On-line estimation of ARMA models using Fisher-scoring†

Abdelhamid Ouakassea and Guy Mélardb

aFederal Public Service Justice of Belgium, Brussels, Belgium; bUniversité libre de Bruxelles (ULB), ECARES, Brussels, Belgium

(Received 19 November 2013; final version received 3 April 2014)

Recursive estimation methods for time series models usually make use of recurrences for the vector of parameters, the model error and its derivatives with respect to the parameters, plus a recurrence for the Hessian of the model error. An alternative method is proposed in the case of an autoregressive-moving average model, where the Hessian is not updated but is replaced, at each time, by the inverse of the Fisher information matrix evaluated at the current parameter. The asymptotic properties, consistency and asymptotic normality, of the new estimator are obtained. Monte Carlo experiments indicate that the estimates may converge faster to the true values of the parameters than when the Hessian is updated. The paper is illustrated by an example on forecasting the speed of wind.

Keywords: time series; ARMA processes; recursive estimation; Fisher information matrix

AMS 2000 MSC: Primary: 62M10, 62F10; Secondary: 93E10

1. Introduction

The development of estimation methods of the parameters of statistical and econometric models was influenced by the availability of more powerful computers. Numerical calculations are lighter and faster with the increased speed of computers, and bigger data bases can be used. For nonlinear models, it is generally not possible to find the estimator analytically so numerical optimisation procedures are applied to obtain the maximum likelihood or even the least squares estimator. These procedures are iterative and make use of all the data at each iteration. They are called off-line because they are applied when all the data are available. Each time we have a new observation the whole estimation procedure has to be repeated. This is not a problem with quarterly or monthly data but availability of large capacity memory also implies that much more data are stored and more frequently. Instead of collecting data at a yearly, quarterly or monthly level, data are more and more collected in real time, starting with financial markets. Also, new fields of applications have appeared, like mobile telecommunications or fluid flow management, where quick automated decisions are required.

When the interval of time between two observations is very short, working with past, off-line, methods becomes inefficient if all data need to be used at high frequency rates and doing huge calculations, because of the expensive computation power needed as well as the memory space. Instead of being used by humans on their desks, the work should be done ‘on the spot’ by computer systems and in an automated way. The idea is to use on-line or recursive methods. They make use of a very small subset of data at each time. These methods appeared first in linear models (Plackett, 1950, who referred to Gauss) when computation was a major annoyance. In statistics they reappeared later (Brown, Durbin, & Evans, 1975) as a way to check the stability of model specification with respect to time. In the discussion of that paper, the influence of Kalman (1960) is clear. Recursive methods became particularly interesting in the context of time series models, see Young (1985). These methods were indeed developed mainly in engineering, under the name of Recursive Identification, for data available on-line in telecommunications, transmissions, management of fluids, etc. For some recent contributions to recursive estimation methods, see Benveniste, Métivier, and Priouret (1990), Guo (1994), Kushner and Yin (1997), Chen (2002), Moulines, Priouret, and Roueff (2005), Dahlhaus and Subba Rao (2007), Gerencsér and Prokaj (2010), Kirshner, Maggio, and Unser (2011) and Marelli, You, and Fu (2013).

Among these recursive methods there is the RML (recursive maximum likelihood) method which was introduced by Söderström (1973), see also Young (1984). We know that, under general conditions, the (off-line) maximum likelihood method gives an estimator which is asymptotically efficient, i.e. it is distributed asymptotically like a normal law whose asymptotic variance-covariance matrix is equal

*Corresponding author. Email: gmelard@ulb.ac.be
†This work was done while the first author was with Université libre de Bruxelles, Département de Mathématique.
to the Cramèr–Rao upper bound. Under certain conditions, Ljung and Söderström (1983) have shown that the RML estimator has the same asymptotic properties as the maximum likelihood estimator. But they noticed that for a finite series, \( \{y_1, \ldots, y_N\} \), the maximum likelihood estimator is always better than the RML estimator. The RML estimator is based on a first-order approximation of the Taylor expansion of the sum of squares of the errors. Let \( \beta \) be the vector of parameters of the model. As we will see in Section 2, the estimate at time \( t \), \( \hat{\beta}_t \), makes use of the value at the previous time, \( \hat{\beta}_{t-1} \), but also of a matrix \( R_t \) which is an approximation of the Hessian of the sum of squares of errors. A recurrence for the residual is used but also a recurrence for the derivative of the error with respect to the parameters and an updating recurrence for the Hessian.

Mélard (1989) and Zahaf (1999) observed that the latter recurrence, with highly variable successive values of \( R_t \), is often the cause for wild variations in the estimates and proposed a modified RML estimator for autoregressive-moving average (ARMA) models. While keeping the spirit of the algorithm, instead of the recurrence for the Hessian \( R_t \), Zahaf (1999) proposed to use the evaluation of the asymptotic Fisher information at the current value of the estimator, \( \beta = \hat{\beta}_{t-1} \). This is a recursive analogue of the well-known Fisher-scoring algorithm, e.g. Kennedy and Gentle (1980, p. 450), Sen and Singer (1993, p. 205). Intuitively this should not change the asymptotic properties.

Zahaf (1999) noticed that the asymptotic theory developed by Ljung (1977), Solo (1981) and Ljung and Söderström (1983) no longer applies. He outlined an asymptotic theory based on the stochastic approximation of Robbins–Monro following Duflo (1997) but it was not complete. Moreover convergence in law of the estimator rested on a conjecture which was later proved to be wrong. For these reasons, after vain attempts including with the alternative approach of Kushner and Huang (1979), we preferred to adapt the approach of Ljung and Söderström (1983). An alternative which is discussed later should be to use Chen (2002) stochastic approximation theory with expanding truncations.

In Section 2, we remind the necessary concepts of RML estimation in order to be able to introduce our version at the beginning of Section 3. The remaining of Section 3 is devoted to the main theorems in order to establish consistency and asymptotic normality of the new estimator. This is done under the very general condition that fourth-order moments are finite instead of assuming that the observations are bounded, like Ljung and Söderström (1983) or that moments of order \( 4/(1-\delta) \), for some strictly positive \( \delta \), are finite, like Solo (1981). This is a clear improvement with respect to the literature. In Section 4, we show small samples results obtained by Monte Carlo simulations. They indicate that the new estimator can be an improvement over the classical RML estimator. Section 5 will present an example of wind forecasting.

2. RML estimation

Let us first describe the RML estimator before introducing how we have modified it. The algorithm for that estimator is derived from the off-line maximum likelihood estimator, see Ljung (1978) and Åström (1980). We assume for simplicity that the observations \( \{y_t; t = 1, \ldots, N\} \) follow a univariate ARMA\((p,q)\) model defined by the equation:

\[
y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p} = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q},
\]

where the roots of the autoregressive and moving average polynomials \( \Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p \) and \( \Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q \) are outside of the unit circle, \( \phi_p \neq 0 \) and \( \theta_q \neq 0 \), and \( e_t \)'s are i.i.d. random variables with \( E(e_t) = 0 \) and \( E(e_t^2) = \sigma^2 > 0 \). Let \( \beta = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)^T \) be the vector of the parameters of interest, where \( ^T \) denotes transposition, and let \( \beta^* \) be the true value of \( \beta \). Let also \( \Phi^*(B) \) and \( \Theta^*(B) \), respectively, the polynomials \( \Phi(B) \) and \( \Theta(B) \) when \( \beta = \beta^* \). The estimator at time \( t \) will be denoted \( \hat{\beta}_t = (\hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q)^T \).

For a given \( \beta \), the forecast \( \hat{y}_{t|t-1}(\beta) \) for time \( t \) can be computed at time \( t-1 \), provided we replace the true errors \( e_{s, t} \), \( s < t \), also called innovations, by the residuals \( e_t(\beta) = y_t - \hat{y}_{t|t-1}(\beta) \), computed by recurrence. This requires suitable initial values whose effect can be neglected because of the assumption on the polynomials. In off-line estimation, under the Gaussian assumption on innovations \( e_t \)'s, the maximum likelihood estimator is obtained, for large \( N \), by minimising the sum of squares of the residuals

\[
V_N(\beta) = \frac{1}{2} \sum_{t=1}^{N} e_t^2(\beta).
\]

Example 1: Specific parts will be illustrated with the ARMA\((1,1)\) model defined by

\[
y_t - \phi y_{t-1} = e_t - \theta e_{t-1},
\]

with \( \beta^T = (\phi, \theta) \). Note that Equation (3) implies

\[
\hat{y}_{t|t-1}(\beta) = \phi y_{t-1} - \theta e_{t-1}(\beta)
\]

and

\[
y_t - \hat{y}_{t|t-1}(\beta) = y_t - \phi y_{t-1} + \theta (y_{t-1} - \hat{y}_{t-1|t-2}(\beta)),
\]

where the starting value \( \hat{y}_{1|0}(\beta) \) can be taken equal to 0. Indeed the effect of a starting value decreases like \( |\theta|^t \), and the assumption made implies that \( |\theta| < 1 \). This recurrence allows computing \( e_t(\beta) \).

For ARMA models, \( V_N(\beta) \) is a nonlinear function of \( \beta \), so \( V_N(\beta) \) cannot be minimised analytically but well using numerical procedures, requiring many iterations on basis of the data from \( t = 1 \) to \( t = N \). An on-line or recursive
algorithm requires a vector of fixed size, preferably small with respect to $N$. Therefore we want an approximation of the off-line maximum likelihood estimator $\hat{\beta}_N$ that can be obtained by recurrences.

Given $\hat{\beta}_{t-1}$, we want to obtain $\hat{\beta}_t$ close to the minimum of $V_t(\beta)$. By a Taylor expansion of $V_t(\beta)$ around $\hat{\beta}_{t-1}$ limited to the second order we obtain

$$V_t(\beta) \simeq V_t(\hat{\beta}_{t-1}) + \left( \frac{\partial V_t(\beta)}{\partial \beta^T} \right)_{\beta=\hat{\beta}_{t-1}} [\beta - \hat{\beta}_{t-1}] + \frac{1}{2} [\beta - \hat{\beta}_{t-1}]^T \left( \frac{\partial^2 V_t(\beta)}{\partial \beta \partial \beta^T} \right)_{\beta=\hat{\beta}_{t-1}} [\beta - \hat{\beta}_{t-1}].$$

(6)

Minimising the right-hand side with respect to $\beta$ leads to

$$\hat{\beta}_t = \hat{\beta}_{t-1} - \left( \frac{\partial^2 V_t(\beta)}{\partial \beta \partial \beta^T} \right)_{\beta=\hat{\beta}_{t-1}}^{-1} \left( \frac{\partial V_t(\beta)}{\partial \beta^T} \right)_{\beta=\hat{\beta}_{t-1}}.$$  

(7)

Denoting $\psi_t(\beta) = -[\partial \epsilon_t(\beta)/\partial \beta^T]^T$, the opposite of the derivative of $\epsilon_t$ with respect to $\beta$, we have

$$\left[ \frac{\partial V_t(\beta)}{\partial \beta^T} \right]^T = - \sum_{k=1}^{t} \psi_k(\beta) \epsilon_k(\beta) = \left[ \frac{\partial V_{t-1}(\beta)}{\partial \beta^T} \right]^T - \psi_t(\beta) \epsilon_t(\beta),$$

(8)

and a further differentiation yields the Hessian:

$$\frac{\partial^2 V_t(\beta)}{\partial \beta \partial \beta^T} = \frac{\partial^2 V_{t-1}(\beta)}{\partial \beta \partial \beta^T} + \psi_t(\beta) \psi_t^T(\beta) + \frac{\partial^2 \epsilon_t(\beta)}{\partial \beta \partial \beta^T} \epsilon_t(\beta).$$

(9)

In order to evaluate Equation (7), the following approximations are made.

1. We assume that $\hat{\beta}_t$ is close to $\hat{\beta}_{t-1}$, a quite reasonable approximation for large $t$, justifying Equation (6) and

$$\left( \frac{\partial^2 V_t(\beta)}{\partial \beta \partial \beta^T} \right)_{\beta=\hat{\beta}_t} \simeq \left( \frac{\partial^2 V_{t-1}(\beta)}{\partial \beta \partial \beta^T} \right)_{\beta=\hat{\beta}_{t-1}}.$$  

(10)

2. We proceed as if $\hat{\beta}_{t-1}$ were optimal at time $t-1$, i.e.

$$\left( \frac{\partial V_{t-1}(\beta)}{\partial \beta^T} \right)_{\beta=\hat{\beta}_{t-1}} \simeq 0.$$  

(11)

3. Since, for $\beta$ close to $\beta^*$, $\{\epsilon_t(\beta)\}$ will almost behave like a white noise process, i.e. $\epsilon_t(\beta)$ will have a mean close to 0 and be nearly independent from the observations and residuals before time $t$, allowing to neglect the last term of Equation (9).

Then, inserting Equation (10) in Equation (9) evaluated at $\beta = \hat{\beta}_{t-1}$, we have an approximation of the Hessian, $\tilde{R}_t$, which can be computed recursively by

$$\tilde{R}_t = \tilde{R}_{t-1} + \psi_t(\hat{\beta}_{t-1}) \psi_t^T(\hat{\beta}_{t-1}).$$

(12)

Insertion of Equation (11) in Equation (8) evaluated at $\beta = \hat{\beta}_{t-1}$, yields

$$\left( \frac{\partial V_{t-1}(\beta)}{\partial \beta^T} \right)_{\beta=\hat{\beta}_{t-1}}^T = -\psi_t(\hat{\beta}_{t-1}) \epsilon_t(\hat{\beta}_{t-1}).$$

Using the approximation $\tilde{R}_t$ in Equation (7), we have

$$\hat{\beta}_t = \hat{\beta}_{t-1} - \tilde{R}_t^{-1} \psi_t(\hat{\beta}_{t-1}) \epsilon_t(\hat{\beta}_{t-1}).$$

(13)

Denoting $tR_t = \tilde{R}_t$, we have the two equations

$$R_t = R_{t-1} + \frac{1}{t} \{\psi_t(\hat{\beta}_{t-1}) \psi_t^T(\hat{\beta}_{t-1}) - R_{t-1}\}$$

(14)

$$\hat{\beta}_t = \hat{\beta}_{t-1} + \frac{1}{t} R_t^{-1} \psi_t(\hat{\beta}_{t-1}) \epsilon_t(\hat{\beta}_{t-1}).$$

There remains to derive equations for computing $\epsilon_t(\hat{\beta}_{t-1})$ and $\psi_t(\hat{\beta}_{t-1})$. Let us first look at the ARMA(1,1) example (3).

Example 2 We have $\psi_t^T(\beta) = \partial \hat{y}_{t-1}(\beta)/\partial \beta^T$ and differentiation of $\hat{y}_{t-1}(\beta) - \partial \hat{y}_{t-1}(\beta)/\partial \beta^T = (\phi - \theta) y_{t-1}$, which is also deduced from Equation (3), gives the two equations:

$$\frac{\partial \hat{y}_{t-1}(\beta)}{\partial \phi} - \frac{\partial \hat{y}_{t-1}(\beta)}{\partial \theta} = y_{t-1},$$

(15)

$$\frac{\partial \hat{y}_{t-1}(\beta)}{\partial \theta} - \frac{\partial \hat{y}_{t-1}(\beta)}{\partial \phi} = -y_{t-1}.$$  

(16)

The latter can also be written

$$\frac{\partial \hat{y}_{t-1}(\beta)}{\partial \theta} - \theta \frac{\partial \hat{y}_{t-1}(\beta)}{\partial \theta} = -\epsilon_{t-1}(\beta).$$

(17)

Grouping Equations (15) and (17) gives

$$\psi_t(\beta) - \theta \psi_{t-1}(\beta) = \left( \begin{array}{} y_{t-1} \\ -\epsilon_{t-1}(\beta) \end{array} \right).$$

(18)

We can compute $\epsilon_t(\hat{\beta}_{t-1})$ and $\psi_t(\hat{\beta}_{t-1})$ by using equations like (4) and (18) but this requires all the observations $y_s, s = 1, \ldots, t - 1$. Let us derive approximations of $\epsilon_t(\hat{\beta}_{t-1})$ and $\psi_t(\hat{\beta}_{t-1})$ that can be computed by recurrence using additional approximations. A natural approximation consists in using only the current estimator and max($p, q$) previous values of $\epsilon, y$ and $\psi$ as initial values.
Example 3 In the case of Equation (3), $e_t(\hat{\beta}_{t-1})$ is approached by $e_t$, computed by

$$e_t = y_t - \hat{y}_{t|t-1} = y_t - \hat{\beta}_{t-1}(y_{t-1} - \hat{y}_{t-1|t-2}).$$

Let us introduce $\varphi_{t-1}^T = (y_{t-1}, -e_{t-1})$. Using Equation (4), we can write

$$e_t = y_t - \hat{\beta}_{t-1}^T\varphi_{t-1}. \quad (19)$$

Similarly, Equation (18) leads to a natural approximation $\psi_t$ of $\psi_t(\hat{\beta}_{t-1})$

$$\psi_t = \hat{\theta}_{t-1}\psi_{t-1} + \varphi_{t-1}. \quad (20)$$

At time $t$ we only need to know $\varphi_{t-1}$, $\psi_{t-1}$ et $\hat{\beta}_{t-1}$. Adding these equations to those of Equation (14) and performing substitutions, we obtain the system

$$\psi_t = \hat{\theta}_{t-1}\psi_{t-1} + \varphi_{t-1},$$

$$\tilde{R}_t = \tilde{R}_{t-1} + \psi_t\psi_t^T,$$

$$e_t = y_t - \hat{\beta}_{t-1}^T\varphi_{t-1},$$

$$\hat{\beta}_t = \hat{\beta}_{t-1} + \tilde{R}_t^{-1}\psi_t e_t.$$  

Note that Equation (21) is not computationally efficient because of the need to invert $\tilde{R}_t$. There exists a more computationally efficient algorithm (Ljung & Söderström, 1983, Chap. 2, p. 19), where $P_t = \tilde{R}_t^{-1}$ is updated instead of $\tilde{R}_t$. The equation makes use of the inversion lemma and can be written

$$P_t = P_{t-1} - \frac{P_{t-1}\psi_t\psi_t^TP_{t-1}}{1 + \psi_t^TP_{t-1}\psi_t}. \quad (22)$$

Let us now go back to the general case (1). To improve the behaviour of the algorithm, we replace the factor $1/t$ by a sequence $\gamma_t$ of positive scalars decreasing to 0 such that $\sum \gamma_t$ is divergent. If we now denote $\psi_t^T = (y_t, \ldots, y_{t-p+1}, -e_t, \ldots, -e_{t-q+1})$, with a due generalisation of Equation (20), the RML algorithm can now be written:

$$\psi_t = \sum_{k=1}^q \hat{\theta}_{k,t-1}\psi_{t-k} + \varphi_{t-1},$$

$$R_t = R_{t-1} + \psi_t\psi_t^T - R_{t-1}, \quad (23)$$

$$e_t = y_t - \hat{\beta}_{t-1}^T\varphi_{t-1},$$

$$\hat{\beta}_t = \hat{\beta}_{t-1} + \psi_t^T\psi_{t-1}^{-1}e_t.$$  

3. Estimation by the RMLMZ method

Let us now consider a modification of the method of Section 2 called the RMLMZ method. From a theoretical point of view, under some assumptions, the RML algorithm (23) provides a consistent estimator with a rate of convergence $\sqrt{t}$. However, Mélard (1989) and Zahaf (1999) have observed huge variations of $R_t$ with respect to time, which produce disturbances in the RML estimator. While keeping the recursive nature of the algorithm, they have tried to improve its accuracy by replacing the central recurrence (12) for the Hessian $\partial^2 V(\beta)/\partial \beta \partial \beta^T$, by the computation of its expectation at the current value of the estimator. Indeed, $\sigma^2 R_t^{-1}$ is an approximation of the asymptotic covariance matrix $\Gamma(\beta^*)$ of the maximum likelihood estimator. But, $\beta^*$ being unknown, they suggest to replace $\Gamma(\beta^*)$ by the asymptotic covariance matrix evaluated at the last value of the estimator, $\Gamma(\hat{\beta}_{t-1})$. If $\hat{\beta}$ converges to $\beta^*$, which will be shown later, then $\Gamma(\hat{\beta}_{t-1})$ converges to $\Gamma(\beta^*)$. Moreover, $\Gamma(\hat{\beta}_{t-1})$ is the inverse $F^{-1}(\hat{\beta}_{t-1})$ of the Fisher information matrix $F(\beta)$ computed at $\beta = \hat{\beta}_{t-1}$. At each time, we will compute $\sigma^2 F(\hat{\beta}_{t-1})$ and then its inverse $\sigma^2 F^{-1}(\hat{\beta}_{t-1})$ which will replace $R_t^{-1}$ in Equation (23). This is thus a recursive analogue of the well-known Fisher-scoring algorithm, e.g. Kennedy and Gentle (1980, p. 450) and Sen and Singer (1993, p. 205). For a given $\sigma^2$, the algorithm is written:

$$\psi_t = \sum_{k=1}^q \hat{\theta}_{k,t-1}\psi_{t-k} + \varphi_{t-1}, \quad (24)$$

$$e_t = y_t - \hat{\beta}_{t-1}^T\varphi_{t-1},$$

$$\hat{\beta}_t = \hat{\beta}_{t-1} + \gamma \sigma^2 F^{-1}(\hat{\beta}_{t-1})\psi_t e_t,$$

where $\varphi_t$ is like before. Therefore the recurrence for $R_t$ in Equation (23) will no longer be needed. Note that

$$F(\beta) = \sigma^2 E(\psi_t^T(\beta)\psi_t^T(\beta)),$$

where

$$\psi_t^T(\beta) = \sum_{k=1}^q \hat{\theta}_{k,t-1}\psi_{t-k} + \varphi_{t-1},$$

and $\varphi_t^T = (y_t, \ldots, y_{t-p+1}, -e_t, \ldots, -e_{t-q+1})$. Note also that, under a stationarity assumption, $F(\beta)$ does not depend on $t$. For simple models, an analytic expression does exist for $F^{-1}(\beta)$, see Box, Jenkins, and Reinsel (2008). Otherwise, there are simple algorithms for computing $F(\beta)$, see e.g. Klein and Mélard (1989).

But $\sigma^2$ is generally unknown so the algorithm (24) is modified as follows

$$\psi_t = \sum_{k=1}^q \hat{\theta}_{k,t-1}\psi_{t-k} + \varphi_{t-1}, \quad (26)$$

$$e_t = y_t - \hat{\beta}_{t-1}^T\varphi_{t-1}, \quad (27)$$

$$\hat{\beta}_t = \hat{\beta}_{t-1} + \gamma \sigma^2 F^{-1}(\hat{\beta}_{t-1})\psi_t e_t,$$

$$\hat{\sigma}^2_{t+1} = \hat{\sigma}^2_t + \gamma \sigma^2 (e_t^2 - \hat{\sigma}^2_t).$$

In order to avoid problems with uncontrolled recursions we will later change the estimator by a projection mechanism.

Remark 1 Since $F^{-1}(\hat{\beta}_{t-1})$ is to be computed at each time $t$, a matrix inversion is needed at each time. Consequently
the RMLMZ method is not as computationally efficient as the RML method using Equation (22). In principle it should be possible to exploit the properties of the Fisher information matrix \(F(\beta)\) to improve efficiency. In the pure AR(p) model, \(F(\beta)\) is a Toeplitz matrix. Hence there are algorithms like Trench (1964) for inverting it in a number of operations proportional to \(p^2\) instead of \(p^3\). For an ARMA(p, p) model, \(F(\beta)\) can be put under the form of a block Toeplitz matrix with \(p \times p\) blocks of size 2, hence an inversion algorithm can also exploit that structure, e.g. Akaike (1973). For other ARMA(p, q) models with \(p\) different from \(q\) and \(p\) and \(q\) different from 0, some computational improvements can be found, for example by considering matrices of order \(2(p + q)\). We will not further discuss improved algorithms in this paper.

To show convergence of the algorithm to the optimal value, we make two assumptions; the first one is about the errors. 

**Assumption 1 (on the model)** The autoregressive and moving average polynomials have no common root and their roots are all outside of the unit circle (satisfying the causality or stationarity condition and the invertibility condition of the process), and \(\phi_p \neq 0\) and \(\theta_q \neq 0\).

**Assumption 2 (on the errors)** The errors \(\{e_t\}\) have finite fourth-order moment: \(\forall t, E(e_t^4) < \infty\).

Instead of Assumption 2, it is usually supposed, see e.g. Ljung and Söderström (1983), that the sequence of observations \(\{y_t\}\) has a uniform upper bound in absolute value, more precisely \(|y_t| < Y\), where \(Y\) is a random variable with a finite variance. That supposition is not really on the data but rather both on the parameter set and on the probability distribution of the errors. Our assumption is clearer on this respect. Also Ljung and Söderström (1983, p. 191) explicitly exclude errors which are not bounded, such as errors with a normal distribution, which is not the case here. Solo (1981) has slightly stronger assumptions than ours, by assuming that moments of order \(4/(1 - \delta)\), for some strictly positive \(\delta\), are finite.

Let \(D_S = \{ (\beta^T, \sigma^2)^T \in \mathbb{R}^{p+q+1}/\{ the eigenvalues of \(A(\beta)\) are in the unit circle \}\}, hence \(D_S = \{ (\beta^T, \sigma^2)^T \in \mathbb{R}^{p+q+1}/\{ the roots of the moving average polynomial are outside of the unit circle \}\}. Because of the Fisher information matrix, the definition of the set \(D_R\) is also different: 

\(D_R = \{ \beta \in \mathbb{R}^{p+q+1}/\{ the roots of the autoregressive and moving average polynomials are outside of the unit circle, \(F(\beta)\) is invertible, \(\|F^{-1}(\beta)\| < k\} \) for some constant \(k > 0\) large enough. 

\(D_R = \{ (\beta^T, \sigma^2)^T \in \mathbb{R}^{p+q+1}/\{ \beta \in D_B \text{ and } \sigma^2 \geq \delta \text{ and } \sigma^2 \leq \delta' \} \) for some constant \(\delta > 0\) small enough and \(\delta' > \delta\) large enough.

More care is needed than in the original RML algorithm because \(F^{-1}(\beta)\) can become very large. This is especially serious in the analysis where we will need a suitable projection mechanism built into the recursion, simpler than the one used in practice. As will be seen that projection mechanism makes use of \(\beta^*\). This is quite disturbing for a statistician but it follows the suggestion of Benveniste et al. (1990, p. 56) to analyse a simplified algorithm. Example 4 below will show that the analysis is nevertheless very interesting.

In spite of that, the study the statistical properties of the RMLMZ recursive estimator is very technical so most of the details will be given in Appendices 1 (Lemmas A.1–A.8) and 2 (Lemmas A.9–A.24). In Section 4, we will discuss small sample results obtained by Monte Carlo simulation, this time with a more realistic projection mechanism, and show a comparison with the original RML algorithm.

### 3.1. Almost sure convergence

Zahaf (1999) has used results from Duflo (1997) about Robbins-Monro stochastic approximation in order to obtain asymptotic properties for a Newton approximation to the RMLMZ estimator, called the RMLNE estimator. The algorithm has the form \(\hat{\beta}_{t+1} = \hat{\beta}_t + \gamma Y_{t+1}, \) where the conditional expectation of \(Y_{t+1}\) given the past information fulfills \(E[Y_{t+1}/F_t]\) is a measurable function of \(\hat{\beta}_t\). But here \(E[Y_{t+1}/F_t]\) depends on both \(\hat{\beta}_t\) and \(t\), and it is even difficult to deduce convergence of the RMLMZ estimator from its Newton version. Also Kushner and Yin (1997, p. 94) cannot be applied because their assumption (A2.2) is not valid in our case.

The theory contained in Ljung and Söderström (1983) is based on writing the algorithm under the following form

\[
\begin{align*}
\hat{h}_t = A(\hat{x}_{t-1}) h_{t-1} + B(\hat{x}_{t-1}) z_t,
\hat{x}_t = \hat{x}_{t-1} + \gamma Q(t, \hat{x}_{t-1}, h_t),
\end{align*}
\]

(30)

where \(A(\cdot), B(\cdot),\) and \(Q(\cdot, \cdot, \cdot)\) are functions, \(\gamma_t\) is like in Section 2 and \(z_t\) makes use of the data. Like in Ljung (1977), the idea is to associate an ordinary differential equation (ODE) to the algorithm and obtain the attraction domain of an invariant set of that ODE. For the original RML estimator, \(\hat{x}_t = (\hat{\beta}_t^T, \text{vec}(R_t)^T)^T\) and it appears that \(A(\cdot)\) and \(B(\cdot)\) depend only on \(\hat{\beta}_t\). Here we have to consider the same but where \(\hat{x}_t = (\hat{\beta}_t^T, \hat{\sigma}^2_{t+1})^T\)

\[
\begin{align*}
\hat{h}_t = A(\hat{\beta}_{t-1}) \hat{h}_{t-1} + B(\hat{\beta}_{t-1}) z_t,
\left(\hat{\beta}_t, \hat{\sigma}^2_{t+1}\right) = \left(\hat{\beta}_{t-1}, \hat{\sigma}^2_{t}, \gamma Q(t, \hat{\beta}_{t-1}, \hat{\sigma}^2_{t}, h_t)\right),
\end{align*}
\]

(31)

where \(h_t\) is \(q(p + q + 1) \times 1\), \(Q(t, \beta, \sigma^2, h)\) is \((p + q + 1) \times 1,\) and

\[
\begin{align*}
h_t = (e_t, e_{t-1}, \ldots, e_{t-q+1}, \psi_T, \psi_{t-1}^T, \ldots, \psi_{t-q+1}^T)^T,
\end{align*}
\]

(32)
\(Q(t, \beta, \sigma^2, h) = [\sigma^{-2}[F^{-1}(\beta)(h_{q+1}, h_{q+2}, \ldots, h_{2q+p})]^T h_1, \\
h_1^2 - \sigma^2]^T, \)  
\tag{33}

so that \(h_1\) represents \(e_1, (h_{q+1}, h_{q+2}, \ldots, h_{2q+p})^T\) represents \(\psi_1\), and

\[Q(t, \beta_{1-}, \sigma^2, h_1) = [\sigma_{\beta}^{-2}[F_{\sigma}^{-1}(\beta_{1-})\psi_1]^T e_t, e_t^2 - \sigma_{\beta}^2]^T.\] Notice that \(R_t\), obtained by the Fisher information matrix evaluated at \(\beta = \beta_{1-}\), appears in the second term of the right-hand side of the second equation of (31), making derivations very different from Ljung and Söderström (1983). Their theory cannot be applied directly for the RMLMZ algorithm. However, the first equation of (31) still holds with the same choice for the matrices \(A\) and \(B\) as in Equation (30). For an ARMA\((p, q)\) model, it can be seen that

\[\det(A(\beta) - \lambda I) = (-1)^{(q-1)(p+q-1)}/(-\lambda^q - \lambda^{-q} \theta_1) - \lambda^{-q-2} \theta_2 - \ldots - \theta_q)^p + \sigma^2.\]  
\tag{34}

In the following, \(x = (\beta^T, \sigma^2)^T \in D_R\) and we write sometimes \(Q\) with three arguments instead of four. We will make use of Ljung (1977, Theorems 1 and 4), summed up as Lemma A.8 in Appendix 1. Here is the third subset of his conditions, denoted by C, without C7 which is not needed:

\begin{enumerate}
\item[C1:] \(Q(t, x, h)\) is Lipschitz continuous in \(x\) and \(h:\)
\[\|Q(t, x_1, h_1) - Q(t, x_2, h_2)\| < C_1(x, h, \rho, \upsilon) \]
\[\{\|x_1 - x_2\| + \|h_1 - h_2\|\} \]

for \(x_1 \in B(x, \rho)\), an open ball of centre \(x\) and diameter \(\rho\), for \(\rho = \rho(x) > 0\), where \(x \in D_R\), \(h_1 \in B(h, \upsilon)\) for \(\upsilon \geq 0\);

\item[C2:] matrices \(A(\cdot)\) and \(B(\cdot)\) are Lipschitz continuous functions over \(D_R\);

\item[C3:] \(H(\bar{x}) = \lim_{t \to \infty} (1/t) \sum_{k=1}^t Q(k, \bar{x}, \bar{h}_k(\bar{x}))\) exists for all \(\bar{x} \in D_R\);

\item[C4:] for all \(x \in D_R\), \(0 < \lambda < 1\) and \(c < \infty\), the random variable \(k_c(t, \bar{x}, \lambda, c)\) defined by

\[k_c(t, \bar{x}, \lambda, c) = k_c(t - 1, \bar{x}, \lambda, c) + \gamma_1[K_1(\bar{x}, h, \rho(\bar{x}), v(t, \lambda, c)) + \gamma_2(\gamma_1 + \gamma_2)]
\]

and \(k_c(0, \bar{x}, \lambda, c) = 0\) and \(v(t, \lambda, c) = c \sum_{k=1}^t \lambda^{k-1} \gamma_1 \eta(t,k)\), converges to a finite limit when \(t \to \infty\);

\item[C5:] \(\sum_{t=1}^\infty \gamma_t = \infty\);

\item[C6:] \(\lim_{t \to \infty} \gamma_t = 0\).
\end{enumerate}

According to Ljung (1977), these conditions are used in the deterministic case, but the results are valid with probability 1 as far as \(z_i\) defined by Equation (32) is such that the conditions C3 and C4 are satisfied with probability 1.

On the basis of Equations (27) and (29), let us define \(e_i(\beta) = y_i - \beta^T \psi_{i-1}(\beta)\), where \(\psi_{i-1}(\beta) = (y_{i-1}, y_{i-2}, \ldots, -y_{i-q})(\beta), \ldots, - \psi_{i-q}(\beta))^T\), and \(\sigma_i^2(\beta) = \sigma_i^{-2}(\beta) + (1/t)(e_i^{-2}(\beta) - \sigma_i^{-2}(\beta)).\) Hence

\[\sigma_i^2(\beta) = \frac{1}{t} \sum_{k=1}^t e_i^{-2}(\beta).\]  
\tag{35}

Define also

\[\psi_t(\beta) = \sum_{k=1}^q \theta_k \psi_{t-k}(\beta) + \psi_{t-1}(\beta),\]  
\tag{36}

\[R_t(\beta) = R_{t-1}(\beta) + \frac{1}{t} (\psi_{t}(\beta)\psi_t^T(\beta) - R_{t-1}(\beta)) \]

\[= \frac{1}{t} \sum_{k=1}^t \psi_k(\beta)\psi_k^T(\beta).\]  
\tag{37}

**Theorem 1** Under Assumptions 1 and 2, conditions C1–C6 of Ljung (1977, Theorem 4) are satisfied.

The proof is given in Appendix 1 on the basis of Lemmas A.1–A.3.

We will consider the following ODE

\[\frac{\partial x(t)}{\partial t} = \frac{\partial (\beta^T(t), \sigma^2(t))^T}{\partial t} = H(\beta(t), \sigma^2(t)),\]

where

\[H(\beta, \sigma^2) = [\sigma^{-2}[F^{-1}(\beta)E(\psi(\beta)e(\beta)))]^T, E[e^2(\beta)] - \sigma^2]^T.\]

Letting \(f(\beta) = E[\psi(\beta)e(\beta)]\) and \(V(\beta) = E[e^2(\beta)]\), the ODE can be put under the form

\[\frac{\partial \beta(t)}{\partial t} = \sigma^{-2}(t)[F^{-1}(\beta(t))f(\beta(t)),\]  
\tag{38}

\[\frac{\partial \sigma^2(t)}{\partial t} = V(\beta(t)) - \sigma^2(t).\]  
\tag{39}

**Theorem 2** Let

\[c(\beta^*) = \sup_{D \in K_D} \inf_{\beta \in Fr(D)} V(\beta),\]

where \(K_D\) is a set of convex parts of \(D_\beta\) containing \(\beta^*\) and \(Fr(D)\) is the frontier of \(D\). Let \(D_2 = \{\beta^T, \sigma^2]^T \in D_R/V(\beta) \leq c(\beta^*) - g\text{ with a very small positive constant } g\). Under Assumptions 1 and 2, let the recursive RMLMZ.
estimator (26)–(28) replaced by the following recurrences
\[
\left( \hat{\beta}_t, \hat{\sigma}^2_{t+1} \right) = \left[ \left( \hat{\beta}_{t-1}, \hat{\sigma}^2_{t-1} \right) + \frac{1}{t} \left( \hat{\sigma}^2_{t-1}F^{-1}(\hat{\beta}_{t-1})\psi_t \varepsilon_t \right) \right]_{D_{k,D_2}}, \tag{40}
\]
where
\[
[z]_{D_{k,D_2}} = \begin{cases} 
  z & \text{if } z \in D_R \\
  a \text{ point in } D_2 & \text{if } z \notin D_R,
\end{cases}
\]
and
\[
\varepsilon_t = y_t - \hat{\beta}_t^T \psi_{t-1},
\]
\[
\psi_t = \sum_{k=1}^{q} \delta_{k,t-1} \psi_{t-k} + \varphi_{t-1},
\]
\[
(\varepsilon_t, \psi_t)^T \text{ a point in } K \quad \text{if } (\hat{\beta}_t^T, \hat{\sigma}^2_t)^T \notin D_R,
\]
where \(K\) is a compact subset of \(\mathbb{R}^{1+q+1}\) defined in advance. Then \(\hat{\beta}_t\) converges to \(\beta^*\) almost surely when \(t \to \infty\).

**Proof** Given that
\[
\lim_{t \to \infty} \frac{1}{t} \sum_{k=1}^{t} Q(k, \beta, \sigma^2, h_k(\beta)) = E((\sigma^2 F^{-1}(\beta) \psi_t(\beta) \varepsilon_t(\beta))^T, \varepsilon_t^2(\beta) - \sigma^2)^T \tag{41}
\]
(proved in Lemma A.3 when checking C3, see Appendix 1), we have to analyse (38)–(39). We need to check some assumptions on that differential equation. We have by Lemma A.4
\[
V(\beta(t)) = E[\varepsilon^2(\beta(t))] \geq \sigma^2 > 0,
\]
and
\[
\frac{\partial V(\beta(t))}{\partial t} = \frac{\partial V(\beta(t))}{\partial \beta(t)} \frac{\partial \beta(t)}{\partial t} = -2f(\beta(t))^T \sigma^{-2}(t)F^{-1}(\beta(t))f(\beta(t)) \leq 0,
\]
since \(F^{-1}(\beta(t))\) is a symmetric positive definite matrix in \(D_R\) and \(\sigma^{-2}(t)\) is positive. Let \(V(\beta(t)) = \partial V(\beta(t))/\partial t\). Hence an invariant set of the ODE is \(E = \{ (\beta^*, \sigma^2)^T \in D_R / V(\beta) = 0 \} = \{ (\beta^*, \sigma^2)^T \in D_R / f(\beta) = 0 \} = \{ (\beta^*) \times [\delta; \delta'] \}. \) By Lemma A.5(a), there is a solution of the ODE (38)–(39) over some interval \([t_0, t_1]\). Let \(\beta(t_0) = \beta^*, \sigma^2(t_0) = \sigma^2 \in D_2\), since \(V(\beta(t)) \) is decreasing in \(t\) then, for all \(t > t_0\), \(V(\beta(t)) < V(\beta(t_0)) \leq c(\beta^*) - \varrho\), by Lemma A.6, \((\beta(t))^T, \sigma^2(t))^T \in D_2\), \(\forall t \in [t_0, t_1]\). Hence by Lemma A.5(b), \(\forall t > t_0\), \((\beta(t))^T, \sigma^2(t))^T \in D_2\). Finally, like in Ljung and Söderström (1983), \(D_2\) is a part of the attraction domain for the invariant set \(E = \{ \beta^* \} \times [\delta; \delta']\) because it fulfills the conditions of Lemma A.7. Then, we can apply Lemma A.8 which summarises Theorems 1 and 4 of Ljung (1977), by letting \(D_1 = D_R\).

**Example 4** Let the ARMA(1,1) model defined by Equation (3). Let \(\beta^* = (\phi^*, \theta^*)^T\) and assume that \(\phi^* \neq \theta^*\). We know that
\[
F^{-1}(\beta) = E \left[ \frac{\partial \varepsilon_t(\beta)}{\partial \beta} \frac{\partial \varepsilon_t(\beta)}{\partial \beta} \right]^{-1} = \frac{1 - \phi\theta}{(1 - \phi\theta)^2} \times \frac{1 - \phi^2(1 - \theta^2)}{(1 - \phi^2)(1 - \theta^2)}.
\]
The RMLMZ algorithm can be written
\[
\begin{cases}
  h_t = \begin{pmatrix} \hat{\theta}_{t-1} & 0 & 0 \\ 0 & \hat{\theta}_{t-1} & 0 \\ -1 & 0 & \hat{\theta}_{t-1} \end{pmatrix} h_{t-1} + \begin{pmatrix} 1 & \hat{\omega}_{t-1} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_t \varepsilon_t \end{pmatrix}, \\
  \left( \hat{\beta}_t, \hat{\sigma}^2_t \right) = \left( \hat{\beta}_{t-1}, \hat{\sigma}^2_{t-1} \right) + \frac{1}{t} Q(t, \hat{\beta}_{t-1}, \hat{\sigma}^2_{t-1}, h_t),
\end{cases}
\]
with \(Q(t, \hat{\beta}_{t-1}, \hat{\sigma}^2_{t-1}, h_t) = (\hat{\sigma}^2_t \psi_t F^{-1}(\hat{\beta}_{t-1}) \varepsilon_t - \hat{\sigma}^2_t)^T\) and \(h_t = (\varepsilon_t, \psi_t)^T\). Hence
\[
A(\beta) = \begin{pmatrix} \theta & 0 & 0 \\ 0 & \theta & 0 \\ -1 & 0 & \theta \end{pmatrix}, \quad \det(A(\beta) - \lambda I) = (\theta - \lambda)^3.
\]
Let \(U = \{ \beta = (\phi, \theta) \in \} = (0, 1], T\). For that model, we have
\[
D_2 = \{ (\beta^*, \sigma^2)^T \in D_R / (\beta, \sigma^2) = \} = \{ (\beta^*) \times [\delta; \delta'] \}. \]
\(D_R = \{ (\beta, \sigma^2)^T \in D_R / (\beta, \sigma^2) = [0, 1], D_R = \{ (\beta, \phi, \theta) \in U / \|F^{-1}(\beta)\| < k \}, D_R = \{ (\beta, \sigma^2)^T \in \mathbb{R}^2 / (\beta, \sigma^2) = \} = (\beta^*) \times [\delta; \delta'] \}
\]
\(\kappa = 0\) is a very small positive real number.

Let us compute \(E(\varepsilon_t^2(\beta))\). We have \(\varepsilon_t = \Theta_{t-1}^{-1}(\beta) \Phi_1(B) \varepsilon_t = \Theta_{t-1}^{-1}(\beta) \Phi_1(B) \Phi_{t-1}^{-1}(B) \Theta_{t}^{-1}(B) \varepsilon_t\). Let \(\phi(\omega)\) the spectral density of \(\varepsilon_t\). We have
\[
\phi(\omega) = \frac{1}{2\pi} |\Theta_1(e^{i\omega})|^{-2} |\Phi_1(e^{i\omega})| |\Phi_{t-1}^{-1}(e^{i\omega})|^{-2} |\Theta_{t}^{-1}(e^{i\omega})| \\
\]
hence
\[
E(\varepsilon_t^2(\beta)) = \int_{-\pi}^{\pi} \phi(\omega) d\omega \\
= \frac{1}{2\pi i} \int_{-\pi}^{\pi} \Theta_{t-1}^{-1}(z) \Theta_{t}^{-1} \left( \frac{1}{z} \right) \Phi_1(z) \Phi_{t-1}^{-1}(z) \Theta_{t}^{-1} \left( \frac{1}{z} \right) \frac{dz}{z} \\
\times \Phi_{t-1}^{-1}(z) \phi(\omega) \Theta_{t}^{-1} \left( \frac{1}{z} \right) \Theta_{t}^{-1} \left( \frac{1}{z} \right) \frac{dz}{z} \\
= \frac{1}{2\pi i} \int_{-\pi}^{\pi} \left( 1 - \phi z + \phi (1 - \theta z) (1 - \phi z) (1 - \phi z) \right) dz.
\]
For all \(\beta\) such that \(\phi = \theta\),
\[
E(\varepsilon_t^2(\beta)) = \frac{(1 - \theta^* \phi^*)(\phi^* - \theta^*)}{(1 - \phi^* \phi^*)} + \phi^*.
\]
and the Fisher information matrix is not invertible. It is obvious that when \(\theta\) comes close to 1 or \(-1, E(\varepsilon_t^2(\beta))\) converges to infinity except when \(\phi = \theta\). Let us consider the
For several ARMA(1, 1) processes characterised by \((\phi^*, \theta^*)\) values of \(\rho\) part \(D^2\) is shown where we can project \(\hat{\beta}_t\) in order to achieve convergence. For graph (b), level curves for several values of \(V(\beta)\) are shown instead. (a) \((\phi^*, \theta^*) = (-0.2, 0.7)\), (b) \((\phi^*, \theta^*) = (-0.2, 0.7)\), (c) \((\phi^*, \theta^*) = (-0.1, -0.6)\), (d) \((\phi^*, \theta^*) = (0.2, 0.7)\), (e) \((\phi^*, \theta^*) = (0.5, 0.7)\) and (f) \((\phi^*, \theta^*) = (-0.9, 0.9)\).

ODE
\[
\begin{align*}
\dot{\beta} &= \sigma^{-2}(t) F^{-1}(\beta) f(\beta) \\
\dot{\sigma}^2(t) &= V(\beta) - \sigma^2(t)
\end{align*}
\]

where
\[
f(\beta) = E[\epsilon_t(\beta) \psi_t(\beta)] \quad \text{and} \quad V(\beta) = E[\epsilon_t^2(\beta)].
\]

Let
\[
c(\phi^*, \theta^*) = \sup_{D \in K_D} \inf_{\beta \in Fr(D)} E(V(\beta)),
\]
where \(K_D\) is a set of connex parts of \(D_B\) containing \(\beta^*\).

Let \(D_2 = \{V(\beta) \leq c(\beta^*) - \varrho\}\) with a very small positive \(\varrho\). If we know a value in \(D_2\), we can use it for estimation by the algorithm of Ljung (1977, Theorem 4). In that case we have almost sure convergence.

In the cases shown in Figure 1, the part \(D_2\) where we can project the estimator to achieve convergence is the crossed surface.

As said before, the admissible region is the square \(U\) except the diagonal joining the points \((-1, -1)\) and \((1, 1)\) so it is composed of two half squares. For each of the six cases \(V(\beta)\) has a unique minimum located in one of the half squares. The part \(D_2\), shown in Figure 1 except in case (b), is always in the half square where the minimum is located. If the initial value of the ODE is in that half square, and better in \(D_2\), the solutions will turn towards that minimum when \(t\) goes to infinity, hence also the estimator \((\hat{\phi}_t, \hat{\theta}_t)\). Convergence is faster in \(D_2\). If the initial value is in the other half square, the solutions of the ODE will turn towards the frontier formed by the diagonal joining the points \((-1, -1)\) and \((1, 1)\) but will stop before reaching it. Similarly, the estimator \((\hat{\phi}_t, \hat{\theta}_t)\) will turn towards the minimum but will
have to jump over the diagonal since we may not have \( \hat{\phi}_t = \hat{\theta}_t \), because the Fisher information matrix is not invertible there. Convergence will also be slower than in the other half square and much slower than in \( D_2 \). This is well illustrated by (b) which shows the contour levels of \( V(\beta) \) for the same parameter values as (a). The level corresponding to \( D_2 \) can be seen and even smaller areas where convergence will be still faster. The other levels are higher in the upper half square and much higher in the lower half square.

Figure 1 shows that \( D_2 \) is sometimes lenticular, like in (c, d, e) but not always. Its size depends on the true values of the parameters and is smaller when they are close one from the other. Case (c) shows a situation where the point corresponding to the true values of the parameters is in the lower half square. The surface of \( D_2 \) is small when the point is close to the boundary, like in (e) and (f).

Remark 2 A possible alternative would have been to apply Chen (2002) stochastic approximation theory with expanding truncations. Indeed the assumptions made there can be verified. But Chen (2002, p. 291) assumption A6.1.3 that the norm of \( |z| \), defined by Equation (32), is bounded by a random variable with a finite variance implies here that the observations \( \{y_t\} \) are bounded by a random variable with a finite variance, ruling out again a Gaussian process. Therefore, we have preferred analysing a slightly simplified algorithm.

### 3.2. Convergence in law

Fabian (1968) has studied asymptotic normality of the algorithm

\[
\hat{\beta}_{t+1} = (I - t^{-\alpha} \Gamma_t) \hat{\beta}_t + t^{-(\alpha + \frac{1}{2})} \Phi_t V_t + t^{-\alpha - \frac{1}{2}} T_t,
\]

where \( \hat{\beta}_t = \hat{\beta}_t - \beta^* \) in our case, \( \Gamma_t \), \( \Phi_t \), and \( T_t \) are vectors, by letting conditions on the components of that algorithm. He has shown that \( t^{1/2} \hat{\beta}_t \) converges in law to the normal distribution. In our case, we let \( \alpha = \delta = 1 \), \( \Gamma_t = 0 \), \( T_t = 0 \), \( \Phi_t V_t = F^{-1}(\hat{\beta}_t) \psi_t e_t \), but one of the conditions of Fabian (1968) is that \( \Gamma_t \) is definite positive.

Ljung, Pflug, and Walk (1992) have studied a special case of that algorithm by letting \( T_t = T \), \( \alpha = 1 \) and \( \Phi_t = I \). They have shown convergence in law under other conditions. We have tried to verify the conditions of Kushner and Huang (1979) which are more general but they are not satisfied for the RML-MZ estimator. Zahaf (1999) has tried to show convergence in law of the RML-MZ estimator by using a theorem from Duflo (1997, p. 52). To achieve this goal, Zahaf (1999) makes use of an unproven conjecture which is only valid in some special cases and it is also supposed that \( F^{-1}(\beta) \psi_t(\beta) \psi_t^T(\beta) \) is (strictly) positive definite which is not true. Therefore, we have preferred to adapt the approach of Ljung and Söderström (1983).

**Theorem 3** Consider an ARMA model defined by Equation (1) and the algorithm (40), according to Assumptions 1 and 2 of Section 3.1. Then, \( \sqrt{t}(\hat{\beta}_t - \beta^*) \) converges in law to a normal distribution \( N(0, F^{-1}(\beta^*)) \) when \( t \to \infty \).

**Proof** By Theorem 2, we know that the algorithm (40), which includes a projection mechanism, leads to convergence. Hence, for \( t > t_0 \) for some finite \( t_0 \), that algorithm behaves like the algorithm (26)–(28) without projection. To simplify, we can change the origin of time at \( t_0 \) and omit the contribution of terms before \( t_0 \) which are asymptotically negligible. We know that the algorithm (26)–(28) can be written under the form (31). We have already shown in Theorem 2 that, under the same assumptions, the estimator converges almost surely to the true value of the parameter.

Let \( \varepsilon_0 = \varepsilon_0(\beta^*) = 0 \), and consider \( t \geq 1 \). Denote \( \hat{\sigma}_t = \hat{\sigma}^2_{t+1} \). Using Equations (29) and (35), we have \( \hat{\sigma}_t = \hat{\sigma}_{t-1} + e_t^2 \) and

\[
\hat{\sigma}^2_{t+1} = \frac{1}{t} \sum_{k=1}^{t} e_k^2, \quad \sigma^2_{t+1}(\beta^*) = \frac{1}{t} \sum_{k=1}^{t} e_k^2(\beta^*) = \frac{1}{t} \sum_{k=1}^{t} e_k^2(\beta^*)
\]

because \( \forall k > 0, e_k(\beta^*) = e_k \). Denote

\[
\tilde{R}_t(\beta^*) = tR_t(\beta^*) = \tilde{R}_{t-1}(\beta^*) + \psi_t(\beta^*) \psi_t^T(\beta^*),
\]

and let \( \tilde{\beta}_t = \tilde{\beta}_t - \beta^* \). From Equation (28), we have

\[
\tilde{\beta}_t = \tilde{\beta}_{t-1} + \frac{1}{t} \hat{\sigma}_t^{-2} F^{-1}(\tilde{\beta}_{t-1}) \psi_t e_t.
\]

According to Lemma A.9, \( K_t = \hat{\sigma} \hat{\beta} \) can be decomposed in a sum of terms (A3). Using that decomposition, we need Lemma A.20 to show that \( \forall \delta > 0, t^{1/2-\delta} ||\tilde{\beta}_t|| \to 0 \) a.s. when \( t \to \infty \). The proof of that Lemma A.20 makes use of Lemmas A.11–A.19. We will use that result in Lemmas A.21 and A.24. All these lemmas are in Appendix 2.

From Equations (43) and (44), we can write

\[
\tilde{R}_t(\beta^*) \tilde{\beta}_t = \tilde{R}_{t-1}(\beta^*) \tilde{\beta}_{t-1} + \frac{1}{t} \tilde{R}_t(\beta^*) \hat{\sigma}_t^{-2} F^{-1}(\tilde{\beta}_{t-1}) \psi_t e_t,
\]

\[
= \tilde{R}_{t-1}(\beta^*) \tilde{\beta}_{t-1} + \psi_t(\beta^*) \psi_t^T(\beta^*) \tilde{\beta}_{t-1} + R_t(\beta^*) \hat{\sigma}_t^{-2} F^{-1}(\tilde{\beta}_{t-1}) \psi_t e_t.
\]

But \( \psi_t e_t \) is equal to

\[
\psi_t(e_t - e_t(\tilde{\beta}_{t-1})) + (\psi_t - \psi_t(\tilde{\beta}_{t-1})) e_t(\tilde{\beta}_{t-1}) + \psi_t(\tilde{\beta}_{t-1})(e_t(\tilde{\beta}_{t-1}) - e_t) + \psi_t(\tilde{\beta}_{t-1}) e_t,
\]

and, using a Taylor expansion,

\[
e_t(\tilde{\beta}_{t-1}) - e_t = -\psi_t^T(\beta^*) \tilde{\beta}_{t-1} - \frac{1}{2} \tilde{\beta}_{t-1} \left( \frac{\partial \psi_t(\beta)}{\partial \beta T} \right)_{\beta = \beta^*} \tilde{\beta}_{t-1}.
\]

\[
(45)
\]
where $x_t$ is a point on the segment joining $\hat{\beta}_{t-1}$ and $\beta^*$. Letting $U_t = R_t(\beta^*)\hat{\sigma}_k^{-2}F^{-1}(\hat{\beta}_{t-1})$, we have
\[
\tilde{R}_t(\beta^*)\tilde{\beta}_t = \tilde{R}_{t-1}(\beta^*)\tilde{\beta}_{t-1} + \psi_t(\beta^*)\psi_t^T(\beta^*)\tilde{\beta}_{t-1} - U_t\psi_t(\beta^*)\psi_t^T(\beta^*)\tilde{\beta}_{t-1} - U_t(\psi_t(\hat{\beta}_{t-1}) - \psi_t(\beta^*))\psi_t^T(\beta^*)\tilde{\beta}_{t-1} - \frac{1}{2}U_t\psi_t(\hat{\beta}_{t-1})\tilde{\beta}_{t-1}^T \left( \frac{\partial \psi_t(\beta)}{\partial \beta^T} \right)_{\beta=\hat{x}_t} \tilde{\beta}_{t-1} + U_t\psi_t(e_t - e_t(\hat{\beta}_{t-1})) + U_t\psi_t(\hat{\beta}_{t-1})e_t.
\] We can write
\[
(I_p+u_t)\psi_t(\beta^*)\psi_t^T(\beta^*)\tilde{\beta}_{t-1} = (\hat{\sigma}_k^2F(\beta^*) - R_t(\beta^*))\hat{\sigma}_k^{-2}F^{-1}(\hat{\beta}_{t-1})\psi_t(\beta^*)\psi_t^T(\beta^*)\tilde{\beta}_{t-1} + (F(\hat{\beta}_{t-1}) - F(\beta^*))F^{-1}(\hat{\beta}_{t-1})\psi_t(\beta^*)\psi_t^T(\beta^*)\tilde{\beta}_{t-1}.
\] Letting
\[
B_{1,t} = (F(\hat{\beta}_{t-1}) - F(\beta^*))F^{-1}(\hat{\beta}_{t-1})\psi_t(\beta^*)\psi_t^T(\beta^*)\tilde{\beta}_{t-1} - U_t(\psi_t(\hat{\beta}_{t-1}) - \psi_t(\beta^*))\psi_t^T(\beta^*)\tilde{\beta}_{t-1} - \frac{1}{2}U_t(\psi_t(\hat{\beta}_{t-1})\tilde{\beta}_{t-1}^T \left( \frac{\partial \psi_t(\beta)}{\partial \beta^T} \right)_{\beta=\hat{x}_t} \tilde{\beta}_{t-1},
\]
and
\[
B_{2,t} = U_t(\psi_t - \psi_t(\hat{\beta}_{t-1}))e_t(\hat{\beta}_{t-1}) + U_t(\psi_t(e_t - e_t(\hat{\beta}_{t-1})),
\]
we have
\[
\tilde{R}_t(\beta^*)\tilde{\beta}_t = \tilde{R}_{t-1}(\beta^*)\tilde{\beta}_{t-1} + (\hat{\sigma}_k^2F(\beta^*) - R_t(\beta^*))\hat{\sigma}_k^{-2}F^{-1}(\hat{\beta}_{t-1})\psi_t(\beta^*)\psi_t^T(\beta^*)\tilde{\beta}_{t-1} + B_{1,t} + B_{2,t} + U_t\psi_t(\hat{\beta}_{t-1})e_t,
\]
where
\[
H_t = R_t^{-1}(\beta^*)\frac{1}{\sqrt{t}} \sum_{k=1}^t (\sigma_e^2F(\beta^*) - R_k(\beta^*)\hat{\sigma}_k^{-2}F^{-1}(\hat{\beta}_{k-1})\psi_k(\beta^*)\psi_k^T(\beta^*)\hat{\beta}_{k-1} + R_t^{-1}(\beta^*)\hat{\sigma}_k^{-2}F^{-1}(\hat{\beta}_{k-1})\psi_k(\beta^*)\psi_k^T(\beta^*)\hat{\beta}_{k-1} + R_t^{-1}(\beta^*)\hat{\sigma}_k^{-2}F^{-1}(\hat{\beta}_{k-1}) \times \psi_k(\beta^*)\psi_k(\beta^*)\psi_k^T(\beta^*)\hat{\beta}_{k-1} + R_t^{-1}(\beta^*) \times \frac{1}{\sqrt{t}} \sum_{k=1}^t B_{1,k} + R_t^{-1}(\beta^*) \frac{1}{\sqrt{t}} \sum_{k=1}^t B_{2,k},
\]
and
\[
L_t = R_t^{-1}(\beta^*) \frac{1}{\sqrt{t}} \sum_{k=1}^t R_k(\beta^*)\hat{\sigma}_k^{-2}F^{-1}(\hat{\beta}_{k-1}) + F^{-1}(\beta^*)\psi_k(\hat{\beta}_{k-1})e_k + R_t^{-1}(\beta^*) \frac{1}{\sqrt{t}} \times \sum_{k=1}^t R_k(\beta^*)\hat{\sigma}_k^{-2}F^{-1}(\beta^*)(\psi_k(\hat{\beta}_{k-1}) - \psi_k(\beta^*))e_k.
\]
In Lemma A.24, based on Lemma A.21, we show convergence a.s. to 0 of $H_t$ and $L_t$.

From Equation (49) and according to Lemma A.23, based on Lemmas A.21 and A.22, we have that $\sqrt{t}\hat{\beta}_t$ converges in law to a normal distribution with mean 0 and variance
\[
V = E(\psi_t(\beta^*)\psi_t^T(\beta^*))^{-1} \sigma_e^4F(\beta^*)E(\psi_t(\beta^*)\psi_t^T(\beta^*))^{-1} = \sigma_e^4F^{-1}(\beta^*)^\top\sigma_e^2F^{-1}(\beta^*) = F^{-1}(\beta^*),
\]
when $t \to \infty$ since $R_t(\beta^*)$ converges a.s. to $E(\psi_t(\beta^*)\psi_t^T(\beta^*))$ which is equal to $\sigma_e^2F(\beta^*)$ by Equation (25).

4. Finite sample properties

A computer program in Fortran 90 was written in order to experiment with the new method. It is a part of a bigger project described in Ouakasse and Mélard (2014) for general single input single output models. We will compare the results of our algorithm with Ljung (2007) System Identification Toolbox in Matlab version 7.0 (R2007a), and more specifically function RARMAX. Before that, let us discuss some implementation aspects that are essential for the interpretation.

First let us note that inspection of the RARMAX code reveals that the effective iterations are slightly different from Equation (23). At each iteration, only the elements $e_t$, $\psi_t(1)$
and \(\psi_1(p_1 + 1)\) of \(\psi_t\) are computed

\[
\psi_t(1) = \sum_{k=1}^{q} c_{tk} \psi_t(1 + k) + \eta_{t-1},
\]

(52)

\[
\psi_t(p_1 + 1) = \sum_{k=1}^{q} c_{tk} \psi_t(p_1 + 1 + k) + \epsilon_{t-1},
\]

with \(p_1 = \max(p, q)\). After that, a sliding operation is performed by

\[
\psi_{t+1}(2) = \psi_t(1), \ldots, \psi_{t+1}(p_1) = \psi_t(p_1 - 1),
\]

\[
\psi_{t+1}(p_1 + 2) = \psi_t(p_1 + 1), \ldots, \psi_{t+1}(p_1 + q)
\]

\[
= \psi_t(p_1 + q - 1).
\]

Our program follows the same sliding operations.

Like Ljung (2007), we introduce a second estimate of the forecast error \(\tilde{\epsilon}_t\), so that the algorithm (24) becomes

\[
\psi_t = \sum_{k=1}^{q} \hat{\theta}_{k,t-1} \psi_{t-k} + \tilde{\psi}_{t-1},
\]

(53)

\[
\epsilon_t = \eta_t - \hat{\beta}^T_{t-1} \tilde{\psi}_{t-1},
\]

\[
\hat{\beta}_t = \hat{\beta}_{t-1} + \gamma_t \hat{\sigma}_t^{-2} F^{-1}(\hat{\beta}_{t-1}) \psi_t \epsilon_t;
\]

(54)

\[
\hat{\sigma}_t^2 = \hat{\sigma}_t^2 + \gamma_t (\epsilon_t^2 - \hat{\sigma}_t^2),
\]

\[
\tilde{\epsilon}_t = \eta_t - \hat{\beta}^T_t \tilde{\psi}_{t-1},
\]

where this time \(\tilde{\psi}_t^T = (\psi_t, \ldots, \psi_{t-p+1}, -\tilde{\psi}_t, \ldots, -\tilde{\psi}_{t-q+1})\).

Also by code inspection, the method in RARMAX makes use of a projection using the function ‘FSTAB’ in Matlab but, as can be seen in the code, only for the parameters of the moving average polynomial \(\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q\) as follows. Let \(\hat{\theta}_t(B) = 1 - \hat{\theta}_1 B - \cdots - \hat{\theta}_q B^q\) be the polynomial estimated at time \(t\). The roots should be outside of the unit circle. Therefore, those roots which are inside of the unit circle are inverted and the others are unchanged. Then, the polynomial is computed again. Because computation of the Fisher information matrix requires that the roots of the autoregressive polynomial are also outside of the unit circle, we make use of a projection mechanism for the two polynomials. Also, we have preferred a different, smoother, projection mechanism. Let us illustrate the case of an AR polynomial. If \(\hat{\phi}_1, \ldots, \hat{\phi}_q\) is not admissible, let \(\rho < 1\), consider instead \((\rho \hat{\phi}_1, \rho \hat{\phi}_2, \ldots, \rho^q \hat{\phi}_q)\) and iterate by using successive powers of \(\rho\) until the subset of parameters becomes admissible.

Another implementation aspect is about the initial values for \(\beta_0\) and other variables subject to recurrences. Initial \(\psi_t\) and \(\epsilon_t\) are zeroes. For the RML method, an initial matrix \(R_0\) is also needed and Ljung and Söderström (1983) recommend to use \(R_0 = 10000 I\), expressing thereby a large amount of uncertainty. For our method, we have to choose \(\sigma^2_0\) instead. Clearly it should depend on the dispersion of the data. Intuitively, it is better to use a bigger initial value than needed with the hope that \(\sigma^2_t\) will converge to the (supposedly constant) innovation variance. In the present illustrations, we have taken \(\hat{\sigma}_{0}^2 = 10\) or \(\hat{\sigma}_0^2 = 1000\).

Finally, we have used a factor \(\psi_t\) in Equations (23) or (26)–(28) although this was taken as \(1/t\) in the theory. In practice it should be selected in order to improve convergence. It is often based on the forgetting factor defined by \(\lambda_t = \gamma_t/(1 - \gamma_t)\), which corresponds to \(\gamma_t = \gamma_{t-1}/(\gamma_t + \gamma_{t-1})\) and \(\lambda_t = 1\), \(\forall t\), corresponds to \(\gamma_t = 1/t\). We have followed the recommendation

\[
\lambda_t = \lambda^0 \lambda_{t-1} + (1 - \lambda^0),
\]

(55)

where typically \(\lambda^0 \approx 0.95, 0.99\), or \(1\). We have often chosen \(\lambda^0 = 1\) and \(\gamma_0 = 1\).

Since Ljung (2007) makes use of the standard notation in engineering, i.e. \(-\phi_t\) and \(-\theta_t\) instead of \(\phi_t\) and \(\theta_t\), we will report the results for \(-\hat{\phi}_t\) and \(-\hat{\theta}_t\). We will compare our estimator (solid line) with the RML estimator of Ljung as implemented in Matlab with a forgetting factor (dashed line: \(\lambda = 1\), or dot-dashed line: \(\lambda = 0.99\)) on the same Gaussian time series. They were produced in Matlab using simple recurrences and omitting the first 150 observations. They were immediately treated with the RARMAX procedure and then exported and treated by the Fortran program. For our RMLMZ algorithm, we have used a different forgetting factor for the variance, denoted with a subscript \(\sigma\), characterised by \(\lambda_\sigma^0 = 1\) and \(\gamma_{\sigma 0} = 1\). It will be noticed that, even for RML and \(t = 1000\), estimates are quite different according to \(\lambda\).

### 4.1. ARMA(1,1) model

Let us consider the ARMA(1,1) model with Equation (3) with \(\phi^* = -0.5\) and \(\theta^* = 0.5\), with \(\sigma^2 = 1\). We have generated 10,000 series of length 1000 for which we have computed the estimates of \(\phi\) and \(\theta\), for each time \(t = 1, \ldots, 1000\). The following initial values were used: \(\hat{\sigma}_0^2 = 10\), \(\hat{\phi}_0 = -0.25\), \(\hat{\theta}_0 = 0.25\), and constant forgetting factors \(\lambda_\phi = 1\), \(\lambda_{\theta 0} = 0.99\). The averages and standard deviations across the experiments are shown in function of time. For each plot, the true value of the parameter is given. It is even displayed in the plot of the averages as a dotted horizontal line. The averages should be as close as possible of the true value and the standard deviations should be close to 0.

The plots for averages indicate that the new estimator seems to converge faster than the RML estimator. On the plot for standard deviations, we observe that those of the RML estimator decrease more slowly than ours.
4.2. ARMA(2, 2) model

Let us consider the ARMA(2, 2) model with equation

\[(1 + 0.8B + 0.25B^2) y_t = (1 + 1.378B + 0.5B^2) \varepsilon_t,\]

with \(\sigma^2 = 1\). We have generated 10,000 series of length 1000 for which we have computed the estimates of \(\phi_1, \phi_2, \theta_1, \theta_2\) for each time \(t = 1, \ldots, 1000\). The following initial values were used: \(\hat{\sigma}_0^2 = 1000, \hat{\phi}_{1,0} = -0.5, \hat{\phi}_{2,0} = -0.8, \hat{\theta}_{1,0} = -0.69, \hat{\theta}_{2,0} = -0.14, \lambda_0 = 1, \gamma_0 = 1, \lambda_{\sigma} = 0.9, \gamma_{\sigma} = 1\). The results that were obtained, presented like for the previous example, are shown in Figures 4–7.

The graphs show that the averages for our method converge faster than for the RML estimator, and also that the dispersion across simulations is smaller.

4.3. AR(1) model

To consider a more critical situation, with a parameter close to the unit root, let us consider the AR(1) model with Equation (3) with \(\phi^* = 0.9\), and \(\sigma^2 = 1\). We have generated 10,000 series of length 1000 for which we have computed the estimates of \(\phi\), for each time \(t = 1, \ldots, 1000\). But, instead of using Gaussian errors, we have used a Student distribution with five degrees of freedom, with fatter tails than the normal distribution. Note that Assumption 2 is therefore fulfilled since the fourth-order moments of the process are finite. The following initial values were used: \(\hat{\sigma}_0^2 = 10, \hat{\phi}_0 = 0.25, \) with variable forgetting factors determined by \(\lambda_0 = 0.95, \lambda^0 = 0.99, \gamma_0 = 100,\) and the same for the variance. Figure 8 shows the results which are very satisfactory for the RML\textsubscript{MZ} method.

5. An example

We will illustrate the procedure on the following example. Windmills produce electricity in a way which is cleaner for the environment than with thermal or nuclear power stations. Electricity is however irregular because it depends of wind irregularity. When the wind is strong, more electricity is produced. Conversely, when the wind is weak, the quantity of electricity is very small. In order to maintain the offer of electricity at the level of demand, it is required to adapt production from traditional power stations in function of the amount of electricity produced by
a park of windmills. Response time of a power station can go from a few minutes to several hours according to the technology being used. It is therefore useful to forecast wind speed a few hours in advance. The data come from speed of wind measurements at the top of a windmill. They are available every 10 minutes, hence 144 observations per day. We have used about 12 days of measurements, more precisely 1728 observations. The data are shown in Figure 9.

We have specified an ARMA(1,2) model with a constant, described by the equation:

\[(1 - \phi_1 B)(y_t - \mu) = (1 - \theta_1 B - \theta_2 B^2)\epsilon_t.\]

Here the vector of parameters is composed of \(\beta = (\phi_1, \theta_1, \theta_2, \mu)\). A statistician or an econometrician would probably select a model with a unit root. For that reason, we have used both forgetting factors equal to 1 and an initial
The estimates at the end of the series are $\hat{\theta} = (0.925, 0.112, 0.066, 4.933)$ and the final value of the innovation variance is 0.739. Note that the exact maximum likelihood method (Mélard, 1984) of SPSS gives the following model using the whole data set:

$$(1 - 0.976B)(y_t - 4.894) = (1 - 0.216B - 0.196B^2)e_t,$$

with an estimate of the innovation variance equal to 0.400.

To illustrate the effect of having parameters which are not inside the stability region, we have restarted the on-line estimation but with an autoregressive parameter $\phi_1$ set to 1.9 (instead of 0.5 in the case of Figure 10, top left plot). The projection described at the beginning of Section 4 is performed immediately. Since we have taken $\rho = 0.5$, the initial $\phi_1$ becomes 0.95 which is inside the stability region. It can be seen that the successive values of $\phi_1$ stay in the stability region. We have obtained the following estimates at the end of the series $\hat{\theta} = (0.957, 0.014, -0.009, 4.939)$ and 0.740 for the final value of the innovation variance. Of course there is little chance that using another set of initial values will give the same estimates. It is not guaranteed even with off-line method but it is more likely to happen.
6. Conclusion

In Section 2, we have recalled the RML method proposed by Ljung (1977) and Ljung and Söderström (1983). That method provides recursive estimates using a system of equations. In one of the equations, the Hessian matrix of the error is updated. An improved RML method called RML_MZ is the subject of the present paper. It has been described in Section 3. It is based on using the Fisher information matrix, evaluated at the current value of the estimator, in order to update the estimator, instead of updating the Hessian. The asymptotic statistical properties of the new method have been studied in Sections 3.1 and 3.2. Under fairly general assumptions, it was proved that the RML_MZ estimator is consistent in the almost sure sense and also asymptotically normally distributed. This is done by following Ljung (1977) but the details are very different from those of the Ljung and Söderström (1983) approach. It is based on a result that the mathematical expectation of the errors, $E(e^2)$, has an absolute minimum obtained at the true value of the parameter. We have obtained a part of the attraction domain around that minimum of the differential equation associated to the algorithm. In Section 4 we have shown Monte Carlo simulations (for some ARMA models and using 10,000 series of length 1000) for the comparison between the RML_MZ estimator and the original RML method. This suggests that indeed the RML_MZ estimator often does converge more quickly in practice. There is however no guarantee that it will always give a better performance. In Section 5 we show an example on real data which shows the usefulness of the new method.

One important capability of recursive estimation is when dealing with a possibly time invariant process. In that case, of course, asymptotic properties are less crucial than tracking ability. According to Ljung (1985), the latter can be achieved by using a sequence $\gamma(t)$ which converges to a strictly positive value instead of 0. Alternatively $\lambda(t)$, defined in Equation (55), is chosen as a constant $\lambda_0$.

Of course the requirement to be able to compute the Fisher information matrix implies restrictions. It has been shown in Section 5 that working outside of the stability region is not really a problem provided the projection mechanism is appropriate. If there is a unit root equal to one exactly, it is of course possible to handle the differenced series, $y_t - y_{t-1}$. But, clearly the RML_MZ method will not work if there is a root equal to one exactly on parts of the series and different from one on other parts. Cases where
autoregressive and moving average roots come close are particularly challenging as discussed in Section 3.1.

Acknowledgements
This paper has benefited from an IAP-network in Statistics, contract P5/24, financially supported by the Belgian Federal Office for Scientific, Technical and Cultural Affairs (OSTC) and a FNRS grant. We thank Anestis Antoniadis, Marc Hallin, André Klein and Bas Werker, for their useful suggestions, and an anonymous referee of a previous version who suggested Chen’s approach. We also thank Tractebel which kindly provided the data used in Section 5, and Nicolas Standaert, Denis Devleeschouwer and Bogdan Petrescu who performed an off-line analysis of these data. Finally, we would like to thank the two referees on this version whose remarks have improved the paper, like recent references, remarks on computational efficiency, discussion on Solo (1981) assumptions, and examples close or outside of the stability region.

References
Akaike, H. (1973). Block Toeplitz matrix inversion. SIAM Journal of Applied Mathematics, 24, 234–241.
Åström, K. J. (1979). Maximum likelihood and prediction error methods. Automatica, 16, 457–476.
Åström, K. J., & Hågkvist, T. (1974). Uniqueness of the maximum likelihood estimates of the parameters of an ARMA Model. IEEE Transactions on Automatic Control AC, 19, 769–773.
Benveniste, A., Métivier, M., & Priouret, P. (1990). Adaptive algorithms and stochastic approximation. Berlin: Springer-Verlag.
Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). Time series analysis, forecasting and control (4th ed.). New York, NY: Wiley.
Brown, B. M. (1971). Martingale central limit theorems. Annals of Statistics, 1, 1011–1021.
Brown, R. L., Durbin, J., & Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time. Journal of the Royal Statistical Society Series B, 37, 149–163 (with discussion: 163–192).
Cartan, M. H. (1967). Calcul différentiel. Paris: Hermann.
Chen, H.-F. (2002). Stochastic approximation and its applications. Dordrecht: Kluwer.
Chung, K. L. (1968). A course in probability theory. New York, NY: Harcourt, Brace and World.
Dahlhaus, R., & Subba Rao, S. (2007). A recursive online algorithm for the estimation of time-varying ARCH parameters. Bernoulli, 13, 389–422.
Duflo, M. (1997). Random iterative models. Berlin: Springer-Verlag.
Fabian, V. (1968). On asymptotic normality in stochastic approximation. Annals of Statistics, 29, 1327–1332.
Gerencsér, L., & Prokaj, V. (2010). Recursive identification of continuous-time linear systems – an off-line approximation. In A. Edelmayer (Ed.), Proceedings of the 19th international symposium on mathematical theory of networks and systems, MTNS 2010, Budapest, Hungary, 5–9 July, pp. 375–380. Retrieved from http://www.conferences.hu/mtns2010/proceedings/start.pdf
Guo, L. (1994). Stability of recursive stochastic tracking algorithms. SIAM Journal of Control and Optimization, 32, 1195–1225.
Harville, D. A. (1997). Matrix algebra from statistician’s perspective. New York, NY: Springer.
Kalman, R. E. (1960). A new approach to general filtering and prediction problems. Transactions ASME Series D. Basic Engineering, 83, 95–108.
Kennedy, W. J. Jr., & Gentle, J. E. (1980). Statistical computing. New York, NY: Marcel Dekker.
Kirchner, H., Maggio, S., & Unser, M. (2011). A sampling theory approach for continuous ARMA identification. IEEE Transactions on Signal Processing SP, 59, 4620–4634.
Klein, A., & Mélard, G. (1989). On algorithms for computing the covariance matrix of estimates in autoregressive-moving average models. Computational Statistics Quarterly, 5, 1–9.
Klein, A., & Spreij, P. J. C. (1997). On Fisher’s information matrix of an ARMA process. In I. Csiszár & Gy. Michaletzky (Eds.), Stochastic differential and difference equations. Progress in systems and control theory (Vol. 23, pp. 273–284). Boston: Birkhäuser.
Kushner, H. J., & Huang, H. (1979). Rates of convergence for stochastic approximation type algorithms. SIAM Journal of Control and Optimization, 17, 607–617.
Kushner, H. J., & Yin, G. (1997). Stochastic approximation algorithms and applications. Berlin: Springer-Verlag.
Ljung, L. (1977). Analysis of recursive stochastic algorithms. IEEE Transactions on Automatic Control AC, 22, 551–575.
Ljung, L. (1978). Convergence analysis of parametric identification methods. IEEE Transactions on Automatic Control AC, 23, 770–783.
Ljung, L. (1985). Estimation of parameters in dynamical systems. In E. J. Hannan, P. R. Krishnaiah, & M. M. Rao (Eds.), Handbook of statistics, Vol 5: Time series in the time domain (pp. 189–211). Amsterdam: North Holland.
Ljung, L. (2007). System identification toolbox user’s guide: For use with MATLAB version 7.0 (Release 2007a). Natick, MA: The MathWorks Inc.
Ljung, L., Pflug, G., & Walk, H. (1992). Stochastic approximation and optimisation of random systems. Boston: Birkhäuser.
Ljung, L., & Söderström, T. (1983). Theory and practice of recursive identification. Cambridge, MA: MIT Press.
Löwe, M. (1977). Probability theory I (4th ed.). New York, NY: Springer.
Lukacs, E. (1975). Stochastic convergence (2nd ed.). New York, NY: Academic Press.
Marelli, D., You, K., & Fu, M. (2013). Identification of ARMA models using intermittent and quantized output observations. Automatica, 49, 360–369.
Mélard, G. (1984). Algorithm AS197: A fast algorithm for the exact likelihood of autoregressive-moving average models. Journal of the Royal Statistical Society Series C Applied Statistics, 33, 104–114.
Mélard, G. (1989). Méthode d’identification récurrente pour la modélisation des données chronologiques économiques. In Actes du Salon International de l’Informatique de la Félemtique (Séance III: Intelligence artificielle et génie logiciel, 10 pp.), Rabat, Maroc: Association Marocaine des utilisateurs de l’informatique.
Moulines, E., Priouret, P., & Roueff, F. (2005). On recursive estimation for locally stationary time varying autoregressive processes. Annals of Statistics, 33, 2610–2654.
Ouakasse, A., & Mélard, G. (2014). A new recursive estimation method for single input single output models. In preparation.
Plackett, R. L. (1950). Some theorems in least squares. Biometrika, 37, 149–157.
Pottor, M. H., & Morrey, C. B. (1977). A first course in real analysis. New York, NY: Springer-Verlag.
Rouche, N., & Mawhin, J. (1980). Ordinary differential equations. Stability and periodic solutions. London: Pitman.
Appendix 1

Here are a few lemmas needed for the proofs in Section 3.1. Invertibility of the Fisher information matrix is satisfied by Assumption 1, given the following lemma.

**Lemma A.1** (Klein & Spreij, 1997) The Fisher information matrix $F(\beta)$ is invertible if and only if the autoregressive and moving average polynomials have no common root.

**Lemma A.2** (Harville, 1997, p. 307) Let $F$ be a matrix function of $\mathbb{R}^m \to \mathbb{R}^{n \times n}$, let $x \in \mathbb{R}^m$ be a point where $F$ is invertible and continuously differentiable, then

$$
\frac{d F^{-1}(x)}{dx} = -F^{-1}(x) \left( \frac{d F(x)}{dx} \right) F^{-1}(x).
$$

**Lemma A.3** The limit (41), see Section 3.1 needed in condition C3 of Ljung does exist.

**Proof** On the basis of Equation (33), let $Q(t, \beta, \sigma^2, h_t(\beta)) = (\psi_1^T(\beta) F^{-1}(\beta) e_t(\beta), e_t^T(\beta) - \sigma^2)^T$. We have to prove that $(1/t) \sum_{t=1}^{k} \psi_4(\beta) e_t(\beta)$ converges to $E(\psi_4(\beta) e_t(\beta))$ and that $(1/t) \sum_{t=1}^{k} e_t^2(\beta)$ converges to $E(e_t^2(\beta))$. By Lemma A.11, we have that $(1/t) \sum_{t=1}^{k} \psi_4(\beta) e_t(\beta) = E(\psi_4(\beta) e_t(\beta))$ and $(1/t) \sum_{t=1}^{k} e_t^2(\beta) - E(e_t^2(\beta))$ converge a.s. to 0.

**Proof of Theorem 1** We have to prove the conditions C for the algorithm (31). Condition C2 on matrices A and B which are the same as in the RML method is of course valid. Condition C3 is proved in Lemma A.3. Of course conditions C5 and C6 are satisfied since $\gamma_t = 1/t$. There remains to check conditions C1 and C4.

Let us first check condition C1. Suppose $(\beta_1^T, \sigma_1^2)^T$ and $(\beta_2^T, \sigma_2^2)^T$ in a ball $B(0, \rho)$ with $\rho > 1$. Then for any $\ell > 0$, $\ell \in \mathbb{N}$, $\beta_t = (\beta_1, \beta_2, \sigma_1^2, \sigma_2^2)^T$ belongs to $D_t$. Let $h$ and $h'$ be two vectors in a ball $B(0, \rho)$ of $\mathbb{R}^{p+q+1}$ with $h = (h_1^{(p+q+1)})^T$ and $h' = (h'_1, ..., h'_{(p+q+1)})^T$. Let $k_1 = (h_0^{(p+q+2)}, h_1^{(p+q+2)}, ..., h_{(p+q+2)})$. By Equation (33) we have

$$
Q(t, \beta_1, \sigma_1^2, h) = Q(t, \beta_2, \sigma_2^2, h').
$$

Since $\sigma_2^2 \leq \delta, \sigma_1^2 \leq \delta - 1, (h_1 - h_0^2) \leq v$ and $(k_1 - k_0^2) \leq v$, there exists a constant $C$ such that

$$
\|Q(t, \beta_1, \sigma_1^2, h) - Q(t, \beta_2, \sigma_2^2, h')\| \leq C(\|h_1\|^2 + \|h_2\|^2 + v^2 + v)\|h - h'\| + \|F^{-1}(\beta_1) - F^{-1}(\beta_2)\| + \|\sigma_1^2 - \sigma_2^2\|.
$$

According to Lemma A.2, $\delta F^{-1}(\beta)/\delta \beta$ is continuous on the ball $B(0, \rho)$ hence it is bounded, then there exists a constant $C_1 > 0$ such that $\|F^{-1}(\beta_1) - F^{-1}(\beta_2)\| \leq C_1 \|\beta_1 - \beta_2\|$, and the preceding expression can be written

$$
\|Q(t, \beta_1, \sigma_1^2, h) - Q(t, \beta_2, \sigma_2^2, h')\| \leq C(\|h_1\|^2 + \|h_2\|^2 + v^2 + v)\|h - h'\| + C_1(\|\beta_1 - \beta_2\|^2 + \|\sigma_1^2 - \sigma_2^2\|) + \|\sigma_1^2 - \sigma_2^2\| + \|\sigma_1^2 - \sigma_2^2\| + \|h - h'\|.
$$

Solving condition C4 means to prove that, for all $\tilde{z} \in D_t$, $0 < \lambda < 1$ and $c \in \mathbb{R}$, the random variable $k_t(\tilde{z}, \tilde{\lambda}, \lambda, c)$ defined by

$$
\sum_{t=1}^{N} C(\|h_t(\beta)\|^2 + \|h_t(\beta)\|^2 + v(\tilde{z}, \lambda, c) + v(\tilde{z}, \lambda, c) + C_1(1 + v(\tilde{z}, \lambda, c))
$$

with $k_t(0, \tilde{z}, \lambda, c) = 0$ and $v(\tilde{z}, \lambda, c) = c \sum_{t=1}^{k} \lambda^{t-k} \|z_t\|$, converges to a finite limit when $t \to \infty$. In Lemma A.12, we will show that

$$
\lim_{t \to \infty} (k_t(\tilde{z}, \lambda, c)) = 0.
$$

According to Ljung and Söderström (1983) there exists a constant $C_3$ such that

$$
\sum_{t=1}^{N} C(\|h_t(\beta)\|^2 + \|h_t(\beta)\|^2 + v(\tilde{z}, \lambda, c) + v(\tilde{z}, \lambda, c) + C_1(1 + v(\tilde{z}, \lambda, c)) \leq C_3 \frac{1}{N} \sum_{t=1}^{N} (1 + \|z_t\|)^3.
$$

Hence we need only to prove that $(1/N) \sum_{t=1}^{N} (1 + \|z_t\|)^3$ converges when $N \to \infty$. Let take the norm $\|z_t\| = |\tilde{z}_t| + |\tilde{z}_{t-1}| +$: 

$$\text{Downloaded by [Archives & Bibliothèques de l'ULB], [Guy Mélard] at 02:26 24 June 2014}$$
We need to prove that $(1/N)\sum_{i=1}^{N} |y_i|^3$ is finite which is true for an ARMA process. Indeed we can write $y_i = \sum_{t=0}^{\infty} \theta_i y_{t-1}$, with $\sum_{t=0}^{\infty} |\theta_i| < \infty$, so by using Lemma A.10, part 7, $E[|y_i|^3]$ exists and ergodicity of the process $[|y_i|]$ (Taniguchi & Kazikawa, 2000) implies the result.

**Lemma A.4** (Åström & Söderström, 1974) For the ARMA($p,q$) model defined by Equation (1), $\beta^*$ is the unique solution of $E_{[\varepsilon(\beta)\psi(\beta)]} = 0$.

**Lemma A.5** (Cartan, 1967, p. 122) (a) Let $g : U \rightarrow \mathbb{R}^n : (t,x) \rightarrow g(t,x)$, where $U \subset \mathbb{R} \times \mathbb{R}^n$, be a continuously locally Lipschitz function. If $(t_0, x_0) \in \overline{U}$, an open set included in $U$, then there exists a largest interval $J$ such that $t_0 \in J$ in which there exists a unique differentiable solution $x(t) = g(t,x(t))$, with the initial condition $x(t_0) = x_0$. That function $x(t)$ is called the maximal solution. (b) Let $g(x) = a$ be a class $C^1$ function $\omega \rightarrow \mathbb{R}^n$, $\omega$ being an open set of $\mathbb{R}^n$. Let $A_0$ be a compact of $\omega$. We suppose that any solution of $\dot{x}(t) = g(x(t))$, with the initial condition $x(t_0) = x_0$, defined over $[t_0, t_1]$ is such that $\forall t \in [t_0, t_1], x(t) \in A_0$. Then the upper bound of the maximal interval of existence of the ODE is $+\infty$.

**Lemma A.6** (Rouche & Mawhin, 1980, p. 12) Consider the ODE in Lemma A.5 where $g$ is a continuously locally Lipschitz function: $g : I \times B_{\rho} \rightarrow \mathbb{R}^n$, $B_{\rho} = B(0, \rho) \subset \mathbb{R}^n$. Let $\Gamma$ be a part in $\mathbb{R}^n$ such that $\Gamma \subset B_{\rho}$. Let $V : I \times B_{\rho} \rightarrow \mathbb{R}^n$ a function of class $C^1$, and $a$, a positive constant. If

(a) $x_0 \in \Gamma$, $t_0 \in I$,
(b) $V(t_0, x_0) < a$,
(c) $\forall (t,x) \in I \times Fr(\Gamma)$, $V(t,x) \geq a$,
(d) $\forall (t,x) \in I \times \Gamma$, $\dot{V}(t,x) \leq 0$,

then the solution of the ODE is such that $\forall t \geq t_0$, $x(t) \in \Gamma$.

Consider the following ODE: $\dot{x} = g(x)$, $x(t_0) = x_0$ where $g : \Omega \rightarrow \mathbb{R}^n$ is a continuously locally Lipschitz function. Let $y^+(x_0) = \{x(\tau,x_0), \tau \geq 0\}$ be the trajectory of $x(\tau,x_0)$.

**Lemma A.7** (Rouche & Mawhin, 1980, p. 50) Let $\Psi$ a compact of $\Omega$, $\Omega$ an open set of $\mathbb{R}^n$, and $V : \Omega \rightarrow \mathbb{R}^n$ a function of class $C^1$ such that $\forall x \in \Psi$, $\dot{V}(x) \leq 0$. Let $E_{\Psi} = \{x \in \Psi / \dot{V}(x) = 0\}$ and $M$ the largest invariant subset of $E$. Then for any $x_0$ such that $y^+(x_0) \subset \Psi$, $x(t, x_0) \rightarrow M$ as $t \rightarrow \infty$.

**Lemma A.8** (Ljung, 1977, Theorems 1 and 4) Under conditions C, let us consider the algorithm (30) modified as follows:

$$h_t = \begin{cases} A(\hat{x}_t - h_{t-1} + B(\hat{x}_{t-1})\hat{z}_t) & \text{if } \hat{x}_t \in D_1 \\ \text{a point in } D_3 & \text{if } \hat{x}_{t-1} \notin D_1, \end{cases}$$

$$\hat{x}_t = \begin{cases} \hat{x}_{t-1} + \gamma Q(t,\hat{x}_{t-1},h_t) & \text{if } \hat{x}_t \in D_1, \\ \text{a point in } D_2 & \text{if } \hat{x}_t \notin D_1, \end{cases}$$

where $D_1 \subset D_R \subset \mathbb{R}^m$ is a bounded open part containing the compact $D_2$, $D_3$ is a compact of $\mathbb{R}^m$, and $m$ is the dimension of $\hat{x}_t$.

$$||x||_{D_1 \cup D_2} = \begin{cases} ||z||_{D_1} & \text{if } z \in D_1, \\ ||z||_{D_2} & \text{if } z \notin D_1. \end{cases}$$

Let $\hat{D}$ be a compact part of $D_R$ such that the trajectories of the following ODE $\dot{\hat{x}}(t)/\hat{x}(t) = f(\hat{x}(t))$ starting from a point in $D$ stay in a closed part $\hat{D}_R$ of $D_R$. Suppose that the ODE possesses an invariant set $D_A$ with its domain of attraction $D_A$ such that $D \subset D_A$. Let $\hat{D} = D_1 \setminus D_2$ and suppose there exists a twice differentiable function $U(x) \geq 0$ defined over a neighbourhood of $\hat{D}$ and such that:

$$\sup_{x \in \hat{D}} U'(x)f(x) < 0,$$

$$U(x) \geq c_1 \text{ for } x \notin D_1,$n

$$U(x) \leq c_2 < c_1 \text{ for } x \in D_2.$$n

Then $\hat{x}_t \rightarrow D_A$ almost surely when $t \rightarrow \infty$.

**Appendix 2**

**Lemma A.9** Consider $K_t = \hat{\alpha}_t \hat{\beta}_t$, defined in Section 3.2. Then

$$K_t = \sum_{k=0}^{t} A_{1,k} + \sum_{k=1}^{t} A_{2,k} + \sum_{k=2}^{t} F^{-1}(\hat{\beta}_{k-1})\psi_k e_k + T_k \hat{\beta}_{k-1}$$

$$- \sum_{k=1}^{t-1} T_k (\hat{\beta}_k - \hat{\beta}_{k-1}) - F^{-1}(\beta^*)S_k \hat{\beta}_{k-1} + F^{-1}(\beta^*)$$

$$\times \sum_{k=1}^{t-1} S_k (\hat{\beta}_k - \hat{\beta}_{k-1}) + \sum_{k=1}^{t-1} \frac{1}{k}(\sigma^2 - \sigma^2_k)F^{-1}(\hat{\beta}_{k-1})\psi_k e_k,$$

$$\text{where}$$

$$A_{1,t} = (e^2 - e^2_t)\hat{\beta}_{t-1} - F^{-1}(\hat{\beta}_{t-1})(\psi_t(\hat{\beta}_{t-1})$$

$$- \psi_t(\beta^*)\psi_t^T(\beta^*)\hat{\beta}_{t-1} - (F^{-1}(\hat{\beta}_{t-1})$$

$$- F^{-1}(\beta^*)\psi_t(\beta^*)\psi_t^T(\beta^*)\hat{\beta}_{t-1}$$

$$- \frac{1}{2}F^{-1}(\beta_{t-1})\psi_t^T(\beta^*)\psi_t F^{-1}(\hat{\beta}_{t-1})\psi_t e_k,$$

$$\hat{\beta}_t$$

is a point on the segment joining $\hat{\beta}_{t-1}$ and $\beta^*$, and

$$A_{2,t} = (e^2 - e^2_t)(\hat{\beta}_{t-1} - F^{-1}(\hat{\beta}_{t-1})\psi_t e_t - e_t(\hat{\beta}_{t-1})$$

$$- F^{-1}(\hat{\beta}_{t-1})\psi_t(\beta^*)\psi_t^T(\beta^*)\hat{\beta}_{t-1}$$

$$+ \frac{1}{t}(e^2 - \sigma^2_t)F^{-1}(\hat{\beta}_{t-1})\psi_t e_t,$$

$$\text{with}$$

$$S_k = \sum_{j=1}^{k}(\sigma^2 - \sigma^2_j)F^{-1}(\beta_{t-1})\psi_j e_j,$$

$$T_k = \sum_{j=1}^{k}(e^2 - e^2_j),$$

$$\text{with}$$

$$T_0 = 0.$$
We can use Equation (45) so that
\[ K_i = K_{i-1} + (e_i^2 - \sigma_i^2)\tilde{b}_i - 1 + (e_i^2 - \sigma_i^2)\tilde{b}_i - 1 \]
\[ + e_i^2\tilde{b}_i - 1 + F^{-1}(\tilde{b}_i - 1)\psi_i(\epsilon_i - \epsilon_i(\tilde{b}_i - 1)) \]
\[ - F^{-1}(\tilde{b}_i - 1)(\psi_i - \psi_i(\tilde{b}_i - 1))\psi_i^{T}(\beta^*)\tilde{b}_i - 1 \]
\[ - F^{-1}(\tilde{b}_i - 1)(\psi_i(\tilde{b}_i - 1) - \psi_i(\beta^*))\psi_i^{T}(\beta^*)\tilde{b}_i - 1 \]
\[ - (F^{-1}(\tilde{b}_i - 1) - F^{-1}(\beta^*))\psi_i(\beta^*)\psi_i^{T}(\beta^*)\tilde{b}_i - 1 \]
\[ - F^{-1}(\beta^*)\psi_i(\beta^*)\psi_i^{T}(\beta^*)\tilde{b}_i - 1 \]
\[ - \frac{1}{2} F^{-1}(\tilde{b}_i - 1)\psi_i^{T}(\beta^*) \frac{\partial \psi_i(\beta)}{\partial \beta^{T}} \psi_i^{T}(\beta^*) \tilde{b}_i - 1 \]
\[ + F^{-1}(\tilde{b}_i - 1)\psi_i x_i \]
\[ + \frac{1}{t}(e_i^2 - \sigma_i^2)F^{-1}(\tilde{b}_i - 1)\psi_i x_i . \]

Moving some terms leads to
\[ K_i = K_{i-1} + A_{i,1} + A_{i,2} + F^{-1}(\tilde{b}_i - 1)\psi_i x_i + (e_i^2 - \sigma_i^2)\tilde{b}_i - 1 \]
\[ - F^{-1}(\beta^*)[\psi_i(\beta^*)\psi_i^{T}(\beta^*) - \sigma_i^2 F(\beta^*)]\tilde{b}_i - 1 \]
\[ = \sum_{k=1}^{i} A_{k,1} + \sum_{k=1}^{i} A_{k,2} + \sum_{k=2}^{i} F^{-1}(\tilde{b}_k - 1)\psi_k x_k \]
\[ + \sum_{k=1}^{i} (e_k^2 - \sigma_k^2)\tilde{b}_k - 1 - F^{-1}(\beta^*) \]
\[ \times \sum_{k=1}^{i} [\psi_k(\beta^*)\psi_k^{T}(\beta^*) - \sigma_k^2 F(\beta^*)]\tilde{b}_k - 1 \]
\[ + \sum_{k=1}^{i} \frac{1}{t}(e_k^2 - \sigma_k^2)F^{-1}(\tilde{b}_k - 1)\psi_k x_k . \]

since \( K_0 = 0 \). Introducing \( S_i \), defined by Equation (A6), yields after some algebra
\[ \sum_{k=1}^{i} [\psi_k(\beta^*)\psi_k^{T}(\beta^*) - \sigma_k^2 F(\beta^*)]\tilde{b}_k - 1 \]
\[ = \sum_{k=1}^{i} (S_k - S_{k-1})\tilde{b}_k - 1 = S_i\tilde{b}_i - 1 - \sum_{k=1}^{i-1} S_k(\tilde{b}_k - \tilde{b}_{k-1}) . \]

Similarly using \( T_k \) defined by Equation (A7),
\[ \sum_{k=1}^{i} (e_k^2 - \sigma_k^2)\tilde{b}_k - 1 = \sum_{k=1}^{i} (T_k - T_{k-1})\tilde{b}_k - 1 \]
\[ = T_i\tilde{b}_i - 1 - \sum_{k=1}^{i-1} T_k(\tilde{b}_k - \tilde{b}_{k-1}) . \quad (A8) \]

We need several classical lemmas which are collected here for convenience, see e.g. Loeve (1977, pp. 250–251), Lukacs (1975, p. 80) and Chung (1968, p. 307), plus two other lemmas.

**Lemma A.10**

1. **(Kolmogorov’s proposition)** Let \( x_n \) be independent random variables, \( b_n \not\sim \infty \), and \( S_n = \sum_{k=1}^{n} x_k \). Suppose that \( \sum_{k=1}^{n} \text{Var}(x_k)/b_k^2 < \infty \), then \( (S_n - E(S_n))/b_n \to 0 \) a.s. when \( n \to \infty \).

2. **(Kronecker)** Let \( x_k \) be a sequence of real numbers, \( a_k \) a sequence of positive numbers which converges to \( \infty \). Then \( \sum_{k=1}^{\infty} (x_k/a_k) < \infty \) implies \((1/a_k) \sum_{k=1}^{\infty} x_k \to 0 \) when \( n \to \infty \).

3. **(Toepplitz)** Let \( a_{nk}, k = 1, 2, \ldots \), be numbers such that for every fixed \( k, a_{nk} \to 0 \) when \( n \to \infty \) and, for all \( n, \sum_{k=1}^{\infty} |a_{nk}| \leq c < \infty \). Let \( x_k \) be a sequence of real numbers. If \( x_k < \infty \), then \( \sum_{k=1}^{n} a_{nk} \to \infty \) when \( n \to \infty \).

4. **(Lukacs)** Let \( \{x_k\} \) be a sequence of random variables with finite expectations and suppose that \( \sum_{k=1}^{\infty} E(|x_k|) \) is finite. Then the infinite series \( \sum_{n=1}^{\infty} |x_n| \) is a.s. convergent.

5. **(Chung)** Let \( \{x_n\} \) be a submartingale satisfying the condition \( \lim_{n \to \infty} E(\max(x_n, 0)) < \infty \). Then \( \{x_n\} \) converges a.s. to a finite limit when \( n \to \infty \).

6. Let \( \{x_t, t \in Z\} \) and \( \{z_t, t \in Z\} \) be ARMA(\( p, q \)) and ARMA(\( p_2, q_2 \)) stationary processes, respectively, in terms of the same errors. Then, for all \( \varepsilon > 0 \) and all \( s \in Z \),
\[ t^{1/2-\varepsilon}((1/t) \sum_{k=1}^{t} x_k z_{k-s} - E(x_k z_{k-s})) \to 0 \] when \( n \to \infty \).

7. Let \( \{x_t, t \in Z\}, i = 1, 2, 3, 4 \), be four processes defined in terms of a sequence of i.i.d. random variables \( \{\varepsilon_t, t \in Z\} \), with finite fourth-order moment, by \( x_t(i) = \sum_{k=1}^{i} \varepsilon_{t-k} |x_{t-k}| \) with \( \sum_{k=i}^{\infty} |x_{t-k}| < \infty \). Then \( E(x_t(i)^4) \to \infty \) is finite.

**Proof** To prove part 6, let
\[ x_k = \phi_1 x_{k-1} - \cdots - \phi_{p_1} x_{k-p_1} \]
\[ + e_k - \theta_1 e_{k-1} - \cdots - \theta_{q_1} e_{k-q_1} \]
and
\[ z_k = \phi_2 z_{k-1} - \cdots - \phi_{p_2} z_{k-p_2} \]
\[ + e_k - \theta_2 e_{k-1} - \cdots - \theta_{q_2} e_{k-q_2} . \]

We have
\[ x_k z_{k-s-i} - \phi_1 x_{k-1} z_{k-s-i} - \cdots - \phi_{p_1} x_{k-p_1} z_{k-s-i} \]
\[ = (e_k - \theta_1 e_{k-1} - \cdots - \theta_{q_1} e_{k-q_1}) z_{k-s-i} . \quad (A9) \]

with \( i = 0, 1, \ldots, p_2 - 1 \) and
\[ z_k x_{k+s-j} - \phi_2 z_{k-1} x_{k+s-j} - \cdots - \phi_{p_2} z_{k-p_2} x_{k+s-j} \]
\[ = (e_k - \theta_2 e_{k-1} - \cdots - \theta_{q_2} e_{k-q_2}) z_{k+s-j} . \quad (A10) \]

with \( j = 1, \ldots, p_1 \). Denote \( Y_s = \lim_{t \to \infty} t^{-1/2-\varepsilon} \sum_{k=1}^{t} (x_k z_{k-s} - E(x_k z_{k-s})) \) hence the equation of (A9) for \( i = 0 \) yields
\[ Y_s = \phi_1 Y_{s-1} - \cdots - \phi_{p_1} Y_{s-p_1} \]
\[ = \lim_{t \to \infty} t^{-1/2-\varepsilon} \sum_{k=1}^{t} \{e_k - \theta_1 e_{k-1} - \cdots - \theta_{q_1} e_{k-q_1}) z_{k-s} - E((e_k - \theta_1 e_{k-1} - \cdots - \theta_{q_1} e_{k-q_1}) z_{k-s})) \} . \quad (A11) \]

Writing \( z_{k-s} \) as function of \( e_k - \cdots - e_{k-p_2} - z_{k-q_1} - \cdots - z_{k-q_2-p_2} \), we have in the right-hand side a finite sum of expressions
of the form

$$\lim_{t \to \infty} t^{-1/2-\varepsilon} \left( \sum_{k=1}^{t} (e_k z_{k-h} - E(e_k z_{k-h})) \right)$$

$$= \lim_{t \to \infty} t^{-1/2-\varepsilon} \sum_{k=1}^{t} e_k z_{k-h}$$

(A12)

for $h > 0$, or similar expressions with $e_k e_{k-h}$ instead of $e_k z_{k-h}$, and also an expression with $e_k^2$. The latter provides a sum of independent variables

$$\lim_{t \to \infty} t^{-1/2-\varepsilon} \sum_{k=1}^{t} (e_k^2 - E(e_k^2)).$$

Since $E(e_k^2)$ exists, this limit is 0 by applying Lemma A.10 (Kolmogorov’s proposition). Let us take for example (A12) and prove that it equals 0. Consider $r_t = \sum_{k=1}^{t} k^{-1/2-\varepsilon} e_k z_{k-h}$ and the $\sigma$-algebra $\mathcal{F}_{t-1}$, spanned by the $s_i$, $i \leq t-1$. For $h > 0$, we have

$$E(r_t | \mathcal{F}_{t-1}) = E\left( \sum_{k=1}^{t} k^{-1/2-\varepsilon} e_k z_{k-h} \bigg| \mathcal{F}_{t-1} \right) = r_{t-1},$$

hence $\{r_t\}$ is a martingale with respect to $\mathcal{F}_{t-1}$. Moreover

$$E[r_t]^2 \leq \sum_{k=1}^{t} k^{-2-2\varepsilon} E[z_{k-h}]^2 E[e_k^2] \leq C \sum_{k=1}^{t} k^{-2-2\varepsilon} < \infty,$$

where $C$ denotes a constant. Hence $r_t$ is a martingale with a bounded variance, and according to Lemma A.10 (Chung, 1968), $r_t$ converges a.s. to a finite limit $r_\infty$. Hence by Lemma A.10 (Kronecker’s lemma), $t^{-1/2-\varepsilon} \sum_{k=1}^{t} e_k z_{k-h} \to 0$ a.s. when $t \to \infty$, for $h > 0$, with a similar reasoning for terms in $e_k e_{k-h}$. Hence the right-hand side of Equation (A11) converges to 0. Proceeding in the same way for all the equations of (A9) and (A10), and denoting the $(p_2 + p_1) \times (p_2 + p_1)$ matrix

$$A = \begin{pmatrix}
1 & -\phi_{1,1} & \ldots & -\phi_{1,p_1} & \ldots & 0 \\
0 & 1 & -\phi_{1,1} & \ldots & -\phi_{1,p_1} & \ldots & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & 0 & 1 & -\phi_{1,1} & \ldots & -\phi_{1,p_1} \\
-\phi_{2,p_2} & -\phi_{2,p_2} - 1 & \ldots & 1 & 0 & \ldots & 0 \\
0 & -\phi_{2,p_2} & -\phi_{2,p_2} - 1 & \ldots & 1 & \ldots & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & 0 & 0 & -\phi_{2,p_2} & -\phi_{2,p_2} - 1 & 1 \\
\end{pmatrix},$$

and

$$X' = (\gamma_{p+p_2-1} \ldots \gamma_t \ldots \gamma_{p-1}),$$

we obtain

$$AX = 0.$$

From van der Waerden (1937), $A$ is invertible because the zeroes of the polynomials $1 - \phi_{1,1} B - \ldots - \phi_{1,p_1} B^{p_1}$ and $1 - \phi_{2,1} B - \ldots - \phi_{2,p_2} B^{p_2}$ are outside the unit circle so the polynomials $1 - \phi_{1,1} B - \ldots - \phi_{1,p_1} B^{p_1}$ and $B^{p_2} - \phi_{2,1} B^{p_2-1} - \ldots - \phi_{2,p_2} B^{p_2}$ have no common zeroes. Hence $X = 0$, which implies the result.

To prove part 7, we write

$$E[X_i^{(1)} X_i^{(2)} X_i^{(3)} X_i^{(4)}] = \sum_{v=1}^{\infty} \sum_{v_1}^{\infty} \sum_{v_2}^{\infty} \sum_{v_3}^{\infty} E[(\varepsilon_t - \varepsilon)^4]$$

$$+ \sum_{v_1 \neq v_2 \neq v_3} \sum_{v_1}^{\infty} \sum_{v_2}^{\infty} \sum_{v_3}^{\infty} E[(\varepsilon_t - \varepsilon)^4]$$

$$+ \sum_{v_1 \neq v_2 \neq v_3} \sum_{v_1}^{\infty} \sum_{v_2}^{\infty} \sum_{v_3}^{\infty} E[(\varepsilon_t - \varepsilon)^4] E[(\varepsilon_t - \varepsilon)^4]$$

$$+ \sum_{v_1 \neq v_2 \neq v_3} \sum_{v_1}^{\infty} \sum_{v_2}^{\infty} \sum_{v_3}^{\infty} E[(\varepsilon_t - \varepsilon)^4] E[(\varepsilon_t - \varepsilon)^4]$$

$$X_i^{(1)} E[(\varepsilon_t - \varepsilon)^4]$$

$$+ \sum_{v_1 \neq v_2 \neq v_3} \sum_{v_1}^{\infty} \sum_{v_2}^{\infty} \sum_{v_3}^{\infty} E[(\varepsilon_t - \varepsilon)^4]$$

$$X_i^{(1)} E[(\varepsilon_t - \varepsilon)^4]$$

Denoting $m_j = E[(\varepsilon_t - \varepsilon)^4]$ (the absolute value of the left-hand side is bounded by the maximum of $m_1, m_2, m_3$ and $m_4$ which is finite, multiplied by the $\sum_{v=1}^{\infty} \sum_{v_1}^{\infty} \sum_{v_2}^{\infty} \sum_{v_3}^{\infty} \sum_{v_4}^{\infty}$, which is also finite).

**Lemma A.11** Let $\beta \in \mathcal{D}_B$ and consider $R_1(\beta)$ defined by Equation (37). Then for all $\varepsilon > 0$, $t^{1/2-\varepsilon} R_1(\beta) - \alpha^2 t\beta^2$ converges to 0 a.s. when $t \to \infty$, $t^{1/2-\varepsilon} (1/t) \sum_{s=1}^{t} \psi_s(\beta)$ $\varepsilon_s(\beta) - E[\psi_s(\beta) \varepsilon_s(\beta)]$ and $t^{1/2-\varepsilon} (1/t) \sum_{s=1}^{t} \sigma_s(\beta)$ $\varepsilon_s(\beta) - E[\sigma_s(\beta) \varepsilon_s(\beta)]$, where $\alpha^2(\beta)$ is defined by Equation (35). Similarly, considering $S_t$ and $T_t$ defined by Equations (A6) and (A7), respectively, we have that, for all $\varepsilon > 0$, $t^{-1/2-\varepsilon} S_t$ and $t^{-1/2-\varepsilon} T_t$ converge to 0 a.s. when $t \to \infty$.

**Proof** Let $\psi_1(\beta) = (\psi_1, \ldots, \psi_{p+q})(\beta)^T$. From Equation (36) we have $\Theta(B) \psi_1(\beta) = \phi_{1,1}(\beta)$, hence

$$\psi_{i,1}(\beta) = \Theta(B)^{-1} \varepsilon_i, \quad i = 1, \ldots, p,$$

$$\psi_{p+i,1}(\beta) = -\Theta(B)^{-1} \varepsilon_{i-1}, \quad i = 1, \ldots, q,$$

but, since $\Phi^*(B) \psi_i = \Theta^*(B) \varepsilon_i$, we know that $\varepsilon_i = -\Phi^*(B)^{-1} \Theta^*(B) \varepsilon_i$, and that $\varepsilon_{i-1} = \Theta(B)^{-1} \Phi(B) \Phi^*(B)^{-1} \Theta^*(B) \varepsilon_i$, hence

$$\psi_{i,1}(\beta) = \Theta(B)^{-1} \Phi(B)^{-1} \Theta^*(B) \varepsilon_i, \quad i = 1, \ldots, p,$$

$$\psi_{p+i,1}(\beta) = -\Theta(B)^{-1} \Phi(B)^{-1} \Theta^*(B) \varepsilon_i, \quad i = 1, \ldots, q.$$**Consequently** the $\varepsilon_i(\beta)$ and $\{\psi_{i,1}(\beta), i = 1, 2, \ldots, p\}$ are ARMA processes in terms of the errors $\varepsilon_i$ of the original process. The proof is completed by using three times Lemma A.10, part 6. The last part of the lemma is immediate given the definition of $S_t$ and $T_t$ with $\beta \in \mathcal{D}_B$ replaced by $\beta^p$.
Lemma A.12. Let us consider the algorithm (31). First, \( \|h_t\| \), \( \|h_t(\beta)\|_{\beta^*} \), and \( \|\partial h_t(\beta)\|_{\beta^*T} \), where \( \beta = \gamma_t \) is sufficiently close to \( \beta^* \), are bounded by a process \( \sum_{i=1}^{\infty} |v_i| |v_i - 1| \) where \( \sum_{i=1}^{\infty} |v_i| \) is finite. Then, for all random variable \( h_t \) such that, \( h_t = \sum_{i=1}^{\infty} \xi_i |e_{t-i}| \) with \( \sum_{i=1}^{\infty} |\xi_i| < \infty \), there exists a positive constant \( M \) such that \( E(\|h_t - h_t(\beta)\|) \leq M/t \) where \( \forall \beta \in \mathcal{D}_B \), \( h_t(\beta) = A(\beta) h_{t-1}(\beta) + B(\beta) z_t \). Moreover \( h_t - h_t(\beta^*) \to 0 \) a.s. when \( n \to \infty \).

Proof. From Equation (1) we have the infinite moving average representation

\[
y_t = \sum_{i=1}^{\infty} \mu_i e_{t-i},
\]

where \( \sum_{i=1}^{\infty} |\mu_i| < \infty \). \( \forall \beta \in \mathcal{D}_B \) we may write \( h_t(\beta) \) under the form

\[
h_t(\beta) = \sum_{i=0}^{t-1} A(\beta)' B(\beta) z_{t-i}(\beta) + A(\beta)' h_0(\beta),
\]

where \( v = q(p + q + 1) \). Let

\[
G_t(\beta) = (h_{t-1}(\beta) \otimes I_v) \frac{\partial \text{vec}(\hat{A}(\beta))}{\partial \beta^T} + (z_t \otimes I_v) \frac{\partial \text{vec}(B(\beta))}{\partial \beta^T}.
\]

We can write \( \partial h_t(\beta)/\partial \beta^T \) under the form

\[
\frac{\partial h_t(\beta)}{\partial \beta^T} = \sum_{i=0}^{\infty} A(\beta)' G_{t-i}(\beta).
\]

We know that \( \beta^* \in \mathcal{D}_B \), so \( \|A(\beta)'\| \leq C \lambda^T \) for some \( \lambda < 1 \). Furthermore, for \( \beta_t \beta \) belonging to a neighbourhood of \( \beta^* \), we have also \( \|\prod_{k=1}^{t} A(\beta_k)\| \leq C \lambda^T \) for some \( \lambda < 1 \) since \( \prod_{k=1}^{t} A(\beta) \) is a continuous function of \( \beta \). In Section 3.1, we have proved that \( \beta_t \to \beta^* \) a.s. when \( t \to \infty \), then for a large enough \( t, 3T > 0 \), such that \( \forall \beta, T > \lambda, \|A(\beta(t))\| < \lambda_1 \), so \( \forall \beta, T \),

\[
\left| \prod_{k=1}^{t} A(\hat{\beta}_k) \right| \leq \prod_{k=1}^{t-1} A(\hat{\beta}_k) \left| \prod_{k=1}^{t} A(\hat{\beta}_k) \right| \leq C_0 \lambda_1^{t-1} = C_2 \lambda^T,
\]

(17)

where \( C_0 \) can be taken as \( (C_1)^T \), for example, where \( C_1 = \sup_{\beta \in \mathcal{D}_B} \|\hat{A}(\beta)\| \).

According to Equation (31) we have

\[
h_t = A(\hat{\beta}_{t-1}) h_{t-1} + B(\hat{\beta}_{t-1}) z_t.
\]

Since \( h_t \) contains \( e_t \) and \( \psi_t, \psi_{t-1}, \ldots, \psi_{t-q} \), we can suppose that \( h_0 = 0 \) hence

\[
h_t = A(\hat{\beta}_{t-1}) A(\hat{\beta}_{t-2}) h_{t-2} + A(\hat{\beta}_{t-1}) B(\hat{\beta}_{t-2}) z_{t-1}
+ B(\hat{\beta}_{t-2}) z_t = \ldots
\]

\[
= \sum_{k=1}^{t} \left[ \prod_{j=k}^{t-1} A(\hat{\beta}_j) \right] B(\hat{\beta}_{t-k}) z_k + \left[ \prod_{j=k}^{t} A(\hat{\beta}_j) \right] h_0,
\]

(18)

with the convention \( \prod_{j=1}^{0} A(\hat{\beta}_j) = I_p \). Since \( \forall \beta \in \mathcal{D}_B \), there exists a positive real \( C \) such that \( \|B(\beta)\| < C \), so Equations (17) and (18) imply

\[
\|h_t\| \leq C \sum_{k=1}^{t} \lambda_1^{t-k} \|z_k\|.
\]

(19)

To conclude the proof of the first statement, let us first show that \( \|h_t\| \) is bounded by some process. Indeed, we deduce from Equation (A19) that

\[
\|h_t\| \leq C \sum_{k=0}^{t-1} \lambda_1^{t-k} \|z_{t-k}\| \leq C \sum_{k=0}^{\infty} \lambda_1^k \sum_{i=1}^{\infty} |\alpha_i| |e_{t-k-i}|,
\]

where the coefficients \( \alpha_i \) are related to the \( \mu_i \) in Equation (A13).

Hence \( \|h_t\| \) is dominated by a process such as \( \sum_{i=1}^{\infty} |v_i| |e_{t-i}| \) where \( \sum_{i=1}^{\infty} |v_i| \) is finite. The derivatives are similar for \( \|h_t(\beta)\| \) and \( \|\partial h_t(\beta)/\partial \beta^T\| \), where \( \beta = \gamma_t \) is sufficiently close to \( \beta^* \), using Equations (A14) and (A16), respectively, with appropriate \( \lambda_1, \alpha_t \) and \( v_t \).

For the second part of the proof, let \( \hat{h}_t(\beta) = h_t - h_t(\beta), \hat{A}(\hat{\beta}_k) = A(\hat{\beta}_k) - A(\beta), \hat{B}(\hat{\beta}_k) = B(\hat{\beta}_k) - B(\beta) \). For \( k \leq t \), we have similarly to Equation (A18)

\[
\hat{h}_t(\beta) = \sum_{k=1}^{t} \left( \prod_{j=k}^{t} A(\hat{\beta}_j) \right) \left[ \hat{A}(\hat{\beta}_{t-k-1}) h_{t-1}(\hat{\beta}_{t-1}) + \hat{B}(\hat{\beta}_{t-k-1}) z_k \right],
\]

(20)

thus

\[
\|\hat{h}_t(\hat{\beta}_{t-1})\| \leq \sum_{k=1}^{t} \lambda_1^{t-k} \|\hat{A}(\hat{\beta}_{t-k}) h_{t-1}(\hat{\beta}_{t-1}) + \hat{B}(\hat{\beta}_{t-k}) z_k\|.
\]

Since \( \hat{A}(\beta) \) and \( B(\beta) \) are Lipschitz continuous, there exists a constant \( C_{AB} \) such that

\[
\|\hat{A}(\hat{\beta}_{t-k-1})\| \leq C_{AB} \|\hat{\beta}_k - \hat{\beta}_{t-1}\| + \|\hat{B}(\hat{\beta}_{t-k-1})\| \leq C_{AB} \|\hat{\beta}_k - \hat{\beta}_{t-1}\|.
\]

We have

\[
t \|\hat{h}_t(\hat{\beta}_{t-1})\| \leq t C_{AB} \sum_{k=1}^{t} \lambda_1^{t-k} \|\hat{\beta}_k - \hat{\beta}_{t-1}\|
+ t \|\hat{\beta}_k - \hat{\beta}_{t-1}\| \leq C_{AB} \|\hat{\beta}_k - \hat{\beta}_{t-1}\| + \|\hat{\beta}_k - \hat{\beta}_{t-1}\|.
\]

Let us prove that, for all random variable \( h_t \) such that \( h_t = \sum_{i=1}^{\infty} \xi_i |e_{t-i}| \) with \( \sum_{i=1}^{\infty} |\xi_i| < \infty \), \( tE(\|h(t)\|) \) is
bounded. Indeed, by using two times Lemma A.10, part 7,
\[ tE(\|\hat{h}_t(\hat{b}_{1-})\|_1) \leq Ct \sum_{l=1}^{t} \frac{h_{l-1}^2}{l^2} \sum_{k=1}^{l} \lambda_{l-k} \]
\[ \times E[b_t \|\psi_T\|_1 \|\hat{h}_{k-1}(\hat{b}_{1-})\| + b_t \|\hat{\psi}_T\|_1 \|\epsilon_2\|_1] \]
\[ \leq Ct \sum_{l=1}^{t} \frac{h_{l-1}^2}{l^2} \sum_{k=1}^{l} \lambda_{l-k} \]
\[ = Ct \sum_{l=1}^{t} \lambda_{l-k} - \lambda_1 \]
\[ = \frac{\lambda_1}{1 - \lambda_1} C \left( \sum_{l=1}^{t} \lambda_{l-k} - \sum_{l=1}^{t} \lambda_{l-1} \right) \]
which is bounded.

For the third part of the proof, let \( \tilde{h}_t(\beta^*) = h_k - h_k(\beta^*), \tilde{A}(\hat{p}_k, \beta^*) = A(\hat{p}_k) - A(\beta^*), \tilde{B}(\hat{p}_k, \beta^*) = B(\hat{p}_k) - B(\beta^*). \) We have
\[ \tilde{h}_t = \sum_{k=1}^{t} \left( \prod_{j=0}^{k-1} \tilde{A} \right) [\tilde{A}(\hat{p}_{k-1}, \beta^*) h_k(\beta^*) + \tilde{B}(\hat{p}_{k-1}, \beta^*) \epsilon_2]. \]
Because \( \tilde{A}(\hat{p}_{k-1}, \beta^*), \tilde{B}(\hat{p}_{k-1}, \beta^*) \) converge to 0, hence \( \tilde{h}_t \to 0 \) a.s., when \( n \to \infty. \)

**Lemma A.13** There exists a process \( a_t \) such that \( (1/t) \sum_{k=1}^{t} a_k \) is bounded and \( \forall t \geq 1, \|A_{1,t}\| < a_t \|\hat{b}_{1-}\|^2 \), where \( A_{1,t} \) is defined by (A4).

**Proof** We know that \( \hat{b}_t \to \beta^* \) a.s., and \( F^{-1}(\beta) \) being continuous over \( D_B, F^{-1}(\hat{b}_{1-}) \to F^{-1}(\beta^*) \) a.s., so there exists a positive constant \( C_1 \) such that
\[ \frac{1}{2} F^{-1}(\hat{b}_{1-}) \psi(T_{\beta}) \sum_{k=1}^{t} \frac{\partial \psi(\beta)}{\partial \beta^T} \beta_{x_1} \]
\[ < C_1 \|\psi_T\|_1 \|\tilde{h}_{1-}\|^2 \]
where \( x_1 \) is a point on the segment joining \( \hat{b}_{1-} \) and \( \beta^* \) near to \( \beta^* \). Also, \( \partial F(\beta)/\partial \beta \) and \( F^{-1}(\beta) \) are bounded over \( D_B \) so that, by Lemma A.2, \( \partial F^{-1}(\beta)/\partial \beta \) is bounded, there exist positive constants \( C_2 \) and \( C_3 \) such that
\[ \|F^{-1}(\hat{b}_{1-}) - F^{-1}(\beta^*)\|_1 \psi(T_{\beta}) \beta_{x_1} \]
\[ < C_2 \|\psi_T(\beta^*)\|^2 \|\tilde{h}_{1-}\|^2 \]
and
\[ \|F^{-1}(\hat{b}_{1-}) \psi(T_{\beta}) - F^{-1}(\beta^*) \beta_{x_1} \]
\[ < C_3 \|\psi_T(\beta^*)\|^2 \|\tilde{h}_{1-}\|^2. \]
and there exists a constant \( C_4 \) such that
\[ \|\varepsilon_2^2(\hat{b}_{1-}) - \varepsilon_2^2(\beta^*)\|_1 \]
\[ < C_4 \|\psi_T(x_1)\|_1 \|\tilde{h}_{1-}\|^2. \]
We have to prove that
\[ \frac{1}{t^{1-\theta}} \sum_{k=1}^{t} \left( \|\psi_T\|_1 \left\| \frac{\partial \psi_T(\beta)}{\partial \beta^T} \right\|_{\beta=x_1} \right) + \|\psi_T(\beta^*)\|^2 + \|\tilde{h}_{1-}\|^2 \]
\[ \times \left( \left\| \frac{\partial \psi_T(\beta)}{\partial \beta^T} \right\|_{\beta=x_1} \right) \] (A21)
is finite. Since \( \psi_T \), both \( \psi_T(\beta) \) and \( \varepsilon_1(\beta) \), and \( \partial \psi_T(\beta)/\partial \beta^T \) are components of \( h_k, h_k(\beta) \), and \( \partial h_k(\beta)/\partial \beta^T \), respectively, from Lemma A.12, since \( x_1 \) is near \( \beta^* \), we have for each term of (A21) an upper bound of the form \( \sum_{j=1}^{\infty} |n|_{\|\epsilon_2\|_1} \sum_{j=1}^{\infty} |\eta_j|_{\|\epsilon_2\|_1} \) which is an ergodic process, hence we have the result.

**Lemma A.14** For any \( \vartheta < 1 \), and any \( a_t \) such that \( (1/t) \sum_{k=1}^{t} a_k \) is finite, \( t^{-1+\vartheta} \sum_{k=1}^{t} k^{-\vartheta} a_k \) is finite. If \( \vartheta > 1 \) then \( \sum_{k=1}^{t} k^{-\vartheta} a_k \) is finite.

**Proof** We have
\[ \frac{1}{t^{1-\vartheta}} \sum_{k=1}^{t} k^{-\vartheta} a_k \]
\[ = \frac{1}{t^{1-\vartheta}} \sum_{k=1}^{t-1} a_k - \frac{1}{t^{1-\vartheta}} \sum_{k=1}^{t-1} a_k \]
\[ \times \sum_{k=1}^{t-1} |(k+1)^{-\vartheta} - k^{-\vartheta}| \sum_{l=1}^{k} a_l \]
\[ \leq \frac{1}{t} \sum_{k=1}^{t-1} a_k + \frac{1}{t^{1-\vartheta}} \sum_{k=1}^{t-1} \sum_{l=1}^{k} (k^{-\vartheta} - (k+1)^{-\vartheta}) \]
\[ \times \sum_{k=1}^{t} a_k \]
Let us prove that \( t^{-1+\vartheta} \sum_{k=1}^{t} k^{-\vartheta} a_k \) is finite when \( t \to \infty \) in order to apply Toeplitz lemma (see Lemma A.10). Denote
\[ \Sigma_t \]
\[ = \sum_{k=1}^{t} k^{-\vartheta} - (k+1)^{-\vartheta} \]
and
\[ \Sigma(x) = x(x^{-\vartheta} - (x+1)^{-\vartheta}). \]
Since \( \Sigma(x) \) is a decreasing function of \( x \), we have by using Cauchy theorem on comparisons between series and integrals (e.g. Protter & Morrey, 1977)
\[ \int_{t}^{t+\vartheta} \Sigma(x) \]
\[ \leq \frac{1}{t^{1-\vartheta}} \int_{t}^{t+\vartheta} \Sigma(x) \]
and, integrating by parts and rearranging terms,
\[ \int_{t}^{t+\vartheta} \Sigma(x) \]
\[ = \frac{1}{1-\vartheta} t \left( t^{1-\vartheta} \left( 1 - (1 + \frac{1}{1-\vartheta}) \right) \right) \]
\[ \times \left( 1 - (1 - 2^{1-\vartheta}) - (1 - 2^{-1-\vartheta}) \right) \]
\[ = \frac{1}{t^{1-\vartheta}} \left( 1 - (1 + \frac{1}{1-\vartheta}) \right) \]
\[ \times \left( 2^{1-\vartheta} - (1 - 1) \right) \]
\[ \leq \left( 1 - (2^{1-\vartheta}) \right) \].
Hence \( \Sigma_t/t^{1-\vartheta} \) is bounded by above by the integral divided by \( t^{1-\vartheta} \) which is finite. If \( \vartheta > 1 \), we will have \( \int_{t}^{t+\vartheta} \Sigma(x) \) dx < \( \infty. \)
LEMMA A.15  Consider $A_{2,k}$ defined by Equation (A5), $\forall \delta > 0$, $t^{-1/2-\delta} \sum_{k=1}^{t} A_{2,k}$ and $t^{-1/2-\delta} \sum_{k=1}^{t} (1/k) (e_{k}^{2} - \hat{\sigma}_{k}^{2}) F^{-1}(\hat{\beta}_{k-1}) \psi_{k} e_{k}$, converge to 0 a.s. when $t \to \infty$.

Proof  A sketch of the proof is given by Ljung and Söderström (1983, p. 444), Lemma 4.8.4 but the proof given here differs because of Assumption 2. Let us consider the series $Z_{t} = \sum_{k=1}^{t} k^{-1/2-\delta} A_{2,k}$ obtained by

$$Z_{t} = \sum_{k=1}^{t} \left( k^{-1/2-\delta} (e_{k} - e_{k} (\hat{\beta}_{k-1})) (e_{k} + e_{k} (\hat{\beta}_{k-1})) \hat{\beta}_{k-1} \right)$$

$$+ \sum_{k=1}^{t} k^{-1/2-\delta} F^{-1}(\hat{\beta}_{k-1}) \psi_{k} (e_{k} - e_{k} (\hat{\beta}_{k-1}))$$

$$- \sum_{k=1}^{t} k^{-1/2-\delta} F^{-1}(\hat{\beta}_{k-1}) (\psi_{k} - \psi_{k} (\hat{\beta}_{k-1})) \psi_{k}^{T} (\beta^{*}) \hat{\beta}_{k-1}.$$  

By using four times Lemma A.12, we know that

$$kE(e_{k} - e_{k} (\hat{\beta}_{k-1})) || e_{k} + e_{k} (\hat{\beta}_{k-1})) E(e_{k} - e_{k} (\hat{\beta}_{k-1})) || e_{k} (\hat{\beta}_{k-1})),$$

$$kE(|| \psi_{k} || || e_{k} - e_{k} (\hat{\beta}_{k-1})) + kE(|| \psi_{k} - \psi_{k} (\hat{\beta}_{k-1}) || \psi_{k} (\beta^{*}) || e_{k} (\hat{\beta}_{k-1})).$$

are bounded, as well as $F^{-1}(\beta)$ over $D_{G}$, hence by using Lemma A.10 (Lukacs, 1975) we prove that $Z_{t}$ converges a.s. and by Lemma A.10 (Kronecker’s lemma) $t^{-1/2-\delta} \sum_{k=1}^{t} A_{2,k} \to 0$ a.s. when $t \to \infty$. Similarly, this time with $Z_{t} = \sum_{k=1}^{t} k^{-1/2-\delta} (1/k) (e_{k}^{2} - \hat{\sigma}_{k}^{2}) F^{-1}(\hat{\beta}_{k-1}) \psi_{k} e_{k}$, since $E(|| \psi_{k} - \hat{\sigma}_{k}^{2}) F^{-1}(\hat{\beta}_{k-1}) \psi_{k} e_{k} ||)$ is bounded, then $Z_{t}$ is finite and the proof proceeds as before.

LEMMA A.16  $\forall \delta > 0$, $t^{-1/2-\delta} \sum_{k=1}^{t} F^{-1}(\hat{\beta}_{k-1}) \psi_{k} e_{k} \to 0$ a.s. when $t \to \infty$.

Proof  The proof is given by Ljung and Söderström (1983, p. 442), Lemma 4.8.3 and is repeated here for completeness. Consider the series

$$s_{t} = \sum_{k=1}^{t} k^{-1/2-\delta} F^{-1}(\hat{\beta}_{k-1}) \psi_{k} e_{k}.$$ 

That random vector is a martingale with respect to the $\sigma$-algebra $\mathcal{F}_{t-1}$, spanned by the $e_{t}$, $i \leq t - 1$. Indeed

$$E(s_{t} | \mathcal{F}_{t-1}) = s_{t-1} + E(t^{-1/2-\delta} F^{-1}(\hat{\beta}_{k-1}) \psi_{k} e_{k} | \mathcal{F}_{t-1}) = s_{t-1} + t^{-1/2-\delta} F^{-1}(\hat{\beta}_{k-1}) \psi_{k} E(e_{t} | \mathcal{F}_{t-1}) = s_{t-1},$$

since $F^{-1}(\hat{\beta}_{k-1})$ and $\psi_{t}$ do not depend of $e_{t}$, $i \leq t - 1$. Moreover

$$E(s_{t}^{2}) \leq \sum_{k=1}^{t} k^{-1/2-\delta} E((F^{-1}(\hat{\beta}_{k-1}) \psi_{k} e_{k})^{2}) \leq C \sum_{k=1}^{t} k^{-1/2-\delta} < \infty,$$

where $C$ denotes a constant. Hence $s_{t}$ is a martingale with a bounded variance, and according to Lemma A.10 (Chung, 1968), $s_{t}$ converges a.s. to a finite limit $\bar{s}_{\infty}$. Hence by Lemma A.10 (Kronecker’s lemma), $t^{-1/2-\delta} \sum_{k=1}^{t} F^{-1}(\hat{\beta}_{k-1}) \psi_{k} e_{k} \to 0$ a.s. when $t \to \infty$.

LEMMA A.17  Let $S_{t}$ defined by Equation (A6) and $T_{k}$ defined by Equation (A7), then $\forall \delta > 0$,

$$G_{t} = t^{-1/2-\delta} \left( T_{t \beta_{t-1}} - \sum_{k=1}^{t-1} T_{k} (\hat{\beta}_{k} - \hat{\beta}_{k-1}) - F^{-1}(\beta^{*}) S_{t} \hat{\beta}_{t-1} \right)$$

converges to 0 a.s. when $t \to \infty$.

Proof  Replacing $\beta$ by $\beta^{*}$ in Lemma A.11 implies that, for all $\delta > 0$, $t^{-1/2-\delta} S_{t}$ and $t^{-1/2-\delta} T_{t}$ converge to 0 a.s. when $t \to \infty$, which implies that $t^{-1/2-\delta} T_{k} \hat{\beta}_{k-1}$ and $t^{-1/2-\delta} F^{-1}(\beta^{*}) S_{t} \hat{\beta}_{t-1}$ converge to 0 a.s. when $t \to \infty$. Let us now prove that

$$t^{-1/2-\delta} F^{-1}(\beta^{*}) \sum_{k=1}^{t-1} S_{k} (\hat{\beta}_{k} - \hat{\beta}_{k-1}) \to 0, \quad \text{a.s. when } t \to \infty.$$ 

Consider the series

$$F^{-1}(\beta^{*}) \sum_{k=1}^{t-1} k^{-1/2-\delta} S_{k} (\hat{\beta}_{k} - \hat{\beta}_{k-1}).$$

From Equation (44), we know that $(\hat{\beta}_{k} - \hat{\beta}_{k-1}) = (1/k) \delta_{k}^{2} F^{-1}(\hat{\beta}_{k-1}) \psi_{t,1}$ and $\forall \mu > 0$, $k^{-1/2-\mu} S_{k} = o(1)$. Let $0 < \mu < \delta$, so

$$k^{-1/2-\delta} || S_{k} (\hat{\beta}_{k} - \hat{\beta}_{k-1}) || = k^{-1/2-\mu} \delta_{k}^{2} || S_{k} (\hat{\beta}_{k} - \hat{\beta}_{k-1}) || \leq C k^{-1/2-\mu} || \psi_{t,1} ||$$

hence, by Lemmas A.12 and A.14, we conclude that $F^{-1}(\beta^{*}) \sum_{k=1}^{t-1} k^{-1/2-\delta} S_{k} (\hat{\beta}_{k} - \hat{\beta}_{k-1})$ is finite, and by Lemma A.15, when $t \to \infty$, $t^{-1/2-\delta} F^{-1}(\beta^{*}) \sum_{k=1}^{t-1} S_{k} (\hat{\beta}_{k} - \hat{\beta}_{k-1}) \to 0$ a.s. Similarly, we show that $t^{-1/2-\delta} \sum_{k=1}^{t-1} T_{k} (\hat{\beta}_{k} - \hat{\beta}_{k-1}) \to 0$, a.s.

LEMMA A.18  Using notations in the proof of Theorem 3 and the sequence $a_{k}$ from Lemma A.13, there exist two constants positive $M_{1}$ and $M_{2}$ such that

$$\forall t > 0, \quad t || \hat{\beta}_{t} || < M_{1} \sum_{k=1}^{t} a_{k} || \hat{\beta}_{k} ||^{2} + M_{2} t^{1/2+\delta} \quad \text{a.s. (A22)}.$$ 

Proof  From Equation (A3) and Lemma A.9, we have

$$t \hat{\sigma}_{t+1}^{2} \psi_{t} = \sum_{k=1}^{t} A_{1,k} + \sum_{k=1}^{t} A_{2,k} + \sum_{k=2}^{t} F^{-1}(\hat{\beta}_{k-1}) \psi_{k} e_{k}$$

$$+ T_{t-1} \hat{\beta}_{t-1} - \sum_{k=1}^{t-1} T_{k} (\hat{\beta}_{k} - \hat{\beta}_{k-1})$$

$$- F^{-1}(\beta^{*}) S_{t} \hat{\beta}_{t-1} + F^{-1}(\beta^{*}) \sum_{k=1}^{t-1} S_{k} (\hat{\beta}_{k} - \hat{\beta}_{k-1})$$

$$+ \sum_{k=1}^{t} \frac{1}{k} (e_{k}^{2} - \hat{\sigma}_{k}^{2}) F^{-1}(\hat{\beta}_{k-1}) \psi_{k} e_{k} \quad \text{(A23)}$$

Downloaded by [Archives & Bibliothèques de l'ULB], [Guy Mélard] at 02:26 24 June 2014
hence
\[ t \tilde{\beta}_t = \tilde{\sigma}_{t+1}^2 \sum_{k=1}^t A_{1,k} + \tilde{\sigma}_{t+1}^2 \sum_{k=1}^t A_{2,k} \]
\[ + \tilde{\sigma}_{t+1}^2 \sum_{k=2}^t F^{-1}(\tilde{\beta}_{k-1}) \psi_k e_k + \tilde{\sigma}_{t+1}^2 T_t \tilde{\beta}_{t-1} \]
\[ - \tilde{\sigma}_{t+1}^2 \sum_{k=1}^{t-1} T_k (\tilde{\beta}_k - \tilde{\beta}_{k-1}) - \tilde{\sigma}_{t+1}^2 F^{-1}(\beta^*) S_t \tilde{\beta}_{t-1} \]
\[ + \tilde{\sigma}_{t+1}^2 F^{-1}(\beta^*) \sum_{k=1}^{t-1} S_k (\tilde{\beta}_k - \tilde{\beta}_{k-1}) \]
\[ + \frac{1}{C} (\epsilon^2 - \tilde{\sigma}_k^2) F^{-1}(\tilde{\beta}_{k-1}) \psi_k e_k, \]
which implies
\[ \|t \tilde{\beta}_t\| \leq \left| \tilde{\sigma}_{t+1}^2 \right| \left( \sum_{k=1}^t \|A_{1,k}\| + \|A_{2,k}\| \right) \]
\[ + \left| \tilde{\sigma}_{t+1}^2 \right| \left( \sum_{k=1}^t F^{-1}(\tilde{\beta}_{k-1}) \psi_k e_k \right) \]
\[ + \left| \tilde{\sigma}_{t+1}^2 \right| \left( T_t \tilde{\beta}_{t-1} + \sum_{k=1}^{t-1} T_k (\tilde{\beta}_k - \tilde{\beta}_{k-1}) \right) \]
\[ + F^{-1}(\beta^*) S_t \tilde{\beta}_{t-1} + F^{-1}(\beta^*) \sum_{k=1}^{t-1} S_k (\tilde{\beta}_k - \tilde{\beta}_{k-1}) \]
\[ + \left| \tilde{\sigma}_{t+1}^2 \right| \left( \sum_{k=1}^t \frac{1}{C} (\epsilon^2 - \tilde{\sigma}_k^2) F^{-1}(\tilde{\beta}_{k-1}) \psi_k e_k \right) \]  
(A24)

By Lemma A.13, there exists a constant C such that \( \forall k \geq 1, \|A_{1,k}\| < C \|\tilde{\beta}_{k-1}\|^2 \), hence \( \sum_{k=1}^t \|A_{1,k}\| < C \sum_{k=1}^t a_k \|\tilde{\beta}_{k-1}\|^2 \), so there exists a constant \( C_1 > 0 \) such that \( \forall k > 1 \),
\[ \left| \tilde{\sigma}_{t+1}^2 \right| \sum_{k=1}^t \|A_{1,k}\| \leq C_1 \sum_{k=1}^t a_k \|\tilde{\beta}_{k-1}\|^2. \]

By Lemma A.16, we know that for all \( \delta > 0 \), \( r^{1/2-\delta} \sum_{k=1}^t F^{-1}(\tilde{\beta}_{k-1}) \psi_k e_k \to 0 \) a.s. when \( t \to \infty \), so there exists a constant \( C_2 > 0 \) such that \( \forall k > 1 \),
\[ \left| \tilde{\sigma}_{t+1}^2 \right| \sum_{k=1}^t F^{-1}(\tilde{\beta}_{k-1}) \psi_k e_k \leq C_2 t^{1/2+\delta}. \]

By Lemma A.17, we have that
\[ G_t = r^{1/2-\delta} \left( T_t \tilde{\beta}_{t-1} - \sum_{k=1}^t T_k (\tilde{\beta}_k - \tilde{\beta}_{k-1}) - F^{-1}(\beta^*) S_t \tilde{\beta}_{t-1} \right) \]
\[ + F^{-1}(\beta^*) \sum_{k=1}^{t-1} S_k (\tilde{\beta}_k - \tilde{\beta}_{k-1}) \]
converges to 0 a.s. when \( t \to \infty \) and, by Lemma A.15, \( r^{-1/2-\delta} \sum_{k=1}^t A_{2,k} \to 0 \) a.s. when \( t \to \infty \), so that there exists a constant \( C_3 > 0 \) such that \( \forall k \geq 1 \),
\[ \left| \tilde{\sigma}_{t+1}^2 \right| \sum_{k=1}^t A_{2,k} \leq \left| \tilde{\sigma}_{t+1}^2 \right| \sum_{k=1}^t T_k (\tilde{\beta}_{k-1} - \tilde{\beta}_{k-2}) \]
\[ + F^{-1}(\beta^*) S_t \tilde{\beta}_{t-1} + F^{-1}(\beta^*) \sum_{k=1}^t S_k (\tilde{\beta}_k - \tilde{\beta}_{k-2}) \]
\[ + \left| \tilde{\sigma}_{t+1}^2 \right| \left( \sum_{k=1}^t \frac{1}{C} (\epsilon^2 - \tilde{\sigma}_k^2) F^{-1}(\tilde{\beta}_{k-1}) \psi_k e_k \right) \leq C_3 t^{1/2+\delta}. \]

From Equation (A24), we can conclude that for each \( \delta > 0 \), there exists a constant \( M > 0 \), such that
\[ t \| \tilde{\beta}_t \| \leq M \sum_{k=1}^t a_k \|\tilde{\beta}_{k-1}\|^2 + M t^{1/2+\delta}. \]  
(A25)

The following lemma is more general than a result of Ljung and Söderström (1983, p. 445) and the proof is very different.

**Lemma A.19**  Let \( b_t \) and \( a_t \) sequences of positive numbers such that \( b_t \) converges to 0, and \( (1/t) \sum_{k=1}^t a_k \) converges. If for all \( t \)
\[ t b_t < M \sum_{k=1}^t a_k b_k^2 + M t^{1/2+\delta} \]  
(A26)

then for all \( \gamma < \frac{1}{2} - \delta \), \( t^\gamma b_t \) converges to 0.

**Proof**  Let us take \( \theta < (1 - 2\delta)/4 \) and suppose that \( t^\theta b_t \) is not bounded. From Lemma A.14, \( r^{-1+\theta} \sum_{k=1}^t \epsilon_k^2 a_k \) is finite. Let
\[ \epsilon < \left( \frac{2 M_1 \sup_{t \to 0} \sum_{k=1}^t a_k k^{-\theta}}{r^{-1+\theta} \sum_{k=1}^t \epsilon_k^2 a_k} \right)^{-1}. \]  
(A27)

There exists \( t_0 \) such that \( b_t < \epsilon \) for all \( t \geq t_0 \). Let \( M \) a positive real number such that \( \sup_{t \leq t_0} t^\theta b_t < M \), hence for all \( t < t_0, t^\theta b_t < M \). Let \( t_0 \geq t_0 \) such that \( t_0^\theta b_t > M \) and for all \( t < t_1, t^\theta b_t < M \) thus \( t_0^\theta > M/\epsilon \). To have \( t_1 \) larger it suffices to take \( M \) larger. We have
\[ t_1 b_{t_1} < M_1 M^2 \sum_{k=1}^{t_1-1} a_k k^{-2\theta} + M t_1^{1/2+\delta} \]
\[ < M_1 M^2 \sum_{k=1}^{t_1-1} a_k k^{-2\theta} + M t_1^{1/2+\delta} \]
hence
\[ b_{t_1} < M_1 b_{t_1}^{1-2\theta} \sum_{k=1}^{t_1-1} a_k k^{-2\theta} + M t_1^{1/2+\delta}. \]  
(A28)

Let us consider the corresponding second degree equation in \( b_t \) and its roots
\[ b_{t_1}^2 = \frac{1 \pm \sqrt{1 - 4M_1 M_2 (t_1^{-1+2\theta} \sum_{k=1}^{t_1-1} a_k k^{-2\theta}) t_1^{1/2+\delta}}}{2M_1 (t_1^{-1+2\theta} \sum_{k=1}^{t_1-1} a_k k^{-2\theta})}. \]  
(A29)

These roots are real for a large enough \( t_1 \), hence for large values of \( M \). Inequality (A28) implies either \( b_{t_1} > b_{t_1}^\uparrow \) or \( b_{t_1} < b_{t_1}^\uparrow \). The former is impossible because we would have \( b_{t_1} > \epsilon \) in contradiction
with the choice of $t_1 \geq t_0$. Hence $b_{t_1} < b_{t_0}^-$ which implies

$$t_0^* b_{t_1} < \sqrt{t_0^*} \left( - \sqrt{t_0^*} - 4M_1 \left( t_1^{1+2\theta} \sum_{k=1}^{t_1} a_k^{-1/2} \right)^{1/2} \right).$$

Since $-1 + 2\theta < 0$, when $M$ is large the right-hand side becomes smaller, consequently we will have $t_0^* b_{t_1} < M$ which is absurd since $t_0^* b_{t_1} > M$. Hence for all $\theta < (1 - 2\theta)/4$, $t_0^* b_{t_1}$ is bounded.

Let now take $\gamma = 1/2 - \delta - \epsilon_0$, where $\epsilon_0$ is small. From Equation (A26), we have

$$tb_t < M_1 \sum_{k=1}^{t-1} a_k^{-1/2} \frac{1}{2}(k^{1/2} + a_k^{-1/2} b_k^2) + M_2 t^{1/2 + \delta}$$

$$< M_1 C \sum_{k=1}^{t-1} a_k^{-1/2} + M_2 t^{1/2 + \delta}$$

since $\gamma/2 + \epsilon_0/4 < (1 - 2\delta)/4$ and where $C$ is a constant, hence

$$t^\gamma b_t < M_1 C \sum_{k=1}^{t-1} a_k^{-1/2} + M_2 t^{1/2 + \gamma}$$

$$= t^{-\epsilon_0/2} M_1 C \sum_{k=1}^{t-1} a_k^{-1/2} + M_2 t^{-\epsilon_0}$$

which implies the result.

**Lemma A.20** For all $\gamma < 1/2$, $t^\gamma \| \mathbf{\hat{B}}_t \|$ converges to 0 almost surely.

**Proof** It is an immediate consequence of Lemmas A.18 and A.19.

**Lemma A.21** For all $\gamma$ such that $0 < \gamma < 1/2$, then $t^\gamma (\hat{\sigma}_t^2 - \sigma_t^2) \longrightarrow 0$ a.s. when $t \longrightarrow \infty$.

**Proof** From Equation (35) we have

$$t^\gamma (\hat{\sigma}_t^2 - \sigma_t^2) = \frac{1}{t-1} \sum_{k=1}^{t} \left( (\varepsilon_{k-1} - \varepsilon_{k-1}(\hat{\beta}_{k-2})) \times (\varepsilon_{k-1} + \varepsilon_{k-1}(\hat{\beta}_{k-2})) \right) + \frac{1}{t-1} \sum_{k=1}^{t} (\hat{\beta}_{k-2} - \varepsilon_{k-1}(\beta^*)) \times (\hat{\beta}_{k-2} + \varepsilon_{k-1}(\beta^*)) + \frac{1}{t-1} \sum_{k=1}^{t} (\hat{\beta}_{k-2} + \varepsilon_{k-1}(\beta^*)) \times (\hat{\beta}_{k-2} + \varepsilon_{k-1}(\beta^*))$$

By Lemma A.12 we know that $kE(|\varepsilon_{k-1} - \varepsilon_{k-1}(\hat{\beta}_{k-2})|) |\varepsilon_{k-1}| + |\varepsilon_{k-1}(\hat{\beta}_{k-2})|$ is bounded, hence

$$\sum_{k=1}^{t} |(\varepsilon_{k-1} - \varepsilon_{k-1}(\hat{\beta}_{k-2}))| \leq C \sum_{k=1}^{t} k^{-2+\gamma}$$

and, by Lemma A.10 (Lukacs, 1975),

$$\sum_{k=1}^{t} k^{-1+\gamma} (|\varepsilon_{k-1} - \varepsilon_{k-1}(\hat{\beta}_{k-2})|) \leq C \sum_{k=1}^{t} k^{-2+\gamma}$$

converges, and, by Lemma A.10 (Kronecker’s Lemma), Equation (A30) $\rightarrow 0$.

There exists a constant $C$ such that

$$\sum_{k=1}^{t} k^{-1+\gamma} (|\varepsilon_{k-1} - \varepsilon_{k-1}(\hat{\beta}_{k-2})|) (|\varepsilon_{k-1} - \varepsilon_{k-1}(\hat{\beta}_{k-2})|) \leq C \sum_{k=1}^{t} k^{-1+\gamma} (|\varepsilon_{k-1} - \varepsilon_{k-1}(\hat{\beta}_{k-2})|)$$

converges to a finite limit, and, by Kronecker’s Lemma, Equation (A31) $\rightarrow 0$ a.s. By Lemma A.11, Equation (A32) $\rightarrow 0$ a.s. when $t \rightarrow \infty$.

**Lemma A.22** (Brown, 1971, Theorem 1) Let $\{S_t, F_t, t = 1, \ldots\}$ be a martingale. Let

$$X_t = S_t - S_{t-1}, \quad V_t = \sum_{k=1}^{t} E(X_k^2 | F_{k-1}), \quad s_t^2 = EV_t^2 = ES_t^2.$$

Suppose that $V_t^2 s_t^{-2} \text{ converges with probability } 1 \text{ to } 0 \text{ when } t \rightarrow \infty$, and that the following Lindeberg condition is satisfied: $\forall \delta > 0, s_t^{-2} \sum_{j=1}^{t} E X_j^2 I(X_j \geq \delta s_t) \text{ converges in probability to } 1 \text{ when } t \rightarrow \infty$. Then $S_t/s_t$ converges in law to the normal distribution with mean 0 and variance 1.

**Lemma A.23**

$$\left( \frac{1}{\sqrt{t}} \right) \sum_{k=1}^{t} R_k (\beta^*) \hat{\sigma}_k^{-2} F^{-1}(\beta^*) \psi_k(\beta^*) \epsilon_k \leq C \sum_{k=1}^{t} k^{-2+\gamma}$$

converges to the normal distribution $N(0, \sigma_0^2 F(\beta^*))$.

**Proof** A sketch of the proof is given by Ljung and Söderström (1983, p. 448), Lemma 4.4.7 but there are again differences here because of Assumption 2.
We use Lemma A.22 with the Cramér–Wold device. Let
\[ Y_t^2 = \sum_{k=1}^{t} E[R_k(\beta^*) \delta_k^{-2} F^{-1}(\beta^*) \psi_k(\beta^*) e_k] \]
\[ \times (R_k(\beta^*) \delta_k^{-2} F^{-1}(\beta^*) \psi_k(\beta^*) e_k)^T [\mathcal{F}_{t-1}] \]
and \( M_t^2 = EY_t^2 \). We have
\[ Y_t^2 = \sum_{k=1}^{t} (R_k(\beta^*) - \sigma_k^2 F(\beta^*)) \delta_k^{-4} F^{-1}(\beta^*) \psi_k(\beta^*) \]
\[ \times \psi_k(\beta^*) F^{-1}(\beta^*) R_k(\beta^*) \sigma_k^2 \]
\[ + t \sigma_k^4 \delta_k^{-4} \psi_k(\beta^*) \psi_k(\beta^*) F^{-1}(\beta^*) [R_k(\beta^*) - \sigma_k^2 F(\beta^*)] \]
\[ + \sum_{k=1}^{t} \sigma_k^6 (\sigma_k^2 - \delta_k^2) (\sigma_k^2 + \delta_k^2) \delta_k^{-4} \sigma_k^{-4} \psi_k(\beta^*) \psi_k(\beta^*) \]
\[ + \sum_{k=1}^{t} \sigma_k^2 \psi_k(\beta^*) \psi_k(\beta^*). \quad (A34) \]

Let \( x_k \) be the first component of \( R_k(\beta^*) \delta_k^{-2} F^{-1}(\beta^*) \psi_k(\beta^*) \). To prove that \( S_t = \sum_{k=1}^{t} x_k e_k \) converges in law to the normal distribution, we check the two conditions of Lemma A.22, with \( V_t^2 = \sum_{k=1}^{t} E[(x_k e_k)^2 | \mathcal{F}_{k-1}] \) and \( \sigma_t^2 = E(V_t^2) \).

First, by Lemma A.21, we have \( (\sigma_t^2 - \delta_t^2) = o(t^{-\gamma}) \), hence
\[ \frac{1}{t} \sum_{k=1}^{t} (\sigma_k^2 - \delta_k^2) (\sigma_k^2 + \delta_k^2) \delta_k^{-4} \sigma_k^{-4} \psi_k(\beta^*) \psi_k(\beta^*) \]
\[ \leq C \frac{1}{t} \sum_{k=1}^{t} k^{-\gamma} \| \psi_k(\beta^*) \psi_k(\beta^*) \| \]
and by Lemma A.11,
\[ (R_t(\beta^*) - \sigma_k^2 F(\beta^*)) = o(t^{-1/2+\delta}), \]
so by applying Kronecker’s Lemma A.10 to Equation (A34), the first three terms of \( Y_t^2 / t \) converge to 0 as \( t \to \infty \), hence \( Y_t^2 / t \to \sigma_k^2 F(\beta^*) \). It is obvious that \( M_t^2 / t \to \sigma_k^2 F(\beta^*) \) so \( V_t^2 / \sigma_t^2 \to 1 \) in probability when \( t \to \infty \), where \( s_t / \sqrt{t} \) is the square root of the element \((1,1)\) of \( M_t^2 / t \).

To prove the Lindeberg condition, let us consider
\[ W_t = \frac{1}{t} \sum_{k=1}^{t} E[x_k e_k]^2 I(|x_k e_k| > \delta s_t). \]
We have \( E(x_k e_k)^2 = E(V_k^2) = kC \), where \( C \) is a constant. By Tchebychev inequality we obtain
\[ P(|x_k e_k| > \delta s_k) \leq \frac{1}{|\delta s_k|^2} E(x_k e_k)^2 \leq \frac{1}{\delta^2 kC} E(x_k e_k)^2. \]
Therefore, since for all \( k < t, s_k < s_t \)
\[ W_t \leq \frac{1}{t} \sum_{k=1}^{t} \left( \frac{1}{s_k^2} \right) E(x_k e_k)^2 E(I(|x_k e_k| > \delta s_k)) \]
\[ \leq \frac{1}{\delta^2 kC} C^2 (E(x_k e_k)^2), \]
hence, by Lemma A.10 (Lukacs, 1975) and Kronecker’s Lemma A.10, \( W_t \to 0 \) when \( t \to \infty \). Hence Lemma A.22 implies convergence in law of \((S_t/\sqrt{t})/(\delta_t/\sqrt{t})\). In a similar way, we can show that any linear combination of the components of Equation (A33) converges in law to a normal distribution.

**Lemma A.24** The series \( H_t \) and \( L_t \) defined by Equations (50) and (51) converge to 0 a.s. when \( t \to \infty \).

**Proof** Let us show that \( L_t \) converges to 0 a.s. when \( t \to \infty \). Consider the series
\[ L_t^1 = \sum_{k=1}^{t} \frac{1}{\sqrt{k}} R_k(\beta^*) \delta_k^{-2} F^{-1}(\beta_k-1) \psi_k(\beta_k-1) e_k \]
\[ + \sum_{k=1}^{t} \frac{1}{\sqrt{k}} R_k(\beta^*) \delta_k^{-2} F^{-1}(\beta^*) (\psi_k(\beta_k-1) - \psi_k(\beta^*)) e_k. \]
\( L_t^1 \) is a martingale and
\[ E\|L_t^1\|^2 \leq \sum_{k=1}^{t} \frac{1}{k} \| R_k(\beta^*) \delta_k^{-2} F^{-1}(\beta_k-1) \]
\[ - F^{-1}(\beta^*) \| \psi_k(\beta_k-1) e_k \|^2 \]
\[ + \sum_{k=1}^{t} \frac{1}{k} \| R_k(\beta^*) \delta_k^{-2} F^{-1}(\beta^*) (\psi_k(\beta_k-1) - \psi_k(\beta^*)) e_k \|^2. \]
We know that \( F^{-1}(\beta) \) and \( \delta F(\beta)/\beta \) are bounded, then by Lemma A.2, we have that \( \beta F^{-1}(\beta)/\beta \) is bounded, then there exists a constant \( C \) such that
\[ E\|L_t^1\|^2 \leq C \sum_{k=1}^{t} \frac{1}{k} E\|\beta_k-1\|^2 E|e_k|^2. \]
Using Equation (A22) and Lemma A.20, for all \( \gamma \) such that \( 0 < \gamma < 1/2 - \delta, \|B_\beta\|^2 = t^{3/2} \|B_\beta\|^2 \) converges to 0, hence \( E\|L_t^1\|^2 \) is bounded. Then \( L_t^1 \) is a martingale with a bounded variance, and using Chung (1968, p. 310) \( L_t^1 \) converges a.s. to a finite limit when \( t \to \infty \), hence \( R^{-1}(\beta^*) L_t^1 \) converges, and by Kronecker’s Lemma A.10, \( L_t \) converges a.s. to 0 when \( t \to \infty \).

Let us now show that \( H_t \to 0 \) a.s. when \( t \to \infty \). Let
\[ H_t^1 = R_t^{-1}(\beta^*) \sum_{k=1}^{t} \frac{1}{\sqrt{k}} (\sigma_k^2 F(\beta^*)) \]
\[ - R_k(\beta^*) \delta_k^{-2} F^{-1}(\beta_k-1) \psi_k(\beta^*) \psi_k(\beta^*) \beta_k \]
\[ + R_t^{-1}(\beta^*) \sum_{k=1}^{t} \frac{1}{\sqrt{k}} (\delta_k^2 - \sigma_k^2) F(\beta^*) \delta_k^{-2} F^{-1}(\beta_k-1) \]
\[ \times \psi_k(\beta^*) \psi_k(\beta^*) \beta_k \]
\[ + R_t^{-1}(\beta^*) \sum_{k=1}^{t} \frac{1}{\sqrt{k}} B_{1,k} + R_t^{-1}(\beta^*) \sum_{k=1}^{t} \frac{1}{\sqrt{k}} B_{2,k}. \]
By Lemma A.11, we know that \( \forall \mu > 0 \),
\[ t^{1/2-\mu} (R_t(\beta^*) - \sigma_k^2 F(\beta^*)) = o(1), \]
and since for every \( \gamma \) positive such that \( 0 < \gamma < \frac{1}{2} \), \( t^\gamma \hat{\beta}_k \)

converges to 0, then

\[
\|k^{-1/2}(\hat{\sigma}_k^2 F(\beta^*) - R_k(\beta^*)\hat{\sigma}_k^{-2}F^{-1}(\hat{\beta}_{k-1})(\beta^*)\psi_k^T(\beta^*)\hat{\beta}_k)\| \\
\leq Ck^{-1+\mu-\gamma}\|\psi_k^T(\beta^*)\| \\
\]

hence with \( \mu < \gamma \) by Lemmas A.12 and A.14

\[
R_t^{-1}(\beta^*) \sum_{k=1}^{t} \frac{1}{\sqrt{k}}(\hat{\sigma}_k^2 F(\beta^*) - R_k(\beta^*)\hat{\sigma}_k^{-2}F^{-1}(\hat{\beta}_{k-1})(\beta^*)\psi_k^T(\beta^*)\hat{\beta}_k) \\
- R_k(\beta^*)\hat{\sigma}_k^{-2}F^{-1}(\hat{\beta}_{k-1})(\beta^*)\psi_k^T(\beta^*)\hat{\beta}_k \\
\]

is convergent. In the same manner, we have

\[
\frac{1}{\sqrt{k}}\| (\hat{\sigma}_k^2 - \sigma_k^2)F(\beta^*)\hat{\sigma}_k^{-2}F^{-1}(\hat{\beta}_{k-1})(\beta^*)\psi_k^T(\beta^*)\hat{\beta}_k \| \\
\leq Ck^{-1/2}\| \hat{\sigma}_k^2 - \sigma_k^2 \| \| \psi_k(\beta^*)\psi_k^T(\beta^*) \| \\
\]

by Lemma A.21, \( t^{1/2-\mu}(\hat{\sigma}_t^2 - \sigma_t^2) \to 0 \) a.s. when \( t \to \infty \), hence

\[
\frac{1}{\sqrt{k}}\| (\hat{\sigma}_k^2 - \sigma_k^2)F(\beta^*)\hat{\sigma}_k^{-2}F^{-1}(\hat{\beta}_{k-1})(\beta^*)\psi_k^T(\beta^*)\hat{\beta}_k \| \\
\leq Ck^{-1+\mu-\gamma}\| \psi_k(\beta^*)\psi_k^T(\beta^*) \| \\
\]

hence

\[
R_t^{-1}(\beta^*) \sum_{k=1}^{t} \frac{1}{\sqrt{k}}(\hat{\sigma}_k^2 - \sigma_k^2)F(\beta^*)\hat{\sigma}_k^{-2}F^{-1}(\hat{\beta}_{k-1}) \\
\times \psi_k(\beta^*)\psi_k^T(\beta^*)\hat{\beta}_{k-1} \\
\]

is convergent.

In the same way as in Lemma A.13, we can show from Equation (46) that \( \| B_{1,t} \| < C\| \hat{\beta}_t \|^2 \), where \( (1/t) \sum_{k=1}^{t} a_t \) is finite a.s., and since \( \forall \gamma < 1/2, \psi_t^2 \| \hat{\beta}_t \|^2 \) is bounded, then with \( \gamma \) such that \( \gamma > \frac{1}{4} \),

\[
\sum_{k=1}^{t} \frac{1}{\sqrt{k}}B_{1,k} \leq C \sum_{k=1}^{t} a_t k^{-1/2-2\gamma}, \\
\]

thus \( R_t^{-1}(\beta^*) \sum_{k=1}^{t} B_{1,k}/\sqrt{k} \) converges a.s. Let us now consider \( B_{2,t} \) defined by Equation (47).

From Lemma A.12 we know that \( tE(\| \psi_t - \psi_t(\hat{\beta}_{t-1}) \| \| E_t(\hat{\beta}_{t-1})) \) and \( tE(\| E_t(\hat{\beta}_{t-1}) \| \| \psi_t \|) \) are bounded and since \( \hat{\sigma}_t^{-2} \) is bounded, then \( R_t^{-1}(\beta^*) \sum_{k=1}^{t} E\| B_{2,k} \|/\sqrt{k} \) converges a.s. hence by Lukacs’ Lemma A.10 \( R_t^{-1}(\beta^*) \sum_{k=1}^{t} B_{2,k}/\sqrt{k} \) converges a.s. Finally \( H_1 \) is finite. According to Kronecker’s Lemma A.10, \( H_1 \to 0 \) a.s. when \( t \to \infty \).