be an i.i.d. sequence of discrete random variables with distribution $\boldsymbol{\pi}$ (see Definition 1.1). Then y_t can be written as

$$y_t = \phi_{z_t,0} + \sum_{i=1}^{p} \phi_{z_t,i} y_{t-i} + \sigma_{z_t} \epsilon_{z_t}(t), \qquad (1.1.2)$$

where $\{\epsilon_{z_t}\}$ are jointly independent and independent of past ys and the probability density of $\{\epsilon_{z_t}\}$ is $f_k(.)$ (Boshnakov, 2009; Boshnakov, 2011, for further details see). Let $\{z_t\}$ be an iid sequence of random variables with distribution π such that $Pr\{z_t = k\} = \pi_k, \quad k = 1, \dots, g$, define a vector $Z_t = [Z_{t,1}, \dots, Z_{t,g}]^{'}$ such that,

$$Z_{t,k} = \begin{cases} 1 & \text{if } z_t = k \\ 0 & \text{otherwise} \end{cases}$$

Then, the process y_t can be written as (Boshnakov, 2009),

$$y_t = \mu_{z_t}(y_t^{'}) + \sigma_{z_t}\epsilon_{z_t}(t) \tag{1.1.3}$$

where

$$\mu_{z_t}(y_t^{'}) = \phi_{z_t,0} + \sum_{i=1}^{p}(\phi_{z_t,i}y_{t-i}) \qquad (p = \max_{1\le k\le g} p_k). \tag{1.1.4}$$

The conditional density of y_t given both the past values of y_t and z_t is,

$$f_\theta(y_t \mid y_t^{'}, z_t) = \frac{1}{\sigma_{z_t}} f_{z_t}\left(\frac{y_t - \phi_{z_t,0} - \sum_{i=1}^{p_{z_t}} \phi_{z_t,i}y(t-i)}{\sigma_{z_t}}\right), \tag{1.1.5}$$

$\{Z_t, t > 0\}$ is a simple case of a hidden Markov chain on a finite state space $S \in [0, 1]$ with stationary k-step transition probability matrix. $\{Z_t, t > 0\}$ drives the dynamics of $Y_t = (y_t, \dots, y_{t-p+1})^{'}$. Thus, we can write a chain,

$$Q_t = (Z_t, Y_t), \tag{1.1.6}$$

where, Q_t is an aperiodic $S \times \mathbb{R}^p$-valued Markov chain.

Let A be a non -negative $g \times g$ matrix such that $A = (a_{ij})$ and $\sum_j a_{ij} = 1$.
Let θ be the vector of all the free parameters of the model. We assume that θ belongs to a compact subset of R^d denoted by Θ.

1.2 Notations and assumptions

- Given $(Y_t, t \ge 0)$, $Y_t = (y_t, \dots, y_{t-p+1})^{'}$ each y_t is an MAR process defined by Equation (1.1.3) with conditional distribution function defined in Equation (1.1.1).
- Denote by θ^0 the true value of the parameters to be estimated and $\hat{\theta}$ the maximum likelihood estimate. Let $f_\theta(\cdot \mid y, k)$ denote the conditional density of y_t given $y_{t-1}, \dots, y_{t-p}, Z_{t,k}$, defined in Equation (1.1.5). We write $\{Y_m, \dots, Y_n\} = Y^{(m,n)}$.
- By the markov property, the filtering distribution of the unknown state given past information is given by,

$$\begin{aligned}\mathbb{P}(z_t = k \mid z_s, Y_s, s = 0, \dots, t-1) &= \mathbb{P}(z_t = k \mid Y_0, Y_s, s = 0, \dots, t-1) \\ &= \mathbb{P}(z_t = k \mid z_{t-1}) \quad \text{for } k = 1, \dots, g.\end{aligned} \tag{1.2.1}$$

- The conditional likelihood function of $Y^{(1,n)}$ given both Y_0 and $Z_0 = z_0$ is given as,

$$p_\theta(Y^{(1,n)} \mid Y_0, Z_0 = z_0) = \sum_{z_n=1}^{g} \cdots \sum_{z_1=1}^{g} \prod_{t=1}^{n} a_{z_{t-1},z_t} f_\theta(Y_t \mid Y_{t-1}, z_t) \tag{1.2.2}$$

  where a_{ij} is the transition probability matrix such that $P(Y_t = i \mid Y_{t-1} = j)$.

- The corresponding conditional log-likelihood function is,

$$l_n(\theta, z_0) = \log p_\theta(Y^{(1,n)} \mid Y_0, Z_0 = z_0) = \sum_{t=1}^{n} p_\theta(Y_t \mid Y^{(0,t-1)} Y_0, Z_0 = z_0). \tag{1.2.3}$$

- Similarly, the conditional log-likelihood function given Y_0 only is,

$$l_n(\theta) = \sum_{t=1}^{n} p_\theta(Y_t \mid Y^{(0,t-1)}) \tag{1.2.4}$$

- where

$$\begin{aligned} & p_\theta(Y_t \mid Y^{(0,t-1)}, Z_0 = z_0) \\ &= \sum_{z_{t-1}=1}^{g} \sum_{z_t=1}^{g} f_\theta(Y_t \mid Y_{t-1}, z_t) a_{z_{t-1}, z_t} \mathbb{P}(Z_{t-1} = z_{t-1} \mid Y^{(0,t-1)}, Z_0 = z_0) \end{aligned} \tag{1.2.5}$$

- and

$$p_\theta(Y_t \mid Y^{(0,t-1)}) = \sum_{z_{t-1}=1}^{g} \sum_{z_t=1}^{g} f_\theta(Y_t \mid Y_{t-1}, z_t) a_{z_{t-1}, z_t} \mathbb{P}(Z_{t-1} = z_{t-1} \mid Y^{(0,t-1)}) \tag{1.2.6}$$

(Douc et al., 2004, Corollary 1) show that the total variation distance between the filtering probabilities $\mathbb{P}_\theta(Z_{t-1} = z_{t-1} \mid Y_0)$ and $\mathbb{P}_\theta(Z_{t-1} = z_{t-1} \mid Y_0, Z_0 = z_0)$ tends to zero exponentially fast as $t \to \infty$ uniformly with respect to z_0.

The following assumptions are made on the chain Q_t in Equation1.1.6.

Assumptions

1. The true parameter value which we represent by θ^0 lies in the interior of Θ.
2. For each $k \in \{1, \ldots, g\}$, $\{Z_{t,k} : t \geq 0\}$ is an irreducible, aperiodic Markov chain on a finite space S with probability distribution $\pi_1, \ldots, \pi_g$ and transition probability matrix $A = (a_{ij})$, so that $Z_{t,k}$ inherits the properties of $\{Z_t\}$.
3. The chain $\{Z_t\}$ is independent of the ϵ_t, also, for $\mathcal{F}_{t-1} = \sigma\{Y_r,\ r \leq t-1\}$ and all i, j,

$$P(z_t = j \mid z_{t-1} = i, \mathcal{F}_{t-1}) = P(z_t = j \mid z_{t-1} = i). \tag{1.2.7}$$

4. $\{\epsilon_t\}$ are jointly independent and are independent of past ys.
5. $\{\epsilon_t\}$ has a probability density function that is continuous and positive everywhere.
6. $f_{z_t}(y)$ is non periodic and bounded on all compacts sets for all k and $z_t \in S$.

Furthermore, we require that the following conditions be satisfied.

Condition A

- $\{Y_t, t \geq 0\}$ is geometrically ergodic
- For all $y, y^{'} \in \mathbb{R}^p$, y' is a vector of past values of y.

$$\inf_{\theta} f_\theta(y \mid y^{'}) > 0, \qquad \sup_{\theta} f_\theta(y \mid y^{'}) < \infty \tag{1.2.8}$$

-

$$b_+ = \sup_{\theta} \sup_{y,y^{'}} f_\theta(y \mid y^{'}, k) < \infty \tag{1.2.9}$$

and

$$E|\log \inf_{\theta} f_\theta(Y_1 \mid Y_0)| < \infty \tag{1.2.10}$$

We propose the following Lemma on y_t

Lemma 1.1. Let (y_t) be an MAR process and $Y_t = (y_t, \ldots, y_{t-p+1})^{'}$. Then Conditon1.2 holds.

Proof 1.1. The geometric ergodicity of Y_t has been established in (Akinyemi, 2013). Assume

$$f_k(\frac{y_t - \phi_{k,0} - \sum_{i=1}^{p_k} \phi_{k,i} y_{t-i})}{\sigma_k} \leq 1. \tag{1.2.11}$$

Choose a positive constant M such that for $\sigma_k^2 > 0$, let $\sigma_k^2 \geq M^2$. Then,

$$\begin{aligned} f_\theta(y \mid y^{'}, k) &\leq \sum_{k=1}^{g} \frac{\pi_k}{M} \\ &= \frac{1}{M} \sum_{k=1}^{g} \pi_k = \frac{1}{M} \qquad (\text{since } \textstyle\sum_{k=1}^{g} \pi_k = 1) \\ &\text{this implies that } f_\theta(y \mid y^{'}, k) \leq \frac{1}{M} \\ &\text{which then implies that } f_\theta(y \mid y^{'}, k) \leq \frac{1}{\sigma_k} \end{aligned}$$

so that for all $y, y^{'} \in \mathbb{R}$, we can write $f_\theta(y, \mid y^{'}) \leq \frac{1}{\sigma_k}$. Furthermore by the compactness of Θ, we can choose $M > 0$ such that for $k = 1, \ldots, g$, $\phi_{k,0}^2, \phi_{k,i}^2, \sigma_k^2 \leq M^2$. Then,

$$(y - \phi_{k,0}^2 - \phi_{k,i}^2 y^{'})^2 \leq (|y + \phi_{k,0}^2 + \phi_{k,i}^2 y^{'})|^2 \leq (|y| + M|y^{'}|)^2, \tag{1.2.12}$$

and

$$0 \leq \sigma_k \leq M(1 + |y^{'}|). \tag{1.2.13}$$

So that for all $\theta \in \Theta$,

$$\begin{aligned} f_\theta(y \mid y^{'}) &\geq \max_{k=1,\ldots,g} \frac{\pi_k}{\sigma_k} f_k(\frac{y_t - \phi_{k,0} - \sum_{i=1}^{p_k} \phi_{k,i} y_{t-i})}{\sigma_k} \qquad (1.2.14) \\ &\geq \max_{k=1,\ldots,g} \frac{\pi_k}{M(1 + |y^{'}|)} f_k(\frac{y_t - \phi_{k,0} - \sum_{i=1}^{p_k} \phi_{k,i} y_{t-i})}{\sigma_k} \\ &\geq \frac{1}{g} \frac{1}{M(1 + |y^{'}|)} f_k(\frac{|y| + M|y^{'}|}{\sigma_k}) > 0. \end{aligned}$$

$\max_{k=1,\dots,g} \pi_k \geq \frac{1}{g}$ and $\sum_{k=1}^{g} \pi_k = 1$ so that the second part of Condition1.2 follows.

By the definition of $f_\theta(\cdot)$ and proof of the second part of Condition1.2 , Equation b_+ is trivially dominated by a positive constant thus the first part of the third part of Condition1.2 holds. To prove the second part,

$$\frac{1}{\sigma_k} \geq \inf_\theta f_\theta(Y_1 \mid Y_0) \geq \frac{1}{gM} \cdot \frac{1}{1+|Y_0|} f_k(\frac{(|Y_1| + M|Y_0|}{\sigma_k}) > 0 \tag{1.2.15}$$

So that

$$\log(\inf_\theta f_\theta(Y_1 \mid Y_0)) \geq \log|\frac{1}{gM}| + \log(\frac{1}{1+|Y_0|}) + \log(f_k(\frac{|Y_1| + M|Y_0|}{\sigma_k})) > 0 \tag{1.2.16}$$

$$= -\log|gM| - \log(1+|Y_0|) + \log(f_k(\frac{|Y_1| + M|Y_0|}{\sigma_k}))$$

$$\geq -\log|gM| - 0 + \log(f_k(\frac{|Y_1| + M|Y_0|}{\sigma_k})) > -\infty$$

using the fact that $EY_t^2 < \infty$, $E\log(1+|Y_0|) \leq E|Y_0| < \infty$, hence the second part of of the third part of Condition1.2 follows.

2. Main Results: Aymptotic properties of the MAR model

Our first result in this section is concerned with the consistency of the MLE of the MAR model.

2.1 Consistency of the maximum likelihood estimator of the MAR model

Theorem 2.1. *Let* $Y_t = (y_t, \dots, y_{t-p+1})'$*, each* y_t *be an MAR model as defined in Equation* (1.1.3)*. Under some mild assumptions, for any* $z_0 \in 1, \dots, g$

$$\lim_{n\to\infty} \hat{\theta}_{n,z_0} = \theta^0 \text{ a.s.}, \tag{2.1.1}$$

where, $\hat{\theta}_{n,z_0} = arg\max_{\theta\in\Theta} l_n(\theta, z_0)$ *is the maximum likelihood estimator of* θ*.*

Proof 2.1. The proof of the above theorem largely uses the results in (Douc et al., 2004).Their assumptions/proofs are hinged on the paper by (Wald, 1949) which says that there exists a deterministic asymptotic criterion function $l(\theta)$ such that $n^{-1}l_n(\theta, z_0) \to l(\theta)$ a.s. uniformly with respect to $\theta \in \Theta$.

The conditional form of the log likelihood function that is $l_n(\theta, z_0)$ is considered instead of $l(\theta)$. So that proving consistency of the maximum likelihood estimator of the MAR model involves checking that the limit of the normalized log-likelihood is only maximized at the true value of the parameter (θ^0) that is, $l(\theta) \leq l(\theta^0)$.

Now, by the first and second part of Condition1.2 above, for $(Y_t, t \geq 0)$, $Y_t = (y_t, \dots, y_{t-p+1})'$ each y_t being an MAR process. Then the following holds for all $\theta \in \Theta$.

$$\sup_{\theta\in\Theta} |l_n(\theta, z_0) - l_n(\theta)| \leq \frac{1}{(1-\rho)^2} \qquad \text{a.s for some } 0 \leq \rho < 1t \tag{2.1.2}$$

where

$$l_n(\theta, z_0) = \log P_\theta(Y_t \mid Y^{(0,t-1)}, Z_0 = z_0), \qquad l_n(\theta) = \log P_\theta(Y_t \mid Y^{(0,t-1)}), \tag{2.1.3}$$

$$\rho = 1 - \frac{\mu_-}{\mu_+}, \qquad 0 \leq \mu_- = \inf_\theta \inf_{i,j} a_{i,j} \quad \text{and} \qquad \mu_+ = \sup_\theta \sup_{i,j} a_{i,j} < 1.$$

So that, $\frac{1}{n} l_n(\theta) = \frac{1}{n} \sum_{t=1}^n \log P_\theta(Y_t \mid Y^{(0,t-1)})$ can be approximated by $\frac{1}{n} \sum_{t=1}^n \log P_\theta(Y_t \mid Y^{(-\infty,t-1)})$, where $\frac{1}{n} \sum_{t=1}^n \log P_\theta(Y_t \mid Y^{(-\infty,t-1)})$ is the sample mean of observations from a two-sided stationary ergodic sequence of random variables in $L^{'}$. We summarize this in the following corollary.

Corollary 2.1. Given that the process Y_t satisfies Condition1.2. Then for all z_0 and $\theta \in \Theta$, the following holds,

$$\lim_{n\to\infty} \frac{1}{n} l_n(\theta, z_0) = l(\theta). \tag{2.1.4}$$

Proof 2.2. We adapt the following notation from (Douc et al., 2004),

$$\begin{aligned} \Delta_{t,m,z}(\theta) &= \log P_\theta(Y_t \mid Y^{(-m,t-1)}, Z_{-m} = z_{-m}) \qquad \text{and} \\ \Delta_{t,m}(\theta) &= \log P_\theta(Y_t \mid Y^{(-m,t-1)}) \end{aligned} \tag{2.1.5}$$

so that

$$l_n(\theta) = \sum_{t=1}^n \Delta_{t,0}(\theta) \tag{2.1.6}$$

(Douc et al., 2004, Lemma 3) show that $\Delta_{t,m,z}(\theta)$ and $\Delta_{t,m}(\theta)$ are uniform Cauchy sequence and converge uniformly with respect to θ a.s. They also show that they are uniformly bounded in L^1 for all m and that $\lim_{m\to\infty} \Delta_{t,m,z}(\theta) = \Delta_{t,\infty}(\theta)$.
They say that the inequality does not depend on z and is a stationary ergodic process such that the following inequalities hold,

$$\begin{aligned} \sup_\theta \sup_z |\Delta_{t,m,z}(\theta) - \Delta_{t,m^{'},z^{'}}| &\leq \frac{\rho^{t+(m\wedge m^{'})-1}}{1-\rho} \qquad \text{and} \\ \sup_\theta \sup_z |\Delta_{t,m,z}(\theta) - \Delta_{t,m}| &\leq \frac{\rho^{t+m-1}}{1-\rho} \end{aligned} \tag{2.1.7}$$

Setting $m = 0$ and $m^{'} \to \infty$ in the system of Equations (2.1.7) gives

$$\begin{aligned} \sup_\theta |\Delta_{t,0,z}(\theta) - \Delta_{t,\infty}| &\leq \frac{\rho^{t-1}}{1-\rho} \qquad \text{and} \\ \sup_\theta |\Delta_{t,0,z}(\theta) - \Delta_{t,0}| &\leq \frac{\rho^{t-1}}{1-\rho}. \end{aligned} \tag{2.1.8}$$

Pulling them together and summing over all t we have,

$$\sum_{t=1}^n \sup_\theta |\Delta_{t,0}(\theta) - \Delta_{t,\infty}| \leq \frac{2}{(1-\rho)^2} \text{ a.s.} \tag{2.1.9}$$

Thus by Equation (2.1.9) $\frac{1}{n} l_n(\theta)$ can be approximated by the sample mean of a stationary ergodic sequence, uniformly with respect to $\theta \in \Theta$ ((Douc et al., 2004)), so that by the ergodic theorem we can write,

$$\frac{1}{n} l_n(\theta) \to l(\theta) = \mathbb{E}\Delta_{0,\infty}(\theta) \quad \text{a.s.} \tag{2.1.10}$$

This together with Equation 2.1.3 imply that for $\theta \in \Theta$,

$$\lim_{n\to\infty} \frac{1}{n} l_n(\theta, z_0) = l(\theta) \quad \text{a.s.} \tag{2.1.11}$$

For the MAR process, at any initial point z_0, $\frac{1}{n}(l_n(\theta, z_\theta) - l_n(\theta)) \to 0$ uniformly with respect to $\theta \in \Theta$ due to the uniform forgetting of the conditional Markov chain ((Douc et al., 2004)).

Hence, $\hat{\theta}_{n,z_0}$ and $\hat{\theta}_n$ are asymptotically equivalent and are the maximum of $l(\theta_{n,z_0})$ and $l(\theta_n)$ respectively.

By Condition1.2 above, we can have

$$\sup_{\theta} \sup_{1\leq z_0 \leq k} \left| \frac{1}{n} l_n(\theta, z_0) - l(\theta) \right| \to 0 \quad \text{as } n \to \infty \tag{2.1.12}$$

So that given the following conditions,

Condition B

- The equality

$$\theta = \theta^0 \text{implies that} \quad E\left[\log \frac{P_\theta(y^{(1,p)} \mid y_0)}{P_{\theta^0}(y^{(1,p)} \mid y_0)}\right] = 0 \qquad \text{for all } p \geq 1 \tag{2.1.13}$$

.

- The expectation

$$E\left[E_{\theta^0}\left[\log \frac{P_\theta(y^{(1,p)} \mid y_0)}{P_{\theta^0}(y^{(1,p)} \mid y_0)} \mid y_0\right]\right] = 0 \qquad \text{for all } p \geq 1 \tag{2.1.14}$$

.

We are able to show that the stationary laws of the observed process associated with two different values of the parameters (say $\mathbb{P}^Y_\theta, \mathbb{P}^Y_{\theta^0}$) do not coincide unless the parameters do ((Douc et al., 2004)).

Hence for $Y_t = (y_t, \ldots, y_{t-p+1})'$, by Condition1.2 and Condition2.1, we can write

$$\mathbb{P}^Y_\theta = \mathbb{P}^Y_{\theta^*} \quad \text{implies} \quad E\left[\log \frac{P_{\theta^0}(y^{(1,p)} \mid y_0)}{P_{\theta^0}(y^{(1,p)} \mid y_0)}\right] = 0 \qquad \text{for all } p \geq 1 \tag{2.1.15}$$

Then

$$l(\theta) = l(\theta^0) \qquad \text{implies that} \qquad \theta = \theta^0 \tag{2.1.16}$$

For the MAR model, at any initial point z_0, $\frac{1}{n}(l_n(\theta, z_0) - l_n(\theta)) \to 0$ uniformly with respect to $\theta \in \Theta$ this follows from Proposition 2.1.12. The proposition also establishes the consistency of the conditional log-likelihood of the model. Furthermore, the geometric ergodicity of the chain Y_t (and by implication the process y_t) establishes the β-mixing property and hence absolute regularity of the process y_t so that Equation (**??**) is established. This together with Propositions 2.1.12 and Proposition**??** as well as the identifiably condition established by Condition2.1 and Lemma **??** prove the consistency of the maximum likelihood estimators of the MAR model.

2.2 Asymptotic normality of the maximum likelihood estimator of the MAR model

Theorem 2.2. *Let* $Y_t = (y_t, \ldots, y_{t-p+1})'$, *each* y_t *be an MAR model as defined in Equation* (1.1.3)*. Given the theorem above holds and assume that* $E(\epsilon_t^4) < \infty$ *and that the Fisher information matrix* $(I(\theta^0))$ *is positive definite, then for all* $z_0 \in 1, \ldots, g$ *then,*

$$\sqrt{n}(\hat{\theta}_{n,z_0} - \theta^0) \to \mathcal{N}(0, (I(\theta^0))^{-1}) \tag{2.2.1}$$

where

$$I(\theta^0) = -E_{\theta^0} \frac{\partial^2 \log p_{\theta^0}(Y_t \mid Y_{-\infty}, \ldots, Y_{t-1})}{\partial\theta\partial\theta'}. \tag{2.2.2}$$

Proof 2.3. The proof of asymptotic normality makes use of the following,

- A central limit theorem (CLT) for the Fisher score function $\frac{1}{\sqrt{n}} \frac{\partial l_n(\theta^0, z_0)}{\partial\theta}$.
- A local uniform law of large numbers for the observed Fisher information $\frac{1}{n} \frac{\partial^2 l_n(\theta^0, z_0)}{\partial\theta\partial\theta'}$ in the neighborhood of θ^0.

(Douc et al., 2004) express the score function and the observed fisher information as functions of conditional expectations of the complete score function and the complete Fisher information.

2.2.1 A central Limit theorem for the score function

The method here for the Fisher identity is due to (Louis, 1982) (see also (Tanner, 1993)). The Louis *Missing Information Principle* says that,

Observed Information=Complete Information - Missing Information.

Now, for all z_0 and $\theta \in \Theta$,

$$\begin{aligned} \frac{1}{\sqrt{n}} \frac{\partial l_n(\theta^0, z_0)}{\partial\theta} &= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \log p_{\theta^0}(Y_t \mid Y^{(0,t-1)}, Z_0 = z_0)}{\partial\theta} \\ &= \frac{1}{\sqrt{n}} \frac{\partial \sum_{t=1}^{n} \Delta_{t,0,z_0}(\theta^0)}{\partial\theta}. \end{aligned} \tag{2.2.3}$$

Using the notation in the proof of Corollary 2.1, write,

$$\begin{aligned} \frac{\partial \Delta_{t,0,z_0}(\theta)}{\partial\theta} &= E_\theta \left[\sum_{i=1}^{t} \phi(\theta, Q_{i-1}, Q_i) \mid Y^{(0,t)}, Z_0 = z_0 \right] \\ &- E_\theta \left[\sum_{i=1}^{t-1} \phi(\theta, Q_{i-1}, Q_i) \mid Y^{(0,t-1)}, Z_0 = z_0 \right] \end{aligned} \tag{2.2.4}$$

where

$$\begin{aligned} \phi(\theta, Q_{i-1}, Q_i) &= \phi(\theta, (Z_{i-1}, Y_{i-1}), (Z_i, Y_i)) \\ &= \frac{\partial \log(a_{Z_{i-1}, Z_i} f_\theta(Y_i \mid Y_{i-1}, Z_i))}{\partial\theta} \end{aligned} \tag{2.2.5}$$

is the conditional score function of (Z_i, Y_i) given (Z_{i-1}, Y_{i-1}).

Similarly, for $m \geq 0$,

$$\frac{\partial \Delta_{t,m}(\theta)}{\partial \theta} = E_\theta \left[\sum_{i=1}^{t} \phi(\theta, Q_{i-1}, Q_i) \mid Y^{(-m,t)} \right] - E_\theta \left[\sum_{i=1}^{t-1} \phi(\theta, Q_{i-1}, Q_i) \mid Y^{(-m,t-1)} \right], \tag{2.2.6}$$

consider the filtration $\mathcal{F}_t = \sigma(Y_s, s \leq t)$ for all, $t \in \mathbb{Z}$. By the dominated convergence theorem, we can write,

$$E_{\theta^0}\Bigg[\sum_{i=-\infty}^{t-1} \Big(E_{\theta^0} \left[\phi(\theta^0, Q_{i-1}, Q_i) \mid Y^{(-\infty,t)} \right] - E_{\theta^0} \left[\phi(\theta^0, Q_{i-1}, Q_i) \mid Y^{(-\infty,t-1)} \right] \Big) \mid Y^{(-\infty,t-1)} \Bigg] = 0, \tag{2.2.7}$$

where

$$E_{\theta^0} \left[\phi(\theta^0, Q_{i-1}, Q_i) \mid Y^{(-\infty,t-1)} \right] = E_{\theta^0} \left[E_{\theta^0} \left[\phi(\theta^0, Q_{i-1}, Q_i) \mid Y^{(-\infty,t-1)}, Z_{t-1} \right]) \mid Y^{(-\infty,t-1)} \right] = 0 \tag{2.2.8}$$

So that $\{\frac{\partial \Delta_{t,\infty}(\theta^0)}{\partial \theta}\}_{t=-\infty}^{\infty}$ is an $\mathcal{F}_t = \sigma(Y_s, s \leq t)-$ adapted, stationary, ergodic and square integrable martingale increment sequence for which the CLT for sums of such sequences (see (Durrett, 1996)) can be applied to show that,

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \Delta_{t,\infty}(\theta^0)}{\partial \theta} \rightarrow \mathcal{N}(0, I(\theta^0)), \tag{2.2.9}$$

where

$$I(\theta^0) = E_{\theta^0}[\frac{\partial \Delta_{0,\infty}(\theta^0)}{\partial \theta} \frac{\partial \Delta_{0,\infty}(\theta^0)}{\partial \theta}^T] \tag{2.2.10}$$

is the asymptotic Fisher information matrix defined as the covariance matrix of the asymptotic score function ((Douc et al., 2004)).

So that

$$\lim_{n \to \infty} E \left\| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (\frac{\partial \Delta_{t,0}(\theta^0)}{\partial \theta} - \frac{\partial \Delta_{t,\infty}(\theta^0)}{\partial \theta}) \right\|^2 = 0 \tag{2.2.11}$$

and

$$\lim_{n \to \infty} E \left\| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (\frac{\partial \Delta_{t,0,z}(\theta^0)}{\partial \theta} - \frac{\partial \Delta_{t,0}(\theta^0)}{\partial \theta}) \right\|^2 = 0 \tag{2.2.12}$$

Hence,

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \Delta_{t,0}(\theta^0)}{\partial \theta} \text{ and } \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \Delta_{t,0,z}(\theta^0)}{\partial \theta} \text{ have the same limiting distribution.} \tag{2.2.13}$$

Therefore , $\frac{\partial \Delta_{t,0}(\theta^0)}{\partial \theta}$ can be approximated in L^2 by a stationary martingale increment sequence.

Thus

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \Delta_{t,0,z}(\theta^0)}{\partial \theta} \to \mathcal{N}(0, I(\theta^0)). \tag{2.2.14}$$

and

$$\frac{1}{\sqrt{n}} \frac{\partial l_n(\theta^0, z_0)}{\partial \theta} \to \mathcal{N}(0, I(\theta^0)). \tag{2.2.15}$$

2.2.2 *Uniform Law of Large numbers for the observed Fisher information*

A locally uniform law of large numbers is explored for the observed Fisher information that is, for all possibly random sequences $\{\theta_n^0\}$ such that $\theta_n^0 \xrightarrow{a.s.} \theta^0$ and

$$-\frac{1}{n} \frac{\partial^2 l_n(\theta_n^0, z_0)}{\partial \theta \partial \theta'} \tag{2.2.16}$$

converges a.s. to the Fisher information matrix at θ^0.

First express the observed Fisher information in terms of the hessian of the complete log-likelihood, we do this by leaning on the Louis missing information principle [see (Louis, 1982),(Tanner, 1993), (Wong and Li, 2000)]. The basic idea in the principle leads to,

$$\begin{aligned} &\frac{\partial^2 \log p_\theta(Y^{(1,n)} \mid Y_0, Z_0 = z_0)}{\partial \theta \partial \theta'} \\ &\qquad = E_\theta \left[\sum_{i=1}^{n} \psi(\theta, Q_{i-1}, Q_i) \mid Y^{(0,n)}, Z_0 = z_0 \right] \\ &\qquad\qquad + var_\theta \left[\sum_{i=1}^{n} \phi(\theta, Q_{i-1}, Q_i) \mid Y^{(0,n)}, Z_0 = z_0 \right], \end{aligned} \tag{2.2.17}$$

where

$$\begin{aligned} \psi(\theta, Q_{i-1}, Q_i) &= \psi(\theta, (Z_{i-1}, Y_{i-1})(Z_i, Y_i)) \\ &= \frac{\partial^2 \log(a_{Z_{i-1}, Z_i} f_\theta(Y_i \mid Y_{i-1}, Z_i))}{\partial \theta \partial \theta'}. \end{aligned} \tag{2.2.18}$$

Also,

$$\begin{aligned} &E_\theta \left[\sum_{i=1}^{n} \psi(\theta, Q_{i-1}, Q_i) \mid Y^{(0,n)}, Z_0 = z_0 \right] \\ &\qquad = \sum_{t=1}^{n} \left(E_\theta \left[\sum_{i=1}^{t} \psi(\theta, Q_{i-1}, Q_i) \mid Y^{(0,t)}, Z_0 = z_0 \right] \right. \\ &\qquad\qquad \left. - E_\theta \left[\sum_{i=1}^{t-1} \psi(\theta, Q_{i-1}, Q_i) \mid Y^{(0,t-1)}, Z_0 = z_0 \right] \right) \end{aligned} \tag{2.2.19}$$

and

$$var_\theta\left[\sum_{i=1}^{n}\phi(\theta,Q_{i-1},Q_i)\mid Y^{(0,n},Z_0=z_0\right]$$
$$=\sum_{t=1}^{n}\left(var_\theta\left[\sum_{i=1}^{t}\phi(\theta,Q_{i-1},Q_i)\mid Y^{(0,t)},Z_0=z_0\right]\right.$$
$$\left.-\,var_\theta\left[\sum_{i=1}^{t-1}\phi(\theta,Q_{i-1},Q_i)\mid Y^{(0,t-1)},Z_0=z_0\right]\right). \quad (2.2.20)$$

As $t\to\infty$ the initial condition on Y_0 becomes more trivial.

Thus for $t\geq 1$ and $m\geq 0$,define,

$$\frac{\partial\Delta_{t,m}(\theta)}{\partial\theta}=E_\theta\left[\sum_{i=-m+1}^{t}\psi(\theta,Q_{i-1},Q_i)\mid Y^{(-m,t)}\right]$$
$$-\,E_\theta\left[\sum_{i=-m+1}^{t-1}\psi(\theta,Q_{i-1},Q_i)\mid Y^{(-m,t)}\right] \quad (2.2.21)$$

and

$$\Gamma_{t,m}(\theta)=var_\theta\left[\sum_{i=-m+1}^{t}\phi(\theta,Q_{i-1},Q_i)\mid Y^{(-m,t)}\right]$$
$$-\,var_\theta\left[\sum_{i=-m+1}^{t-1}\phi(\theta,Q_{i-1},Q_i)\mid Y^{(-m,t-1)}\right] \quad (2.2.22)$$

Now, $\frac{\partial\Delta_{t,m}(\theta)}{\partial\theta}$ and $\Gamma_{t,m}(\theta)$ both converge to $\frac{\partial\Delta_{t,\infty}(\theta)}{\partial\theta}$ and $\Gamma_{t,\infty}(\theta)$ respectively in L^1 as $m\to\infty$. It also follows that $\{\frac{\partial\Delta_{t,m}(\theta)}{\partial\theta}\}_{t=1}^{\infty}$ and $\{\Gamma_{t,\infty}(\theta)\}_{t=1}^{\infty}$ are stationary and ergodic.

Thus, the observed Fisher information will converge to

$$-E_{\theta^0}\left[\frac{\partial\Delta_{t,m}(\theta^0)}{\partial\theta}+\Gamma_{t,\infty}(\theta^0)\right] \text{ (see (Douc et al., 2004).)} \quad (2.2.23)$$

For all z_0, the Fisher Information identity implies that

$$\frac{1}{n}E_\theta\left[\frac{\partial l_n(\theta,z_0)}{\partial\theta}\frac{\partial l_n(\theta,z_0)^T}{\partial\theta}\mid Y_0,Z_0=z_0\right]$$
$$=-\frac{1}{n}E_\theta\left[\frac{\partial^2 l_n(\theta,z_0)}{\partial\theta\partial\theta'}\mid Y_0,Z_0=z_0\right] \quad (2.2.24)$$

Finally, the Louis missing information principle ((Louis, 1982) and (Tanner, 1993)) show that the limits in n of the two quantities in Equation (2.2.24) both coincide with the Fisher information at θ^0 which completes the proof.

3. Applications

3.1 An MAR approach to measuring VaR and ES

We apply the MAR (3;2,2,1) models with both Gaussian and Student-t innovations to estimating 1% and 5% one step ahead out of sample VaR and ES for the Standard and Poor

(S&P500) log returns. The data covered the period between 2002-06-24 and 2012-06-22. Descriptive statistics computed gave a Kurtosis and skewness of 1.3761 and 8.4558 respectively indicating that the data is far from Gaussian. The MAR(3;2,2,1) model is a MAR model with three AR components. The first two AR components are of order two and the third one is of order one, that is, $p_1 = p_2 = 2$, $p_3 = 1$ and $k = 3$. The model is such that,

$$y_t = \begin{cases} \phi_{1,0} + \phi_{1,1}y_{t-1} + \phi_{1,2}y_{t-2} + \sigma_1\epsilon_1(t) & \text{with probability } \pi_1 \\ \phi_{2,0} + \phi_{2,1}y_{t-1} + \phi_{2,2}y_{t-2} + \sigma_2\epsilon_2(t) & \text{with probability } \pi_2 \\ \phi_{3,0} + \phi_{3,1}y_{t-1} + \sigma_3\epsilon_3(t) & \text{with probability } \pi_3, \end{cases}$$

with conditional distribution

$$\begin{aligned} F_{t|t-1}(x) = \pi_1 F_1 &\left(\frac{y_t - \phi_{11}y_{t-1} - \phi_{12}y_{t-2}}{\sigma_1}\right) \\ + \pi_2 F_2 &\left(\frac{y_t - \phi_{21}y_{t-1} - \phi_{22}y_{t-2}}{\sigma_2}\right) \\ + \pi_3 F_3 &\left(\frac{y_t - \phi_{31}y_{t-1}}{\sigma_3}\right). \end{aligned} \tag{3.1.1}$$

3.1.1 The approach

- The parameters of the model is estimated by the Maximum (conditional) likelihood method using the EM algorithm of (Dempster et al., 1977). The standard errors of this parameter estimates can be computed using (Louis, 1982) (see (Wong and Li, 2001) for a more detailed description).
- One step ahead predictive distribution is then computed for the returns series based on the MAR model (see (Boshnakov, 2009)).
- VaR is computed as the $100_\alpha\%$ quantile of the predictive distribution and ES is computed as, $E[r_t \mid r_t > VaR_\alpha]$.

3.1.2 VaR and ES estimation results

Table 3.1.2 below shows the results of the estimated VaR and ES. The results as interpreted thus; an investor that holds a long position worth 100,000GBP in S&P500, then estimates 1-day VaR based on the MAR(3;2,2,1) model with Gaussian innovations at 1% is computed as, 100,000X0.0349 =3,490GBP with corresponding ES as 5,490GBP.

The results were compared to the Empirical Quantile method, celebrated Riskmetrics method, the AR(2)-GARCH(1,1) models with both Gaussian and students t-innovations.

1% and 5% VaR/ES for daily S&P500 log returns (returns are in percentages)

3.2 Interpretation of the results

Leveraging (Tsay, 1997)'s suggestion that for daily returns, the empirical quantiles of 5% and 1% are decent estimates of the quantiles of the return distribution, we treat the results based on empirical quantiles as conservative estimates of the true VaR (i.e., lower bounds, We do not backtest) in this literature, we simply comment on the range of values across all the methods considered. We find that the approaches based on MAR models give values close to the empirical quantiles, while the approaches based on AR-GARCH models tend to underestimate VaR and ES, these results agree with the results in (Tsay, 1997).

	VaR		ES	
	1%	5%	1%	5%
Riskmetrics	0.0254	0.0180	0.0291	0.0226
GARCH(1,1)-norm	0.0264	0.0187	0.0302	0.0234
AR(2)-GARCH(1,1)-norm	0.0261	0.0184	0.0299	0.0231
GARCH(1,1)-t	0.0265	0.0187	0.0302	0.0235
AR-GARCH(1,1)-t	0.0263	0.0186	0.0302	0.0234
Empirical Quantile	0.0409	0.0216	0.0573	0.0341
EVT Threshold (0.019)	0.0391	0.0192	0.0541	0.0319
EVT -GEV	0.0411	0.0217	0.0579	0.0341
MAR(3;2,2,1)-norm	0.0358	0.0215	0.0455	0.0306
MAR(3;2,2,1)-t	0.0349	0.0169	0.0549	0.0294

4. Summary

We have considered the asymptotic properties of the MLE of the MAR process. We considered a vector of MAR processes (Y_t) as a markov regime autoregressive process with a compact and finite hidden space.

We considered the conditional form of the log likelihood function that is $l_n(\theta, z_0)$ instead of $l(\theta)$ and show that the Maximum Likelihood Estimate of the MAR model is both consistent and asymptotically normal.

We also considered the out of sample VaR and ES measures, at tail probabilities $\alpha = 1\%$ and $\alpha = 5\%$ for each of S&500,PWe treated the results based on empirical quantiles as conservative estimates of the true VaR (i.e., lower bounds), and find that the approaches based on EVT and MAR models give significantly better results as they give values close to the empirical quantiles while the approaches based on GARCH models tend to underestimate VaR and ES.

We found that the MAR(3;2,2,1) models with Gaussian and Student-t innovations consistently perform well at both $\alpha = 5\%$ and $\alpha = 1\%$.

References

Akinyemi, M. I. (2013). *Mixture autoregressive models: asymptotic properties and application to financial risk*. PhD thesis, Probability and Statistics Group, School of Mathematics, University of Manchester.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(2):pp. 283–301.

Boshnakov, G. N. (2009). Analytic expressions for predictive distributions in mixture autoregressive models. *Stat. Prob. Let*, 79(15):1704–1709.

Boshnakov, G. N. (2011). On first and second order stationarity of random coefficient models. *Linear Algebra Appl.*, 434(2):415–423.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Douc, R., Moulines, E., and Ryden, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with markov regime. *The Annals of Statistics*, 32(5):pp. 2254–2304.

Durrett, R. (1996). *Stochastic calculus: a practical introduction*. CRC Press, Boca Raton, London.

Jin, S., Yang-Wallentin, F., and Christoffersson, A. (2015). Asymptotic efficiency of the pseudo-maximum likelihood estimator in multi-group factor models with pooled data. *British Journal of Mathematical and Statistical Psychology*.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):pp. 226–233.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *J. Math. Psychol.*, 47(1):90–100.

Nguyen, H. and McLachlan, G. (2015). Maximum likelihood estimation of gaussian mixture models without matrix operations. *Advances in Data Analysis and Classification*, pages 1–24.

Tanner, M. A. (1993). *Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions*. Springer-Verlag, New York ,London, 2nd edition edition. Springer series in statistics.

Tsay, R. S. (1997). *Analysis of Financial Time Series*. John Wiley & Sons, Inc., New Jersey, second edition. Wiley Series in Probability and Statistics.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, 20(4):pp. 595–601.

Wong, C., Chan, W., and Kam, P. (2009). A student t-mixture autoregressive model with applications to heavy-tailed financial data. *Biometrika*, 96(3):751–760.

Wong, C. and Li, W. (2000). On a mixture autoregressive model. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 62(1):95–115.

Wong, C. S. and Li, W. K. (2001). On a mixture autoregressive conditional heteroscedastic model. *Journal of the American Statistical Association*, 96(455):982–995.