The Selection of ARIMA Models with or without Regressors

Søren Johansen*
Department of Economics, University of Copenhagen and
CREATES, University of Aarhus, Denmark
Marco Riani†
Dipartimento di Economia, Università di Parma, Italy
Anthony C. Atkinson‡
Department of Statistics,
London School of Economics, London WC2A 2AE, UK

November 8, 2012

Abstract

We develop a C_p statistic for the selection of regression models with stationary and nonstationary ARIMA error term. We derive the asymptotic theory of the maximum likelihood estimators and show they are consistent and asymptotically Gaussian. We also prove that the distribution of the sum of squares of one step ahead standardized prediction errors, when the parameters are estimated, differs from the chi-squared distribution by a term which tends to infinity at a lower rate than χ^2_n. We further prove that, in the prediction error decomposition, the term involving the sum of the variance of one step ahead standardized prediction errors is convergent. Finally, we provide a small simulation study. Empirical comparisons of a consistent version of our C_p statistic with BIC and a generalized RIC show that our statistic has superior performance, particularly for small signal to noise ratios. A new plot of our time series C_p statistic is highly informative about the choice of model.

*e-mail: sjo@math.ku.dk
†e-mail: mriani@unipr.it
‡e-mail: a.c.atkinson@lse.ac.uk.

On the way we introduce a new version of AIC for regression models, show that it estimates a Kullback-Leibler distance and provide a version for small samples that is bias corrected. We highlight the connections with standard Mallows C_p.

Keywords: AIC; ARMA models; bias correction; BIC; C_p plot; generalized RIC; Kalman filter; Kullback-Leibler distance; state-space formulation

1 Introduction

There is a vast literature on methods for selection of non-nested models. In AIC (Akaike, 1974) the maximized log-likelihood is penalized by the number of parameters in the model. For Gaussian regression models, AIC becomes Mallows' C_p (Mallows, 1973) when the nuisance parameter σ^2 is estimated from a full model containing sufficiently many terms to ensure that the estimator is unbiased. For such regression models the distribution of C_p is a linear function of an F random variable (Gilmour, 1996). There is no simple non-asymptotic expression for the distribution of AIC. Book length treatments of the properties and applications of these and other procedures include McQuarrie and Tsai (1998), Burnham and Anderson (2002) and, more recently, Konishi and Kitagawa (2008) and Claeskens and Hjort (2008).

Tong (2001, §9) gives references to methods solely for time series. More recent contributions include So *et al.* (2006) who consider the case of, possibly lagged, exogenous variables and GARCH errors. Wang *et al.* (2007) extend the least absolute shrinkage and selection operator ('lasso') to regression models with autocorrelated errors. Claeskens *et al.* (2007) emphasize the mean squared forecast error and suggest an alternative to AIC and BIC in autoregressive time series models. Finally, Shi and Tsai (2004) obtain a residual information criterion (RIC) for joint selection of regression variables and the order of autoregressive errors. However, there appears to be no extension of the standard results to the class of models of interest here, that of all models which contain explanatory variables, and have a reduced form ARIMA representation whose AR and MA roots lie outside the unit circle.

The structure of the paper is as follows. In §2 we propose new versions of AIC (or its consistent version BIC) and of C_p which can be applied to the choice of regression models with independent or stationary error terms as well as to some non-stationary error processes. We provide arguments for the asymptotic distribution of our time series C_p statistic. We also show, for regression with independent errors, that our new version of AIC is a consistent unbiased estimator of the expected Kullback-Leibler information between the true model and the fitted candidate model. We derive a small sample correction factor to make this new AIC

unbiased and we explore its close relationship with traditional AIC and C_p.

We start the derivation of the asymptotic distribution of our C_p statistic for time series in §3 by proving a series of theorems in the context of linear regression with ARIMA errors. The asymptotic distribution of the Gaussian maximum likelihood estimator of the ARMA parameter vector θ and the regression parameters β, has been given by Hannan (1973), for stationary regressors using frequency domain methods. Yao and Brockwell (2006) analyse the model without regressors and show that $\hat{\theta}$ is consistent and asymptotically Gaussian using time domain methods. In §3.2 we show how the results of Yao and Brockwell can be applied and extended to the regression model with deterministic regressors. Our new results cover the strong consistency of maximum likelihood estimators, and we find expressions for the score and the information matrix and use them to find the asymptotic distribution of the maximum likelihood estimator. In §3.3 we analyse the likelihood ratio test of a linear hypothesis on the regression parameters, both with the same ARMA parameters, and provide a new stochastic expansion for the likelihood ratio test of a reduced model.

In §3.4 we recall the prediction error decomposition and prove that the distribution of sum of squares of one step ahead standardized prediction errors, when the ARMA parameters are estimated, differs from the chi-squared distribution by a term which tends to infinity at a lower rate than χ^2_n. We also prove that the term involving the sum of logarithms of the variances of one step ahead standardized prediction errors is convergent, leading to a simplified asymptotic form for the likelihood ratio test. In §4 we sketch the extension of our results to non-stationary models.

In §5 we address the issue of model selection for linear regression with ARIMA errors. We propose our new C_p for time series and use the results of §3 to find its asymptotic distribution. Theoretical arguments and simulations indicate that the distribution of the new statistic is well approximated by an F distribution.

In §6 we provide a small simulation study and compare a consistent version of our statistic with BIC and a generalized RIC (Shi and Tsai, 2004) extended to ARMA models. An important feature of our new statistic is that, as with C_p in regression, an error variance is estimated from a large model. Our calculations show that, as expected, use of this estimate provides a statistic with superior performance, particularly for small signal to noise ratios. Finally we suggest a new plot for our time series C_p statistic which is highly informative about the choice of model. Theory and simulations show that the plot comes with a banded structure which easily enables us to appreciate the effect of the introduction of additional explanatory variables and/or stochastic parameters. Section 7 concludes and provides food for thought for further research.

2 AIC, C_p and Likelihood Ratio Tests

2.1 Regression with i.i.d. Errors

It is often stated that the use of AIC in the selection of Gaussian models is equivalent to the use of C_p (for example, Hastie *et al.*, 2009, p. 231) and a Taylor series justification is sometimes given: Venables and Ripley (2002, §6.6), Davison (2003, problem 8.16). We now establish an exact relationship for normal theory regression.

We first consider regression without a time-series structure. For the linear multiple regression model we assume that y has been generated by the unknown model $y = X_0\beta_0 + \epsilon$ where $\epsilon \sim N(0\ \sigma^2 I)$, $X_0 = (x_{10}, ..., x_{n0})'$. We estimate the model $y = X\beta + \epsilon$, where X is an $n \times d$ full-rank matrix of known constants, with ith row x_i'. The normal theory assumptions are that the errors ϵ_i are i.i.d. $N(0, \sigma^2)$. To estimate σ^2 we may also regress y on all d^+ columns of the $n \times d^+$ matrix X^+.

Assumption 1. *Let $\mathcal{C}(X)$ be the column space of X. Null distributional results are obtained when $\mathcal{C}(X_0) \subseteq \mathcal{C}(X) \subset \mathcal{C}(X^+)$.*

The log-likelihood of n observations y, a function of σ^2 and of the $d \times 1$ vector of parameters β is $L_n(\beta, \sigma^2)$. If $\hat{\beta}$ is the maximum likelihood estimator of β, AIC is often defined, particularly in the time series context, as

$$AIC = -2L_n(\hat{\beta}; y) + 2(d + 1), \tag{1}$$

since we have estimated σ^2 as well as β. That model is selected for which AIC is a minimum. It would be natural to use p as the number of parameters, but this is a paper about the analysis of time series and our notation is intended to allow for the discussion of general ARMA(p, q) models with regressors.

The residual sum of squares from fitting this model to the data is R_d and, for known σ^2,

$$AIC_\sigma = n \log(2\pi) + n \log \sigma^2 + R_d/\sigma^2 + 2d. \tag{2}$$

If, as is usually the case, σ^2 is not known, the maximum likelihood estimator is

$$\hat{\sigma}^2 = R_d/n. \tag{3}$$

With this internal estimate of σ^2 the criterion (2) is replaced by

$$AIC_I = n \log(2\pi) + n \log\{R_d/n\} + n + 2(d + 1), \tag{4}$$

a form frequently used in the selection of non-nested time series models with normally distributed errors (Tong, 2001, §9).

In the selection of regression variables, the unbiased estimator of σ^2 comes from regression on all d^+ columns of X^+ and can be written

$$s^2_{d^+} = R_{d^+}/(n - d^+). \tag{5}$$

With this estimate the criterion (2) is

$$AIC_{d^+} = n\log(2\pi) + n\log\{R_{d^+}/(n-d^+)\} + (n-d^+)R_d/R_{d^+} + 2(d+1), \tag{6}$$

although (4) is the standard form for AIC. The main difference between AIC_{d^+} and AIC_I is in the estimate of σ^2 which is used. In the context of model choice, both n and $s^2_{d^+}$ are fixed, the variable factors being the value of d and the regressors that are being considered. Then choice of the model minimizing (6) is identical to the choice of model minimizing

$$C_p = R_d/s^2_{d^+} - n + 2d = (n - d^+)R_d/R_{d^+} - n + 2d. \tag{7}$$

One interpretation of C_p (Mallows, 1973) is that its value provides an estimate of the scaled mean squared error of prediction at the n observational points from the model of interest, provided Assumption 1 holds. Then $\mathrm{E}\{R_d\} = (n-d)\sigma^2$, $\mathrm{E}(s^2_{d^+}) = \sigma^2$ and $\mathrm{E}(C_p)$ is approximately d.

If Assumption 1 holds, an assumption we make for the rest of the paper, both $R_{d^+}/(n-d^+)$ and $R_d/(n-d)$ provide consistent estimates of σ^2. However, if $\mathcal{C}(X_0) \not\subset \mathcal{C}(X)$, R_d contains a non-centrality parameter and, because $s^2_{d^+}$ is unbiased, C_p (7) will be large and the model will be rejected.

As we illustrate in §6, it helps not merely to select models with small values of C_p but also to calibrate those values against their distribution. The distribution of C_p, under the null Assumption 1, is given, for example, by Mallows (1973) and by Gilmour (1996). From (7) we require the distribution of the ratio of two nested residual sums of squares. It is straightforward to show that the required distribution is

$$C_p \sim (d^+ - d)F + 2d - d^+, \qquad \text{where} \qquad F \sim F_{d^+-d,n-d^+}. \tag{8}$$

In short, if $F^* \sim F_{\nu_1,\nu_2}$, $\mathrm{E}\,(F^*) = \nu_2/(\nu_2 - 2)$ and, from (8),

$$\mathrm{E}\,(C_p) = d + 2\frac{d^+ - d}{n - d^+ - 2}. \tag{9}$$

As $n \to \infty$, $\mathrm{E}\,(C_p) \to d$. Hurvich and Tsai (1989) find corrections for the bias $2(d^+ - d)/(n - d^+ - 2)$ for very small n while Fujikoshi and Satoh (1997) give modifications to AIC and C_p to reduce bias when the candidate models may be under- or over-specified and so Assumption 1 fails.

In Appendix A we prove similar results for our new version of AIC and show that it retains the same interpretation as traditional AIC in providing a consistent estimator of the expected Kullback-Leibler information between the true model and the fitted candidate model. We also give in the following theorem the expression for the bias corrected version of our new AIC – say $AIC_d^+(u)$ – which is

Theorem 1. *The unbiased version of the new AIC suggested in equation* (6) *is given by*

$$AIC_d^+(u) = n\log(2\pi)+n\log\{R_{d^+}/(n-d^+)\}+(n-d^+)R_d/R_{d^+}+2(d+1)\frac{n-d^+}{n-d^+-2}. \tag{10}$$

The model minimizing AIC or C_p has a fixed probability of being too large, even as $n \to \infty$. Schwarz (1978) shows that replacing $2(d+1)$ in (6) by $(d+1)\log n$ provides consistent model selection. The alternative of $(d+1)\log(n-d^+)$ used in the RIC of Shi and Tsai (2004) is obviously also consistent. Such consistency factors can also be applied to our new AIC.

3 Advances in linear regression with ARIMA errors

We now provide the necessary distributional results to justify an asymptotic F distribution for our time series C_p, thus providing a natural extension of (8). Proofs of the results are in Appendix B.

3.1 The model and the assumptions

We consider the regression model with stationary invertible ARMA errors u_t and deterministic regressors

$$y_t = \beta' x_t + u_t,\ t = 1, \ldots, n. \tag{11}$$

$$A_p(L)u_t = B_q(L)\varepsilon_t,\ \varepsilon_t \text{ i.i.d. } N(0,\sigma^2),\ t = 1, \ldots, n. \tag{12}$$

The polynomials are $A_p(z) = 1-\sum_{i=1}^p \phi_i z^i, \phi_0 = 1$ and $B_q(z) = \sum_{i=0}^q \psi_i z^i, \psi_0 = 1$, and we define the parameters $\theta = (\phi_1, \ldots, \phi_p, \psi_1, \ldots, \psi_q)$. The matrices of regressors are as in §2.1. We use the notation $\Omega = \Omega_n(\theta) = \sigma^{-2}Var(y)$, and define the autocovariance function $\gamma(h) = \sigma^{-2}Cov(y_t, y_{t+h}) = \Omega_{t,t+h}$. Note that Ω is $n \times n$ and therefore depends on n and the dynamic parameters θ.

The Gaussian log likelihood function is

$$-2\log L_n(\beta,\theta,\sigma^2) = n\log\sigma^2 + \log|\Omega| + \sigma^{-2}(y-X\beta)'\Omega^{-1}(y-X\beta). \quad (13)$$

In accordance with Yao and Brockwell (2006) we make the following assumption.

Assumption 2. *Let the roots of $A_p(z)$ and $B_q(z)$ be denoted $z_1(\theta),\ldots,z_{p+q}(\theta)$ and let $\theta \in \mathcal{D}$ be a compact subset of the set where $A_p(z)$ and $B_q(z)$ have no common factors and where $A_p(z)$ and $B_q(z)$ are invertible, that is, there is a $\rho<1$ for which*

$$\min_{\theta\in\mathcal{D}} \min_{1\le i\le p+q} |z_i(\theta)| \ge \rho^{-1}\,. \quad (14)$$

Together with model (11) we also consider the regression with nonstationary error term where (12) is replaced by

$$A_p(L)\Delta^d u_t = B_q(L)\varepsilon_t,\ \varepsilon_t \text{ i.i.d. } N(0,\sigma^2),\ t=1,\ldots,n,\ d=1,2 \quad (15)$$

with Assumption 2 still satisfied.

Under Assumption 2 we apply the representation

$$u_t = \frac{B_q(L)}{A_p(L)}\varepsilon_t = \sum_{n=0}^{\infty}\eta_n\varepsilon_{t-n},\ \eta_0=1,$$

see Lemma 1 in Appendix B for the properties of the coefficients and for various evaluations of Ω and its derivatives.

In what follows we give our results in terms of asymptotic inference, hypothesis testing and prediction error decomposition.

3.2 Asymptotic inference

The asymptotic distribution of the Gaussian maximum likelihood estimator of the ARMA parameter vector θ and the regression parameters $\hat\beta$, has been given by Hannan (1973) for stationary regressors using frequency domain methods. Yao and Brockwell (2006) analyse the model without regressors and show that $\hat\theta$ is consistent and asymptotically Gaussian using time domain methods. We show here how the results of Yao and Brockwell can be applied and extended to the regression model with deterministic regressors and possibly nonstationary ARMA errors.

We denote the true values of the parameters by β_0, σ_0^2 and Ω_0 and introduce

$$\kappa = (n^{-1}X'\Omega_0^{-1}X)^{1/2}(\beta-\beta_0), \quad (16)$$

with true value $\kappa_0 = 0$.

Theorem 2. *Under Assumption 2, the maximum likelihood estimators exist and are strongly consistent*

$$(\hat{\sigma}^2, \hat{\theta}, \hat{\kappa}) \xrightarrow{a.s.} (\sigma_0^2, \theta_0, 0).$$

It follows that if $\lambda_{\min}(n^{-1}X'\Omega_0^{-1}X) \geq a > 0$, *then* $\hat{\beta} \xrightarrow{a.s.} \beta_0$.

Now let $\mathsf{D}^m f(\theta)$ denote the m'th derivatives of f with respect to the arguments of θ, with the $n \times n$ derivatives $\dot{\Omega}_s = \mathsf{D}_{\theta_s}\Omega$ and $\ddot{\Omega}_{sk} = \mathsf{D}^2_{\theta_s\theta_k}\Omega$. Further, in the following we denote by $\{A, B, C\}$ the block diagonal matrix with A, B, C in the diagonal. Then we can write the following theorem about asymptotic distributions, see Hannan (1973) Theorem 2, and Yao and Brockwell (2006) Theorem 2.

Theorem 3. *Let* $\lambda = (\sigma^2, \theta, \kappa)$. *Under Assumption 2, the score function* $S_{n\lambda} = n^{-1/2}\mathsf{D}\log L(\lambda_0) = (S_{n\sigma^2}, S'_{n\theta}, S'_{n\kappa})'$ *is asymptotically Gaussian with covariance*

$$\lim_{n\to\infty} E(-n^{-1}\mathsf{D}^2 \log L(\lambda_0)) = \{\frac{1}{2}\sigma_0^{-4}, \Sigma_0, \sigma_0^{-2}I_d\}, \tag{17}$$

where Σ_0 *is* $(p+q) \times (p+q)$ *with elements*

$$\Sigma_{0sk} = \lim_{n\to\infty} \frac{1}{2} n^{-1} tr\{\Omega_0^{-1}\dot{\Omega}_{0k}\Omega_0^{-1}\dot{\Omega}_{0s}\}.$$

Moreover for a sequence $\varepsilon_n \to 0$ *we have for the information per observation* $I_{n\lambda\lambda}(\lambda) = -n^{-1}\mathsf{D}^2 \log L(\lambda)$,

$$\max_{|\lambda-\lambda_0|\leq\varepsilon_n} ||I_{n\lambda\lambda}(\lambda) - I_{n\lambda\lambda}(\lambda_0)||_2 = O(\varepsilon_n). \tag{18}$$

The logical implication of Theorem 3 is that $\hat{\beta}$ is asymptotically independent of $(\hat{\sigma}^2, \hat{\theta})$ and therefore estimates of θ under two different models for β satisfy $\hat{\theta} - \hat{\theta}_* = O_P(n^{-1})$, $\hat{\sigma}^2 - \hat{\sigma}^2_* = O_P(n^{-1})$. These two key facts are stated more formally in the two following corollaries.

Corollary 4. *Under Assumption 2, the maximum likelihood estimators*

$$(n^{1/2}(\hat{\sigma}^2 - \sigma_0^2), n^{1/2}(\hat{\theta} - \theta_0), (X'\Omega_0^{-1}X)^{1/2}(\hat{\beta} - \beta_0))$$

are asymptotically Gaussian and asymptotically independent with an asymptotic variance given by

$$\{2\sigma_0^4, \Sigma_0^{-1}, \sigma_0^2 I_d\}.$$

Corollary 5. *For two models* $E(y) = X\beta$ *and* $E(y) = XA\xi$, *or* $\beta = A\xi$, *with estimates* $(\hat{\sigma}^2, \hat{\theta}, \hat{\beta})$ *and* $(\hat{\sigma}^2_*, \hat{\theta}_*, \hat{\beta}_* = A\hat{\xi}_*)$ *respectively we have* $(\hat{\sigma}^2 - \hat{\sigma}^2_*, \hat{\theta} - \hat{\theta}_*) = O_P(n^{-1})$.

3.3 Hypothesis testing

In the model with $E(Y) = X\beta, \Omega_n = \Omega_n(\theta), \beta \in \mathbb{R}^d, \theta \in \mathbb{R}^{p+q}$ with estimators $(\hat{\sigma}^2, \hat{\theta}, \hat{\beta})$, we want to test the hypothesis $E(Y) = XA\xi$, or equivalently $\beta = A\xi$, $\xi \in \mathbb{R}^{d_*}$, $d_* < d$. Under this hypothesis the estimators are $(\hat{\sigma}^2_*, \hat{\theta}_*, \hat{\beta}_* = A\hat{\xi}_*)$.

Let $\hat{\beta}, \hat{\theta}, \hat{\sigma}^2, \hat{\Omega}$ denote the maximum likelihood estimators in model (11). We want to test the hypothesis that $\beta = A\hat{\xi}$ and denote the maximum likelihood estimators under this restriction by $\hat{\beta}_* = A\hat{\xi}_*, \hat{\theta}_*, \hat{\sigma}^2_*, \hat{\Omega}_*$.

The theorem below shows that under the null hypothesis $\beta = A\xi$ the fact of estimating the covariance matrix with and without this restriction leads to an error which is of order $O_P(n^{-1})$.

Theorem 6. *For Ω and $\hat{\Omega}$ we have*

$$(y - X\hat{\beta}_*)'(\hat{\Omega}^{-1} - \hat{\Omega}_*^{-1})(y - X\hat{\beta}_*) = O_P(n^{-1}). \tag{19}$$

The theorem below provides a similar result for the log of the ratio of the estimates of the scale parameter of the covariance matrix.

Theorem 7. *It follows from asymptotic theory of the maximum likelihood estimator that $-2\log LR(\beta = A\xi) \xrightarrow{D} \chi^2(d - d_*)$, but we also have the stochastic expansion*

$$-2\log LR(\beta = A\xi) = n\log\frac{\hat{\sigma}^2_*}{\hat{\sigma}^2} + O_P(n^{-1}) = \hat{\sigma}^{-2}(\hat{\beta}_* - \hat{\beta})'X'\hat{\Omega}^{-1}X(\hat{\beta}_* - \hat{\beta}) + O_P(n^{-1}).$$

3.4 The prediction error decomposition

In the extension of results on C_p and its distribution to time series with a Gaussian structure it is convenient to use the state-space representation (Anderson and Moore, 1979; Durbin and Koopman, 2012) and the Kalman filter for calculation of the likelihood associated with each model. In this section we briefly recall the prediction error decomposition and provide new results and insight about the distribution of its terms.

The best linear prediction of y_t is

$$\hat{y}_t = E_{t-1}(y_t) = x_t'\beta + E_{t-1}(u_t),\ t = 2, \ldots, n, \text{ and } E(y_1) = x_1'\beta,$$

where the subscript indicates expectation conditional on $y_1, \ldots, y_{t-1}$. The prediction error v_t is

$$v_t = y_t - \hat{y}_t,\ t = 2, \ldots, n, \text{ and } y_1 - x_1'\beta$$

and the variance of the prediction error, with σ^2 concentrated out, defines the factor $f_t^c = f_t^c(\theta)$ by

$$Var_{t-1}(y_t) = \sigma^2 f_t^c(\theta),\ t = 2, \ldots, n, \text{ and } Var(y_1) = \sigma^2\gamma(0).$$

Decomposing the density in successive conditional densities we get the prediction error decomposition

$$p(y_1,\ldots,y_n) = \prod_{t=1}^{n} p(y_t|y_1,\ldots,y_{t-1}) \tag{20}$$
$$= \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2 f_t^c}} \exp(-\frac{1}{2\sigma^2}\frac{(y_t-\hat{y}_t)^2}{f_t^c})$$

and hence the identities, see (13),

$$\sum_{t=1}^{n} \frac{(y_t-\hat{y}_t)^2}{f_t^c} = (y-X\beta)'\Omega^{-1}(y-X\beta) \text{ and } \sum_{t=1}^{n} \log f_t^c = \log|\Omega_n|.$$

The log likelihood can therefore be rewritten as:

$$-2\log L(\beta,\hat{\theta},\hat{\sigma}^2) = n\log(2\pi) + n\log\hat{\sigma}^2 + \sum_{t=1}^{n}\log f_t^c(\hat{\theta}) + \hat{\sigma}^{-2}\sum_{t=1}^{n}\frac{v_t^2}{f_t^c(\hat{\theta})}$$

The expression for the log likelihood can be further simplified to

$$-2\log L(\beta,\hat{\theta},\hat{\sigma}^2) = n\log(2\pi) + n\log\hat{\sigma}^2 + \sum_{t=1}^{n}\log f_t^c(\hat{\theta}) + n, \tag{21}$$

using the estimate

$$\hat{\sigma}^2 = n^{-1}\sum_{t=1}^{n}\frac{v_t^2}{f_t^c(\hat{\theta})}.$$

Equation (21) is known in the literature as the concentrated profile log likelihood (see for example Francke *et al.* (2010), for a discussion of alternative specifications for the likelihoods). The prediction error decomposition (Harvey, 1989, eq. 3.4.7) yields a particularly simple form for the likelihood since the quantities v_t and f_t^c can easily be calculated putting model (11) in the so-called state space form and applying the Kalman filter.

Computational remark: when β is estimated we need to run the Kalman filter with an additional set of recursions which are usually referred to in the literature as the diffuse Kalman filter (de Jong, 1991).

The new result of this subsection refers to the behaviour of the two terms

$$\sum_{t=1}^{n}\log f_t^c(\hat{\theta}) \quad \text{and} \quad \sum_{t=1}^{n} v_t^2/f_t^c(\hat{\theta}). \tag{22}$$

The second term is the sum of squares of independent one-step ahead prediction errors for normal random variables, scaled by their variances. If all dynamic parameters in the model are known and the current model is correct, this sum has exactly a $\sigma^2\chi^2_{n-d}$ distribution, because the v_t are linear functions of the observations and are independent with mean zero and variance $\sigma^2 f_t^c(\theta)$ (e.g. Harvey, 1989).

However, when the model also requires estimation of the parameters θ in the ARIMA model or, equivalently, the estimation of the variances of the disturbances in the structural framework, the distribution, as we state in the theorem below, is asymptotically chi-squared, differing from the χ^2_{n-d} distribution by a term that tends to infinity at a lower rate than χ^2_{n-d}.

In what follows we use the symbol $\tilde{\cdot}$ to denote an estimate in which only the regression parameters are estimated whilst using $\hat{\cdot}$ when both the regression coefficients and the stochastic parameters are unknown and estimated.

Assumption 3. *In addition to Assumption1 of §2.1, we also require that the models have $p \geq p_0$ and $q \geq q_0$.*

Theorem 8. *Let the data be generated with additive errors u_t such that $a_{\phi_0,p_0}(L)u_t = b_{\theta_0,q_0}(L)\varepsilon_t$. We fit model (11). Under Assumptions 1 and 3 this will contain the true model, so that $\beta_0' z_{0t} = \beta' x_t$ for some β. Then the residual sum of squares satisfies*

$$\hat{u}'\hat{\Omega}^{-1}\hat{u} = (y - X\hat{\beta})\hat{\Omega}^{-1}(y - X\hat{\beta}) = \tilde{u}'\Omega_0^{-1}\tilde{u}\{1 + O_P(n^{-1/2})\},\ n \to \infty. \quad (23)$$

The result shows that the sum of squares of one step ahead prediction errors divided by their scaled variances when all parameters are unknown is equal to the sum of squares of the residuals which we obtain when only the regression parameters are estimated, apart from a term which is of order $n^{1/2}$. That is

$$\hat{u}'\hat{\Gamma}^{-1}\hat{u} = \sum_{t=1}^{n} \frac{\hat{v}_t^2}{\hat{f}_t^c} = \sum_{t=1}^{n} \frac{\tilde{v}_t^2}{\tilde{f}_t^c}\{1 + O_P(n^{-1/2})\} = \tilde{u}'\Gamma_0^{-1}\tilde{u}\{1 + O_P(n^{-1/2})\}. \quad (24)$$

The effect of estimation of the stochastic parameters can accordingly be summarized as

$$(y - X\hat{\beta})'\hat{\Omega}^{-1}(y - X\hat{\beta}) = \chi^2(n - d) + O_P(n^{1/2}). \quad (25)$$

For the first term in (22) we have

Theorem 9. *The limits of $\log|\Omega_n(\theta)| = \sum_{t=1}^{n} \log f_t^c$ and its derivatives $\partial \log|\Omega_n(\theta)|/\partial\theta_s, s = 1, \ldots, p+q$ exist as continuous functions*

$$0 \leq \lim_{n\to\infty} \log|\Omega_n(\theta)| < \infty,$$
$$\lim_{n\to\infty} \mathsf{D}_s \log|\Omega_n(\theta)| = \mathsf{D}_s \lim_{n\to\infty} \log|\Omega_n(\theta)|.$$

4 Extensions to non stationary models

For ARIMA models the error term u_t and hence y_t is nonstationary but both can be differenced until stationarity is obtained. The differenced model then includes differenced x_t. In addition we cover the class of "structural" models, so-called because they have easily interpreted parameters for modelling economic times series. Harvey (1989, Appendix 1 and equation (2.4.26)) summarizes the ARIMA form of these models by use of a multivariate error term. Our results for asymptotic inference require a univariate error term, in which case the $p+q$ parameters θ in (12) can be restricted so it becomes a nonlinear function of the structural parameters: $\theta = \theta(\xi)$, where $\dim \xi \leq \dim \theta$. Using the methods developed in this paper it is possible to obtain the asymptotic theory for the maximum likelihood estimator in the structural model which is just the maximum likelihood estimator in the restricted ARIMA model.

To see this more formally, we start by noticing that, for example for $d = 1$, equations (15) can be written in the equivalent form

$$\begin{aligned} y_1 &= \beta' x_1 + u_0 + v_1, t = 1, \\ \Delta y_t &= \beta' \Delta x_t + v_t, t = 2, \ldots, n, \end{aligned} \tag{26}$$

where $v_t = \Delta u_t$, is a stationary ARMA(p, q) process and $u_t = u_0 + \sum_{i=1}^{t} v_i$ where u_0 is an initial value independent of $v_1, \ldots, v_t$. The transformation of the data to the last $n-1$ equations, that is equations for $\Delta y_t, t = 2, \ldots, n$ means ignoring the first equation because we can write the density of $y_1, y_2, \ldots, y_T$ conditional on u_0 as

$$\begin{aligned} p(y_1, y_2, \ldots, y_T | u_0) &= p(y_1, \Delta y_2, \ldots, \Delta y_T | u_0) \\ &= p(\Delta y_2, \ldots, \Delta y_T) p(y_1 | \Delta y_2, \ldots, \Delta y_T, u_0) \end{aligned}$$

In the Gaussian case the last factor is a Gaussian distribution with mean

$$E(y_1 | \Delta y_2, \Delta x_t \ldots, \Delta y_T, u_0) = \beta' x_t + u_0 + E(v_1 | \Delta y_2, \ldots, \Delta y_T),$$

and variance $Var(v_1 | \Delta y_2, \ldots, \Delta y_T)$. This reduces the problem of inference in the regression model with nonstationary ARIMA errors (15) to the case of regression of Δy_t on Δx_t with stationary ARMA errors in (26).

If instead we consider (15) for $d = 2$ we define $v_i = \Delta^2 u_i$ and find $u_t = u_0 + t\Delta u_0 + \sum_{i=1}^{t} (t + 1 - i) v_i$. Then the equations are

$$\begin{aligned} y_1 &= \beta' x_1 + u_0 + \Delta u_0 + v_1 \\ y_2 &= \beta' x_2 + u_0 + 2\Delta u_0 + 2v_1 + v_2 \\ \Delta^2 y_t &= \beta' \Delta^2 x_t + v_t, t = 3, \ldots, n, \end{aligned}$$

and a similar argument can be made for focussing on the last $n-2$ equations.

5 Model selection for linear regression with stationary ARIMA errors

5.1 Model selection

In this section we extend the criteria from §2 developed for i.i.d. observations to ARMA models with explanatory variables. We continue to consider models of the form (11) under Assumptions 1 and 3. Then the model defined by $\mathcal{M}_{X_+} = (\theta \in \mathbb{R}^{p+q}, E(y) = X_+\beta_+, \beta_+ \in \mathbb{R}^{d_+}, \sigma^2 > 0)$ contains the data generating process because for $X_0 = X_+A$ we can take $\beta_{+0} = A\beta_0$.

We search over regression models defined by $\mathcal{M}_X = (\theta \in \mathbb{R}^{p+q}, E(y) = X\beta, \xi \in \mathbb{R}^d, \sigma^2 > 0)$, for which $\mathcal{C}(X) \subseteq \mathcal{C}(X^+)$. We denote the maximum likelihood estimators by $(\hat\theta, \hat\beta, \hat\sigma^2)$. We also fit the model $\mathcal{M}_{X_+}$ and denote the estimators by $(\hat\theta_+, \hat\beta_+, \hat\sigma^2_+)$.

If σ^2 and θ are not estimated we get the analogue of (2)

$$AIC_{\sigma,\theta} = n\log(2\pi) + n\log\sigma^2 + \log|\Omega| + \sigma^{-2}(y - X\hat\beta)'\Omega^{-1}(y - X\hat\beta) + 2c,$$

where $c = d$ and $\hat\beta = (X'\Omega^{-1}X)^{-1}X'y$.

If σ^2 and θ are estimated we get the analogue of (4)

$$AIC_I = n\log(2\pi) + n(\log\hat\sigma^2 + 1) + \sum_{t=1}^{n}\log f_t^c(\theta(\hat\kappa)) + 2c, \tag{27}$$

where $c = 1 + d + k, k = p + q$ and

$$\hat\sigma^2 = (n-d)^{-1}(y - X\hat\beta)'\Omega(\theta(\hat\kappa))^{-1}(y - X\hat\beta). \tag{28}$$

Finally the analogue of AIC_{d_+} (6) uses an estimator of σ^2 and Ω based upon the model with d_+ regressors X_+ and unrestricted θ :

$$AIC^T_{d+} = n\log(2\pi) + n\log\hat\sigma^2_+ + \log|\hat\Omega_+| + n\hat\sigma^2\hat\sigma_+^{-2} + 2c, \tag{29}$$

where $c = 1 + d + k$, and

$$\begin{aligned}\hat\sigma^2 &= (n-d)^{-1}(y - X\hat\beta)'\hat\Omega^{-1}(y - X\hat\beta),\\ \hat\sigma^2_+ &= (n-d^+)^{-1}(y - X_+\hat\beta_+)'\hat\Omega_+^{-1}(y - X_+\hat\beta_+).\end{aligned} \tag{30}$$

Our simulations show, in complete agreement with the theorems developed in the previous section, that (29) has a distribution close to that of AIC_I.

We now derive a statistic with an asymptotically known distribution. The choice of the model minimizing (29) is identical to the choice of model minimizing

$$C_p^T = \log|\hat\Omega_+| + (n-d)\hat\sigma^2\hat\sigma_+^{-2} - n + 2c. \tag{31}$$

In this criterion the term $\log|\hat{\Omega}_+|$ is $O_P(1)$, so for selection purposes we focus on

$$C_p = \frac{(y - X\hat{\beta})'\hat{\Omega}^{-1}(y - X\hat{\beta})}{\hat{\sigma}^2_+} - n + 2c, \tag{32}$$

where $c = 1 + d + k$, is the number of fitted parameters $(\beta, \sigma^2, \theta)$ in the model with d regressors X and dynamic parameters θ of dimension k.

The simulation results of Tables 1 and 2 show the effect of this omission, which is beneficial when regression is relatively weak. For the full model with c^+ parameters, C_p has the exact value $2c^+ - d^+$, since we have the same sum of squares in the numerator and denominator of (32). Distributional results about C_p rely on the asymptotic χ^2 distribution of the residual sums of squares (25) for models satisfying our assumptions. From (32) the approximate distribution of C_p is given by

$$C_p \ \sim \ (c^+ - c)F + 2c - d^+ \qquad \text{where} \qquad F = F_{c^+-c,n-d^+}. \tag{33}$$

and that

$$E(C_p) = c + \frac{(c^+ - d^+)(n - d^+)}{n - d^+ - 2} + \frac{2(d^+ - c)}{n - d^+ - 2}.$$

As $n \to \infty$ we obtain

$$E(C_p) \to c + (c^+ - d^+) = c + p^+ + q^+.$$

Thus the expected value of the statistic, for large n, depends on the total number of parameters in the reduced model and on the number of stochastic parameters in the full model. This, however, is a constant when comparing different reduced models, so that the penalty in comparisons is just c, as it is d for regression models. In neither case does the parameter for the error variance, which is concentrated out in the time series application, affect the F distribution of the statistic. However, a consequence of taking $c = 1 + d + k$ in the time series formulation is that the two statistics differ by a constant value of 2 when the errors are independent.

Application of the prediction error decomposition to (12) of Shi and Tsai (2004) yields the generalization

$$\begin{aligned} RIC^G = (n-c)\log\sigma^2_* + \sum_{t=d+1}^{n} \log \hat{f}^c_t + c\log\{n - (p+q)\} - \{2(p+q) + d\} \\ + \ \frac{4}{n - 2(p+q) - d - 2}, \end{aligned} \tag{34}$$

where σ^2_* is estimated from the current model.

6 C_p Plots and Empirical Performance Using Simulated Data

To compare the performance of our new statistic with those in the literature we report the results of a small simulation experiment. We included our new C_p statistic (32) and C_p^T (31) from which it was derived in both their original and consistent forms in which $2c$ is replaced by $c\log(n-c)$. Also included are AIC, BIC and our generalization of RIC, called RIC^G, to general state-space models. The seven statistics are

C_p. The F-distributed statistic (32).
B_p. Consistent C_p; (32) with $2c$ replaced by $c\log(n-c)$.
C_p^T. C_p including the sum of terms $\log \hat{f}_t^c$ (31).
B_p^T. Consistent C_p^T; (31) with $2c$ replaced by $c\log(n-c)$.
AIC_I. Equation (27) with $\hat{\sigma}^2$ given by (28).
BIC_I. Consistent AIC_I; (27) with $2c$ replaced by $c\log(n-c)$.
RIC^G. Generalized RIC; (34).

The comparative performance of the seven statistics depends on the signal to noise ratio. The signal comes from the matrix of explanatory variables X^+ generated, once for each table, from standard normal random variables with the values of β equal to one. Following Shi and Tsai (2004), we take as the numerator of the ratio the average variance of the mean function of the data-generating model. Here, with k variables each with variance one, the mean equals one. We vary the variance of the errors var(ϵ_t) from one to 200, so that the range of the signal to noise ratio is 1 to 0.005. The results for an MA model are in Table 1 and, for an AR model, in Table 2.

Table 1: Percentage of time true model (MA(1) + 2 explanatory variables) is chosen. Results of 1,000 simulations with $n = 200$. Full model ARMA(2,2) + 4 explanatory variables. Signal to noise ratio = 1/var(ϵ_t)

var(ϵ_t)	C_p	B_p	C_p^T	B_p^T	BIC	RIC^G	AIC
100	16	5	0	0	7	0	7
50	37	25	8	2	16	1	22
25	45	71	34	27	54	23	43
10	48	91	64	93	92	65	74
1	51	90	64	94	92	94	80

The results in the two tables are surprisingly similar. When the signal to noise ratio is large, var(ϵ_t) $= 1$, the best performance in terms of automatic model selection is for B_p^T, BIC and RIC^G. If the signal to noise ratio decreases to 0.1

Table 2: Percentage of time true model (AR(2) + 1 explanatory variable) is chosen. Results of 1,000 simulations with $T = 200$. Full model ARMA(2,2) + 3 explanatory variables. Signal to noise ratio = 1/var(ϵ_t)

var(ϵ_t)	C_p	B_p	C_p^T	B_p^T	BIC	RIC^G	AIC
200	18	10	0	1	1	0	1
150	24	12	0	1	1	0	2
100	31	20	2	1	3	0	4
50	45	44	10	4	17	2	10
20	56	76	36	28	47	21	32
10	57	91	73	92	89	76	67
1	57	92	74	95	94	94	85

(var(ϵ_t) $= 10$), while B_p^T and BIC still show high values, those of RIC^G rapidly decrease. On the other hand, when the ratio is small (say not greater than than 0.02, that is var(ϵ_t) $= 50$), C_p significantly outperforms all other statistics, even though it is not consistent. The consistent version of our statistic, B_p, is good, although not best, over the whole range. A conclusion from the tables is that there is much to be gained, over the whole range of signal to noise ratios, from using statistics based on $\hat{\sigma}_+^2$ (30).

We now show how our C_p statistic can be used to provide an informative plot for the selection of time series models keeping in mind that, in our opinion, the mechanical use of C_p is to be avoided.

Table 3: Notation used in the figures for ARMA models with regressors

Notation	Model and Regressors
$a0$	AR(1)
$0a$	MA(1)
$0b$	MA(2)
$aa34$	ARMA(1,1) x_3 x_4

We look at the structure of plots of C_p for time series in a simulated example. We label the models with a notation of the form "$p\,q\,i_1\,i_2...$", where p and q denote the order of the autoregressive and moving average models and the i_j denote those regression variables that are included in the model. Further, we denote the increasing values of p and q as $0, a, b$ etc. Some examples are in Table 3.

We simulated 100 observations from an MA(1), that is $0\,a$, with $\theta = 0.9$, that

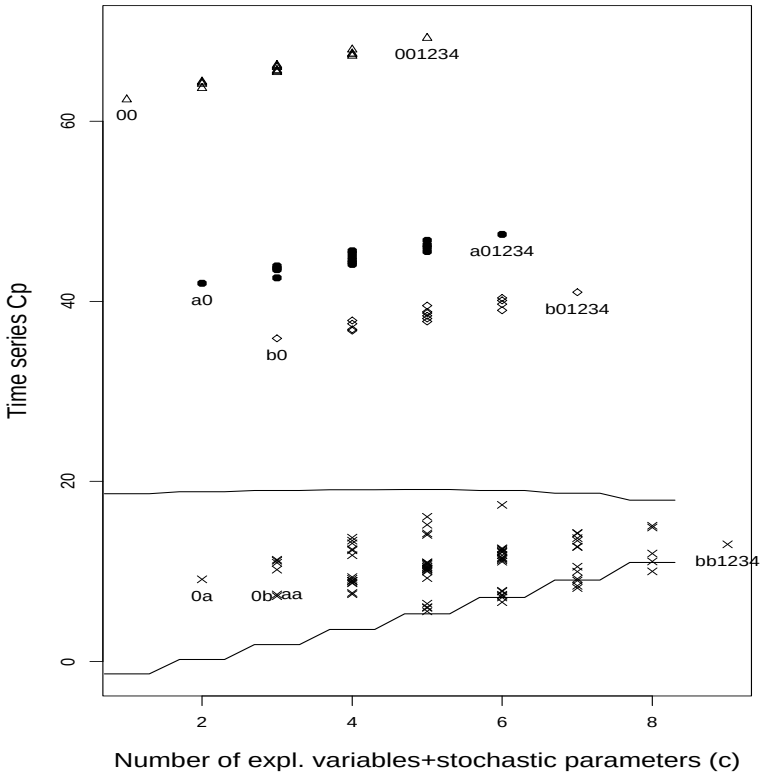

Figure 1: Plot of C_p statistic for time series for simulated MA(1) process (0 a) with $\theta = 0.9$. Large model $b\,b\,1\,2\,3\,4$ plus a constant (c^+ = 10). $\times$ MA and ARMA models with explanatory variables; $\diamondsuit$ AR(2) models with explanatory variables; $\bullet$ AR(1) models with explanatory variables; $\triangle$ regression models. See Table 3 for notation. Bands 1% and 99% points of (33). The importance of selecting the correct stochastic model is evident

included a constant, equal to 5, and four explanatory variables that were unrelated to the time series. The explanatory variables were independently distributed $N(0, 1)$ and σ^2 = 1. In our calculation of C_p the maximum model was $b\,b\,1\,2\,3\,4$, all models containing a constant. So the maximum number of parameters $c^+ = 9$. Figure 1 shows the resulting plot of C_p for all models containing at least two parameters.

The times series C_p plot shows four bands of values corresponding to different families of stochastic models. The family with smallest C_p values, marked with crosses in the figure, falls within the band of the 1% and 99% points of the F distribution (33). The simplest model is the MA(1) without explanatory variables, that is 0 a, from which the data were simulated. Reading upwards, the next two models in this band, with three parameters, are $a\,a$ and 0 b, with C_p values around

two lower than that of their special case $0\,a$. The remaining four models, with higher C_p values, are $0\,a\,i$, i = 1,...,4; that is MA(1) models including one of the explanatory variables. The models with more parameters in this group ($c \geq 4$) are all at least ARMA(1,1) or MA(2) with explanatory variables. For the maximum model with $c^+ = 9$, the value of C_p is 13, agreeing with the special case of (32).

The second series of C_p values in the plot, shown by diamonds, are for AR(2) models including explanatory variables. The next band is for AR(1) models also including such variables. The highest band of all, the triangles, is for pure regression models without any time series component.

An informative feature of this plot is that the bands sort the models into clear groups with differing stochastic structure. It is clear from the figure that we need at least an MA(1) model and that the improvements from including explanatory variables are negligible. In the provision of this information the time series C_p plot is very different from the C_p plot for regression, for example Figure 1 of Atkinson and Riani (2008), in which the form is that of the series of values for one of the sets of models with the same stochastic structure in Figure 1.

We have found the banded structure of Figure 1 to be typical for the analysis of time series. An example for data on one-day-ahead electricity prices is in Riani and Atkinson (2010b). We trust that this extension of the plot will be as useful as the customary C_p plot in regression.

7 Discussion

In our calculations we used the latest version of the library SSFPACK (Koopman *et al.*, 2008) in conjunction with the Ox programming language of Doornik (2001).

The procedure we have developed applies to a wide class of models. However, it has not escaped our attention that C_p is an aggregate statistic, based on all the data. For regression, Atkinson and Riani (2008) and Riani and Atkinson (2010a) use the forward search (Hadi, 1992; Atkinson and Riani, 2000; Atkinson *et al.*, 2004) to determine how the choice of a regression model using C_p is affected by groups of observations. Although the numerical procedure is more complicated, related methods can be applied to our C_p statistics for time series to illuminate the dependence of model choice on individual observations, breaks in structure and on anomalous patches of observations in the time series.

Appendix A: properties of new AIC

We show that AIC_{d+}, the new version of AIC, is a consistent estimator of the Kullback-Leibler (KL) information for the fitted model. We derive the small sample correction factor to make our new AIC unbiased.

Let $f(y,\theta)$ be the model for the data and $g(y)$ the true density. In the KL distance

$$KL(g, f(\cdot,\hat{\theta})) = \int g(y)\log\frac{g(y)}{f(y,\hat{\theta})}dy = \int g(y)\log g(y)dy - S_n.$$

Given that $\int g(y)\log g(y)dy$ is constant across models, the AIC strategy is in essence to estimate Q_n,

$$Q_n = E_g S_n = E_g \int g(y)\log f(y,\hat{\theta})dy, \tag{35}$$

for each candidate model and then to select the model with the highest estimated Q_n; this is equivalent to searching for the model with the smallest estimated KL distance.

Let the regression model $E(Y) = X\beta$ satisfy Assumption 1. Then, in the standard AIC, we use the mle of β and σ (see, for example Hurvich and Tsai, 1989) and obtain

$$E_g(\hat{Q}_n - S_n) \;=\; \frac{d+1}{n}\frac{n}{n-d-2}. \tag{36}$$

Thus AIC is an asymptotically unbiased estimator of S_n. The bias-corrected AIC can be written as

$$AIC_c = -2L(\hat{\beta};\hat{\sigma}^2;y) + 2(d+1)n/(n-d-2).$$

When we estimate σ^2 by $s^2_{d^+}$ from the full model (5) we obtain

$$E_g(\hat{Q}_n - S_n) = -\frac{1}{2n}E(C_p + n - 2d) + \frac{1}{2}E_g\left[\frac{\sigma^2}{s^2_{d^+}}\left\{(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})/(n\sigma^2)+1\right\}\right].$$

Taking the various expectations under Assumption 1 on both models yields

$$E_g(\hat{Q}_n - S_n) = -\frac{1}{2n}\left\{2\frac{d^+ - d}{n-d^+-2} + n - d - \frac{n-d^+}{n-d^+-2}(n+d)\right\}.$$

After boring calculations we find

$$E_g(\hat{Q}_n - S_n) = \frac{(n-d^+)(d+1)}{n(n-d^+-2)}.$$

Our new AIC is also an asymptotically unbiased estimator of $Q_n = E(S_n)$. In addition, if we want a more precise penalty for the loglikelihood we obtain

$$AIC^+_d(u) = -2L(\hat{\beta};s^2;y) + 2(d+1)(n-d^+)/(n-d^+-2),$$

which leads to (10).

8 Appendix B: proofs

For any $n \times m$ matrix $B_{n\times m}$ we define the norms $||B||_1$ and $||B||_2$ by

$$||B||_1 = \max_{1\le i\le n} \sum_{j=1}^{m} |B_{ij}| \text{ and } ||B||_2^2 = tr(B'B) = \sum_{i=1}^{n}\sum_{j=1}^{m} B_{ij}^2 \tag{37}$$

and note that

$$||AB||_1 \le ||A||_1||B||_1 \qquad \text{and} \qquad ||B||_1 \le ||B||_2 \le n^{1/2}||B||_1, \tag{38}$$

$$||C'AD||_1 \le ||C'||_1||D||_1||A||_1 \le ||C||_2||D||_2||A||_1, \tag{39}$$

$$tr\{A\} \le n||A||_1, \tag{40}$$

and if A is symmetric then

$$||A||_1 \le \max_i |\lambda_i(A)|. \tag{41}$$

We collect some technical results about the coefficients. We call an $n \times n$ matrix A exponentially decreasing if $|A_{ij}| \le c\rho^{|i-j|}$ for some $\rho < 1$

Lemma 1. *Under Assumption 2 it holds that*

$$u_t = \frac{B_q(L)}{A_p(L)}\varepsilon_t = \sum_{n=0}^{\infty} \eta_n \varepsilon_{t-n},\ \eta_0 = 1 \tag{42}$$

$$\varepsilon_t = \frac{A_p(L)}{B_q(L)}u_t = \sum_{n=0}^{\infty} \xi_n u_{t-n},\ \xi_0 = 1. \tag{43}$$

Then $Var(u_t) \ge \sigma^2$ and equality holds only for $u_t = \varepsilon_t$. Moreover

$$\max(\mathsf{D}^m\xi_n(\theta), \mathsf{D}^m\eta_n(\theta)) \le c\rho^n, m = 0, 1, 2. \tag{44}$$

It follows that Ω and its $n \times n$ derivatives $\dot{\Omega}_s = \mathsf{D}_{\theta_s}\Omega$ and $\ddot{\Omega}_{sk} = \mathsf{D}^2_{\theta_s\theta_k}\Omega$ are exponentially decreasing and therefore have bounded 1-norm, and furthermore

$$||\Omega(\theta) - \Omega(\theta_0)||_1 \le c|\theta - \theta_0|, \tag{45}$$

$$||\dot{\Omega}_s(\theta) - \dot{\Omega}_s(\theta_0)||_1 \le c|\theta - \theta_0|, \tag{46}$$

$$||\ddot{\Omega}_{sk}(\theta) - \ddot{\Omega}_{sk}(\theta_0)||_1 \le c|\theta - \theta_0|. \tag{47}$$

The eigenvalues of $\Omega(\theta)$ and $\Omega(\theta)^{-1}$ are bounded away from zero and infinity uniformly in $\theta \in \mathcal{D}$ and n, so that

$$||\Omega^{-1}(\theta) - \Omega^{-1}(\theta_0)||_1 \le c|\theta - \theta_0|, \tag{48}$$

$$c_1\Omega_0 \le \Omega \le c_2\Omega_0. \tag{49}$$

Proof of Lemma 1. *Proof of (42), (43), and (44):*

Condition (14) shows that uniformly for $\theta \in \mathcal{D}$, the power series

$$\sum_{n=0}^{\infty} \eta_n z^n = \frac{B_q(z)}{A_p(z)}, \ \sum_{n=0}^{\infty} \xi_n z^n = \frac{A_p(z)}{B_q(z)},$$
$$\sum_{n=0}^{\infty} \mathsf{D}_{\phi_s} \eta_n z^n = -\frac{B_q(z)}{A_p(z)^2} z^s, \ \sum_{n=0}^{\infty} \mathsf{D}_{\phi_s} \xi_n z^n = \frac{z^s}{B_q(z)},$$
$$\sum_{n=0}^{\infty} \mathsf{D}_{\psi_s} \eta_n z^n = -\frac{z^s}{A_p(z)}, \ \sum_{n=0}^{\infty} \mathsf{D}_{\psi_s} \xi_n z^n = -\frac{A_p(z)}{B_q(z)^2} z^s,$$

$$\sum_{n=0}^{\infty} \mathsf{D}^2_{\phi_s \phi_k} \eta_n z^n = 2\frac{B_q(z)}{A_p(z)^3} z^{s+k}, \sum_{n=0}^{\infty} \mathsf{D}^2_{\phi_s \psi_k} \eta_n z^n = -\frac{1}{A_p(z)^2} z^{s+k}, \sum_{n=0}^{\infty} \mathsf{D}^2_{\psi_s \psi_k} \eta_n z^n = 0,$$
$$\sum_{n=0}^{\infty} \mathsf{D}^2_{\psi_s \psi_k} \xi_n z^n = 2\frac{A_q(z)}{B_p(z)^3} z^{s+k}, \sum_{n=0}^{\infty} \mathsf{D}^2_{\phi_s \psi_k} \xi_n z^n = -\frac{1}{B_q(z)^2} z^{s+k}; \sum_{n=0}^{\infty} \mathsf{D}^2_{\phi_s \phi_k} \xi_n z^n = 0,$$

are convergent for $|z| < \rho^{-1}$, and hence the coefficients are bounded by $c\rho^t$ for some $\rho < 1$, see Yao and Brockwell (2006, page 867). Therefore the representations (42) and (43) hold and it follows that $Var(u_t) = \sigma^2 \sum_{n=0}^{\infty} \eta_n^2 \geq \sigma^2 \eta_0^2 = \sigma^2$ where equality holds only if $\eta_n = 0, n \geq 1$, that is, $u_t = \varepsilon_t$.

Proof of (45), (46), and (47): We then find $\Omega_{ij}(\theta) = \sigma^{-2} \sum_{n=0}^{\infty} \eta_n(\theta)\eta_{n+|i-j|}(\theta)$ is bounded by $\sum_{n=0}^{\infty} c\rho^n \rho^{n+|i-j|} \leq c\rho^{|i-j|}$, so that Ω is bounded in 1-norm:

$$||\Omega||_1 = \max_i \sum_j |\Omega_{ij}| \leq \max_i \sum_j c\rho^{|i-j|} \leq c.$$

The same result holds for the derivatives, so they are all exponentially decreasing and therefore bounded in 1-norm. Moreover by a Taylor's expansion

$$\Omega_{ij}(\theta) - \Omega_{ij}(\theta_0) = \sum_{s=1}^{p+q} (\theta_s - \theta_{0s}) \dot{\Omega}_{s,ij}(\theta_*),$$

which is bounded by

$$c \sum_{s=1}^{p+q} |\theta_s - \theta_{0s}| \rho^{|i-j|} \leq c|\theta - \theta_0| \rho^{|i-j|},$$

so that

$$||\Omega(\theta) - \Omega(\theta_0)||_1 \leq c \max_i \sum_j |\sum_{s=1}^{p+q} |\theta_s - \theta_{0s}| \rho^{|i-j|} \leq c|\theta - \theta_0|.$$

This shows (45), and the same argument works for the derivatives in (46) and (47).

Proof of (48) and (49): We next need the result, see Hannan and Kavalieris (1984, p. 539), that any eigenvalue, $\lambda(\Omega)$, of the $n \times n$ matrix Ω is bounded between two constants $0 < c(\theta) \leq \lambda(\Omega_n(\theta)) \leq C(\theta) < \infty$ independently of n, and Assumption 2 shows that $0 < c \leq \lambda(\Omega_n(\theta)) \leq C < \infty$, so the bound is uniform in $\theta \in \mathcal{D}$. Finally this holds for Ω^{-1} because $\lambda(\Omega^{-1}) = 1/\lambda(\Omega)$, and (48) follows from

$$\begin{aligned} ||\Omega(\theta)^{-1} - \Omega(\theta_0)^{-1}||_1 &= ||\Omega(\theta)^{-1}(\Omega(\theta) - \Omega(\theta_0)\Omega(\theta_0)^{-1}||_1 \\ &\leq ||\Omega(\theta)^{-1}||_1 ||\Omega(\theta) - \Omega(\theta_0)||_1 ||\Omega(\theta_0)^{-1}||_1. \end{aligned}$$

Finally the uniform bound on the eigenvalues implies (49). □

Proof of Theorem 2. The normalized log likelihood, using $\Omega = \Omega_n(\theta)$, is

$$\ell_n(\sigma^2, \theta, \beta) = \log \sigma^2 + n^{-1} \log |\Omega| + \sigma^{-2} n^{-1} (y - X\beta)' \Omega^{-1} (y - X\beta)$$

Minimizing over β gives $\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$, and the profile loglikelihood

$$\begin{aligned} &\ell_n(\sigma^2, \theta, \hat{\beta}(\theta)) \\ &= \log \sigma^2 + n^{-1} \log |\Omega| + \sigma^{-2} n^{-1} u' \Omega^{-1} u - \sigma^{-2} n^{-1} u' \Omega^{-1} X (X'\Omega^{-1}X)^{-1} X'\Omega^{-1} u \\ &= \ell_n(\sigma^2, \theta) - \sigma^{-2} n^{-1} u' \Omega^{-1} X (X'\Omega^{-1}X)^{-1} X'\Omega^{-1} u = \ell_n(\sigma^2, \theta) - \sigma^{-2} n^{-1} u' A u, \end{aligned}$$

where $\ell_n(\sigma^2, \theta)$ is the loglikelihood function analysed by Yao and Brockwell (2006) in the model without regressors. They prove that in the model without regressors the maximum likelihood estimators of σ^2 and θ, obtained by minimizing $\ell_n(\sigma^2, \theta)$ are consistent. The maximum likelihood estimators of σ^2 and θ in the model with regressors are found by minimizing $\ell_n(\sigma^2, \theta, \hat{\beta}(\theta))$, but the same result holds in this case because the difference tends to zero almost surely, that is, $n^{-1} u' A u \stackrel{a.s.}{\to} 0$. To see this, we apply the inequality (49), and find that uniformly for $\theta \in \mathcal{D}$:

$$A = \Omega^{-1} X (X'\Omega^{-1}X)^{-1} X'\Omega^{-1} \leq c\Omega_0^{-1} X (X'\Omega_0^{-1}X)^{-1} X'\Omega_0^{-1} = A_0. \qquad (50)$$

Then $u'Au \leq u'A_0u$, which is distributed as $\sigma_0^2 \chi^2(d)$ so that $-n^{-1}u'Au \stackrel{a.s.}{\to} 0$ and therefore $(\hat{\theta}, \hat{\sigma}^2) \stackrel{a.s.}{\to} (\theta_0, \sigma_0^2)$.

We thus obtain

$$\hat{\kappa} = n^{-1/2} (X'\Omega_0^{-1}X)^{1/2} (X'\hat{\Omega}^{-1}X)^{-1} X'\hat{\Omega}^{-1} u$$

and find from (49) that

$$\hat{\kappa}'\hat{\kappa} = n^{-1} u' \hat{\Omega}^{-1} X (X'\hat{\Omega}^{-1}X)^{-1} (X'\Omega_0^{-1}X) (X'\hat{\Omega}^{-1}X)^{-1} X'\hat{\Omega}^{-1} u \leq c u' A_0 u \stackrel{a.s.}{\to} 0.$$

Finally

$$\hat{\kappa}'\hat{\kappa} = (\hat{\beta}-\beta_0)'n^{-1}X'\Omega_0^{-1}X(\hat{\beta}-\beta_0) \geq \lambda_{\min}(n^{-1}X'\Omega_0^{-1}X)(\hat{\beta}-\beta_0)'(\hat{\beta}-\beta_0)$$

which shows that if $\lambda_{\min}(n^{-1}X'\Omega_0^{-1}X) \geq c > 0$ then $\hat{\beta} \overset{a.s.}{\to} \beta_0$. □

Proof of Theorem 3. We find the normalized scores

$$\begin{aligned}
S_{n\sigma^2} &= \frac{1}{2}n^{-1/2}\sigma_0^{-4}(u'\Omega_0^{-1}u - n\sigma_0^2),\\
S_{n\theta_s} &= \frac{1}{2}n^{-1/2}tr\{\Omega_0^{-1}\dot{\Omega}_{0s}\Omega_0^{-1}(\sigma_0^{-2}uu' - \Omega_0)\},\\
S_{n\kappa} &= \sigma_0^{-2}u'\Omega_0^{-1}X(X'\Omega_0^{-1}X)^{-1/2}.
\end{aligned}$$

The observed information per observation is found from the negative second derivative of the loglikelihood function, $u = Y - X\beta$,

$$\begin{aligned}
I_{n\theta_s\theta_k} &= \frac{1}{2}n^{-1}tr\{\Omega^{-1}\dot{\Omega}_k\Omega^{-1}\dot{\Omega}_s\} \qquad (51)\\
&+ \frac{1}{2}n^{-1}tr\{(\Omega^{-1}(\ddot{\Omega}_{sk} - 2\dot{\Omega}_k\Omega^{-1}\dot{\Omega}_s)\Omega^{-1}(\Omega - \sigma^{-2}uu')\},
\end{aligned}$$

$$I_{n\sigma^2\sigma^2} = -\frac{1}{2}\sigma^{-4} + n^{-1}\sigma^{-6}u'\Omega^{-1}u, \qquad (52)$$

$$I_{n\theta_s\sigma^2} = \frac{1}{2}n^{-1}\sigma^{-4}u'\Omega^{-1}\dot{\Omega}_s\Omega^{-1}u, \qquad (53)$$

$$I_{n\kappa\kappa} = \sigma^{-2}I_d, \qquad (54)$$

$$I_{n\theta\kappa} = n^{-1/2}\sigma^{-2}u'\Omega^{-1}\dot{\Omega}_s\Omega^{-1}X(X'\Omega_0^{-1}X)^{-1/2}, \qquad (55)$$

$$I_{n\sigma^2\kappa} = n^{-1/2}\sigma^{-4}u'\Omega^{-1}X(X'\Omega_0^{-1}X)^{-1/2}. \qquad (56)$$

The expected information per observation for $\lambda = \lambda_0$ is

$$E\begin{pmatrix} I_{n\sigma^2\sigma^2} & I_{n\sigma^2\theta_k} & I_{n\sigma^2\kappa} \\ I_{n\theta_s\sigma^2} & I_{n\theta_s\theta_k} & I_{n\theta_s\kappa} \\ I_{n\kappa\sigma^2} & I_{n\kappa\theta_k} & I_{n\kappa\kappa} \end{pmatrix} = \qquad (57)$$

$$\begin{pmatrix} \frac{1}{2}\sigma_0^{-4} & \frac{1}{2n}\sigma_0^{-2}tr\{\Omega_0^{-1}\dot{\Omega}_{0s}\} & 0 \\ \frac{1}{2n}\sigma_0^{-2}tr\{\Omega_0^{-1}\dot{\Omega}_s\} & \frac{1}{2n}tr\{\Omega_0^{-1}\dot{\Omega}_{0k}\Omega_0^{-1}\dot{\Omega}_{0s}\} & 0 \\ 0 & 0 & \sigma_0^{-2}I_d \end{pmatrix}, \qquad (58)$$

and it follows from the formula

$$Cov(\sigma_0^{-2}u'Au, \sigma_0^{-2}u'Bu) = tr(B\Omega_0A'\Omega_0) + tr(B\Omega_0A\Omega_0) \text{ for } u \sim N_n(0,\Omega_0) \qquad (59)$$

that the variance of the score $S_{n\lambda}$ is (58). This expression has the same limit as the block diagonal matrix

$$\{\frac{1}{2}\sigma_0^{-4}, (\frac{1}{2n}tr\{\Omega_0^{-1}\dot{\Omega}_{0k}\Omega_0^{-1}\dot{\Omega}_{0s}\})_{s,k=1}^{p+q}, \sigma_0^{-2}I_d\},$$

because Theorem 9 shows that

$$\frac{1}{2n}\sigma_0^{-2}tr\{\Omega_0^{-1}\dot{\Omega}_{0s}\} = \frac{1}{2n}\sigma_0^{-2}\mathsf{D}_{\theta_s}\log|\Omega(\theta_0)| = O(n^{-1}).$$

Proof of (17): We need to show that the main term

$$\frac{1}{2n}tr\{\Omega_0^{-1}\dot{\Omega}_{0k}\Omega_0^{-1}\dot{\Omega}_{0s}\}$$

converges to a limit which we call Σ_{0sk}. We notice that the score and expected information for (θ, σ^2) are the same as for the model without regressors analysed by Yao and Brockwell (2006). They prove asymptotic normality of the maximum likelihood estimator in this model, and find a nice representation of the limiting variance in terms of two AR Gaussian processes generated by $A_p(L)$ and $B_q(L)$, see Yao and Brockwell (2006), Theorem 2 and Hannan (1973) Theorem 2. It was part of their proof to show that the expected information converges, so the result follows.

Proof of (18): Next we analyse the terms of the information matrix $I_{n\lambda\lambda}(\lambda)$, see (51-56). It is seen that all components of $I_{n\lambda\lambda}(\lambda) - I_{n\lambda\lambda}(\lambda_0)$, except for some trivial factors, have one of the forms

$$\begin{aligned}
n^{-1}tr(A - A_0) \text{ for } A &= \Omega^{-1}\dot{\Omega}_k\Omega^{-1}\dot{\Omega}_s \text{ or } \Omega^{-1}(\ddot{\Omega}_{sk} - 2\dot{\Omega}_k\Omega^{-1}\dot{\Omega}_s),\\
n^{-1}u'\Omega_0^{-1/2}(A - A_0)\Omega_0^{-1/2}u \text{ for } A &= \Omega_0^{1/2}\Omega^{-1}(\ddot{\Omega}_{sk} - 2\dot{\Omega}_k\Omega^{-1}\dot{\Omega}_s)\Omega^{-1}\Omega_0^{1/2}\\
&\quad \text{or } \Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2} \text{ or } \Omega_0^{1/2}\Omega^{-1}\dot{\Omega}_s\Omega^{-1}\Omega_0^{1/2},\\
n^{-1/2}u'\Omega_0^{-1/2}A\Omega_0^{-1/2}X(X'\Omega_0^{-1}X)^{-1/2} \text{ for } A &= \Omega_0^{1/2}\Omega^{-1}\dot{\Omega}_s\Omega^{-1}\Omega_0^{1/2} \text{ or } \Omega_0^{1/2}\Omega^{-1}\Omega_0^{1/2},
\end{aligned}$$

where

$$u = Y - X\beta = Y - X\beta_0 - X(\beta - \beta_0) = u_0 + n^{1/2}X(X'\Omega_0^{-1}X)^{-1/2}\kappa.$$

For all such matrices A it follows from Lemma 1 that, for $|\lambda - \lambda_0| \leq \varepsilon_n$,

$$||A - A_0||_1 \leq c|\theta - \theta_0| \leq c\varepsilon_n,$$

and we can evaluate

$$\begin{aligned}
n^{-1}u'\Omega_0^{-1}u &= n^{-1}(u_0 + n^{1/2}X(X'\Omega_0^{-1}X)^{-1/2}\kappa)'\Omega_0^{-1}(u_0 + n^{1/2}X(X'\Omega_0^{-1}X)^{-1/2}\kappa)\\
&\leq 2(n^{-1}u'\Omega_0^{-1}u + \kappa'\kappa).
\end{aligned}$$

Moreover we have from inequality (40)

$$n^{-1}|tr(A - A_0)| \leq ||A - A_0||_1 = O(\varepsilon_n),$$

and from inequality (39) with $C = D = \Omega_0^{-1/2}u$:

$$n^{-1}|u'\Omega_0^{-1/2}(A - A_0)\Omega_0^{-1/2}u| \leq n^{-1}u'\Omega_0^{-1}u||A - A_0||_1 = O_P(\varepsilon_n)$$

and finally from (39) with $C' = u'\Omega_0^{-1/2}$ and $D = \Omega_0^{-1/2}X(X'\Omega_0^{-1}X)^{-1/2}$, for which $D'D = I_d$, we find

$$\begin{aligned} n^{-1/2}||u'\Omega_0^{-1/2}(A - A_0)\Omega_0^{-1/2}X(X'\Omega_0^{-1}X)^{-1/2}||_1 \quad &= \\ \leq (n^{-1}u'\Omega_0^{-1}u)^{1/2}||I_d||_2||A - A_0||_1 = O_P(\varepsilon_n). & \end{aligned}$$

□

Proof of Corollary 4. Note that $n^{1/2}\hat{\kappa} = (X'\Omega_0^{-1}X)^{1/2}(\hat{\beta} - \beta_0)$; we continue with the parameter κ. Again $\lambda = (\sigma^2, \theta, \kappa)$ denotes the $m = 1 + p + q + d$ parameters. In order to find the asymptotic distribution, we consider the Taylor expansion of the score function around $\lambda = \lambda_0$

$$n^{1/2}(\hat{\lambda} - \lambda_0) = (-n^{-1}\mathsf{D}^2 \log L(\lambda_*))^{-1}n^{-1/2}\mathsf{D} \log L(\lambda_0), \tag{60}$$

where the notation λ_* indicates that row s is evaluated at an intermediate point $\lambda_*^{(s)}$ which satisfies $|\lambda_*^{(s)} - \lambda_{0s}| \leq |\hat{\lambda} - \lambda_0|$. The result now follows from Theorem 3 because $n^{-1/2}\mathsf{D} \log L(\lambda_0) \xrightarrow{D} N_{1+p+q+d}(0, \{\frac{1}{2}\sigma_0^{-4}, \Sigma_0, \sigma_0^{-2}I_d\})$ and from (18) get that

$$-n^{-1}\mathsf{D}^2 \log L(\lambda_*) \xrightarrow{P} \{\frac{1}{2}\sigma_0^{-4}, \Sigma_0, \sigma_0^{-2}I_d\}.$$

□

Proof of Corollary 5. We let $\tau = (\sigma^2 - \sigma_0^2, \theta' - \theta_0')'$ and find the equation

$$I_{n\tau\tau}(\lambda_*)n^{1/2}\hat{\tau} + I_{n\tau\kappa}(\lambda_*)n^{1/2}\hat{\kappa} = S_{n\tau},$$

which shows that when $I_{n\tau\kappa}(\lambda_*) = O_P(n^{-1/2})$ we have

$$n^{1/2}\hat{\tau} = E(I_{n\tau\tau})^{-1}S_{n\tau} + O_P(n^{-1/2}).$$

The same result holds in the model $\beta = A\xi$ with estimator $\hat{\tau}_*$ and, as a consequence,

$$n^{1/2}(\hat{\tau} - \hat{\tau}_*) = O_P(n^{-1/2}),$$

which proves the result. □

Proof of Theorem 6. We write $\hat{u}_* = y - X\hat{\beta}_*$ and expand $\hat{u}_*'\Omega(\hat{\theta})^{-1}\hat{u}_*$ as a function of $\hat{\theta}$ around $\hat{\theta}_*$ and find, using $\tilde{\theta}$ for an intermediate point for which $|\tilde{\theta} - \hat{\theta}_*| \leq |\hat{\theta} - \hat{\theta}_*|$, that

$$\hat{u}_*'(\hat{\Omega}^{-1} - \hat{\Omega}_*^{-1})\hat{u}_* = -\sum_{s=1}^{p+q}(\hat{\theta}_s - \hat{\theta}_{s*})\hat{u}_*'\hat{\Omega}_*^{-1}\widehat{\dot{\Omega}}_{*s}\hat{\Omega}_*^{-1}\hat{u}_* + R_n, \tag{61}$$

where

$$R_n = \frac{1}{2}\sum_{s=1}^{p+q}\sum_{k=1}^{p+q}(\hat{\theta}_s - \hat{\theta}_{s*})(\hat{\theta}_k - \hat{\theta}_{k*})\hat{u}_*'[\tilde{\Omega}^{-1}\widetilde{\ddot{\Omega}}_{sk}\tilde{\Omega}^{-1} - \tilde{\Omega}^{-1}\widetilde{\dot{\Omega}}_s\tilde{\Omega}^{-1}\widetilde{\dot{\Omega}}_k\tilde{\Omega}^{-1}]\hat{u}_*. \tag{62}$$

Here $\widetilde{\dot{\Omega}}_s = \mathsf{D}_{\theta_s}\Omega(\tilde{\theta})$ and $\widetilde{\ddot{\Omega}}_{sk} = \mathsf{D}^2_{\theta_s\theta_k}\Omega(\tilde{\theta})$. The main term of (61) can be simplified using the first order condition at $\hat{\theta}_*$, that is, $\partial\ell_n/\theta_s|_{\theta=\hat{\theta}_*} = 0$:

$$\hat{u}_*'\hat{\Omega}_*^{-1}\widehat{\dot{\Omega}}_{*s}\hat{\Omega}_*^{-1}\hat{u}_* = \hat{\sigma}_*^2\mathsf{D}_{\theta_s}\log|\Omega(\theta)|, s = 1, \ldots, p+q.$$

This implies that

$$\sum_{s=1}^{p+q}(\hat{\theta}_s - \hat{\theta}_{s*})\hat{u}_*'\hat{\Omega}_*^{-1}\widehat{\dot{\Omega}}_{*s}\hat{\Omega}_*^{-1}\hat{u}_* = \sum_{s=1}^{p+q}(\hat{\theta}_s - \hat{\theta}_{s*})\hat{\sigma}_*^2\mathsf{D}_{\theta_s}\log|\hat{\Omega}(\hat{\theta}_*)|.$$

It follows from Theorem 9 that $\mathsf{D}_{\theta_s}\log|\hat{\Omega}(\hat{\theta}_*)| \xrightarrow{P} \mathsf{D}_{\theta_s}F_t^c(\theta_0)$, so that the main term is $O_P(\hat{\theta} - \hat{\theta}_*) = O_P(n^{-1})$.

The remainder term (62) is also $O_P(n^{-1})$ because it is bounded by

$$\begin{aligned}
&c|\hat{\theta} - \hat{\theta}_*|^2[|\hat{u}_*'\tilde{\Omega}^{-1}\widetilde{\ddot{\Omega}}_{sk}\tilde{\Omega}^{-1}\hat{u}_*| + |\hat{u}_*'\tilde{\Omega}^{-1}\widetilde{\dot{\Omega}}_s\tilde{\Omega}^{-1}\widetilde{\dot{\Omega}}_k\tilde{\Omega}^{-1}\hat{u}_*|] \\
&\leq c|\hat{\theta} - \hat{\theta}_*|^2\hat{u}_*'\hat{u}_*(||\tilde{\Omega}^{-1}\widetilde{\ddot{\Omega}}_{sk}\tilde{\Omega}^{-1}||_1 + ||\tilde{\Omega}^{-1}\widetilde{\dot{\Omega}}_s\tilde{\Omega}^{-1}\widetilde{\dot{\Omega}}_k\tilde{\Omega}^{-1}||_1)
\end{aligned}$$

The first factor is $O_P(n^{-2})$ and $||\tilde{\Omega}^{-1}\widetilde{\ddot{\Omega}}_{sk}\tilde{\Omega}^{-1}||_1 + ||\tilde{\Omega}^{-1}\widetilde{\dot{\Omega}}_s\tilde{\Omega}^{-1}\widetilde{\dot{\Omega}}_k\tilde{\Omega}^{-1}||_1 \leq c$ because of Lemma 1. Finally $\hat{u}_*'\hat{u}_* = O_P(n)$. □

Proof of Theorem 7. The likelihood ratio test is

$$-2\log LR(\beta = A\xi) = n(\log\hat{\sigma}_*^2 - \log\hat{\sigma}^2) + \log|\Omega_n(\hat{\theta}_*)| - \log|\Omega_n(\hat{\theta})|$$

where $\hat{\sigma}_*^2 = n^{-1}(y - X\hat{\beta}_*)'\hat{\Omega}_*^{-1}(y - X\hat{\beta}_*)$ and $\hat{\sigma}^2 = n^{-1}(y - X\hat{\beta})'\hat{\Omega}^{-1}(y - X\hat{\beta})$ both converge to σ^2 and $(\hat{\sigma}^2 - \sigma_0^2, \hat{\sigma}_*^2 - \sigma_0^2) = O_P(n^{-1/2})$, but it follows from (5) that $\hat{\sigma}^2 - \hat{\sigma}_*^2 = O(n^{-1})$. A Taylor expansion gives

$$\begin{aligned}
n(\log\hat{\sigma}_*^2 - \log\hat{\sigma}^2) &= n\log(\hat{\sigma}_*^2\hat{\sigma}^{-2}) \qquad (63)\\
&= n(\hat{\sigma}_*^2\hat{\sigma}^{-2} - 1) + nO_P((\hat{\sigma}_*^2 - \hat{\sigma}^2)^2) = n(\hat{\sigma}_*^2\hat{\sigma}^{-2} - 1) + O_P(n^{-1}).
\end{aligned}$$

Similarly, $\hat{\theta}_* - \hat{\theta} = O_P(n^{-1})$ and $F_t^c(\theta) = \sum_{t=1}^{\infty} \log f_t^c(\theta)$, see Theorem 9, so that

$$\log|\Omega_n(\hat{\theta}_*)| - \log|\Omega_n(\hat{\theta})| = F_t^c(\hat{\theta}_*) - F_t^c(\hat{\theta}) + O_P(\rho^{-2n}) = O_P(n^{-1}). \quad (64)$$

Thus from (63) and (64) we find

$$-2\log LR(\beta = A\xi) = n\hat{\sigma}^{-2}(\hat{\sigma}_*^2 - \hat{\sigma}^2) + O_P(n^{-1}).$$

The first order condition for β is $(y - X\hat{\beta})'\hat{\Omega}^{-1}X = 0$ so that

$$y - X\hat{\beta}_* = (y - X\hat{\beta}) + (X\hat{\beta} - X\hat{\beta}_*)$$

is an orthogonal decomposition with respect to $\hat{\Omega}^{-1}$, that is, $(y - X\hat{\beta})'\hat{\Omega}^{-1}(X\hat{\beta} - X\hat{\beta}_*) = 0$, and

$$(y-X\hat{\beta}_*)'\hat{\Omega}^{-1}(y-X\hat{\beta}_*) = (y-X\hat{\beta})'\hat{\Omega}^{-1}(y-X\hat{\beta}) + (\hat{\beta}_*-\hat{\beta})'X'\hat{\Omega}^{-1}X(\hat{\beta}_*-\hat{\beta}).$$

Therefore

$$\begin{aligned} n(\hat{\sigma}_*^2 - \hat{\sigma}^2) &= (y - X\hat{\beta}_*)'\hat{\Omega}_*^{-1}(y - X\hat{\beta}_*) - (y - X\hat{\beta})'\hat{\Omega}^{-1}(y - X\hat{\beta}) \\ &= (y - X\hat{\beta}_*)'\hat{\Omega}^{-1}(y - X\hat{\beta}_*) - \\ &\quad (y - X\hat{\beta})'\hat{\Omega}^{-1}(y - X\hat{\beta}) + (y - X\hat{\beta}_*)'(\hat{\Omega}_*^{-1} - \hat{\Omega}^{-1})(y - X\hat{\beta}_*) \\ &= (\hat{\beta}_* - \hat{\beta})'X'\hat{\Omega}^{-1}X(\hat{\beta}_* - \hat{\beta}) + (y - X\hat{\beta}_*)'(\hat{\Omega}^{-1} - \hat{\Omega}_*^{-1})(y - X\hat{\beta}_*). \end{aligned}$$

The first term measures the deviation between the estimators using the variance estimate from the larger of the two models, and the second term tends to zero because

$$R_n = (y - X\hat{\beta}_*)'(\hat{\Omega}^{-1} - \hat{\Omega}_*^{-1})(y - X\hat{\beta}_*) = O_P(n^{-1}),$$

see Lemma 6. This proves (7). □

Proof of Theorem 8. For $\hat{\beta} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y$ we find

$$(y-X\hat{\beta})'\hat{\Omega}^{-1}(y-X\hat{\beta}) = y'X_\perp(X_\perp'\hat{\Omega}X_\perp)^{-1}X_\perp'y = u'X_\perp(X_\perp'\Omega_0X_\perp)^{-1}X_\perp'u + R_n,$$

where

$$R_n = u'X_\perp[(X_\perp'\hat{\Omega}X_\perp)^{-1} - (X_\perp'\Omega_0X_\perp)^{-1}]X_\perp'u.$$

The first term is distributed as $\sigma^2\chi^2(n-d)$ because $X_\perp'u \sim N_n(0, X_\perp'\Omega_0X_\perp)$.

For the remainder term we find, using inequality (39), for $C' = u'X_\perp(X_\perp'\Omega_0X_\perp)^{-1}X_\perp'\Omega_0^{1/2}$, $D = \hat{\Omega}^{1/2}X_\perp(X_\perp'\hat{\Omega}X_\perp)^{-1}X_\perp'u$, and $A = \Omega_0^{-1/2}(\hat{\Omega} - \Omega_0)\Omega_0^{-1/2}$ that

$$\begin{aligned} |R_n| &= |C'AD| \\ &\leq (u'X_\perp(X_\perp'\Omega_0X_\perp)^{-1}X_\perp'u)^{1/2}(u'X_\perp(X_\perp'\hat{\Omega}X_\perp)^{-1}X_\perp'u)^{1/2}||\Omega_0^{-1/2}(\hat{\Omega} - \Omega_0)\hat{\Omega}^{-1/2}||_1 \\ &\leq c(u'X_\perp(X_\perp'\Omega_0X_\perp)^{-1}X_\perp'u)|\hat{\theta} - \theta_0| = O_P(n^{1/2}). \end{aligned}$$

□

Proof of Theorem 9. The process $u_t = \sum_{n=0}^{\infty} \eta_n \varepsilon_{t-n}$ is a linear invertible process with exponentially decreasing coefficients, so that

$$\varepsilon_t = \frac{A_p(L)}{B_q(L)} u_t = \sum_{n=0}^{\infty} \xi_n u_{t-n}, \text{ or } u_t = -\sum_{n=1}^{\infty} \xi_n u_{t-n} + \varepsilon_t$$

where $|\xi_n| \leq c\rho^n$. It follows that the prediction variance for Gaussian variables satisfies

$$Var_{t-1}(y_t) = Var_{t-1}(\sum_{n=t}^{\infty} \xi_n u_{t-n} | \mathcal{F}_{t-1}) + \sigma^2 \leq \sigma^2 + E(\sum_{n=t}^{\infty} \xi_n u_{t-n})^2.$$

Then

$$E(\sum_{n=t}^{\infty} \xi_n u_{t-n})^2 = \sigma^2 \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \xi_{n+t} \xi_{m+t} \gamma(n-m) \leq c\rho^{2t} \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \rho^{n+m} |\gamma(n-m)| \leq c\rho^{2t}.$$

Thus $0 \leq \log f_t^c(\theta) \leq f_t^c(\theta) - 1 = \sigma^{-2}(Var_{t-1}(y_t) - \sigma^2) \leq c\rho^{2t}$, and therefore $\log f_t^c(\theta)$ is summable uniformly in θ and hence the limit is continuous

$$0 \leq \log |\Omega_n(\theta)| = \sum_{t=1}^{n} \log f_t^c(\theta) \to \sum_{t=1}^{\infty} \log f_t^c(\theta).$$

It follows from Brockwell and Davis (1991, p 394-395) that also

$$|\frac{\partial}{\partial \theta_s} f_t^c(\theta)| \leq c\rho^t,$$

for $\theta \in \mathcal{D}$. Because $f_t^c(\theta) \geq 1$ the same argument shows that $\partial \log f_t^c(\theta)/\partial\theta_s$ is uniformly dominated by $c\rho^t$ and hence the sum $\sum_{t=1}^{\infty} \partial \log f_t^c(\theta)/\partial\theta_s$ exists as a continuous function. This shows that

$$\frac{\partial \log |\Omega_n(\theta)|}{\partial \theta_s} = \sum_{t=1}^{n} \frac{\partial \log f_t^c(\theta)}{\partial \theta_s} \to \sum_{t=1}^{\infty} f_t^c(\theta)^{-1} \frac{\partial f_t^c(\theta)}{\partial \theta_s}$$

is finite and the uniform convergence shows continuity. □

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**, 716–723.

Anderson, B. D. O. and Moore, J. B. (1979). *Optimal Filtering*. Prentice Hall, Englewood Cliffs, N. J.

Atkinson, A. C. and Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer–Verlag, New York.

Atkinson, A. C. and Riani, M. (2008). A robust and diagnostic information criterion for selecting regression models. *Journal of the Japanese Statistical Society*, **38**, 3–14.

Atkinson, A. C., Riani, M., and Cerioli, A. (2004). *Exploring Multivariate Data with the Forward Search*. Springer–Verlag, New York.

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multi-Model Inference. A Practical Information-Theoretic Approach*. Springer, New York.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge University Press, Cambridge.

Claeskens, G., Croux, C., and Van Kerckhoven, J. (2007). Prediction-focused model selection for autoregressive models. *Australian and New Zealand Journal of Statistics*, **49**, 359–379.

Davison, A. C. (2003). *Statistical Models*. Cambridge University Press, Cambridge.

de Jong, P. (1991). The diffuse Kalman filter. *The Annals of Statistics*, **19**, 1073–1083.

Doornik, J. A. (2001). *Ox 3.0: Object-oriented matrix programming language (4th ed.)*. Timberlake Consultants Press, London.

Durbin, J. and Koopman, S. J. (2012). *Time Series Analysis by State Space Models, 2nd edition*. Oxford University Press, Oxford.

Francke, M., Koopman, S. J., and De Vos, A. F. (2010). Likelihood functions for state space models with diffuse initial conditions. *Journal of Time Series Analysis*, **31**, 407–414.

Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and Cp in multivariate linear regression. *Biometrika*, **84**, 707–716.

Gilmour, S. G. (1996). The interpretation of Mallows's C_p-statistic. *The Statistician*, **45**, 49–56.

Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, Series B*, **54**, 761–771.

Hannan, E. (1973). The asymptotic theory of linear time-series models. *Journal of Applied Probability*, **10**, 130–145.

Hannan, E. and Kavalieris, L. (1984). Multivariate linear time series models. *Advances in Applied Probability*, **16**, 492–561.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, Cambridge.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference and Prediction, 2nd edition*. Springer, New York.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

Konishi, S. and Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer-Verlag, New York.

Koopman, S. J., Shephard, N., and Doornik, J. A. (2008). *Statistical Algorithms for Models in State Space Form: SsfPack 3.0*. Timberlake Consultants Press, London.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

McQuarrie, A. D. R. and Tsai, C. L. (1998). *Regression and Time Series Model Selection*. World Scientific, Singapore.

Riani, M. and Atkinson, A. C. (2010a). Robust model selection with flexible trimming. *Computational Statistics and Data Analysis*, **54**, 3300–3312. doi: 10.1016/j.csda.2010.03.007.

Riani, M. and Atkinson, A. C. (2010b). The selection of time series models perhaps with regressors. Technical Report LSERR147, London School of Economics, Department of Statistics, London WC2A 2AE, UK.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Shi, P. and Tsai, C.-H. (2004). A joint regression variable and autoregressive order selection criterion. *Journal of Time Series Analysis*, **25**, 923–941.

So, M. K. P., Chen, C. W. S., and Liu, F.-C. (2006). Best subset selection of autoregressive models with exogenous variables and generalized autoregressive conditional heteroscedasticity errors. *Journal of the Royal Statistical Society Series C*, **55**, 923–941.

Tong, H. (2001). A personal journey through time series in Biometrika. *Biometrika*, **88**, 195–218.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S (4th Edition)*. Springer-Verlag, New York.

Wang, H., Li, R., and Tsai, C.-L. (2007). Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, **69**, 63–78.

Yao, Q. and Brockwell, P. J. (2006). Gaussian maximum likelihood estimation for ARMA models I: time series. *Journal of Time Series Analysis*, **27**, 857–875.